Bona Fide Predictions of Protein Secondary Structure Using Transparent Analyses of Multiple Sequence Alignments

Steven A. Benner,* Gina Cannarozzi, Dietlind Gerloff, Marcel Turcotte, and Gareth Chelvanayagam

Department of Chemistry, University of Florida, Gainesville, Florida 32611-7200

Received July 5, 1996 (Revised Manuscript Received September 19, 1997)

Contents

* Author to whom all correspondence should be addressed.

S0009-2665(94)00469-3 CCC: $28.00

I. Introduction

By any measure, the 1990s is the decade of the genome. Sequences of the chromosomes of two eubacteria (*Haemophilus influenzae* and *Mycoplasma genitalium*),[1,2] one archaebacterium (*Methanococcus jannaschii*),[3] and one eukaryote (*Saccharomyces cerevisiae*, bakers' yeast)[4] have appeared, and several other completed microbial genomes will be announced while this review is in press. Before the decade is out, the genome of the worm *Caenorhabiditis elegans* will be added to this collection,[5] as will perhaps several dozen further genomes of microorganisms. The genomes for a plant and man will be complete soon thereafter. These will supplement sequences from dozens of other organisms whose genomes are not being comprehensively sequenced, but are being studied in laboratories around the world.

Organic chemistry has always been driven by the discovery of new natural products, elucidation of their structures, and exploration of their behaviors. The genome sequence database provides an enormous new collection of natural products to study. These display every behavior important in chemistry: conformation, supramolecular organization, combinatorial assembly, and catalysis are just a few. Every branch of chemistry will therefore be advanced as the chemistry of the natural products in the genomic databases is explored in the postgenomic world. Further, through an evolutionary picture of how these molecules arose, an understanding of biological function will come from the chemical structure of molecules, allowing natural history to join coherently the physical and life sciences.

This review focuses on the first of the "chemical" behaviors displayed by these natural products: conformation. Conformation defines how a molecule exists in three dimensions when it has achieved a (presumably global) energy minimum after searching through all rotational degrees of freedom. In protein chemistry, conformation is referred to variously as the fold, secondary and tertiary structure, or sometimes simply "structure". From conformation comes many other physical and physiological properties of proteins. The review is directed toward the nonspecialist, a chemist or biochemist who knows something about structural biology in general and wishes to understand more about how conformational analysis for proteins is developing in light of genomic data.

Steven Benner received a B.S.-M.S. at Yale University in Molecular Biophysics and Biochemistry, and a Ph.D. in Chemistry at Harvard University under the joint sponsorship of Frank H. Westheimer and R. B. Woodward. After two years as a Junior Fellow of the Harvard Society of Fellows, he became an Assistant Professor in the Department of Chemistry at Harvard University. In 1985, he became Professor of Bioorganic Chemistry at the Swiss Federal Institute of Technology, and in 1995, Professor of Chemistry, Anatomy, and Cell Biology at the University of Florida. His research covers the chemistry, biology, and evolution of proteins and nucleic acids.

Gina M. Cannarozzi received her B.S. in chemistry from the University of Central Florida and her M.S. and Ph.D. in physical chemistry from the University of California, San Diego, studying deuterium relaxation methodology with Professor Regitze Vold. After investigating questions of membrane asymmetry while a postdoctoral researcher with Philippe Devaux at the Institut de Biologie Physico-Chimique in Paris, France, she joined the laboratory of Professor Steven Benner at the University of Florida in 1996 as a postdoctoral associate to work on protein structure prediction. Her research interests include the relationships between protein structure and function and their implications for evolution.

A. Why Is the Protein Conformation Problem Hard?

The "protein structure prediction problem" is the classical unsolved problem in protein chemistry. It

Dietlind Gerloff received a Diplom in Chemistry from the Swiss Federal Institute of Technology and a Ph.D. at the same institution under the direction of Professor Steven A. Benner. She then joined the laboratory of Professor Fred E. Cohen at the University of California, San Francisco, where she is presently. Her postdoctoral research is supported by Fellowships of the Swiss National Science Foundation (1995) and the Leukemia Society of America (1996 to present). Dietlind Gerloff's research interests are in approaches toward protein structure and function which involve biochemistry, bioinformatics, and molecular evolution. She was selected to present predictions at the second Critical Assessment of Protein Structure Prediction (CASP2) meeting in Asilomar, in December 1997.

Marcel Turcotte received his B.Sc., M.Sc., and Ph.D. in computer science from the *Université de Montréal*. His thesis advisors were Guy Lapalme (*informatique et recherche opérationnelle*) and Robert Cedergren (*biochimie*). He joined the laboratory of Professor Steven Benner at the University of Florida in 1995 as a postdoctoral associate while receiving a fellowship from *Fond pour la Formation de Chercheurs et l'Aide à la Recherche du Québec*. His research interests are macromolecular structures, evolution, and programming paradigms.

Gareth Chelvanayagam received his B.Sc. and Ph.D. in Computer Science from the University of Western Australia, with the research for his doctoral thesis being done at the EMBL in Heidelberg, Germany, under the supervision of Patrick Argos. After two years as a postdoctoral associate with Professors Steven Benner and Gaston Gonnet at the ETH in Zurich, he then joined the group of Simon Easteal, at the John Curtin School of Medical Research at the Australian National University where he is now a Research Fellow, supported by the Australian Research Council. His research interests include protein structure, function, and evolution.

is difficult for many reasons, all of which are important as we consider how it might be solved.

First, proteins are big, especially when compared with the molecules that have long been the focus of conformational analysis in organic chemistry. Proteins typically contain 100−1000 amino acids, or 1000−20000 atoms. Every peptide unit in the polypeptide chain has two rotational degrees of freedom (Figure 1), assuming that the amide bond itself is planar and lies exclusively in the "trans" conformation. One degree of rotational freedom is around the bond joining the carbonyl carbon and the α carbon of the amino acid. The second is around the bond joining the α carbon and the nitrogen (Figure 1). These are often known as the ψ and ϕ angles.[6] Flexibility in the side chains adds additional rotational degrees of freedom to the molecule. Together, these make the conformational energy surfaces associated with protein sequences enormous, especially when compared with those of molecules traditionally studied by chemists. It is difficult to search a surface this large, and considerable effort has been devoted to developing ways to do so.[7−9]

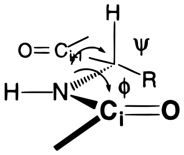


Figure 1. The two rotational degrees of freedom in an amino acid, designated by the dihedral angles ϕ and ψ, give a peptide chain its flexibility.

Second, understanding conformation is difficult in proteins because it is difficult in *all* molecules, even molecules much smaller than a typical protein. The protein conformation problem is intricately connected with questions that lie at the heart of physical chemistry: How do we describe the interaction of two molecules with each other? How do we describe the interaction of ensembles of molecules? Answers for these questions for simpler systems have not yet been found, although impressive progress has been made in this area in the past few years.[10−14] There is today no method, automated or manual, parameterized or *ab initio*, that precisely predicts the conformation of any organic molecule in solution. Conformation is especially poorly understood in strongly interacting solvents such as water, the environment where most globular proteins exist physiologically.

If this were not sufficient, evolutionary issues unique to biological molecules such as proteins suggest that predicting conformation should be especially difficult.[15] Natural selection seeks biomolecules that contribute to survival, mate selection, and reproduction in their host organism. A protein with extreme conformational stability is rarely desired by natural selection, if only because a cell living in a changing environment is continually degrading proteins to reuse their constituent amino acids to make new proteins. Thus, natural selection typically seeks a protein that unfolds at a temperature only a few degrees higher than the physiological temperature for an organism.[15]

If a protein obeys all of the "rules" of folding, excessive conformational stability is possible, however.[15] The conformational stability of proteins from thermophiles, the ease with which point mutation can increase conformational stability, and the insolubility of a typical peptide (remembering that precipitation, where a peptide interacts with other peptides rather than with solvent, is a "folding" process) is evidence for this.

Thus, selective pressures create proteins that are conformationally unstable relative to the stability that could be achieved if a protein were to exploit all of the stabilizing interactions available to a typical polypeptide chain.[15] This implies that natural proteins violate folding "rules" to achieve a desired level of instability. This, in turn, implies that even if the chemist learns the "rules" that confer conformational stability on molecules, and can apply them to large molecules such as proteins, natural protein sequences will deceive the chemist attempting to apply these rules to predict their conformations.

B. The Focus of This Review. Evolution-Based Structure Prediction

The fact that natural proteins are the products of divergent evolution creates opportunities as well as problems when developing tools for predicting conformation from sequence.[10,15−21] Proteins in the modern world almost never come alone. Rather, Nature presents sets of *homologous* proteins (proteins related by common ancestry) performing analogous functions in different organisms. As long as their genes have continuously performed a function since they divergently evolved, homologous proteins retain their overall conformation. Indeed, this conformation can be retained long after sequence similarity has been lost in statistical noise.[22,23] This is quite different from the conformational behavior of a "homologous series" of compound in organic chemistry, a set of compounds differing in the length of a chain, where conformation between members need have no similarities. Natural selection acting on homologous proteins divergently evolving under functional constraints is the reason for this difference.

For this reason, a set of sequences of proteins within a family of homologous proteins contains more information about conformation than a single sequence or a single member of the family.[15,21,24−29] The set of protein sequences is a set of different molecular structures that achieve (more or less) the same conformation.

This review begins with this fact and will focus on methods that build models for the conformation of a protein family from a set of homologous protein sequences. These are by necessity *consensus models* of protein conformation, those that describe features of conformation that are conserved among all of the members of the protein family. We will focus in particular on *secondary structure*, the local conformation of a protein. The α helix and the *β* strand are the standard elements of secondary structure.

Second, this review focuses on ways of building consensus models of conformation that exploit an increased understanding of how functioning proteins suffer point mutation, insertion, and deletion during divergent evolution. This insight has come from the revolution in genomics. Advances have come in many sectors, including Web sites that provide access to sequences,[30] improved tools for comparing the sequences of proteins related by common ancestry,[31−33] new schemes for classifying organisms,[34] new ideas relating the *in vitro* behavior of proteins to their physiological function *in vivo*,[35] and experiments that have reconstructed in the laboratory ancient biological macromolecules from extinct organisms to permit experimental evaluation of evolutionary models.[36−39] From these studies have come improved models describing the divergent evolution of proteins at the molecular level. These models permit an approach to predict protein conformation that is "transparent" to the user.

The concept of transparency in structure prediction has an analogy in conventional conformational analysis in chemistry. In small molecules, conformation can be studied by using a computationally intensive tool based on quantum mechanics or molecular mechanics. Or it can be studied by hand. The latter approach is very familiar to students of organic chemistry, who build ball-and-stick models of molecules, inspect these by eye for steric interactions (for example), and use the process to understand molecular conformation. The quantum mechanical calculation is arguably more fundamental than an analysis that involves a physical model of a molecule and human intervention. Yet the ball-and-stick model is ultimately more satisfying to the chemist, who feels that it yields more of an explanation of molecular behavior. Further, the history of chemistry has shown that transparent approaches for analyzing conformation (as well as other properties of organic molecules) have been more powerful as a way to generate new ideas than purely computational ones.

Computationally intensive approaches to model protein conformation are also available, generally based on molecular mechanics tools and a variety of force fields. These are reviewed elsewhere,[40−44] and will not be discussed here. Rather, we will focus on the "ball-and-stick" approach for modeling protein structure, an approach made possible by our improved understanding of the molecular details of evolution at the level of the protein molecule. These allow the user to understand why a prediction is made, how it might fail, and why it works (when it works). Such transparent analyses of protein conformation also allow a more rational design of prediction heuristics.[45]

The third focus of this review is a recent trend toward testing methods for predicting protein conformation using *bona fide* predictions, those made and announced before an experimental conformation has been determined.[46−49] The term *bona fide* (meaning "genuine" [50] without pejorative overtones) reflects the widespread practice in the field of using the word "prediction" to denote "retrodiction",[51] where a tool is used to build a model of the conformation of a protein whose structure was known at the time that the tool was applied. Certainly in the late 1980s and early 1990s, a typical title in the field that reported a method for "prediction" of protein secondary structure at (for example) 70% accuracy meant a method that was developed and tested by retrodiction.

As discussed below, *bona fide* predictions are an integral tool of any scientific analysis of molecular conformation. *Bona fide* predictions have proven to be important to the field for sociological reasons as well, however, and these require some comment. Many experimental biochemists have come to find unpersuasive any evaluation of structure prediction methods tested retrodictively.[52] Over several decades, methods that performed well when tested retrodictively were found to perform worse when tested on new proteins.[53] This was especially the case in structure prediction "contests", where knowledge of the conformation of the target structure was explicitly withheld from those making predictions. With a notable exception of the first such contest,[54] results were largely disappointing in comparison with expectations based on retrodictions of protein conformation using the same methods.[55−57]

As discussed below, this phenomenon can arise in many ways, many of which are innocuous. However, by the early 1990s, many experimental biochemists came to believe, correctly or incorrectly, that procedures for predicting features of protein conformation from sequence data will *always* perform substantially worse than they perform in retrodictive tests. In many circles, it came to be feared that they might never work at a level to make them useful.[58]

As a result, a relatively small number of *bona fide* predictions that later proved, in the opinion of independent judges, to have been "remarkably accurate",[59−62] has transformed the outlook of the field in a way that would have been impossible by any other approach. The resulting impact has been especially important to scientists not directly involved in the structure prediction field.

The review will combine these three elements: evolutionary analysis, *bona fide* prediction, and transparency. The review attempts to be comprehensive up until January 1, 1996. Further, during the period of time that elapsed since this review was first prepared, a second "Critical Assessment of Structure Prediction" (CASP)[49] project was completed. The results of this project are included where they meet the scope of the review. The review therefore covers all *bona fide* predictions made to that date that relied on transparent prediction methods applied to a set of homologous sequences. We have erred on the side of inclusiveness. Many predictors are now combining transparent and nontransparent methods in their analysis; we have attempted to include these as well.

The review is set in four parts.

(a) First, approaches to evaluate the quality of predictions of secondary structure will be discussed. Predictions made by prediction tools must be evaluated to learn whether the tools are being improved, of course. The evaluation problem itself raises important scientific issues, however, and it is essential to sort these out before we attempt to evaluate the output of transparent prediction methods.

(b) Next, the introduction of evolutionary ideas into the field of protein structure prediction will be traced. This will require an abbreviated discussion of classical prediction methods that incorporate no evolutionary models, starting in the 1970s. We cannot duplicate the many excellent reviews of the field; an especially valuable collection of reviews to the end of the 1980s was edited by Fasman.[63] This review will instead present classical methods in a way that allows the reader to understand how they have contributed to evolution-based methods that are the focus of this review, and how their procedures and results differ from evolution-based methods.

(c) We will then show how the availability of massive amounts of sequence data emerging from genome projects has yielded an improved understanding of how sequences evolve subject to "functional constraints", that is, how amino acid substitutions, insertions, and deletions take place in real proteins that must fold and perform functions in real organisms. We will show how improved models of molecular evolution have guided the development of tools for secondary structure prediction in proteins.

(d) Last, we will illustrate how transparent methods based on evolutionary analysis have been tested through *bona fide* prediction by bringing together examples where evolutionary analysis has been used to predict the secondary structure of proteins.

Finally, the average chemist or biochemist is not as computer literate as the average informaticist working in the field of structure prediction. The past few years has seen a proliferation of computer programs and tools, some commercial, some available on the Web, some simply reported in journal articles. We present a selective compilation of these in a "Glossary" and "Appendix" at the end of this review, chosen to include those that will be the most interesting to the nonspecialist. The reader should recognize that this list is out of date even as it is being prepared. But it is a start.

II. Evaluating Predictions. How Do We Recognize Progress?

We must first address an issue that appears technical, but actually contains an important unsolved scientific problem: What tools should be used to evaluate prediction methods? As it turns out, this apparently simple question contains many levels of complexity.

Consider a simple task, to evaluate a secondary structure prediction made for a single protein. Let us assume that the secondary structure prediction assigns to segments of the protein sequence one of three secondary structural types: α helix, β strand, and coil (a conformation of the backbone that is neither a helix nor a strand). Such a prediction could, it seems, be evaluated by comparing the predicted secondary structure, residue-by-residue, with an experimental secondary structure. Comparing the experimental secondary structure, residue-by-residue, with the predicted secondary structure should yield a "three-state residue-by-residue score", sometimes known as "Q_3", the percentage of residues correctly assigned to one of three states (helix, strand, or neither). Q_3 would seem to be an objective measure of the quality of a prediction.[64,65]

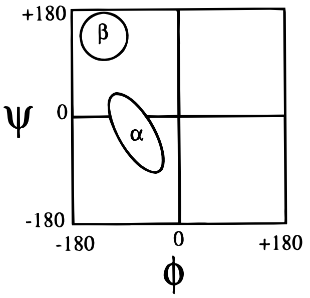


Figure 2. Ramachandran plot showing the (arbitrary) boundaries between values of ϕ and ψ that indicate β strands (β) α helices (α), and coils (the remainder of the diagram).

A. Scoring Problem 1: The Definition of Secondary Structural Units (Helix and Strand) Is Subjective

More detailed consideration shows that the Q_3 score is subjective in several important ways. First, there is no such thing as an "experimental secondary structure". The experimental data produced by X-ray crystallography (or by NMR) are a set of coordinates for atoms in a protein. Secondary structure is an abstraction of these coordinates. Converting the primary experimental data into an assignment of secondary structure requires definitions (What is an "α helix" or a "β strand"?). These definitions are themselves subjective.

Consider three different ways to define secondary structure in terms of coordinates. In one, secondary structure is defined by the two dihedral angles in the polypeptide backbone that undergo free rotation (Figure 1). The ϕ and ψ angles of amino acids in natural proteins are conveniently presented on a Ramachandran diagram (Figure 2).[6] In natural proteins, certain combinations of dihedral angles are more populated than others, and certain regions of the Ramachandran diagram are defined as holding amino acids in "α helices", and others hold "β strands". Amino acids with dihedral angles lying outside of these regions are defined as "coil". Thus, arbitrarily placed regions on the Ramachandran diagram defines "three states" that might be used to score a secondary structure prediction, where the dihedral angles of individual amino acids are extracted from crystallographic coordinates.

This definition of secondary structure is inadequate for evaluating a prediction, however. A single amino acid may have ϕ and ψ angles squarely in the middle of the region of the Ramachandran diagram that defines an α helix, but still not be a part of a helix. An α helix is stabilized by hydrogen bonding between backbone atoms coming from amino acids four positions removed in a chain. In a β sheet, the N—H and C═O groups of the backbone participate in hydrogen bonds to C═O and N—H groups in other strands still more distant in the sequence. Whether or not a particular residue is part of a helix or strand depends, therefore, in part on the conformation of *other* amino acids in the polypeptide chain, and their ability to form hydrogen bonds to the residue in question.

Instead, helices and sheets might be defined by the presence of these hydrogen bonds. For idealized data, this is a powerful tool for assigning secondary structure. Indeed, a more detailed description of secondary structural types, including 3_{10} helices, π helices, and various types of bends and turns can be obtained by a careful analysis of hydrogen bonding patterns.[66] Crystal structures of proteins generally do not have the resolution needed to see hydrogens, however, meaning that the positions of hydrogens and hydrogen bonding patterns must be inferred from the positions of heavy atoms. Further, the dynamic behavior of protein structures, together with the occurrence of distorted secondary structural elements, means that not all helices and strands evident to a human eye inspecting a crystal structure are identified using programs that search for hydrogen bonding. In the discussion below, we will see specific examples where β hairpins and α helices are missed by the automated assignment program, even though these structures are evident by visual inspection of the structure and are conserved throughout the evolution of a protein family.

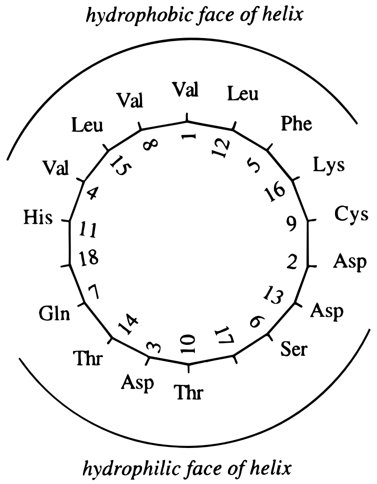


Figure 3. Schiffer—Edmundson helical wheel showing the position of hydrophobic and hydrophilic amino acids in the C-terminal α helix of adenylate kinase. This particular relative orientation of the side chains can be used as a definition of a helix.[53]

A third way to define secondary structure relies on the relative orientation of the side chains in a polypeptide chain. In an α helix, the side chain of an amino acid protrudes from a cylinder approximately 1.5 Å along the helix axis, and ~100° around the helix axis, relative to the side chain of the amino acid preceding it in the chain. This relationship is graphically described by a Schiffer—Edmundson helical wheel,[67] which is a projection of a helix down its long axis to view the relative disposition in space of the amino acid side chains (Figure 3). The side chains in a β strand alternate above and below the sheet. As the side chains of all amino acids (except, of course, glycine) contain heavy atoms, the relative orientation of side chains is easily seen in crystal structures with satisfactory resolution.

As "secondary structure" is an abstraction of the human intellect, no one of these definitions is more correct than another. What is clear, however, is that the different definitions need not yield the same experimental secondary structure assignments from the same set of experimental coordinates.[63] The subjective nature of experimental secondary struc-

```
                                                      Sequences
        sub       0    1    1    2    2    3    3    4    4    5    5    6    6    7    7
       family     5    0    5    0    5    0    5    0    5    0    5    0    5    0    5
PI3K-1    d   AEGYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETTGERGDFPGTYVEYIGRK human
PI3K-2    d   AEGYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEAKPEEIGWLNGYNETTGERGDFPGTYVEYIGRK ox

                                               Experimental Structures
PI3K Expt 1      EEEEEeeEE      EEEEEE  EEEEEHHHHHHHH   3333333333EEEEEE    EEEEE3333EEEEE  ref.72
PI3K Expt 2       EEEE                  EEEE                HHHH  EEEEEE    EEEEEEHHHEEEEEE ref.71

Hypothetical 1    HHHH                  HHHH                HHHH  HHHHHH    HHHHHH3333EEEEE
Hypothetical 2    EEEE  EE      EEEEEE       HHHHHHHH   3333333333EEEEEE    EEEEE
```

Figure 4. Alignment of sequences and experimentally assigned secondary structures[71,72] for two Src homology 3 (SH3) domains. Key: H, α helix; E, β strand; e, weakly assigned β strand; 3, 3_{10} helix. The sequences of the two proteins differ by a single amino acid (at position 47). The proteins give the visual appearance of having the same overall fold. Yet the sequences have the same assignment at only 73% of the positions, if "e" is treated as a coil and a 3_{10} helix is assumed to match equally an α helix or a strand. The segment scores are either 50% or 70%, depending on how a 3_{10} helix is treated.

ture assignments was quantitated by Colloc'h *et al.*,[68] who compared three automated tools (DSSP,[66] P-curve,[69] and Define[70]) that assign secondary structure to crystallographic data. The P-curve program identifies regularities along the helicoidal axis in a polypeptide in assigning secondary structure, DSSP considers hydrogen-bonding patterns, while Define measures distances between C-α atoms. Colloc'h *et al.* asked what percentage of the residues in the protein received the same secondary structural assignment by all three methods applied to the very same coordinate data. The answer was a strikingly low 63%.[68] This number is especially relevant considering that current secondary structure prediction heuristics are routinely yielding three-state Q_3 scores of approximately 70% (see below).

One specific example of this problem is shown in Figure 4. The figure shows two published *experimental* secondary structures determined for the same protein, the src homology 3 (SH3) domains of the phosphatidyl-inositol-3-kinase (PI3K) from ox and man.[71,72] Both experimental structures were determined by NMR spectroscopy. Except for a single amino acid, the sequences of the two proteins are identical. By eye, the folds are indistinguishable. Yet the two experimental secondary structures (Figure 4), taken directly from the papers reporting those structures, agree at only 73% of the positions.

This means that if experimental structure 1 in Figure 4 were to be judged using experimental structure 2 as a reference, the resulting Q_3 would be only 73%, even though the target and reference secondary structural assignments being compared are experimental, are obtained on proteins with essentially the same sequence, and the conformations of the two proteins are essentially identical.

This is bad enough. Still worse is the fact that we can construct an entirely hypothetical secondary structural model (the line labeled "Hypothetical 1" in Figure 4) that completely obliterates the fact that the core fold of the SH3 domain is built from β strands; Hypothetical 1 models the protein instead as largely helical. This hypothetical model is quite wrong. But it *also* gives a Q_3 score of 73%.

An alternative approach is to score segment-by-segment instead of residue-by-residue.[73-75] This approach would eliminate the Hypothetical model 1 for the SH3 domain (Figure 4) as a plausible prediction, and therefore represents an advance. Even so, the two experimental structures in Figure 4 agree in their assignments for only 50% or 70% of the segments (depending on whether one counts a 3_{10} helix as an equivalent of an α helix; see below).

In the context of the modern literature, a "prediction" for one structure based on the experimental secondary structural model from the other would be "wrong", again despite the fact that the conformations of the two proteins are identical within any plausible level of resolution. To make the point completely, Hypothetical model 2 (Figure 4) has the same segment score, but does not represent either structure accurately.

Both of these examples and the more comprehensive study by Colloc'h *et al.*[68] make the general statement: *One cannot score a secondary structure prediction objectively if the experimental secondary structure that serves as a reference is subjective.* At the very least, the subjectivity in assigning secondary structure to crystallographic data sets an upper limit on the Q_3 score that a prediction can have. The lack of objectivity associated with defining secondary structure from experimental coordinates alone makes it impossible for the residue-by-residue score of a secondary structure assignment to be routinely higher than ~75−85%.[75] Higher scores obtained by predictions judged against an experimental assignment generated by one method imply lower scores when judged against scoring obtained by another.

One solution to this problem is to distinguish between "serious" and "not serious" mistakes.[73] Different methods, while assigning secondary structure differently to the same set of coordinates, generally do not disagree in their assignments in any way that is significant to the overall perception of the fold. Thus, a segment that is assigned as a helix by one method is virtually never assigned as a strand by another, and a segment that is assigned as a strand by one method is virtually never assigned as a helix by another. Rather, the different assignment tools disagree about the precise beginning and end of helices and strands, the assignments given to distorted secondary structural elements, and the assignments of short elements, often on the surface of the fold. Each of these differences changes the score; none change the overall perception of the fold.

This suggests that mistakes (in this discussion, the word "error" is reserved for experimental error) made by a prediction fall into two classes, "serious" and "not

```
                                    Sequences
        sub     0    1    1    2    2    3    3    4    4    5    5    6    6    7    7
     family     5    0    5    0    5    0    5    0    5    0    5    0    5    0    5
src       a GGVTTFVALYDYESRTETDLSFKKGERLQIVNNTRKVDVR---------EGDWWLAHSLSTGQTGYIPSNYVAPSD
Fyn       a --VTLFVALYDYEARTEDDLSFHKGEKFQILNSS---------------EGDWWEARSLTTGETGYIPSNYVAPVD
H PLC     b TFKCAVKALFDYKAQREDELTFIKSAIIQNVEKQ---------------EGGWWRGDYGG-KKQLWFPSNYVEEMV
C spec    c TGKELVLALYDYQEKSPREVTMKKGDILTLLNST---------------NKDWWKVEV--NDRQGFVPAAYVKKLD
PI3K-1    d AEGYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETTGERGDFPGTYVEYIGRK
PI3K-2    d AEGYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEAKPEEIGWLNGYNETTGERGDFPGTYVEYIGRK

                              Experimental Structures
src       a     EEEEeEEE      EEEE    EEEEE   ---------------   EEEEEEE    EEEE3333EEEE
Fyn-1     a     EEEE                  EEEEEE  ---------------   EEEEEE     EEEE    EEE
Fyn-2     a     EEEE                  EEE     ---------------   EEEEEE     EEEE    EEE
H PLC     b    EEEEE EEE         EEE EEEEE EEE---------------   EEEEEE   EEEEEE     EEE
C spec    c     EEEE                  EEEEEE  ---------------   EEEEEE--  EEEEE 3333EEE
PI3K-1    d     EEEE                  EEEE                HHHH  EEEEEE    EEEEEE3333EEEEE
PI3K-2    d    EEEEEeeEE      EEEEEE  EEEEEHHHHHHHH   3333333333EEEEEE    EEEEE3333EEEEE

Ideal pred.     EEEE EEE       EEEE   EEEEE                     EEEEEE     EEEE     EEE
```

Figure 5. Alignment of sequences and experimentally assigned secondary structures[71,72,76,85−87] for a family of distantly related Src homology 3 (SH3) domains. Different SH3 domains are specified using standard nomenclature; see references. Dashes in the sequence are deleted amino acids. Key: H, α helix; E, β strand, 3; 3_{10} helix.

serious", the first being a difference between the prediction and the experimental assignments of secondary structure where all methods agree, the second being a difference between the prediction and the secondary structural assignment where the methods disagree. While a prediction must be described by more than a single score to give an accurate view of its success, if a single score *must* be constructed, the most valuable may well be the number of helices mistaken for strands and strands mistaken for helices.

B. Scoring Problem 2: Predictions for a Set of Homologous Proteins Are "Consensus Models"

The evaluation of predictions is still more problematical when the prediction applies to a family of proteins rather than to a single protein. Such a model is a "consensus prediction". Experimental structures are determined for single proteins, not for families of proteins. When building a model for a single protein, one clearly can use an experimental structure of the individual protein as a reference when evaluating the prediction. But what experimental structure should one use when evaluating a prediction for a family of proteins?

Consensus modeling assumes, of course, that homologous proteins have identical conformations.[22,23] This is only true as an approximation, of course, especially for proteins whose sequences have diverged substantially. For example, some 30% of the side chains in a pair of proteins with 40% sequence identity have different orientations.[77] By definition, a consensus model should predict the orientation of the 70% of the residues whose orientation is conserved throughout the protein family, and leaves the remainder unassigned. To evaluate the model generally requires comparing it with a single experimental structure where *all* of the side chain orientations are defined, however. Thus, in a family of proteins that has diverged to 40% sequence identity, a perfect consensus description of side chain orientation cannot have a score higher than 70% when evaluated using a single experimental structure. If one is interested simply in boosting the score, one might assign orientations ("inside" and "outside", for example) randomly to the residues that are unassigned in the consensus model. This would (on average) boost the score to 85%. But this increase in the score would have no particularly interesting scientific meaning.

Secondary structure also diverges during divergent evolution. A consensus model for secondary structure is one that identifies the secondary structural elements that are conserved and leaves unassigned segments of the protein whose secondary structure is not conserved. Again, the consensus model is generally evaluated using a single protein as a reference, where all of the amino acids are assigned to some secondary structural state (helix, strand, or coil). Thus, the regions of the reference protein that correspond to segments in the consensus model that are unassigned will all be scored as "wrong". Again, one might boost the score by randomly assigning secondary structure to these nonconserved regions, again without coherent scientific meaning.[73−75,78]

The SH3 domain can be used again to illustrate these points. Figure 5 shows now a set of aligned sequences of SH3 domains from different "subfamilies". Clearly, the sequence of SH3 domains has diverged substantially, with the gain and loss of some secondary structural elements. Thus, the long helix in the PI3K SH3 domain is not conserved in the family, and a consensus model of secondary structure of the family might not be expected to report it. If that consensus model were evaluated using the PI3K SH3 domain as a reference structure, however, the score would be lower to reflect the "omission" of the nonconserved helix.

These considerations add a layer of complexity to that introduced in earlier discussions of the limitations of three-state scores.[73−75,78] When building a consensus model of secondary structure to be evaluated using a reference structure subjectively assigned to experimental coordinates, it is not possible to resolve the flaws in three-state scores, either residue-by-residue or segment-by-segment, simply by setting the goal lower (for example, to 80%). The three-state score of a perfect consensus prediction can be made arbitrarily low simply by selecting a reference protein that has an arbitrarily large number of noncore segments inserted relative to the core.

The past five years of *bona fide* prediction projects has provided many examples where this has distorted evaluations of predictions. An excellent example is offered by the *bona fide* prediction of phospho-β-D-galactosidase, discussed in greater detail below. The transparent prediction[79] successfully identified every conserved secondary structural element in the core, successfully identified the noncore regions, and generated a correct tertiary structural model for the core, an 8-fold $\alpha-\beta$ barrel. Because the consensus model was scored using a reference protein that had elements of an additional, nonconserved domain interspersed with the core secondary structural units, the Q_3 score for that prediction was only ~65%, both by residue and by segment, a score that might be considered to indicate that little progress has been made in structure prediction in the past 20 years.[80] In reality, the prediction of the core secondary structural units was sufficiently accurate to identify the core fold overall, one of the first times that this has been done in a *bona fide* prediction environment.

Analogous cases discussed below include threonine deaminase and fibrinogen. In each case, Q_3 scores (for example, of 68%) could not be used even as cutoffs to separate models worthy of further examination from those not worthy of further examination without creating artifacts in the evaluation. The reference proteins simply contained too much polypeptide chain that was not part of a core fold.

C. Progress in Evaluating Secondary Structure Predictions

The inadequacy of three-state scores is now widely appreciated, and many groups have produced important new ideas on how to evaluate predictions.[73−75,78] These are increasingly being applied.[81,82] Nevertheless, many papers in the recent literature continue to use small (one to three percentage points is typical) increases in Q_3 scores as evidence for an improvement in a prediction heuristic in the 70−75% range.[83,84]

Without making the effort to reexamine the original data from which these scores are constructed, it is impossible to know whether these increased scores reflect meaningful improvements in the prediction tool. If the improvement in three-state score represents a decrease in the number of strands misassigned as helices, or helices misassigned as strands ("serious mistakes"), then the improved score indicates a more useful heuristic. It is also possible, however, that the score has increased without any useful improvement in the predictions themselves. Future investments in the detailed analysis of protein structure must adopt more sophisticated methods for scoring, so that these investments can pay the highest dividend in information.

Steps have also been taken to improve the tools used to automatically assign secondary structure to experimental coordinates. For example, Frishman and Argos recently reported a tool named "STRIDE" for assigning secondary structure to experimental coordinates.[88] STRIDE uses both hydrogen bonding and main chain dihedral angles as input, parameterizes this information against secondary structures assigned by crystallographers, and optimizes the relative contributions of the two with the specific goal of producing assignments that are in closer agreement with the assignments that crystallographers make. The propensities of amino acid residues with specific ψ and ϕ angles to be part of helices and strands are also considered, so the method depends on the nature of the amino acids involved. While no independent evaluation of the method is presently available, anecdotal experience in these laboratories suggests that the tool improves assignments in regions where they are critical for structure prediction (see below).

Another approach to circumventing the problems associated with scoring is to score only those regions of the core fold that are conserved in the protein family.[78] The disadvantage of this approach is that it normally requires at least two experimental structures within a protein family, preferably themselves quite distant in the evolutionary tree, to identify the elements of the consensus fold. These are not always available, especially for *bona fide* predictions. In the case of the SH3 domain, however, where multiple experimental structures of domains distant in the evolutionary tree are available, this approach is clearly viable (Figure 5). The β strands that define the character of the core of the SH3 domain are conserved. Strands 2 and 3 in the src SH3 domain are assigned in some domains but not in others; thus the ambiguity arises from the subjectivity of strand assignment. This too is evident by looking at several homologous structures. The approach clearly identifies Hypothetical prediction 2 as bad. Further, it defines more precisely an ideal consensus prediction, shown as the last line in Figure 5. Indeed, Figure 5 shows that three or four experimental structures from members of a protein family widely dispersed in an evolutionary tree are sufficient to generate a solid picture of the secondary structural elements of a protein that are important to predict.

In the absence of multiple experimental structures for a protein family, a scoring system must identify noncore regions by inspecting a single experimental structure in light of the multiple alignment itself. Core elements might be defined geometrically; a core element is one where a substantial fraction is buried. Thus, a core strand is one that forms strand−strand interactions, is central to a β sheet, and forms backbone hydrogen-bonding interactions with two other strands on both of its edges. By this definition, a core strand is distinct from an edge strand, which forms backbone hydrogen bonds to only one other strand on only one of its edges. In a number of evaluations discussed below, edge and core strands are distinguished.

A more general definition of a core secondary structural unit focuses on the evolutionary stability of the secondary structural unit. Non-core regions generally suffer multiple insertions and deletions after ~100 point mutations per 100 amino acids.[89] This procedure can be used to rule out some segment of a target peptide sequence as contributing to the core. If a segment is deleted in some homologs (and if the deletion is not a database error), then it is not a core. The procedure was used, for example, to identify noncore regions in the phospho-β-galactosidase structure.

Another method for identifying a core segment of a protein sequence is applicable to any set of sequences containing three sequences or more. In the tool, a pairwise alignment is constructed for each pair of sequences in the set using a dynamic programming tool. Consider for example a set of sequences with three proteins, A, B, and C. A core segment of the multiple alignment is defined as those regions where the alignment of A with B and the alignment of B with C is consistent with the alignment of A with C. This approach is generally useful only in the absence of an experimental structure and needs further experimental support. Thus, it is not yet empirically established that segments that are noncore by this rule are also more likely to suffer insertions and deletions after protracted divergent evolution, or whether they lie predominantly on the surface of a protein. It is clear (see below), that predictions in such segments are difficult to make reliably.

A final method for identifying core elements relates to the reconstructed ancestral sequences for a protein family. In general, the part of the ancestral sequence that is reconstructed with high probability is the "core" of the protein.

Regardless of the definition of the core, the distinction between serious and nonserious mistakes is helpful in determining how well a prediction has done in identifying core secondary structural elements in the absence of more than one structure within the protein family. During divergent evolution, strands are rarely converted to helices and *vice versa*. Rather, short helices and strands, usually not in the core of the folded structure, are distorted or replaced by coils or gaps during divergent evolution, and small numbers of residues are added to or removed from helices and strands at the core. Again, none of these changes change the overall fold. Therefore, a score that focuses closely on mispredictions that confuse strands and helices has proven to be a useful, if incomplete, tool for evaluating consensus predictions.[78,90]

This discussion is especially timely as the best *bona fide* structure predictions (see below) are achieving Q_3 scores in the 70−75% range. As this is also the level of ambiguity in secondary structural assignments and in the divergence of secondary structure commonly found in a prediction dataset, an improved scoring system is needed, and this almost certainly requires a focus on core secondary structural elements.

D. Scoring Predictions in This *Chemical Review*

Recognizing that a scoring problem exists with conventional tools for scoring predictions is the first step toward resolving the problem. Fortunately, the problem is easily understood by those trained in chemistry. Tradition in chemistry has long recognized that molecular structures have complexity, that this complexity is interesting, and that this complexity is not easily abstracted by a single number. When examining a prediction, a chemist is interested in the details of the experimental structure.

With secondary structure, these details are relatively accessible, within the limits noted above. In this review, complete secondary structure predictions are presented, together with one or more experimental assignments of secondary structure. These are accompanied by the sequences of proteins in the family containing the "target" protein, the protein whose conformation is sought. From this detailed presentation, the reader can gain his/her own perception of the prediction by inspection. Commentary is then provided to point out why specific mistakes were made.

E. Scoring Predictions of Secondary Structures in the Future

Few experimental biochemists find a secondary structure prediction useful in itself. Rather, a secondary structure prediction is a starting point for further work. Most important from a structural perspective, a secondary structure model is the starting point for building a model of tertiary structure. This requires assembling the predicted secondary structural elements in three-dimensional space. Alternative uses include detecting long-distance homologs,[91−93] antigenic sites,[94,95] active sites,[15,21,91] defining quaternary structure,[15] or proposing mechanistic hypotheses for how the protein might catalyze a reaction.[96] The ultimate value of tools for predicting secondary structure will be defined by their value in these and other applications.

When assembling a tertiary structural model from a set of predicted secondary structural elements, mistakes that misassign a core helix as a strand or a core strand as a helix will both generally be fatal to an effort to build a tertiary structural model. Misassignment of an element that is not in the core, or that has undergone divergence during divergent evolution, generally will not be. Omission of a secondary structural element is generally fatal when that element is at the core of the folded structure. Omission of a peripheral secondary structural element is generally not. Thus, evaluations that focus on serious mistakes, and that weight mistakes more seriously when they are in the core of the fold, are likely to be more relevant to understanding the value of secondary predictions than those that do not.

To date relatively few predicted models for secondary structure that have been placed in the public domain have been applied. This makes it difficult to do a comprehensive evaluation of prediction methodology using these tests. They are, however, enough to support the comments below, where tertiary structural models built on predicted secondary structural units in a *bona fide* prediction setting are discussed.

III. Background: Classical Structure Prediction

Discussions of conformation in proteins began immediately after the first proteins were sequenced. A daring attempt by Scheraga to predict the conformation of ribonuclease as early as 1960, based on a variety of experimental and theoretical considerations, is especially noteworthy, if only because it illustrates how difficult the problem is.[97] Not until the early 1970s, however, did the search for methods to predict conformation begin in earnest. Work of Anfinsen and others showed that denatured proteins could refold spontaneously,[98] at least in certain cases, providing experimental support for the paradigm that the protein sequence alone determines the conformation of a protein. This paradigm remains dominant today, despite the discovery of chaperonins,[99] evidence that some proteins form metastable structures,[100] and renewed interest in protein folding pathways,[101] all of which suggest that protein folding has a kinetic as well as a thermodynamic component.

A discussion of classical methods is necessary to prepare the reader for a discussion of modern methods. As this Review is intended in part for chemists, biochemists, and students not directly involved in structure prediction research, we provide a summary of these methods. Consistent with the nature of this audience, the summary focuses on the underlying philosophy and strategy of classical approaches, rather than providing a comprehensive review of their technical details. Greater technical exposition is found in many excellent reviews, both those mentioned above and those cited below. Especially helpful is a compendium of reviews edited by Fasman,[63] published in 1989. It remains a timely volume, and the reader is referred to it for a more comprehensive coverage of the classical aspects of the problem. This book also contains a list of earlier reviews on proteins structure prediction.[63]

Most of the heuristics developed during the first three decades of the field attempt to predict protein conformation from a single protein sequence, without embedding that sequence within a family of homologous protein sequences. Approaches of this type for predicting the conformation of a protein sequence are generally classified as either “probabilistic” or “physicochemical”.[53] We will comment on these separately below.

A. Probabilistic Methods for Predicting Secondary Structures

Probabilistic methods tabulate from known crystal structures the propensity of each of the amino acids to form secondary structures of each type. Early work with myoglobin and hemoglobin found, for example, that proline lies more frequently in a coil or a turn than the average amino acid.[102] More comprehensive analyses showed that different amino acids have different propensities for different types of secondary structure. Propensities for individual amino acids to lie in particular secondary structural types can be expressed numerically.[103] These propensities are generally small. Thus, the best “helix-forming” amino acids (Ala and Glu) are only ~50% more likely to lie in a helix than the average amino acid. The worst “helix-forming” amino acids are only ~50% less likely.

Propensities for individual amino acids to adopt particular secondary structures have been used for predicting secondary structure for 25 years. In their simplest form, probabilistic prediction tools assign secondary structure (helices, strands, or neither) to segments of polypeptide chains that are rich in amino acids with propensities for the particular structural type. Often, a model for how proteins fold underlies the assignment tool. The Chou−Fasman method, for example, looks for a nucleation site for a helix, a segment of four amino acids with high propensities to form a helix.[104] The GOR method of Garnier, Osguthorpe, and Robson treats a string of amino acids as a message that is translated by the folding mechanism into another message, a string of conformational states, and applies information theory methods to deduce the “code” for converting one message into the other.[64,105]

Probabilistic methods are well known in the literature.[106] The Chou−Fasman method and the GOR method are probably the most frequently cited and used. The methods are easily automated and are frequently implemented (sometimes incorrectly)[107] in standard computer software packages for protein sequence analysis. This makes secondary structure prediction tools readily available to the nonspecialist. Indeed, in the 1980s, a Chou−Fasman or GOR prediction of secondary structure was routinely reported for new protein sequences.

It is quite difficult to evaluate these, however, as both valid and invalid implementations of various standard methods have been used to make these predictions,[107] and it is difficult to determine which were used to assign secondary structure to any particular sequence. Nevertheless, probabilistic methods have been the subject of many excellent reviews,[63] and their strengths and weaknesses are well known. The most prominent weakness is their underlying strategy of assuming that local conformation (secondary structure) is predominantly determined by local sequence. The tools assign secondary structure to a polypeptide segment by examining a sliding window (generally 1−10 consecutive amino acid residues) and ignoring the influence of the rest of the protein on secondary structure.

Unfortunately, much information shows that long-distance interactions in a protein dominate local sequence in determining local conformation.[108] For example, Kabsch and Sander,[109] Argos,[110] and Presnell and Cohen[111] identified specific pentapeptides and hexapeptides that form a helix in one protein context and a strand in another. This shows convincingly for these sequences that secondary structure is not determined by local sequence and raises the possibility that no probabilistic method fashioned in the classical sense could possibly assign both structures correctly. None of this surprises the chemist, of course; local conformation is frequently influenced by long-distance interactions in many classes of natural products.

This work does not, however, prove that *no* sequences exist that have secondary structures independent of tertiary interactions. Nor does it exclude the possibility that small propensities exist. Indeed, many of the “parsing” tools (see below) used by contemporary prediction methods identify specific sequences that, with high probability, form coils.[91] Further, short (5−15 residue) polypeptide sequences that adopt specific secondary structures in the absence of tertiary interactions can be found.[112,113]

Other difficulties encountered by statistical methods arise from biases in the crystallographic database used to parameterize them. Anecdotally, it has been suggested that probabilistic methods generally perform better on proteins that adopt a class of fold that is well represented in the database upon which the method is parameterized, and poorly on classes of fold that are poorly represented in the same database. Nine folds represent over 30% of the structures contained in the 1994 database ($\alpha-\beta$ doubly wound, the eight-fold barrel analogous to that found in triose phosphate isomerase, split $\alpha-\beta$ sandwich, Greek key immunoglobulin, α up−down, globin, jelly roll, trefoil, and $\alpha-\beta$ roll).[114] In particular, $\alpha-\beta$ proteins with the β sheet buried seem to be predicted better than all β proteins using classical methods.[65,115,116] Buried β sheets are heavily represented in the database.

Inspection of the statistical parameters themselves shows evidence of this bias. For example, the GOR parameter for a coil structure correlates both with the hydrophobicity index[117] and with observed side chain accessibility of the individual amino acids (Figure 6). This correlation presumably reflects the fact that both coils and hydrophilic amino acids are found preferentially on the surface of proteins within the set of proteins used to parameterize the GOR method.[118] Similarly, the strongest predictor of the GOR strand propensity is hydrophobicity and interior position. This is expected given the fact that strands lie preferentially inside the globular structures found in the databases used to parameterize the GOR method. Only the helix parameter lacks a correlation with hydrophobicity. This might be interpreted as reflecting the fact that in the crystallographic database, a majority of the helices lie on the surface of globular folds, with part of their residue side chains pointing out to solvent and part pointing in toward a hydrophobic core. These correlations suggests at least the possibility that the observed propensities reflect in part tertiary structural influences on secondary structure rather than intrinsic propensities of specific side chains to force the backbone to adopt specific ϕ and ψ angles. This does not mean that all propensities can be explained in this way, of course.[112,113,119]

For example, Pro lies (as expected) off of the correlation in Figure 6 for coil parameters; this almost certainly reflects an intrinsic propensity of Pro to be disfavored in helices and strands. Further, the correlation between hydrophilicity and the propensity to form coils may reflect the fact that hydrophilic side chains have functionality able to form hydrogen bonds, which in turn can form hydrogen bonds to the backbone atoms, thereby disrupting helices and strands, which are stabilized by backbone−backbone interactions.

Whatever the true interpretation of the statistical propensities, this discussion illustrates the complexities of the problem, and the potential for systematic errors in predictions made using probabilistic methods. These will become important below when we discuss methods that extend statistical methods using evolutionary analyses.

B. Physicochemical Methods

Physicochemical methods rely on physical and chemical principles to rationalize and predict protein conformation. For example, hydrophobic side chains are more likely to be buried in a protein that folds in water than are hydrophilic side chains,[53] and this fact can be used to predict secondary structure. Lim noted many years ago that a helix might be identified in a polypeptide sequence from a characteristic 3.6-residue periodicity in the placement of hydrophilic and hydrophobic residues.[120] Such periodicity is easily visualized by use of a Schiffer−Edmundson helical wheel (Figure 3). The hydrophobic face of the amphiphilic helix is often found to be buried within the fold.

The notion of amphiphilicity has been generalized to include hydrophobic moments of secondary structural elements.[121] The hydrophobic moment is an analog of the electric dipole moment, except that it measures the asymmetry of the hydrophobicity in a structure rather than the asymmetry of the electrical charge. Thus, a helix with hydrophobic residues on one side and hydrophilic resides on the other has a large “hydrophobic moment” and is expected to be stable at (for example) an interface between oil and water.

Physicochemical methods for predicting secondary structure have also been the subject of excellent reviews.[63] These tools have shown promise when applied to single sequences in some cases but not in others. These are discussed in greater detail below. Further, physicochemical analyses have proven to be important in many evolution-based prediction tools, as they appear to be more readily “averagable” than statistical methods (see below).

In individual cases, failures of physicochemical methods to make correct secondary structure predictions can often be related to violations of “folding rules” by proteins (see above). When such violations are observed, they often offer the biochemist an opportunity to engineer the protein to improve its stability. For example, if a natural protein places a hydrophobic residue on its surface, a glycine in a helix, or an acyclic amino acid at a position in a protein where a proline would fit the backbone configuration,[122,123] a more stable protein can often be obtained by replacing the hydrophobic residue by a hydrophilic residue, the glycine by an alanine, or the flexible residue by a proline. In each case, the mutation makes the sequence obey the folding “rules” better. Examples where improved stability is engineered into a protein via a single amino acid substitution offer additional evidence that natural selection does not seek proteins with maximized stability.[15] Were increased stability a goal of natural selection and achievable by simple point mutation, evolutionary processes would have already introduced the changes made by the protein engineer.

Physicochemical methods of increased sophistication use energy minimization, molecular dynamics, or even quantum mechanical tools. These tools have been reviewed in detail elsewhere.[124−126] Here, the limitations of the methods relate directly to the complexity of the computations involved, the difficulties associated with finding optima on an energy

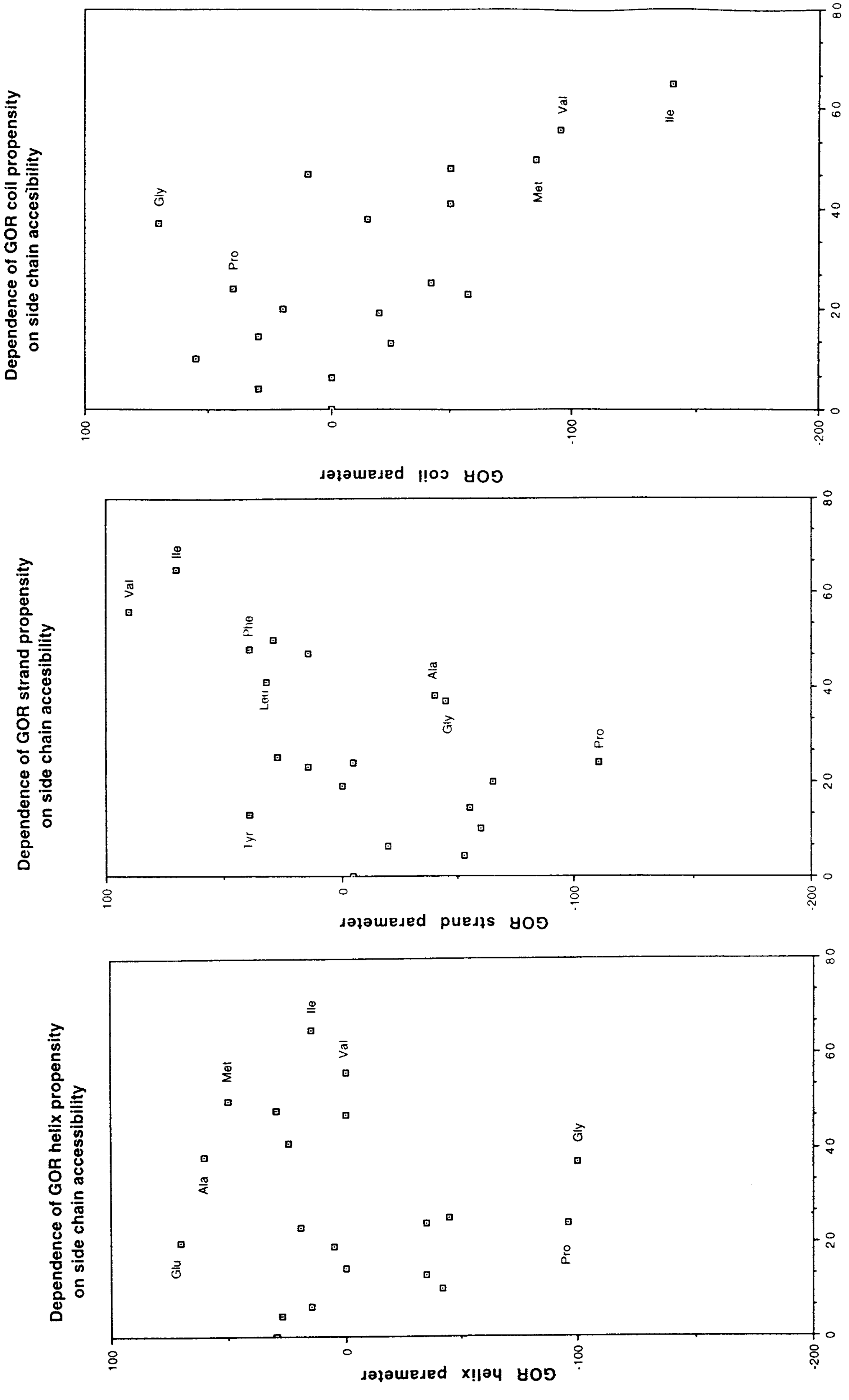


Figure 6. Correlation between GOR parameters[64] for coil, strand, and helix and surface accessibility, suggesting that the parameters might reflect database bias at least in part in addition to intrinsic propensity of individual amino acids to form specific secondary structures.

Table 1. Summary of the Results of Six Classical Joint *Bona Fide* Predictions[134,a]

protein	α %	β %	coil %	H−H %	E−E %	C−C %	H−C %	C−H %	E−C %	C−E %	H−E %	E−H %	correct %	serious mistakes %
ribonucleotide reductase	70.9	3.5	25.6	40.0 57.1	1.2 34.3	19.4 75.8	18.8 55.7	3.5	2.4	2.6	12.1	1.2	60.6	13.3
nitrogenase (Fe)	41.1	13.2	45.6	31.7 77.1	9.4 71.2	19.2 42.1	6.6 63.5	16.7	1.7	9.8	3.1	2.1	59.9	5.2
renin	17.4	47.0	35.6	4.6 26.4	24.9 53.0	21.7 61.0	8.5 46.8	5.3	18.9	8.5	4.3	3.2	51.2	7.5
avidin	2.5	51.2	46.3	0.0 0.0	20.7 40.4	36.4 78.6	2.5 39.7	0.8	22.3	9.1	0.0	8.3	57.0	8.3
enolase	42.9	17.4	39.7	30.5 71.1	5.7 32.8	25.0 63.0	7.8 55.6	12.6	5.0	2.1	4.6	6.6	61.2	11.2
soyabean proteinase inhibitor	0.0	29.6	70.4	0.0 -.-	9.9 33.4	56.3 80.0	0.0 56.7	8.4	19.7	5.6	0.0	0.0	66.2	0.0
average													59.3	7.6

[a] The α, β, and coil columns contain the percentage of residues assigned to each of these secondary structural units. The H−H, E−E, and C−C columns contain the percentage of residues in the alignment that are correctly assigned as helices, strands, and coils (respectively); underneath is the percentage of the helix, strand, and coil positions (respectively) correctly identified. The H−C, C−H, E−C, C−E, H−E, and E−H columns contain the percentage of residues in the alignment that are incorrectly assigned, with the first index indicating the experimental assignment, the second indicating the prediction. The percent correct is calculated from (H−H + E−E + C−C)/(total number of positions in the protein), and represents a classical three state residue-by-residue score. A serious mistake is defined as one where a residue in a helix in the experimental structure is predicted to be in a strand, or vice versa. Figure 7 should be inspected to obtain a more comprehensive view of the quality of the predictions.

surface, and the difficulties in obtaining accurate models for water, side chain−solvent interactions, and side chain−side chain interactions. Together, these have often defeated direct computation of protein conformation, although some interesting cases where quite good conformational models have been built.[127] Further, the increase in computational power is encouraging many groups to make a direct assault on the *de novo* computation of protein conformation.[128,129] Some of these have now been shown to fail in specific cases in a *bona fide* prediction setting.[130]

C. Joint Methods

Many have hoped that a prediction can be improved by merging different classical prediction methods to obtain “joint” predictions.[131] For example, the COMBINE method joins the GOR III method with the SIMPA[132] tool and a heuristic known as Bit Pattern, which is a physicochemical tool that searches for hydrophobicity.[133] Joint methods are reviewed elsewhere in detail;[63] specific examples of *bona fide* predictions made using them with homologous sequences are discussed below. To give the reader a general view of how joint methods perform when applied to a single sequence, however, Figure 7 presents a collection of *bona fide* predictions made using a joint method of Nishikawa and Ooi.[134] These authors combined Chou−Fasman and GOR predictions for 10 individual proteins for which no structure was known at the time. Table 1 collects scores of various types for several of these.

This collection of predictions is representative of those made by many others using classical statistical methods, individually or jointly, on single sequences. It is clear from Figure 7 that the results are not useful for tertiary structure modeling. Too many strands are mistaken for helices; too many helices are mistaken for strands. It is this type of data that the editors of the journal *Trends in Biochemical Sciences* were undoubtedly thinking of when they summarized the status of the structure prediction field in 1992 as part of a celebration of the 200th issue of their magazine. They wrote: “The ability to predict folding patterns from amino acid sequences is still, we understand, more a matter for soothsayers than scientists, despite lavish support from optimistic protein and drug designers.” [52]

IV. Introducing Evolution into Classical Prediction Methods

Proteins diverging from a common ancestor retain a core structural fold, as long as the proteins have served a selected function during the period of divergent evolution. This generalization was first adumbrated in the 1970s, when Rossman and his coworkers noted that dehydrogenases acting on different substrates have similar folds.[22] In the mid 1980s, Chothia and Lesk published a quantitative relationship between the extent of identity in two protein sequences and the extent of divergence in their respective conformations.[23]

By almost any perspective, the conservation of fold is remarkable. Sequences that have changed over 70% of their amino acids still have backbone chains that are superimposable with a root mean squared deviation of ∼2 Å. This is not greatly different from the 0.7 Å rms deviation for the identical protein crystallized in two different crystal forms,[23] and not greatly higher than the nominal resolution of many crystal structures in the database. Further, only a modest extrapolation suggests that the core fold will remain after 80−90% of the amino acids have been substituted. At this level of substitution, it is impossible to tell by simple sequence analysis that the two proteins are related by common ancestry. This implies that similar fold is the strongest indicator of common ancestry, stronger than sequence, mechanism, stereospecificity, or any other “wet” biochemical trait.[35,135]

It should be emphasized that conservation of tertiary fold is not an intrinsic property of a protein, but rather an evolutionary property of a protein

```
avidin
ARKCSLTGKWTNDLGSNMTIGAVNSRGEFTGTYTTAVTATSNEIKESPHL  sequence
                 EEEEEE          EEEEEE              predicted
       EEEEETT  EEEE    TT   EEEEE              EEEE  experimental

GTENTINKRTQPTFGFTVNWKFSESTTVFTCQCFIDRNGKEVLKTMWLLR  sequence
            EEEEEE EEEE   EEEEEEEEE     HHHHHHHHH   predicted
  E     TT  EEEEEEE      EEEEEEEEEE     EEEEEEEEEE  experimental

SSVNDIGDDWKATRVGINIFTRLRTQKE                        sequence
    CCCCCCCCCC  EEEEHHHHHHH                         predicted
     HHHHHH EEEEEEEEEE                              experimental

Proteinase  Inhibitor  (soyabean)
DDESSKPCCDQCACTKSNPPQCRCSDMRLNSCHSACKSCI            sequence
                                      EE            prediction
  TT        EEEE      EEEE  EE        EE            experiment

CALSYPAQCFCVDITDFCYEPCKPSEDDKEN                     sequence
EEE     EEEEEE           HHHHHH                     prediction
EE     EEEE                                         experiment

Enolase
AVSKVYARSVYDSRGNPTVEVELTTEKGVFRSIVPSGASTGVHEALEMRD sequence
hhhhhhhhhhCCCCCCC hhhhhhhh  eeee   CCCC   HHHHHHH   predicted
     EEEEEE TT  EEEEEEEEETTEEEEEE  E     TT         experimental

GDKSKWMGKGVLHAVKNVNDVIAPAFVKANIDVKDQKAVDDFLISLDGTA sequence
 hhhhhhhhhhhhhheeeeeehhhhhhhhhhhhhhhhhhheeee CCCC predicted
  TT HHHT  HHHHHHHHHTHHHHHHHH    TT HHHHHHHHHHHT    experimental

NKSKLGANAILGVSLAASRAAAAEKNVPLYKHLADLSKSKTSPYVLPVPF sequence
CCCCCC EEEEEEEhhhhhhhhhhh      HHHHHH CCCCC EEEEEEE predicted
  TTT HHHHHHHHHHHHHHHHHHHT   HHHHHHHHTT      EEE EEE experimental

LNVLNGGSHAGGALALQEFMIAPTGAKTFAEALRIGSEVYHNLKSLTKKR sequence
EEEECCCCCCC hhhhhhhhhhCCCHHHHHHHHHHH    HHHHHHHHHH   predicted
EEEEE  TT          EEEEE       HHHHHHHHHHHHHHHHHHHHHH experimental

YGASAGNVGDEGGVAPNIQTAEEALDLIVDAIKAAGHDGKVKIGLDCASS sequence
  CCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHH   eeeeeeee   h predicted
T HHHH E TT   E        HHHHHHHHHHHHHHTT TTT EEEEE  HH experimental

EFFKDGKYDLDFKNPNSDKSKWLTGPQLADLYHSLMKRYPIVSIEDPFAE sequence
hhhh         CCCCCC            HHHHHHHHHHHHeeeee     HH predicted
HHEETTEE TTTT      TTT E HHHHHHHHHHHHHHH   EEEEE     T experimental

DDWEAWSHFFKTAGIQIVADDLTVTNPKRIATAIEKKAADALLLKVNQIG sequence
HHHHHHHHH     hhhhhh    CCCCCHHHHHHHHHHHHHHeeee       predicted
T HHHHHHHHTT    EEEE TTTTT HHHHHHHHHTT    EEEE HHHH  experimental

TLSESIKAAQDSFAAGWGVMVSHRSGETEDTFIADLVVGLRTGQIKTGAP sequence
  HHHHHHHHHHHHH          CCCCCC  hhhhhhEEEEE    CCC   predicted
 HHHHHHHHHHHHTTT EEEEE    E      HHHHHHHHTT   EEE     experimental

ARSERLAKLNQLLRIEEELGDNAVFAGENFHHGDKL                sequence
 HHHHHHHHHHHHHHHHHH     hhhhhh  CCCCC                predicted
  HHHHHHHHHHHHHHHHHHHHHHEEE HHH TT TT               experimental
```

```
ribonucleotide  reductase  B2  subunit
AYTTFSQTKNDQLKEPMFFGQPVNVARYDQQKYDIFEKLIEKQLSFFWRP sequence
eeeeeCCCCCCCCCCCCCCCCCC         HHHHHHHHHHHHCCCCCC  predicted
           HHH   E                HHHHHHHHHHHHT    H experimental

EEVDVSRDRIDYQALPEHEKHIFISNLKYQTLLDSIQGRSPNVALLPLIS sequence
      CCCCCCC hhhhhhhhhhh    hhhhhhh CCCC EEEEEEEE predicted
HH   HHHHHHHHH  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHTHHHE  experimental

IPELETWVETWAFSETIHSRSYTHIIRNIVNDPSVVFDDIVTNEQIQKRA sequence
   hhhhhhhhhh   CCCC   eeeeee    eeeeeeHHHHHHHHHHH predicted
 HHHHHHHHHHHHHHHHHHHHHHHHHHHHH   THHHHHHHHH HHHHHHH experimental

EGISSYYDELIEMTSYWHLLGEGTHTVNGKTVTVSLRELKKKLYLCLMSV sequence
HCCC   HHHHHHHH eeee                 eeeeHHHHHHHhhhhhhh predicted
TTHHHHHHHHHHHHHHHHHHH  EEEEETTEEEEE HHHHHHHHHHHHHHH experimental

NALEAIRFYVSFACSFAFAERELMEGNAKIIRLIARDEALHLTGTQHMLN sequence
hHHHHHHHHHHHHHHHHHHHHHHHH   hhhhhhHHHHHHHH    eeee predicted
HHHHHTHHHHHHHHHHHHHHHTT  HHHHHHHHHHHHHHHHHHHHHHHHHH experimental

LLRSGADDPEMAEIAEECKQECYDLFVQAAQQEKDWADYLFRDGSMIGLN sequence
e   CCCCHHHHHHHHHHHHHHHHHHHHHHHHHHhhhhhhh CCC       predicted
HHHTT   HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHTTT  ETTE experimental

KDILCQYVEYITNIRMQAVGLDLPFQTRSNPIPWINTWLVSDNVQVAPQE sequence
  eeeeeeeeee eeeeee         CCCCC         CCCCC    h predicted
HHHHHHHHHHHHHHHHHHTTT                TTHHHHH          experimental

VEVSSYLVGQIDSEVDTDDLSNFQL                          sequence
hhhh  eeeee    CCCC                                predicted
                                                   experimental


Ribosomal  protein  S5
        ELEERVVAVNRVAKVVKGGRRLRFSALVVVGDKNGHVGFGTG sequence 1
AHIEKQAGELQEKLIAVNRVSKTVKGGRIFSFTALTVVGDGNGRVGFGYG sequence 2
hhhhhhhHHHHhhhhhhhhhh CCCEEEEEEEEEEEECCCCCeeee CC predicted
          EEEEEEEE    E         EEEEEEEE    EEEEEEE experimental

KAQEVPEAIRKAIEDAKKNLIEVPIVGTTIPHEVIGHFGAGEIILKPASE sequence 1
KAREVPAAIQKAMEKARRNMINVALNNGTLQHPVKGVHTGSRVFMQPASE sequence 2
CCCHHHHHHHHHHHHHHHHHHEEEEECCCCC      CCCCCEEEEECCCC predicted
EE  HHHHHHHHHHHHTT   EE    ETTE    EEEEETTEEEEEEE  T experimental

GTGVIAGGPARAVLELAGISDILSKSIGSNTPINMVRATFDGLKQLKRAE sequence 1
GTGIIAGGAMRAVLEVAGVHNVLAKAYGSTNPINVVRATIDGLENMNSPE sequence 2
C EEEE  HHHHHHHHHHHHHHHHHH  CCCCeeeee    CCCCCC hh predicted
T  EE  HHHHHHHHHHT    EEEEE     HHHHHHHHHHHHHH      experimental

DVAKLRGKTVEELLG                                    sequence 1
MVAAKRGKSVEEILG                                    sequence 2
hhhhhh HHHHHHHH                                    predicted
                                                   experimental
```

Figure 7. Comparison of *bona fide* joint predictions made by Nishikawa and Ooi[134] from a single protein sequence with subsequently determined experimental structures. Experimental secondary structure assignments taken directly from the crystallographic database. Key: E, β strand; H, α helix; T, turn; C, coil. In the prediction, "e" refers to a weakly predicted strand, while "E" refers to a strongly predicted strand; "h" refers to a weakly predicted helix, while "H" refers to a strongly predicted helix.

evolving under functional constraints. Changing randomly 70% of the amino acids in a polypeptide chain will, with extraordinarily high probability, greatly change the conformation of the protein. The fact that it has not done so in natural proteins arises from the fact that before they enter our databases, proteins that have undergone random variation have been filtered by natural selection to remove polypeptides that do *not* retain the same overall tertiary fold, at least to the extent that they can help their host organism survive, select a mate, and reproduce.[35]

A. Homology Modeling

For structure prediction, the conservation of conformation after substantial sequence divergence has an important corollary: if one knows the conformation of one member of a protein family, one knows (more or less) the conformation of all other members of the family. This corollary has generated the field of "homology modeling". In this field, the conformation of a target sequence is modeled by extrapolation of an experimental conformation of a homolog with known structure. It has also created the impetus to develop methods for detecting very distant homologs of proteins, as these are the starting points for homology modeling.

Homology modeling is one type of approach that uses evolutionary analyses to predict protein conformation. The second type, often referred to as *ab initio* structure prediction, seeks structure of a family of proteins where no member of the family has a known experimental conformation.

Ab initio prediction is the primary focus of this review. Homology modeling does, however, introduce concepts that are valuable for all evolution-based structure prediction tools. Further, as discussed below, the goal of an *ab initio* structure prediction exercise is often a consensus model for a protein family that needs optimization for a specific protein sequence that is a member of this family. This might, at least in principle, be done using procedures that have been developed for homology modeling. Therefore, we summarize briefly the approach of homology modeling and provide leading references for the reader who wishes to delve deeper.

1. Homology Modeling with a Clearly Identifiable Homolog

Homology modeling is the process of creating a model of the conformation of a target protein by comparing it to a homolog with known conformation.[136,137] It is difficult to identify the origins of homology modeling. In 1969, Brown *et al*. built a three-dimensional model of bovine α-lactalbumin starting from the known structure of hen's egg white lysozyme, which was believed to be a homolog.[138] Argos and Rossman were concerned in the mid-1970s with comparing structures of homologous proteins, following the discovery that dehydrogenases acting on different substrates had similar folds.[139] An excellent example of homology modeling was provided by Greer for serine proteases.[140] Homology modeling has become still more widespread with the increase in computational power and the refinement of molecular dynamic tools. The approach has recently been covered in a number of excellent reviews.[42,137,141,142]

Homology modeling requires four steps:

(i) First, a protein must be found in the crystallographic database that can be shown to be a homolog of the target protein. Generally, this is done by a computer, which attempts to align the sequence of the target protein with the sequences of every protein in the crystallographic database. The criteria for a match are discussed in greater detail below. Generally, however, if a protein in the crystallographic database can be found that matches the sequence of the target protein with 30% identity (or more) over a segment length of 100 amino acids (or more), a homolog with known structure has been found and homology modeling can begin.

(ii) Next, an alignment must be constructed to pair specific amino acids in the sequence of the target to specific amino acids in the sequence of the reference protein. This process is obviously easier if the reference and target proteins are more similar in sequence than if they are not. After substantial amounts of sequence divergence (see below), the alignment requires placement of gaps. This is difficult, and undermines many examples of homology modeling exercises,[143] as discussed below.

(iii) Next, amino acids in the reference protein must be replaced in the crystal structure of the reference protein by the amino acids found at the corresponding position in the target protein. The orientation of the side chains can come from a variety of sources, including the original structure,[77,144] from matching protein segments,[10] from a library of rotamer conformations,[145] or from similar local residue environments found in the protein database.[146] The details of the approach are reviewed elsewhere.[42]

(iv) Last, the conformation of the resulting model, having the coordinates of the reference protein but the sequence of the target protein, must be optimized. This is generally done by molecular mechanics processes, which in turn rely on force fields. The goals, aside from minimizing the potential energy of the model, include removing unfavorable contacts, filling in holes in the structure, or modeling loops that appear in the target sequence without a corresponding element in the reference structure.

A variety of computer packages are now available to do homology modeling. These include Composer (from Tripos), Look (from Molecular Applications Group), Modeller (from Molecular Simulations Inc.), and Insight-Homology.

2. Does Homology Modeling "Work"?

Evaluation of a homology model presents different problems than evaluation of a secondary structure prediction. First, homologs share essentially all of the core secondary structural elements. Therefore, if one has truly identified a homolog with a known crystal structure, and if the sequence identity is greater than 30%, it is difficult not to correctly place the core secondary structural elements. When scoring a homology model, the structural features at issue are those in which the target and reference structure differ. This is conveniently measured by a root mean squared (rms) deviation of atoms in the target sequence and atoms in the model for the target structure as built by analogy to the reference structure.

Not surprisingly, homology modeling of secondary structure is successful by most of the standards used to judge prediction methods; it could hardly be otherwise. Further, it is most successful when the target protein and the reference protein are relatively similar in sequence. The less the sequence of the target protein has diverged from that of the reference homolog, the more similar the conformation of the target sequence will be to the known structure. For example, Harrison *et al.* examined six comparative modeling targets predicted in a procedure that relied on energy minimization alone to position all new atoms.[147] Root mean squared deviations between the calculated and experimental structure on C α atoms in the polypeptide backbone ranged from 0.69 to 1.73 Å in protein pairs whose sequence identities decreased from 60 to 20%. Similar results have been seen in other examples.

In *bona fide* predictions, homology modeling has done less well in modeling those regions (generally external loops) where homologs have different conformations. In the CASP1 project in 1994,[148] for example, six models were built for the eosinophil-derived neurotoxin, a homolog of ribonuclease A with approximately 35% sequence identity. A range of modeling methods and force fields were used; each started with a high resolution crystal structure of ribonuclease A. Using root mean squared deviations as a guide, all six of the models computed by energy optimization were more different from the target structure than the starting structure. In other words, a better model of eosinophil-derived neurotoxin would have been obtained by using the coordinates of ribonuclease A directly without *any* energy optimization than the coordinates produced by any of the refinement packages tested. This disappointing result undoubtedly reflects the immature status of force fields, and difficulties inherent in detailed modeling of interactions between solutes and water.

3. Homology Modeling with Distant Homologs: Profile Methods and Threading

Homology modeling faces an obvious limitation: it works only if a homolog can be found in the crystallographic database. With only a few hundred folds in the database, this is by no means certain with any particular target sequence. What happens if a homolog is not readily discernible in the database?

The first approach is to relax the criteria used to identify the homolog. While proteins sharing 30% sequence identity are certainly homologs, proteins with a 25% identity are likely to be homologs as well. Below this level, one enters the "twilight zone" of protein structure sequence comparison,[149] a region where nebulous similarities between sequences can be seen, each suggestive of distant homology, but none adequate to make a statistically significant case for it. Considerable effort has been devoted to developing tools to identify long-distance homologs in a database, in particular, by expanding the tools needed to compare protein sequences directly.[92,150,151] Many of these have been reviewed recently.[152]

A more comprehensive class of tools that combines sequence and structural information has been developed to detect long-distance homologs. These come under the titles of "profile methods", "threading", or occasionally as approaches to "the inverse folding problem".[153,154] The inverse folding concept aims to reformulate the prediction challenge to change the question from "What conformation does this sequence adopt?" to the question "What sequences adopt this conformation?" The philosophies and strategies underlying these approaches are discussed below.

Early work by Eisenberg and his co-workers developed "profile" methods for detecting distant homologs in a database of known structures. In its first version, protein sequences related to a protein with known conformation were aligned, and the probabilities of each of the 20 amino acids appearing at each position in the alignment were deduced from the sequences. The result is a "profile" of the protein family, a position-by-position statement of what residues might be accepted by functional constraints on the divergent evolution of the family. The sequence of a target protein can then be examined to see whether it fits the profile.[155,156] If it does, then the protein with known conformation is a possible homolog of the target protein. The alignment generated by the profile analysis is then used as the starting point for homology modeling as described above.

The profile method was extended by Bowie *et al.* to include information directly related to the conformation of the reference protein, available from the crystallographic database.[28] Here, the environment of each amino acid in the reference crystal structure is assigned to one of a number of classes, for example, the local secondary structure, the extent to which the side chain is buried, or the nature of other atoms in contact with the side chain. This provides more information, this time from the known conformation of the reference protein, that can be used to better assess the probability that the target protein might have the same fold. Blundell and his group have developed in parallel a set of structure-based substitution matrices that has the same effect.[157] In each case, the goal is to glean as much information about the proteins in the crystallographic database that might be extrapolatable to very distant homologs, the target protein in particular.

A third approach reconstructs a maximum likelihood representation of the most recent common ancestor of all proteins in a family.[92] This ancestor stands at the head of an evolutionary tree and represents the most ancient protein in the family. The ancestral protein is the closest in geological time to the divergence point of any long-distance homolog, and therefore resembles it most closely. Thus, if the target sequence is to align with any sequence clearly homologous to a protein with a known conformation, then it will be to this ancestral sequence.

In a prediction setting, threading is to date the most popular way to use such methods to identify distant homologs.[158,159] A threading heuristic attempts to fit, or "thread" a sequence of a target protein onto the coordinates of another protein of known structure. Threading may use profiles or may attempt to combine molecular mechanics with a reference conformation to learn how easily the sequence of the target protein can be "fit" on top of the reference crystal structure. In this case, force fields are important to evaluate the fit of the threaded sequence from the target protein on the reference protein structure. Especially influential have been pairwise potentials derived by examining crystal structures directly.[160,161] Last, although a technical detail to those not working in the field, threading and homology modeling can be treated within the mathematical framework known as "hidden Markov models", a field that concerns strategies for rigorously defining and optimizing models on the basis of a large number of probability tables.[162]

Threading asks whether the target sequence *might* adopt the same fold as the sequence with a crystal structure. It is, in this way, an "inverse folding" approach to structure prediction. It relies again on the database having a protein that is homologous or, in its broadest interpretation, simply analogous in structure, to the target protein. In the first case, threading is simply long-distance homology modeling, with selective pressure conserving the functional aspects of the fold during long periods of divergent evolution. In the second, threading implies the convergence of tertiary fold, which reflects underlying propensities of amino acids in the two proteins to form the same conformation.

4. Does Threading Work?

Unlike with homology modeling with clearly identifiable homologs, threading can be judged in two ways. We first may ask whether the reference protein in the crystallographic database identified by the threading procedure is indeed a homolog. Obviously, if the overall fold of the reference protein from the database proves to be radically different from that of the target protein, the threading exercise has failed.

If the reference protein from the database proves to have the same overall fold, the threading tool has successfully passed its first test. Next, the threading must produce a correct alignment between the target and reference sequences. Secondary structural elements in the target structure must be matched with the homologous elements in the reference structure. This matching is critical for the next step: replacing amino acids in the reference structure by amino acids from the target structure. If the alignment is incorrect, the homology model will be incorrect. A threading result can therefore be judged by how well the alignment has succeeded.

A large number of reviews have appeared recently assessing the outcome of threading exercises, both tested retrodictively and in *bona fide* prediction settings.[29,58,163−165] Perhaps the earliest significant concentration of *bona fide* predictions came through the threading test performed in the context of the "Critical Assessment of Structure Prediction" (CASP1) project consummated in Asilomar in December 1994.[166,167] Here, the results were intriguing.[168,169] Nine different teams of predictors submitted 86 threading predictions covering 21 target proteins, chosen to have little or no sequence similarity to proteins of known structure. Of these, 44 predictions were submitted for 11 target proteins that were later found to adopt known folds. The predictions for the remaining 10 proteins were not analyzed, as the fold adopted by these proteins displayed no strong similarity to any fold known in the database (making it impossible for even the best threading tool to succeed).

In many cases, threading identified a protein in the database having a similar fold. Indeed, every team predicted correctly some target structures, and virtually all targets were assigned a correct fold by at least one team. One team identified the correct homolog in five of the nine test cases. Common folds such as the eight-fold $\alpha-\beta$ barrel were recognized more readily than folds with only a few examples known in the database.

Surprisingly, however, the quality of the alignments generated by the threading tools turned out to be quite poor in many cases. This was true even in the cases where the threading method had correctly identified the fold in the crystallographic database that resembles the fold in the target protein. In other words, the threading had identified in the database a protein having the same fold as the target protein, but not for the correct reasons. Further, the alignment generated by the threading tool could not be used to superimpose the target protein sequence on the reference protein structure. Lemer *et al.* concluded from this result that "threading can presently not be relied upon to derive a detailed three dimensional model from the amino acid sequence",[168] and offered some suggestions for why incorrect alignments might identify correct homologs.

Others have provided additional evaluations of the results.[29,44] agreeing about both the "good news" (it is likely that a correct homolog will be identified by at least one threading tool) and the "bad news" (no single tool is able to identify a correct homolog with a correct alignment in most of the challenges). This combination of good and bad news might, of course, indicate that each of the threading tools is making a small contribution toward a larger solution to the problem. Unfortunately, it is also consistent with the conclusion that the tools are randomly identifying homologs in the database. As the database is finite, and as the evaluation considers only those prediction targets that have a homolog in the database, the results obtained when a large number of prediction tools produce random assignments will also be distributed so that at least one tool will get the correct answer for every individual case, but no tool will get the correct answer in many cases. Distinguishing the two interpretations of the CASP1 threading project depends on the precise number of tools, targets, and reference structures and is complicated by difficulties in finding a controlled set of proteins to test threading methods.

This being said, threading methods remain intriguing, and several threading predictions are included in the figures associated with *bona fide* predictions discussed below. In part, the approach will undoubtedly be improved by new force fields, and many groups continue to work in this area.[170] One encouraging recent example, also made in a *bona fide* prediction setting, concerns the protein leptin derived from the obesity gene. Bryant applied a threading tool to propose that leptin may be a helical cytokine.[171] The receptor for leptin was later identified and shown to belong to the family of cytokine receptors.[172] Very recently, the crystal structure of a variant of human leptin was solved, showing a good correlation between the model based on threading and the experimental structure.[173] The CASP2 threading project in December 1996 produced additional results, which will be reviewed elsewhere.[130,148,174]

B. Knowledge-Based Modeling

Homology modeling is best defined strictly as the process of identifying a protein with known conformation that is a homolog of a target, where the conformation of the homolog is used as a starting point to model the conformation of the target.[136,137] A process that appears similar, but is different in terms of its underlying philosophy, is “knowledge-based” modeling. Here, a database of peptide fragments with known conformations is assembled from the crystallographic database (the “knowledge”). Similar sequences in the target protein are then identified, and modeled on the basis of the conformational information in the database.[142]

Although somewhat similar in form, homology modeling and knowledge-based modeling differ fundamentally in theory. Homology modeling assumes that the conformation of the target protein is similar to the conformation of the homolog in the databank because the proteins are homologs. Knowledge-based modeling assumes that the conformation in the target protein and the protein in the databank are similar because of intrinsic tendencies of similar polypeptide segments to adopt similar folds.[175] Thus, knowledge-based modeling assumes that long-range “tertiary” interactions are not important, while homology modeling relies upon them. Knowledge-based modeling is therefore best considered as an *ab initio* approach, provided that the protein that is providing the knowledge is not a homolog of the target protein.

An interesting illustration of the distinction between homology and knowledge-based secondary structure prediction is provided by the SIMPA software package developed by Levin and Garnier.[132] The package assigns secondary structure on the basis of sequence similarity between a stretch of amino acids (17 amino acids long) in the target sequence and the sequences in a database of known structure. Similarity in the two amino acid sequences might, of course, indicate that the entire target protein is a homolog of the entire reference protein. If so, this secondary structure can be said to have been obtained by homology modeling, and is accurate with a Q_3 of 87%. Alternatively, the target and reference proteins might not be homologs. In this case, the similarities in the sequences in the 17 amino acid segment arose convergently. If the segments have similar secondary structure, then the secondary structure also arose convergently, and reflects in part the intrinsic propensity of the amino acids particular segment to adopt the specific secondary structure; this is knowledge-based prediction. In this case, however, the Q_3 score drops to 63%.[132]

C. *Ab Initio* Approaches

Even should homology modeling work, it does not address the larger challenge, *ab initio* prediction, where a full conformational model is built without reference to any experimental conformation of any homolog. *Ab initio* prediction methods come in many forms. As these are the principal focus of this review, we will review each in some detail. At the outset we should note, however, that one conclusion that might be drawn from this discussion is that the distinction between *ab initio* and homology modeling tools is not always clear.

As with homology modeling, *ab initio* prediction tools that assign secondary structure to a protein using evolutionary information begin with an alignment. Again, the alignment shows the evolutionary relationship between individual amino acids in two or more homologous protein sequences. As before, amino acids matched in the alignment are encoded by codons in their respective genes that are presumably descendants of a single codon in a single ancestral protein.

Given an alignment, one way of extracting conformational information is simply to apply the same secondary structure prediction tool to each of the homologous sequence individually and then extract a “consensus” secondary structure prediction for the whole family by averaging these individual predictions. For example, a “consensus Chou−Fasman” prediction is obtained by applying the Chou−Fasman heuristic to each member of a protein family and then by averaging the individual predictions. A “consensus GOR” prediction is obtained in the same way using the GOR heuristic.

Alternatively, the alignment might be inspected residue-by-residue, with patterns of variation and conservation used to infer information about the conformational environment for each individual position. This process, occasionally known as looking “down” an alignment (as opposed to looking “across” an alignment), is different in its implementation from the “consensus” approach noted above.

Both approaches have been explored in the past decade, and both must consider the way in which homologous protein sequences are averaged, or weighted, in the analysis. It is generally incorrect to make a numerical average (or “majority rule”) to obtain a consensus prediction. Ten closely similar proteins with the same conformation should not carry 10 times the weight of one distantly homologous protein in a consensus prediction. When averaging any property across a family of homologous proteins, the relationships between members of the family must be considered. The most effective use of evolutionary information comes with a *per stirpes* analysis that weights lineages (branches in a tree) according to their priority of divergence. This will be discussed in greater detail below.

D. *Bona Fide* Predictions Made with Consensus Classical Methods

A simple method for exploiting the similarities in the conformations of homologous proteins in a prediction, but without the need to identify a homologous protein whose structure has already been solved, is to simply apply a classical prediction method to each member of a protein family, obtain separate predictions, and then average the individual predictions in some way to obtain a consensus model. This approach assumes that the mistakes made by a classical method using a single sequence represent "noise".[176] Should this be the case, averaging secondary structural predictions over a set of sequences that differ in their details but which fold to give the same secondary structure overall should filter out the noise, leaving behind the signal.

The "consensus classical" approach was identified first by Lenstra *et al.*, who applied three classical methods individually to each member of a family of pancreatic ribonucleases.[26] Two probabilistic tools (Chou−Fasman[104] and Burgess−Scheraga[177]) and the physicochemical tool developed by Lim[120] were used. The results were then compared with a known crystal structure for the ribonuclease family.

Overall, the results obtained by averaging these particular classical prediction tools were disappointing, despite the use of evolutionary information. The secondary structure assignments made for the ribonuclease homologs by the Burgess−Scheraga method were not consistent, and it was difficult to obtain a sensible average secondary structural model over the entire protein family. The Chou−Fasman method provided more consistent assignments for individual sequences in the protein family, but the overall retrodiction was disappointing. This suggested that the Chou−Fasman parameterization contained systematic errors, which cannot be removed by averaging. Only the Lim method showed promise. Lenstra *et al.* also pointed out that hydrophobic side chains are not only frequently found inside globular structures, but that hydrophobicity is often conserved at critical interior positions during divergent evolution.[26]

The notion of averaging predictions made by classical tools for individual members of a protein family over a set of homologous protein sequences has recurred often in the literature. Maxfield and Scheraga noted that small improvements could be made in predictions by averaging predictions made on individual sequences over a set of homologous sequences.[25] Similarly, Garnier *et al.* suggested that predictions made with the use of their method might be improved by averaging predictions obtained from homologous sequences.[64] These suggestions have recently been analyzed systematically. Adding homologous protein sequences over a set of homologous sequences appears to improve the three state residue-by-residue score (Q_3) of an average prediction by 5−10 percentage points.[178] Regrettably, this approach has not been evaluated with more useful scoring methods, and has not been quantified in detail with respect to different parameters of the evolution of sequence families. It would be interesting to know whether improvements obtained when classical methods are applied to a family of homologous sequences arise disproportionately in core regions of the fold, and reflect fewer serious errors. A recent paper takes the first steps in this direction.[179]

Table 2. Consensus Classical Prediction

predictions made with input from circular dichroism data
interferon[24,180]
aspartate receptor[184]
annexin[187]
predictions made without input from circular dichroism data
tryptophan synthase[27]
glutamine amidotransferase[197]

Nevertheless, the consensus classical approach has been used frequently to make *bona fide* predictions that have an element of transparency. These are therefore the first that we will discuss that fall directly within the scope of this review. Many of these predictions can now be analyzed by a subsequently determined experimental structure. These are listed in Table 2, and discussed individually below.

1. All Helical Proteins

Because helical proteins have a distinctive signature in their circular dichroism spectra, they are easy to recognize with relatively little experimental effort. Therefore, helix bundles were among the first challenges to classical methods averaged over a set of aligned homologous protein sequences. As the examples below illustrate, the effort met with considerable success.

Interferons were among the first proteins examined in this way using the "consensus classical" approach.[24,180] Sternberg and Cohen applied classical prediction heuristics to make secondary structure predictions for four homologous interferons, and then averaged the predictions to generate a consensus prediction for the interferon family. This was then used as the starting point for tertiary structural modeling. Although no crystal structures were known for any member of the interferon family when the prediction was made (making it a *bona fide* prediction), the prediction was not based solely on sequence data. Circular dichroism data suggested that the polypeptide chain adopted only helical secondary structures,[181] and this information was used to guide the prediction. Much later, an experimental structure became available for the interferon family.[182] When analyzed in detail in light of an evolutionary alignment,[183] four of the five helices in the protein were correctly predicted (Figure 8).

Circular dichroism data also indicated a helical structure for much (90%) of the extracellular domain of the aspartate receptor from *Escherichia coli*.[184] This information was combined with information derived from patterns of hydrophobicity and hydrophilicity, suggesting helical conformations. The amino acid sequences in each of these regions was correlated with similar regions in other bacterial receptors. Chou−Fasman analysis was used to identify turns in the structure, and a crude energy minimization was done to evaluate possible packings (Figure 9).[185,186] As Figure 9 shows, the positions of the helices as assigned from experimental data were predicted quite well, even though their lengths were significantly underestimated.

Likewise, Taylor and Geisow[187] and, later, Barton *et al.*,[188] exploited circular dichroism data that suggested that the annexins formed largely helical

```
    0    0    0    0    0    0    0    0    0    0    0    0
    0    1    1    2    2    3    3    4    4    5    5    6
    5    0    5    0    5    0    5    0    5    0    5    0
INYKQLQLQERTNIRKCQELLEQLNGKI--NLTYRADFKIPEEMTEKMQ--KSYTAFAIQ sequence (expt)
MSYNLLGFLQRSSNFQCQKLLWQLNGRLEYCLKDRMNFDIPEEIKQLQQFQKEDAALTIY sequence (pred)
              HHHHHHHHHHHHH                             HHHH predicted
       HHHHHHHHHHHHHHHHHH                              HHHHH experimental


    0    0    0    0    0    0    0    1    1    1    1    1
    6    7    7    8    8    9    9    0    0    1    1    2
    5    0    5    0    5    0    5    0    5    0    5    0
EMLQNVFLVFRNNFSSTGWNETIVVRLLDELHQQTVFLKTVLEEKQE-ERLTWEMSSTAL sequence (expt)
EMLQNIFAIFRQDSSSTGWNETIVENLLANVYHQINHLTKVLEEKLEKEDFTRGKLMSSL sequence (pred)
HHHHHHHHHHHH        HHHHHHHHHHHHHHHHHHHHHHHHHH               predicted
HHHHHHHHHHH           HHHHHHHHHHHH                        HH experimental


    1    1    1    1    1    1    1    1
    2    3    3    4    4    5    5    6
    5    0    5    0    5    0    5    0
HLKSYYWRVQRYLKLMKYNSYAWMVVRAEIFRNFLIIRRLTRNFQN               sequence (expt)
HLKRYYGRILHYLKAKEYSHCAWTIVRVEILRNFYFINRLTGYLRN               sequence (pred)
                      HHHHHHHHHHHH                           predicted
HHHHHHHHHHHHHHH        HHHHHHHHHHHHHHHH                      experimental
```

Figure 8. Representative sequences, *bona fide* consensus prediction,[24] and experimental[182] secondary structure for the interferon α family. The target protein used in the prediction was different from the protein whose crystal structure was ultimately solved. Both protein sequences are shown for comparison. Key: H, α helix.

```
a MFNRIRVVTMLMMVLGVFALLQLVSGGLLFSSLQHNQQGFVISNELRQQQSELTSTWDLMLQTRINLSRSAARM Asp receptor
b MINRIRVVTLLVMVLGVFALLQLISGSLFFSSLHHSQKSFVVSNQLREQQGELTSTWDLMLQTRINLSRSAVRM Asp receptor
c MLKRIKIVTSLLLVLAVFGLLQLTSGGLFFNALKNDKENFTVLQTIRQQQSTLNGSWVALLQTRNTLNRAGIRY Ser receptor
  *..**..** * .**.**.**** **.*.*..* . .   *.. . .*.**..*...* ..****..*.*...*.
                                              HHHHHHHHHHHHHHHHH            predicted
  start aspartate binding domain HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH experimental

a MMDASNQQSSAKT_DLLQNAKTTLAQAAAHYANFKNMTPLP____AMAEASANVDEKYQRYQAALAELIQFLDN Asp receptor
b MMDSSNQQSNAKV_ELLDSARKTLAQAATHYKKFKSMAPLP____EMVATSRNIDEKYKNYYTALTELIDYLDY Asp receptor
c MMDQNNIGSGSTVAELMESASISLKQAEKNWADYEA___LPRDPRQSTAAAAEIKRNYDIYHNALAELIQLLGA Ser receptor
  *** .* .* .....*...*  .*.**. ... .. . .**.... . .........*  *. **.***..*.
                HHHHHHHHHHHHHHHHHHH                 HHHHHHHHHHHHHHHHHH       predicted
  HH         HH HHHHHHHHHHHHHHHHHHHHHH          HHHHHHHHHHHHHHHHHHHHHHHHHHHH experimental

a GNMDAYFAQPTQGMQNALGEALGNYARVSENLYRQTFDQSAHDYRFAQWQLGVLAVVLVLILMVVWFGIRHALL Asp receptor
b GNTGAYFAQPTQGMQNAMGERFAQYALSSEKLYRDIVTDNADDYRFAQWQLAVIALVVVLILLVAWYGIRRMLL Asp receptor
c GKINEFFDQPTQGYQDGFEKQYVAYMEQNDRLHDIAVSDNNASYSQAMWILVGVMIVVLAVIFAVWFGIKASLV Ser receptor
  *.  ..*.*****.*.. .. .. *.  .. *..  . ... .*..*.* *.  . *..... ..*.**.  *.
             HHHHHHHHHHHHHHHHHHHHHH                                          predicted
  H HHHHHHH HHHHHHHHHHHHHHHHHHHHHHHHHHH end of aspartate binding domain      experimental
```

Figure 9. Representative sequences, *bona fide* consensus prediction,[186] and experimental[185] secondary structure for the extracellular domain of the aspartate receptor from *E. coli*. Residues 31−188 constitute this extracellular aspartate binding domain. Sequences of three homologous receptors are given: (a) (P02941, MCP2_SALTY) methyl-accepting chemotaxis protein II (MCP-II)(aspartate chemoreceptor protein) *Salmonella typhimurium*; (b) (P07017, MCP2_ECOLI) methyl-accepting chemotaxis protein II (MCP-II) (aspartate chemoreceptor protein) *Escherichia coli*; (c) (P02942, MCP1_ECOLI) methyl-accepting chemotaxis protein I (MCP-I) (serine chemoreceptor protein) *Escherichia coli*. Key: asterisks (*) indicate amino acids that are conserved; periods (.) indicate amino acids that have undergone conservative substitution. Key: E, *β* strand; H, α helix; t, turn.

structures. Taylor and Geisow subjected each annexin to a secondary structural analysis using the GOR tool,[64] with the decision constants preferentially chosen to favor helix predictions. This analysis found each of the helices later assigned to the experimental structure (Figure 10), with the third, fourth, and fifth helices incorrectly joined into one long helix.

Using an unbiased set of parameters, these helices were separated, but one helix was mispredicted as a strand by the consensus GOR tool (a “serious” mistake, see above). The two predictions with biased and unbiased decision constants are shown in Figure 10. Taylor and Geisow combined the two to obtain a model that corresponded closely to the subsequently determined experimental secondary structure, both in the position and length of the predicted helices. Barton *et al.* did a similar analysis using a variety of methods, including that of Zvelebil *et al.*,[21] Chou and Fasman,[104] and GOR.[64] In this respect, theirs was a joint prediction guided by circular dichroism data. Their predictions, also collected in Figure 10, are largely consistent with those of Taylor and Geisow.

Taylor and Geisow took the next step, using their secondary structure prediction as the starting point for modeling tertiary structure. They began by searching the crystallographic database for an experimental structure built from a set of secondary

```
  AQFDADELRAAMKGLGTDEDTLIEILASRTNKEIRDINRVYREELKRDLAKDITSDTSGDFRNALLSLAKG sequence (expt)
FDERADAETLRKAMKGLGTDEESILTLLTSRSNAQRQEISAAFKTLFGRDLLDDLKSELTGKFEKLIVALMKP sequence (pred)
 HHHHHHHHHHHHHHH    HHHHEEEE      HHHHHHHHHH    HHHHHHHHHHH  HHHHHHHHHHHH predict 1 ref.187
 HHHHHHHHHHHHHHH  HHHHHHHHHHH   HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH  predict 2 ref.187
    HHHHHHHHHHH      HHHHHHHH     HHHHHHHHHHHH HHHHHHHHHHHHH HHHHHHHHH    predict 3 ref.188
  HHHHHHHHHHHH      HHHHHHHHtt  HHHHHHHHHHHHHHH   HHHHHHHH  HHHHHHHHHHH   experimental
```

Figure 10. Representative sequences, *bona fide* consensus prediction, and experimental secondary structure for annexin. Prediction 1 was made from a multiple alignment using a consensus GOR method with unbiased decision constants. Prediction 2 was made from a multiple alignment using a consensus GOR method with decision constants biases to favor all helices to reflect circular dichroism data. Predictions 1 and 2 are adapted from ref 187. Prediction 3 was made analogously (see ref 188). Experimental secondary structure is taken for ANX5_HUMAN annexin V (lipocortin V, endonexin II). The target protein used in the prediction was different from the protein whose crystal structure was later solved. Both protein sequences are shown for comparison. Key: E, β strand; H, α helix; t, turn.

```
MERYENLFAQLNDRREGAFVPFVTLGDPGIEQSLKIIDTLIDAGADALELGVPFSDPLADGPTIQNANLRAFAA sequence
HHHHHHHHHHHHHH    EEEEEE     HHHHHHHHHHH     EEEEE             eeeee       prediction
 HHHHHHHHHHtttt  EEEEEEEtt   HHHHHHHHHHHHHtt   EEEE           HHHHHHHHHHHt experimental
     α1            β1             α2           β2             (non-core)

GVTPAQCFEMLALIREKHPTIPIGLLMYANLVFNNGIDAFYARCEQVGVDSVLVADVPVEESAPFRQAALRHNI
    HHHHHHHHHHH     EEEEEEEEEE       HHHHHHHHHHHHHEEEEEE   HHHHHHHHHHH     prediction
t  HHHHHHHHHHHHtt     EEEEE HHHHtt  HHHHHHHHHHHt  EEEEtt  HHH HHHHHHHHHtt  experimental
        α3             β3                α4           β4          α5

APIFICPPNADDDLLRQVASYGRGYTYLLSRSGVTGAENRGALPLHHLIEKLKEYHAAPALQGFGISSPEQVSA
EEEEEE     HHHHHHHH     EEEEEEE        HHHHHHHHHHHHHHHHH   EEEEE    HHHHHH prediction
EEE    tt   ttHHHHHHH    EEE                HHHHHHHHHHtt   EEE      HHHHHH experimental
 β5            α6         β6                   α7           β7         α8

AVRAGAAGAISGSAIVKIIEKNLASPKQMLAELRSFVSAMKAASR
HHHHH         EEEEE       HHHHHHHHHHHHHHHHHHH pred
HHHHt  EEEEttHHHHHHHH tt HHHHHHHHHHHHHHHHH    expt
        β8     α9               α10
```

Figure 11. Representative sequence, *bona fide* consensus prediction,[27] and experimental[191] secondary structure for tryptophan synthase (α chain). Experimental secondary structural assignments are taken directly from SwissProt entry TRPA_SALTY, tryptophan synthase α chain (EC 4.2.1.20) from *Salmonella typhimurium*. Key: E, β strand; H, α helix; t, turn. In the prediction, "e" refers to a weakly predicted strand, while "E" refers to a strongly predicted strand; "H" refers to a strongly predicted helix.

structural elements similar to those that they had predicted for the annexins. The bovine intestinal vitamin D-dependent calcium-binding protein (ICaBP) met their specifications and served as a template for tertiary structural modeling of annexin. This superimposition made no direct presumption of homology and might be viewed as knowledge-based modeling.

While the annexin prediction was not an explicit search for homologous structures, secondary structure predictions could clearly be used to identify long-distance homologs where secondary structure, but not sequence, had been sufficiently conserved. For example, Pearl and Taylor[189] and Bazan and Fletterick[190] were able to interpret a secondary structure prediction made by consensus GOR prediction for viral proteases with unknown structure to confirm the speculation that these proteases are homologs of aspartic proteases with known experimental structures. This is a form of threading, where predicted secondary structural information is used to help in the detection of long-distance homologs (see below).

2. Moving Up to α–β Barrels

No prediction method can be considered to be general if it is successful only with helix bundles, especially if circular dichroism data are required to bias decision parameters to favor an all-helical structure. The first to use a "consensus classical" strategy in a fully *a priori* sense without supporting circular dichroism data were Kirschner and his colleagues.[27] The GOR method[64] was applied to individual sequences of the α domain of tryptophan synthase (Figure 11). A preliminary prediction used unbiased decision constants. After an α–β structure was inferred from the results, decision constants optimized for α/β proteins were used. The predictions were then averaged in a non-tree-weighted procedure to yield a consensus model.

A consensus Chou–Fasman[104] prediction was also obtained, as was a hydropathy index profile using the Kyte–Doolittle tool.[192] Finally, the average chain flexibility was predicted using the algorithm of Karplus and Schulz.[193] Significantly (see below), the prediction also used gaps in the sequence alignment to place breaks in secondary structure.

The results of these combined analyses suggested that tryptophan synthase folds to give an eight-fold α–β barrel, a class of protein well known in the database.[194] The crystal structure[191] showed this prediction to be correct, although with a noncore secondary structural element mispredicted and the final β strand shifted (Figure 11). Subsequent analysis suggested that the "consensus GOR" prediction method might be generally useful in predicting such barrels.[195] As the GOR program is parameterized on

```
a- NIHKHRILILDFGSQYTQLVARRVRELGVYCELWAW______DVTEAQIRDFNPSGIILSGGPESTTEENSPRAPQY
b- NIHYHKILILDFGSQYTQLIARRVREIGVYCELWAW______DVTEQQIREFAPTGIILSGSPESTTEENSPRAPEY
d-  -----ILIIDNYDSFTYNLVQYVGVLT___DVAVVKND___DDSLGNMAEK_ADALIFSPGPGWPADAGKMETLIQ
c-  ---KRVIVIDNYDSFVYNIVQYIGEVEPDCEIEVFRND___EITIEEIERKNPTHIVISPGPGRPEEAGISVDVVR
e-  -----ILLLDNVDSFTYNLVDQLRA__SGHQVVIYRNQIGAEVIIERLQHMEQPVLMLSPGPGTPSEAGCMPELLQ
        EEEEE      HHHHHHHHHH  H   EEEE       HHHHHHHHH   EEEE             HHHHH prediction
        EEEEEE    HHHHHHHHHHHH  EEEEEEE______  HHHHHHHHH  EEEEE              HHH experimental
          β1           α1          β2            α2         β3

a- VFEAGVPVFGVCYGMQTMAMQLGGHVEASNEREFG_YAQVEVVNDSALVRGIEDALTADGKPLLDVWMSHGDKVTA
b- VFNAGVPVLGICYGMQTMAMQLGGLTETSDHREFG_YASVSLENSTALFANLNDNLTAS_EPKLDVWMSHGDKVTR
c- QFAGKVPILGVCLGHQVIGYAFGGKIVHAKRILHGKTSKIVH_NGKGVFSGVKNPLVATRYHSLVV_____EEA_S
d- HFSGQKPILGICLGFQAIVEVFGGKLRLAHQVMHGKNSQVRQTSGNLIFNHLPSKFLVMRYHSIVM_____DEAVA
e- RLRGQLPIIGICLGHQAIVEAYGGQVGQAGEILHGKASAIAH_DGEGMFAGMANPLPVARYHSLV________GSN
   HHHH   EEEE     HHHHHHH      EEEE            HHHHHHH   EEEE                H prediction
   HHHH  EEEEEHHHHHHHHHHH  EEE     EEE_EEEEEEE           EEE   EEEEEEEEEE  EEEE experimental
    α3    β4       α4       β5          β6                β7       β8        β9

a- IPS_DFITVASTESCPFAIMANEEKRFYGVQFHPEVTHTRQGMRMLERFVRDICQCEALWTPAKIIDDAVARIREQV
b- LPE_NFKVTGTTLTCPIAAMSDESRRFYGVQFHPEVTHTKKGLELLTNFVVNICGCETKWTAENIIEDAVARIKEQV
c- LPEVLEITAKSDDGE_IMGLQHKEHPTFGVQFHPESVLTEEGKRIIKNFL---------------------------
d- LPDFAITAVATDDGE_IMAIENEKEQIYGLQFHPESIGTLDGMTMIENFV---------------------------
e- IP__ADLTVNARSGEMVMAVRDDRRRVCGFQFHPESILTTHGARLLEQTLWAWLAK---------------------
   HHHHHHHHHHH     EEEE      EEEE            HHHHHHHHHHH   end of prediction     prediction
         EEEEE      EEEEE   EEEEEE         HHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHHH experimental
          β10        β11      β12               α5                     α6
```

Figure 12. Representative sequences, *bona fide* consensus prediction,[197] and experimental[198] secondary structure for glutamine amidotransferase: (a) GMP synthase (glutamine-hydrolyzing) (AC=P04079, GUAA_ECOLI) *Escherichia coli*; (b) GMP synthase (glutamine-hydrolyzing) (AC=P44335, GUAA_HAEIN) *Haemophilus influenzae*; (c) anthranilate synthase component II (AC=Q08654, TRPG_THEMA) *Thermotoga maritima*; (d) anthranilate synthase component II (AC=Q02003; TRPG_LACLA) *Lactococcus lactis*; and (e) anthranilate synthase component II (AC=P00900,TRPG_SERMA) *Serratia marcescens*. Key: E, β strand; H, α helix.

a database containing many such folds,[64] this success is perhaps not surprising.

A parallel prediction was made for tryptophan synthase by Hurle *et al.*[196] These authors exploited circular dichroism data, which suggested that the protein adopted an $\alpha-\beta$ structure. They then applied a turn heuristic to a multiple alignment of eight homologous sequences. Secondary structure was assigned by using a pattern based method. The resulting secondary structural model was used to build a tertiary structural model. A biochemical experiment caused the predictors to exclude (incorrectly) a barrel structure in favor of a β-sheet structure. Otherwise, the prediction had the same merits as the prediction by Kirschner and his group.

Looking to extend this success, Niermann and Kirschner applied a similar analysis to the G-type glutamine amidotransferase family of proteins, and again detected an $\alpha-\beta$ pattern of secondary structure (Figure 12).[197] They then suggested that the predicted secondary structure was again compatible with an eight-fold $\alpha-\beta$ barrel topology. Here, the prediction method made several mistakes, as shown in Figure 12, which records the secondary structural assignments made on a similar domain in GMP synthetase.[198] Most notably, β strands 5, 6, 8, and 9 were missed, a helix between strands 6 and 7 was overpredicted, and strand 10 was mispredicted as a helix. As a result, what was a largely β domain in the experimental structure was mispredicted by the consensus GOR methods to be an $\alpha-\beta$ structure.

The consensus GOR has overpredicted $\alpha-\beta$ structures elsewhere. Poulter and his group used a consensus GOR method to predict the secondary structure of a family of enzymes that synthesize isoprenyl diphosphates, starting from a set of homologous protein sequences. Again, the consensus GOR analysis predicted a structure built from eight helices interrupted by four strands (Figure 13).[199] A subsequently determined crystal structure found a fully helical structure.[200] Helix 3 was mispredicted as a strand, while helix 9 was misassigned in part as a strand. Two shorter predicted helices were found in the experimental structure as one long helix, while one long predicted helix was assigned in the experimental structure as two shorter helices.

These three *bona fide* prediction results seem to confirm what is suggested anecdotally by retrodiction-based studies with known structures using the consensus GOR approach. Consensus GOR approaches appear to be biased in their predictions to favor $\alpha-\beta$ proteins. This bias may reflect the fact that such structures are richly represented in the database upon which the GOR tool is parameterized. Averaging over a set of homologous sequences evidently tends to amplify rather than eliminate this bias, leading to the prediction of $\alpha-\beta$ conformations even where they do not exist. Parameters may be deliberately altered to favor a structure that is suspected based on circular dichroism or other data (as was done with annexin, see above). However, consensus classical approaches were unable to identify any important secondary structure feature of the Src homology 3 domain (see below),[65] which adopts a fold that was underrepresented in the crystallographic databases at the time it was predicted.

This discussion is unfortunately clouded by a recent report that the GOR heuristic is not imple-

```
a- EKQDFVQHFSQIVRVLTEDEMGHPEIGDAIARLKEVLEYNAIGGKYNRGLTVVVAFRELVEPRKQDADSLQRAW
   |!::||..|.|||| ||||.:||||!|||!|||||||:||| |||.|||||||.|!|||..|...||:||. |.
b- EREEFVGFFPQIVRDLTEDGIGHPEVGDAVARLKEVLQYNAPGGKCNRGLTVVAAYRELSGPGQKDAESLRCAL
        HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH      EEEEEEEEEEEE               HH prediction
   HHHHHHHHHHHHHHHHHHHH      HHHHHHHHHHHHHH      HHHHHHHHHHHHH        HHHHHHH experimental

                  117
                   |
a- TVGWCVELLQAFFLVADDIMDSSLTRRGQTCWYQKPGVGLDAINDANLLEACIYRLLKLYCREQPYYLNLIELF
   .||||!||.||||||||||||.|||||||.|||.|.|||||||||. |||..!||.|| |||:.|||..|:|||
b- AVGWCIELFQAFFLVADDIMDQSLTRRGQLCWYKKEGVGLDAINDSFLLESSVYRVLKKYCRQRPYYVHLLELF
   HHHHHHHHHHHHHHHHHH                  HHHHHHHHHHHHHHHHHHHHH         HHHHHHHH prediction
   HHHHHHHHHHHHHHHHHHH                     HHHHHHHHHHHHHHHHHH        HHHHHHHH experimental


a- LQSSYQTEIGQTLDLLTAPQGNVDLVRFTEKRYKSIVKYKTAFYSFYLPIAAAMYMAGIDGEKEHANAKKILLE
   ||:.||||:||.|||:|||...|||..|:|.|||.||||||||||||||!||||||.|||...||.|||.||||
b- LQTAYQTELGQMLDLITAPVSKVDLSHFSEERYKAIVKYKTAFYSFYLPVAAAMYMVGIDSKEEHENAKAILLE
   HHHH      HHHHHH                          EEEEEHHHHHHHHHHH            HHHH prediction
   HHHHHHHHHHHHHHHHH             HHHHHHHHHHHHHHHHHHHHHHHHH       HHHHHHHHHHHH experimental


a- MGEFFQIQDDYLDLFGDPSVTGKIGTDIQDNKCSWLVVQCLQRATPEQYQILKENYGQKEAEKVARVKALYEEL
   |||!||||||||| ||||..|||!|||||||||||||||||||.||||.|:|.:|||.||.||||!||.|||..
b- MGEYFQIQDDYLDCFGDPALTGKVGTDIQDNKCSWLVVQCLQRVTPEQRQLLEDNYGRKEPEKVAKVKELYEAV
   HHHHEEEE                         EEEEEE              HHHHHHHHHHHHHHHHHHHHH prediction
   HHHHHHHHHHHHHH                     HHHHHHHHHHHHH             HHHHHHHHHHH   experimental

a- DLPAVFLQYEEDSYSHIMALIEQYAAPLPPAVFLGLARKIYKRRK
   .:.|.|.||||.||..:..|||.:...||..!|||||.|||||.|
b- GMRAAFQQYEESSYRRLQELIEKHSNRLPKEIFLGLAQKIYKRQK
   HHHHHHHH                                       prediction
     HHHHHHHHHHHHHHHHHHHHHH     HHHHHHHHHHH       experimental
```

Figure 13. Representative sequences, *bona fide* consensus prediction,[199] and experimental[200] secondary structure for farnesyl diphosphate synthase. The overprediction of an α−β structure is noteworthy: (a) farnesyl pyrophosphate synthetase (AC=P14324; ID=FPPS_HUMAN, (EC 2.5.1.10) *Homo sapiens*; and (b) farnesyl diphosphate synthase (PDB 1fps). Key: E, β strand; H, α helix; t, turn. An active-site residue at position 117 is indicated.

mented as originally described in the original papers, in at least some computer packages.[107] With different packages available in different versions, it has proven difficult to determine for any individual prediction exactly what implementation is used. Nor is it possible to learn the impact of the incorrect implementation on this discussion, short of repeating all of the predictions using an authorized implementation of the program. This has not yet been done.

By the time that the CASP1 project began, many variants of consensus classical methods were available. As CASP1 brought together predictions made by many methods, these are discussed in detail in section VI of this review.

E. Consensus Probabilistic Tools Combined with Consensus Physicochemical Methods

The next step in the development of the consensus classical heuristics involved coupling probabilistic and physicochemical tools to make joint predictions, but where a multiple sequence alignment is used as an input. This approach has now been successful in several instances within a *bona fide* prediction setting. For example, Bazan recently applied GOR methods to individual members of the cytokine receptor superfamily to obtain an average secondary structure prediction for the family (Figure 14).[201] Information concerning amphiphilicity and predicted β turns was then added. From this analysis, the cytokine receptor was proposed to be an all-β structure, with a folding topology similar to that found in immunoglobulin molecules. A subsequently determined crystal structure shows the close correspondence between the placement of the strands in the model and the position of the strands in the experimental structure (Figure 14), even though the tertiary structure proposed to assemble the β strands proved to be slightly different from that found experimentally.[202] In addition to being a powerful demonstration of the approach, the prediction shows the importance of expert involvement in a prediction exercise, in particular, an expert who knows something about the biochemistry of the target protein and uses what he/she knows while making the prediction.[91] This truism is now becoming more widely appreciated, even by workers in the area whose research is predominantly computational.[203]

F. Nontransparent Parameterized Methods To Predict Secondary Structure

Physicochemical analyses are transparent, as they are based on chemical principles that are relevant for protein and nonprotein molecules alike, and understandable to anyone trained in chemistry. Probabilistic methods are less so, as they are derived by parameterization processes that are not general to other classes of molecules, and may not be general to other types of proteins (for example, membrane

```
  Ile 128 \
 a - FSVDEIVQPDPPIALNWTLLNVSLTGIHADIQVRWEAPRNADIQKGWMVLEYELQYKEVNETKWKMMDPILTTS GHR_HUMAN
 b - FTVDEIVQPDPPIGLNWTLLNISLTGIRGDIQVSWQPPPNADVLKGWIILEYEIQYKEVNESKWKVMGPIWLTY GHRH_MOUSE
     * ***********.******* ***** .**** *  * ***  ***  **** ******* *** * **  *
prediction starts EEEEEE          EEEEEE              EEEEEEEE       EEEE       prediction
     E           EEEEEEE  EE   EEEEEEEEEE              EEEEEEEE      EEE     EE experimental


 a - VPVYSLKVDKEYEVRVRSKQRNSGNYGEFSEVLYVTLPQMSQF_TCEEDFYFPWLLIIIFGIFGLTVMLFVFLF GHR_HUMAN
 b - CPVYSLRMDKEHEVRVRSRQRSFEKYSEFSEVLRVIFPQTNILEACEEDIQFPWFLIIIFGIFGVAVMLFVVIF GHRH_MOUSE
      *****. ***.******.**    *.****** *  **    . ****  *** *********  *****  *
     EEEE      EEEEEEEEE            EEEEEE        prediction stops              prediction
     EEEEEEEE  EEEEEEEEE            EEE                                         experimental
```

Figure 14. Representative sequences, *bona fide* consensus prediction,[204] and experimental[202] secondary structure for the cytokine receptor family. The experimental structure is for the complex between human growth hormone and extracellular domain of its receptor: (a) growth hormone receptor GHR_HUMAN; and (b) growth hormone receptor GHRH_MOUSE (Ile 128 at the start of the domain is marked). Key: E, β strand; *, conserved amino acid.

proteins). Nevertheless, they gain a degree of transparency through analyses such as that above, which provide possible physicochemical reasons underlying the propensities.

In recent years, fully nontransparent methods have also emerged that exploit the fact that homologous protein sequences have similar conformations. These have been dominated by neural networks, suggested some time ago as tools for predicting the secondary structure of proteins.[205,206] A neural network is a computer construct that connects many nodes, each of which operates on data that comes to it from other nodes (or from the outside). The neural network is "trained", a process in which the weights of connections are adjusted on the basis of data so that the network generates a known output from input data in a "training set". In this manner, the neural network "learns" on the basis of examples and can then apply the rules that it has learned to new problems.

When applied to predicting secondary structure from single sequences, the first generation of neural networks gave little improvement over classical methods, at least as far as can be judged from classical scoring tools (see below).[207] Very recently, however, neural networks trained on multiple alignments have been shown to perform better.[19,208,209] Average, cross-validated three-state scores have been improved from 60% to 72% in retrodictive tests.[19,209] Again, the three-state scores do not reveal many important details of the retrodiction. It is conceivable that the modest improvement in the three-state score hides a dramatic improvement in performance concentrated in core secondary structural elements.

For example, an early report suggested that the Heidelberg neural network (the "PHD" tool) might be able to detect internal helices,[176] a type of secondary structural element that is at the core of a fold, and is often difficult to detect (see below). This suggestion arose from a retrodiction of a secondary structure for the protein kinase family of proteins. It was later noted that this retrodiction was not repeatable.[210] The reason for this remains unclear; it appears that in an early implementation of the PHD server, when a target sequence submitted to the network was a duplicate of a sequence already in the database, that sequence was counted twice, and the ability of neural network methods to identify internal helices has not yet been systematically explored.

Neural networks were first applied in a *bona fide* prediction setting in a project designed to compare transparent predictions, consensus classical predictions, and PHD predictions. The developers of the PHD tool had twice claimed that the neural network performed better than transparent methods. Both involved comparison of a *bona fide* prediction made transparently with a retrodiction made by PHD, however, which is not a fair comparative test of two methods.[176,211] Therefore, it was decided to allow all methods competing on equal grounds. The target, suggested by Professor Edgar Meyer (Texas A&M), was the family of proteins that includes the metallohemorrhagic proteinase from snake venom.[90] Experimental structures from two groups subsequently emerged.[212−215,219]

The results are shown in Figure 15. The three-state score Q_3 for the transparent prediction is 70% (Table 3), slightly higher than the consensus neural network prediction (66%) and much higher than the consensus GOR and Chou−Fasman predictions (Table 3). However, the differences between the predictions can be best seen by examining the misassignments. Of 202 positions in the alignment, the transparent prediction makes α-for-β misassignments at only two positions. The other predictions make considerably more. This is not because the transparent prediction made fewer α and β assignments overall; in fact, the transparent prediction makes the most. Rather, the transparent prediction made essentially no serious residue misassignments, while the neural net predictions did. Two of the three misassignments in helical regions would have been particularly problematic when assembling a tertiary structural model. Mistakes made in the transparent prediction are discussed below.

The PHD neural network has undergone revision subsequent to this test, and its output has improved. The first large-scale test of the PHD neural network in a *bona fide* prediction setting was done as part of the CASP1 project. As CASP1 brought together predictions made by many methods, these are discussed in detail in section VI of this review. An assessment of these predictions, both by the predictors themselves and by independent judges,[62] provides an overall view of the tool as applied in a *bona fide* prediction setting. The CASP1 predictions are discussed in greater detail below in the section that focuses on *bona fide* predictions. To illustrate the

```
                                               Sequence (Hemorrhagic Metalloproteinases)
    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    1
    0    1    1    2    2    3    3    4    4    5    5    6    6    7    7    8    8    9    9    0
    5    0    5    0    5    0    5    0    5    0    5    0    5    0    5    0    5    0    5    0
QNLPQRYIELVVVADHRVFMKYNSDLNTIRTRVHEIVNFINGFYRSLNIHVSLTDLEIWSNEDQINIQSASSDTLNAFAEWRETDLLNRKSHDNAQLLTAI  a
QNLPQRYIELVVVADRRVFMKYNSDLNIIRTRVHEIVNIINGFYRSLNIRVSLTDLEIWSGQDFITIQSSSSNTLNSFGEWRERVLLIWKRHDNAQLLTAI  b

                                                   Experimental Secondary Structures
     EEEEEEEEE HHHHHHTTT HHHHHHHHHHHHHHHHHHHGGGTEEEEEEEEEE SS  SS   SSHHHHHHHHHHHHHHTHHHH   SEEEEEE S  DSSP
     EEEEEEEEEHHHHHHHH HHHHHHHHHHHHHHHHHHHHH   EEEEEEEEEEEE          HHHHHHHHHHHHHHHHHH     EEEEEEE    a
     EEEEEEEEE HHHHHHTTT HHHHHHHHHHHHHHHHHHHGGGTEEEEEEEEEE SS  SS   S HHHHHHHHHHHHHHTTTTSS  SEEEEEE S  DSSP
     EEEEEEEEEHHHHHHHH  HHHHHHHHHHHHHHHHHHHH   EEEEEEEEEEEE          HHHHHHHHHHHHHHHHHH     EEEEEEE    b

                                                         Consensus Predictions
       EEEEEEE  HHHHHHHHHHHHHHH HHHHHHHHHHHHH   EEEEEEEEEE     EEEEEE  HHHHHHHHHHHHH          EEEEEEE  FL
      EEEEEEEEE  EEEEE    HHHHHHHHHHHHHHHHHH   EEEEEE EEEEEE   EEEEEE  HHHHHHHHHHHHHHHH      HHHHEEEE  RS
        EEE EHHHHHHTTTT     EEEEEEEEE    TTT   EEEEEEEE     TTT          EEE    HHHHHHHTTT     EEEEEE  GOR
    TTEEEEEEEETT EEEEETTT  EEEEEEEEEEEEEE     EEEEEEEEEEHHH TT EEEEE     EEEEE HHHHHHHH  T TTTHHHHHHH  CF


                                               Sequence (Hemorrhagic Metalloproteinases)
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    2
   0    1    1    2    2    3    3    4    4    5    5    6    6    7    7    8    8    9    9    0
   5    0    5    0    5    0    5    0    5    0    5    0    5    0    5    0    5    0    5    0
ELDEETLGLAPLGTMCDPKLSIGIVQDHSPINLLMGVTMAHELGHNLGMEHDGKDCLRGASLCIMRPGLTKGRSYEFSDDSMHYYERFLKQYKPQCILNKP a
NFEGKIIGKAYTSSMCNPRSSVGIVKDHSPINLLVAVTMAHELGHNLGMEHDGKDCLRGASLCIMRPGLTPGRSYEFSDDSMGYYQKFLNQYKPQCILNKP b

                                                   Experimental Secondary Structures
  TTT  EE  TT TT TTTSEEEEE   S HHHHHHHHHHHHHHHTT     TT EETTEEETT SS    S EE  HHHHHHHHHHHHHH  STTS    DSSP
      EEEEE         EEEEEEE   HHHHHHHHHHHHHHHH                    EEEEE      HHHHHHHHHHHHHHHH         a
  GGG  EEE SS TT TTTSEEEEE   SSHHHHHHHHHHHHHHHTT     TT EETTEEETTSSS    S EE  HHHHHHHHHHHHHH          DSSP
      EEEEE         EEEEEEE   HHHHHHHHHHHHHHHHHH                  EEEEE   EEEHHHHHHHHHHHHHHHH         b

                                                         Consensus Predictions
EEE  ??????         EEEEEE     HHHHHHHHHHHHH  EEEEE           EEEE EEEE HHHHHHHHHHHHHHHHHHH   EEEEEE  FL
E          E         EEEEE    EEEEEEEEEEHHH                  EEEEE         EE    HHHHHEEEE    EEEE    RS
      EEEEETTTT  TTT EEEEE       HHHHHHHHHHHH  HHHHHH TTTTTT  EEE        TTTTTTTTTTT EEETTTT TTEEEE   GOR
  TTT     TE  TT TTT EEEEETTT EEEEEHHHHHHHH     HHHHH  TTT E  EEEE    TTTT   TTTTTTEEEEEEETT  EEEEEE  CF
```

Figure 15. Representative sequences, experimental secondary structures, and *bona fide* consensus predictions[90] for the hemorrhagic metalloproteinase family. Key: E, β strand; H, α helix; T, turn; G, 3_{10} helix; B, β bridge; S, bend. Lines designated as follows: (a) Atrolysin;[213] (b) Adamalysin;[212] Z, prediction made by transparent method; RS, prediction made by PHD server; GOR, consensus GOR prediction; and CF, consensus Chou−Fasman prediction, as implemented in the GCG package.

Table 3. Summary of the Results of the Prediction Contest for Hemorrhagic Metalloproteinase[90,214,a]

	three-state residue score, %	no. of assignments (total): $\alpha + \beta$	no. of correct assignments: $\alpha + \beta$	no. of seriously incorrect assignments: α vs β
Florida	69.8	131	97	5
Heidelberg neural network (RS)	63.8	114	70	24
GOR	54.9	81	51	16
Chou−Fasman	45.0	122	38	43

[a] Three-state residue scores are calculated by dividing the number of correct assignments ($\alpha + \beta +$ coil) by the total length of the alignment, following the classical scoring paradigm. A seriously incorrect assignment is one where a residue in a helix in the experimental structure is predicted to be in a strand, or vice versa. Figure 15 should be inspected to obtain a more comprehensive view of the quality of the predictions. Slightly different values are obtained when using different experimental structures.[215]

application of the PHD tool, we discuss here briefly the prediction for urease generated by Hubbard and Park using the PHD neural network server.[216]

Urease has three subunits.[217] Hubbard and Park made predictions for the β and γ subunits. The γ subunit is largely helical, while the β subunit is largely strand. The PHD program produced an essentially perfect prediction for the γ domain (Figure 16), although evidently after some manual adjustment of the multiple alignment that it produced.[216] The prediction for the β domain missed only one of the core strands, assigning it as part of a long helix. Thus, this prediction can be judged as being very good.

In the CASP2 project (see below), a neural network developed by Rost and his co-workers performed well, both as applied by Rost (21 predictions, mean Q_3 score of 74, with 13 predictions having a $Q_3 > 68\%$), or as applied by others (for example, Flohil, de Hoop, and Freitman, with a mean Q_3 score of 71, with 12 predictions having a $Q_3 > 68\%$). Similar scores were obtained by the method of Solovyev and Salamov,[81] and by the method of King and Sternberg.[106] These are reviewed elsewhere[130,174] and in greater detail below.

V. Models for Molecular Evolution and Their Role in Structure Prediction

To this point in this review, three ways evolutionary information might be used to assist protein structure prediction have been discussed. First, evolutionary information may identify a reference protein having a known structure as a homolog of the target protein. Second, evolutionary information may be used to average single predictions made classically, in the hope of filtering out noise. Last, a set of homologous proteins might be used to train a neural network, with the additional information exploited in a way hidden within the network.

```
Beta  domain
MIPGEYHVKPGQIALNTGRATCRVVVENHGDRPIQVGSHYHFAEVNPALKFDRQQAAGYR
    EEEE    EEEE    EEEEEEE     EEEE   EEHHHHHHHHHH HHHHH  E  Hubbard
EE         EEEEEE     EEEEEEEEE      EEEEE    EEEE      EEEE  Matsuo
     EE    EEE      EEEEEEEE     EEEE           EE        EE  experimental DSSP

LNIPAGTAVRFEPGQKREVELVAFAGHRAVFGFRGEVMGPLEVNDE
E     EEEEE     EEEEEEEE   EEEEEE  EE                         Hubbard
EEE  EEEEE    EEEEEE     EEEEEE      EEEEEE                   Matsuo
E      EEEE    EEEEEEEE     EE        EEE                     experimental DSSP

Gamma  domain
MELTPREKDKLLLFTAALVAERRLARGLKLNYPESVALISAFIMEGARDGKSVASLMEEG
    HHHHHHHHHHHHHHHHHHHHHH      HHHHHHHHHHHHHHH     HHHHHHHH   Hubbard
EEEE        EEEE EE    EE  EEE     EEE HHHHHHHH      EEE       Matsuo
    HHHHHHHHHHHHHHHHHHHHHH     HHHHHHHHHHHHHHHHHH    HHHHHH    experimental DSSP

RHVLTREQVMEGVPEMIPDIQVEATFPDGSKLVTVHNPII
HHHEE  HHHHHHHHHHH  EEEEEE    EEEEEE                           Hubbard
EEEE        E  EE HHHH  E      EEEE                            Matsuo
            HHHH  EEEEEEEE  EEEEEEEE                           experimental DSSP
```

Figure 16. Predicted[216] and experimental[217] structures for urease from *Klebsiella aerogenes* (P18314, 1kau). The predicted structures were submitted for the CASP1 prediction project.[148] The prediction of Hubbard was built using the PHD neural network server.[218] The prediction of Matsuo was based on threading to macromomycin (2mcm) for the β domain and to endathiapepsin (PDB 2ert) for the gamma domain. Key: E, β strand; H, α helix.

None of these approaches considers explicitly the underlying processes by which proteins themselves diverge under functional constraints and how an understanding of these processes might be used to design prediction tools. The explosion in the size of the protein sequence database made possible a detailed study of these processes.[220] These studies have identified a different general approach for using homologous protein sequences to make structure predictions. The primary advantage of the approach is that it is quite transparent. A prediction for protein conformation can be analyzed just as a conformational analysis can be done with smaller molecules. The approach has been used to make over two dozen predictions to date, many of which have been remarkably accurate. Further, the mistakes made in these predictions have been instructive, and much has been learned both about protein folding and methods for making predictions as a result.

A. Understanding the Details of Molecular Evolution

1. The Alignment

To have a transparent view of evolutionary analysis as a tool for making secondary structure predictions, we must begin by understanding the key element of an evolutionary analysis: the sequence alignment.[221,222] As noted above, an alignment attempts to represent the evolutionary relationship between two protein sequences by placing them side-by-side so that codons encoding amino acids paired in an alignment have arisen from a single codon in a single ancestral gene, at least with the highest probability. An example of an alignment of two protein sequences, here chosen from two homologous protein kinases, is given in Figure 17. Let us ask how this alignment was constructed and what is shows.

An alignment shows what amino acid substitutions have been accepted since two proteins diverged from their common ancestor. These substitutions are not random if the descendent proteins have served functions in the descendent organisms (that is, assuming that the proteins have "diverged under functional constraints"). Most proteins have a function that contributes to the ability of their host organism to survive, select a mate, and reproduce. To perform this function, proteins adopt a fold, or tertiary structure, a structure that is conserved much more highly than the sequence itself.

```
DLYTYLSRRLNPLGRPQIAAVSRQLLSAVDYIHRQGIIHRD
||! !!!!|   |    ! !!  |!| ||!! |! |!!|||
DLFDFITERGA-LQEDLARGFFWQVLEAVRHCHNCGVLHRD

1    1    1    1    1    1    1    1    1
1    2    2    3    3    4    4    5    5
5    0    5    0    5    0    5    0    5
```

Figure 17. Part of an alignment of two protein kinase sequences, used in the text to illustrate how transparent tools for predicting elements of tertiary and secondary structure work. A vertical line (|) indicates an identical match in the alignment. An exclamation point (!) indicates a mutation with high probability.

Function therefore constrains what amino acid substitutions are accepted during divergent evolution; some substitutions are never observed because they are lethal to the host organisms. Other substitutions help the protein perform its selective function (positive, or adaptive substitutions) and will be incorporated at a high rate, especially when a new function is emerging. Still other substitutions represent neutral drift in the structure,[223,224] having no selectable impact on the fitness of the protein.

In principle, an alignment is "correct" if it correctly represents actual events in the historical past; a correct alignment matches amino acid codons that are descendent from a single codon in an ancestral protein, correctly reconstructs ancestral sequences, and indicates substitutions, insertions, and deletions as they actually occurred during historical evolution. Proving that an alignment is correct is essentially impossible, of course. In some cases, the ancestral genes have been synthesized, in part to test this premise.[36-39] In general, however, the accuracy of an alignment is judged by a score that represents the probability that an alignment has done what it should do.

C	11.5																			
S	0.1	2.2																		
T	-0.5	1.5	2.5																	
P	-3.1	0.4	0.1	7.6																
A	0.5	1.1	0.6	0.3	2.4															
G	-2.0	0.4	-1.1	-1.6	0.5	6.6														
N	-1.8	0.9	0.5	-0.9	-0.3	0.4	3.8													
D	-3.2	0.5	0.0	-0.7	-0.3	0.1	2.2	4.7												
E	-3.0	0.2	-0.1	-0.5	0.0	-0.8	0.9	2.7	3.6											
Q	-2.4	0.2	0.0	-0.2	-0.2	-1.0	0.7	0.9	1.7	2.7										
H	-1.3	-0.2	-0.3	-1.1	-0.8	-1.4	1.2	0.4	0.4	1.2	6.0									
R	-2.2	-0.2	-0.2	-0.9	-0.6	-1.0	0.3	-0.3	0.4	1.5	0.6	4.7								
K	-2.8	0.1	0.1	-0.6	-0.4	-1.1	0.8	0.5	1.2	1.5	0.6	2.7	3.2							
M	-0.9	-1.4	-0.6	-2.4	-0.7	-3.5	-2.2	-3.0	-2.0	-1.0	-1.3	-1.7	-1.4	4.3						
I	-1.1	-1.8	-0.6	-2.6	-0.8	-4.5	-2.8	-3.8	-2.7	-1.9	-2.2	-2.4	-2.1	2.5	4.0					
L	-1.5	-2.1	-1.3	-2.3	-1.2	-4.4	-3.0	-4.0	-2.8	-1.6	-1.9	-2.2	-2.1	2.8	2.8	4.0				
V	0.0	-1.0	0.0	-1.8	0.1	-3.3	-2.2	-2.9	-1.9	-1.5	-2.0	-2.0	-1.7	1.6	3.1	1.8	3.4			
F	-0.8	-2.8	-2.2	-3.8	-2.3	-5.2	-3.1	-4.5	-3.9	-2.6	-0.1	-3.2	-3.3	1.6	1.0	2.0	0.1	7.0		
Y	-0.5	-1.9	-1.9	-3.1	-2.2	-4.0	-1.4	-2.8	-2.7	-1.7	2.2	-1.8	-2.1	-0.2	-0.7	0.0	-1.1	5.1	7.8	
W	-1.0	-3.3	-3.5	-5.0	-3.6	-4.0	-3.6	-5.2	-4.3	-2.7	-0.8	-1.6	-3.5	-1.0	-1.8	-0.7	-2.6	3.6	4.1	14.2
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Figure 18. A "log odds" scoring matrix, which reports 10 times the common logarithm of the probability of two amino acids being matched in a pairwise alignment by reason of common ancestry divided by the probability that these are matched by random chance.[220] This matrix is optimized to align protein pairs ~150 PAM units apart.

The basic element of score is the probability that the proteins whose sequences are being aligned are in fact related by common ancestry. This score is often expressed as logarithm of the probability that the similarities in the two sequences seen in the alignment arose by reason of common ancestry, divided by the probability that these similarities arose by random chance. This probability is generally obtained by comparing the aligned sequences one position at a time. Under this procedure, a score is first given to each pair of amino acids matched in the alignment.[225] This pairwise score is the logarithm of a probability that the two amino acids will be paired in a protein by reason of common ancestry, divided by the probability that they would be paired by random chance. This probability is derived from one of the many "log odds" matrices that provide pairwise probabilities for the 210 possible amino acid pairs (Figure 18).[226] Gaps in the alignment are penalized, the pairwise terms are summed for the entire alignment, and the resulting score reported. The evolutionary distance between the two sequences is then measured in PAM units,[225] the number of point accepted mutations that the two protein sequences have suffered (per 100 amino acids) since they diverged an unspecified number of years ago.

Underlying these processes for constructing and evaluating an alignment is a model for the way amino acids undergo substitution during divergent evolution.[224] The model is "first-order" Markovian in that it assumes that subsequent amino acid substitutions in a protein occur with a probability independent of previous substitutions, that substitutions occur independently at different positions in the polypeptide chain, and that a single substitution matrix can represent the probability of amino acid substitution at any and all positions in a protein.

This model is, of course, an approximation. Real proteins adopt three-dimensional conformations where amino acids distant in the sequence come in contact and therefore interact. Thus, residues in a protein sequence need not undergo substitution independent of substitution at other positions in the protein. Likewise, biological function constrains the types of amino acid substitutions that are acceptable to natural selection. Therefore, amino acids need not suffer mutation independently, either in sequence or in time. The Markov model should fail when applied to real proteins.

This failure, of course, contains information about the "nonlinear" part of protein structure, that is, conformation. Accordingly, non-Markovian behavior during the divergent evolution of protein sequences can be sought as a source of information for predicting protein conformation.

For example, it has been well recognized that amino acids near an active site are more conserved than expected under the Markov model.[21,26] Conversely, positions on the surface of a folded structure tolerate more variation than positions inside.[15,26,227−233] Thus, it is widely assumed that if an amino acid carrying a functional group is conserved over a wide PAM distance, it lies at or near an active site.

With the growth in the protein sequence databases and improvements in tools for organizing them,[220] systematic studies have been made to identify features of divergent sequence evolution where the Markov model fails and to use these failures systematically to develop heuristics for predicting protein structure.[15,91,234] In brief, the approach allows one to do a residue-by-residue analysis that does not assume that local sequence determines local conformation. This insight has allowed the development of an important class of transparent structure prediction tools.[91,234]

2. Understanding Divergent Evolution: Substitution Matrices

To extract conformational information from non-Markovian behavior in protein sequences undergoing divergent evolution, we must first learn to identify and understand the behavior expected from the Markovian model. Central to the first-order Markovian model is a matrix describing the probabilities of each amino acid undergoing replacement by each of the 19 other proteinogenic amino acids. These symmetrical 20 × 20 matrices have indices that are the 20 amino acids, and elements that are the logarithms of probabilities that the index amino acids will be paired in an alignment divided by the probability that the pairing would occur by chance.[225] Thus, the diagonal elements of the matrix represent the probabilities that the indexed amino acid will be conserved (i.e., that the amino acid will be matched against itself), while the off-diagonal elements represent the probabilities that an amino acid will be replaced by one of the other 19 amino acids (Figure 18).

A scoring matrix is defined for a specific PAM distance. This is most easily seen by considering the diagonal and off-diagonal terms. In a matrix describing the alignment of two closely related proteins, the diagonal terms are large relative to the off-diagonal terms; more amino acids have been conserved than have been replaced. In contrast, in a matrix describing two distantly related proteins, the diagonal terms are small relative to the off-diagonal terms; many more amino acids have been replaced than have been conserved. Indeed, the PAM distance between two protein sequences is the PAM distance of the scoring matrix that best describes the pairing. Thus, the score of an alignment of the sequences of two proteins that have diverged by one point accepted mutation per 100 amino acids is highest when the alignment is scored using the 1 PAM scoring matrix. The alignment of two sequences that have diverged by 10 PAM units receives the highest score with the 10 PAM scoring matrix.

These scoring matrices can, of course, be constructed directly from empirical data. To do this, a statistically large collection of pairwise alignments must be collected for protein pairs that have diverged (for example) 1 PAM unit. From these, the number of times each of the 210 possible pairings occurs in the alignments must be tabulated, and normalized to give logarithms of probabilities. To get a scoring matrix appropriate for proteins that have diverged 10 PAM units, the process must be repeated, but with protein pairs that have diverged by 10 PAM units.

This is not how the matrices have generally been calculated, however. Under the Markov assumption, subsequent amino acid substitutions are independent of earlier substitutions. If this assumption is correct, the matrix describing an alignment of two protein sequences 10 PAM units distant can be obtained by multiplying the 1 PAM matrix by itself 10 times. This process (raising the 1 PAM matrix to the 10th power) is equivalent (given the Markovian assumption) to evolving a protein sequence through 10 successive evolutionary steps, each 1 PAM unit in length. This process assumes that substitutions occurring at the nth step occur independently of the substitutions in the $(n - 1)$th step.

In the original work of Dayhoff,[225] a scoring matrix applicable for proteins 250 PAM units distant was calculated this way. Empirical substitution data were collected from alignment pairs of proteins only 5−10 PAM units distant. A matrix containing the logarithm of the probability of each amino acid being replaced by each of the others in these similar pairs of proteins was then constructed, normalized by the probability that each substitution would occur by random chance. The PAM 250 matrix was then obtained by multiplying the PAM 5−10 matrix by itself the requisite number of times, a process that assumes that subsequent mutations follow the same pattern as earlier mutations.

Table 4. Ten Times the Logarithm of the Probabilities That the Indicated Amino Acids Will Be Matched in a Pairwise Alignment at the Indicated Evolutionary Distance[a]

evolutionary distance	probability of Trp−Arg pairing	probability of Trp−Phe pairing
5.5	1.5	−3.9
10.2	0.5	−0.9
42.5	−1.3	1.3
86.5	−1.8	3.0

[a] Evolutionary distances are measured in PAM units, the number of point accepted mutations separating two sequences per 100 amino acids. Ten times the logarithm of the probability is reported.

Extrapolating from PAM 5−10 to PAM 250 is substantial and requires that the Markov model for amino acid substitution be valid over a considerable evolutionary distance. We can, of course, test this assumption by comparing a 250 PAM scoring matrix obtained by normalizing data collected from protein aligned protein pairs 5−10 PAM units with a 250 PAM matrix obtained by normalizing data collected from protein pairs at longer evolutionary distances. To the extent that these matrices are the same, the Markov assumption that future and past substitutions are independent holds. To the extent that they are different, the differences measure the extent to which amino acid substitutions in real proteins deviate from the pattern predicted by the Markov model.

This comparison has in fact been made, and the deviation is large.[235] Consider just two possible replacements for the amino acid Trp (Table 4). In proteins that have diverged only slightly, replacement by Arg is probable (the logarithm of the probability of the pairing is positive), while replacement by Phe is improbable (the logarithm of the probability of pairing is negative). This empirical fact is chemically counterintuitive. The physical chemistry of Arg, which has a positively charged side chain, is quite different from that of Trp (which has a large hydrophobic aromatic side chain). Arg would not be expected to be a good replacement for Trp to maintain folding and function. In contrast, the physical chemistry of Phe is similar to that of Trp; both have aromatic rings in their side chain. Therefore, natural selection should tolerate a Phe-for-Trp substitution frequently.

Only at high evolutionary distances does the chemically more reasonable substitution of Trp by Phe (which conserves the physicochemical properties of the side chain) become probable, and the chemically unreasonable substitution of Trp by Arg become improbable.

Why are the physical chemical properties of the Trp, Arg, and Phe side chains reflected in amino acid substitutions only after long evolutionary distance? The genetic code provides a possible explanation. At short evolutionary distances, enough time has elapsed to change only a single base in the triplet codon. For the Trp codon (UGG), nine codons arise by single point mutation (AGG, CGG, GGG, UAG, UCG, UUG, UGA, UGC, and UGU). Two of these encode Arg (AGG and CGG); none encode Phe. Thus, it appears that at low evolutionary distances, the genetic code

```
                      s
DLYTYLSRRLNPLGRPQIAAVSRQLLSAVDYIHRQGIIHRD
||! !!!!|   |    ! !!  |!| ||!! |! |!!|||
DLFDFITERGA-LQEDLARGFFWQVLEAVRHCHNCGVLHRD

1    1    1    1    1    1    1    1    1
1    2    2    3    3    4    4    5    5
5    0    5    0    5    0    5    0    5
```

Figure 19. Part of an alignment of two protein kinase sequences, with an assignment of a single underlined position in the protein to the surface of the folded structure to reflect the code-driven substitution of an Arg by a Trp. A vertical line (|) indicates an identical match in the alignment. An exclamation point (!) indicates a mutation with high probability.

constrains amino acid substitution to enforce substitutions that do not conserve the chemical properties of the amino acid side chains. Examination of all of the elements of the substitution matrix shows that this conclusion is general for other pairs of amino acids.[150]

Trivially, the genetic code should influence amino acid substitution. One does not expect, however, that the code will influence *accepted* amino acid substitution, substitution that does not compromise the ability of the protein to contribute to survival and reproduction. Remembering that a substitution must be accepted by natural selection before it can appear in a database, code-driven substitutions, especially those that do not conserve physical chemical properties, are consistent with continued biological function when they occur on the surface of the folded protein. Thus, if a Trp−Arg pairing (for example) is observed in an alignment, the position containing it can be assigned to the surface of the folded structure. The fragment of the alignment of protein kinase shown in Figure 19 contains a Trp−Arg pairing. Therefore, we conjecture that this position lies on the surface of the folded structure.

3. Adjacent Covariation

By assuming that any substitution at position i in a protein sequence is independent of the substitution at position j, the Markov model also assumes that adjacent amino acids undergo independent substitution. This is true only as an approximation. Enough sequence data are now available to generate a dipeptide substitution matrix showing the probabilities for each of the 380 possible dipeptides to be substituted by each of the 380 possible dipeptides, normalized by the probabilities expected if adjacent positions undergo independent substitution.[235]

Again, substitution in real proteins deviates from that expected from the Markov model. In particular, if residue i is conserved, then the adjacent residue $i + 1$ is in general more likely to be conserved. Conversely, if residue i is variable, then residue $i + 1$ is more likely to be variable (Table 5). This empirical observation is a violation of the Markovian assumption that substitutions occur independently at adjacent positions in a protein sequence, but is not unexpected from standard models of protein structure. If residue i lies on the surface of the globular structure, it is likely that residue $i + 1$ also lies on the surface. If residue i lies inside, then residue $i + 1$ is also likely to lie inside. Residues inside the folded structure of a protein are more likely to be conserved; residues on the surface are less likely to be conserved. The empirically observed breakdown of the Markov model is expected.

Table 5. Correlation between Conservation and Variation at Adjacent Positions in a Protein Sequence[a]

conserved amino acid	10 log(probability that adjacent residue is conserved) − 10 log(probability that adjacent residue is not conserved)
Pro	−12.5
Gly	−3.9
Glu	−2.1
Lys	0.0
Asp	0.6
Ser	1.2
Leu	1.5
Ala	1.5
Asn	3.8
Arg	4.8
Gln	5.0
Thr	5.4
Phe	5.7
Ile	7.1
Tyr	8.0
Val	8.3
Cys	8.5
Trp	10.5
His	16.3
Met	16.8

[a] Values represent 10 times the logarithm of the probability that the amino acid adjacent to the conserved amino acid will also be conserved minus 10 times the logarithm of the probability that the adjacent amino acid will not be conserved.

```
                      s          turn
DLYTYLSRRLNPLGRPQIAAVSRQLLSAVDYIHRQGIIHRD
||! !!!!|   |    ! !!  |!| ||!! |! |!!|||
DLFDFITERGA-LQEDLARGFFWQVLEAVRHCHNCGVLHRD

1    1    1    1    1    1    1    1    1
1    2    2    3    3    4    4    5    5
5    0    5    0    5    0    5    0    5
```

Figure 20. Assignment of a turn in the alignment of two protein kinases. A vertical line (|) indicates an identical match in the alignment. An exclamation point (!) indicates a mutation with high probability.

Surprising, however, are the exceptions to the generalization (Table 5). If Pro or Gly is conserved at position i, then position $i + 1$ is more likely to have undergone *variation*. A structural conjecture might explain these exceptions. If a Pro or Gly is conserved when it induces a turn in the folded structure of the protein, and if turns generally occur on the surface of a folded structure,[236] a conserved Pro or Gly is likely to be adjacent to a surface position, which in turn is more likely to tolerate amino acid substitution. Each of these steps implies deviation from patterns of amino acid substitution expected from the Markov model, deviations that can be detected in analyzing sequence alignments and used to predict conformation in a polypeptide chain. For example, the fragment of the alignment of protein kinase contains a conserved Gly adjacent to a substituted position that might lie on the surface, and we might conjecture that the polypeptide chain turns at this point in the sequence (Figure 20).

4. Gaps in an Alignment

During divergent evolution, portions of genes may be added (inserted) or removed (deleted). This results in homologous proteins that contain different numbers of amino acids. This implies, in turn, that an alignment of sequences within a family of proteins where insertions and deletions ("indels") have taken place will have unmatched amino acids, which form "gaps" in the alignment. In an alignment of just two homologous sequences, it is impossible to tell whether the gap arose from an insertion event in the lineage leading to the protein with additional amino acids (implying that the ancestral protein had fewer amino acids), or whether the gap arose from an deletion event that removed amino acids from the ancestral sequence in the lineage leading to the protein with fewer amino acids. Therefore, the term "indel", a contraction of "insertion" and "deletion", has been adopted to refer to evolutionary events that place gaps in an alignment.

The placement of gaps is a critical step when constructing an alignment, and considerable research has been devoted toward understanding how gaps should be placed.[89,237] In practice, one does not know which amino acids have been inserted/deleted. Gaps are placed to optimize a score associated with an alignment. But if gaps are introduced without limit, even two random sequences can be aligned to give a perfect score. Therefore, gaps must be penalized to enforce their judicious use. The most common scheme for penalizing gaps charges a price for introducing a gap, and an incremental price for each additional amino acid that is added to the gap. This scheme is conveniently incorporated into the dynamic programming tools that implement the Markov model for scoring amino acid alignments using substitution matrices[238,239] and implies that the probability of a gap decreases exponentially with its length.

Analysis of real proteins shows that the probability of a gap does not decrease exponentially with its length.[237] Rather, the probability of a gap in a pairwise alignment is inversely proportional to its length raised to the 1.7 power.[89] The structural basis for this empirical relationship is unknown, but some hypotheses can be formulated to explain it. We may assume that a polypeptide paired with a gap forms a coil, that the ends of inserted or deleted segments lie close in space, and that the laws governing the conformation of free coils are followed by coils in a polypeptide chain. The probability that the two ends of a coil lie together in three dimensions is inversely proportional to the mean volume occupied by the coil. For a linear, unidimensional polymer, volume is proportional to the length of the polymer chain raised to the 1.5 power.[240] Thus, the probability that the two ends of a polypeptide will be near in space (and therefore that the peptide segment can be deleted without major change in the overall fold of the protein) is inversely proportional to the length of the polypeptide chain raised to the 1.5 power. From this, the probability of a gap of length k occurring in a pairwise alignment varies with $k^{-1.5}$ follows.

Real polypeptides are not, of course, idealized unidimensional polymers. Rather, the polypeptide chain itself fills a volume. This excluded volume raises the exponent in the formula relating volume to length. This exponent is experimentally measurable, and depends to some extent on the composition of the polymer. For a typical polypeptide, however, the volume of a random coil is a function of length raised to the 1.7−1.8 power.[241] This exponent is remarkably close to that needed to explain the empirical gap−length distribution in terms of the hypotheses outlined above.

```
         coil         s          turn
DLYTYLSRRLNPLGRPQIAAVSRQLLSAVDYIHRQGIIHRD
||! !!!!|   |    ! !!  |!| ||!! |! |!!|||
DLFDFITERGA-LQEDLARGFFWQVLEAVRHCHNCGVLHRD


1    1    1    1    1    1    1    1    1
1    2    2    3    3    4    4    5    5
5    0    5    0    5    0    5    0    5
```

Figure 21. Assignment of a coil in a gapped segment in the alignment of two protein kinase sequences. A vertical line (|) indicates an identical match in the alignment. An exclamation point (!) indicates a mutation with high probability. An indel (insertion or deletion) is indicated by a dash (−).

If these hypotheses are true, gaps can convey structural information. Whenever a gap is found, we can assume that it indicates a "parse", a point in the polypeptide chain where secondary structure is broken.[27] The fragment of the alignment of protein kinase that we have been discussing itself contains a gap (Figure 21). On the basis of this hypothesis, we might conjecture that secondary structure preceding this gap is independent of secondary structure that follows.

5. Understanding the Behavior of Coils: Parsing Strings

As discussed in greater detail below, much of the success of transparent tools for predicting helices and strands arises from tools that predict regions that are *not* helices or strands. Parsing tools divide a protein sequence into segments that form standard secondary structure independently. By parsing a sequence, secondary structure predictions need consider at any one time only short segments of the polypeptide chain, which is intrinsically easier than considering the polypeptide chain as a whole. Thus, understanding the evolution of loops is an important step toward developing tools for predicting secondary structure in proteins.

As discussed above, many polypeptide segments adopt different secondary structures when embedded in different tertiary structural contexts. Fortunately, this does not appear to be the case for many sequences involved in loops. Strings (consecutive positions in a polypeptide chain) of Pro, Gly, Asp, Asn, or Ser prove to be good indicators of a break, or parse, in standard secondary structural elements.[104] In general, a longer parsing string is more reliable than a shorter parsing string, and a string containing more prolines is better than one containing fewer prolines. Thus, a single Pro in a sequence is not a reliable indicator of a parse. However, a Pro-Gly sequence nearly always indicates a parse, while a Gly-Ser-Asn-Ser sequence nearly always does as well.[79]

A large number of parsing strings have been identified, especially those that combine information

Table 6. Accuracy of Surface Assignments made with and without Concurrent Variation[a]

	variation observed in	
	one subbranch, %	more than one subbranch, %
aspartate aminotransferase	82	93
alcohol dehydrogenase	69	86
lactate dehydrogenase	78	86
myoglobin	85	91
plastocyanin	91	100
phospholipase A	74	79
Cu/Zn superoxide dismutase	81	98
average	80	90

[a] In protein families diverging up to PAM 200.

concerning the position of surface residues (see below). Four consecutive surface residues indicate a parse with high reliability.[242] Parsing heuristics based on strings are available through a server accessible on the World Wide Web (URL http://cbrg.inf.ethz.ch) and have been used in making the transparent predictions described below.

6. Neutral vs Adaptive Variation

To this point, three pieces of tertiary structural information have been collected regarding the protein kinase fragments aligned in Figure 17 using a transparent evolutionary analysis of homologous protein sequences. At three points, the segment is near the surface of the fold, at positions 126 and 149 because turns and breaks in secondary structure are generally on the surface, and at position 137 because of the code-driven Trp-Arg substitution at this position. This is tertiary structural information, because it relates the positions of these residues in three dimensions to the overall fold. It is, however, only a limited amount of tertiary structural information.

To get more information, we might exploit other deviations in the Markov pattern of amino acid substitutions. In particular, the well-known fact that surface positions on a protein generally tolerate more variation than positions inside[227−230] suggests a simple heuristic for assigning surface positions. In this heuristic, positions in an alignment that contain one or more "surface-indicating" amino acids (for example, Lys, Arg, Glu, Asp, or Asn) and that are variable, in particular, at low PAM distance, are assigned to the surface.[234]

This heuristic is disappointing in its accuracy (Table 6).[234] On average, only 80% of the surface assignments made using this heuristic are correct. In some proteins (e.g., alcohol dehydrogenase), the accuracy is as low as 69%. Considering that approximately 50% of the side chains of a typical protein of this size lie on the surface of the folded structure, this performance is not impressive.

Why is the performance so bad? Here, conjectures concerning mechanisms of divergent evolution are suggestive. Two types of variation occur as protein sequences divergently evolve. Neutral variation involves substitutions that do not influence the ability of an organism to survive and reproduce.[223,244] These are variations that have little impact on behavior in a protein. From a structural viewpoint, such variations should lie predominantly on the surface of the folded structure. Thus, neutral variation is sought when attempting to identify surface positions by seeking variation in an alignment.

```
          coil                     turn
DLYTYLSRRLNP--LGRPQIAAVSRQLLSAVDYIHRQGIIHRDIK
||! !!!!|     |    ! !!  |!| ||!! |! |!!|||
DLFDFITERGA---LQEDLARGFFWQVLEAVRHCHNCGVLHRDIK
          parse                    parse

SIISiIssSSSS  ISisSIssIIssiISIIsII sSSIiI  I
1    1    1    1    1    1    1    1    1
1    2    2    3    3    4    4    5    5
5    0    5    0    5    0    5    0    5
```

Figure 22. Protein kinase fragment with complete surface and interior assignments. S and s indicate strong and weak surface assignments, respectively. I and i indicate strong and weak interior assignments, respectively. A vertical line (|) indicates an identical match in the alignment. An exclamation point (!) indicates a mutation with high probability. The gap in the alignment is indicated by a dash (−).

Adaptive substitutions accumulate as well during divergent evolution, however. Adaptive substitutions alter the behavior of the protein, often to make it better suited for a new environment or a new function. Mutations that alter function or create new function are the opposite, structurally, of mutations that do not influence function, and adaptive variation need not lie on the surface of a protein. Indeed, it may lie near an active site, a regulatory site, or inside the folded structure of a protein.[15,245]

Unfortunately, neutral and adaptive variation appear the same at first inspection of a multiple alignment. To use variation to identify surface positions, therefore, heuristics must be developed that separate (as much as possible) adaptive variation from neutral variation. No filter is known that reliably distinguishes between neutral and adaptive variation, as a rich literature in the field shows.[244] However, a filter built on the notion of "concurrent variation" has proven to be rather effective for the purpose of structure prediction.[234] To apply this filter, positions are identified in a multiple alignment where variation is observed simultaneously in different subbranches of the evolutionary tree. A position is assigned to the surface of the folded structure only if it is variable in more than one subbranch of an evolutionary tree relating the sequences.

Surface assignments made by heuristics based on concurrent variation in several subbranches of an evolutionary tree are significantly more accurate than those obtained by heuristics that search for variation in a single subbranch (Table 6). This improved accuracy has a cost, however. Several sets of homologous sequences are needed to extract conformational information using this heuristic. For the protein kinase alignment shown in Figure 17, 77 additional sequences were available in 1989. Adding the surface assignments obtained from these additional sequences to the larger multiple alignment, together with assignments obtained from analogous heuristics that identify interior positions in a protein fold,[234] the amount of tertiary structural information available for the fragment increases remarkably (Figure 22).

The step from pairwise alignments to multiple alignments is not trivial, either methodologically or

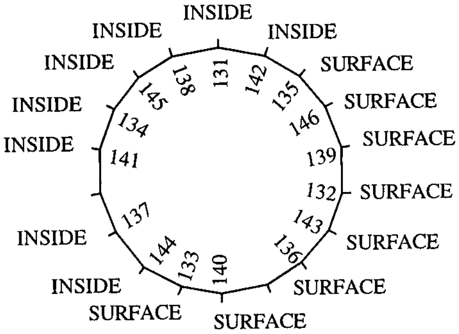


Figure 23. Schiffer–Edmundson[67] helical wheel showing 3.6-residue periodicity in surface and interior assignments of the protein kinase segment presented first in Figure 11. This helix was predicted as part of a *bona fide* prediction of the secondary and supersecondary structure of the protein kinase family.[91]

from the point of view of structure prediction. It is the complexion of a multiple alignment, how many sequences it contains, how much they have diverged, and how they are interrelated, that ultimately determines how much conformational information it will yield. A discussion of tools for constructing multiple alignments is, however, beyond the scope of this review. In the discussion that follows, we will assume that the multiple alignment exists.[246] We will identify cases where problematic multiple alignments cause mistakes in predictions made using tools that analyze variation and conservation in homologous protein sequences.

B. Selecting the Hierarchy

Protein structure prediction has its own "chicken-or-egg" paradox. This paradox arises because tertiary structural interactions are often stronger than local sequence interactions in determining secondary structure.[109–111] This implies that predicting secondary structure from primary structure is essentially impossible without having at least some tertiary structural information. At the same time, a reliable model for secondary structure appears to be necessary before a tertiary structural model can be built. Thus, it appears that neither secondary nor tertiary structure can be predicted before the other is in hand. This paradox must be resolved before satisfactory prediction tools can be developed.

Surface and interior assignments are, of course, a type of tertiary structural information. Further, the heuristics that examine "down" an alignment to extract this information work without the need to have any secondary structural model at all. The heuristics can therefore provide the information needed to resolve the chicken-or-egg paradox.

To illustrate this, we need to assign surface and interior positions more fully for the segment of protein kinase between the two "parses" (the gap and the turn) in Figure 22. The reader can then use this tertiary structural information to make his/her own assignment of secondary structure to this region. One can then proceed to Figure 23, which shows a Schiffer–Edmundson helical wheel that provides a diagram showing the relative positions in space of the surface and interior positions.[67]

The helical wheel suggests that the segment between the parses forms an α helix; this is the only conformation that places the side chain at position 131 on the inside, 132 on the surface, 133 on the surface, and so on. This approach to using tertiary structural information to assign secondary structure proves to be rather general; a 3.6 residue pattern of surface and interior residues nearly always indicates a surface α helix (see below). Similarly, alternating periodicity in interior and surface assignments should indicate a β strand, while four or more consecutive surface positions should indicate a surface turn or coil.[15,245]

This approach for predicting secondary structure is, of course, analogous to the approach suggested many years ago by Schiffer and Edmundson[67] and Lim.[120] In this classic work, however, side chain hydrophilicity and hydrophobicity were used as indicators of surface and interior position, respectively. Side-chain hydrophilicity and hydrophobicity are good, but certainly not excellent, indicators, of surface and interior positions; as discussed above, natural proteins will occasionally place hydrophobic residues on the surface and hydrophilic residues inside, if only to destabilize the protein. This limits the reliability of secondary structure assignments made using the classical approach. The analysis of non-Markovian substitution of amino acids during divergent evolution provides a more reliable indicator of surface and interior position and makes the approach workable.

Even given perfect interior and surface assignments, however, it is clear that this method works best for secondary structural elements that lie on the surface. Secondary structural elements that lie entirely within the fold of a globular protein are more difficult to assign using this strategy. Empirically, short (3–7 positions) segments of internal positions between parses are generally interior β strands.[91] A longer segment (eight or more positions) that is entirely interior could be a long interior strand, two or more internal strands where a parse is not indicated, or an internal helix. Without surface assignments interspersing the interior segments in a defined pattern, it is difficult to distinguish between these.

Further, distinguishing short (1–3 positions) β strands that lie on the surface from surface coils should prove difficult. Because consecutive side chains in a β strand alternate "in–out" in the structure, short surface strands might be indicated by "surface–interior–surface" assignments. However, such assignments are also expected as part of surface coils, making the interior and surface assignments too few to make a statistically reliable case favoring one particular secondary structure over another.

How good (or bad) are such approaches to assigning secondary structure? The helix that the reader has "predicted" for the segment of protein kinase in Figure 22 was in fact part of a *bona fide* prediction of secondary structure for the protein kinase family.[91] The helix was found in the subsequently determined experimental structure. Indeed, the crystallographers pointed out that overall, the prediction was "remarkably accurate, particularly for the small lobe".[247] Subsequent reviewers noted that the protein kinase prediction was much better than that achieved by standard methods,[59] while others commented that the prediction was a "spectacular achievement" that might "come to be recognized as a major breakthrough".[61]

To answer this question in a way that is convincing to experimental and computational biochemists alike, transparent tools must be used to make more *bona fide* predictions. We therefore turn to examples of transparent *bona fide* predictions of secondary structure based on evolutionary analyses.

VI. Transparent Bona Fide Prediction as a Tool for Developing Secondary Structure Prediction Methods

Bona fide predictions are those made and announced before experimental knowledge of a structure is available. They are different from *blind predictions*, which are made without the predictor having knowledge of the experimental structure that is available to others, and *retrodictions*, which are made with the information concerning the correct answer available and known to the predictor. The literature generally refers to all three as "predictions". Because of the very different roles these three different processes have played in the development of the field, it is important to maintain the distinction with some rigor.

While *bona fide* prediction was recognized very early as a useful tool in the field, it was infrequently used in the 1980s, either as a method for developing or as a method for testing new prediction tools. Indeed, the resurrection in the late 1980s of *bona fide* predictions as a general tool for developing and testing methods was criticized, at times harshly. Some scientists asserted that tools developed through *bona fide* predictions could not be subjected to rigorous testing.[176] Others argued that transparent methods are intrinsically not reproducible.[65] Still others argued that *bona fide* predictions, because they were made one at a time, could never be made in sufficient number to permit a statistically valid test of a method. Still others rejected *bona fide* predictions as simply being unscientific. These issues have been discussed in detail elsewhere.[210,248] Even today, despite the evident fact that *bona fide* predictions have been an important force driving both development and testing of prediction methods, many still have reservations.[249]

To understand the importance of *bona fide* predictions, we must consider briefly how prediction tools are developed in their absence. Tools for predicting the conformation of proteins invariably include at least some parameters derived from experimental data. To avoid having those parameters biased to reproduce a specific test set, most computational biochemists divide available data into two sets, a development set to generate the parameters and a test set to evaluate the parameterized tool. A process of "cross validation", where the elements in the development data set and the test data set are permuted, is often used as well.

As important as this procedure is, it does not guarantee an objective test of a parameterized theoretical tool, as appreciated in other areas of theoretical chemistry.[250] Various mechanisms allow the test set to influence the parameters derived from the development set even when the development set and test set are different. Most simply, knowledge of the correct "answer" from the test set determines when parameterization ends. Also, knowledge of the correct "answer" determines which papers are accepted for publication and which ones are rejected. As a consequence, a parameterized method that reaches the published literature will perform better on average when evaluated against the test set than when evaluated against new structures. This is expected even if the test set is explicitly excluded from the data used for the parameterization.

This bias cannot be avoided as long as knowledge of the correct structure can intervene at any time between the time the prediction is made and the time the prediction is announced. The impact of the bias can, however, be minimized by combining retrodictive tests of prediction tools with tests that make *bona fide* predictions, those announced before experimental data are known. This procedure is well known in protein chemistry. It was used, for example, by Georg Schulz (for adenylate kinase) and Brian Matthews (for T4 phage lysozyme) in two well-known prediction "contests" in the 1970s.[54,55] In a *bona fide* prediction, knowledge of a specific test case cannot possibly influence the parameterization of the prediction tool. Nor can it filter the prediction results, favoring publication of successful predictions and removing unsuccessful predictions. The experimental biochemist is therefore more likely to credit a published *bona fide* prediction than a retrodiction. One disadvantage is that *bona fide* predictions must be made and tested one at a time.

Further, a *bona fide* prediction is typically made in a different way from a retrodiction. First, it is generally made singly, for a single protein; retrodictions are generally made against a database of structures. This means that the scientist making the prediction encounters molecular structure *as a chemist*, rather than as a statistician. A single structure can be examined individually; mistakes in the prediction can be discussed individually in terms of real atoms and bonds. The audience for a *Chemical Review* has little difficulty appreciating the value of the approach to developing chemical theory. Even those who are not chemists, however, can understand how the results are different.

Further, a *bona fide* prediction is made with a sense of urgency and focus that does not normally characterize retrodictions. Mistakes in a *bona fide* prediction are obvious, specific, and, in many ways, personal, not buried in the anonymity of a three-state score for a set of proteins. This brings a certain focus to a prediction exercise that is not present in nonpredictive work, as recent prediction projects have indicated. This again forces the predictor to encounter the molecule as an individual, to search, at times frantically with a deadline, for new ideas and new approaches that are fundamentally chemical.

This ultimately leads to the strongest advantage of *bona fide* predictions: they allow transparent tools to be developed more freely. As with the development of other transparent theories of conformation in chemistry, the development of prediction tools from an understanding of molecular evolution required human involvement. Human involvement creates a problem, as it does throughout science. It would be very unusual indeed if the humans involved in the enterprise could separate entirely their understanding of theory, the development of prediction tools, and their hopes for success, and keep these from influencing their judgement about their own work. As any of those involved in chemistry can attest, it is always easy to explain experimental results post hoc, regardless of what these results are.

Bona fide prediction has proven to be a useful tool for overcoming these problems. By making and announcing a prediction before it is known whether the prediction is correct, the predictor is free to hypothesize, speculate, or even guess as to why the prediction worked when it did, and why it failed when it did. In some circles, this may be regarded as “excuse making” and was suggested to be such by one referee of this review. This process is, however, most appropriately characterized as “learning” and, therefore, an essential step in improving prediction tools.

In one sense, prediction “contests” are critical to *bona fide* prediction strategies, as they allow a substantial number of protein targets to be assembled at one time in one place. Otherwise, prediction comparisons must be made one at a time.[90] Their disadvantage is sociological; they represent the prediction exercise as a competition between individuals rather than as the learning exercise that it could (and should) be. This makes the “score” more important than the “analysis”, which in turn is not the optimum use of the exercise. The organizers of the CASP1 project were especially helpful in directing the discussion in this way, encouraging presentations to explain what went wrong, what went right, and why. To contribute to this trend, we refer to prediction “projects” rather than “contests”.

We attempt to review here every example of a transparent *bona fide* prediction based on evolutionary analyses that does not rely on the identification of a homolog whose structure is already solved experimentally. In practice, this goal is not easily achieved. First, many predictions are “joint”, combining transparent and nontransparent tools (for example, where a neural network has been used to assist in the prediction). Neural networks based on multiple sequence alignments have come in many respects to reproduce transparent prediction methods, often making the same mistakes as these (see below). Therefore, such “joint” predictions have been included here where the “transparent” component was significant.

Further, many *a priori* prediction efforts have generated a secondary structure assignment that immediately suggested that the protein folds in the same overall structure as a protein whose structure is known. The known structure has frequently been used to construct a tertiary structure model for the target protein following an approach that is similar to the homology modeling discussed above. The use of predicted secondary structures to establish “long-distance” homologies is becoming frequent.[92,251] Further, the growth in the size of the protein crystallographic database suggests that the most common use of secondary structure prediction tools will be to identify long-distance homologs as a starting point for modeling. We have included a secondary structure prediction here if the homology modeling was dependent on a secondary structure prediction that was made *a priori*, without knowledge of the homolog.

Table 7. Some Predictions Made by Transparent Analysis of Multiple Sequences

protein kinase
Src homology 2 domain
Src homology 3 domain
MoFe nitrogenase
hemorrhagic metalloproteinase
protein tyrosine phosphatase
Pleckstrin homology domain
Von Willebrand factor
proteasome
isopenicillin N synthase
protein serine phosphatase
factor XIIIa
6-phospho-β-galactosidase
synaptotagmin
cyclin
heat shock protein 90 (HSP90)
NK lysin
calponin
fibrinogen

Table 7 lists the predictions discussed here. For those cases where the published literature contains residue-by-residue assignments, and where subsequently determined crystal structures are available for a member of the protein family being examined, prediction and experimental secondary structures are presented in figures associated with the discussion of each.

A. Early Transparent Predictions and Their Mistakes

The most interesting parts of a *bona fide* prediction are their mistakes. These convey the insights, not only into how prediction heuristics might be improved, but also into protein structure and evolution. Therefore, with apologies to the many individuals who have made *bona fide* predictions using methods that analyze patterns of conservation and variation among homologous protein sequences, we focus on the mistakes, and what was learned from these mistakes as we review the *bona fide* predictions made in the past 10 years.

Again, we must emphasize that a discussion of errors is not an *apologia*. It is a learning exercise. One of the great strengths of transparent prediction tools coupled with *bona fide* prediction is that it facilitates, indeed encourages, what has come to be called *post mortem* analyses of mistakes made by predictions. The predictors gather around the prediction and discuss the mistakes, ask what went wrong, and propose ways in which the mistakes might have been avoided. This is, of course, a common exercise in the experimental sciences, where it is viewed as a way of improving methods, models, and theories.

Sequences (Protein Kinases)

```
   001      010       020 024   030       040       050    57/60      070     78
    |        |          |   |     |         |         |      ||         |      |
1 - DQFDRIKTLGTGSFGRVMLVKHKE---------------SGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAV----
2 - TDFNFLMVLGKGSFGKVMLSERKG---------------TDELYAVKILKKDVVIQDDDVECTMVEKRVLALPG---
3 - DEYQLYEDIGKGAFSVVRRCVKLC---------------TGHEYAAKIINTKKLSAR-DHQKLEREARICRLL----
4 - GVWRLGKTLGTGSTSCVRLAKHAK---------------TGDLAAIKIIPIR-------YASIGMEILMMRLL----
5 - ENYQKVEKIGEGTYGVVYKARHKL---------------SGRIVAMKKIRLEDESEG-VPSTAIREISLLKEVNDEN
6 - NEYKLIDKIGEGTFSSVYKAKDITGKITKKFASHFWNYGSNYVALKKIYVTS-----SPQRIYNELNLLYIMT---
7 - SEVQLLKRIGTGSFGTVFRGRWHG------------------DVAVKVLKVSQPTAE-QAQAFKNEMQVLRKT----
```

Experimental Secondary Structure[247]

```
      EEEEEE        EEEEEEE --------------  EEEEEEEEEHHHHHHH HHHHHHHHHHHHHH ----
```

Consensus Prediction[91]

```
      ******   EEEEEEEEEEE  --------------   EEEEEEE ******** HHHHHHHHHHHHHH----
```

Consensus Retrodictions made by the Heidelberg Neural Network[211]

```
1 -    EEEEEEE    EEEEEEEE  --------------   EEEEEEEHHHHH     HHHHHHHHHHHHH ----
2 -    EEEEEEE    EEEEEEEE  --------------   EEEEEEE HHHEEE     HHHHHHHHHHHH ---
3 -     EEEEEE  EEEEEEEEE   --------------   EEEEEEEHHH     -  HHHHHHHHHHHH ----
4 -    EEEEEEE   EEEEEEEEE  --------------   EEEEEEHHHH-------   HHHHHHHHHH ----
5 -     EEEEEE   EEEEEEEEE  --------------  EEEEEEEEE       -  HHHHHHHHHHHHH
6 -     EEEE      EEEEEEE     HHHHHHHHHHH    EEEEEEEEE  -----  HHHHHHHHHHHHH ---
7 -   EEEEEEE      EEEEEEE  ----------------- EEEEEE      HH-HHHHHHHHHHHHHHH----
```

Sequences (Protein Kinases)

```
   081     090   95/101      113/115 120      130       140  146/147        160
    |        |    ||            ||    |         |         |     ||            |
1 - NFPFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGR---FSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPEN
2 - KPPFLTQLHSCFQTMDRLYFVMEYVNGGDLMYHIQQVGR---FKEPHAVFYAAEIAIGLFFLQSKGIIYRDLKLDN
3 - KHSNIVRLHDSISEEGFHYLVFDLVTGGELFEDIVAREY---YSEADASHCIQQILEAVLHCHQMGVVHRDLKPEN
4 - RHPNILRLYDVWTDHQHMYLALEYVPDGELFHYIRKHGP---LSEREAAHYLSQILDAVAHCHRFRFRHRDLKLEN
5 - NRSNCVRLLDILHAESKLYLVFEFLDM-DLKKYMDRISETGALDPRLVQKFTYQLVNGVNFCHSRRIIHRDLKPQN
6 - GSSRVAPLCDAKRVRDQVIAVLPYYPHEEFRTFYRD------LPIKGIKKYIWELLRALKFVHSKGIIHRDIKPTN
7 - RHVNILLFMGFMTR-PGFAIITQWCEGSSLYHHLHVADTR--FDMVQLIDVARQTAQGMDYLHAKNIIHRDLKSNN
```

Experimental Secondary Structure[247]

```
           EEEEEE   EEEEEE       HHHHHHHH  ---  HHHHHHHHHHHHHHHHHHHH EEEE      E
```

Consensus Prediction[91]

```
       EEEEEEEEEE    EEEEEEEEE  HHHHHHHHHH ---  HHHHHHHHHHHHHHHHHHHHHHEEEE......
```

Consensus Retrodictions made by the Heidelberg Neural Network[211]

```
1 -     EEEEEEEEE    EEEEEEE    HHHHHHHH   ---  HHHHHHHHHHHHHHHHHHHH  EEEE    HH
2 -    EEEEEEEEEE    EEEEEEEE    HHHHHHH   ---   HHHHHHHHHHHHHHHHHHH   EEE    HH
3 -     EEEEEEEEEE   EEEEEEE     HHHHHH    ---  HHHHHHHHHHHHHHHHHHHH  EEEEE
4 -     EEEEEEEEE    EEEEEEE    HHHHHHHH   ---  HHHHHHHHHHHHHHHHHHHH   EEE
5 -    EEEEEEEEEE    EEEEEEEE  -HHHHHHHHH       HHHHHHHHHHHHHHHHHHHH  EEEEE
6 -      EEEEE E     EEEEEEEE   HHHHH   ------  HHHHHHHHHHHHHHHHHHEEEEEEEEE
7 -     EEEEEEEEE -  EEEEEEEE    HHHHHHH    --  HHHHHHHHHHHHHHHHHHHHH  HHHHHHHHH
```

```
                           Sequences (Protein Kinases)
           172/177        190       200         215/217 220      230      240/242
             ||            |         |              ||  |         |         ||
1 - LLIDQQG---YIQVTDFGFAKRVKGRT----WTLCGTPEYLAPEII---LSKGYNK-AVDWWALGVLIYEMAAGYP
2 - VMLDSEG---HIKIADFGMCKENIWDGVTT-KTFCGTPDYIAPEII---AYQPYGK-SVDWWAFGVLLYEMLAGQA
3 - LLLASKCKGAAVKLADFGLAIEVQGDQQAW-FGFAGTPGYLSPEVL---RKEAYGK-PVDIWACGVILYILLVGYP
4 - ILIKVNEQ--QIKIADFGMATVEPNDSCL--ENYCGSLHYLAPEIV---SHKPYRGAPADVWSCGVILYSLLSNKL
5 - LLIDKEG---NLKLADFGLARSFGVPLRNY-THEIVTLWYRAPEVLL--GSRHYST-GVDIWSVGCIFAEMIRRSP
6 - FLFNLELG--RGVLVDFGLAERQMDYKSMISANRAGTRGFRAPEVLM--KCGAQST-KIDIWSVGVILLSLLGRRF
7 - IFLHEGL---TVKIGDFGLATVKTRWSGAQPEQPSGSVLWMAAEVIRMQDPNPYSF-QSDVYAYGVVLYELMTGSL

                       Experimental Secondary Structure247
    EEEE  ---EEEEEE    EEEE    ----               ---       -HHHHHHHHHHHHHHHH

                              Consensus Prediction91
    EEEE   ---EEEEEEEEEE       ----   EEEEEEEEEEEE          - EEEEEEEEEEEEEEE

         Consensus Retrodictions made by the Heidelberg Neural Network211
1 - HE     --- EEEEEE   EE     ----EEEE      HHHHH---H      - EEEEEEEEEEEEEEE
2 - HHH    --- EEEEE    EEEE      -EEEE    HHHHHHH---HH     -   EEEEEEEEEEEEE
3 -  EE        EEEEEE  EEE      EE-EEEE       HHHH---H      - EEEEEEEEEEEEE
4 - EEE     -- EEEEE   EEE      E--EEEE       HHHH---H        EEEEEEEEEEEEEEE
5 -  EE    --- EEEEEHHHHHHH      E-EEEEEEEEE  HHHHH--       - EEEEHHHHHHHHHH
6 - EEEE    -- EEEEHHHHHHHHHHHHHHHEHHHH       HHHHH--       -EEEEEEEHHHHHHHHHH
7 - HHH    ---EEEEEE  HHHEEE             EEE  HHHHH       EE-EEEEEEEEEEEEEEEE


                           Sequences (Protein Kinases)
         252/260     270       280       290       300       310       320  325
           ||         |         |         |         |         |         |    |
1 - PFFA----DQPIQIYEKIVSG-KVRFPSH-----FSSDLKDLLRNLLQVDLTKRFGNLKNGVNDIKNHKWFATT
2 - PFEG----EDEDELFQSIMEH-NVAYPKS-----MSKEAVAICKGLITKHPGKRLGCGPEGERDIKEHAFFRYI
3 - PFWD----EDQHKLYQQIKAG-AYDFPSPEWDT-VTPEAKNLINQMLTINPAKRITAHEALK-----HPWVCQR
4 - PFGG----QNTDVIYNKIRHG-AYDLPSS-----ISSAAQDLLHRMLDVNPSTRITIPEFFS-----HPFLMGC
5 - LFPGDSEIDEIFKIFQVLGTPNEEVWPGVTLLQDGEEDAIELLSAMLVYDPAHRISAKRALQ-----QNYLRDF
6 - PMFQSL--DDADSLLELCTIFGWKELRKCAALHGDHYWCFQVLEQCFEMDPQKRSSAEDLLK-----TPFFNEL
7 - PYSHI---GCRDQIIFMVGRGYLSPDLSKISSN-CPKAMRRLLSDCLKFQREERPLFPQILATI---ELLQRSL

                       Experimental Secondary Structure247
        ----   HHHHHHHHH         -----  HHHHHHHHHH               HHHHHH

                              Consensus Prediction91
        ----  HHHHHHHHHHH-H      -----HHHHHHHHHHHHHHHH     EEEEEE *****

         Consensus Retrodictions made by the Heidelberg Neural Network211
1 -     ----  HHHHHHHHHH  -      -----HHHHHHHHHHHHHH              HHHH   EEE
2 -     ----  HHHHHHHHHH  -      -----HHHHHHHHHHHHHH              HHHHHHHHEE
3 -     ----  HHHHHHHHHHH-           -HHHHHHHHHHHHHH     HHHHHHHH -----   EE
4 -     ----  HHHHHHHHH   -      -----HHHHHHHHHHHHHH        HHHH  -----  EEE
5 -       HHHHHHHHHHHH          EE      HHHHHHHHHHHH      HHHHHH  -----   E
6 -       -- HHHHHHHHHHHH           E HHHHHHHHHHHHHH    HHHHHHHHHH-----HHHHH
7 -      ---  HHHHHHHHH              -  HHHHHHHHHHHH        HHHHHHHH---HHHHH
```

Figure 24. Sequences, experimental secondary structure, prediction, and neural network retrodictions for the protein kinase family.[45] The inconsistencies of the retrodictions obtained from the PHD neural network are especially noteworthy. Key: E, β strand; H, α helix; the interior helix is underlined; 1, cAMP-dependent protein kinase (mouse); 2, protein kinase C (ox); 3, protein kinase type II (rat); 4, protein kinase CDR1, *S. pombe*); 5, CDC28-cdc 2 protein kinase (*S. pombe*); 6 CDC Protein 7 (*S. cerevisiae*); 7, Human Raf protooncogene kinase.

1. Protein Kinases (Catalytic Domains)

While not the first *bona fide* prediction to be made with tools that transparently analyzed patterns of variation and conservation within homologous protein sequences,[15] the protein kinase prediction was the first to be tested by a subsequently determined crystal structure (Figure 24).[59,91,247] The protein kinase prediction illustrated several points. First, it illustrated how surface and interior assignments can be joined with parsing assignments to identify most of the important secondary structural elements in the fold, especially surface helices and internal β strands.

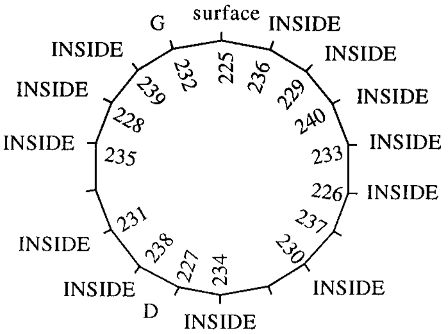


Figure 25. Interior helix in protein kinase, showing the absence of 3.6-residue patterns that might indicate a consensus secondary structure.

Further, the predicted secondary structure proved to be sufficiently accurate that, when combined with assignments of positions near the active site and some covariation analysis, an *antiparallel* β sheet at the center of the first domain was correctly surmised.[59,91,247] The crystal structure also confirmed that the covariation analysis obtained by inspecting homologous sequences indeed reflected real contacts between the relevant residues in the folded structure. Here, the Markov model broke down at positions distant in the sequence in a way that was useful to identify packing.[91] This is, we believe, one of the first times that the general nature of a tertiary fold has been correctly predicted in a fully *a priori* sense without an explicit model of a fold drawn from a crystallographic database and without exploiting circular dichroism data.

Within the context of classical evolution-based methods discussed above, protein kinase provides an example as well. Predicting the antiparallel β sheet required special confidence in the secondary structure prediction, as it contradicted the conjecture that a *parallel* sheet lay at the core of the first domain of the protein kinase structure. Many groups applied the homology modeling, joint prediction methods, and knowledge-based techniques outlined above to support this conjecture.[252−256] This conjecture was based in part on a conserved sequence motif, Gly-Xxx-Gly-Xxx-Xxx-Gly, found in many kinases and dinucleotide binding domains, where it is part of a parallel sheet in an $\alpha-\beta$ fold. This conjecture was wrong. The many predictions of the structure of protein kinase are not included in Figure 24 because they cannot be coherently aligned with the correct structure. Nevertheless, this may be the first case where a transparent secondary structure prediction overrode an assignment based on sequence motifs.[257]

The mistaken assignments in the prediction are especially instructive, however. The most serious mistake was the misassignment of an internal helix between positions 225 and 240 as a strand (Figure 24). Such misassignments are expected from an approach that assigns secondary structure based on patterns in surface and interior assignments (see above). Every position in the 226−240 segment was assigned to the interior of the protein kinase structure (Figure 25). The assignments were correct. But because the helix was entirely buried, no 3.6 residue pattern of periodicity in surface and interior assignments indicated a helical geometry. The mistake had a double impact. Not only was the internal helix missed, but the misassignment of this core secondary structural element prevented the construction of a tertiary structural model for the second domain in the protein.

Accordingly, efforts have been devoted to developing tools to distinguish internal helices from internal strands. The simplest heuristic is, of course, the length of the internal segment, where long internal stretches are marked as possible internal helices. When the internal helices pass near an active site, 3.6-residue periodicity of active-site assignments is also observed. Using these tools, an internal helix was correctly predicted in the hemorrhagic metalloproteinase family (Figure 15);[90] another has been predicted in the structure of the serine/threonine protein phosphatases (see below).[96] Very often, interior helices can be identified through efforts to build a supersecondary structure from a set of predicted secondary structural units in the problematic region. This constitutes a “refinement” of secondary structural units in light of additional tertiary structural information extracted from the multiple alignment.

The secondary structure assignments near the active site (segment 177−193, Figure 24) and the autophosphorylation site in protein kinase (segment 198−212, Figure 24) were also problematic. In the first region, the experimental structure identified two β strands, while the prediction assigned one long strand with a break at position 182. In the second, the prediction placed a long β strand (positions 201−212) with breaks between positions 203−204 and 208−209. The crystallographers assigned no defined secondary structure in this region. The first part of the segment forms an extended structure, while the second and third segments are best viewed as coils.

Regions near an active or regulatory site play unique functional roles in a polypeptide chain. They are the least likely to conform to expectations based on an analysis of protein sequences overall. Markov rules fail severely in these regions. However, altered patterns of variation and conservation in these regions generally reflect catalytic function rather than secondary structure. Thus, predicting secondary structure in these regions is the most difficult for any modeling tool. However, identification of non-Markov behavior in divergent evolution can identify active-site regions (see below), and prediction tools can be designed to alert the biochemist to the existence of the problematic region.

Further, the difficulties in predicting the secondary structure of segment 198−212 in protein kinase prompted efforts to improve heuristics to parse, or divide, the multiple alignment into units that form independent secondary structures. One of the most powerful tools to have resulted from this effort are parsing strings, consecutive combinations of Pro, Gly, Ser, Asp, and Asn in a polypeptide that break secondary structures with a high probability, as discussed above.[73] These tools became part of a growing set of heuristics for assigning secondary structures.

Misassignments were also made in regions where secondary structure in the protein kinase homologs has diverged: at the beginning of the multiple

Table 8. Types of Mistakes in the Prediction for the MoFe Nitrogenase Family[a]

position	mistakes	comments
serious mistakes		
internal helix		
076−080	mistaken strand for helix	internal helix
bad multiple alignment		
147−154	underpredicted strand	bad parse at 148−149: misplaced gap or sequence error in the database
164−174	underpredicted helix	helix shortened by a badly placed gap; weak α predicted
370−374	helix too short	bad alignment and misplaced gap
392−395	mistaken strand for helix	bad alignment leading to bad parse
434−451	underpredicted helix	bad alignment
461−466	underpredicted strand	bad alignment
491−504	underpredicted helix	bad alignment
active site		
068−072	overpredicted strand	active site
094−107	underpredicted helix	a weak helix assignment was made, active site
122−125	mistaken strand for helix	active site
155−160	mistaken strand for helix	active site helix (helix bundle with 122−125 and two from α-subunit)
less serious mistakes		
various definitions of secondary structure type		
112−118	underpredicted strand	DSSP assigns a 2 residue edge strand; parsing strings limit β to 114−115
186−194	underpredicted strand	DSSP also does not assign a strand here
280−283	underpredicted helix	DSSP does not assign a helix here, but rather a turn
335−344	underpredicted strand	DSSP does not assign a strand here; an edge strand in the publication
523−526	underpredicted helix	DSSP does not assign a helix here
529−532	overpredicted strand	strand assigned in the databank, but not the published version of the structure
short secondary structural element with mistaken surface/interior assignments		
272−278	underpredicted strand	incorrect surface assignment at position 274
521−523	overpredicted strand	incorrect interior assignments

[a] DSSP indicates an assignment made by the "define secondary structure of proteins" program.[66] See Figure 26 to obtain a more comprehensive view of the quality of the predictions.

alignment (Figure 24), at the end of the multiple alignment, and in a short segment between positions 050 and 057. At the beginning of the alignment, the experimental structure for a cAMP-dependent protein kinase assigned an edge strand; the prediction proposed a coil. At the end of the alignment, the divergence was so severe that the multiple alignment misplaced a gap and, therefore, missed a noncore helix assigned in the crystal structure. The model overpredicted a strand at positions 307−312; the experimental structure places a coil in this region. Finally, the cAMP-dependent kinases contain a short helix at positions 050−056 not present in other kinases. Because of this gap, the consensus model assigned a coil in this region. In the refinement process, however, the conformation of the cAMP-dependent kinase subfamily was examined separately, and the possibility of a helix in this region in this particular subfamily was noted.[91]

As noted above, misassignments of secondary structure in regions where secondary structure has diverged rarely present serious obstacles in the use of a predicted secondary structural model. Thus, the last three misassignments are not serious, in contrast to the misassignment of the internal helix.

2. The β Subunit of MoFe Nitrogenase

The MoFe nitrogenase challenge was issued just days before a crystal structure appeared in print. The prediction was therefore unrefined;[73] the multiple alignments generated by the automated computer tool DARWIN[258] were not separately adjusted, secondary structural elements were not evaluated within possible supersecondary structural models, and problematic assignments near the active site were not addressed. Even so, long surface helices were readily identified.[73,259] Ten surface helices were predicted (Figure 26); all were found in the experimental structure.

The prediction could not, of course, have supported tertiary structural modeling, as it contained too many serious mistakes. Indeed, the MoFe nitrogenase prediction provided examples of five different ways where patterns in surface and interior assignments might be unreliable indicators of secondary structure (Table 8). Thus, the mistakes proved to be more instructive than the successes.

Two classes of misassignments were clearly not serious. The first set, accounting for six "misassignments" when comparing the predicted and experimental structures, arose from differing experimental definitions of secondary structure. The details are instructive. The "underpredicted" strands at positions 112−118, 186−194, and 335−344 and the "underpredicted" helices at positions 280−283 and 523−526, all listed as standard secondary structural units in the paper where the crystal structure was published,[259] are not assigned as such by DSSP,[66] one of the standard tools discussed above for automatically assigning secondary structures to coordinate data. All of the strands with uncertain experimental secondary structure assignments are at the edge of their respective sheets, and both of the controversial helices contain only four residues. The "overpredicted" strand (positions 529−532), missing in the published structure, was later assigned as a strand in the databank version of the structure. Thus, each of these misassignments provides an illustration of the discussion of scoring methods above; different

```
   65   70   75   80   85   90   95   100  105  110  115  120  125  130    alignment numbering
    .    |    .    |    .    |    .    |    .    |    .    |    .    |
   TVNPAKACQPLGAVLCALGFEKTMPYVHGSQGCVAYFRSYFNRHFREPVSCVSDSMTEDAAVFGGQQ   sequence
       EEEEE  CEEEEE CC EEE     CCC              CCC  CCC    EEEECC  H   prediction
 ..EEE     HHHHHHHHHHH   EEEEEEE HHHHHHHHHHHHHH    EEEEEEE  HHHHHH  HH   exp (crystallographer)
     S S   HHHHHHHHHTTBTTEEEEEES HHHHHHHHHHHHHHHSS    EE    TTHHHH SHH   experiment (DSSP)
     .    |    .    |    .    |    .    |    .    |    .    |    .
         70        80        90        100       110       120           crystal numbering

         140  145  150  155  160  165  170  175  180  185  190  195  200 alignment numbering
     .    |    .    |    .    |    .    |    .    |    .    |    .    |
 NMKDGLQNCKATY-KPDMIAVSTTCMAEVIGDDLNAFINNSKKEGFI-----PDEFPVPFAHTPSFVGSH  sequence
 HHHHHHHHHHH   CC       EEEE   CCCCCCC         CCCCC      CC   CC  CCH   prediction
 HHHHHHHHHHHHH   EEEEEEEEHHHHHH   HHHHHHHHHHH           EEEEEEEEE     H  exp (crystallographer)
 HHHHHHHHHHHHH   SEEEEEE HHHHHHT  HHHHHHHHHHTTSS      TTS   B    TTSS H  experiment (DSSP)
 |    .    |     .    |    .    |    .    |    .         |    .    |
130       140        150       160       170            180       190    crystal numbering

    205  210  215  220  225  230  235  240  245  250  255  260  265  270 alignment numbering
     .    |    .    |    .    |    .    |    .    |    .    |    .    |
 VTGWDNMFEGIARYF----T---------LKSMDDKVVGSNKKINIVPGFETYL--GNFRVIKRMLSEMG  sequence
 HHHHHHHHHHHHHHH  unassigned due to gaps   EEEEE CCCCCCCCC HHHHHHHHHHHH  prediction
 HHHHHHHHHHHHHHH    H                      EEEEEEE    H  HHHHHHHHHHHHH   exp (crystallographer)
 HHHHHHHHHHHHHHH    H         GGGGGG  TTTT  EEEE  S   H  HHHHHHHHHHHHTT  experiment (DSSP)
  .    |    .                 |    .    |    .    |      .    |    .
      200                    210       220       230         240         crystal numbering

    275  280  285  290  295  300  305  310  315  320  325  330  335  340 alignment numbering
     .    |    .    |    .    |    .    |    .    |    .    |    .    |
 VGYSLLSDPEEVLDTPADGQ-FRMYA-GGTTQEEMKDAPNALNTVLLQPWHLEKTKKFVEGTWKHEVPKL  sequence
            CCCCCCCCCCCC  CCCC  HHHHHHH  CC EEEEE  HHHHHHHHHHHHH  CCCCC  prediction
  EEEEEEE HHHH                  HHHHHHHHH  EEEEE     HHHHHHHHHHH EEEEEE  exp (crystallographer)
   EEESS  TTTTS   SS   S      B HHHHHSGGGSSEEEES GGG HHHHHHHHHHT         experiment (DSSP)
  |    .    |     .    |     .    |    .    |    .    |    .    |    .
 250       260        270        280       290       300       310       crystal numbering

    345  350  355  360  365  370  375  380  385  390  395  400  405  410 alignment numbering
     .    |    .    |    .    |    .    |    .    |    .    |    .    |
 ---NIPMGLDWTDEFLMKVSEISG-QPIPASLTK----ERGRLVDMMTD-SHTWLHGKRFALWGDPDFVM  sequence
    CC        HHHHHHHHHHH CCCCCC       C HHHHHHHHHCCEEEEE   EEEE CCC     prediction
    E     HHHHHHHHHHHHHH      HHHHH    HHHHHHHHHHH HHHHH  EEEEEEE HHHHH  exp (crystallographer)
      S HHHHHHHHHHHHHHHHT     HHHHH    HHHHHHHHHHH HHHHHTT EEEE   HHHHH  experiment (DSSP)
       |    .    |    .     |    .        |    .     |    .    |    .
      320       330        340           350        360       370        crystal numbering

    415  420  425  430  435  440  445  450  455  460  465  470  475  480 alignment numbering
     .    |    .    |    .    |    .    |    .    |    .    |    .    |
 GLVKFLLELGCEPVHILCH-NGNKRWKKAVDAI-----LAASP----YGKNATVYIGKDLWHLRSLVFTD  sequence
 HHHHHHHHHHHH EEEEE  CCCC  unassigned    due    to    gaps   HHHHHHHH C  prediction
 HHHHHHHHH EEEEEEEE     HHHHHHHHHH     HHH         EEEEEE    HHHHHHHHH   exp (crystallographer)
 HHHHHHHHTT EEEEEEET T  HHHHHHHHHH     HHT G    GGTT EEEES  HHHHHHHHHHS  experiment (DSSP)
   |    .    |    .     |    .         |        .    |    .    |    .
  380       390        400            410           420       430        crystal numbering

    485  490  495  500  505  510  515  520  525  530  535  540  545      alignment numbering
     .    |    .    |    .    |    .    |    .    |    .    |    .
 --KPDFMIGNSYGKFIQRDTLHKGKEFEVPLIRIGFPIFDRHHLHRSTTLGYEGAMQILTTLVNSILE    sequence
 CCC  EEEE      unass. gaps    EEEEEE    EEE     EEEE HHHHHHHHHHHHHHH    prediction
     EEEEEEHHHHHHHHHHHHHH   EEEEEEE        HHHH      HHHHHHHHHHHHHHHH    exp (crystallographer)
     SEEEE TTHHHHHHHHHHH TGG   EEE SS   SSSSGGGS  SHHHHHHHHHHHHHHHHHH    experiment (DSSP)
     |    .    |    .    |    .    |    .    |    .    |    .   |
        440       450       460       470       480       490      500   crystal numbering
```

Figure 26. Representative sequence, experimental secondary structure,[259] and secondary structure prediction[73] for the MoFe nitrogenase family. Key: E, β strand; H, α helix; T, turn; C, coil; G, 3_{10} helix; S, bond; B, bridge. Underlined segments in the sequence are residues near the active site. Top numbering is the alignment number; bottom numbering is the numbers in the experimentally determined crystal structure. Especially noteworthy are the differences in secondary structural assignments obtained by the crystallographers and by application of the program DSSP[66] to the coordinates provided by the crystallographers.

ways of looking at the same experimental structure yield different secondary structure assignments, and this fact must be considered in analyzing each prediction.

Misassignments in noncore regions account for two additional mistakes in the prediction. Two short noncore segments (positions 272−278 and 521−523) were underassigned and overassigned, respectively, because of the small number of surface and interior assignments applying to the segments overall. Here, the seriousness of the misassignments is less easily determined. It will be interesting to see if these assignments are conserved in homologous nitrogenase proteins.

A third class of misassignments arose from a failure to align gaps with other gaps in the unrefined multiple alignment. Together with the substantial sequence divergence in the MoFe nitrogenase family, the multiple alignment was poor. Positions 215 and 240 illustrate this (see, for example, the floating Thr 220). Mistaken gap insertions obliterated three helices, at positions 164−174, 434−451, and 491−504 (Figure 26), and three strands, at positions 147−154, 335−344, and 461−466. Further, a β strand was overpredicted at alignment positions 392−396 when a misplaced gap in the alignment disrupted a helix that would otherwise have been propagated to include this region. These are serious mistakes. However, the prescription for avoiding them is clear; the multiple alignment must be refined. This conclusion has again been noted very recently.[179]

Two classes of misassignments are more serious and more difficult to avoid. First, an internal helix was misassigned as an internal strand (positions 076−080), as in protein kinase (see above), for much the same reasons.

Second, the MoFe nitrogenase has an extended active site with two metal binding sites. Cys 71, Cys 96, Cys 155, and Ser 195 serve as cluster ligands, Pro 73, Phe 100, Tyr 99, Met 156, and Phe 196 form a hydrophobic environment around the cluster, and Gly 95, Gln 94, and Thr 154 are conserved hydrophilic amino acids in the vicinity of the cluster (all underlined in Figure 26).[259] Further, two short helices (alignment positions 121−126 and 155−160) are oriented in parallel from one metal cluster toward the surface, forming a four helix bundle with two helices from the other subunit. The 4Fe:4S cluster binds on the top surface of these helices. As noted above, secondary structure is especially difficult to assign in active-site regions, and these regions contained virtually all of the instances where the prediction confused helices and strands.

The mistakes made in the MoFe nitrogenase prediction suggested a particular hierarchical procedure for structure prediction to avoid similar mistakes. The procedure must start with tertiary structural assignments, parses, and active-site assignments. These are jointly used to assign secondary structure to “easy” regions first, where the multiple alignment is good, which are distant from the active site, and where periodicity in the assignments is obvious. Where the multiple alignment is evidently bad, it must be refined, possibly with the help of secondary structural assignments made for subfamilies of the evolutionary tree. Two potentially problematic regions then remain: internal helices and active-site regions. The first are identified by a stretch of continuous interior assignments. The second are identified by their distinctive conservation of functionalized amino acids. Efforts to improve the prediction tools must focus on these regions.

3. The Hemorrhagic Metalloproteases

The prediction effort that for the first time compared on an equal footing consensus classical prediction tools with transparent methods and the (then) new PHD neural network was discussed above (Figure 15). The transparent prediction performed significantly, if modestly better than the PHD tool, while the PHD tool performed considerably better than both the classical GOR and classical Chou−Fasman tools averaged over the multiple sequence alignment. Further, the transparent prediction avoided one of the principal errors noted above; an internal helix was correctly assigned to positions 133−145.

The remaining misassignments in the transparent prediction included an overprediction of a strand at positions 064−069, the underprediction of an edge strand at positions 108−112, the overprediction of a strand at positions 148−152, the overprediction of a strand at positions 169−172, and the misassignment of the final two-residue element of the β meander at positions 176−177. In several cases, these problems can be traced directly toward a problem of definition. For example, the 169−172 strand, although not technically a β strand, does form an extended structure. Several residues in strands in the β meander at positions 176−177, missed in the prediction, are also missed in some experimental secondary structural assignments (Figure 15). The edge strand at positions 108−112 is not a core structure. The overprediction at positions 148−152 is near the active site and appears to be an extended structure as well. Thus, none of the misassignments would seem to be fatal to a tertiary structure modeling effort. Indeed, the antiparallel nature of the central β sheet might have been identified.

B. Predicting Small Domains

Intracellular signal transduction is mediated in higher cells by small domains, usually containing approximately 100 amino acids, that interact with other domains. The rapid emergence of experimental structures (both by crystallography and NMR) for these offered an opportunity to test many of the prediction methods. As illustrated below, these demonstrated much of the power of transparent prediction tools.

1. The Src Homology 3 (SH3) Domain

The Src homology 3 (SH3) domain, an independent unit found in many proteins involved in intracellular signal transduction, was an unrefined prediction.[260] Because the SH3 domain family had undergone considerable sequence divergence, the prediction was in fact three predictions, made for each of the major subfamilies of the SH3 domain. Six experimental structures later became available for various SH3 domains. These include two structures solved by crystallographic methods,[85,261] and four solved by NMR methods.[71,76,86,87]

The prediction proved to be controversial, as indicated by the difficulty various commentators have

```
                                    Sequences
              sub       0    1    1    2    2    3    3    4    4    5    5    6    6    7    7
             family     5    0    5    0    5    0    5    0    5    0    5    0    5    0    5
src           a     GGVTTFVALYDYESRTETDLSFKKGERLQIVNNTRKVDVR---------EGDWWLAHSLSTGQTGYIPSNYVAPSD
Fyn           a       VTLFVALYDYEARTEDDLSFHKGEKFQILNSS---------------EGDWWEARSLTTGETGYIPSNYVAPVD
H PLC         b     TFKCAVKALFDYKAQREDELTFIKSAIIQNVEKQ---------------EGGWWRGDYGG-KKQLWFPSNYVEEMV
C spec        c     TGKELVLALYDYQEKSPREVTMKKGDILTLLNST---------------NKDWWKVEV--NDRQGFVPAAYVKKLD
PI3K-1        d     AEGYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETTGERGDFPGTYVEYIGRK
PI3K-2        d     AEGYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETTGERGDFPGTYVEYIGRK

                              Experimental Structures
src           a         EEEEeEEE      EEEE    EEEEE   ---------------   EEEEEEE    EEEE3333EEEE
Fyn-1         a         EEEE                  EEEEEE  ---------------   EEEEEE     EEEE    EEE
Fyn-2         a         EEEE                  EEE     ---------------   EEEEEE     EEEE    EEE
H PLC         b        EEEEE EEE         EEE EEEEE EEE---------------   EEEEEE   EEEEEE     EEE
C spec        c         EEEE                  EEEEEE  ---------------   EEEEEE--  EEEEE 3333EEE
PI3K          d         EEEE                  EEEE                HHHH  EEEEEE    EEEEEE3333EEEEE
PI3K          d        EEEEEeeEE      EEEEEE  EEEEEHHHHHHHH   3333333333EEEEEE    EEEEE3333EEEEE

                               Bona Fide Predictions
              a         EEEEEE EEEE   EEEEE   HHHHHHHH                 EEEEEE        EEE EEEE
              b          EEEEEEEEE    EEEEE HHHHHHHHHH                 EEEEEEE       EE EEEE
              c           EEEEEEE     EEEEE  HHHHHHHH                   EEEE        EEEE EEE
              d                       EEEE    EEEEEE                   EEEEEEEE

                              Consensus Retrodictions
                              PHD Neural Network[208]
M nsrc        a        EEEEEEE           E   EEEEEE         --------- HHHHHHHHH      EEEE EEEEE
H PLC1        b        HHHEEHHHH         E    EEEEEE  ---------------    EEHHH  -    EEE    EEE
C spec        c         EEEEEE         EEE   EEEEEEE  ---------------   HHHHHH--    EEEE  HEEE
H PI3K        d      HHHHHHHHH      HHHHHH    EEEEE  EEEEEE                               EEEEE

                        Garnier-Osguthorpe-Robson (GOR)[105]
C csrc        a     TTEEEEEEE     HHHHHHHHHHHHEEEE    ---------------TTTT      TTTT     TT
C spec        c      HHHHHHHHHHHTT   HHHHHHHHHEEEEE   ---------------TTTT    T--TTTTT  HHHHHHHH
H PI3K        d      TTTEEEEETT   HHHHHHHHHHHHEEEETTT EEEEE          TTTT      TTTTTT   EEEEEE

                                   COMBINE[262]
CONSENSUS +C        CCCHHHHHHHCCCCCHHHHHHHHHCCHHHHHHCCCCCEE--CCCCCCCCHCCCCECCCCCCCCCCCCCCCCCCE--CC
CONSENSUS +S        CCCHHHHHHHCCCCCHHHHHHHHHCCHHHHHHCCCCCEECCCCCCCCCCHCCCCECCCCCCCCCCCCCCCCCCCEECCC
```

Figure 27. Representative sequences, experimental secondary structures,[71,76,85−87] predictions, and retrodictions for the Src homology 3 (SH3) domain family and subfamilies. Key: E, β strand; H, α helix; T, turn; C, coil; 3, 3_{10} helix. The *bona fide* predictions[260] marked a, b, and c are for three separate subfamilies of the SH3 domain. The prediction marked d is for a specific alignment.[263] The differences in the output of the PHD neural network[208] tested blind with different homologs are especially noteworthy, as is the poor quality of the prediction made by a consensus GOR approach. It should be noted that the GOR tool used is implemented in the GCG package[107] and may differ from the implementation proposed by Garnier *et al.*[105]

had in agreeing upon its three-state score (Table 9). Rost and Sander proposed a three state Q_3 score of 56%.[211] Robson and Garnier proposed a score of 46%.[65] Barton *et al.* calculated Q_3 scores ranging from 42% to 58%, after disregarding one experimental structure.[74] This divergence has more to say about the scoring methods than about the prediction itself; hence, the SH3 domain served as an excellent example for our discussion of scoring methods.

Figure 27 presents the transparent predictions for the SH3 domain with retrodictions made by the 1993 PHD neural network, the GOR program found in the GCG package, and the COMBINE program.[262] The transparent prediction contains two problematic aspects, one serious and the other less so. First, a helix is predicted near the middle of the domain. The predicted helix obliterates an important β strand present in all of the structures. This misassignment illustrates the difficulty in obtaining a statistically significant pattern of surface and interior assignments in a short strand. In the SH3 domain, the helix was assigned because positions 27, 39, and 30 were placed on the inside of the folded structure while positions 26, 28, 31, and 32 were placed on the surface (Figure 27). In the spectrin crystal structure, the actual side-chain exposures of residues 30 and 28 are reversed. Nonetheless, it is intriguing to note that a helix does appear in this region in some members of the SH3 domain family.

The second problematic region is the shift in the placement of the final β strand. This can be attributed in part to divergence in sequence and a resulting bad alignment. The final strands in the experimental structures fall in a region where DARWIN does not consider the overall alignment to be significant.

The remaining problem is difficulties in identifying by DSSP the β hairpin in the first part of the structure. Visual inspection of the experimental structures makes it certain that the structure is there. Some automated tools for assigning secondary structure to coordinate data find it; others do not. Because the protein is small, this creates large variation in the Q_3 scores.

```
sequence             WYFGKITRRESERLLLNPENPRGTFLVRESETTKGAYCLSVSDFDNAKG
experimental 1       EEEE   HHHHHHHH       EEEEEEEEE     EEEEEEEE
experimental 2       tE     HHHHHHHHt tt  tt EEEEE    tt EEEEEEEEEtttE
prediction 1 ref 264 EEE    HHHHHHHH      EEEEEEE      EEEEEEE
prediction 2 ref 265 EEEEEEEHHHHHHHH       EEEEEEEE      EEEEE


sequence             LNVKHYKIRKLDSGGFYITSRTQFSSLQQLVAYYSKHADGLCHRLTNVC
experimental 1         EEEEEEEEEE   EEE   EEE  HHHHHHHHHH         EEEE
experimental 2       EEEEEEEEEE tt  EE  tt EE  HHHHHHHHtt  tt       E
prediction 1 ref 264  EEEEEEE      EEEE      HHHHHHHHH
prediction 2 ref 265   EEEEEE       EEE     HHHHHHHHHHH
```

Figure 28. Representative sequence, *bona fide* consensus prediction, and experimental secondary structure for the Src homology 2 (SH2) domain. Experimental structure 1 is from the paper describing Brookhaven database PDB 1sha (ref 266); experimental 2, for Swiss Port (P00524, SRC_RSVSR) tyrosine-protein kinase transforming protein Src (EC 2.7.1.112) (P60-SRC), from the Rous sarcoma virus. Key: E, β strand; H, α helix; t, turn.

Table 9. What is the Correct Three-State Score for the SH3 Domain Prediction?[a]

experimental structure used as reference	correct	incorrect	seriously incorrect	total residues	three-state score (in %)
C csrc	43	16	5	64	67
PI3K	52	22	5	79	66
FYN-1	37	21	6	64	58
FYN-2	38	24	3	65	58
H PLC	34	24	6	64	53
C spec	32	24	6	62	52

[a] All numbers represent residues, except the percentage three-state score. Correct assignments indicate residues assigned in the experimental structure as part of helices paired with residues predicted to lie in helices, plus residues assigned in the experimental structure as part of strands paired with residues predicted to lie in strands, plus residues assigned in the experimental structure as part of coils paired with residues predicted to lie in coils. The 3_{10} helices are treated as coils. Seriously incorrect assignments are those that mistake residues assigned to a helix for those predicted to be part of a strand and *vice versa*. The calculated scores for the same consensus prediction range from 52 to 67%, depending only on which member of the protein family is chosen as the reference structure.

In Heidelberg, Musacchio *et al.* made a transparent prediction for part of the secondary structure of the SH3 domain using an analysis of conservation and variation within the protein family.[263] First, they constructed a multiple alignment for the family. They then positioned three strands in the SH3 domain, and surmised that the protein would form five or six strands overall. The three predicted strands are indeed found in the experimental structure (Figure 27).

2. The Src Homology 2 (SH2) Domain

Two *bona fide* predictions of the Src homology 2 (SH2) domain were published, one by Blundell's group,[264] the other by Barton's group.[265] Both predictions are essentially perfect (Figure 28).[266] The strand missed is an edge strand, and the underprediction is not serious to an overall perception of the fold.

Missed strand 6 lies in a region where substantial divergence of sequence has taken place, including some gapping, implying that it is not present in all of the SH2 domain homologs. Not surprisingly, it is also an edge strand. Thus, all core secondary structural elements were correctly identified, no elements were predicted that were not later found to be part of the core fold, and no region of helix was misassigned as a strand (or vice versa).

3. The Pleckstrin Homology Domain

Two *bona fide* predictions were made for the pleckstrin homology domain,[267,268] another domain putatively involved in signal transduction and identified by sequence similarities in a variety of proteins.[269,270] The predictions are compared with two experimental structures in Figure 29;[271,272] the comparison was reviewed by Russell and Sternberg.[60] In both cases, the sequence was first parsed, and secondary structure was assigned to separate elements. A single helix and six or seven strands were predicted in each case. A subsequently determined experimental structure showed that the core elements were correctly predicted in terms of number, type, and location. Within the pleckstrin homology domain family, considerable divergence of secondary structure is seen; indeed, the residue-by-residue three-state correspondence between any two sequences can be as low as 73%.[271,272] Both predictions achieve this three-state score and differ from a consensus model only in the precise start and end points of the helices (something that depends on the crystallographic assignments in any case) and in overlooking a short helix found in only one branch of the pleckstrin homology domain family tree. Thus, these predictions are essentially perfect as consensus models.

Russell and Sternberg examined the possibility of predicting the pleckstrin homology domain structure using a consensus GOR method.[60] In this particular case, the outcome was considerably worse than the published transparent predictions. The PHD neural network did considerably better than the consensus GOR tool, however, replicating the nearly perfect performance of the transparent methods.

4. The Cyclin Family

Two independent predictions of secondary structure were made for the cyclins.[204,273] These are shown in Figure 30, together with experimental assignments of secondary structure.[274] The two predictions are quite similar, and correspond well with the experimental structure, except for a pair of strands

```
experimental sequence 1    MEPKRIREGYLVKKGSV--------FNTWKPMWVVLLEDGIEFYKKK------SDNSPK
experimental sequence 2          MEGFLNRKHEWEAHNKKASSRSWHNVYCVINNQEMGFYKDAKSAASGIPYHSE

consensus prediction 1              EEEEE                 EEEEE    EEEE
consensus prediction 2             EEEE                   EEEEEE  EEEEE

experimental structure 1         EEEEEEEE              EEEEEEEEE  EEEEE
experimental structure 2         EEEEEEEEEE            EEEEEEEEE  EEEE  HHHHHHH


experimental sequence 1    GMIPLKGSTLTSPCQDFGKRMFVFKITTTKQQDHFFQAAFLEERDAWVRDINKAIKCIEG
experimental sequence 2    VPVSLKEAICEVALDYKKKK-HVFKLRLSDGNEYLFQAKDDEEMNTWIQAISSAISSDKH

consensus prediction 1      EEEE  EEE            EEEEEE     EEEEE    HHHHHHHHHHHHH
consensus prediction 2      EEEE EEEEE           EEEEE      EEEEE    HHHHHHHHHHHHHHH

experimental structure 1   EEE    EEEEE         EEEEEEE   EEEEEEE   HHHHHHHHHHHHHHHHH
experimental structure 2    EEE   EEEEEE         EEEEEE    EEEEEE   HHHHHHHHHHHHH
```

Figure 29. Representative sequences, *bona fide* consensus predictions, and experimental secondary structures[271,272] for the pleckstrin homology domain family. Key: E, β strand; H, α helix. Prediction 1 is from ref 267. Prediction 2 is from ref 268.

mispredicted in one but not in the other. This misprediction illustrates the interplay of experimental data and prediction.

The cyclin structure as solved shows an internal repeat, where two halves have equivalent chain topology built from five helices. This internal repeat had been detected on the basis of weak sequence similarities before the experimental structure was solved.[275] Bazan used this repeat in his secondary structure prediction.

At the time that the predictions were made, experimental results with deletion mutants were available that suggested that a portion of the protein could be deleted with only modest effect on function.[276] These deletions would disrupt a portion of a predicted internal helix, a disruption that would be expected to have far greater impact on performance.

Bazan chose to ignore the experimental data (mentioning nevertheless the problematic conclusions that might be drawn from these experiments in light of his model) and predicted a helix that extended through the deletion. Gerloff and Cohen chose to modify their prediction in light of the experimental data. Interestingly, ignoring the experimental data provided the better prediction. This is not the first time that Bazan has used an analysis of aligned homologous sequences to draw correct inferences that contradicted conclusions presumed to be supported by experiment.[277]

C. Predictions of Large Proteins

The results obtained from *bona fide* prediction efforts for the SH2, SH3, and pleckstrin homology domains, synaptotagmin (see below) and cyclin show that transparent approaches to structure prediction can reliably predict secondary structure over the entire length of a protein. The *bona fide* nature of these predictions makes this conclusion convincing even to the most skeptical experimental biochemist. Further, it is possible to venture that transparent methods produce results that are superior to those obtained using consensus classical prediction methods, at least for these domains. Finally, predicting secondary structure is no longer a limiting step in the modeling of tertiary structure for such domains. Improved tools that help assemble tertiary structural models from a set of predicted secondary structural elements would be useful, as would be tools that distinguish between alternative packings of predicted secondary structural elements. These could be used to retrospectively evaluate alternative secondary structure models, the preferred model being the one that provides the most convincing tertiary structural modeling. This is currently done routinely by hand. Were the second class of tools available, it is conceivable that both the pleckstrin homology domain (see below) and cyclin structures could have been built entirely *de novo*.

These small domains might be expected to be the best targets for these tools, however. The polypeptide chains form soluble single domain structures that are ideal for modeling, the ratio of surface area to volume is large, and many sequences are available in the databases. Attention therefore returned to predicting secondary structure in larger proteins. The experience discussed above with protein kinase, MoFe nitrogenase, and the hemorrhagic metalloproteinases showed that secondary structure can be accurately predicted for many secondary structural elements of such proteins using transparent methods. However, experience also showed that certain types of secondary structural elements are difficult to identify: internal helices, regions near the active site, edge strands, and regions where the core fold is not conserved (in decreasing order of seriousness).

“Perfection” in a secondary structural model is very important. A single serious mistake in the assignment of secondary structural elements normally prevents modeling tertiary structure for an entire domain. A “perfect” prediction is one that misassigns no core helices as strands (or vice versa), misses no core secondary structural elements, and misassigns no noncore region in a way that obstructs modeling of a tertiary structure.

1. Isopenicillin N Synthase

Isopenicillin N synthase lies within a family of homologous proteins that includes enzymes involved

```
              NEVPDYHEDIHTYLREMEVKCKPKVGYMKKQPD    sequence
                                                   prediction 1
                HHHHHHHHHHHHHHHHHHHHH              prediction 2
                     HHHHHHHHHHHHHHH    hhhhhh     subprediction helix only
                                        eee        subprediction strand only
                                                   experimental assignment[274]
                 HHHHHHHHHHHHHHHHHH                experimental assignment by DSSP

ITNSMRAILVDWLVEVGEEYKLQNETLHLAVNYIDRFLSSMSVLRGKLQL sequence
----  HHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHH    HHHHHH prediction 1
   HHHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHH     EEEEEEE prediction 2
   hhhhhHHHHHHHHHHHHhh hhhhHHHHHHHHHHH         hhh subprediction helix only
                        eee eeeeee         eeeeeee subprediction strand only
  HHHHHHHHHHHHHHHHHH   HHHHHHHHHHHHHHH         HHH experimental assignment[274]
  HHHHHHHHHHHHHHHHHH   HHHHHHHHHHHHHH       HHHHHH experimental assignment by DSSP

VGTAAMLLASKFEEIYPPEVAEFVYITDDTYTKKQVLRMEHLVLKVLTFD sequence
HHHHHHHHHHHHHH    HHHHHHHHHHH HHHHHHHHHHHHHHHHHHHH prediction 1
  HHHHHHHHHHHHHH     HHHHHHH       HHHHHHHHHHHHHH  prediction 2
hhhhhhhhhhhhhhhh     ???????       HHHHHHHHHHhhhh  subprediction helix only
                                     eeeeeee eeeee subprediction  strand only
HHHHHHHHHHHHH      HHHHHHH      HHHHHHHHHHHHHHH    experimental assignment[274]
HHHHHHHHHHHHH      HHHHHH       HHHHHHHHHHHHHHH    experimental assignment by DSSP

LAAPTVNQFLTQYFLHQQPANCKVESLAMFLGELSLIDADPYLKYLPSVI sequence
    HHHHHHHHHHHHH    HHHHHHHHHHHHHHHHH        HHHH prediction 1
   HHHHHHHHHHHHHHH    EEEEEEEHHHHHHHHHHHEEEEEE HHH prediction 2
   hhHHHHHHHHHHHHH     hhhhhhhhhhhhhh          HHH subprediction helix only
                      eeeeeee  eeeeee              subprediction strand only
  HHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHH          HHHH experimental assignment[274]
     HHHHHHHHH       HHHHHHHHHHHHHH   HHHH    HHHH experimental assignment by DSSP

AGAAFHLALYTVTGQSWPESLIRKTGYTLESLKPCLMDLHQTYLKAP    sequence
HHHHHHHHHHHHH  HHHHHHHHHH HHHHHHHHHHHHHHHHHHH      prediction 1
HHHHHHHHHHHHHH  HHHHHHHHHhHHHHHHHHHHHHHHHHHHHHH    prediction 2
HHHHHHHHHHHHHh  HHHHHHHHHhhHHHHHHHHHHHHHHHHHHHH    subprediction helix only
                                                   subprediction strand only
HHHHHHHHHHHH      HHHHHH    HHHHHHHHHHHHHHH        experimental assignment[274]
HHHHHHHHHHHHHH    HHHHHHH   HHHHHHHHHHHHHHHHH      experimental assignment by DSSP

QHAQQSIREKYKNSKYHGVSLLNPPETLNL                     sequence
                                                   prediction 1
HHHHHHHHHHHHHHHHH      -------                     prediction 2
HHHHHHHHHHHHHhhhh                                  subprediction helix only
                                                   subprediction strand only
                                                   experimental assignment[274]
     HHHHH                                         experimental assignment by DSSP
```

Figure 30. Representative sequences, *bona fide* consensus prediction, and experimental secondary structure[274] for the cyclin family. Prediction 1 is adapted from ref 204. Prediction 2 is adapted from ref 273. Analysis is adapted in part from appendix to ref 273. Key: E, β strand; H, α helix. In the prediction, "e" refers to a weakly predicted strand, while "E" refers to a strongly predicted strand; "h" refers to a weakly predicted helix, while "H" refers to a strongly predicted helix; ??? indicates a region of unpredicted secondary structure for cyclin A. The underlined predicted segment of strand was based on an interpretation of an experimental result involving deletion mutations (see text). The inference from the experimental results proved to be incorrect. Subpredictions are for prediction 2.

in ethylene biosynthesis, oxidizing enzymes (flavanone 3-β-hydroxylase, flavanone 3-dioxygenase, hyoscyamine 6-dioxygenase), and anthocyanidin synthases. Divergence in function often makes a prediction challenging, as it implies difficulties in detecting active-site residues. The prediction exercise was rendered especially challenging by the substantial divergence in sequence that is associated with the divergence in function. A global consensus prediction was made from separate predictions for three subfamilies of proteins. These are compared with an experimental structure of isopenicillin N synthase (Figure 31).[278]

A comparison of the predicted and experimental secondary structure assignments identifies examples of the misassignments discussed above. Least serious is the omission of noncore structures in a consensus model. For example, the consensus pre-

```
              10        20        30        40        50        60        70
          .    |    .    |    .    |    .    |    .    |    .    |    .    |
           sisipsi i  i psssss ssiississii s  iiii sa isississiissi ssi ss
a.   MPIPMLPAHVPTIDISPLSGGDADDKKRVAQEINKACRESGFFYASHHGIDVQLLKDVVNEFHRTMTDEEK
b.   MPVLMPSADVPTIDISPLFGTDPDAKAHVARQINEACRGSGFFYASHHGIDVRRLQDVVNEFHRTMTDQEK
c.        PKANVPKIDVSPLFGDNMEEKMKVARAIDAASRDTGFFYAVNHGVDVKRLSNKTREFHFSITDEEK
d.         SAHVPTIDISPLFGTDAAAKKRVAEEIHGACRGSGFFYATNHGVDVQQLQDVVNEFHGAMTDQEK
e.   MPILMPSAEVPTIDISPLSGDDAKAKQRVAQEINKAARGSGFFYASNHGVDVQLLQDVVNEFHRNMSDQEK
f.       MPSAEVPTIDVSPLFGDDAQEKVRVGQEINKACRGSGFFYAANHGVDVQRLQDVVNEFHRTMSPQEK
g.          ADVPVIDISGLSGNDMDVKKDIAARIDRACRGSGFFYAANHGVDLAALQKFTTDWHMAMSAEEK
h.        PVANVPRIDVSPLFGDDKEKKLEVARAIDAASRDTGFFYAVNHGVDLPWLSRETNKFHMSITDEEK
i.    MGSVSKANVPKIDVSPLFGDDQAAKMRVAQQIDAASRDTGFFYAVNHGINVQRLSQKTKEFHMSITPEEK
pred          eeEEEEe      hhHHHHHHHHHHHHHHhh eeeee aaa hhHHHHHHHHHHHHHh HHH
expt            EEE HHHH   HHHHHHHHHHHHHHHH   EEEEE     HHHHHHHHHHHHHH  HHHH
                β1   α1          α2            β2            α3          α4
              core not core     core          core          core

             80        90        100       110       120       130       140
         .    |    .    |    .    |    .    |    .    |    .    |    .    |
     isiiisii sssp si  piis spss     i ii psiss  psisss p      ip  ss  siss
a.   YDLAINAYNKNNP_RTRNGYYMAVKGKKAVESWCYLNPSFSEDHPQIRSGTPMHEGNIWPDEKRHQRFRP
b.   HDLAIHAYNENNS_HVRNGYYMARPGRKTVESWCYLNPSFGEDHPMIKAGTPMHEVNVWPDEERHPDFRS
c.   WDLAIRAYNKEHQDQIRAGYYLSIPEKKAVESFCYLNPNFKPDHPLIQSKTPTHEVNVWPDEKKHPGFRE
d.   HDLAIHAYNPDNP_HVRNGYYKAVPGRKAVESFCYLNPDFGEDHPMIAAGTPMHEVNLWPDEERHPRFRP
e.   HDLAINAYNKDNP_HVRNGYYKAIKGKKAVESFCYLNPSFSDDHPMIKSETPMHEVNLWPDEEKHPRFRP
f.   YDLAIHAYNKNNS_HVRNGYYMAIEGKKAVESFCYLNPSFSEDHPEIKAGTPMHEVNSWPDEEKHPSFRP
g.   WELAIRAYNPANP_RNRNGYYMAVEGKKANESFCYLNPSFDADHATIKAGLPSHEVNIWPDEARHPGMRR
h.   WQLAIRAYNKEHESQIRAGYYLPIPGKKAVESFCYLNPSFSPDHPRIKEPTPMHEVNVWPDEAKHPGFRA
i.   WDLAIRAYNKEHQDQVRAGYYLSIPGKKAVESFCYLNPNFTPDHPRIQAKTPTHEVNVWPDETKHPGFQD
pred HHHHHHHHhh         eeeee      ac site       e?e?    act site      hhhH
exp  HHH               EEE         EEEEE         HHHH                   HHH
                        β3          β4            α5
      core             core        core        not core

            150        160       170       180        190       200       210
        .    |    .     |    .    |    .    |    .     |    .    |    .    |
     ii siissiisi si i s iiiii ss siisssisss  i  s i isipiissip   i   s  ssi
a.   FCEQYYRDVFSLSKV_LMRGFALALGKPEDFFDASLSLADTLSAVTL_IHYPYLEDYP__PVKTGPDGTKLS
b.   FGEQYYREVFRLSKVLLLRGFALALGKPEEFFENEVTEEDTLSCRSLMIRYPYLDPYPEAAIKTGPDGTRLS
c.   FAEQYYWDVFGLSSA_LLRGYALALGKEEDFFSRHFKKEDALSSVVL_IRYPYLNPIPPAAIKTAEDGTKLS
d.   FCEGYYRQMLKLSTV_LMRGLALALGRPEHFFDAALAEQDSLSSVSL_IRYPYLEEYP__PVKTGPDGQLLS
e.   FCEDYYRQLLRLSTV_IMRGYALALGRREDFFDEALAEADTLSSVSL_IRYPYLEEYP__PVKTGADGTKLS
f.   FCEEYYWTMHRLSKV_LMRGFALALGKDERFFEPELKEADTLSSVSL_IRYPYLEDYP__PVKTGPDGEKLS
g.   FYEAYFSDVFDVAAV_ILRGFAIALGREESFFERHFSMDDTLSAVSL_IRYPFLENYP__PLKLGPDGEKLS
h.   FAEKYYWDVFGLSSA_VLRGYALALGRDEDFFTRHSRRDTTLSSVVL_IRYPYLDPYPEPAIKTADDGTKLS
i.   FAEQYYWDVFGLSSA_LLKGYALALGKEENFFARHFKPDDTLASVVL_IRYPYLDPYPEAAIKTAADGTKLS
pred HHHHHHHHHHHHHhh_hh  eeeeee  e?e?        eeeeeee eeeeee
expt HHHHHHHHHHHHHHH_HHHHHHHH        HHH        EEEE EEE       HHH EE     EE
           α6                         α7            β5         α8  β6     β7
          core                       not core     core      not core not core
             220       230       240       250       260       270       280
         .    |    .    |    .    |    .    |    .    |    .    |    .    |
     i s s i ii iii  si  i is  s ississsssiii     ii  i ssiip psa i ii s
a.   FEDHLDVSMITVLFQTEVQNLQVETADGWQDLPTSGENFLVNCGTYMGYLTNDYFPAPNHRVKFINAERL
b.   FEDHLDVSMITVLFQTEVQNLQVETVDGWQSLPTSGENFLINCGTYLGYLTNDYFPAPNHRVKYVNAERL
c.   FEWHEDVSLITVLYQSDVANLQVEMPQGYLDIEADDNAYLVNCGSYMAHITNNYYPAPIHRVKWVNEERQ
d.   FEDHLDVSMITVLFQTQVQNLQVETVDGWRDIPTSENDFLVNCGTYMAHVTNDYFPAPNHRVKFVNAERL
e.   FEDHLDVSMITVLYQTEVQNLQVETVDGWQDIPRSDEDFLVNCGTYMGHITHDYFPAPNHRVKFINAERL
f.   FEDHFDVSMITVLYQTQVQNLQVETVDGWRDLPTSDTDFLVNAGTYLGHLTNDYFPSPLHRVKFVNAERL
g.   FEHHQDVSLITVLYQTAIPNLQVETAEGYLDIPVSDEHFLVNCGTYMAHITNGYYPAPVHRVKYINAERL
h.   FEWHEDVSLITVLYQSDVQNLQVKTPQGWQDIQADDTGFLINCGSYMAHITDDYYPAPIHRVKWVNEERQ
i.   FEWHEDVSLITVLYQSNVQNLQVETAAGYQDIEADDTGYLINCGSYMAHLTNNYYKAPIHRVKWVNAERQ
pred  act site              h?h?h?h?h?  internal---helix      act site
expt EEEE     EEEEEE     EEEEE  EEEE       EEEEE HHHHHH         EEEE     EE
      β8        β9        β10    β11        β12   α9             β13
     core    not core     not   core       core  core           core
```

```
             290       300            310       320       330
       .      |    .    |         .    |    .    |    .    |
      ipiiis   ss
a.   SLPFFLHAGHTTVMEPFSPE_____DTRGKELNPPVRYGDYLQQASNALIAKNGQT
b.   SLPFFLHAGQNSVMKPFHPE_____DTGDRKLNPAVTYGEYLQEGFHALIAKNVQT
c.   SLPFFVNLGFNDTVQPWDPS_____KEDGKTDQRPISYGDYLQNGLVSLINKNGQT
d.   SLPFFLNGGHEAVIEPFVPE_____GASEEVRNEALSYGDYLQHGLRALIVKNGQT
e.   SLPFFLNAGHNSVIEPFVPE_____GAAGTVKNPTTSYGEYLQHGLRALIVKNGQT
f.   SLPFFFHAGQHTLIEPFFP____DGAPEGKQGNEAVRYGDYLNHGLHSLIVKNGQT
g.   SIPFFANLSHASAIDPFAP_____PPYAPPGGNPTVSYGDYLQHGLLDLIRKNGQT
h.   SLPFFVNLGWEDTIQPWDPATAKDGAKDAAKDKPAISYGEYLQGGLRGLINKNGQT
i.   SLPFFVNLGYDSVIDPFDPR_____EPNGKSDREPLSYGDYLQNGLVSLINKNGQT
pred   eEEEEee                           hhhhHHHHHHHHHHHhhhhh
expt EEEEEE      EE                     EEHHHHHHHHHHH
      β14        β15                    β16     α10
      core     not core               not core  core
```

Figure 31. Representative sequences, *bona fide* consensus prediction,[248] and experimental secondary structure[278] for the isopenicillin N synthase superfamily. Key: E, β strand; H, α helix; t, turn; C, coil. In the prediction, "e" refers to weakly predicted strand; E, strongly predicted strand; h, weakly predicted helix; H, strongly predicted helix. Predicted surface and interior assignments are indicated by "s" and "i" above the sequences; "p" indicates parse; "a" indicates active site; ? indicates uncertain prediction. The crystal structure is for enzyme i from *Aspergillus nidulans*. Sequences are labeled as follows: (a) isopenicillin N synthase from *S. griseus*; (b) (P12438) isopenicillin N synthase from *S. lipmanii*; (c) (P08703) isopenicillin N synthase from *P. chrysogenum*; (d) (P10621) isopenicillin N synthase from *S. clavuligerus*; (e) (P18286) isopenicillin N synthase from *S. jumonjinensis*; (f) (X57310) isopenicillin N synthase from *N. lactamdurans*; (g) (P16020) isopenicillin N synthase from *Flavobacterium* sp.; (h) (P05189) isopenicillin N synthase from *C. acremonium*; (i) (P05326) isopenicillin N synthase from *A. nidulans*.

diction does not identify two segments that the crystallographers assigned as helices (α7 and α8) built from only three residues. Nor does it predict a helical conformation for two segments that are assigned by the crystallographers as helices (α1 and α5) built from only four residues. The prediction also does not assign strand conformations to four segments assigned as β strands (β6, β7, β15, and β16) built from only two residues. None of these secondary structural elements is important for the overall fold, least of all α8, which comes in a region that is a gap in many of the homologous proteins. Further, it is likely that such short secondary structural elements are not uniformly found by different experimental methods examining the same coordinates (see above). For example, as a typical α helix requires four residues before the first intrahelix hydrogen bond can be formed, a helix built from only three residues can be equally well described as a coil. Therefore, these underpredictions have no impact on the overall structural model. Further, two strands (β10 and β11) form an external hairpin that is also not a core element of the fold, and were mispredicted.

More serious, and therefore more interesting, are the misassignments of secondary structure near active-site residues. Four strands (β4, β8, β9, and β13) were underpredicted because of their proximity to a segment of the protein that was assigned to the active site. Three of these are actually near the active site; β4 is not. Normally, active-site segments are identified more successfully. Here, the difficulties in finding active-site residues can be directly attributed to the enormous divergence in catalytic function of members of the protein families, which in turn implies that functionalized amino acids that are normally conserved at active-site positions are not conserved within the isopenicillin N synthase superfamily.

Last, the prediction noted the difficulty in assigning the segment comprising residues 246−260, which it was noted could be built either from two β strands or an internal helix. In reality, the segment forms one strand and an internal helix (β12 and α9). The prediction itself discussed this ambiguity and indicated how it must be handled. When building a tertiary structure model, it would be necessary to model both alternative secondary structural assignments in this region.

Thus, the prediction for isopenicillin N synthase provides an excellent catalog of problems needing to be solved, with an understanding of why they exist. It was not, however, adequate as a starting point for modeling tertiary structure.

2. Factor XIIIa

The Oxford group undertook a prediction of Factor XIIIa in response to a challenge from the crystallographers. The protein is very large (some 730 amino acids). An experimental structure recently emerged,[279] and the predicted and experimental structures are compared in Figure 32. In independent work, the Chou−Fasman method[104] was also applied in a routine fashion to a single protein sequence in the family.[280] The details of the prediction are not available, but the secondary structural model built from a single sequence evidently predicted considerably more helix than the consensus model.

As with isopenicillin N synthase, the prediction was good, if not outstanding. A large number of β strands, 27 in all, were assigned correctly, with the usual variation in length, but with remarkably little shifting (Figure 32). Two additional helices were correctly assigned. Many of the underpredictions were not serious. For example, several short helices, assigned in the experimental structure but not assigned in the prediction (at positions 59−63, 176−178, 478−481, and 593−597), do not appear to be critical to the fold.

```
        010       020       030       040       050
    .    |    .    |    .    |    .    |    .    |
SETSRTAFGGRRAVPPNNSNAAEDDLPTVELQGVVPRGVNLQEFLNVTSV sequence
                                            xxxxxx prediction
                                             EEEEE experimental

        060       070       080       090       100
    .    |    .    |    .    |    .    |    .    |
HLFKERWDTNKVDHHTDKYENNKLIVRRGQSFYVQIDFSRPYDPRRDLFR 100
xx                     eeee     eeeeee         eee prediction
EEE     HHHHH        EEEE      EEEEEEE          EE experimental

        110       120       130       140       150
    .    |    .    |    .    |    .    |    .    |
VEYVIGRYPQENKGTYIPVPIVSELQSGKWGAKIVMREDRSVRLSIQSSP 150
eeeee         eeeeeee        eeeeee     eeeeeee    prediction
EEEE           EEEEE          EEEEEEEE  EEEEEEEE   experimental

        160       170       180       190       200
    .    |    .    |    .    |    .    |    .    |
KCIVGKFRMYVAVWTPYGVLRTSRNPETDTYILFNPWCEDDAVYLDNEKE 200
     eeeeee                  eeeee       eeeee     prediction
    EEEEEEEEEEE  EEEE    HHH EEEE              HHH experimental

        210       220       230       240       250
    .    |    .    |    .    |    .    |    .    |
REEYVLNDIGVIFYGEVNDIKTRSWSYGQFEDGILDTCLYVMDRAQMDLS 250
  eeee    eeee                   eeeeeeeeee        prediction
HHHHHHH EEEEEEEE  EEEEEEEE       HHHHHHHHHHH       experimental

        260       270       280       290       300
    .    |    .    |    .    |    .    |    .    |
GRGNPIKVSRVGSAMVNAKDDEGVLVGSWDNIYAYGVPPSAWTGSVDILL 300
     eeeeeeeeeee       eee                  hhhhhh prediction
    HHHHHHHHHHH         EEE                 HHHHHH experimental
near active site

        310       320       330       340       350
    .    |    .    |    .    |    .    |    .    |
EYRSSENPVRYGQCWVFAGVFNTFLRCLGIPARIVTNYFSAHDNDANLQM 350
hh           eeee  eeeeeeeee    eeee           eee prediction
HHHH    EEEE HHHHHHHHHHHH      EEEEEE              experimental

        360       370       380       390       400
    .    |    .    |    .    |    .    |    .    |
DIFLEEDGNVNSKLTKDSVWNYHCWNEAWMTRPDLPVGFGGWQAVDSTPQ 400
eeeee    eee      eeeee    eee                     prediction
EEEEE                    EEEEE          EEEEEEEEEE experimental

        410       420       430       440       450
    .    |    .    |    .    |    .    |    .    |
ENSDGMYRCGPASVQAIKHGHVCFQFDAPFVFAEVNSDLIYITAKKDGTH 450
     eeee    eeee   eeeeee   eeee     eeee      ee prediction
      EEEEEEEHHHHHH        HHHHHHHHH  EEEEE        experimental

        460       470       480       490       500
    .    |    .    |    .    |    .    |    .    |
VVENVDATHIGKLIVTKQIGGDGMMDITDTYKFQEGQEEERLALETALMY 500
eeee       eeeeeeee                 hhhhhhhhhh     prediction
EEEEEE       EEEE      EEE HHHH      HHHHHHHHHHHH  experimental

        510       520       530       540       550
    .    |    .    |    .    |    .    |    .    |
GAKKPLNTEGVMKSRSNVDMDFEVENAVLGKDFKLSITFRNNSHNRYTIT 550
                 eeeeeee        eeeeeeeeee    eeee prediction
                 EEEEEEE         EEEEEEEE          experimental

        560       570       580       590       600
    .    |    .    |    .    |    .    |    .    |
AYLSANITFYTGVPKAEFKKETFDVTLEPLSFKKEAVLIQAGEYMGQLLE 600
eeeeee                              eeeee      eee prediction
             EEEEEEEEEEEEEE     EEEEEE   HHHHHH    experimental
        610       620       630       640       650
    .    |    .    |    .    |    .    |    .    |
          edge strand
QASLHFFVTARINETRDVLAKQKSTVLTIPEIIIKVRGTQVVGSDMTVTV 650
e      eeee   hhhhhhhhhhh     eeeeee         eeeee prediction
  EEEEEEEEEE    EEEEEEEEEE      EEEEE        EEEEE experimental


        660       670       680       690       700
    .    |    .    |    .    |    .    |    .    |
EFTNPLKETLRNVWVHLDGPGVTRPMKKMFREIRPNSTVQWEEVCRPWVS 700
ee           eeeeee                 eeeeeeeeee     prediction
EEE      EEEEEEEEE        EEEEEEE    EEEEEEEE      experimental

        710       720       730
    .    |    .    |    .    |
GHRKLIASMSSDSLRHVYGELDVQIQRR
   eeeeeeee    EEEEEEE            prediction
   EEEEEEE    EEEEEEEE            experimental
```

Figure 32. Representative sequences, *bona fide* consensus prediction,[281] and experimental secondary structure[279] for the blood coagulation factor XIIIa family. An "x" indicates a region that was assigned "A1" in the prediction. Numbers correspond to residue numbers in the crystal structure. Key: E, β strand; H, α helix. In the prediction, "e" refers to a weakly predicted strand, while "E" refers to a strongly predicted strand; "h" refers to a weakly predicted helix, while "H" refers to a strongly predicted helix.

The most informative aspects of the prediction are again the mistakes. Most prominent are several serious misassignments of helices as strands, including the helices at positions 198−207, 234−244, 255−265, 314−325, 415−419, and 428−436. Further, toward the carboxyl end, a strand (positions 617−626) is misassigned as a helix, and some secondary structural elements between residues 580 and 600 are missed or shifted.

Some of the mistakes are very interesting. For example, helix 198−207 is missed because positions 196−202 were all strongly assigned to the surface, and position 203 holds a conserved Glu (E), which might also be assigned to the surface, except for the fact that it is so highly conserved. The following three interior positions (204−206) are canonically assigned as a strand. The helix formed by this segment is not reflected in any 3.6 residue periodicity. A part of the helix appears to be buried, while a part appears to be fully exposed. Close inspection of the experimental structure shows that this Glu forms a salt bridge with Lys 467. This long-distance tertiary contact undoubtedly has something to do with the unusual behavior of this secondary structure segment during divergent evolution.

Other mispredictions are rather surprising. For example, helix 234−244 is found on the surface of the protein. A clean 3.6-residue pattern of periodicity is identified using surface and interior predictions (such as those generated as outlined above)[234] across positions 232−239. This pattern extends to position 244 if a weak surface assignment at position 240 is accepted. Thus, this helix would have been assigned correctly had contemporary transparent prediction methods been used. However, the joint prediction method used by Barton allowed a misprediction made by classical methods to outweigh a correct prediction made by contemporary methods.

Several mispredictions reflect mistakes that are commonly made by all methods. For example, the helix between positions 255 and 265 is near an active site, as is the helix between positions 314 and 325. Both of these are mispredicted as strands. Finally, helix 428−436 is an internal helix, also difficult to find by transparent methods.

As an exercise in the learning curve, the Factor XIIIa structure is a milestone. Some 30% larger than the MoFe nitrogenase (see above), it is the largest protein to have been modeled to date using evolutionary information. Further, few proteins exist with more than 1000 amino acids in a single polypeptide chain. Thus, successful modeling of proteins of this size will bring to a close an important phase in the development of prediction methodology.

3. The von Willebrand Factor A Domain

The von Willebrand factor is a large glycoprotein found in blood plasma, where mutant forms are associated with bleeding disorders. Edwards and Perkins applied unbiased GOR and Chou−Fasman tools to each of 75 homologous protein sequences within the family to obtain an average prediction.[282] To resolve ambiguities in the averaging, the PHD and SAPIENS programs[17] were applied. The protein was predicted to fold in an $\alpha-\beta$ conformation, with six β strands identified. The crystallographic database was then searched to find possible templates for homology modeling. The $\alpha-\beta$ TIM barrel was not considered, because too few strands were predicted, while the six predicted strands were consistent with a doubly wound β sheet. A search through the crystallographic database found 38 proteins that have a doubly wound $\alpha-\beta$ core. These were used as threading targets for the predicted secondary structure. The GTP-binding domain of the ras protein was found to give the best score using the THREAD[159] and QSLAVE[283] programs, and was used to model the tertiary fold. The crystal structure of the protein has now been published,[284] and Figure 33 compares it with the predicted structure.

The experimental structure of the von Willebrand factor turned out to differ from that of ras only in the orientation of two β strands. This template was then used to search the database for proteins with similar secondary structures ($\alpha-\beta$). The ras-p21

```
PGEQQKRKIVLDPSGSMNIYLVLDGSDSIGASNFTGAKKCLVNLIEKVASYGVKPRYGLYTY   prediction sequence
            |....:|.::!|||.||:..:|...|:.:.::!|:...  .|..!:|..|
            PQQESDIVFLIDGSGSINNIDFQKMKEFVSTVMEQFKK__SKTLFSLMQY   experimental sequence
  TTTTTTTTTTTTCHHEEEEEEETCCCCCCCCHHHHHHHHHHHHHHTCCCTCEEEEEEEET   prediction
                  EEEEEEE      HHHHHHHHHHHHHHHH      EEEEEEEEE   experimental
                  strand 1

ATYPKIWVKVSEADSSNADWVTKQLNEINYEDHKLKSGTNTKKALQAVYSMMSWPDDVPPEGWN prediction sequence
:.  !|....:!.:.:.:.  ..::::.|:. :.!.|::::.!|:::.:!...:       .:.:
SDEFRIHFTFNDFKRNPSP__RSHVSPIKQLNGRTKTASGIRKVVRELFHKTN_______GARE experimental sequence
CCCCCEEEHHCCCCCHHHHHHHHHBBTTTCC______CCCCCEHHHHHHHEEECCCCCCCC___ prediction
   EEE   HHHHHHHH    HHHHH          HHHHHHHHHHHH                 experimental

RTRHVILLMTDGLHNMGGDPITVIDEIRDLLYIGKDRKNPREDYLDVYVFGVGPL          prediction sequence
.:.:!!::!||| :   |||!...|.|.!. ..|. |          ||:|||..
NAAKILVVITDGEKF__GDPLDYKDVIPEADRAGVIR__________YVIGVGNA          experimental sequence
TCHEEEEEEEECCCCCCCHHHHHHHH_____________HTTEEEEEEEECCCC           prediction
  EEEEEEEE            HHHHHHHHHH   EE          EEEE              experimental
```

Figure 33. Representative sequences, *bona fide* consensus prediction,[282] and experimental secondary structure[284] for the von Willebrand factor type A domain. Key: E, β strand; H, α helix; T, turn; C, coil. In the comparison of sequences, vertical lines (|) indicate identical amino acids, exclamation points (!) indicate conservative substitution, colons (:) indicate less conservative substitution; underscoring (_) indicates insertion or deletion (indel). Prediction sequence is part of human complement factor B (EC 3.4.21.47, P00751, CFAB_HUMAN). The experimental sequence is the mouse cell surface glycoprotein MAC-1 α subunit (P05555, ITAM_MOUSE). The two proteins are 180 PAM units distant.

came closest to the predicted pattern, and this was used as a template for threading. An alignment with the ras-p21 was then made. As with the prediction for the cytokine receptor, the strand orientation was not precisely as predicted, with an edge and an internal strand swapped.[284] However, a metal binding site was predicted.

Several interesting features of the errors are instructive. For example, both the first helix and the first strand were correctly predicted. In the alignment with ras-p21, however, both secondary structural elements were predicted to be disrupted by an indel. This raised questions about the correlation between the von Willebrand structure and the ras-p21 structure. Further, helix 2 in the prediction is found in the experimental structure as two helices. Nevertheless, the overall packing was not affected by this mistake, illustrating a degree of tolerance of errors in noncore regions.

4. Protein Tyrosine Phosphatase

Livingstone and Barton applied methods similar to those used with Factor XIII to construct a prediction for the protein tyrosine phosphatase family.[281] The crystallographers subsequently noted that the prediction was very good when evaluated against their experimental structure, in particular with respect to the core secondary structural units.[285] Figure 34 shows the predicted and experimental structures. With the exception of a single β strand assigned to a region that is helical (224−227), the prediction is free of serious mistakes.

5. Protein Serine/Threonine Phosphatases

Both the Florida group and the Oxford group performed evolutionary analyses of the protein serine/threonine phosphatases.[96,286] The predicted secondary structures, together with the experimental structure assigned to coordinates from calcineurin,[287] are shown in Figure 35. The Florida prediction identified correctly every helix and strand in the core domain, with the exception of a single region that passes near the active site. The treatment of this active-site region in the prediction was important. By the time that this prediction was made, tools for identifying active-site regions had been developed, and the implications of an active-site region on the accuracy of a secondary structure prediction were understood. In the prediction paper, this ambiguity was presented with the following discussion:[96]

"The helix assigned to segment (246−262) contains a conserved tripeptide RxH that is plausibly (but not definitely, see below) placed at the active site. The segment displays convincing 3.6 residue periodicity if residue 254 is assigned to the surface. To observe this periodicity requires, however, assignment of a conserved R (251) to the surface and a conserved H (253) and a conserved G (259) to the inside. Further, the DG element at positions (258−259) is a weak parsing element. Thus, a second, weaker, assignment separates this segment into two shorter elements separated by an active site coil. In tertiary structure modelling, this alternative assignment must be considered."

In the experimental structure, two shorter strands, separated by an active-site coil were observed (Figure 35). Thus, the prediction provided two alternative secondary structure models, one entirely correct. This underscores again the need to develop accurate tools for evaluating alternative packings that might allow selection between a small number of alternative secondary structural models. Further, the example provides another illustration of the difficulty in predicting secondary structure near an active site.

Similar difficulties are seen at positions 071−074. The experimental structure starts helix 3 in this region. The Florida prediction, recognizing the active site, did not. The Oxford group mispredicted this as a strand. Again, the reason is that patterns of conservation that reflect active sites mask patterns that would normally be used to assign secondary structure. From an analysis of the structure overall,

```
         0         0         0         0         0         0         0         0
         1         2         3         4         5         6         7         8
         0         0         0         0         0         0         0         0
MEMEKEFEQIDKSGSWAAIYQDIRHEASDFPCRVAKLPKNKNRNRYRDVSPFDHSRIKLHQEDNDYINASLIKMEESQRS sequence
     HHHHHHH   HHHHHHHHHHH     HHHHH                   EE          EEEEEEE    EE experimental
not part of the multiple alignment|          EEEEE      EEEEEEE   EEEEEEE      E prediction


         0         1         1         1         1         1         1         1
         9         0         1         2         3         4         5         6
         0         0         0         0         0         0         0         0
YILTQGPLPNTCGHFWEMVWEQKSRGVVMLNRVMEKGSLKCAQYWPQKEEKEMIFEDTNLKLTLISEDIKSYYTVRQLEL sequence
EEEEE     HHHHHHHHHHHH   EEEE    EE  EE             EEE    EEE    EEEE  EEEEEEEE experimental
EEEEE     HHHHHHHHHHH    EEEEE        EEEEEEE             EEEEEEE      EEEEEEEEE prediction


         1         1         1         2         2         2         2         2
         7         8         9         0         1         2         3         4
         0         0         0         0         0         0         0         0
ENLTTQETREILHFHYTTWPDFGVPESPASFLNFLFKVRESGSLSPEHGPVVVHCSAGIGRSGTFCLADTCLLLMDKRKD sequence
EE     EEEEEEEEE           HHHHHHHHHHHHHH         EEEE    HHHHHHHHHHHHHHHHHHH    experimental
E       EEEEEEEE              HHHHHHHH            EEEEE        EEEE HHHHHHH      prediction


         2         2         2         2         2         3
         5         6         7         8         9         0
         0         0         0         0         0         0
PSSVDIKKVLLEMRKFRMGLIQTADQLRFSYLAVIEGAKFIMGDSSVQDQWKELSHEDLE sequence
     HHHHHHHHH         HHHHHHHHHHHHHHHH                      experimental
     HHHHHHHH      EEEE   HHHHHHHHH | end of alignment       prediction
```

Figure 34. Representative sequences, *bona fide* consensus prediction,[281] and experimental secondary structure[288] for the protein tyrosine phosphatase family. Key: E, β strand; H, α helix.

the Florida group built a mechanistic model for the phosphatase (see below) based on two active-site metals. This required the identification of a larger number of active-site functionality than would normally be found in an enzyme of this type.

As in the phospho-β-galactosidase prediction (see below), the most significant mistakes identified by comparison of the prediction with a crystal structure for a single member of the family lay in the nonconserved extra domain. In Figure 35, the divergence of secondary structure is most obvious at positions 220−222. Here, the three residues assigned to a strand in the protein whose crystal structure was solved are missing in the alignment of almost all other proteins. This is almost certainly a problem with the multiple alignment.

Despite these issues, the secondary structure prediction was adequate to allow the prediction of the central features of the supersecondary structure. A parallel core β sheet was correctly predicted, as was the packing of separate β−α−β units.

The prediction proved to have more than academic implications. The Florida prediction was prepared for an industrial collaborator, who was interested in the mechanistic implications of the structure. The model predicted that the phosphatase would have two metals in the active site and catalyze the hydrolysis of the phosphate using a two-metal mechanism. The crystal structure is consistent with this proposal. The model also allowed the identification of loops as appropriate targets for peptide-based epitopes. Thus, consensus predictions can have practical value, even when they are at low resolution.

This prediction further illustrates the need to build preliminary tertiary structural models as a first step toward evaluating the plausibility of a prediction. This is similar to using a secondary structural model to evaluate the plausibility of parses and surface/interior assignments.

Last, it is interesting to compare the Oxford and Florida predictions for the protein serine/threonine phosphatases. Both groups are using similar methods, even though the underlying conceptual basis for the two approaches differ somewhat. The predictions are different only in their details, and even these can be understood if one understands the differences in the approach. For example, the Oxford prediction misassigns strand 1 (positions 023−027) as an extension of helix 1; the Florida prediction terminated the helix at the correct point. The transparency of the prediction allows us to understand the difference. The PN dipeptide found at positions 021 and 022 in many of the homologs is a “dipeptide parse” (see above), and caused the predictors in Florida to terminate the helix. The dipeptide parsing tool is implemented in Florida, but not in Oxford.

6. The Proteasome

The proteasome is the central enzyme of nonlysosomal protein degradation, and its 20S core is conserved from archaebacteria to humans. A low-resolution model shows that the protein is cylindrical and is built from two subunits (in the archaebacterium), termed α and β. The α and β subunits are themselves homologous (Figure 36), with approximately 26% overall sequence identity.

Using a set of aligned sequences, Lupas *et al.* predicted a consensus secondary structure for the α

Position		Conser	Inte-	Sur-		Multiple		Alignment			Predicted		Experi
Align	Targ	ved	rior	face	tr	JxuzFBCAGDEyvKL	h	fdnolmkjiec	IHgw	qspa	Sec	Struc	mental
											Florida	Oxford	
001	56		0.30		ll	lLLLLLlLLLxLLlf	L	LLLLLLLLLLL	IIIl	lvvv			
002	57			3.56	ss	kSSELSsTSSxTTkh	P	SSSTNNNSSSS	KKPt	eddd			
003	58			0.90	ak	eDEEEEgEEExEEns	E	EEEEEEEEEEE	EEEe	eeee	H		
004	59			4.00	ae	eADNSGyNSGxAAnw	V	DDQTNNNSSAE	SSEn	seee	H		H
005	60			0.33	qq	lEEEEEEEEExEEEE	T	DDQQQQQQQDD	EETe	viii	H	H	H
006	61		0.30		aa	iIIIIIIIIIxVVII	V	VVVVVVVVVVV	VVVm	aaaa	H	H	H
007	62			5.35	ai	eRRRRRRRKRxRRLQ	R	AARRRRRKKEI	KKFk	llll	H	H	H
008	63			3.01	rk	rYFYYGGGQQxWWQL	A	RRATTTTSSMQ	AARQ	rrrr	H	H	H
009	64		0.45		ii	lLLLLLLLLLxLLII	L	LLLLLLLLLLM	LLLL	iiii	H	H	H
010	65		0.43		vl	iCCCCCCCCCxVVCC	C	CCCCCCCCCCC	CCCC	iiii	H	H	H
011	66			2.63	tn	qTNSTLLLAAxMMIY	F	KKEDEEEEEDD	AALE	tnnn	H	H	H
012	67			0.91	lm	qTKKKKKKAVxEEKH	K	MMKKKKKKKKL	KKNM	eeee	H	H	H
013	68		0.30		AS	tSAAASSSASxSSAA	L	AAAAAAAAAAA	AASV	gggg	H	H	H
014	69			0.83	TT	rRRRRRRRKKxRRRR	K	VVKKKKKKKRK	RRQK	aaaa	H	H	H
015	70			1.75	EV	eSESEEEEEExAAEE	E	DDEEEEEEEEE	EEEE	saaa	H	H	H
016	71		0.45		LA	vIIIIIIIIIxLLII	M	VVIIIIIIIVV	IILL	IIII	H	H	H
017	72		0.45		FL	iFFFFFFFFFxFFFF	L	LLLLLLLLLLL	LLLL	LLLL	H	H	H
018	73			0.81	SS	kLIIILLLLLxMMLL	V	QQMSTTTTTCS	VVMM	RRRR	H	H	H
019	74			1.75	KK	wSSKSSSSQQxSSSN	K	FFDKKKKKKQV	EENE	QRRR	H	H	H
020	75			0.03	EE	QQQQQQQQQQxQQQQ	E	EEEEEEEEEEE	EEEE	EEEE		H	
021	76			0.41	PP	PPPPPPPPPPxPPPP	S	EESSSSSSSNS	SSGS	KKKK		H	
022	77			0.22	NN	MMIIIIIINNxMMST	N	NNNNNNNNNNN	NNNN	NTTT		H	
023	78		0.45		LL	LLLLLLLLLLxLLLL	V	VVVVVVVVVVV	VVVI	LMMM		H	E
024	79			0.17	IL	LLLLLLLLLLxVVLL	I	KKQQQQQQQQQ	QQTQ	LIII		H	E
025	80			0.93	SK	EEEEEEEEEExEEER	H	PPPEEEEEEPS	RRQP	DEEE		H	E
026	81		0.45		VL	LLLLLLLLLLxIILL	I	IIVVVVVVVVV	VVVV	IVVV		H	E
027	82			2.65	PK	QEEEEEEEEExAASQ	Q	NNKKRRRRRRR	DDDQ	DEEE			E
028	83			0.19	AA	AAAAAAAAAAxAAPA	T	VVSCCCCCCNC	SSTT	AAAA			
029	84				PP	PPPPPPPPPPxPPPP	P	PPPPPPPPPPP	PPPP	PPPP			
030	85		0.45		II	VLLIILLLIIxVVVI	V	VVVVVVVVVVV	VVVV	VIII	E	E	E
031	86		0.08		TT	NKKKKKKKKKxRRKK	T	TTTTTTTTTTT	TTTT	TTTT	E	E	E
032	87		0.45		VI	IIIIIIIIIIxIIIV	V	IIIVVVVVVVV	VVIV	VVVV	E	E	E
033	88		0.43		CC	CCCCCCCCCCxCCVV	V	CCCCCCCCCCC	CCCC	CCCC	E	E	E
034	89	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGGG	GGGG	GGGG	E	E	E
035	90	D			DD	DDDDDDDDDDDDDDD	D	DDDDDDDDDDD	DDDD	DDDD	$		
036	91		2.60		II	IIIIIIIIVIIVVVV	M	VVIVVVVVVII	IIII	IIII	$		
037	92	H			HH	HHHHHHHHHHHHHHH	H	HHHHHHHHHHH	HHHH	HHHH	$		
038	93	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGGG	GGGG	GGGG	$		
039	94	Q			QQ	QQQQQQQQQQQQQQQ	Q	QQQQQQQQQQQ	QQQQ	QQQQ	$		
040	95		2.60		YY	FYYYYYYYYYYYYYF	F	FFFFFFFFFFF	FFLF	FFFF	$	H	H
041	96			0.22	FY	TSYYYYYYSSQTTGN	H	HHHHHHHHHHH	YYHH	FFFF	$	H	H
042	97	D			DD	DDDDDDDDDDDDDDD	D	DDDDDDDDDDD	DDDD	DDDD	$	H	H
043	98		2.60		LL	LLLLLLLLLLLLLLL	M	LLLLLLLLLLL	LLLL	LLLL	H	H	H
044	99		0.19		LL	LLLLLLLLLLLLLLL	L	LLAMMMMMMMM	KKLL	MMMM	H	H	H
045	100			0.23	KK	RRRRRRRRRRRRRRR	E	EEEEEEEEEEE	EETE	KKKK	H	H	H
046	101		2.60		LL	ILLLLLLLLLLLLLI	I	LLLLLLLLLLL	LLLL	LLLL	H	H	H
047	102		2.60		FF	FFFFFFFFFFFFFFL	F	FFFFFFFFFFF	FFFF	FFFF	H	H	H
048	103			0.96	EE	KEEEEEEEDEEDDTK	Q	KKRRRRRRRKN	RRER	EEEE	H	H	H
049	104			0.13	VV	AYYYYYYYYYYLLKL	I	IIIIIIIIIII	VVKT	VVVV	H	H	H
050	105				GG	CGGGGGGGGGGGGCS	G	GGGGGGGGGGG	GGSA	GGGG			H
051	106	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGGG	GGGG	GGGG			
052	107			0.96	DD	_______________	P	PPMKKKKKKDP	DDGG	SSSS			
053			0.42		__	FYFFFFFFYFYFFFV	V	CCCSSSSSSVS	VVVF	____			
054	108				PP	PPPPPPPPPPPPPPP	P	PPPPPPPPPPP	PP P	PPPP			
055	109			0.14	AA	PPPPPPPPPPPPPPS	_	___________	EEED	AAAA			
056	110			1.31	TE	KDEEEEEEQESDDSD	D	DDDDDDDDDDD	__KD	NNNN			
057	111			21.36	TI	AAASASSSAAAAAST	T	TTTTTTTTTMT	TRTI	TTTT			
058	112			0.46	SD	NNNNNNNNNNNNNNN	N	NNNNNNNNNNN	NNRN	RRRR			E
059	113		2.60		YY	YYYYYYYYYYFYYYY	Y	YYYYYYYYYYY	YYYY	YYYY	E	E	E
060	114		2.60		LL	LLLLLLLLLLLIILL	L	LLLLLLLLLLL	LLII	LLLL	E	E	E
061	115	f	2.60		FF	FFFFFFFFFFFFFFF	F	FFFFFFFFFFF	FFFF	FFFF	E	E	E
062	116		2.60		LL	LLLLLLLLLLLLLLL	L	MMMMMMMMMMM	MMLL	LLLL	E	E	E
063	117	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGGG	GGGG	GGGG	$		
064	118	D			DD	DDDDDDDDDDDDDDD	D	DDDDDDDDDDD	DDDD	DDDD	$		
065	119		2.60		YY	YYYYYYYYYYYYYYY	Y	YYYYYYYYYYY	FFFY	YYYY	$		
066	120	v	2.60		VV	VVVVVVVVVVVVVVV	V	VVVVVVVVVVV	VVVV	VVVV	$		

067	121	D			DD	DDDDDDDDDDDDDDD	D	DDDDDDDDDDD	DDDD	DDDD	$		
068	122	R			RR	RRRRRRRRRRRRRRR	R	RRRRRRRRRRR	RRRR	RRRR	$		
069	123	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGGG	GGGG	GGGG	$		
070	124		0.17		SA	KKKKKKKKKKKDDKK	L	YYYYYYYYYYY	FFFY	YYYY	$		
071	125			0.09	FF	QQQQQQQQQQQQQQN	Y	YYYYYYYYYHH	YYYY	FFFF	$	E	H
072	126	S			SS	SSSSSSSSSSSSSSS	S	SSSSSSSSSSS	SSSS	SSSS	$	E	H
073	127		2.60		FF	LLLLLLLLLLLLLLL	V	VVVVVVVVVVV	VVLL	IIII	$	E	H
074	128	E			EE	EEEEEEEEEEEEEEE	E	EEEEEEEEEEE	EEEE	EEEE	$	E	H
075	129			0.33	CC	TVVTTTTTTTTTTTT	T	TTTTTTTTTTT	TTST	CCCC	H		H
076	130		2.60		LL	IIIIIIIIIIIIIII	I	VVVVVVVVVVV	FFFF	VVVV	H		H
077	131		0.05		II	CCCCCCCCCCCCCLL	M	SSTTTTTTTSS	LLLT	LLLL	H	?	H
078	132		2.60		YY	LLLLLLLLLLLLLLL	L	YYLLLLLLLLL	LLLL	YYYY	H	?	H
079	133	l	2.60		LL	LLLLLLLLLLLLLLL	L	LLLLLLLLLLL	LLLL	LLLL	H	?	H
080	134		2.60		YY	FFLLLLLLLLLLLFL	I	VVVVVVVVVVI	LLLM	WWWW	H	?	H
081	135		0.40		SS	AAAAAAAAAAAAACC	V	AAGAAAAAAAA	AACC	AVVV	H	?	H
082	136		2.60		LL	YYYYYYYYYYYYYYY	L	MMLLLLLLLMF	LLYL	LLLL	H	?	H
083	137	K			KK	KKKKKKKKKKKKKKK	K	KKKKKKKKKKK	KKKK	KKKK	H	?	H
084	138		2.60		LL	VIIIIIIIVIILLII	L	VVVVVVVVVLI	VVLV	IIII	H	?	H
085	139			0.37	NN	KKKKKKKKKKRSSKK	R	RRRRRRRRRRR	RRRK	LLLL	H	?	H
086	140		0.67		FN	YYYYYYYYYYYFFYY	Y	YYYYYYYYYYY	YYYY	YYYY	H	?	H
087	141			5.40	NL	PPPPPSPPPPPPPPK	P	PPPRPPPRRPP	PPPP	PPPP		?	
088	142			1.12	DG	LEEEEEEEEESEEED	S	HHQEEEEEENQ	DDDA	KSSS			
089	143			0.62	HR	NNNNNNNNNNKTTNN	R	RRRRRRRRRRR	RRRK	TTTT			
090	144		2.60		FF	FFFFFFFFFFIFFFF	I	IIIIIIIIIII	IIII	LLLL	E		E
091	145		1.92		WW	FFFFFFFFFFYFFFF	H	TTTTTTTTTTT	TTTT	FFFF	E	E	E
092	146		2.60		LM	LLIIVLLLLLLLLLM	L	IIIIIIIIIII	LLLL	LLLL	E	E	E
093	147		2.60		LL	LLLLLLLLLLLLLLL	L	LLLLLLLLLLL	IIIV	LLLL	E	E	
094	148	R			RR	RRRRRRRRRRRRRRR	R	RRRRRRRRRRR	RRRR	RRRR	$	E	
095	149	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGGG	GGGG	GGGG	$		
096	150	N			NN	NNNNNNNNNNNNNNN	N	NNNNNNNNNNN	NNNN	NNNN	$		
097	151	H			HH	HHHHHHHHHHHHHHH	H	HHHHHHHHHHH	HHHH	HHHH	$		
098	152	E			EE	EEEEEEEEEEEEEEE	E	EEEEEEEEEEE	EEEE	EEEE	$		
099	153			0.22	CC	CFCCCCCCCSDCCCS	S	SSSSSSSSSSS	SSTS	CCCC			
100	154		0.04		KK	AAAAAAAAAAASSAA	R	RRRRRRRRRRR	RRRR	RRRR			H
101	155			0.01	HH	SSSSSSSSSSKSSNN	Q	QQQQQQQQQQQ	QQQQ	HHHH	H		H
102	156		2.60		LL	IIIIIIIIVIIIIVV	I	IIIIIIIIIII	IIII	LLLL	H		H
103	157		0.16		TT	NNNNNNNNNNNNNTT	T	TTTTTTTTTTT	TTTT	TTTT	H		H
104	158			0.46	SS	KRRRRRRRRRRRRRK	Q	QQQQQQQQQQQ	QQKQ	EEEE	H	H	H
105	159		0.42		YY	IIIIIIIIIIIIIVM	S	VVVVVVVVVVV	VVVV	YYYY	H	H	H
106	160		2.60		FF	YYYYYYYYYYYYYYY	Y	YYYYYYYYYYY	YYYY	FFFF	H	H	
107	161			0.08	TT	GGGGGGGGGGGGGGG	G	GGGGGGGGGGG	GGGG	TTTT	H	H	
108	162	f	2.60		FF	FFFFFFFFFFFFFFF	F	FFFFFFFFFFF	FFFF	FFFF	H	H	H
109	163		0.75		KK	YYYYYYYYYYYFFYY	Y	YYYYYYYYYYY	YYYY	KKKK	H	H	H
110	164			0.16	NN	DDDDDDDDDDDDDDD	T	DDDDDDDDDDD	DDDE	QQQQ	H		H
111	165	E			EE	EEEEEEEEEEEEEEE	E	EEEEEEEEEEE	EEEE	EEEE	H		H
112	166		0.40		MM	ICCCCCCCCCCCCCC	S	CCCCCCCCCCC	CCVC	CCCC	H		H
113	167		0.24		LL	KKKKKKKKKKKKKKK	L	LLLLLLLLLLL	LLVL	KKKK	H		H
114	168			0.14	HH	RRRRRRRRRRRRRRR	N	RRRRRRRRRRR	RRRN	IIII	H		H
115	169			0.21	KK	RRRRRRRRRRRRRRR	K	KKKKKKKKKKK	KKKK	KKKK	H		H
116	170		1.90		YY	HYYYYYYYFFFYYCL	Y	YYYYYYYYYYY	YYYY	YYYY	H		
117					__	_______________	G	___________	____	____			
118					__	_______________	G	GGGGGGGGGGG	GGGG	____			
119	171			1.80	ND	TSNNNSNNSNNSSNS	N	SSNNNNNNNSN	SSNS	SSSS	H		
120	172		0.15		LM	VIIIIIIIVVVVVIS	S	AAAAAAAAAAA	VVST	EEEE	H		
121	173			0.62	DE	RKKKKKKKKRRRRKK	R	NNNNNNNNNNN	TTNT	RRRR	H		H
122	174		2.60		IV	LLLLLLLLLLLLLIV	V	VVVVVVVVVVV	VVVV	VVVV	H		H
123	175		2.60		YY	WWWWWWWWWWWWWWW	W	WWWWWWWWWWW	WWWW	YYYY	H		H
124	176			0.93	ED	HKKKKKKKKKKKKKK	Q	KKKKKKKKKKQ	RRRK	DEEE	H		H
125	177			0.42	KA	NTTTTTTTTIIQQTM	Y	MMYYYYYYYHY	YYYY	AAAA	H		H
126	178		2.37		CC	FFFFFFFFFFFFFFF	L	FFFFFFFFFFF	CCCC	CCCC	H		H
127	179		1.77		CC	TTTTTTTTTTTTTIV	T	TTTTTTTTTTT	TTCC	MMMM	H		H
128	180			0.70	ER	DDDDDDDDDDDDDDD	D	DDDDDDDDDND	EEEQ	DEEE	H		H
129	181		0.36		SS	CCCCCCCCCCCTTTV	I	LLLLLLLLLLL	IIVV	AAAA	H		H
130	182	f	2.60		FF	FFFFFFFFFFFFFFF	F	FFFFFFFFFFF	FFFF	FFFF	H		H
131	183			0.21	NN	NNNNNNNNNNNNNNN	D	DDDDDDDDDDD	DDDD	DDDD			
132	184		0.40		NV	WCCCCCCCCCCCCTT	Y	YYYYYYYYYYY	YYYF	CSSS			
133	185		2.60		LL	LMLLLLLLLLLMMLL	L	FFLLLLLLLFL	LLLL	LLLL			
134	186			0.18	PP	PPPPPPPPPPPPPPP	V	PPPPPPPPPPP	SSST	PPPP			
135	187		2.60		LL	VVIIIVIIVVVVVLL	L	VILLLLLLLLL	LLLL	LLLL	E		
136	188		0.06		AA	AAAAAAAASAAAAAA	C	TTTTTTTTTTT	SSGA	AAAA	E		E
137	189		0.43		AA	AAAAAAAAAAAGGAA	C	AAAAAAAAAAA	AAAA	AAAA	E		E

138	190		2.60		LL	LVIIIIIILLLLLII	I	LLLLLLLLLL	IIII	LLLL	E		E
139	191		2.60		MM	VIIIIVVVIIIVVVI	I	VVVVVVVVII	IIII	MLLL	E		E
140	192			0.74	NN	GDDDDDDDDDDEEAQ	D	DDDDDDDDEE	DDND	NNNN			E
141	193			0.77	GG	EEEEEEEEEDDGGGD	D	NNSGGGGGDD	GGNG	QQQQ			
142	194			0.84	QQ	RKKKKKKKKRKRRKK	E	KKEQQQQQRR	KKSK	QQQQ			
143	195		2.60		YY	IIIIIIIIIIIIIII	I	IIIIIIIIII	IIII	FFFF	E		E
144	196		2.60		LF	FFFFFFFFLLLLLFF	F	FFFFFFFFFF	FFFL	LLLL	E	E	E
145	197		1.32		CC	CCTCTCCCCCCCCCC	C	CCCCCCCCCC	CCCC	CCCC	E	E	E
146	198		2.37		VV	CMMMMCCCMMMMMVV	V	LLLLLLLLLL	VVVV	VVVV	E	E	E
147	199	H			HH	HHHHHHHHHHHHHHH	H	HHHHHHHHHH	HHHH	HHHH	$	E	
148	200	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGG	GGGG	GGGG	$		
149	201	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGG	GGGG	GGGG	$		
150	202		2.60		II	LLLLLLLLLILLLLI	L	LLLLLLLLLL	LLLL	LLLL	$		
151	203	s			SS	SSSSSSSSSSSSSSS	S	SSSSSSSSSS	SSSS	SSSS	$		
152	204	p			PP	PPPPPPPPPPPPPPP	P	PPPPPPPPPP	PPPP	PPPP	$		
153	205			0.65	EE	SDDDDDDDEEEEEVD	N	MMSSSSSSSS	SSDE	EEEE			
154	206		2.60		LL	LLLLLLLLLLLLLLL	V	IIIIIIIIII	IIMI	IIII	3		
155	207			1.37	NK	RNNNNTQQNKDTTNH	Q	EEEDDDDDDD	QQTR	NHHH	3	H	
156	208			0.93	SS	NSSSSSSSKSNDDSD	T	TTTSTTTTST	TTTM	TTTT	3	H	
157	209		2.60		LV	LLMMMMMMLLLLLMM	I	IILLLLLLLL	LLVL	LLLL	3	H	
158	210			1.35	QE	QDDEEEEEEDNDDDK	D	DDDDDDDDDD	DDDD	DDDD	3	H	H
159	211			0.35	DD	QQQQQQQQQQQQQEQ	Q	QQNHHHHHHH	QQEQ	DDDD	3	H	H
160	212		2.60		IV	IIIIIIIIIIIIIII	I	VVIIIIIIVV	IIII	IIII	3	H	H
161	213			1.35	NN	NQQRRRRRLRRRRRE	K	RRRRRRRRRR	RRRR	RRRR	3	H	H
162	214			10.65	NK	HRRRRRRRNNERRHK	I	DENAAAAATI	TTTV	KRRR	3	H	
163	215		2.60		LI	IIIVVIIILIIIIVV	I	LLFLLLLLLL	IIIL	LLLL	3	H	
164	216			0.46	NN	QIMMMMMMNAQLLVA	D	NNDDDDDDDD	DDDS	DDDD	3	H	
165	217	R			RR	RRRRRRRRRRRRRRR	R	RRRRRRRRRR	RRRR	RRRR	3		
166	218		0.07		FF	PPPPPPPPPPPPPPP	F	IIVLLLLLVV	KKKA	FFFF	3		
167	219			0.09	RR	TTTTTTTTTMTTTTT	R	QQQQQQQQQQ	QQQQ	KKKK			
168	220			0.38	EE	DDDDDDDDDDEDDDD	E	EEEEEEEEEE	EEEE	EEEE			
169	221		0.94		II	IIVIIVVVVIIVVVI	I	VVVVVVVVVV	VVVV	PPPP			
170	222	p			PP	PPPPPPPPPPPPPPP	P	PPPPPPPPPP	PPPP	PPPP			
171	223			0.29	SS	DDDDDDDDDEDDDDE	H	HHHHHHHHHH	HHHH	AAAA			
172	224			0.83	HR	ETTVCQQQTSSSSFS	D	EEGEEEEEEE	DDEE	YFFF			
173	225	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGG	GGGG	GGGG			
174	226		0.42		LL	ILLLLLLLLLLLLLL	A	PPPPPPPPPP	PPAG	PPPP			E
175	227		2.60		MM	MLLLLLLLLVLIIIV	M	MMMMMMMMMI	MMMF	MMMM	E		E
176	228		0.10		CC	CCCCCCCCCCCCCNT	A	CCCCCCCCCC	CCCS	CCCC	E		E
177	229	D			DD	DDDDDDDDDDDDDDD	D	DDDDDDDDDD	DDDD	DDDD	E		E
178	230	l	2.60		LL	LLLLLLLLLLLLLLL	L	LLLLLLLLLL	LLLL	ILLL	E		E
179	231		2.60		LL	LVLLLLLLLLLLLLL	V	LLLLLLLLLL	LLLL	LLLL	E		E
180	232	w	2.60		WW	WWWWWWWWWWWWWWW	W	WWWWWWWWWW	WWWW	WWWW	E		
181	233		0.06		AA	ASSSSSSSSSSSSSS	S	SSSSSSSSSS	SSSS	SSSS			E
182	234	D			DD	DDDDDDDDDDDDDDD	D	DDDDDDDDDD	DDDD	DDDD			E
183	235				PP	LPPPPPPPPPPPPPP	P	PPPPPPPPPP	PPPP	PPPP			
184	236			0.84	IV	NEDDDDDDSSDSSTD	E	DDDDDDDDDD	EEED	L___			
185	237			0.53	EE	HKKKKKKKNGQTTDP	E	__________	____	E___			
186				0.30	EN	_______________	N	__________	____	____			
187				0.11	YY	_______________	N	__________	____	____			
188				0.19	DD	_______________	N	__________	____	____			
189				0.30	ED	_______________	P	__________	____	____			
190				0.09	VA	_______________	T	__________	____	____			
191					L_	_______________	L	__________	____	____			
192					D_	_______________	D	__________	____	____			
193				0.30	KR	_______________	H	__________	____	____			
194				0.19	DD	_______________	P	__________	____	____			
195				0.65	LG	TDDDDDDDEDKNNSQ	D	DDDDDDDDDD	DDDN	____			
196				0.76	TS	TLLIITVVAVIMMPV	N	RRRRRRRRRR	TTVV	____			
197				1.46	EE	KTTVTMQQT_EEENT	S	GGCGGGGGCP	TTDE	_SSS			
198				0.35	EF	GGGGGGGGGGGSS__	G	GGGGGGGGGG	GGTA	_EEE			
199	238			0.26	DD	___________NNED	Q	__________	____	DDDD			
200	239		0.09		I_	_______________	_	__________	____	FFFF			
201	240				V_	_______________	_	__________	____	GGGG			
202	241				N_	_______________	_	__________	____	NNNN			
203	242			0.19	S_	_______________	_	__________	____	EEEE			
204	243				K_	_______________	_	__________	____	KKKK			
205	244				T_	_______________	_	__________	____	T___			
206					M_	_______________	_	__________	____	____			

207					V_	_______________	_	___________	____	____			
208					P_	_______________	_	___________	____	____			
209					H_	_______________	H	___________	____	____			
210			0.78		H_	WWWWWWWWWWWWWWW	F	WWWWWWWWWWW	WWWW	____			
211				0.39	G_	GGGSSGGGAGASSES	Q	GGGGGGGGGGG	GGSQ	____			
212				0.90	K_	HEDEEEEEIMDEEDE	V	IIIIIIIIIII	VVLV	____			
213					M_	_______________	_	___________	____	____			
214					A_	_______________	_	___________	____	____			
215				0.25	PQ	_______________	_	___________	____	____			
216					SS	_______________	_	___________	____	_SSS			
217	245			0.23	RE	_______________	_	___________	____	QQQQ			
218	246			0.19	DD	_______________	_	___________	____	EEEE			
219	247			0.23	ME	_______________	_	___________	____	HHHH			
220	248		0.09		FF	_______________	_	___________	____	FFFF			E
221	249			0.09	VV	_______________	_	___________	____	TSSS			E
222	250			0.09	PP	_______________	_	___________	____	HHHH			E
223	251				NN	_______________	_	___________	____	NNNN			
224	252			0.25	SS	NNNNNNNNNNSNNNN	S	SSSSSSSSSSS	SSSS	TTTT			
225	253			0.17	VL	DDDDDDDDDDDDDED	P	PPPPPPPPPPP	PPPP	VVVV			
226	254	R			RR	RRRRRRRRRRRRRRR	R	RRRRRRRRRRR	RRRR	RRRR			
227	255	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGGG	GGGG	GGGG			
228	256		2.37		CC	VVVVVVVVVVIVVVV	A	AAAAAAAAAAA	AAAA	CCCC	E		
229	257				SS	SSSSSSSSSSSSSSS	G	GGGGGGGGGGG	GGGG	SSSS	E	E	
230	258		2.37		YF	FYFFFFFFFYCWWYY	Y	FFYYYYYYYYY	YYFW	YYYY	E	E	
231	259		1.77		AA	TTTTTTTTTTTTTCT	T	TTTTTTTTTTT	LLLL	FFFF	E	E	E
232	260		2.60		FF	FFFFFFFFFFFFFYF	F	FFFFFFFFFFF	FFFF	YYYY	E	E	E
233	261			0.39	TT	DGGGGGGGGGGSSNS	G	GGGGGGGGGGG	GGGG	SNNN		E	E
234	262			0.45	YF	KAPPPAAAPAAEEKK	R	QQQQQQQQQQP	SSKS	YYYY			E
235	263			1.20	RK	VDDDDEEEDDDSSVR	S	DDDDDDDDDDD	DDRK	PPPP			
236	264			0.32	AA	IVVVVVVVKKKVVAN	V	IVIIIIIIIII	VVEV	AAAA	H		H
237	265		0.42		AS	VVVVVVVVVVVVVIV	V	SSSSSSSSSSA	VVVA	VVVV	H	H	H
238	266			1.26	CC	RSSNSAAASAAKKNL	E	EENEEEEEEEE	AADR	CCCC	H	H	H
239	267			0.96	HK	DRRRRKKKEEESSKD	K	QQQTTTTTTTA	QQQE	DEEE	H	H	H
240	268	f	2.60		FF	FFFFFFFFFFFFFF	F	FFFFFFFFFF	FFFF	FFFF	H	H	H
241	269			0.16	LL	LLLLLLLLLLLNNLC	L	NNNNNNNNNNN	NNLN	LLLL	H	H	H
242	270			3.46	QK	KQHQQQHHEEDKKNA	R	HHHNHHHHHHH	AAEH	QQQQ	H	H	H
243	271			0.90	EA	AKKKKKKKKKKKKKK	M	TTSTAAAAAAN	AAKV	HNNN	H		H
244	272			1.10	TN	FHHQHHHHHHNFFFF	N	NNNNNNNNNNN	NNNN	NNNN			H
245	273			2.70	GG	DDDDDEDDDDDDDGK	D	DDSGGGGGGGG	DDNG	NNNN			
246	274		2.60		LL	LLMMMFLLLMLLLFF	M	LLLLLLLLLLL	IIVL	LLLL	H/E		
247	275			1.68	LL	QDDEDDDDDDDDDDD	N	SSKTTTTTTSD	DDEN	LLLL	H/E	H	
248	276		0.20		SS	LLLLLLLLLLLLLLL	R	LLLLLLLLLLL	MMLL	SSSS	H/E	H	E
249	277		1.79		II	MIVIIIIIIIIIIVI	I	IIIVVVVVVTI	IIII	IIII	H/E	H	E
250	278		0.40		II	VCCCCCCCCCCCCCL	Y	AASSSSSSSAA	CCAA	LIII	H/E	H	E
251	279	R			RR	RRRRRRRRRRRRRRR	R	RRRRRRRRRRR	RRRR	RRRR	H	H	E
252	280		0.45		AA	AAAAAAAAAAGAAAG	A	AAAAAAAAAAA	AAAA	AAAA	H	H	
253	281	H			HH	HHHHHHHHHHHHHHH	H	HHHHHHHHHHH	HHHH	HHHH	H	H	
254	282		0.11		EE	EQQQQQQQQQQQQMM	Q	QQQQQQQQQQQ	QQQQ	EEEE	H	H	
255	283		2.60		AA	VVVVVVVVVVVVVVV	L	LLLLLLLLLLL	LLLL	AAAA	H	H	
256	284		1.95		QQ	VVVVVVVVVVVVVVV	C	VVVVVVVVVVV	VVVV	QQQQ	H		
257	285		0.03		DD	EEEEEEEEEEEDDEE	N	MMMMMMMMMMM	MMMM	DDDD	H		
258	286			0.48	AA	DDDDDDDDDDDAADD	E	EEDEEEEEEEE	EEEE	AAAA	H		
259	287	g			GG	GGGGGGGGGGGGGGG	G	GGGGGGGGGGG	GGGG	GGGG	H/E		
260	288		2.60		YY	YYYYYYYYYYYYYYY	Y	YYYYYYYYYFY	YYYF	YYYY	H/E		E
261	289			0.31	RR	EEEEEEEEEEEEEEE	Q	SANNNNNNNNN	KKKK	RRRR	H/E		E
262	290		0.19		MM	FFFFFFFFFFFFFFF	I	WWWWCWWWWWW	WWEY	MMMM	H/E		E
263	291			0.18	YY	_______________	Y	SSACCCCCCAT	HHIH	YYYY			E
264	292		0.23		KK	FFFFFFFFFFFFFFF	F	___________	FFFF	RRRR			
265	293			0.37	NN	AGSSSAAAAAAAANA	_	HHHHHHHHHHT	NN_P	KKKK			
266	294			1.39	TN	NKKKKKKKSEKAADR	D	QQEDDDDDDDN	EEDE	SSSS			
267	295			0.70	KK	RRRRRRRRRRRRRRK	G	QQARRRRRRGH	TTGK	QQQQ			
268	296			0.15	TV	_______________	_	___________	____	TTTT			
269	297			0.15	LT	_______________	_	___________	____	TTTT			
270	298				GG	_______________	_	___________	____	GGGG			
271	299		0.09		FF	_______________	_	___________	____	FFFF			
272	300				PP	_______________	_	___________	____	PPPP			
273	301			1.26	SS	QQQQQMQQQQRQQSK	L	NNKNNNNNNDN	__GD	SSSS			
274	302		0.42		LL	LLLLLLLLLLLLLLF	V	VVGVVVVVVVV	VVLV	LLLL	E		
275	303		0.25		LI	VVVVVVVVVVVVVVV	T	VVGVVVVVVVV	LLVV	IIII	E	E	E
276	304	T			TT	TTTTTTTTTTTTTTT	T	TTTTTTTTTTT	TTTT	TTTT	E	E	E

277	305		2.60		LM	VILLLLLLIVIVVVI	V	IIIIIIIIIII	VVVV	IIII	E	E	E
278	306		2.60		FF	FFFFFFFFFFFFFFF	W	FFFFFFFFFFF	WWWW	FFFF	E	E	E
279	307	S			SS	SSSSSSSSSSSSSSS	S	SSSSSSSSSSS	SSSS	SSSS	$		
280	308	a	2.60		AA	AAAAAAAAAAAAAAA	A	AAAAAAAAAAA	AAAA	AAAA	$		
281	309	p			PP	PPPPPPPPPPPPPPP	P	PPPPPPPPPPP	PPPP	PPPP	$		
282	310	N			NN	NNNNNNNNNNNNNNN	N	NNNNNNNNNNN	NNNN	NNNN	$		
283	311	y	2.60		YY	YYYYYYYYYYYYYYY	Y	YYYYYYYYYYY	YYYY	YYYY	$		
284	312		0.43		LL	CCCCCCCCCCGCCCC	C	CCCCCCCCCCC	CCCC	LLLL		E	
285	313		0.17		DD	GGGGGGGGGGGDDGG	Y	YYYYYYYYYYY	YYYY	DDDD		E	
286	314			0.16	TT	MEEEEEEEEEEEEEE	R	RRRRRRRRRRR	RRRR	VVVV		E	
287	315		2.37		YY	MFFFFFFFFFFFFFF	C	CCCCCCCCCCC	CCCC	YYYY		E	
288	316			0.66	NH	NDDDDDDDDDDDDDH	G	GGGGGGGGGGG	GGGG	NNNN		E	
289	317	N			NN	NNNNNNNNNNNNNNN	N	NNNNNNNNNNN	NNNN	NNNN			
290	318		0.06		KK	AVAAAAAAAAAAAWW	K	QQMQQQQQQQQ	VVVV	KKKK	E	E/H	
291	319		0.45		AA	GGGGGGGGGGGGGGG	A	AAAAAAAAAAA	AAAA	AAAA	E	E/H	
292	320		0.31		AA	GAAAAAAAAAAAAAA	S	AASAAAAAAAA	AAAS	AAAA	E	E/H	E
293	321		2.60		IV	VMMMMMMMMMLFFVV	I	IIILIIIIIII	IIVV	VVVV	E	E/H	E
294	322		2.60		LL	MMMMMMMMMMLMMMM	L	MMLMMMMMMLM	LLLM	LLLL	E	E/H	E
295	323			0.43	KK	SSSSSSSSSSSCCSS	E	EEEEEEEEEEG	EEKK	KKKK	E	E/H	E
296	324		0.42		YY	VVVVVVVVVIVVVVV	l	VVVLLLLLLVI	LLIV	YYYY		E/H	E
297	325			0.90	EE	SNDDDDDDDDDDDST	y	DDDDDDDDDDD	DDDD	EEEE		E/H	E
298	326			2.70	NE	TEEEEDEEDEEEEET	s	EEDDDDDDDDD	EEDE	NNNN		E/H	
299	327			1.20	NN	DDSSSTTTTSSNNGG	k	NNCSTTTTTTH	HHDD	NNNN		E/H	

Figure 35. Representative sequences, *bona fide* consensus predictions,[96,286] and experimental[287] secondary structure for the protein serine/threonine phosphatases. Protein sequences are read vertically. From left to right, the columns are alignment numbering, position number in 1tco,[287] functional residues conserved across the entire alignment (lower case, almost entirely conserved), interior score (from DARWIN; higher values mean more buried), surface score (from DARWIN; higher values mean more exposed), multiple sequences, secondary structure (key: E, β strand; H, α helix; $, active site; 3, 3_{10} helix) first from the Florida group,[96] then from the Oxford group,[286] then experimental secondary structure.[287] The reader is encouraged as an exercise to build helical wheels to see how a helix can be transparently predicted from the predicted interior and surface assignments.

and β subunits of the 20S proteasome.[289] Information was also obtained by electron microscopy and image processing of the proteasome from the archaebacterium *Thermoplasma acidophilum*, making the prediction not entirely *ab initio*. However, theory was the most important tool in the model building, and virtually every tool available was used. Assignments of surface and interior residues, made as discussed above, were obtained using DARWIN as implemented on the ETH server and used to derive secondary structure predictions. The PHD server was consulted to obtain an independently predicted secondary structure.[208] Consensus Chou−Fasman and GOR predictions were obtained, as were predictions using the Presnell−Cohen tool.[290] Thus, this prediction represents a "state-of-the-art" combination of imaging and modeling.

The predicted and experimental secondary structures are compared in Figure 36.[291] The correspondence between the experimental and predicted structures were very good. No serious mispredictions were made, and only two short strands were missed. It is interesting to note that both the transparent prediction and the PHD server made similar underpredictions in one region. PHD predicts that the third strand in the α subunit is a helix, while the aligned region in the β subunit is predicted to be a strand. The transparent prediction tool identifies this as a surface region, with perhaps one interior hydrophobic residue anchoring the element. As noted above, this could be either a coil or a strand. The crystallographers assign a strand to this region.

D. The Critical Assessment of Structure Prediction (CASP1) Project

The Critical Assessment of Structure Prediction (CASP) project was undertaken to supplement the *bona fide* prediction efforts described above. CASP was organized by John Moult and Jan Pedersen from the Center for Advanced Research in Biology, Krzysztof Fidelis from the Lawrence Livermore Laboratory, and Richard Judson from the Sandia National Laboratory. The first phase of the CASP project (entitled CASP1) was completed in December 1994 with a meeting in Asilomar. The project attracted several dozen participants.[148] A discussion of the project, including the homology modeling, knowledge-based modeling, and threading projects can be found in a special issue of *Proteins: Structure, Function and Genetics*.[48]

In achieving the goal of bringing together large numbers of predictors and exchanging ideas, CASP1 was quite successful. In terms of generating insights, the project was frustrated by a lack of uniformity in the format in which predictions were submitted, the absence of some key individuals in the field from the list of participants, and the difficulty in obtaining contributions from crystallographers. These problems have been largely resolved in the second phase of the project, CASP2, completed in December 1996 (see below).

```
MQQGQMAYDRAITVFSPDGRLFQVEYAREAVKKGSTALGMKFANGVLLIS sequence of alpha subunit
                                  TTTVGITLKDAVIMAT sequence of beta subunit
           eEEe     eEehhHHHHHHH    eEEee     eEEE PHD alpha subunit
                                  eEEEEEe          PHD beta subunit
           EEEE      HHHHHHHHHHHH  EEEEEE    EEEEE consensus prediction
                    HHHHHHHHHHHH   EEEEEEE  EEEEEE experimental for alpha subunit
                                  EEEEEEEE  EEEEEE experimental for beta subunit


DKKV_RSRLIEQNSIEKIQLIDDYVAAVTSGLVADARVLVDFARISAQQE sequence of alpha subunit
ERRVTMENFIMHKNGKKLFQIDTYTGMTIAGLVGDAQVLVRYMKAELELY sequence of beta subunit
e               eeeee    eeee    hhHHHHHHHHHHHHHhh PHD alpha subunit
          ee  hhhHHHh   eeee    hHHHHHHHHHHHhh     PHD beta subunit
        EEE             EEEEE    HHHHHHHHHHHHHHHHH consensus prediction
EE               EEEEE  EEEEEEE HHHHHHHHHHHHHHHHHH experimental for alpha subunit
                 EEEEE  EEEEE   HHHHHHHHHHHHHHHHHH experimental for beta subunit


KVTYGSLVNIENLVKRVADQMQQYTQYGGVRPYGVSLIFAGIDQIGPRLF sequence of alpha subunit
RLQRRVNMPIEAVATLLSNMLNQ____VKYMPYMVQLLVGGID_TAPHVF sequence of beta subunit
        hhhHHHHHHHHhhhhheee      eEEEEEE       eEE PHD alpha subunit
        hhhHHHHHHHHHHHh           eeeEEee      eEE PHD beta subunit
        HHHHHHHHHHHHHHH           EEEEEE       EEE consensus prediction
HHHH     HHHHHHHHHHHHHH           EEEEEEEE    EEEE experimental for alpha subunit
HHHH     HHHHHHHHHHHHHH            EEEEEEE    EEEE experimental for beta subunit


DCDPAGTINEYKATAIGSGKDAVVSFLEREYKENLPEKEAVTLGIKALKS sequence of alpha subunit
SIDAAGGSVEDIYASTGSGSPFVYGVLESQYSEKMTVDEGVDLVIRAISA sequence of beta subunit
Ee        eeee    hhhHHHHHHHHHHh  HHHHHHHHHHHHHHH  PHD alpha subunit
EEe         eee    hHHHHHHHHHh      HHHHHHHHHHHHHH PHD beta subunit
EE         EEEE    HHHHHHHHHHHH     HHHHHHHHHHHHHH consensus prediction
EE     EEE  EEEEE  HHHHHHHHHHH      HHHHHHHHHHHHHH experimental for alpha subunit
EEE   EEEE  EEEEE  HHHHHHHHHHH      HHHHHHHHHHHHHH experimental for beta subunit


SLE_EGEELKAPEIASITVGNKYRIYDQEEVKKFL                sequence of alpha subunit
AKQRDSASGGMIDVAVITRKDGYVQLPTDQIESRIRKLGLIL         sequence of beta subunit
           eEEEEEe       hhh   hhHHHHH             PHD alpha subunit
hh         EEEEEE     eeee       hHHHhhhhh         PHD beta subunit
HH         EEEEEE              HHHHHHH             consensus prediction
HH        EEEEEEEE    EEE     HHHHHHHH             experimental for alpha subunit
HHHH        EEEEEE   EEEE     HHHHHHHHHH           experimental for beta subunit
```

Figure 36. Representative sequences, *bona fide* consensus prediction,[289] and experimental[291] secondary structure for the homologous α and β proteasome subunits. Separate experimental secondary structural assignments are reported for the α and β subunits. Key: E, β strand; H, α helix. In the prediction, "e" refers to a weakly predicted strand, while "E" refers to a strongly predicted strand; "h" refers to a weakly predicted helix, while "H" refers to a strongly predicted helix. The underlined secondary structural element is predicted inconsistently by the PHD server.[218]

Some of the CASP1 results relating to molecular mechanics and threading predictions were discussed above. Below, we discuss the *ab initio* predictions that are based on evolutionary analyses. A more detailed analysis was provided by DeFay and Cohen.[62]

1. 6-Phospho-β-D-galactosidase

Several predictions were prepared for 6-phospho-β-D-galactosidase as part of the CASP1 project. One was fully transparent.[292] A second used directly the PHD neural network.[208] Still others were based on threading heuristics. Figure 37 compares the predicted and experimental structures.[80]

The transparent prediction assigned both secondary structure and the tertiary fold. The protein was predicted to adopt an eight-fold α−β barrel fold as the conserved core, and this prediction was correct. Thus, this prediction is another example of a case where a secondary structural model was put to good use by serving as the starting point for tertiary structure modeling. It should be pointed out, however, that the α−β barrel has proven in many cases to be an easy fold to identify.

This particular barrel was difficult to identify because the core barrel was interrupted in the primary sequence by segments of polypeptide chain that looped out to form a separate domain. In the transparent prediction, this was recognized because the second domain was not conserved in the superfamily of proteins containing phospho-β-galactosidase, and the barrel structure was correctly pre-

```
MTKTLPKDFIFGGATAAYQAEGATHTDGKGPVAWDKYLEDNYWYTAEPAS 50
        core strand 1
         EEEEEEEEE               HHHHHHH            experimental
        EEE   EEEEEE                            HH  transparent prediction
             HHHHHHHHHH       EEE HHHH        HHHH  Livingston
             HHHHHHH           EEE                  Sander
        EEE   HHHHHH               H     HHH        QL State
        HEE    HHHHH           EEEEHH               QL Profile
       HHHH       HHH           EEHHH      E        Combine

DFYHKYPVDLELAEEYGVNGIRISIAWSRIFPTGYGEVNEKGVEFYHKLF 100
    core helix A   core strand 2       core helix B
HHHHHHHHHHHHHHHH  EEEEEEEEEEE            HHHHHHHHH  experimental
HHHHHHHHHHH   EEEEE active site         HHHHHHHHHH  transparent prediction
HH   HHHHHHHHH              HHH         HHHHHHHH    Livingston
     HHHHHHHHHHHH   EEEEEEEEEEE         HHHHHHHHHHH Sander
HH       HHHHHH     EEEEEEEEE         HHHHHHHHHHHH  QL State
  H   HHHHHHHHHH   EEEEEEE             HHHHHHHHHHH  QL Profile
  H   HHHHHHHHHH    EEEEE    E            HHHHHHHH  Combine

AECHKRHVEPFVTLHHFDTPEALHSNGDFLNRENIEHFIDYAAFCFEEFP 150
      core strand 3                  core helix C
HHHHHH EEEEEE      HHHHHH       HHHHHHHHHHHHHHHHH   experimental
HHHHH     EEEEE                HHHHHHHHHHHHHHHHHHH  transparent prediction
    EEE      HHH      HHHH     HHHHHHH      EEE     Livingston
HHHHHH   EEEEEE    HHHHHH       HHHHHHHHHHHHHHHHHH  Sander
HHH       EEEE      HH         H HHH HHHHHHH        QL State
HHH       EEEEEH    HH         HHHHHHHHHHHHHHHHH    QL Profile
HHHH       HEH      HHHHH         HHHHHHHHHHHHHHH   Combine

EVNYWTTFNEIGPIGDGQYLVGKFPPGIKYDLAKVFQSHHNMMVSHARAV 200
core strand 4                      core helix D
  EEEEE  HHHHHHHHHH             HHHHHHHHHHHHHHHHHH  experimental
H   EEEE                         HHHHHHHHHHHHHHH    transparent prediction
HHHH   HHHH                    EEEEEEE    EE   HHH  Livingston
  EEEEEE    EEE                   EEHHHHHHHHHHHHHH  Sander
    EEEEE         EEE       HHHHHHHHHHHHHHHHHHHHHH  QL State
   EEEE      EEEH E             HHHHHHHHHHHHHH  HH  QL Profile
 HHHHE         HH  E           HHHHHHHHHHHHHHHHHHH  Combine

KLYKDKGYKGEIGVVHALPTKYPYDPENPADVRAAELEDIIHNKFILDAT 250
        core strand 5            core helix E
HHHHH    EEEEEEEEEEEE        HHHHHHHHHHHHHHHHHHHHH  experimental
             EEEE             HHHHHHHHHHHHHHH EEEE  transparent prediction
HHHHH           HHHHHHHH    HHHHHHHHHHHHH     HHH   Livingston
HHHHHH    EEEEEEE           HHHHHHHHHHHHHH   EEEE   Sander
HHH        EEEEE              HHHHHHHHHHHHHHHHH HH  QL State
HHHHH     EEEEEE              HHHHHHHHHHH   HE      QL Profile
HHHH        EEEE            HHHHHHHHHHHHH  HHHH     Combine
YLGHYSDKTMEGVNHILAENGGELDLRDEDFQALDAAKDLNDFLGINYYM 300
                                      core stand 6
      HHHHHHHHHHHHHHEEEEEEE HHHHHHHHH    EEEEEEEEE  experimental
       HHHHHHHH                   HHHHHH  EEEEEE H  transparent prediction
EE       HHHHHHHHH               HHHH       HH   H  Livingston
      HHHHHHHHHHHHH           HHHHHH      EEEEEHHH  Sander
H      HHHH HHHHHHH    HHHHHHHHHHHHHHHHHHHEEEEEEEE  QL State
      HHHHHHHHHH           H HHHHHH HH H HEEE       QL Profile
      HHHHHHHHH             HHHHHHHH     HH         Combine

SDWMQAFDGETEIIH________SKYQIKGVGRRVAPDYVP_____WIIY 350
EEEEEE    EE            EEE    EEEEE                experimental
HHHHHHHHHH                                   EEEEE  transparent prediction
HHHHHHHHH                  HHHHHHH            EE H  Livingston
HHHH                                                Sander
HHHHHHH   HHEE          HHHE                EEEEEE  QL State
HHHH                                          EEE   QL Profile
  E                                            HE   Combine
```

```
PEGLYDQIMRVKNDYPNYKKIYITENGLGYKDEFVDNTVYDDGRIDYVKQ 400
 core helix F     core strand 7              core
 HHHHHHHHHHHH      EEE   EEE             HHHHHHHHH  experimental
 HHHHHHHHHHHHHH     EEEE                   HHHHHHH  transparent prediction
HH  HHHHHHHH                   HHHHHH               Livingston
 HHHHHHHHHHHHHH     EEEEE                 HHHHHHHH  Sander
    HHHHEEEH       EEEEEE      EEEE         HHHHHH  QL State
   HHHHHHHHHH      EEEEE                    HHHHHH  QL Profile
   HHHHHHHHH        EEEE                    HHHHHH  Combine

HLEVLSDAIADGANVKGYFIWSLMDVFSWSNGYEKRYGLFYVDFDTQERY 450
helix G      core strand 8
HHHHHHHHHH  EEEEEEEEEE           EEEE  EEEEE  EEEE  experimental
HHHHHHHHHH EEEEE EEEE                 EEEEE         transparent prediction
HHH            EE  HHHHHH                     EE    Livingston
HHHHHHHHHH    EEEEEEEEE   EEE    EEEEEEEEEE         Sander
HHHHHHHHHHH   H EEEEEEEEEEEE         EEEEE          QL State
HHHHHHHHHH    H EEEEEEE   HHHH       EEEEE          QL Profile
HHHHHHHHHH        EHHHHHHHHHHHH       EEEE     H    Combine

PKKSAHWYKKLAETQVIE
core helix H
E HHHHHHHHHHHH      experimental
 HHHHHHHHHHHHH      transparent prediction
 HHHHHHHHHHHHHHH    Livingston
 HHHHHHHHHHHHH      Sander
  HHHHHHHHHH HHHHH  QL State
  HHHHHHHHH         QL Profile
H   HHHHHHHH  H     Combine
```

Figure 37. Transparently predicted[292] and experimental[80] secondary structure for phospho-β-galactosidase (*Lactococcus lactis*) (LACG_LACLA P11546, 1pbg). Key: E, β strand; H, α helix; _, indel. Experimental structure assigned by DSSP. The underlined regions designate the core secondary structural elements in the conserved α–β barrel domain. These are assigned using the DEFINE program. This illustrates the accuracy of the consensus prediction in the assignment of secondary structure to elements of secondary structure that are conserved throughout the protein family, but not (by definition) to those that are not. Other predictions were generated by the following individuals using the tools indicated: Livingston,[293] Sander,[294] Munson Quadratic Logistic,[295,179] and Munson/Garnier Combine.[178]

dicted. Modeling based on the PHD secondary structure prediction favored (incorrectly) a sheet structure.

Di Francesco *et al.*[179] recently commented on possible approaches toward the prediction of the nonconserved domain in this protein. They came to the interesting conclusion that fewer sequences showing less sequence divergence overall might have produced a better prediction for the nonconserved noncore domain, at least when using a consensus GOR analysis. This is an intriguing idea deserving further exploration.

More divergent sequences contain more information of some types (for example, the location of active sites). However, they also differ more in their conformation. At the very least, this makes scoring difficult. However, if substantial modification of secondary structure has taken place in noncore domains, the signal arising from the sequences themselves might be confusing. In these cases, it might be better to make predictions for subfamilies, as has been done now in many cases.[248,260,292]

2. Xylanase

If further evidence were needed to show that eightfold α–β barrels can be identified in 1996 with high reliability, the PHD neural network prediction of xylanase provides it. Figure 38 shows the prediction with the subsequently reported experimental structure.[296] Unlike phospho-β-galactosidase, xylanase is a relatively simple barrel, lacking intervening secondary structural elements. Thus, with the exception of one core strand that the PHD prediction missed, the prediction is essentially perfect.

3. Synaptotagmin

Synaptotagmin is a protein domain involved in membrane fusion, and is also found in protein kinase assemblies. The prediction is presented in two ways, the first in "transparent form" (Figure 39),[297] the second in a form summarizing all of the predictions made in the CASP1 program for the protein (Figure 40). An experimentally derived assignment of secondary structure accompanies each.[298]

The transparent synaptotagmin prediction identifies the first seven β strands of the fold essentially correctly.[79] Further, with the exception of β 4, the beginnings and ends of the predicted strands correspond well with those assigned by DSSP to the experimental coordinates. Further, the assignments of secondary structure in the synaptotagmin family were correct for the correct reasons. Figure 39 shows both the predicted assignment of secondary structure (S and s for strong and weak surface assignments, I and i for strong and weak interior assignment) and the experimental assignments (from DSSP). For β 1, β 2, β 3, β 5, and β 7, surface and interior residues were correctly assigned. From these, the assignment of the β strands is transparent. The reader should inspect Figure 39 to see how the alternating surface/

```
VATGNGLASL ADFPIGVAVA ASGGNADIFT SSARQNIVRA EFNQITAENI MKMSYMYSGS
     HHHHH HHH    EEE E           HHHHHHHHH HH                       Sander
................................................................     Hubbard
    HHHHHH HH   HHHHH HHHH        HHHHHHHHH HHHHHHHHHH HH       E    Sippl
      3333     EEEEE              HHHHHHHHH H  EEEE      3333EE      experimental

NFSFTNSDRL VSWAAQNGQT VHGHALVWHP SYQLPNWASD SNANFRQDFA RHIDTVAAHF
 HHHHHHHHH HHHHHH   E EEEEEEEEE                 HHHHHH HHHHHHHH      Sander
................................................................     Hubbard
EEE         HHHHHH EE EE                    HHHHHHHHHH hhh           Sippl
EE  HHHHHH HHHHHH   E EEEEEEE    333            HHHHHH HHHHHHHHH     experimental


AGQVKSWDVV NEALFDSADD PDGRGSANGY RQSVFYRQFG GPEYIDEAFR RAPRADPTAE
  EEEEEEHH HHH               HHH H  EEEEE    HHHHHHHHH HHH   H EE    Sander
..........................     H HHHEEEEE     HHHHHHHH HHHH     E    Hubbard
      hhhh hhhhhhhhh                         hhhhHHHHH Hh      hh    Sippl
     EEEEE E                        HHHHHH    HHHHHHHH HHHHH    E    experimental


LYYNDFNTEE NGAKTTALVN LVQRLLNNGV PIDGVGFQMH VMNDYPSIAN IRQAMQKIVA
EEE           HHHHHHH HHHHHHH     EE   EEEE EE      HH HHHHHHHHHH    Sander
EEE           HHHHHHH HHHHHHH     EEEE   E          HH HHHHHHHHHH    Hubbard
hhHHhhhHHh hh  hhhhhh HHHHHHHh   eeeeeeee h hhhhh  HHH HHHHHHHHHH    Sippl
EEEEE        HHHHHHHH HHHHHHHH      EEEE  E EE     HHH HHHHHHHHH     experimental

LSPTLKIKIT ELDVRLNNPY DGNSSNDYTN RNDCAVSCAG LDRQKARYKE IVQAYLEVVP
     EEEEE EEE                         HHHH HHHHHHHHHH HHHHH         Sander
H    EEEEE EEE                        HHHHH HHHHHHHHHH               Hubbard
h   eeeEee eeee                     hhhh    hhHHHHHHHH HHHHHHHH      Sippl
     EEEEE EEEEE         333             HH HHHHHHHHHH HHHHHHHH      experimental


PGRRGGITVW GIADPDSWLY THQNLPDWPL LFNDNLQPKP AYQGVVEALS GR
    EEEEEE E                EEEE E       HH HHHHHHHH                 Sander
     EEEEE E                  EE EE       H HHHHHHHHH                Hubbard
hhhhhheEEE e    hhhhH HHHH     e eee   eeeehhhhHHHhhh                Sippl
   EEEEEE     333     EE  EE              H HHHHHHHHH                experimental
```

Figure 38. Sequence, predictions[294] from the CASP1 site (http://PredictionCenter.llnl.gov), and experimental secondary structure for xylanase, (*Pseudomonas fluorescens*)(P14768, 1clx XYNA_PSEFL). Key: E, *β* strand; H, α helix; e, weakly predicted strand; h, weakly predicted helix; 3, 3_{10} helix; ., unpredicted. The Sippl prediction was based on threading onto 1tim-b (triose phosphate isomerase), the Hubbard prediction was based on threading onto 1xla (D-xylose isomerase).

interior assignments allowed prediction of strands in these regions.

β 4 is too short to be analyzed in this fashion with statistical significance. The segment containing *β* 6 was correctly identified as being largely internal, and the secondary structure correctly assigned using a different rule-based approach.

The single mistake made in the transparent prediction was the misassignment of the final strand as a helix. This misassignment was made because of wide divergence of the sequences in this region and an imprecise placement of a parse. It is interesting to note that this misassignment had essentially no impact on the efforts to build a tertiary structure from the assembled secondary structural elements, in part because this was the final secondary structural element in the protein, and in part because this element was not at the core of the folded structure.

The prediction was sufficiently accurate to permit the correct tertiary fold to be proposed as one of three alternative folds. To build a tertiary structural model, a combinatorial approach first assembled all possible sheet structures from the predicted secondary structural elements.[297] A large majority of these were then excluded by enforcing certain connectivity of strands in a *β* sheet, avoiding loop crossovers, and using other rules that have (at least some) empirical basis.[300] Efforts were made to construct a calcium binding "active site" in the protein fold (see below). After this process was complete, three folds remained.

The database was then examined for analogs for the three remaining folds. The first, where the strands were placed consecutively around a *β* sandwich, found its closest analog in the retinol binding protein (where the strands form a consecutive antiparallel *β* sheet defined in the ABCDEFG sequence).[302] This is, of course, a "knowledge-based" approach to modeling. Including a single Greek key element in the fold approximated the fold found in pseudoazurin.[303] To make this analogy "work", the first strand of pseudoazurin was ignored, and a strand was moved from one sheet in the *β* sandwich to the other. The final topology is best described as ABEDCFG. The third remaining fold had the topology similar to that found in the pleckstrin homology domain (ABCDGFE).[271]

Criteria were then considered to distinguish between these three alternative packings. These suggested a weak preference for a fold similar to that found in the pleckstrin homology domain. In fact, the "modified pseudoazurin" fold turned out to be an

```
Pos q w zyAx tsurv DCBEFGH  p nomlkji h fedcba g | K I J   Predicted Experimental    Ca++
                                                            Sec surf  surface Sec  binding
                                                            ETH inter access  Str
  1 C C CCCC CCCCC CCCCCCC|                      |
  2 G G GGGG GGGGG GGGGGGG| S GKEEEEE H QKKKKK K | P F F          s
  3 A C VVVV MMMMM TTTTTTT| E GEEEEEE Q SEEEEE E | T N N          s
  4 D D DDDD DDDDD DDDDDDD| K QAPPQQQ K EEPDEE E | S S S          s
  5 I H HHHH HHHHH HHHHHHH| A EEEEEEE V QEEEEE V | P R R          s     185
  6 S T TTTT TTTTT TTTTTTT| D KKKKKKK N KKNKKK K | E A A          s     192
  7 E E EEEE EEEEE EEEEEEE| L LSLLLLL C LLLLLL L | R L L                 32
  8 V R RRRR KKKKK RRRRRRR| _ _L_____ _ ______ _ | I T A
  9 R R RRRR RRRRR RRRRRRR| _ _______ _ ______ _ | K Q H
 10 G G GGGG GGGGG GGGGGGG  G GGGGGGG G GGGGGG G | S G G                  6
 11 K R RRRR RRRRR RRRRRRR  E DDDDDDD R RKKKKK R | S P P     E    S      69   b1
 12 L I LLLL IIIII IIIIIII  I IIIIIII I LILLLL I | _ W W     E    I       0   b1
 13 L Y QQQQ YYYYY YYYYYYY  N CCCCCCC N NQQQQQ Q | _ W W     E    s      19   b1
 14 L L LLLL LLLLL IIIIIII  F FFTTFFF F FFFFYY Y | _ _ _     E    i       3   b1
 15 Y E EEEE KKKKK QQQQQQQ  S SSSSSSS M KSSSSS K | _ _ _     E    S      32   b1
 16 V I IIII AAAAA AAAAAAA  L LLLLLLL L LLLLLL L | _ _ _     E    I       2   b1
 17 E N RRRR EEEEE HHHHHHH  C RRRRRRR R EDDDDD D | _ _ _     E    S      45   b1      *
 18 L V AAAA VVVVV IIIIIII  Y YYYYYYY Y YYYYYY Y | _ _ _     E    I       4   b1
 19 K K PPPP ATTTT EDDDDDD  L VVVVVVV T DDDDDD D | _ _ _          i      29   b1      *
 20 _ _ ____ _____ _______  P PPPPPPP Y FFFFFF F | _ _ _                 93
 21 _ _ TTTT _____ _______  T TTTTTTT T NQQQQQ Q | K R T          S     164
 22 G E SSAS DDDDD RRRRRRR  A AAAAAAA T SAANNN Q | Y P P          S     106
 23 N N DDDD EEEEE EEEEEDD  G GGGGGGG E NNNNNN G   S E K          S      85
 24 N L EEEE KKKKK VVVVVVV  R KKKKKKK Q SQQQQQ Q   R R R     E    s      36   b2
 25 L L IIII LLLLL LLLLLLL  L LLLLLLL L LLLLLL L   L L L     E    I       0   b2
 26 K T HHHH HHHHH IIIIIII  T TTTTTTT V ATTILL T   I R N     E    s      36   b2
 27 V V VVVI VVVVV VVVVVVV  I VVVVVVV V VVVVVV V   V V V     E    I       0   b2
 28 D Q TTTT TTTTT VVVVVLL  T VVCCVVV K TGGGGG T   N R R     E    s       8   b2
 29 I I VVVV VVVVV VVVVVVV  I IIIIIII I VIVIII V   V I V     E    I       0   b2
 30 K K GGGG RRRRR RRRRRRR  I LLLLLLL L IILIII I   I I I     E           33   b2
 31 E E EEEE DDDDD DDDDDDD  K EEEEEEE K QQQQQQ Q   S S S     e    S      61   b2
 32 A G AAAA AAAAA AAAAAAA  A AAAAAAA A AAAAAA A   A G G     e    i       1   b2
 33 A R RRRR KKKKK KKKKKKK  T KKKKKKK L EAAAAA E   R Q Q     e    S      16   b2
 34 N N NNNN NNNNN NNNNNNN  N NNNNNNN D EEEEEE D   Q Q Q          s     117
 35 L L LLLL LLLLL LLLLLLL  L LLLLLLL L LLLLLL L   L L L          I       7
 36 I I IIII IIIII VVVVVVV  K KKKKKKK P PPPPPP P   P P P                 36
 37 P P PPPP PPPPP PPPPPPP  A KKKKKKK A AAAAAA G   K K K                 47
 38 M M MMMM MMMMM MMMMMMM  M MMMMMMM K LLLLLL M   Y V V          I      77
 39 D D DDDD DDDDD DDDDDDD  D DDDDDDD D DDDDDD D   T N N          S      49
 40 T P PPPP PPPPP PPPPPPP  L VVVVVVV A MMMVMM M   K K K                192
 41 N N NNNN NNNNN NNNNNNN  T GGGGGGG N GGGGGG S   S N N          s      66
 42 _ _ ____ _____ _______  _ _______ _ ______ _   T _ _
 43 _ _ ____ _____ _______  _ _______ _ ______ _   K K K
 44 _ _ ____ _____ _______  _ _______ _ ______ _   G N N
 45 G G GGGG GGGGG GGGGGGG  G GGGGGGG G GGGGGG G   E S S                 44
 46 F L LLLL LLLLL LLLLLLL  F LLLLLLL F TTTTTT T   V I I          I      29
 47 S S SSSS SSSSS SSSSSSS  S SSSSSSS S SSSSSS S   I V V          i       0
 48 D D DDDD DDDDD DDDDDDD  D DDDDDDD D DDDDDD D   D D D          A      21        Ca++
 49 P P PPPP PPPPP PPPPPPP  P PPPPPPP P PPPPPP P   P P P          .       0   b3
 50 Y Y YYYY YYYYY YYYYYYY  Y YYYYYYY Y YYYYYY Y   Y K K     E            6   b3
 51 I V VVVV VVVVV VVVVVVV  V VVVVVVV V VVVVVV V   V V V     E    I       0   b3
 52 A K KKKK KKKKK KKKKKKK  K KKKKKKK K KKKKKK K   T I T     E    s      46   b3
 53 V V LLLL LLLLL LLLLLLL  A IIIIIII I VVVVVV L   L V V     E    I       0   b3
 54 Q K KKKK KKKKK KKKKKKK  S AVHHHHH Y YFFFFF Y   S E E     E    S      48   b3
 55 M L LLLL LLLLL LLLLLLL  L ILLLLLL L LLLVLL L   I I I     E    I       0   b3
 56 H I IIII IIIII IIIIIII  I MMLMMMM L LLLLLL L   V H H     E    i      67   b3
 57 P P PPPP PPPPP PPPPPPP  C QQQQQQQ P PPPPPP P   G G G          .     104
 58 D D DDDD DDDDD DDDDDDD  D NGNNNNN D DDDDDE E   T V V          S      74
 59 R D PPPP PPPPP PPPPPPP  E GGGGGGG R KKKKKK K   H G T          S      32
 60 S K RRRR KKKKK KKKKKKK  R KKKKKKK _ ______ _   F R G          S
 61 G D NNNN NNNNN SSSSSSS  R RRRRRRR _ ______ _   D D D          s
 62 R Q LLLL EEEEE EEEEEEE  L LLLLLLL K KKKKKK K   Q T V                 94
 63 T S TTTT SSSSS SSSSSSS  K KKKKKKK K KKKKKK K   K G A                 38
 64 K K KKKK KKKKK KKKKKKK  K KKKKKKK K KKKKKK K   V S S                 77
 65 K K QQQQ QQQQQ QQQQQQQ  R KKKKKKK F FYYYFF V   E R R          s      56   b4
 66 K K KKKK KKKKK KKKKKKK  K KKKKKKK Q EEEEEE E   K Q Q          s     100   b4
 67 T T TTTT TTTTT TTTTTTT  T TTTTTTT T TTTTTT T   T T T     e    A       8
 68 K R RRRK KKKKK KKKKKKK  S SSTTTTT K KKKKKK K   K A A     e    s     154
 69 T T TTTT TTTTT TTTTTTT  I VIVVIII V VVVVVV V   V V V     e    I      39
```

70	I	I	VVVV	IIIII	IIIIIII	K	KKKKKKK	H	HQHHHH	H		I	I	V	e		53		
71	Q	K	KKKK	RRRRR	KKKKKKK	K	KKKKKKK	R	RKRRRR	R		D	T	T		S	114		
72	K	A	AAAA	SSSSS	CCCCCCC	N	CCNKNNN	K	KKKKKK	K		N	N	N		S	133		
73	N	C	TTTT	TTNTT	SSSSSSS	T	TTTTTTT	T	TTTTTT	T		N	N	N			38		
74	L	L	LLLL	LLLLL	LLLLLLL	L	LLLLLLL	L	LLLLLL	L		G	G	G		I	41		
75	_	_	____	_____	_______	_	_______	_	______	_		F	F	F					
76	N	N	NNNN	NNNNN	NNNNNNN	N	NNNNNNN	N	SNNNNN	N		N	N	N		s	95		
77	P	P	PPPP	PPPPP	PPPPPPP	P	PPPPPPP	P	PPPPPP	P		P	P	P		.	4		
78	V	V	VVVV	QQQQR	EEEEEEE	V	YYYYYYY	I	VTAVVV	V		H	R	W		S	75	b5-	
79	F	W	WWWW	WWWWW	WWWWWWW	Y	YYYFYYY	F	FFFFFF	F		W	W	W		I	12	b5	
80	N	N	NNNN	NNNND	NNNNNNN	N	NNNNNNN	N	NNNNNN	N		G	D	D		S	121	b5	
81	E	E	EEEE	EEEEE	EEEEEEE	E	EEEEEEE	E	EEEEEE	E		E	M	T		A	39	b5	
82	T	T	TTTT	SSSSS	TTTTTTT	A	SSSSSSS	T	TSTSQQ	T		E	E	E	E	S	122	b5	
83	F	L	FFFF	FFFFF	FFFFFFF	L	FFFFFFF	F	FFFFFF	F		F	F	L	E	I	21	b5	
84	T	T	VVVV	TTTTT	RRRRRRR	V	SSSSSSS	Q	TVTITT	I		E	E	E	E	s	48	b5	
85	F	Y	FFFF	FFFFF	FFFFFFF	F	FFFFFFF	F	FFFFFF	F		F	F	F	E	I	9	b5	
86	E	D	NNNN	KKKKK	QQQQQQQ	D	EEEEEEE	N	KKKKKK	K		P	E	E	E	s	180	b5	
87	L	L	LLLL	LLLLL	LLLLLLL	I	VVIIVVV	V	SVVIVV	V		L	V	V	E	I	11	b5	
88	Q	K	KKKK	KKKKK	KKKKKKK	P	PPPPPPP	P	LPPPPP	A		Y	T	A		s	54		
89	_	_	____	_____	_______	_	_______	_	P_____	_		_	_	_					
90	P	P	PPPP	PPPPP	EEEEEEE	N	FFFFFFF	F	YYYYYY	F		N	V	V		s	116	h	
91	Q	E	GGGG	SSSSS	SSSSSSS	E	EEEEEEE	N	AQQSSS	N		S	P	P		s	112	h	
92	D	D	DDDD	DDDDD	DDDDDDD	N	QQQQQQQ	E	DEEEEE	E		Q	D	E		S	117	h	
93	R	K	VVVV	KKKKK	KKKKKKK	M	MIIIIII	L	ALLLLL	I		L	L	L			0	h	
94	D	D	EEEE	DDDDD	DDDDDDD	E	QQQQQQQ	Q	MGGGGG	T		_	_	_		s	37		
95	_	_	____	_____	_______	H	KKKKKKK	N	NGGGGG	A		_	_	_		s	69		
96	K	R	RRRR	RRRRR	RRRRRRR	V	IVVVVVV	R	KKKKKK	K		S	A	A		s	32		
97	R	R	RRRR	RRRRR	RRRRRRR	N	CSQQQQQ	K	TTTTTT	T		M	L	L	E	s	35	b6	
98	L	I	LLLL	LLLLL	LLLLLLL	V	LLVVVVV	L	LLLLLL	L		L	V	V	E	I	0	b6	
99	L	L	SSSS	SSSSS	SSSSSSS	I	VMCVVVV	H	VMVVVV	V		L	R	R	E		8	b6	
100	I	I	VVVV	VVVVE	VVVVVVV	I	VIVVVVV	F	FMMMMM	F		I	F	F	E		0	b6	
101	E	E	EEEE	EEEEE	EEEEEEE	A	TTTTTTT	S	AAAAAA	A		R	M	V	E	s	4	b6	
102	V	V	VVVV	IIIII	IIIIIII	V	VVVVVVV	V	IVIVVV	I		V	V	V	E	I	0	b6	
103	W	W	WWWW	WWWWW	WWWWWWW	M	VMLLLLL	Y	FYYYYY	Y		D	E	E			48	b6	
104	D	D	DDDD	DDDDD	DDDDDDD	D	DDDDDDD	D	DDDDDD	D		D	D	D		A	4	b6	Ca^{++}
105	W	W	WWWW	WWWWW	WWWWWWW	Y	YYYYYYY	F	FFFFFF	F		K	Y	Y			33		
106	D	D	DDDD	DDDDD	DDDDDDD	D	DDDDDDD	D	DDDDDD	D		D	D	D		A	51		Ca^{++}
107	R	R	RRRR	RRRRR	LLLLLLL	C	RKKKKKK	R	RRRRRR	R		K	S	A		s	161		
108	T	T	TTTT	TTTTT	TTTTTTT	I	ILILIII	F	FFFFFF	F		V	S	S		i	187		
109	S	S	SSSS	TTTTT	SSSSSSS	G	GGGGGGG	S	SSSSSS	S		G	S	S			68		
110	R	R	RRRR	RRRRR	RRRRRRR	H	TSKKKKK	R	KKKKKK	K		H	K	K		s	155		
111	N	N	NNNN	NNNNN	NNNNNNN	N	SNNNNNN	H	HHHHHH	H		N	N	N		s	52		
112	D	D	DDDD	DDDDD	DDDDDDD	E	EDDEDDD	D	DDDDDD	D		R	D	D		s	80	b7	
113	F	F	FFFF	FFFFF	FFFFFFF	V	PAAAAAA	L	QCIVII	Q		I	F	F	e	i	80	b7	
114	M	M	MMMM	MMMMM	MMMMMMM	I	IIIIIII	I	IIIIII	I		G	I	I	e	I	13	b7	
115	G	G	GGGG	GGGGG	GGGGGGG	G	GGGGGGD	G	GGGGGG	G		H	G	G	e		2	b7	
116	S	A	AAAA	SSSSS	SSSSSSS	M	RRKKKKK	Q	EQEEEE	Q		H	Q	Q	E	s	46	b7	
117	F	L	MMMM	LLLLL	LLLLLLL	C	CCIIVVV	V	VVVAFF	V		C	S	S	E	i	2	b7	
118	S	S	SSSS	SSSSS	SSSSSSS	R	ILFFFFF	V	KTKKKK	L		I	T	T	E	s	82	b7	
119	F	F	FFFF	FFFFF	FFFFFFF	V	LLVVVVV	L	VVVVVV	I		R	I	I	E		6	b7	
120	S	G	GGGG	GGGGG	GGGGGGG	G	GGGGGGG	D	PLPPPP	P		_	P	P	E		48	b7	
121	_	_	____	_____	_______	_	_______	N	______	_		_	_	_					
122	_	_	____	_____	_______	_	_______	L	______	_		_	_	_					
123	_	_	____	_____	_______	_	CCSSYYY	L	LMMMMM	L	|	V	W	L		i	3		
124	L	I	VVVV	VVVVV	IIIIIII	N	MNNNNNN	E	CTNNNN	G	|	E	N	K			122		
125	E	S	SSSS	SSSSS	SSSSSSS	A	GGAASSS	F	TKTTTT	K	|	N	S	S		s	95		
126	E	E	EEEE	EEEEE	EEEEEEE	T	TTSTTTT	S	IVVVVV	I	|	I	L	L			22		
127	L	I	LLLL	LLLLL	LLLLLLL	D	GGGGAGG	D	DDDDDD	D	|	R	K	K	H		113		*
128	Q	I	LLLL	MMMMM	QQQQQQQ	G	TATTAAA	F	LLLFFF	L	|	P	Q	Q	H		17		
129	K	K	KKKK	KKKKK	KKKKKKK	P	EEEEEEE	S	AGGGGG	G	|	G	G	G	H	s	46		
130	E	N	AAAA	MMMMM	AAAAAAA	_	_______	E	QQQHHH	A	|	_	_	_	H	s	111		
131	P	P	PPPP	PPPPP	GGGGGGS	G	_______	D	TQPVVV	V	|	_	_	_	H		98		
132	V	T	VVVV	AAAAA	VVVVVVV	R	LLLLLLL	T	ILITTT	I	|	Y	Y	Y	H	i	35	b8	
133	D	N	DDDD	SSSSS	DDDDDDD	E	RRRRRRR	T	EEEEEE	E	|	R	R	R	H	s	112	b8	*
134	G	G	GGGG	GGGGG	GGGGGGG	H	HHHHHHH	I	EEEEEE	E	|	I	H	H	H		68	b8	
135	W	W	WWWW	WWWWW	WWWWWWW	W	WWWWWWW	W	WWWWWW	W	|	L	V	I	H	i	102	b8	
136	Y	F	YYYY	YYYYY	FFFFFFF	N	SMSSSSS	R	RRRRRR	K	|	K	H	H	H	s	68	b8	
137	K	K	KKKK	KKKKK	KKKKKKK	E	DDDDDDD	D	DDDDDD	D	|	L	L	L	H		90	b8	
138	F	L	LLLL	LLALL	LLLLLLL	M	MMMMMMM	I	LLLLLL	I	|	K	L	L	H	i	4		
139	L	L	LLLL	LLHLL	LLLLLLL	L	LLLLLLL	L	VEQQQQ	A	|	N	S	S	H		113		
140	S	T	NNNN	NNNNN	SSSSSSS	A	AAAAAAA	E	SSGGSS	P	|	N	K	K	H		115		
141	Q	Q	QQQQ	QQQQQ	QQQQQQQ	N	SSNNNNN	A	VAGAAA	P	|	F	N	N	H	s			
142	V	D	EEEE	EEEEE	EEEEEEE	P	PPPPPPP	T	EEEEEE	P	|	N	G	G					

Figure 39. Representative sequences, transparent consensus prediction,[297] and experimental[298] secondary structure for the synaptotagmin family, presented to show the reader how a transparent prediction works.[79] Protein sequences are read vertically. Key: E, *β* strand; H, α helix; A, active site . In the prediction, "e" refers to a weakly predicted strand, while "E" refers to a strongly predicted strand; "H" indicates a strongly predicted helix. The predicted surface accessibility of each residue side chain is indicated by S and s (strong and weak surface prediction) and I and i (strong and weak interior prediction). Experimental surface accessibility is reported in terms of relative side chain accessibility to solvent. Residues involved in calcium binding are indicated in the right column.[298]

```
EKLGKLQYSLDYDFQNNQLLVGIIQAAELPALDMGGTSDPYVKVFLLPEKKKKFETKVHR
    EEEEEEEE     EEEEEEEEeee            EEEEEEE        eeeee  Benner
 hhhhh               hhhhhhhhhhhhhh     eeEEEeehhhhhhh hhhhh  Sippl
  EEEEEEEEEE     EEEEEEEHHHHHHHH        EEEEEEE      HHHHHHH  Barton
  EEEEEEEEE      EEEEEEE                 EEEEE        EEEEEE  Hubbard
 EE  EEE EE     E   EEE     HHHHHH                     EEEE   Clarke
HHHHHHHHH           EEEE                EE EEEE    EEE EEEE   Matsuo
   EEEEEEEEEE    EEEEEEEEEEE           EEEEEEEE      EE       experimental


KTLNPVFNEQFTFKVPYSELGGKTLVMAVYDFDRFSKHDIIGEFKVPMNTVDFGHVTEEW
         EEEEEE        EEEEEE          eeeEEEEE    HHHHHHHHH  Benner
hhhhhh    eeeEEE      ee hh h  HHHHHh hhhhheEEEEeEEEe  hhHHH  Sippl
HHHH     EEEEEE       EEEEEEEEEE       EEEEEEE           HHH  Barton
EE      EEEEEEE       EEEEEEEEE         EEEEEEEHHHHH     HHH  Hubbard
            EE   HHHHHHH    EEE  EEEEEE    E................  Clarke
           EEEEEE      EEEEEE HHHHHHHHHHHHHHHHHH    EEEEE     Matsuo
     EEEEEEEE   HHHH   EEEEEEEE       EEEEEEEEE         EEEE  experimental


RDLQSAE
HHHHHH    Benner
HHHHHHH   Sippl
HHHH      Barton
EE        Hubbard
          Matsuo
EE        experimental
```

Figure 40. Sequence and predictions from the CASP1 site, and experimental[298] secondary structure for the first C2 domain of synaptotagmin (P21707, 1rsy SYT1_RAT), which forms a Greek key *β* sandwich. Key: E, *β* strand; H, α helix; e, weakly predicted strand; h, weakly predicted helix. Prediction made by Hubbard[216] combines the PHD neural network and hidden Markov models. The prediction of Sippl,[299] Clarke, and Matsuo are based on threading tools.

approximately correct model for the fold of synaptotagmin as determined experimentally. The order of the strands in the *β* sandwich is correctly assigned (with the omission of the first strand of pseudoazurin, which has no counterpart in the model, and the misassigned helix). The closest analog in the database for the fold of synaptotagmin is PapD, which contains the connectivity of the pseudoazurin fold. These results underscore the need to identify rules, perhaps based on contact potentials or real potentials, for identifying a preferred domain from a small number of alternatives.

The most interesting success of the transparent synaptotagmin prediction is the quality of the model built for the calcium-binding active site. In the prediction, Asp 48 (Asp 178 in the synaptotagmin numbering), Asp 104 (Asp 230) and Asp 106 (Asp 232) and Glu 81 (Glu 208) were assigned as calcium-binding ligands. Except for Glu81, these proved to form the putative calcium-binding active site in synaptotagmin.

A collection of transparent, neural network, and threading predictions is presented in Figure 40. The PHD-based prediction[216] is essentially the same as the transparent prediction, misassigning the final strand as well. The reproduction by the PHD neural network (at least in its 1994 version) of mistakes made by transparent methods appeared to be frequent. The prediction by Barton's group contains a serious mistake, misassigning a core strand as a helix. The remaining predictions are less well suited to serve as starting points for tertiary structural modeling.

4. Staufen

The staufen protein provided an opportunity to compare several largely nontransparent prediction tools. Figure 41 collects a variety of predictions made for the protein, together with an experimental secondary structure.[301]

Hubbard[216] evidently submitted the target sequence to the PHD neural network, which retrieves homologous sequences from a database, constructs a multiple alignment, and then makes a secondary structure prediction. The secondary structure was predicted to be α−*β*−*β*−*β*−α (Figure 41). This prediction is essentially correct. This model was then used to search the crystallographic database to identify proteins having a similar fold. Positions 150−222 of cytoplasmic malate dehydrogenase (2cmd) were recovered. A tertiary structure model for

```
DKKSPISQVHEIGIKRNMTVHFKVLREEGPAHMKNFITACIVGSIVTEGE
    HHHHHHHHHHH     HHHHHH         EEEEEEE  EEE     Garnier SIMPA
       HHHHHH        E E           EEEEEEE  EEE     Hubbard
                     HHHHHHHHHHHHH    HHHHHHHHHHHH  Livingston
                EEEEE             EEEEE     EE      Sander
       EEEEEEEE   EEEEEHHH    H HHHEEEEEEEEEEEE     QL State
         HH          EEHHH          EEEEE   EEE     QL Profile
        HHHH         HHHHH         EEEEEEE          Combine
     HHHHHH     EEEE HHHHHH   HHHHHHHHHHH EEEEEE    Matsuo
 e EeeEe e  h   h     HHHHhhhHHHHHhhhh   E eeEEee   Sippl
    HHHHHHHHHHHH    EEEE          EEEEEEEE  EEEEEE  experimental DSSP

GNGKKVSKKRAAEKMLVELQ KLPPLTPTK
   HHHHHHHHHHHHHHHHH                                Garnier SIMPA
    HHHHHHHHHHHHHHHH                                Hubbard
HHHHHHH   HHHHHHH    EEEEEEEE                       Livingston
          HHHHHHEE                                  Sander
    EH HHH HHHHHHHH                                 QL State
   HHHHHHHHHHHHHHHH                                 QL Profile
    HHHHHHHHHHHHHHH                                 Combine
 HHHHH   HHHHH EEEE                                 Matuso
hHH  h HHHHHHHhhhhHH hhee EEeh                      Sippl
EE   HHHHHHHHHHHHHHH                                experimental DSSP
```

Figure 41. Sequence predictions from the CASP1 site, and experimental[301] secondary structure for domain 3 of staufen (STAU_DROME, P25159, 1stu). Key: E, β strand; H, α helix. Predictions were generated by the following individuals using the tools indicated: Garnier Simpa,[132] Hubbard,[216] Livingston,[293] Sander,[294] Munson Quadratic Logistic,[179,295] and Munson/Garnier Combine.[178] The prediction of Sippl is based on a threading tool [299] as is that of Matsuo.

staufen was then based on the experimental structure of this segment of malate dehydrogenase.

Three details of this prediction are remarkable and worth discussion. First, a second prediction was submitted to CASP1 using the PHD neural network (the prediction marked "Sander" in Figure 41). Although it was evidently obtained from the same server, the "Sander" prediction is quite different from the "Hubbard" prediction: at only 67.5% of the positions is secondary structure for the "Sander" PHD prediction the same as the "Hubbard" PHD prediction. In other words, the Q_3 score of one PHD prediction scored using the other output is only 67.5%. We cannot say from information available from the Web site how these two predictions, ended so differently. Different versions of the PHD may have been used. Different sets of homologs might have been retrieved. Hubbard evidently adjusted the multiple alignment by hand, while the Sander group evidently did not. In any case, it is remarkable how different the output was given what presumably were only minor differences in the input, and points out again the need to look closely at the details of each prediction to learn the most from a prediction.

The second thing unusual about the Hubbard prediction is that the HMM identified in the crystallographic database a domain with a fold similar to that of staufan, but different in a critical feature. The domain came from the middle of the cytoplasmic malate dehydrogenase and is almost certainly not homologous to staufen. It is almost inconceivable that the RNA binding domain of staufen evolved by extraction of a segment in the middle of an enzyme. If not, then the conformational similarity between staufen and residues 150−222 arose by convergent evolution.

Third, the crystallographic database evidently does contain a homolog of staufen, the N-terminal domain of the rS5 protein from *Bacillus subtilis*. This was the reference protein found by Sippl in the threading portion of CASP1, which also considered staufen. Further, the crystallographers identify and discuss the homolog. The homolog apparently lacks the first helix. From our understanding of the method used by Hubbard to find analogous structures in the database, the first helix would have been required to find this homolog.

The Garnier SIMPA prediction is also interesting. The tool provides either a homology search or a knowledge-based model, depending on the circumstances. SIMPA searches up to a 17 residue window to find in the crystallographic database the most similar sequence. If this similarity indicates homology, then the tool is doing homology modeling and predicts secondary structure quite well ($Q_3 \approx 86\%$). If the similarity indicates merely analogy, then it is knowledge-based modeling, and the tool does less well ($Q_3 \approx 64\%$).

The Web site does not inform us in this case whether the prediction tool believes that it has identified a homolog. On one hand, the Q_3 for the SIMPA prediction for staufen is a high 82%, which would indicate that SIMPA has found a homolog. On the other hand, the prediction contains a serious misassignment; the first strand of a three strand sheet is assigned as a helix. This implies that SIMPA has *not* found a homolog. The analysis stops here. The perplexities of the three-state score are illustrated well here, as well as the importance to examine closely the details of each prediction to learn the most from a prediction exercise.

5. The L14 Ribosomal Protein

The L14 ribosomal protein is largely built from strands, with a terminal helical region.[304] The

```
MIQQESRLKVADNSGAREVLVIKVLGGSGRRYANIGDVVVATVKDATPGG
      EEEHH      HEEEEEE       H  H HHHHHHHHH        Garnier
EEE  EEEEEE     EEEEEEEEE     EE     EEEEEEE         Hubbard
                 HHHHHEE     HHHHH   EEEE            Livingston
    EEEEEE      EEEEEEEEE     EEEEEEEEEEEEEE         Sander
HHHHH   EE      HHHEEEEE      HHH   EEEEEEE          QL State
HH  H EEEE       HEEEEEE      HH     HEEEEH          QL Profile
    H HEHH       HHEEEEE             HEEHHHH         Combine
EEE   EEEEE       EEEE EE             EEEEE          Matsuo
      EEEEE EE    EEEEEEE    EEEEEE  EEE   EEE       Wilmanns
   EEEEEEEE    EEEEEEEEEE          EEEEEEEEEEEE      experimental

VVKKGQVVKAVVVRTKRGVRRPDGSYIRFDENACVIIRDDKSPRGTRIFG
     HHHHHHHHHHHHHH      EEEH     EEE         EEHH   Garnier
  E   EEEEEEEEE          EEEE    EEEEE        EEE    Hubbard
    HHHHHHHHHHHH                            HHHHH    Livingston
EEEE EEEEE EEEE     E     EEEE   EEEEEE       EEEE   Sander
 E    EEEEEEEEEE  EE     EE      EEEE         EH     QL State
      EEEEEEH            EE     HHEEE         EEE    QL Profile
     HHHHHHEEH           EEEH    HEEE         EEE    Combine
EE       EEEE     HH     EEEE    EEEE        EEEE    Matsuo
                                             EEE     Wilmanns
     EEEEEEEE   EEEEE   EEEEE  EEEEEEE               experimental

PVARELRDKDFMKIISLAPEVI
HHHHHHHHHHHHHHHHH HHH    Garnier
HHHHHHHH    EEEEEEEE     Hubbard
       EEEE EEEEEE       Livingston
HHHHHHHH    EEEEEEEE     Sander
HHHHH   HHHHHHHH   HHH   QL State
HHHHHHHHHHHHHEEH  HHHH   QL Profile
 HHHHHHHHHHHHHHH   H     Combine
 EEE               EEE   Matsuo
           HHHHHHHHHHH   Wilmanns
  HHHHHHH  HHHHHHH       experimental
```

Figure 42. Sequence and predictions from the CASP1 site and experimental[304] secondary structure for the L14 prokaryotic ribosomal protein, (*Bacillus stearothermophilus*) (RL14_BACST, P04450, 1whi). Key: E, β strand; H, α helix. Predictions were generated by the following individuals using the tools indicated: Garnier Simpa,[132] Hubbard (PHD/HMM),[216] Livingston,[293] Sander,[294] Munson Quadratic Logistic (QL),[179,295] Munson/Garnier Combine,[178] Matsuo (thread), and Wilmanns (thread).

predictions based on the PHD neural network identified the critical strand region quite well (Figure 42), although the discrepancies between the "Hubbard" and "Sander" prediction remain. Interestingly, the Sander prediction is marginally better, even though it was evidently built from an unrefined alignment. The QL prediction assigned the terminal helices correctly. The remaining predictions were less successful.

6. The Subtilisin Propiece Segment

Figure 43 shows a collection of predicted secondary structures for the subtilisin propiece segment, compared with the experimental assignment.[305] The figure is self-explanatory. None of the predictions were particularly outstanding, and none were based on a transparent method. Thus, it is difficult to learn from these results.

7. The Replication Terminator Protein

Figure 44 shows a collection of predicted secondary structures for the replication terminator protein compared with the experimental assignment.[306] The figure is self-explanatory. The prediction of Livingston was the best at identifying core secondary structural units. None of the predictions were particularly outstanding, and none were based on a transparent method.

8. Predicting the Conformation of the "Mystery Protein Sequence"

Students in a protein-design course were challenged to design a polypeptide sequence that would fold to form an eight-fold $\alpha-\beta$ barrel. The mystery sequence was synthesized and, evidently, did not form the designed structure.[62] Nevertheless, parameterized prediction tools predicted the designed "structure" well. The extremely accurate secondary structure predictions shown in Figure 45 show that the rules used to predict these barrels are quite similar to the rules taught to students in protein design courses. They are evidently not, however, the rules that Nature uses for folding barrels.

An intriguing paradox is presented if it proves to be easier to predict $\alpha-\beta$ barrels than to design them, as it contrasts with the conventional wisdom that holds presently that design is easier than prediction. From a physical organic chemical perspective, design

```
AGKSNGEKKYIVGFKQTMSTMSAAKKKDVISEKGGKVQKQFKYVDAASATLNEKAVKELK
        eeeeee   eeehhhhhhhhhh     ee    eeeeeehhhhhhhh       Livingston
        EEEEEE   HHHHHHHHH   EE   HHHHHHHHHHHHHHH HHHHHHHHHH  Sander
       EEEEEEEE  HHHHHHHHHHHEEE    HHHHHHHHHHHHHHHHHHHHHHHHH  Hubbard
        EEEE     HHHHHH     HEHH    HHH EEEHHHHHHHHHHHHHHH    QL State
       HEEEE   HHHHHHHH      EHH    HHHHHHHHHHHHHHHHHHHHH     QL Profile
      HHHEEE HHHHHHHHHHH    HEHH     HHHHHHHHHHHHHHHHHHHHHHH  Combine
      eeeEEE eeee   hhHHHHhh     hh e hhhh hhhhhHHHhHHHHHhhh  Sippl
        EEEEEE        HHHHHHHHHH   EEEEE     EEEEE  HHHHHHHH  experimental


KDPSVAYVEEDHVAHAY
    hhhhh    hhhh  Livingston
    EEEEE          Sander
    EEEEHHHHHHH    Hubbard
     EEH H HHHH    QL State
     EE  H HHHHHH  QL Profile
    HHHHHHHHHHHHH  Combine
  hhh   h  hhhhhh  Sippl
    EEEEEE  EEEE   experimental
```

Figure 43. Sequence and predictions from the CASP1 site, and experimental[305] secondary structure for the propeptide of subtilisin BPN′, (*Bacillus subtilis*) (SUBT_BACAM, P00782, 1spb). Key: E, e, β strand; H, h, α helix. Predictions were generated by the following individuals using the tools indicated: Livingston,[293] Sander,[294] Hubbard (PHD/HMM),[216] Munson Quadratic Logistic (QL),[179,295] and Munson/Garnier Combine.[178] The prediction of Sippl was based on threading to ferredoxin (2fxb).

```
MKEEKRSSTGFLVKQRAFLKLYMITMTEQERLYGLKLLEVLRSEFKEIGF
                HHHHHH        EEEEE HHHHHHHHHH       Livingston
          HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH    Sander
LLLL......HHHHHHHHHHHHHHHHHHHHHHH.HHHHHHHHHHHHHHLL   Hubbard
  hhhhhhhh hhHHHHHHHHHH  hhhh  h  hhhhhhhhhHH hhh    Sippl
              HHHHHHHHHHHHHHH EEEE HHHHHHHHHHH       experimental

KPNHTEVYRSLHELLDDGILKQIKVKKEGAKLQEVVLYQFKDYEAAKLYK
       EEEEEEEEE  EEEEEEEE               HHHHHHHHH   Livingston
  HHHHHHHHHHH    HHHHHEHHHHH    HHHHHHHHH HHHHHHHH   Sander
LLL.HHHHHHHHHHH.HHHHHH..H...L..HHHHHHHHH.HHHHHHHHH   Hubbard
  hHHHHHHhHHHHhh   hh     hh  hhhhEEEEee  hhhhhhhh   Sippl
   HHHHHHHHHHHHHH  EEEEEEE       EEEEEEE HHHHHHHHH   experimental

KQLKVELDRCKKLIEKALSDNF
hhhhh                    Livingston
HHHHHHHHHHHHHHHHH        Sander
HH ...HHHHHHHHHHHHH..L   Hubbard
hh hhhh hhhhhhhhhhhhhh   Sippl
HHHHHHHHHHHHHHHHHHHHHH   experimental
```

Figure 44. Sequence and predictions from the CASP1 site and experimental[306] secondary structure for replication terminator protein (RTP), (*Bacillus subtilis*) (RTP_BACSU, P14382). Key: E, e, β strand; H, h, α helix; L, loop; ., unassigned. Predictions were generated by the following individuals using the tools indicated: Livingston,[293] Sander,[294] and Hubbard (PHD/HMM).[216] The Sippl prediction was based on threading to the globular domain of histone-H5 (1hst).

is not subject to the Darwinian process of random modification subject to functional constraints; it is quite clear that a designed peptide that does not fold can be just a few amino acids away from a peptide that does fold. It is interesting to note that much of organic chemistry is presently focused on "combinatorial" methods. These are, of course, the organic chemistry analogy of Darwinian evolution.

VII. Using Evolution-Based Predictions of Secondary Structure

Both the pre-CASP1 predictions and the CASP1 project itself showed that transparent methods could predict the secondary structure of proteins reliably, with neural networks improving over this period to come to match more closely the predictions made by transparent methods. Nevertheless, the prediction methods could not guarantee models free of all serious misassignments. While "perfect" predictions exist, and most of the later models assigned most core secondary structural elements correctly, predictions for a large protein contained on average one core element that was misassigned. This limitation was only partly mitigated by evolutionary analyses that were frequently able to identify in advance the problematical assignment(s) and to alert the user to the possibility that alternative secondary structural

```
....,....1....,....2....,....3....,....4....,....5....,....6
MKAGVFIQGIGPEAKQLAANFAKNGLYVIVAGGKPEACQALAKNGPKIVVIQGIGPEAKQ
   EEEEE    HHHHHHHHHH   EEEEEE   HHHHHHHHH     EEE     HHHH Sander
             HHHHHHHH     HHHHHHHHHH   EEEEE    EEE       HH Livingston
HHH EEEE     HHHHHHHHHH  EEEEE     HHHHHH     EEEEEE     HHH Munson: QL State
HHH EEEE     HHHHHHHHHH  EEEEE     HHHHHH     EEEEEE     HHH Munson: QL Profile
    EEE     HHHHHHHHHHHHHHEEEE     HHHHHH     HEEEEE    HHHH Combine
    EEEE     HHHHHHHHHH   EEEE      HHHHHHH    EEEEE     HHH Barton
    EEEe   HHHHHHHHHHHHh eeEeee   hHHHHHHHHh eeeEEeeeeehhHHH Sippl

....,....7....,....8....,....9....,....10...,....11..,....12
LAANFAKNGLIVIVAGGKPEACQALAKNGPKVVIIQGIGPEAKELAANFAKEGLWVIVAG
HHH      EEEEEE     HHHHHHH    EEEE    HHHHHHHHHHHHH  EEEEE  Sander
HHHHH     HHHHHHHHHH   EEEEE    EEE      HHHHHHHH       HHH  Livingston
HHHHHHH  HEEEE     HHHHHH     EEEEEE     HHHHHHHHHHHHHEEEE   Munson: QL State
HHHHHHH  HEEEE     HHHHHH     EEEEEE     HHHHHHHHHHHHHEEEE   Munson: QL Profile
HHHHHHHHHHEEEE     HHHHHH     HEEEEE    HHHHHHHHHHHHHHEEEE   Combine
HHHHHHH   EEEE      HHHHHHH    EEEEE     HHHHHHHHHH   EEEE   Barton
HHHHhhhh EEEEee   HHHHHHHHHh eeEEEEeee HHHHHHHHHHHhh  eeeeee Sippl

....,....13...,....14...,....15...,....16...,....17...,....
GKPEACEALAKNGPKVVVIQGIGPEAKELAANFAKEGLIVIVAGGKPEACEALAKAAAN
  HHHHHHHHH   EEEEEEE  HHHHHHHHHHHH  EEEEEE   HHHHHHHHHHHH   Sander
HH    EEEEEEE    EEE     HHHHHHHH   EEEEEE                   Livingston
    HHHHH      EEEEE     HHHHHHHHHHHHHEEEE      HHHHHHHHHHH  Munson: QL State
    HHHHH      EEEEE     HHHHHHHHHHHHHEEEE       HHHHHHHHHH  Munson: QL Profile
   HHHHHHHH    EEEEE    HHHHHHHHHHHHHHEEEE     HHHHHHHHHH    Combine
    HHHHHHH    EEEEE     HHHHHHHHHH   EEEE      HHHHHHH      Barton
 hHHHHHHhhhheeeeEEEeee hhhHHHHHHHHhh eeeeee eehhHHHHHHHHHHH  Sippl
```

Figure 45. Sequence and predictions from the CASP1 site for the "Mystery protein", a protein designed in a course to fold as an eight-fold α−β barrel; when the protein was synthesized, it evidently did not form the designed structure.[62] Nevertheless, parameterized prediction tools "predicted" the designed "structure" well. Key: E, e, β strand; H, h, α helix. Predictions were generated by the following individuals using the tools indicated: Livingston,[293] Sander,[294] and Munson Quadratic Logistic (QL),[179,295] and Munson/Garnier Combine.[178] The Sippl prediction was based on threading to 1pgd (platelet-derived growth factor) while that of Barton was to 5rub (ribulose 1,5-bisphosphate).

models must be built (as in the protein serine/threonine phosphatases).[96]

Despite these limitations, the secondary structure predictions made in the 1993−1996 period were of sufficient quality to give them practical value. As this was the first time that this could be said for any prediction methodology, this represents progress. In several cases, predicted secondary structure models have been used to identify antigenic determinants in a protein family,[307] guide and interpret site-directed mutagenesis studies,[308] identify phosphorylation and glycosylation sites in proteins, assist in experiments to immobilize proteins, and bias combinatorial libraries when searching for protein ligands. Two other applications are discussed in detail below.

A. Detecting Long Distance Homologies

Secondary structure predictions may be used to identify long-distance homology between protein families with only marginal sequence similarities.[92] Often, comparison of two protein sequences identifies motifs, short stretches of polypeptide that are suggestive of homology between two protein families.[309] By themselves, common motifs are not proof of homology, as the probability that such sequence motifs emerged by random chance in evolution is high. Thus, after identifying a motif, the issue then becomes whether the motifs are true indicators of homology, or whether they arose by convergent evolution.

Secondary structure predictions allow this question to be addressed in several cases. Most simply, the secondary structural elements flanking the motifs in the two protein families are compared. If the motif truly indicates distant homology, it should be embedded within the same secondary structural elements. Most simply, four embeddings are possible for a motif: helix−motif−helix, strand−motif−helix, helix−motif−strand, and strand−motif−strand. If the motif is not embedded in the same secondary structural elements in two protein families, the motif is not a likely indicator of homology.

Alternatively, the number and sequence of the secondary structure elements can be compared overall. Here, the distinction between core and peripheral secondary structural elements, apparent in a consensus model, is important. Simple segment-by-segment comparison of secondary structural elements will prevent clear identification of homologs if the comparison includes secondary structural elements that are not likely to be conserved.

Perhaps the most striking case where secondary structure predictions were used in this fashion is in the protein kinase prediction.[91] Many had conjectured that because protein kinase shared the sequence motif, Gly-Xxx-Gly-Xxx-Xxx-Gly with other kinases, protein kinases were homologous to these other kinases, and would adopt the same fold as other kinases. Several models of the overall fold of protein kinase were built on the basis of this assumption. In the prediction made using contemporary methodology, it was noted that the motif was not flanked by the same secondary structural elements, and that this implied that protein kinase adopted a fold different from that found in other kinases. The conclusion was that the core domain most likely contained an antiparallel β sheet.[91] The experimental structure proved the prediction to be correct. While many examples are now available where secondary structure predictions have been used to confirm suspicions of long-distance homology, this is (we believe) the first time that a secondary structure prediction has been used to deny long-distance homology.

The use of predicted secondary structural models to assign long-distance homology is now becoming commonplace. Two recent examples involve the assessment of long-distance homology among pyridoxal-dependent enzymes[92] and the ribonucleotide reductases.[251] Such research is in part based on the notion that practical solutions to the structure prediction problem are most likely to come from the recognition of existing (known) structures that fit the sequence of the unknown structure.[310]

At one level, use of predicted secondary structural elements can be viewed as threading, but using predicted secondary structural elements instead of sequence. Russell *et al.* recently extended the ideas outlined above more systematically.[311] This suggests a bright future for applying predicted secondary structures to detect long distance homologs. Already, in the setting of the pharmaceutical industry, these are among the most widespread applications of secondary structural models predicted using transparent tools.

B. Building Supersecondary and Tertiary Structural Models

The second application of a secondary structure prediction is, of course, the prediction of supersecondary and tertiary structure. Virtually all predictions using contemporary methods make an attempt to build such models. In general, the overall features of the core fold have been correctly assigned. Thus, the antiparallel cores of protein kinase, cyclin, and synaptotagmin were all correctly predicted (see above), as were the parallel cores of protein serine/threonine phosphatase, the proteasome, and other structures.

It remains a difficult task to identify the precise orientation of secondary structural elements within an overall model. As discussed above, the synaptotagmin prediction narrowed the possibilities to just three, one of which was correct. Indeed, it is a frequent occurrence for a tertiary structural model to be largely correct, except for the swap of a β strand or the reorientation of a helix.

To facilitate the development of procedures to take this final step in the construction of consensus models for protein folds, improved computational tools are necessary that assemble predicted secondary structural elements into supersecondary and tertiary structural models. No such tools exist today, although some steps in this direction are now being taken.[63,312] As noted throughout this review, such tools would be useful not only in building tertiary structural models, but also in refining secondary structure models in difficult regions (for example, near an active site). Predictors are already attempting to refine secondary structure predictions by determining which of a small number of alternative models is most easily assembled to give a tertiary structure, and computer assistance would be warmly welcomed in this area.

A second obstacle to obtaining better tertiary structural models is the absence of reliable long-distance constraints on the fold. Several approaches are emerging that might help obtain these long distance constraints. Long-distance compensatory covariation, where amino acids not adjacent in the polypeptide chain undergo correlated substitution, may identify supersecondary structural units.[91,313−316] Again, the protein kinase prediction offers a paradigmatic example, where a long-distance charge compensatory covariation was used to orient two strands antiparallel.[91] More recently developed tools were applied in the CASP2 project (see below). Chain connectivity also proves to be a powerful tool for assembling the topology of β sheets, as outlined many years ago by Cohen *et al.*[317] Further rules must be developed to identify different types of connecting loops from patterns of variation and conservation in a family of proteins. Finally, if disulfide bonds are present with known connectivity, many conceivable folds can be excluded. To date, no reliable tools are available for predicting disulfide connectivity from sequence data alone.

For many of the predictions above, secondary structural models were used to generate tertiary structural models with varying degrees of resolution. Surprisingly, virtually all of them were correct, at least as far as they went. The antiparallel sheet in the first domain of protein kinase,[91] and the three folds of synaptotagmin,[297] are cases of *a priori* tertiary structural modeling based on predicted secondary structural units. In some predictions of $\alpha-\beta$ barrels, in the cyclins,[204] and in the cytokine receptors,[201] the tertiary structural modeling perhaps might viewed as threading. However, given that all biochemists have known about helices and strands since their introductory biochemistry courses, all prediction is partly knowledge-based modeling.

A particularly interesting case is the *bona fide* consensus prediction for the chaperonin GroES.[318] Many items of information were brought to bear on the modeling problem, including experimental information from electron microscopy, NMR, and FT infrared spectrometry). As Figure 46 shows, the predicted and experimental structures are quite similar.[319] Because of the input of substantial amount of experimental data concerning conformation, the prediction cannot be regarded as truly *ab initio.* However, it does show how a highly accurate model could be built in 1996 from a combination of biophysical and theoretical data.

VIII. The CASP2 Prediction Project

The successor to the CASP1 project was the CASP2 prediction project, which was completed in December

```
MNIRPLHDRVIVKRKEVETKSAGGIVLTGSAAAKSTRGEVLAVGNGRILE
        EEEEEE          EEEE         EEEEEE        prediction ref 318
EEE     EEEEEE  EEE     EEEE  EEE      EEEEE       prediction ref 129
 EEEEEE EEEEEEEEEEEEE   EEEEEE      EEEEEEEEE EEEE experiment

NGEVKPLDVKVGDIVIFNDGYGVKSEKIDNEEVLIMSESDILAIVEA
   EEEEEEE   EEEEEE    EEEEE   EEEEEEE  EEEEEE     prediction ref 318
      EEEEE  EEEE   EEEE        EEEEE  EEEEEEE     prediction ref 129
 EEEEEE      EEEEEE   EEEEEEE EEEEEEEE  EEEEEEE    experiment
```

Figure 46. Representative sequence, predictions,[129,318] and experimental[319] secondary structure for GroES.

1996. Few events of the year show more convincingly how the field of structure prediction has changed since the early 1990s. Some 70 research groups participated in the project, showing that *bona fide* prediction is now widely accepted by practitioners. The project attracted the attention of those outside the field as well, particularly among experimental biochemists who were encouraged by the rigor of *bona fide* predictions.[130] The protein sequence databases had grown further, making more targets susceptible to evolution-based analysis. And the number of correct secondary structure predictions was higher in CASP2 than in CASP1, as was the number of times correct inferences concerning tertiary structure and distant homology were drawn from a correct secondary structure prediction.

A. Design of the CASP2 Prediction Project

As with the CASP1 project, the targets in the CASP2 prediction project fell into several categories. For the first time, the project included a set of docking problems. Here, the task was to predict how two molecules of known structure would interact.

The remaining tasks were analogous to those presented in the CASP1 project. Comparative modeling targets were chosen to be proteins whose sequences and folds were both similar (but not identical) to those of proteins in the PDB crystallographic database. The challenge was to predict how the structure of the target protein differed from the structure of the homolog with known structure.

The third task concerned “fold recognition targets”, proteins having folds similar overall to proteins in the PDB crystallographic database, but where a typical sequence search would not indicate homology between the target and known protein. The challenge associated with these targets was to identify the structure in the crystal database that had the same fold as the target protein, starting from the assumption that such a structure existed. This challenge was most often approached using tools related to profile analysis or threading.

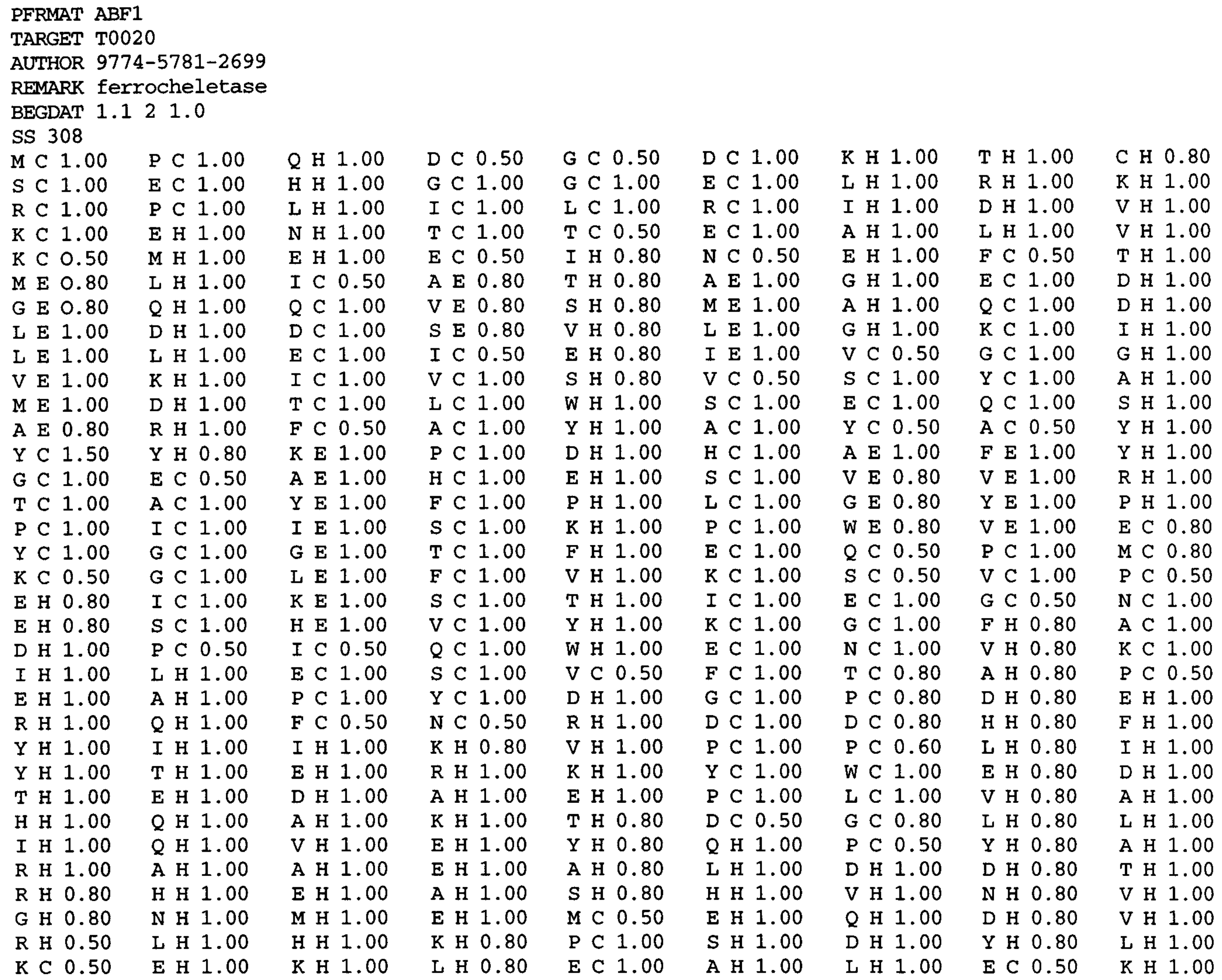
Most relevant to this review were the *ab initio* tasks presented in the CASP2 project. As with the fold recognition tasks, these required conformational predictions to be made for proteins sharing no obvious sequence similarity to proteins with known conformations. The task was distinct from the “fold recognition” challenges in the way in which the predictions were made. Fold recognition methods presume that a similar fold exists in the database, and try to find it. As discussed above, *ab initio* predictions are made with no explicit attempt to identify a fold in the database. The former must fail if the target protein has a unique fold, while the latter need not.

As in CASP1, *ab initio* predictions in CASP2 were approached in two very different ways. The first used force field or simulation methods together with computational search algorithms to find a global energy minimum for the protein sequence. The second approach was evolution based, attempting to extract conformational information from a set of homologous proteins whose sequences had been placed in a multiple alignment.

In CASP2, many of the methods discussed above were applied in their latest form. These included tools that began by predicting features of tertiary structure in the protein (surface residues, interior residues) as discussed above, tools that predicted secondary structure directly (as in a consensus classical approach), and tools for finding contacts between residues by compensatory covariation analysis.[91,313−316] Embolded by successes in CASP1 and elsewhere, several groups then attempted to assemble predicted secondary structural elements to generate models for supersecondary and tertiary structure.

Unlike those in CASP1, where different submission formats from different groups created problems for evaluators, submissions to CASP2 were made using a uniform set of formats, adjusted to allow description of the predicted models at the different levels of resolution implied by different prediction tools. At the lowest level of resolution were predictions that provided a simple secondary structure model for the protein sequence. An example of the format is shown in Figure 47, which contains a prediction for ferrochelatase, one of the CASP2 targets. The sequence is read vertically. The first column is the amino acid of the target protein (one letter code). The second column allows the predictor to assign secondary structure by choosing one of three states (C = coil; H = helix; E = strand). A feature of the submission format allowed the predictor to designate, residue by residue, a reliability of the secondary structure assignment. This was done by providing a number from 0 to 1 to indicate increasing confidence in the assignment. This feature conformed to the output of several automated prediction tools.

The successes in predicting secondary structure, including the correct modeling of the tertiary structure of phospho-β-galactosidase from predicted secondary structural elements, encouraged several groups to attempt to assemble the predicted secondary structural elements into supersecondary structural models and tertiary structural models. This brought



```
PFRMAT ABF1
TARGET T0020
AUTHOR 9774-5781-2699
REMARK ferrocheletase
BEGDAT 1.1 2 1.0
SS 308
M C 1.00   P C 1.00   Q H 1.00   D C 0.50   G C 0.50   D C 1.00   K H 1.00   T H 1.00   C H 0.80
S C 1.00   E C 1.00   H H 1.00   G C 1.00   G C 1.00   E C 1.00   L H 1.00   R H 1.00   K H 1.00
R C 1.00   P C 1.00   L H 1.00   I C 1.00   L C 1.00   R C 1.00   I H 1.00   D H 1.00   V H 1.00
K C 1.00   E H 1.00   N H 1.00   T C 1.00   T C 0.50   E C 1.00   A H 1.00   L H 1.00   V H 1.00
K C O.50   M H 1.00   E H 1.00   E C 0.50   I H 0.80   N C 0.50   E H 1.00   F C 0.50   T H 1.00
M E O.80   L H 1.00   I C 0.50   A E 0.80   T H 0.80   A E 1.00   G H 1.00   E C 1.00   D H 1.00
G E O.80   Q H 1.00   Q C 1.00   V E 0.80   S H 0.80   M E 1.00   A H 1.00   Q C 1.00   D H 1.00
L E 1.00   D H 1.00   D C 1.00   S E 0.80   V H 0.80   L E 1.00   G H 1.00   K C 1.00   I H 1.00
L E 1.00   L H 1.00   E C 1.00   I C 0.50   E H 0.80   I E 1.00   V C 0.50   G C 1.00   G H 1.00
V E 1.00   K H 1.00   I C 1.00   V C 1.00   S H 0.80   V C 0.50   S C 1.00   Y C 1.00   A H 1.00
M E 1.00   D H 1.00   T C 1.00   L C 1.00   W H 1.00   S C 1.00   E C 1.00   Q C 1.00   S H 1.00
A E 0.80   R H 1.00   F C 0.50   A C 1.00   Y H 1.00   A C 1.00   Y C 0.50   A C 0.50   Y H 1.00
Y C 1.50   Y H 0.80   K E 1.00   P C 1.00   D H 1.00   H C 1.00   A E 1.00   F E 1.00   Y H 1.00
G C 1.00   E C 0.50   A E 1.00   H C 1.00   E H 1.00   S C 1.00   V E 0.80   V E 1.00   R H 1.00
T C 1.00   A C 1.00   Y E 1.00   F C 1.00   P H 1.00   L C 1.00   G E 0.80   Y E 1.00   P H 1.00
P C 1.00   I C 1.00   I E 1.00   S C 1.00   K H 1.00   P C 1.00   W E 0.80   V E 1.00   E C 0.80
Y C 1.00   G C 1.00   G E 1.00   T C 1.00   F H 1.00   E C 1.00   Q C 0.50   P C 1.00   M C 0.80
K C 0.50   G C 1.00   L E 1.00   F C 1.00   V H 1.00   K C 1.00   S C 0.50   V C 1.00   P C 0.50
E H 0.80   I C 1.00   K E 1.00   S C 1.00   T H 1.00   I C 1.00   E C 1.00   G C 0.50   N C 1.00
E H 0.80   S C 1.00   H E 1.00   V C 1.00   Y H 1.00   K C 1.00   G C 1.00   F H 0.80   A C 1.00
D H 1.00   P C 0.50   I C 0.50   Q C 1.00   W H 1.00   E C 1.00   N C 1.00   V H 0.80   K C 1.00
I H 1.00   L H 1.00   E C 1.00   S C 1.00   V C 0.50   F C 1.00   T C 0.80   A H 0.80   P C 0.50
E H 1.00   A H 1.00   P C 1.00   Y C 1.00   D H 1.00   G C 1.00   P C 0.80   D H 0.80   E H 1.00
R H 1.00   Q H 1.00   F C 0.50   N C 0.50   R H 1.00   D C 1.00   D C 0.80   H H 0.80   F H 1.00
Y H 1.00   I H 1.00   I H 1.00   K H 0.80   V H 1.00   P C 1.00   P C 0.60   L H 0.80   I H 1.00
Y H 1.00   T H 1.00   E H 1.00   R H 1.00   K H 1.00   Y C 1.00   W C 1.00   E H 0.80   D H 1.00
T H 1.00   E H 1.00   D H 1.00   A H 1.00   E H 1.00   P C 1.00   L C 1.00   V H 0.80   A H 1.00
H H 1.00   Q H 1.00   A H 1.00   K H 1.00   T H 0.80   D C 0.50   G C 0.80   L H 0.80   L H 1.00
I H 1.00   Q H 1.00   V H 1.00   E H 1.00   Y H 0.80   Q H 1.00   P C 0.50   Y H 0.80   A H 1.00
R H 1.00   A H 1.00   A H 1.00   E H 1.00   A H 0.80   L H 1.00   D H 1.00   D H 0.80   T H 1.00
R H 0.80   H H 1.00   E H 1.00   A H 1.00   S H 0.80   H H 1.00   V H 1.00   N H 0.80   V H 1.00
G H 0.80   N H 1.00   M H 1.00   E H 1.00   M C 0.50   E H 1.00   Q H 1.00   D H 0.80   V H 1.00
R H 0.50   L H 1.00   H H 1.00   K H 0.80   P C 1.00   S H 1.00   D H 1.00   Y H 0.80   L H 1.00
K C 0.50   E H 1.00   K H 1.00   L H 0.80   E C 1.00   A H 1.00   L H 1.00   E C 0.50   K H 1.00
```

Figure 47. A transparent *bona fide* prediction prepared by the Benner group for ferrochelatase, showing the new format for the submission of *bona fide* secondary structure predictions used in CASP2. The sequence is read vertically. The first column is the amino acid of the target protein (one letter code). The second column is the secondary structure prediction (C, coil; H, helix; E, strand). The number (0 to 1) allows the predictor to assign a reliability to the assignment. This format standardized submission of secondary structure predictions, facilitating their evaluation.

secondary structural elements into contact with each other. Lesk recently proposed a terminology to describe segment contacts, a terminology that allows a low-resolution description of a model.[320] The terminology provides an excellent way to describe consensus predictions, and CASP2 adopted this terminology for this purpose.

At the highest level of resolution, atomic coordinate sets could be submitted. These were the preferred submission format for those who did energy optimizations. The organizers applied a set of tools to convert these into secondary structural models and contacts.

The targets that were presented for *ab initio* predictions are collected in Table 10, together with data concerning the target and the evolutionary family to which it belongs. The predictors and their "predictor numbers" are collected in Table 11. Primary information on the CASP2 predictions is provided on the Prediction Center Web Page (URL:http://PredictionCenter.llnl.gov/WWW/casp2/evaluation.html).

B. Evaluation of the *ab Initio* Portion of the CASP2 Project

Arthur Lesk judged the *ab initio* portion of CASP2, and his scholarly assessment[174] is its official evaluation. Judging can be the least rewarding part of such projects, and it is a please to note the number of individuals, including the authors of this review, who appreciated the collegiality and intellectual precision that Lesk brought to this task. Lesk cited the neural network of ROST as the best tool for generating secondary structure models, the tool of JONES for producing the 3D structure predictions, the team of Olmea, Pazos and Valencia (VALENCIA) for assigning residue−residue contact patterns the best, and the COBEGETJ team of Cohen, Benner, Gerloff, Turcotte, and Joachimiak (the COHEN and BENNER predictions in the Figures) for making the best segment contact patterns.

Lesk recognized, of course, that his summary could not cover everything that was important in the project and depended on criteria that were, again by necessity, arbitrary. Accordingly, Lesk outlined in some detail his criteria for judging predictions.

Table 10. Summary of Prediction Targets for the CASP2 *ab Initio* Project

target	short name	length	no. of homologs PHD	no. of homologs COBEGET	PAM width of family	major difference DSSP vs STRIDE[174]	other information	fate
T0002	threonine deaminase	514	13	13	130	yes	homolog of Trp synth	8 predictions
T0004	polyribonucleotide nucleotidyltransferase	84	20	16	130	no	homolog with known structure	12 predictions
T0005	fibrinogen	268	22	17	126	no	good target	10 predictions, 1 transparent
T0010	bactericidal permeability protein	456	4	4	100	no	too few sequences	7 predictions
T0011	heat shock protein 90	220	44	31	100	no	good target	11 predictions, 2 transparent
T0012	procaricain	107	24	17	120	no	good target	6 predictions
T0014	dehydroquinase	252	5	4 + 2	80 + 80/100	yes	too few sequences	10 predictions
T0016	peridinin	312	4	1	18	no	too few sequences	7 predictions
T0020	ferrocheletase	320	4	12	215	no	good target	17 predictions, 1 transparent
T0022	fucose isomerase	591	2	2	40	no	too few sequences	8 predictions
T0030	protein G3	66	4	4 (2 + 2)	140	yes	too few sequences	20 predictions
T0031	exfoliative toxin A	242	2	3	140	yes	too few sequences	19 predictions
T0032	cryptogein	98	12 homeobox	11	30	yes	sec. structure published in 1994	8 predictions
T0037	calponin	109	20	18	170	no	good target	20 predictions, 1 transparent
T0038	CBDN1	152	3	2 + 1 (part)	64	no	homolog with known structure	16 predictions
T0042	NK-lysin	78	3	20	200	no	good target	22 predictions

First, Lesk understood that different methods for assigning reference secondary structure to experimental coordinates might not make consistent assignments. An ambiguous reference structure creates ambiguous scores (section II.A). Lesk therefore examined the secondary structural assignments made by both DSSP and STRIDE to the target proteins. Three-state (Q_3) assignments were found to disagree at from 2% to 14% of the residues. Lesk noted that in five of the 16 targets listed in Table 10, DSSP identified one secondary structural element (strand or helix) that was not identified by STRIDE, or vice versa. Considering these differences to be small, Lesk based his assessment of secondary structure predictions based on a comparison with DSSP assignments alone.

Two other features characterized the official evaluation. First, it relied on Q_3 scores to judge secondary structure predictions. A prediction was counted in the official evaluation if and only if it had a Q_3 greater than 68%. A list was prepared of predictors who had contributed a prediction for each target that had a Q_3 score of 68% or greater. A manuscript version of this list is reproduced in Table 12. The predictor producing the highest Q_3 score for each target was also noted. A histogram was prepared that listed, by predictor, the number of predictions that they each made with a Q_3 greater than 68%.

Second, to evaluate the relative performance of different methods, Lesk counted the *total number* of predictions with a Q_3 greater than 68%. No normalization was made for the number of targets predicted by each method. This approach was designed as a way to identify methods that produced a “sustained good performance, rather than good results only occasionally”. This analysis led to the official assessment that the secondary structure prediction tools of ROST, JAAP, SOLOVYEV, and STERNBERG were the most powerful for predicting secondary structure, as these were the tools that generated the largest absolute number of predictions with $Q_3 > 68\%$ for the 16 targets designated by the conference organizers as being appropriate for *ab initio* prediction.

C. Problems Encountered in Judging the CASP2 *ab Initio* Predictions

Earlier sections of this review have discussed some of the problems associated with evaluating predictions of protein conformation. Several points are clear. First, and most important, to compare different methods, predictions of conformation are best made in parallel on the same protein targets. Especially for evolution-based predictions, where the number and divergence of proteins in a family can differ widely by family, some targets are “easier” than others.

Once a uniform set of targets is chosen, it is best to evaluate the predictions using tools that reflect the value of the prediction in addressing further structural and biological questions. Q_3 scores are at best only a crude indicator of this value, and cannot be reliably used even to provide a cutoff to distinguish models that are worthy of further examination from those that are not (see section II). For the purpose

Table 11. The Predictors and Their "Predictor Numbers" in the CASP2 *ab Initio* Project

predictor number	predictor	predictor number	predictor
1	ABAGYAN	60	ROSE
8	AVBELJ	61	ROST
9	BAKER	67	SERVER_DSC_MULT
11	BAZAN	68	SERVER_GOR
12	SOLOVYEV	68	SERVER_GOR
18	COHEN	69	SERVER_NNPREDICT
23	EISENBERG	69	SERVER_NNPREDICT
28	FINKELSTEIN	70	SERVER_NNSSP_MULT
33	GOLDSTEIN	71	SERVER_PREDICTPROTEIN
37	HUBBARD	72	SERVER_PREDICTPROTEIN_SINGLE
38	JAAP	73	SERVER_SSPRED
41	JONES	74	SERVER_SSP_MULT
48	LENGAUER	76	SHESTOPALOV
50	MARSHALL	78	SMITH
51	MOULT	80	STERNBERG
52	MUNSON	81	BENNER
53	MURZIN	83	TAYLOR
55	OSGUTHORPE	88	VALENCIA

Table 12. Predictions for CASP2 Targets[174]

target	no. of attempts	max Q_3, %	group with highest score	groups with $Q_3 \geq 68$
T002	9	76	12	12,38,61
T004	24	83	80	28,38,52,53,61,80,88
T005	15	73	37	18,37
T010	7	70	61	61
T011	14	74	61	11,18,61,80,88
T012	7	92	1	1,12,38,52,61,80
T014	20	80	61	12,38,52,61,80
T016	8	84	80	12,37,38,52,53,61,80
T020	19	80	70	12,18,33,37,61,71
T022	8	72	12,61	12,33,38,52,61,80
T030	33	66	61	
T031	22	66	12	
T032	8	80	88	52,88
T037	20	83	12	9,12,18,37,38,61,67,71,80,88
T038	16	76	70	12,33,61,69,70,74,80
T042	28	90	61	9,12,18,23,38,41,50,51,53,61,67,71,72,12,80,81

of judging a contest, where time is limited, they are acceptable as a way of comparing the quality of different predictions made for the same target. However, as a Q_3 score can be arbitrarily low depending on the extent of noncore elements contained in the reference experimental structure, a cutoff score (for example, 68%) chosen without reference to evolutionary issues will be unsatisfactory in many cases. The prediction discussed above for phospho-β-galactosidase (from CASP1), for example, had a Q_3 score of only 65%, but nevertheless yielded a correct core tertiary structural model.

Last, assessment choices can bias the assessment. For example, the decision in the official assessment in CASP2 to rank different prediction methods relative to each other by counting the absolute number of targets for which each method generated a Q_3 score > 68% favors methods that make more predictions over those that make fewer, without considering why some predictors might choose not to make a prediction for any particular target. Let us look at the details of how these factors make the official assessment of the *ab initio* predictions of CASP2 problematic.

1. Different Participants Made Predictions for Different Targets

To evaluate the relative merits of different prediction methods, the methods must be tested in parallel on the same set of prediction targets. The hope in CASP2 was that a specific list of targets suitable for *ab initio* prediction would provide this set, and that all methods would be applied to all members of this set. This would enable the different methods to be directly compared.

For a variety of reasons, not all participants in the CASP2 project predicted conformation for all targets. Somewhat trivially, participants were constrained by time and resources in their selection of prediction targets, with manual and transparent methods obviously more constrained than automated methods. For example, Bazan provided an outstanding prediction for the secondary structure of target T0011 by a process that involved manual analysis of neural network data and other inputs. He then converted his secondary structure prediction into a largely correct model for the tertiary structure of the protein. It is difficult to imagine a single individual being able to repeat an analysis of such depth on 16 targets. This does not mean that Bazan's approach was inferior to that of the automated approaches. But the official assessment could not rate his approach highly because it generated only a single successful prediction, and multiple successful predictions were required to attract a positive evaluation from the assessors.

Perhaps more trivially, if a tool were applied in a collaboration, where different members of the collaborative team submitted predictions under different predictor numbers, this would diminish the number of predictions any individual participant would be credited for. This would decrease the likelihood that the collaboration would be recognized favorably by an assessment that favored large numbers of predictions submitted under a single predictor. In CASP2, such collaborations existed, for example the collaboration among Cohen, Gerloff, Benner, Turcotte, and Joachimiak (the COBEGETJ team), which involved a work done in San Francisco, Florida, and Switzerland.

In several cases, targets in the CASP2 *ab initio* list were found during the course of the project to be inappropriate for an *ab initio* prediction exercise. For example, cryptogein was entered as a target for the *ab initio* competition (target number T0032) and predictions for it were recorded and officially scored. Gerloff, a member of the COBEGETJ team, realized while considering this target that a secondary structure of the protein had already been published.[321] The conference organizers were informed, the information was distributed via CASP2-Newsflash, and the COBEGETJ team did not submit a prediction for target T0032. Several groups using automated tools did.

Other targets were considered to be inappropriate for an *ab initio* prediction because a homolog was suggested to be in the crystallographic database, making the target more appropriate for homology modeling. For example, the group submitting threonine deaminase (CASP2 target T0002) indicated that it might be a homolog of the β subunit of tryptophan synthase, a protein with a crystal structure in the PDB (PDB entry number 1WSY-B). Several contestants considered this to be an indication that threonine deaminase was not an appropriate target for the *ab initio* effort and did not submit predictions.

Some of the targets for the *ab initio* phase of the CASP2 contest were also poorly suited for evolution-based predictions. As discussed at length in section V, an evolution-based structure prediction will be more accurate for families with more sequences having greater overall evolutionary divergence. If a family has multiple members, but the sequences of those members are all very similar, an evolution-based analysis is little better than a prediction made with a single sequence.

As CASP2 was not an explicit test of evolution-based methods, these considerations did not influence the selection of targets for the *ab initio* portion of the contest. Participants making transparent predictions using evolution-based methods therefore generally examined each of the targets to determine the number and evolutionary divergence of homologs before making a prediction, and did not make a prediction if the family contained too few proteins or proteins with too little divergence. Thus, T0014 had only five homologs, too few to support a strong evolution-based structure prediction. Targets T0010 and T0030 had only three identifiable homologs, T0031 had only two, T0022 had only one, and T0038 had none. Thus, those making transparent predictions using evolutionary analyses generally did not make predictions for these targets.

After excluding CASP2 targets having a homolog with a known structure, targets whose experimental structures had already been published, and targets with few homologs in the database, only six targets remained suitable for *ab initio* prediction using evolution-based analyses: fibrinogen (T0005), heat shock protein 90 (T0011), procaricain (T0012), ferrocheletase (T0020), calponin (T0037), and NK-lysin (T0042). As a rule, those using transparent evolution-based methods made predictions for some set of these targets, while automated tools made predictions for more targets. This gave transparent, evolution-based methods an advantage, as they tended to select targets more suited for their prediction methods. On the other hand, the decision in the assessment to rank methods based on the absolute number of predictions made favored those who made as many predictions as possible. As discussed below, this created artifacts in the evaluation.

2. The Q_3 Score

Another problem in the official evaluation was the heavy reliance on Q_3 to score the predictions. As noted above, use of the Q_3 score is an understandable expedient when judging a prediction project under time constraints. As the project is now completed, we can now at leisure examine the results to see whether the limitations in the Q_3 score had an impact on the overall value of the assessment.

As discussed at length in section II, the Q_3 score for a "perfect" prediction can be arbitrarily low, depending on the fraction of the experimental structure that represents inserted elements relative to the core. The prediction of phospho-β-galactosidase, from the CASP1 project, provides a good illustration of this point.[79] The Q_3 obtained for this prediction was only 65%; it would therefore have been excluded using the official criteria applied in CASP2. Nevertheless, the prediction was adequate to build a correct low-resolution model of the tertiary structure of the conserved core. This was possible because the mistakes that generated the "low" Q_3 score were concentrated in noncore regions.[62,79] Thus, the relevant issue in evaluating a consensus prediction is the number of serious mistakes (mistaking core helices for strands and core strands for helices) it contains.

A similar circumstance arose in the CASP2 project. The BENNER prediction of fibrinogen had a Q_3 score of 65%, again too low to be identified using the official criteria. As with phospho-β-galactosidase, the mistakes were concentrated in noncore regions (see below), making the prediction useful despite its low score (see below).

As discussed in section II, no single number can accurately reflect the value of a secondary structure prediction. If one is desired, the preferable one would count the number of core secondary structural elements that are successfully identified. The overlap of the predicted and experimental secondary structural elements is not especially critical, provided that the correct number and type is obtained. No "overlap" evaluation tool was applied in CASP2; the S_{ov} tool, which scores for the amount of overlap in predicted and experimentally assigned segments,[75] perhaps came the closest. In the CASP2 project, when S_{ov} is used instead of Q_3, the list of "good" methods for predicting secondary structure expands from the four cited in the official evaluation (ROST, JAAP, SOLOVYEV, STERNBERG) to include three more (VALENCIA, BAZAN, and COBEGETJ).

An intriguing phenomenon lies behind this observation. Inspection of the outputs from the neural network automata shows that these tools routinely have Q_3 scores 3−5 percentage points higher than their S_{ov} scores. In contrast, the Q_3 and S_{ov} scores in the transparent COBEGETJ predictions are approximately identical. This phenomenon may arise because the neural network was trained to produce

Table 13. Number of Predictions Having S_{ov} within 7% of Top Score

fraction	counts	average S_{ov}	method
1.000	2 out of 2	77.7	BENNER
0.600	9 out of 15	70.4	ROST
0.500	2 out of 4	73.2	VALENCIA
0.500	8 out of 16	67.6	STERNBERG
0.500	8 out of 16	66.7	SOLOVYEV
0.500	1 out of 2	60.3	BAZAN
0.375	6 out of 16	67.4	JAAP
0.375	3 out of 8	59.4	GOLDSTEIN
0.333	2 out of 6	69.9	COHEN
0.333	2 out of 6	68.8	SERVER_PREDICTPROTEIN
0.333	2 out of 6	67.1	HUBBARD
0.250	1 out of 4	58.5	FINKELSTEIN
0.200	1 out of 5	66.1	SERVER_DSC_MULT
0.200	2 out of 10	64.9	MUNSON
0.167	1 out of 6	62.0	SERVER_NNSSP_MULT
0.140	1 out of 7	50.5	ABAGYAN
0.111	1 out of 9	60.9	MURZIN
0.000	0 out of 1	82.8	EISENBERG
0.000	0 out of 1	69.2	JONES
0.000	0 out of 2	62.8	SMITH
0.000	0 out of 2	60.7	MARSHALL
0.000	0 out of 6	53.9	SERVER_SSPRED
0.000	0 out of 5	53.3	SHESTOPALOV
0.000	0 out of 6	51.8	SERVER_GOR
0.000	0 out of 6	51.2	SERVER_SSP_MULT
0.000	0 out of 6	50.8	SERVER_NNPREDICT
0.000	0 out of 3	49.4	TAYLOR
0.000	0 out of 4	47.5	ROSE
0.000	0 out of 6	44.7	MOULT
0.000	0 out of 1	43.7	BAKER
0.000	0 out of 4	39.1	LENGAUER
0.000	0 out of 1	16.4	OSGUTHORPE
0.000	0 out of 1	15.7	AVBELJ

high Q_3 scores, while the transparent predictors are primarily concerned with getting the number, order, and types of secondary structure segments correct. It is axiomatic that a tool will generate higher scores in tests for which it is optimized.

3. Evolution-Based Assessments of the CASP2 Project

With these considerations in mind, we can offer alternative evaluations of the CASP2 project. The first several differs from the official evaluation simply by using S_{ov} scores rather than Q_3 scores. The tool credits for each target the highest S_{ov} score, together with other tools that produce an S_{ov} score within seven percentage points of the highest score for this target. The second expedient reflects the fact that the highest attainable S_{ov} score depends in part on the extent to which secondary structure has diverged within a family of homologous proteins. The results are collected in Table 13, which shows that prediction tools fell into two categories: those that produce S_{ov} scores that rank highly on occasion and those that do not.

Past this division, little more can be said about the relative merits of different methods from these scores. First, the S_{ov} score does not distinguish between core and noncore secondary structural elements. For this reason, it is possible to have a prediction with a high S_{ov} score that makes all of its mistakes in core segments that is less valuable than an alternative prediction with a lower S_{ov} score that makes its mistakes in noncore regions (see section II above). All of the strong methods provide Q_3 and S_{ov} scores approaching the maximum possible for a consensus prediction given the ambiguities in the reference structure and the fact that secondary structure diverges during divergent evolution (section II). To ascertain whether any individual prediction method scoring in this range is satisfactory for further structural modeling, or as part of a postgenomic analysis of evolution or function, one must learn whether the 25% "mistakes" are serious or not.

Further, the methods evaluated in Table 13 are tested on different sets of targets. As noted above, this can easily generate meaningless evaluations. We can, however, provide an improved evaluation based on a more limited set of target proteins, one where the leading methods all made predictions in parallel. For example, five of the strongest secondary structure prediction tools all made predictions for five targets in common: T0004, T0011, T0020, T0037, and T0042. On these five proteins, the best values of S_{ov} are (in order of decreasing S_{ov}) ROST (75.8) > SOLOVYEV (73.4) > COBEGETJ (72.6) > STERNBERG (67.5) > JAAP (66.5). From this, one draws the conclusion that when the best transparent and nontransparent methods are compared on the same set of targets, they perform equally well.

Of course, one might wish not to exclude those groups that made strong predictions generally, but for some reason omitted one of the five targets that the other methods predicted in parallel. It turned out that there was no predictor who fell in this category. VALENCIA, however, predicted three of these targets (T0004, T0011, and T0037) with an S_{ov} score of 72.3%. For these three targets, the other methods had scores as follows: ROST (72.4), SOLOVYEV (67.9), COBEGETJ (65.4), STERNBERG (71.0), and JAAP (61.6). The difference, of course, reflects a strong score by VALENCIA for T0004 and weak scores by several of the other methods for this target.

Further, results both from CASP1, CASP2, and the literature make clear that secondary structure prediction methods can now provide nearly perfect predictions excluding internal helices, active-site regions, and short surface strands, as well as an understanding of why this must generally be so. As a result, no prediction tool is likely to yield higher scores reliably. The question needing an answer at this point is whether the predictions with this level of mistake can be useful nevertheless. To answer this question, one must attempt to use the predictions in a *bona fide* prediction setting. CASP2 provided several examples where this was done.

D. Examination of Specific Predictions

As in the discussions in previous sections, we provide a set of figures that allows the reader to examine individual predictions individually. For each, an experimental secondary structure was assigned by DSSP. Segment overlap (S_{ov}) and three state residue (Q_3) scores were taken directly from the CASP2 Web site where available; otherwise they were calculated directly. Core strands in the secondary structure were assigned whenever possible using HERA plots;[322] a core strand is defined as one that

```
GAPEGAEYLRAVLRAPVYEAAQVTPLQKMEKLSSRLDNVILVKREDRQPVHSFKLRGAYA    sequence
 |    |            |   | | |              ||||
RAQKDPEFQAQFADLLKNYAGRPTALTKCQNITAGTRTTLYLKREDLLHGGAHKTNQVLG    1wsy
HHHH HHHHHHHHHHHHHH      EEE         EEEEGGG     HHHHHHHHHHH    1wsy DSSP
    HHHHHHHHH            EEE HHHHHHH  EEEEE HHH    E  HHHHHH    DSSP
    HHHHHHHHHH           EEE HHHHHHH  EEEEEE       E HHHHHHH    STRIDE
                       edge/6         core/7     non sheet      Thr deaminase
      HHHHHH       HH    HHHHHHHHHHHHHHHHHHH       EEEE   HH    STERNBERG
HHHHHHHHHHHHHHHHHHHHHH  EEEEE        EEEEEEE         HHHHHHH    ABAGYAN (2)
    HHHHHHHHHH  HHHHHHH   HHHHHHHHH   EEEEE     EEE EEHHHHHH    JAAP
  HHHHHHHHHHH  HHHHHH   HHHHHHHHH     EEEE      HHHHHHHH HHH    FINKELSTEIN
     HHHHHHHH    EEEEE    HHHHH       HHEEE         HHH  HHH    MUNSON
      HHHHHHHHH  EE        EE        EEEEEEE          HHHHHH    SOLOVYEV
     HHHHHHHHHHHHHHHH       HHHHHHH   EEEEE              HHH    ROST
------------------------------------------------------------    MURZIN
    HHHHHHHHHHHHHHHHHH       HHHHH    EEEEE         EE HHHHH    PHD (post CASP)

MMAGLTEEQKAHGVITASAGNHAQGVAFSSARLGVKALIVMPTAT---ADIKVDAVRGFG    sequence
                  || |    |  || || |  | |           |      |
QALLAKRMGKSEIIAETGAGQHGVASALASALLGLKCRIYMGAKDVERQSPNVFRMRLMG    1wsy
HHHHHGGG   EEEEE    HHHHHHHHHHHH   EEEEEE           HHHHHHH     1wsy DSSP
HHHHHHHH      EEE    HHHHHHHHHHHH   EEEEE    --- HHHHHHHHHH     DSSP
HHHH HHH      EEEE  HHHHHHHHHHHHH   EEEE     --- HHHHHHHHHH     STRIDE
             core/4                core/4     ---               Thr deaminase
HHHHHHHHHH   EEEE         HHHHHHH   EEEE     ---      HHHH      STERNBERG
HHHH      EEEEEEE    HHHHHHHHHH   EEEEEEE    ---   HHHHHH       ABAGYAN (2)
HHHHHHHH      EEE   HHHHHHHHHHHHH  EEEEEE    ---HHHHHHHHHH      JAAP
HHHHH HHHHHH EEEE        EEEE    EEEEEEEE    --- EEEE           FINKELSTEIN
HHHHHHHHH    EEE     HHHHHHHHHHHH   EEEE     ---   HHHHHH       MUNSON
HHHH       EEEEE    HHHHHHHHHHHHH  EEEEEE    ---    HHHHHH      SOLOVYEV
HHHHHHHHH    EEEE   HHHHHHHHHHHH    EEEEE    ---HHHHHHHHHH      ROST
------------------------------------------------------------    MURZIN
HHHHHHHHHH    EEEE  HHHHHHHHHHHH   EEEEEE    ---HHHHHHHHHHH     PHD (post CASP)

GEVLLHGANFDEAKAKAIELSQQQG-----FTWVP-----------PFDHPMVIAGQGTL    sequence
 ||          |                                         |  |
AEVIPVHSGSATLKDACNEALRDWSGSYETAHYMLGTAAGPHPYPTIVREFQRMIGEETK    1wsy
 EEEE       HHHHHHHHHHHHHH     EE         HHHHHHHHH HHHHHHHH    1wsy DSSP
 EEEE    HHHHHHHH HHHHHH -----  E  -----------    HHHHHHHHHH    DSSP
  EE     HHHHHHHHHHHHHHH -----  E  -----------    HHHHHHHHHH    STRIDE
edge/4                   -----edge/4----------                  Thr deaminase
EEEEEE   HHHHHHHHHHHHHH  -----     -----------     EEE          STERNBERG
 EE      HHHHHHHHHHHE    -----     -----------HHHHHH     HHH    ABAGYAN (2)
 EEEE     HHHHHHHHHHHHH  ----- EEE -----------     EEEE   HH    JAAP
 EEEEE   HHHHHHH         -----EEEE -----------   EEEEEE    H    FINKELSTEIN
 EEEEE   HHHHHHHHHHHHHH  ----- EEE -----------       EE    H    MUNSON
 EEEE     HHHHHHHHHHH    ----- EEE -----------     HHHHH  HH    SOLOVYEV
---EEE    HHHHHHHHHHHHH  ----- EEEE-----------            HH    ROST
------------------------------------------------------------    MURZIN
 EEEEEE  HHHHHHHHHHHHHH  -----  EE -----------        E   HH    PHD (post CASP)

ALELLQQDAHLDRVFVPVGGGGLAAGVAVLIKQLMPQIKVIAVEAEDSACLKAALDAGHP    sequence
|  |       | |   ||||  | |              | ||             | |
AQILDKEGRLPDAVIACVGGGSNAIGMFADFI-NDTSVGLIGVEPGGHGIETGEH--GAP    1wsy
HHHHHHH     EEEEE    HHHHHHHGGG -     EEEEEEE    GGG   --       1wsy DSSP
HHHHHHH     EEEEE    HHHHHHHHHHHHH    EEEEEEE     HHHHH         DSSP
HHHHHHH     EEEEE    HHHHHHHHHHHHH    EEEEEEE     HHHHHH        STRIDE
           core/6                     core/6                    Thr deaminase
 HHHHH      EEEEE       HHHHHHHHHH    EEEEE    HHHHHHHHH        STERNBERG
HHHHHHHHHHHEEEEEEE   HHHHHHHHHH     EEEEEEEEE                   ABAGYAN (2)
HHHHHHHHH   EEEEE   HHHHHHHHHHHHHH   EEEEEHHH  HHHHHHHHH        JAAP
HHHHHHHHH    EEEEE    HHHHHHHHHHHHHHHEEEEEE    HHHHHHH          FINKELSTEIN
HHHHHHHH    EEEEE      HHHHHHHHHHH    EEEEE H  HHHHHHHH         MUNSON
HHHHHHHH    EEEEE    HHHHHHHHHH        EEEEEEEE      HHHH EE    SOLOVYEV
HHHHHHHHH   EEEEE   HHHHHHHHHHHHHH    EEEEE    HHHHHHHHH        ROST
------------------------------------------------------------    MURZIN
HHHHHHHHH   EEEEE   HHHHHHHHHHHHHH    EEEEE    HHHHHHHHHH       PHD (post CASP)
```

```
VDLPRVGLFAEGVAVKRIG---------------------DETFRLCQEYLDDIITVDSD    sequence
    |||      |                               |      |      |
LKHGRVGIYFGMKAPMMQTADGQIEESYSISAGLDFPSVGPQHAYLNSIGRADYVSITDD    1wsy
    EEEEE  EEEEE                        HHHHHHHHHH EEEEEE HH    1wsy DSSP
                   --------------------- HHHH HHH   EEEEE HH    DSSP
                   ---------------------     HHHH   EEEEE HH    STRIDE
                   ---------------------           edge/6       Thr deaminase
       HHHHHHHHHHHH---------------------HHHHHHHHHH   EEE        STERNBERG
                   ---------------------        HHHHHHH  EEE    ABAGYAN (2)
        HHHHHHHHHHH---------------------HHHHHHHHHHH  EEEE       JAAP
  HHHHHHHHHHHHHH   --------------------- HHHHHHHHHHHHHHH   H    FINKELSTEIN
        EH    E    --------------------- HHHHHHHH    EEE  HH    MUNSON
EEE  EEEE          --------------------- HHHHHHHHH   EEEE HH    SOLOVYEV
EE    HHHH  HHHHHH ---------------------HHHHHHHHHH   EEEE       ROST
------------------------------------------------------------    MURZIN
EEE        HHHHHHHH---------------------HHHHHHHHHHH  EEEE       PHD (post CASP)

AICAAMKDLFEDVRAVAEPSGALALAGMKKYIALHNIRGERLAHILSGANVNFHGLRYVS
    | | |              |||               |   |||                sequence
EALEAFKTLCRHEGIIPALESSHALAHALKMMREQPEKEQLLVVNLSGRGDKDIFTVHDI    1wsy
HHHHHHHHHHHHH     HHHHHHHHHHHHHHHH    EEEEEEEE   HHHHHHHHHHH    1wsy DSSP
HHHHHHHHHHHH      HHHHHHHHHHHHHHHHH     EEEEE  E      HHHHHH    DSSP
HHHHHHHHHHHHH     HHHHHHHHHHHHHHHHH     EEEEEE E     HHHHHHH    STRIDE
                                       core/6  not sheet        Thr deaminase
HHHHHHHHHHH          HHHHHHHHHHHHH       EEEEE      HHHHHHHH    STERNBERG
E HHHHHHHHH     HHHHHHHHHHHHHHHH   EEEEEEE      HHHHHHHHHHHH    ABAGYAN (2)
HHHHHHHHHHHHHHH      HHHHHHHHHHHHHHHH   EEEEEEE  EEEHHHHHHHH    JAAP
HHHHHHHHHHHHHHHHH   HHHHHHHHHHHHHHH     HHHHHHH  EEEE           FINKELSTEIN
HHHHHHHHHHHHHHHHH H HHHHHHHHHHHHHHHHH    EEEEE      HHHHH HH    MUNSON
HHHHHHHHHHHHH     HHHHHHHHHHHHHHHH       EEEE         HHHHHH    SOLOVYEV
HHHHHHHHHHHHHHHHH   HHHHHHHHHHHHHHH     EEEEEE    HHHHHHHHHH    ROST
------------------------------------------------------------    MURZIN
HHHHHHHHHHHHHHH     HHHHHHHHHHHHHHHHH    EEEEEE    HHHHHHHHH    PHD (post CASP)

|start domain 2
ERCELGEQREALLAVTIPEEKGSFLKFCQLLGGRSVTEFNYRFADAKNACIFVGVRLSR     sequence
LKA                                                             1wsy
H                                                               1wsy DSSP
HHHHHH   EEEEEEEE       HHHHH     EEEEEEEE      EEEEEEEE        DSSP
HHHHHH   EEEEEEEE      HHHHHHH    EEEEEEEE      EEEEEEEE        STRIDE
        core/4                  edge/4 edge/5    core/4         Thr deaminase
HHH   HHH   EEEE        HHHHH       HHHHHHHHH    EEEEEE  HH     STERNBERG
HH        EEEEE   HHHHHHHHHHHH  EEEE    HHHHHH EEEEEE           ABAGYAN (2)
HHHHH    EEEEEEE       HHHHHHHH    EEEEEEEEHHH    EEEEEEE       JAAP
           EEEEEE      HHHHHHHH       EEEEEE    EEEEEEEEE       FINKELSTEIN
HHHHHHHHHHHHEEE        HHHHEEE      EHHHH HHH  H EEEEEEEE       MUNSON
HHH       EEEEEE       HHHHHHHH                 EEEEEEEE        SOLOVYEV
HHHHHHHH EEEEEEE      HHHHHHHHH     EEEEEEE       EEEEEEE       ROST
--         EEEEEEE    HHHHHHHHH  EEEEEEEEEE    EEEEEEEEE        MURZIN
HHHHHHHH EEEEEEE       HHHHHHHH      EEEEEEHHH    EEEEEEEE      PHD (post CASP)

                                            |start domain 3
GLEERKEILQMLNDGGYSVVDLSDDEMAKLHVRYMVGGRPSHPLQERLYSFEFPESPGA     sequence
 HHHHHHHHHHH     EE            HHH            EEEEE       H     DSSP
 HHHHHHHHHHH     EE            HHH            EEEEE       H     STRIDE
               edge/4                        core/5             Thr deaminase
HHHHHHHHHHHHHH         HHHHHHHHEEEEE        HHHHHHH             STERNBERG
  HHHHHHHHHHHHHHHHH   EEEEEEEEEE        HHHHHHHHHHHHH           ABAGYAN (2)
  HHHHHHHHHHH     EEE  HHHHHHHHHHHEE  EE    HHHHEEE      HH     JAAP
       HHHHH    EEEEEE  HHHHHHH  EEE      HHHHHHEEEEE   HHH     FINKELSTEIN
 HHHHHHHHHHHH    EEE   HHHHHHHHHEEEE        HHHEEE        H     MUNSON
    HHHHHHHHHH   EEE   HHHHHHHHHHHH            EEEE       H     SOLOVYEV
  HHHHHHHHHHHH    EEE  HHHHHHHHHHHHE           EEEEE     HH     ROST
 HHHHHHHHHHHH    EEEEEE                       _____      EE     MURZIN
  HHHHHHHHHHHH         HHHHHHHHHHHHH      HHHHHHE       HHH     PHD (post CASP)
```

```
LLRFLNTLGTYWNISLFHYRSHGTDYGRVLAAFELGDHEPDFETRLNELGYDCHDETNN     sequence
HHHHHHHH       EEEEE        EEEEE   -------------   EEE         DSSP
HHHHHHHH       EEEEE        EEEEE   -------------   EEE         STRIDE
             core/5        core/5                  edge/5       Thr deaminase
HHHHHHH     EEEEEEE        EEEEEE    ------------               STERNBERG
               _____   EEEEEEEEE  EE-------------  EEEEEEEE     ABAGYAN (2)
HHHHHHHH     EEEEEHHH   HHHHHHHHH    ------------  HHHH         JAAP
HHHHHHHHHHHH EEEEE       HHHHHEEEEE  ------------               FINKELSTEIN
HHHHHH     HHHHHH        HHHEHHH     ------------               MUNSON
HHHHHHH      EEEEEE        EEEE     -------------               SOLOVYEV
HHHHHHHH   EEEEEHHHHH   HHHHHHHEE    ------------               ROST
E     HHHHHHHHHH       EEEEE   EEE   ------------       E       MURZIN
HHHHHHHH   EEEEEEHHHH    HHHEEEEEE     HHHHHHHHH                PHD (post CASP)

PAFRFFLAG                                                       sequence
HHHHHH                                                          DSSP
HHHHHHH                                                         STRIDE
                                                                Thr deaminase
 HHHHH                                                          STERNBERG
EE  HHHHH                                                       ABAGYAN (2)
HHHHHH                                                          JAAP
  EEEEE                                                         FINKELSTEIN
HHHHEEE                                                         MUNSON
HHHHHHHH                                                        SOLOVYEV
 HHHHH                                                          ROST
 HHHHH                                                          PHD (post CASP)
```

Figure 48. Sequence and predictions from the CASP2 site and experimental secondary structure[324] for threonine deaminase, *E. coli* (514 residues), target T0002, THD1_ECOLI, P04968. Experimental secondary structural assignments, calculated with DSSP and STRIDE, were taken from the CASP2 web site. Key: E, β strand; H, α helix; G, 3_{10} helix. Alignment with tryptophan synthase (1wsy) was done using HERA plots of hydrogen bonding in such a manner as to emphasize the similarity in secondary structure motifs. The number in parentheses (*n*) indicates the prediction was a weighted average of *n* predictions. Serious mistakes and omissions are underlined. The prediction with the highest $S_{\rm ov}$–O is shown. For each prediction, $S_{\rm ov}$–O and Q_3 for the residues with no homology to tryptophan synthase are listed in order of descending $S_{\rm ov}$–O: SOLOVYEV, 78.8, 75.8; ROST, 78.0, 69.8; JAAP, 74.8, 69.2; STERNBERG, 73.8, 66.5; MUNSON, 67.9, 61.5; FINKELSTEIN, 59.4, 54.9; MURZIN, 53.7, 60.2 from coordinate model (fold recognition); ABAGYAN (2), 43.4, 43.1.

has hydrogen-bonding interactions to two other strands on both edges.

1. Threonine Deaminase (T0002)

Approximately 15 threonine deaminase homologs with PAM distances less than 150 were available when threonine deaminase was announced as a CASP2 target. Accordingly, *ab initio* evolution-based prediction tools were expected to perform well. Threonine deaminase was announced, however, as a protein that might be homologous to the β subunit of tryptophan synthase. This was based on the knowledge that the protein had a pyridoxal cofactor and the observation of a conserved Lys and a Gly-rich loop at an appropriate position (Travis Gallagher, personal communication). Several *ab initio* prediction groups therefore assumed that the target was more appropriate as a homology modeling target. Nevertheless, a number of other predictors treated this as an *ab initio* target and submitted predictions. In any case, DARWIN failed to identify significant sequence similarity between the two protein sequences, and a CLUSTALW alignment failed to correctly align secondary structural elements. A structure-based alignment yielded only ~15% sequence identity, well into the "twilight zone". It is evident that a secondary structure prediction would have been useful for predicting long distance homology in this case, but no such prediction was explicitly made as part of the CASP2 project.

Figure 48 shows the predictions made for this protein. $S_{\rm ov}$ and Q_3 scores were quite good for the strongest automated neural network and statistical contenders, including the neural network developed by Rost *et al.*,[218] the method of Solovyev and Salamov,[323] and the method of King and Sternberg.[106]

With coordinates now available (we are indebted to Dr. T. Gallagher for sending us coordinates prior to publication), we can apply a more useful scoring system that focuses on core strands that come together to form β sheets in the protein. For a core strand to be "correctly predicted" requires that a strand be assigned between flanking secondary structural elements also assigned correctly, provided that at least one amino acid overlaps in the predicted and experimental secondary structural elements. This reflects the experience with transparent predictions, where successful tertiary structural models can be built if the number and nature of the secondary structural elements are assigned correctly. Segment overlap is less important for this purpose. In the event that both helix and strand residues are predicted for residues assigned to a strand, then the prediction is counted correct if the predicted strand covers ≥50% of the experimental strand. When an edge strand is missed, and a predicted helix intrudes on the strand, it is counted as wrong, except when the helix is part of a correctly assigned adjacent helix, in which case the edge strand is counted as being

```
             1                                                50
predict_h284 MADSQPLSGAPEGAEYLRAVLRAPVYEAAQVTPLQKMEKLSSRLDNVILV
thd1_ecoli   MADSQPLSGAPEGAEYLRAVLRAPVYEAAQVTPLQKMEKLSSRLDNVILV
thd1_salty   MAESQPLSVAPEGAEYLRAVLRAPVYEAAQVTPLQKMEKLSSRLDNVILV
thd1_haein   .MKNLLTNPQPSQSDYINAilGSRVYEAAQVTPLQKMGKLSERLHNNIWI
thd1_burce   ...........ASHDYLKKILTARVYDVAFETELEPARNLSARLRNPVYL
thdh_yeast   .........TDNTPDYVRLVLRSSVYDVINESPISQGVGLSSRLNTNVIL
thd1_lyces   IVNKPTGGDSDELFQYLVDILASPVYDVAIESPLELAEKLSDRLGVNFYI
thd1_soltu   ..................................................
thd1_bacsu   ........................VKDVVIHTPLQRNDRLSERYECNIYL
thd1_myctu   .PSSSPLFSLSGADIDRAAKRIAPVVTP...TPLQPSDRLSAITGATVYL
thd2_ecoli   MHITYDLPVAIDDIIEAKQRLAGRIYK....TGMPRSNYFSERCKGEIFL
ykv8_yeast   .....................SNRIKEYVNKTPVLTSRMLNDRLGAQIYF
thd1_corgl   .........MASGAELIRatAQARISSVIAPTPLQYCPRLSEETGAEIYL
thd1_lacla   LLKAVVTKTPLQLDPYLSNKYQANIYLKEVvtPLQLDPYLSNKYQANIYL

             51                                              100
predict_h284 KREDRQPVHSFKLRGAYAMMAGLTEEQKAHGVITASAGNHAQGVAFSSAR
thd1_ecoli   KREDRQPVHSFKLRGAYAMMAGLTEEQKAHGVITASAGNHAQGVAFSSAR
thd1_salty   KREDRQPVHSFKLRGAYAMMTGLTEEQKAHGVITASAGNHAQGVAFSSAR
thd1_haein   KREDRQPVNSFKLRGAYAMISSLSAEQKAAGVIAASAGNHAQGVALSAKQ
thd1_burce   KREDNQPVFSFKLRGAYNKMAHIPADALARGVITASAGNHAQGVAFSAAR
thdh_yeast   KREDLLPVFSFKLRGAYNMIAKLDDSQRNQGVIACSAGNHAQGVAFAAKH
thd1_lyces   KREDKQRVFSFKLRGAYNMMSNLSREELDKGVITASAGNHAQGVALAGQR
thd1_soltu   ..................................................
thd1_bacsu   KREDLQVVRSFKLRGAYHKMKQLSSEQTENGVVCASAGNHAQGVAFSCKH
thd1_myctu   KREDLQTVRSYKLRGAYNLLVQLSDEELAAGVVCSSAGNHAQGFAYACRC
thd2_ecoli   KFENMQRTGSFKIRGAFNKLSSLTDAEKRKGVVACSAGNHAQGVSLSCAM
ykv8_yeast   KGENFQRVGAFKFRGAMNAVSKLSDEKRSKGVIAFSSGNHAQAIALSAKL
thd1_corgl   KREDLQDVRSYKIRGALNSGAQSPQEQRDAGIVAASAGNHAQGVAYVCKS
thd1_lacla   KEENLQKVRSFKLRGAYYSISKLSDEQRSKGVVCASAGNHAQGVAFAANQ

             101                                             150
predict_h284 LGVKALIVMPTATADIKVDAVRGFGGEVLLHGANFDEAKAKAIELSQQQG
thd1_ecoli   LGVKALIVMPTATADIKVDAVRGFGGEVLLHGANFDEAKAKAIELSQQQG
thd1_salty   LGVKSLIVMPKATADIKVDAVRGLGGEVLLHGANFDEAKAKAIELAQQQG
thd1_haein   LGLKALIVMPQNTPSIKVDAVRGFGGEVLLHGANFDEAKAKAIELSKEKN
thd1_burce   MGVKAVIVVPVTTPQVKVDAVRAHGgeVIQAGESYSDAYAHALKVQEERG
thdh_yeast   LKIPATIVMPVCTPSIKYQNVSRLGSQVVLYGNDFDEAKAECAKLAEERG
thd1_lyces   LNCVAKIVMPTTTPQIKIDAVRALGGDVVLYGKTFDEAQTHALELSEKDG
thd1_soltu   ..................................................
thd1_bacsu   LGIHGKIFMPSTTPRQKVSQVELFGkdIILTGDTFDDVYKSAAECCEAES
thd1_myctu   LGVHGRVYVPAKTPKQKRDRIRYHGGelIVGGSTYDLAAAAALEDVERTG
thd2_ecoli   LGIDGKVVMPKGAPKSKVAATCDYSAEVVLHGDNFNDTIAKVSEIVEMEG
ykv8_yeast   LNVPATIVMPEDAPALKVAATAGYGAHIIRYNRYTEDREQIGRQLAAEHG
thd1_corgl   LGVQGRIYVPVQTPKQKRDRIMVHGGelVVTGNNFDEASAAAHEDAERTG
thd1_lacla   LNISATIFMPVTTPNQKISQVKFFGetIRLIGDTFDESARAAKAFSQDND

             151                                             200
predict_h284 FTWVPPFDHPMVIAGQGTLALELLQQDAHLDRVFVPVGGGGLAAGVAVLI
thd1_ecoli   FTWVPPFDHPMVIAGQGTLALELLQQDAHLDRVFVPVGGGGLAAGVAVLI
thd1_salty   FTWVPPFDHPMVIAGQGTLALELLQQDSHLDRVFVPVGGGGLAAGVAVLI
thd1_haein   MTFIPPFDHPLVIAGQGTLAMEMLQQVADLDYVFVQVGGGGLAAGVAILL
thd1_burce   LTFVHPFDDPYVIAGQGTIAMEILRQHqpIHAIFVPIGGGGLAAGVAAYV
thdh_yeast   LTNIPPFDHPYVIAGQGTVAMEILRQvnKIGAVFVPVGGGGLIAGIGAYL
thd1_lyces   LKYIPPFDDPGVIKGQGTIGTEINRQLKDIHAVFIPVGGGGLIAGVATFF
thd1_soltu   .....PFDAPGVIKGQGTIGTEINRQLKDIHAVFVPVGGGGLISGVAAYF
thd1_bacsu   RTFIHPFDDPDVMAGQGTLAVEILNddTEPHFLFASVGGGGLLSGVGTYL
thd1_myctu   ATLVPPFDDLRTIAGQGTIAVEVLGQLEdpDLVVVPVGGGGCIAGITTYL
thd2_ecoli   RIFIPPYDDPKVIAGQGTIGLEIMEDLYDVDNVIVPIGGGGLIAGIAVAI
ykv8_yeast   FALIPPYDHPDVIAGQGTSAKELLEEVGQLDALFVPLGGGGLLSGSALAA
thd1_corgl   ATLIEPFDARNTVIGQGTVAAEILSQLtsADHVMVPVGGGGLLAGVVSYM
thd1_lacla   KPFIDPFDDENVIAGQGTVALEIFAQAksLDKIFVQIGGGGLIAGITAYS
```

```
             201                                              250
predict_h284 KQLMPQIKVIAVEAEDSACLKAALDAGHPVDLPRVGLFAEGVAVKRIGDE
thd1_ecoli   KQLMPQIKVIAVEAEDSACLKAALDAGHPVDLPRVGLFAEGVAVKRIGDE
thd1_salty   KQLMPQIKVIAVEAEDSACLKAALEAGHPVDLPRVGLFAEGVAVKRIGDE
thd1_haein   KQFMPEIKIIGVESKDSACLKAALDKGEPTDLTHIGLFADGVAVKRIGDE
thd1_burce   KAVRPEIKVIGVQAEDSCAMAQSLQAGKRVELAEVGLFADGTAVKLVGEE
thdh_yeast   KRVAPHIKIIGVETYDAATLHNSLQRNQRTPLPVVGTFADGTSVRMIGEE
thd1_lyces   KQIAPNTKIIGVEPYGAASMTLSLHEGHRVKLSNVDTFADGVAVALVGEY
thd1_soltu   TQVAPHTKIIGVEPYGAASMTLSLYEGHRVKLENVDTFADGVAVALVGEY
thd1_bacsu   KNVSPDTKVIAVEPAGAASYFESNKAGHVVTLDKIDKFVDGAAVKKIGEE
thd1_myctu   AERTTNTAVLGVEPAGAAAMMAALAAGEPVTLDHVDQFVDGAAVNRAGTL
thd2_ecoli   KSINPTIRVIGVQSENVHGMAASFHSGEITTHRTTGTLADGCDVSRPGNL
ykv8_yeast   RSLSPGCKIFGVEPEAGNDGQQSFRSGSIVHINTPKTIADGAQTQHLGEY
thd1_corgl   ADMAPRTAIVGIEPAGAASMQAALHNGGPITLETVDPFVDGAEVKRVGDL
thd1_lacla   KERYPQTEIIGVEAKGATSMKAAYSAGQPVTLEHIDKFADGIAVATVGQK

             251                                              300
predict_h284 TFRLCQEYLDDIITVDSDAICAAMKDLFEDVRAVAEPSGALALAGMKKYI
thd1_ecoli   TFRLCQEYLDDIITVDSDAICAAMKDLFEDVRAVAEPSGALALAGMKKYI
thd1_salty   TFRLCQEYLDDIITVDSDAICAAMKDLFEDVRAVAEPSGALALAGMKKYI
thd1_haein   TFRLCQQYLDDMVLVDSDEVCAAMKDLFENVRAVAEPSGALGLAGLKKYV
thd1_burce   TFRLCKEYLDGVVTVDTDALCAAIKDVFQDTRSVLEPSGALAVAGAKLYA
thdh_yeast   TFRVAQQVVDEVVLVNTDEICAAVKDIFEDTRSIVEPSGALSVAGMKKYI
thd1_lyces   TFAKCQELIDGMVLVANDGISAAIKDVYDEGRNILETSGAVAIAGAAAYC
thd1_soltu   TFAKCQELIDGMVLVRNDGISAAIKDVYDEGRNILETSGAVAIAGAAAYC
thd1_bacsu   TFRTLETVVDDILLVPEGKVCTSILELYNECAVVAEPAGALSVAALDLY.
thd1_myctu   TYAaaAGDMVSLTTVDEGAVCTAMLDLYQNEGIIAEPAGALSVAGL...L
thd2_ecoli   TYEIVRELVDDIVLVSEDEIRNSMIALIQRNKVVTEGAGALACAALLSGK
ykv8_yeast   TFAIIRENVDDILTVSDQELVKCMHFLAERMKVVVEPTACLGFAGAL..L
thd1_corgl   NYTIVEKNQghMMSATEGAVCTEMLDLYQNEGIIAEPAGALSIAGLKE..
thd1_lacla   TYQLINDKVKQLLAVDEGLISQTILELYSKLGIVAEPAGATSVAALE..L

             301                | start of domin 2           350
predict_h284 ALHNIRGERLAHILSGANVNFHGLRYVSERCELGEQREALLAVTIPEEKG
thd1_ecoli   ALHNIRGERLAHILSGANVNFHGLRYVSERCELGEQREALLAVTIPEEKG
thd1_salty   AQHNIRGERLAHVLSGANVNFHGLRYVSERCELGEQREGLLTVTIPEEKG
thd1_haein   KQNHIEGKNMAAILSGANLNFHTLRYVSERCEIGENREALLAVTMPEQPG
thd1_burce   EREGIENQTLVAVTSGANMNFDRMRFVAERAEVGEAREAVFAVTIPEERG
thdh_yeast   STvdHTKNTYVPILSGANMNFDRLRFVSERAVLGEGKEVFMLVTLPDVPG
thd1_lyces   EFYKIKNENIVAIASGANMDFSKLHKVTELAGLGSGKEALLATFMVEQQG
thd1_soltu   EFYNIKNENIVAIASGANMDFSKLHKVTELAELGSDNEALLATFMIEQPG
thd1_bacsu   .KDQIKGKNVVCVVSGGNNDIGRMQEMKERSLIFEGLQHYFIVNFPQRAG
thd1_myctu   EADIEPGSTVVCLISGGNNDVSRYGEVLERSLVHLGLKHYFLVDFPQEPG
thd2_ecoli   LDQYIQNRKTVSIISGGNIDLSRVSQIT......................
ykv8_yeast   KKEELVGKKVGIILSGGNVDMKRYATLISGKEDGP...............
thd1_corgl   .MSFAPGSVVVCIISGGNNDVLRYAEIAERSLVHRGLKHYFLVNFPQKPG
thd1_lacla   IKDEIKGKNIVCIISGGNNDISRMQEIEERALVYEGLKHYFVINFPQRPG

             351                                              400
predict_h284 SFLKFCQLLGGRSVTEFNYRFADAKNACIFVGVRLSRGLEERKEILQMLN
thd1_ecoli   SFLKFCQLLGGRSVTEFNYRFADAKNACIFVGVRLSRGLEERKEILQMLN
thd1_salty   NFPKFCQLLGGRMVTEFNYRFADAKNACIFVGVRVSQGLEERKEIITQLC
thd1_haein   SFLKFAYVLGNRAVTEFSYRYADDKRACVFVGVRTTNE.QEKADIIADLT
thd1_burce   SFKRFCSLVGDRNVTEFNYRIADAQSAHIFVGVQIRRR.GESADIAANFE
thdh_yeast   AFKKMQKIIHPRSVTEFSYRYNEHRhaYIYTSFSVVDREKEIKQVMQQLN
thd1_lyces   SFKTFVGLVGSLNFTELTYRfsERKNALILYRVNVDKE.SDLEKMIEDMK
thd1_soltu   SFKTFAKLVGSMNITEVTYRFTSERKEALVLYRVDVDEKSDLEEMIKKLN
thd1_bacsu   ALREFLDEVLGpdITRFEYTKKNNKSNGPALVGIELQNKADYGPLIERMN
thd1_myctu   ALRRFLDDVLGpdITLFEY...............VKRNNRETGEALVGIE
thd2_ecoli   ..................................................
ykv8_yeast   ..................................................
thd1_corgl   QLRHFLEDILGpdITLFEY...............LKRNNRETGTALVGIH
thd1_lacla   SLRTFVSDILGpdITRFEYIKRADKGKGPCLVGILLSDASDYDSLINRIE
```

```
                401               |start of domin 3            450
predict_h284    DGGYSVVDLSDDEMAKLHVRYMVGGRPSHPLQERLYSFEFPESPGALLRF
thd1_ecoli      DGGYSVVDLSDDEMAKLHVRYMVGGRPSHPLQERLYSFEFPESPGALLRF
thd1_salty      DGGYSVVDLSDDEMAKLHVRYMVGGRPSKPLQERLYSFEFPESPGALLKF
thd1_haein      KNGFDVEDMSDDDIAKTHVRYLMGGRAAND.NERLYTFEFPEQKGALLKF
thd1_burce      SHGFKTADLTHDELSKEHIRYMVGGRSPLALDERLFRFEFPERPGALMKF
thdh_yeast      ALGFEAVDISDNELAKSHGRYLVGGASKVP.NERIISFEFPERPGALTRF
thd1_lyces      SSNMTTLNLSHNELVVDHLKHLVGGSANIS.DEIFGEFIVPEKAETLKTF
thd1_soltu      SSNMKTFNFSHNELVAEHIKHLVGGSASIS.DEIFGEFIFPEKAGTLSTF
thd1_bacsu      KKPFHYVEVNKDE.....................................
thd1_myctu      LGSAADLDGLLARMRaiHVEALEPGSPAY.....................
thd2_ecoli      ..................................................
ykv8_yeast      ..................................................
thd1_corgl      LSEASGLDSLLERMEeiDSRRLEPGTPEYEYLT.................
thd1_lacla      RFDNRYVNLrnDSLYELLV...............................

                451                                            500
predict_h284    LNTLGTYWNISLFHYRSHGTDYGRVLAAFELGDHEPDFETRLNELGYDCH
thd1_ecoli      LNTLGTYWNISLFHYRSHGTDYGRVLAAFELGDHEPDFETRLNELGYDCH
thd1_salty      LHTLGTHWNISLFHYRSHGTDYGRVLAAFELGDHEPDFETRLHELGYECH
thd1_haein      LETLQNRWNISLFHYRAHGADYGNILAGFQIEQReaEFEQGLAQLNYVFE
thd1_burce      LSSMAPDWNISLFHYRNQGADYSSILVGLQVPQAdaEFERFLAALGYPYV
thdh_yeast      LGGLSDSWNLTLFHYRNHGADIGKVLAGISVPPRelTFQKFLEDLGYTYH
thd1_lyces      LDAFSPRWNITLCRYRNQGDINASLLMGFQVPQAedEFKNQADKLGYPYE
thd1_soltu      LEAFSPRWNITLCRYRDQGDINGNVLVGFQVPQSedEFKSQADGLGYPYE
thd1_bacsu      ..................................................
thd1_myctu      ..................................................
thd2_ecoli      ..................................................
ykv8_yeast      ..................................................
thd1_corgl      ..................................................
thd1_lacla      ..................................................

                501        514
predict_h284    DETNNPAFRFFLAG
thd1_ecoli      DETNNPAFRFFLAG
thd1_salty      DESNNPAFRFFLAG
thd1_haein      DVTKSKSYRYFL..
thd1_burce      EESANPAYRLFLS.
thdh_yeast      DETDNTVYQKFL..
thd1_lyces      LDNYNEAFNLVVS.
thd1_soltu      LDNSNEAFNIVVA.
thd1_bacsu      ..............
thd1_myctu      ..............
thd2_ecoli      ..............
ykv8_yeast      ..............
thd1_corgl      ..............
thd1_lacla      ..............
```

Figure 49. Multiple sequence alignment for the threonine deaminase family from the PHD server.[208] Sequences are as follows: thd1_ecoli (P04968), threonine deaminase; thd1_salty (P20506), threonine deaminase; thd1_haein (P46493), threonine deaminase; thd1_burce (P53607), threonine deaminase; thdh_yeast (P00927), threonine dehydratase PRE; thd1_lyces (P25306), threonine deaminase; thd1_soltu (P31212), FRAGMENT; thd1_bacsu (P37946), threonine deaminase; thd1_myctu (Q10766), threonine deaminase; thd2_ecoli (P05792), threonine dehydratase CAT; ykv8_yeast (P36007), hypothetical 34.9 KD prot; thd1_corgl (Q04513), threonine deaminase; thd1_lacla (Q02145), threonine deaminase.

"missed". We recommend that CASP3 use this scoring system for proteins that have β sheets, as it provides an accurate view of the value of the secondary structure model as the starting point for assembling a tertiary structural model.

It is worth looking closely at both the multiple alignment and the structure itself to understand the challenges presented to the evaluator attempting to devise an automated tool for scoring the relative merits of prediction methods. In the structure actually determined, the threonine deaminase fold is constituted into three domains. The first domain includes residues 1−315, and is clearly independent as a folding unit. The second and third include residues 316−418 and 419−493 respectively, with a contact made between the two domains when residues 365−367 form an edge strand of the sheet that forms the core of the third domain.

The domains in threonine deaminase are not only domains in the structural sense. They are also evolutionary modules, able to disassociate and wander freely during divergent evolution. In Figure 49, sequences thd2_ecoli and ykv8_yeast have only the first domain, and are missing the second and third.

```
         core      core      edge                     core
AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKV    sequence
       EEEEEEEEEE  EEEEE       E             HHHH     EEEEEE    experimental
                     FIEVEGGDDVFVHFTAIegdg----yksLEEGQEVSFEI    1CSPD_BACSU
     mlegkvkwfnsekgfgfievegqddvfvhfsaiqgeg----fktkeegvavsfei    1csp (PDB)
      EEEEEEEEE    EEEEE      EEEE                    EEEEEE    1csp (PDB)
        core        core      edge                     core     1csp (hera)
 EEE   HHHHHHHHHH   EEEEE   HHHHHHHHHHH         EEEE   EEEEE    COHEN
  EEEEEEEEEEEEEE  EEEEEEE    EEEEEEEE        HHHHEEE  EEEEEE    ROST
  EEEEEEEEEEEEEE  EEEEEE    EEEEEEEE HHH        EEEE  EEEEEE    PHD (resubmit)
       EEEEEEEE     EEEE       EEEEE                   EEEEE    STERNBERG
  EEE EEEEEEEEEEE EEEEEE       EEEEHHHH  EE    EEEE   EEEEEE    JAAP
               EEE  EEEEE      EEEE                    EEEEE    FINKELSTEIN (2)
HHE    EEE EEEEE    EEEEE      EEEHHHH        H EE     EEEEE    MUNSON (5)
EEEEEEEE   EEEEEE   EEEEE      EEEEEHHHHHHHHHHHHHHH      EEE    ROSE
      EEE     EE     EEEE    HHHH     HHHHHHHH           EEE    SOLOVYEV (2)
        EEEEEEEEE    EEEEE   EEEEE      HHHH          EEEEEE    MURZIN
  EEEEEEEEEEEEEEE  EEEEEEE   EEEEEEEEE          EEE   EEEEEE    VALENCIA
----------   EEE   HHHH           E   E   E  EEE     EEE        ABAGYAN (2)
   EEEEEEEEEE    EEEEEEE      EEEE     HHHH           EEEEE     MOULT
```

```
         core
LEVDRQGRIRLSIKEA                                                sequence
EE       EEEE                                                   experimental
VEGNR                                                           1CSPD_BACSU
vegnrgpqaanvtkea                                                1csp (PDB)
EEE  EEEEEEEEE                                                  1csp (PDB)
         edge                                                   1csp (hera)
EEE    EEEEEEE                                                  COHEN
EEE    EEEEEEE                                                  ROST
EEE    EEEEEEE                                                  PHD (resubmit)
EE      EEEEE                                                   STERNBERG
EEE    EEEEEEE                                                  JAAP
EEEE   EEEEEE                                                   FINKELSTEIN
E      HEEH    H                                                MUNSON (5)
EEE    EEEEEE                                                   ROSE
EE       HHHHHHH                                                SOLOVYEV (2)
E       EEE                                                     MURZIN
EEE    EEEEEEE                                                  VALENCIA
EEE        HHHH                                                 ABAGYAN
         E                                                      MOULT
```

Figure 50. Sequence and predictions from the CASP2 site and experimental secondary structure[329] for polyribonucleotide nucleotidyltransferase, S1 motif, *E. coli* (84 residues), target T0004, 1sro PO5055, PNP_ECOLI. Experimental secondary structural assignments, calculated with DSSP, were taken from the CASP2 web site. Key: E, β strand; H, α helix. The number in parentheses (n) indicates the prediction was a weighted average of n predictions. The prediction with the highest S_{ov}–O is shown. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: ROST, 84.5, 71.1; STERNBERG, 82.5, 82.9; VALENCIA, 78.5, 68.9; MURZIN, from coordinate data, 67.1, 72.4; FINKELSTEIN (2), 66.7, 66.4; MUNSON (5), 62.9, 60.0; COHEN, 61.4, 49.3; JAAP, 60.6, 68.4; MOULT, 60.3, 64.5; SOLOVYEV, 57.0, 55.3; ROSE, 55.7, 54.8; ABAGYAN (2), 39.0, 56.1.

The proteins thd1_myctu and thd1_corg1 have the first two domains but are missing the third. In these two proteins, residues 370−384 in the second domain are deleted; these are the ones that make contact to the third domain, and represent an interesting (if single) case of compensatory covariation. The regulatory issues related to this are beyond the scope of the discussion. For the purposes of predicting structure, however, it should noted that predictions in the first domain are made from 14 sequences with wide evolutionary divergence, the second domain from 12 sequences, and the final domain from 8 sequences. Any method that exploits evolutionary divergence should do better in the first domain than the second, and on the second domain than the third.

Figure 48 shows that this is the case. The first domain contains seven core strands. With seven predictors making assignments, 49 segment assignments were made in all. The seven core strands were identified correctly in every one of these, except one, where a core strand was misassigned as a helix. In the third domain, however, with eight predictors and three core strands in this domain, seven of the assignments seriously mistake a core strand as a helix; two more missed. As discussed in detail above, the quality of an evolutionary model is expected to be based strongly on the nature of the input, the number of homologous sequences, their overall evolutionary divergence, and the quality of the multiple

Pos	jp	g m	nk lo	d cb**a** e f i h	SIA Predict		Sec Expt	Struct Predict
start of target sequence								
618	AN	R H	EQ AK	M AA**A** Q E N P		s		
619	QR	T S	NN KR	S EE**E** S K D S		s		e
620	LL	H H	LL YY	I VI**I** L V L P		i		e
621	GE	A P	QE PP	E EE**E** E K Q V	e	s		e
622	IV	I A	EE VE	V AV**V** V P P L	b	s		
623	GG	G G	GG GG	G GG**G** G G G H	b	s		
624	SE	Q T	MQ KT	S VR**R** S D M K	e	S		
625	VV	I E	EV KK	K II**V** V V I V	b	s	E	h
626	VV	V V	VV IL	L YY**Y** L L L Y	b	I	E	h
627	TV	P E	KE ST	Q KA**T** D E E E	e	s	E	h
628	GG	G G	GG GG	G GG**G** G G G G	b	.	E	h
629	TA	K E	IV TR	K KK**K** K T A K	e	s	E	H
630	VV	V V	VV VV	I VV**V** V V V V	b	I	E	H
631	QR	T K	KK TT	T TT**T** Q Q T R	e	is	E	H
632	SG	K N	NN NN	G RR**R** R R N N	e	S	E	H
633	LI	L K	LI IL	I LI**I** L L V I	b	I	E	H
634	KK	V T	TT TT	T AV**V** T V T T	e	I	E	H
635	PP	P E	DD DD	N DD**D** D S N T	e	S		
636	YY	F F	YY YY	F FF**F** F F F F		I		
637	GG	G G	GG GG	G GG**G** G G G G		.	E	
638	AA	A L	AA AC	A AA**A** A A A C		I	E	E
639	FF	F F	FF FF	F FF**F** F F F F		I	E	E
640	II	V I	VV VV	V VV**V** V V V V		I	E	E
641	DD	R G	DD EE	E AA**A** D E D Q		S	E	E
642	II	V L	LL LI	L II**I** I I I I		I	E	E
643	__	E D	__ EE	P VG**G** _ L G F		s		
644	GG	E G	GG PE	G GG**G** G P V G		S		
645	__	_ _	__ __	_ ___ _ _ _ T		a		
646	__	_ _	__ __	_ ___ _ _ _ R		a		
647	__	_ _	__ __	_ ___ _ _ _ M		.		
648	__	_ _	__ __	_ ___ _ _ _ K		a		
649	GG	G D	GG GG	G NG**G** G G H N		S		
650	IV	I V	VI IV	S KK**K** I V Q C		I		H
651	NS	E D	DD EE	T EE**E** D E D D		S		H
652	GG	G G	GG GG	G GG**G** G G G G		.		H
653	LL	L M	LL LL	L LL**L** L L L L		I		H
654	LL	V V	LL IV	V VV**V** V V V V		I		H
655	HH	H H	HH HH	H HH**H** H H H H		A		H
656	VI	I L	IV IV	I II**I** I I I I		I		H
657	SS	S S	TT SS	S SS**S** S S S S		i		H
658	QE	E D	DD EE	E QQ**Q** Q Q S E		S		H
659	II	L L	MM MM	V II**I** L I L M		I		H
660	SS	A D	AA SD	A AA**A** S S S S		is		h
661	__	_ W	WW WW	_ ___ _ _ _ _		I		
662	__	_ N	__ TT	_ ___ _ _ _ _		i		
663	__	_ R	__ __	_ ___ _ _ _ _		S		
664	__	_ P	__ __	_ ___ _ _ _ _		.		
665	__	_ G	__ __	_ ___ _ _ _ _		.		
666	HH	E E	__ __	D ED**D** H N N D				
667	DD	R Q	KR KN	N EK**K** S K K Q		S		extended
668	RH	H V	RR KK	Y RR**R** H H F R		S		
669	VI	V I	VV NN	V VV**V** V I V T		I		loop
670	SE	E E	KK VI	K EE**E** E G E L		S		
671	DT	V E	HH HH	D KK**K** K T D D		s	H	
672	IP	P F	PP PP	I VV**V** P P P P		i	H	
673	AH	D N	SS GS	N SA**T** S H H H		s	H	
674	TS	Q K	EE KK	D DD**D** D E T D		S	H	
675	VV	V G	II IV	H YY**Y** V V V V		I	H	e alignme
676	LF	V D	VQ LV	L LL**L** V L V V		I	H	e adjuste
677	QN	A V	NN SN	K QQ**Q** E E K R		S		e protein

678	PV	V V	VI TV	V VV**M** E E A Q	i		e
679	GN	G _	GG SG	G GG**G** G G G G	s		
680	DD	D _	DQ QD	D QQ**Q** Q Q D Q	S	E	
681	TE	D _	EQ EV	Q EE**E** E T I H	S	E	
682	LV	A _	IV VV	V VT**V** V V V I	I	E	e
683	KK	M R	TK DE	E NS**P** K K K F	S	E	e
684	VV	V A	VV VV	V VV**V** V V V V	I	E	E
685	MM	K V	KQ VM	K KK**K** K K K E	is	E	E
686	II	V V	VI VV	V VV**V** V V V V	I	E	E
687	LI	I L	LI LL	I VL**L** L L L I	I	E	E
688	SD	D D	KR ED	N EE**E** S D E K	S	E	E
689	HL	I V	FI VI	V II**V** V V V I	I		E
690	DD	D D	DN DD	E DD**D** D N D _	S		
691	RA	L V	RQ PE	K RR**R** R E L Q	S		
692	EE	E D	EE TE	D QQ**Q** D N Q N	S		
693	RR	R K	RT KR	G GG**G** N E R N	S		
694	GG	R E	TH RR	_ ___ E E K G	s		
695	RR	R R	RR RR	K RR**R** R R R K	s		e
696	VI	I I	VI II	I IV**I** I I I I	I		E
697	SS	S S	SS SS	G RR**R** S S A S	s	E	E
698	LL	L L	LL LL	L LL**L** L L L L	I	E	E
699	SS	S G	GG GG	S TS**S** S S T S	s.	E	E
700	TT	L I	LM LL	I MI**I** I M M M	I	E	E
701	KK	K K	KK KK	K KK**K** K R R K	s	E	e
702	KQ	A Q	QQ QQ	K DE**E** D E L N	s		
703	LL	D L	LL TC	A LA**A** T L D I	s		
704	EE	Q G	GE LK	K AT**T** L E E D	s		
	PP	R R	ES EA	D PA**E** P E Q Q	s		
				Q			
				S			
				Q			
				P			
				A			
				A			
				|			

Figure 51. Residue-by-residue secondary structure prediction for polyribonucleotide nucleotidyltransferase S1 motif. The SIA Predict records assignments to the surface (S, s, e), interior (I, i, b), or the "active site" (A, a). Automated assignments from DARWIN are given. Where manual assignments differ, these are indicated to right of the automated assignments. Services of DARWIN are available by server on the Web (URL http://cbrg.inf.ethz.ch/). Where the multiple alignment is adjusted, and at the ends, the surface/interior assignments may no longer correspond precisely to the output generated by the server. Residues participating in parsing strings are underlined. Secondary structure is indicated by E (strong strand assignment), e (weak strand assignment), H (strong helix assignment), and h (weak helix assignment). Sequences, designated using single letters, are from the SwissProt database, as summarized below; sequence "a" is the target sequence: (a) (P05055) Pnp_ecoli polyribonucleotide nucleotidyltransferase (EC 2.7.7.8) (polynucleotide phosphorylase). *Escherichia coli.* Seq# 617−693 = Ali# 618−704 = Target# 1−77. (b) (P41121) Pnp_pholu polyribonucleotide nucleotidyltransferase (EC 2.7.7.8) (polynucleotide phosphorylase) (Cap87k). *Photorhabdus luminescens* (*Xenorhabdus luminescens*). Seq# 617−693 = Ali# 618−704. (c) (P44584) Pnp_haein polyribonucleotide nucleotidyltransferase (EC 2.7.7.8) (polynucleotide phosphorylase). *Haemophilus influenzae.* Seq# 616−692 = Ali# 618−704. (d) (P37560) Yabr_bacsu hypothetical 14.2 kD protein in Divic−Spoiie intergenic region. *Bacillus subtilis.* Seq# 1−77 = Ali# 618−704 (hypothetical protein). (e) (P38494) Rs1h_bacsu 30S ribosomal protein S1 homolog. *Bacillus subtilis.* Seq# 183−259 = Ali# 618−704 (2 repeats are described in SwissProt, both match). (f) (P38494) Rs1h_bacsu 30S ribosomal protein S1 homolog. *Bacillus subtilis.* Seq# 268−345 = Ali# 618−704 (see above). (g) (P46836) Rs1_mycle 30S ribosomal protein S1. *Mycobacterium leprae.* Seq# 289−366 = Ali# 618−704 (best of an unknown number of repeats, SwissProt information is missing). (h) (P24384) Pr22_yeast pre-mRNA splicing factor RNA helicase Prp22. *Saccharomyces cerevisiae* (bakers' yeast). Seq# 173−253 = Ali# 618−704. (i) (P46837) Yhgf_ecoli hypothetical 81.4 kD protein in Greb−Feoa intergenic region. *Escherichia coli.* Seq# 613−690 = Ali# 618−704 (hypothetical protein, conceptual translation) (best of an unknown number of repeats, SwissProt information is missing). (j) (P29344) Rr1_spiol 30S ribosomal protein S1, chloroplast precursor (Cs1). *Spinacia oleracea* (Spinach). Seq# 256−332 = Ali# 618−704 (only match (3rd) of 3 repeats as described in SwissProt). (k) (P14129) Rs1_rhime 30S ribosomal protein S1. *Rhizobium meliloti.* Seq# 193−269 = Ali# 618−704 (4 repeats are described in SwissProt, 1−3 match). (l) (P14129) Rs1_rhime 30S ribosomal protein S1. *Rhizobium meliloti.* Seq# 278−356 = Ali# 618−704 (see above). (m) (P14129) Rs1_rhime 30S ribosomal protein S1. *Rhizobium meliloti.* Seq# 365−443 = Ali# 618−704 (see above). (n) (P02349) Rs1_ecoli 30S ribosomal protein S1. *Escherichia coli.* Seq# 187−263 = Ali# 618−704 (4 repeats are described in SwissProt, 1−2 match). (o) (P02349) Rs1_ecoli 30S ribosomal protein S1. *Escherichia coli.* Seq# 272−350 = Ali# 618−704 (see above). (p) (P46228) Rs1_synp6 30S ribosomal protein S1. *Synechococcus* sp. (strain Pcc 6301). Seq# 191−257 = Ali# 618−704 (only match (3rd) of 3 repeats as described in SwissProt).

alignment. Threonine deaminase illustrates this point within a single prediction target.

One can, of course, calculate an aggregate score for the entire threonine deaminase protein (the CASP2 scores listed in the figure captions). One might set about refining a neural network in an attempt to improve the aggregate. To do so would misunderstand the underlying problem: the reliability of evolution-based methods for predicting conformation of protein depends on the diversity of input. For a score to be informative about the underlying quality of a prediction method applied to threonine deaminase, three scores must be delivered, one for each domain.

2. Polyribonucleotide Nucleotidyltransferase S1 Motif (T0004)

Polyribonucleotide nucleotidyltransferase enhances translation initiation in gram negative bacteria such as *Escherichia coli*. It interacts both with the ribosome and the mRNA. A polypeptide segment ~100 amino acids long is repeated in the polypeptide chain, with the C-terminal segment containing the RNA-binding capacity.[325] The N-terminal region binds to the ribosome.[326] A single copy of the motif is found in other RNA-binding proteins,[327] and the evolution of ribosomal protein S1 and its homologs has been thoroughly analyzed.[328]

Figure 50 collects predictions made within the CASP2 project for the S1 motif of polyribonucleotide nucleotidyltransferase. Over a dozen rather divergent homologous sequences were available for this family (including repeats within a single entry). These have diverged substantially. Accordingly, evolution-based predictions are expected to be good. Figure 50 confirms these expectations.

Within the CASP2 project, Inna Dubchak suggested that the target might have a homolog of known conformation in the crystallographic database, 1csp, the cold shock protein CSP from *Bacillus subtilis*. This was the top fold recognition for T0004 (S1 motif). A BLAST search identified two fragments of the protein (score 35 each) when probed with the target sequence. The sequences of the proteins and the experimentally recorded secondary structure are included in Figure 50. It is clear that the significance of the similarity between the two proteins was insufficient to be more than suggestive of homology, and many (evidently) nonhomologous proteins gave higher BLAST scores. Nevertheless, the secondary structure of the two fragments of 1csp, the PDB entry for the structure of the presumed homolog, was correctly aligned, and the overall fold was quite similar. Thus, T0004 should be viewed as a success for threading methods.

This short fragment was also the target of an *ab initio* prediction using the energy minimization method of Srinivasan and Rose.[129] The secondary structure assignment was not bad, although the overall fold did not resemble the experimental structure closely. The team of Olmea, Pazos, and Valencia also predicted residue−residue contacts in this protein, and the official evaluation for the CASP2 *ab initio* project designated this tool as the most successful for this purpose.[174]

Since the protein is small, we can easily examine the prediction closely to gain insight into evolution-based structure methods. Figure 50 shows the multiple alignment and evolutionary analysis for the protein, as well as the experimentally derived secondary structure for a single protein. With only a single experimental structure, we must guess which elements belong in a consensus model. For example, the experimental structure assigned a four residue helix (Figure 50). Helices so short are rarely conserved, and only rarely an appropriate part of a consensus model. The helix is not conserved in the cold shock protein. Among the high-scoring predictions, only the ROST prediction identified it, although with a low probability. To test the stability of the ROST assignment, the same sequences were submitted to the PHD server six months after the conclusion of the CASP2 project; the PHD server failed to identify the helix (Figure 50, “resubmit”). Thus, the helix should not be part of a consensus model. Nevertheless, it had an impact on the score. The ROST prediction gained five percentage points in its Q_3 score based on its prediction of this segment.

The experimental secondary structure also has a short strand, containing a single residue. When the coordinates were resubmitted to DSSP to generate HERA plots,[322] this strand was not found. In the cold shock protein, however, an edge strand four residues long is found at the corresponding position. Further, the structure for T0004 places an edge strand antiparallel to the previous strand in this region. Thus, if this strand is missed, it will be more difficult to recognize the parallel relationship between the strands preceding it and following it. This implies that a consensus model should contain a strand.

A transparent prediction was made by the COBEGETJ team (listed as COHEN in Figure 50) was made for the S1 motif. The transparency provides clues to why two serious mistakes were made. Each misassigned a strand as a helix. For the first helix, the DARWIN tool identified surface and interior residues in the sequence Is?sI(i/s)SII (Figure 51, positions 626−633, I = strong interior, i = weak interior, S = strong surface; s = weak surface). Placing the residues marked as “?” and “i/s” on the surface yields a region with 3.6 residue periodicity, indicative of a short helix. PHD made different surface and interior assignments for the first part of this segment, designating these as “bebebe” (where “b” means <9% exposed, while e means >36% exposed). Instead of 3.6 residue periodicity indicative of a helix, these assignments give an alternating periodicity indicative of a strand. Thus, the differences in the surface/interior assignments account for the different secondary structure predictions made by the two methods.

Why are the accessibility predictions different for two critical positions, 628 and 631? At position 628, a Gly is conserved in all proteins. The PHD server assigns this pattern as indicative of an interior position. Empirically, a conserved Gly is known not always to be “interior”, but the interior assignment here gives a correct secondary structure prediction. Further, the ROST prediction is based on an alignment containing 21 sequences (Figure 52); the COHEN prediction is based on an alignment that included only 16 sequences.

```
               1         .         .         .         .        50
predict_h284   AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYL
pnp_ecoli      AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYL
pnp_pholu      AEIEVGRIYAGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVADYL
pnp_haein      AEVEAGVIYKGKVTRLADFGAFVAIVGNKEGLVHISQIAEERVEKVSDYL
rs1h_bacsu     QSLEVGSVLDGKVQRLTDFGAFVDI.GGIDGLVHISQLSHSHVEKPSDVV
pnp_bacsu      .EVEVGQLYLGKVKRIEKFGAFVEIFSGKDGLVHISELALERVGKVEDVV
yabr_bacsu     MSIEVGSKLQGKITGITNFGAFVELPGGSTGLVHISEVADNYVKDINDHL
rs1_human      DQIAAGSVLEGTVKRVKDFGAFVEILPGIEGLVHVSQISNKRIENPSEVL
rs1_rhime      AKYPVGKKISGTVTNITDYGAFVELEPGIEGLIHISEMstKKNVHPGKIL
rs1_mycle      .THAIGQIVPGKVTKLVPFGAFVRVEEGIEGLVHISELAERHVEVPDQVV
rr1_spiol      AQLGIGSVVTGTVQSLKPYGAFIDI.GGINGLLHVSQISHDRVSDIATVL
yhgf_ecoli     NDLQPGMILEGAVTNVTNFGAFVDIGVHQDGLVHISSLSNKFVEDPHTVV
pr22_yeast     ....LHKVYEGKVRNITTFGCFVQIFGTrdGLVHISEMSDQRTLDPHDVV
rs1_synp6      NRLEVGEVVVGAVRGIKPYGAFIDI.GGVSGLLHISEISHDHIETPHSVF
rs1_prosp      ENLQEGMEVKGIVKNLTDYGAFVDL.GGVDGLLHITDMAWKRVKHPSEIV
rs1_ecoli      ENLQEGMEVKGIVKNLTDYGAFVDL.GGVDGLLHITDMAWKRVKHPSEIV
rr1_porpu      SNLIVGNIIEGVINQITPYGLFIK.AGNLKGLVHISEINVKQVERIPSQF
rpoe_sulac     ....IHEVIEGEVSQVDNYGVYVNM.GPVDGLVHISQITDDNleKSKKSI
rs1_chltr      SEVQPGAILKGTVVDISKDFVVVDVGLKSEGVIPMSEFIDS.....SEGL
rne_ecoli      HEQKKANIYKGKITRIEpeAAFVDYGAERHGFLPLKEIAREYFpnIKDVL
rne_haein      HEQKKANIYKGKITRVEpeAAFVDYGAERHGFLPLKEIAREYFpnIRDIL

               51        .         .         .   84
predict_h284   QMGQEVPVKVLEVDRQGRIRLSIKEATEQSQPAA
pnp_ecoli      QMGQEVPVKVLEVDRQGRIRLSIKEATEQSQPAA
pnp_pholu      QVGQETSVKVLEIDRQGRVRLSIKEATAGTAVEE
pnp_haein      QVGQEVNVKVVEIDRQGRIRLTMKDLAPKQETE.
rs1h_bacsu     EEGQEVKVKVLSVDRdeRISLSIKDTLP......
pnp_bacsu      KIGDEILVKVTEIDKQGRVNLSRKAVLREEKEKE
yabr_bacsu     KVGDQVEVKVINVEKDGKIGLSIKKAKDRPQARP
rs1_human      KSGDKVQVKVLDIKpeERISLSMKALEEKPERE.
rs1_rhime      STSQEVDVVVLEVDpkRRISLGLKQTLENPWQA.
rs1_mycle      AVGDDAMVKVIDIDLerRISLSLKA.........
rr1_spiol      QPGDTLKVMILSHDRegRVSLSTKKLEP......
yhgf_ecoli     KAGDIVKVKVLEVDLqkRIALTMRLDEQPGETNA
pr22_yeast     RQGQHIFVEVIKIQNNGKISLSMKNIDQHS....
rs1_synp6      NVNDEVKVMIIDLDAegRISLSTKQLEPE.....
rs1_prosp      NVGDEITVKVLKFDRetRVSLGLKQLGEDPWVA.
rs1_ecoli      NVGDEITVKVLKFDRetRVSLGLKQLGEDPWVA.
rr1_porpu      KIGDTIKAVIIHVDkqGRLSLSMK..........
rpoe_sulac     TKGDRVRAMIIssGRLPRIALTMKQP........
rs1_chltr      SVGAEVEVYLDqeDEEGKVVLSREKATRQRQ...
rne_ecoli      REGQEVIVQIDKEERGNK................
rne_haein      VEGQEVIVQVNKEERGNK................
```

Figure 52. Multiple sequence alignment from the PHD server[208] for polyribonucleotide nucleotidyl transferase S1 motif. Organisms are pnp_ecoli (P05055), phosphorylase (PNPASE); pnp_pholu (P41121), phosphorylase (PNPASE); pnp_haein (P44584), phosphorylase (PNPASE); rs1h_bacsu (P38494), 30S ribosomal protein S1; pnp_bacsu (P50849), phosphorylase (PNPASE); yabr_bacsu (P37560), hypothetical 14.2 kD protein; rs1_human (P50889), 40S ribosomal protein S1; rs1_rhime (P14129), 30S ribosomal protein S1; rs1_mycle (P46836), 30S ribosomal protein S1; rr1_spiol (P29344), 30S ribosomal protein S1; yhgf_ecoli (P46837), hypothetical 81.4 kD protein; pr22_yeast (P24384), pre-MRNA splicing factor; rs1_synp6 (P46228), 30S ribosomal protein S1; rs1_prosp (P14128), 30S ribosomal protein S1; rs1_ecoli (P02349), 30S ribosomal protein S1; rr1_porpu (P51345), chloroplast 30S ribosomal; rpoe_sulac (P39466), DNA-directed RNA polymerase; rs1_chltr (P38016), 30S ribosomal protein S1; rne_ecoli (P21513), ribonuclease e; and rne_haein (P44443), ribonuclease E.

The second helix mispredicted by COHEN is discussed at length in a manuscript submitted to *Proteins* as a prediction report (D L. Gerloff, F. E. Cohen, and S. A. Benner, unpublished) prior to the CASP2 project. The manuscript was unpublished on the advice of a referee, who objected to the publication of a prediction for a CASP2 target. The misprediction lies in a region of high conservation of the protein sequence. The conservation extends to the cold shock proteins. This is a region diverging under unusual functional constraints, the “active site” of the protein. Gerloff, Cohen, and Benner recognized this problem and suggested that this was either an internal helix or an active-site segment with unpredictable secondary structure. In fact, the segment is an edge strand involved in binding to RNA. As discussed above, secondary structure prediction in regions of the active site is necessarily difficult by any method, as selection of amino acids is determined in this region by factors other than propensities to create particular secondary structures.

```
 edge                core         core        core
QIHDITGKDCQDIANKGAKQSGLYFIKPLKANQQFLVYCEIDGSGNGWTVFQKRLDGSVD    sequence
 EEEEEE  HHHHHH      EEEEE        EEEEEEE     EEEEEEEE          experimental
EEE      HHHHHHH     EEEEEEE      EEEEEEE      EEEEEE      H    JAAP
                     EEEEEEE   HHHHHHHHHHH    EEEEEEE    HHH    BENNER
                     EEEE         EEEEEE       EEEEEE           STERNBERG
 HHHHH  EEEEEEE      EEEEEEEE       EEEEE EEE  EEEEE EEEEEE     ABAGYAN
EEE      HHHHHH       EEEEE       EEEEE        EEEEEE           SOLOVYEV
       HHHHHHH        EEEEEE     EEEEEEEE      EEEEEE           Doolittle
 EEE    HHHHHHHH     EEEEEEE      EEEEEE       EEEEEE           HUBBARD
 EEE   HHHHHHHHHHHH   EEEEE   HHHHHHHHHHH      EEEEEE    HHH    COHEN
        EEE         HHHHHHH   EEEEE             EEEE            LENGAUER
 E    E  EE HHHEEE     EEE       EEEEE    E    EEEEE      EE    MURZIN
 E      HHHHHHHHH          EE     EE  EE    EE  EE       EE     MOULT  (4)
```

```
           edge     core                core          core
FKKNWIQYKEGFGHLSPTGTTEFWLGNEKIHLISTQSAIPYALRVELEDWNGRTSTADYA    sequence
    HHHHHH EE         EE  HHHHHHHHHHHH  EEEEEEEE     EEEEEEE    experimental
HHHHHHHHHHH         EEEEEE    EEEE      EEEEEEEE     EEEEEEE    JAAP
HHHHHHHHHHHH        EEEEE HHHHHHHHHHH   EEEEEEEEE    EEEEEE     BENNER
HHHHHHHHHH           EEEE   HHHHH       EEEEEEE        EEEE     STERNBERG
       EEEEEEEEE        EEEEEEEEE          EEEEEEE    EEEEEE    ABAGYAN
                     EEE     EEEEE      EEEEEE                  SOLOVYEV
HHHHHHHHHHHH         HHH    HHHHH       EEEEEEE          EEE    Doolittle
HHHHHHHHHHH         EEE    HHHHHHH      EEEEEEEE      EEEEEE    HUBBARD
HHHHHHHHHHHH        EEEEE  HHHHHHHHHH   EEEEEEEE    EEEEEE      COHEN
    EEEEEE  HHHH     EEEE            EEEEEEEEE             E    LENGAUER
E                E     EEEE        E       EEE                  MURZIN
   EE  HHHHHHHH   ------------------------------------------    MOULT (4)
```

```
 edge      core edge                            not core
MFKVGPEADKYRLTYAYFAGGDAGDAFDGFDFGDDPSDKFFTSHNGMQFSTWDNDNDKFE    sequence
  EE  HHH   EE  EEEE     HHH         HHHHH       EE  E          experimental
EEEE     EEEEEEEE     HHHHHHH          EEEEE   EEEEE            JAAP
 EEEE    EEEEEEEE                      EEEEE  EEEEEE            BENNER
          EEEEEEEE                             EEEE             STERNBERG
E          EEEEEEE    EEEEE    EE EEE               EEEEEEEE    ABAGYAN
 EEE      EEEEEEEE                     EEEE    EEE              SOLOVYEV
EEEE      EEEEEEEE                                              Doolittle
EEEE    HHHEEEEEEE   HHHHHHHH          EEEE    EEEE      HHH    HUBBARD
 EEEE    EEEEEEEEEE         HHHHHH            EEEEE             COHEN
EEEE             EEEEEEEEE        EEEEE        EE               LENGAUER
EEEE      EEEE                            E EEEEE     EEEEE     MURZIN
------------------------------------------------------------    MOULT (4)
                         **************                         gapped regions
```

```
       not core hairpin not core          hairpin not core
GNCAEQDGSGWWMNKCHAGHLNGVYYQGGTYSKASTPNGYDNGIIWATWKTRWYSMKKTT    sequence
  HHHHH   E       EE   E     E            E EE      E   EEE     experimental
           HHHHH       EEEE  EEEE          EEEEEE  EEEEE  EE    JAAP
  EEEEE   EEE   EEE    EEEE             EEEEEEEE        EEE     BENNER
          EEEE                              EEEE          EE    STERNBERG
E  HHHH           EEEEEEEEE  EEEEEE        EEEEEEEEEEEEE  EE    ABAGYAN
                        EEE                EEEEE           E    SOLOVYEV
   EEEE   EEEE                             EEEEE        HHHH    Doolittle
HH EEE    HHHHHHHH     EEE   EEE           EEEEEE   EEEEEEEE    HUBBARD
```

```
  EEEEE   EEEEE        EEEE              EEEEEEEE    EEEEEEE    COHEN
 EE  EEEEEEE          EEEEE    HHHHHH                           LENGAUER
  HHHHH   E       EE   E     E            E EE       -------    MURZIN
  EEEEE  EEEEE    ------------------------- EEEE    EEEE        MOULT (4)
                           **************                       gapped regions

core
MKIIPFNR                                                        sequence
EEEEE                                                           experimental
EEEEEE                                                          JAAP
EEEE                                                            BENNER
EEE                                                             STERNBERG
EEEEEEEE                                                        ABAGYAN
EEE                                                             SOLOVYEV
HHHHH                                                           Doolittle
EEEE   H                                                        HUBBARD
EEEE                                                            COHEN
                                                                LENGAUER
--------                                                        MURZIN
    ----                                                        MOULT (4)
```

Figure 53. Sequence and predictions from the CASP2 site and experimental secondary structure[331] for γ-fibrinogen C terminus, human (268 residues), T0005, 1fib, P02679, F1GB_HUMAN. Experimental secondary structure (DSSP) were from the CASP2 site. Key: E, β strand; H, α helix. Number in parentheses (*n*) indicates the prediction was a weighted average of *n* predictions. The prediction with the highest S_{ov}–O is shown. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: HUBBARD, 69.6, 65.9; BENNER, 63.3, 64.7; JAAP, 62.3, 62.9; COHEN, 62.1, 61.5; Doolittle, 54.3, 65.8; STERNBERG, 53.7, 69.4; SOLOVYEV, 50.8, 65.3; ABAGYAN, 47.1, 49.6; MOULT (4), 43.7, 49.5; MURZIN, 43.3, 51.1; LENGAUER, 39.0, 44.5. The Doolittle prediction was independent of CASP2, while the transparent predictions BENNER and COHEN are discussed elsewhere.[330] MOULT and MURZIN were derived from a coordinate model and are fold-recognition based. The line marked with an asterisk (*) shows where the sequence is matched against gaps in a multiple sequence alignments, where secondary structural elements assigned in the experimental structure are presumably not conserved throughout the family.

3. Gamma Fibrinogen C Terminus (T0005)

Figure 53 collects the secondary structure predictions submitted for the CASP2 project for the C-terminal segment of γ-fibrinogen. Independent of the CASP2 project, Doolittle assembled a secondary structure model for fibrinogen in 1992.[131] To do so, he applied the Kyte–Doolittle amphiphilicity tool,[192] the transparent method of Benner and Gerloff[91] and the consensus Chou–Fasman and consensus GOR tools together and produced a joint prediction. Doolittle's prediction is recorded in Figure 53 as well. Further, the BENNER and COHEN predictions, made jointly (the COBEGETJ prediction team) were presented together with a full evolutionary analysis of the family in published form.[330]

With 15–20 homologous protein sequences in the protein family, divergence sufficient to sustain an evolution-based structure prediction, and a variety of predictions from many different methods, the fibrinogen prediction is one of the most useful to come from the CASP2 project.

Inspection of Figure 53 shows that in the first half of the sequence, all of the predictions are quite good. In contrast, all of the predictions appear to be worse in the second half. Inspection of the transparent predictions[131,330] shows why the prediction is so uneven. In the first part of the sequence, the multiple alignment is of high quality. In the second, the multiple alignment is poor. In the second half of the protein, segments found in the target sequence are deleted in homologs. This implies that secondary structural elements assigned in these regions in the experimental structure are not core elements, and cannot be predicted by an evolution-based tool of any kind. These are marked in Figure 53 with asterisks.

The relatively low Q_3 and S_{ov}–O scores for this prediction are attributable to the divergence of secondary structure in this family of proteins and the large amount of coil. As with threonine deaminase, a single score loses the important information in evaluating this target, and the Q_3 score is inadequate, even as a crude measure of prediction quality to be used as a "cutoff". So many of the segments evaluated are not core that a 68% Q_3 score is virtually unattainable for a consensus prediction, even a perfect one. Indeed, the only prediction to make the 68% cutoff is by STERNBERG.

Transparency was especially useful in understanding the assignment of the third strand in the structure (the fourth secondary structural element in line 1 in Figure 53). As pointed out in Gerloff *et al.*,[330] both a strand and a helix are consistent with patterns of predicted exposure in this segment, the first preferred based on simple analysis of the sequence data, the second based on considerations of tertiary packing. Gerloff *et al.* noted that both secondary structural elements must be considered when building a tertiary structural model.[330]

4. Bactericidal Permeability-Increasing Protein (T0010)

Only four homologous sequences could be found in the sequence database for T0010. The four sequences come in two pairs. Each sequence in the pair is separated by 50 PAM units, while the pairs themselves have diverged by ~100 PAM units. Thus, this target should not give good predictions using evolu-

```
VNPGVVVRISQKGLDYASQQGTAALQKELKRIKIPDYSDSFKIKHLGKGHYSFYSMDIRE    sequence
    EEEEEEHHHHHHHHHHHHHHHHHHH       EEEEEEE   EEEEEEEEEEEEEE    experimental
    EEEEEHHHHHHHHHHHHHHHHHHHHHHEE        EEEE    EEEEEEEEE      STERNBERG
    EEEEE   HHHHHHHHHHHHHHHHHHHH        EEEEE     EEEEEEEEHH    JAAP
    EEEEE             HHHHHHHHHH        EEEE      EEEEEEEE E    FINKELSTEIN
    EEEEE HHHHHHHHHHHHHHHHHHHHEE        HHEEHH    E EEHHH  H    MUNSON
    EEEEE            HHHHHHHHHH        EEEEEE       EE          SOLOVYEV
     EEEE  HHHHHHHHHHHHHHHHHHHH EEE     EEEEE   EEEEEEEEEEE     ROST
       EE    HHHHHHHHHHHHHHHHHHHH   E     EEEE  EEEE            MURZIN

FQLPSSQISMVPNVGLKFSISNANIKISGKWKAQKRFLKMSGNFDLSIEGMSISADLKLG    sequence
EE   EEEEEE   EEEEEEEEEEEEEEEEEEEEE  EEEEEEEEEEEEEEEEEEEEEEE    experimental
                EEEE    EEEE     HHHEEE     EEEE   EEEEEEEE     STERNBERG
HH    EEEE      EEEE   EEEE    HHHHHHHHH     EEEE EEEEEEEEE     JAAP
EEE   EEEEE  EEEEEEEE  EEEE                EEEEE  EEEEEEEEE     FINKELSTEIN
H               EEE     EEH   HHHHHHHEE      E EEEEEEEEEEE      MUNSON
               EEEEE   EEEE        HHHH     EE           EE     SOLOVYEV
EE             EEEEE    EEEEEEE HHHHHH     EEEEEE EEEEEE E      ROST
                       EEEEEEEEEE      EEEEEEEEEEE E      E     MURZIN

SNPTSGKPTITCSSCSSHINSVHVHISKSKVGWLIQLFHKKIESALRNKMNSQVCEKVTN    sequence
EE    EEEEEEEEEEEE    EEE      HHHHHHHHHH HHHHHHHHHHHHHHHHHH    experimental
        EEEEE       EEEEEE       HHHHHHHHHHHHHHHHH   HHHHHH     STERNBERG
         EEEEE      EEEEEEE    HHHHHHHHHHHHHHHHHHHHHHHHHHHHH    JAAP
        EEEE         EEEEE    HHHHHHHHHHHHHHHHHHHH   HHHHHHH    FINKELSTEIN
        EEEEE        EEEE     H EHHHHHHHHHHHHHHHHH     EEEEE    MUNSON
        EEE          EEEE      HHHHHHHHHHHHHHHH        HHHHH    SOLOVYEV
        EEEEE        EEEE     HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH    ROST
         E     E EEEEEEE         HHHHHH        E HHHHHHHH HH    MURZIN

SVSSELQPYFQTLPVMTKIDSVAGINYGLVAPPATTAETLDVQMKGEFYSENHHNPPPFA    sequence
HHHHHHHHH      EEE    EEEE  E    EE   EEEEEE  EEEE              experimental
              EEE         EEEE     HHHHHHHH   EEE               STERNBERG
HHHHHH        EEEEE  EEE EEEEE   HHH  HHHHHH  EEE               JAAP
HHHHHHHHHHHHH           EEEEEEE        EEEEE                    FINKELSTEIN
 EE            EEE        HEE       HHHHHHHH   EE               MUNSON
HHHH     HHHHHHHHHHHHH     EEE      HHHHHHH    EE               SOLOVYEV
H HHHHHHHH     EEE   E   EEEEE   HHHHHHHHHHE E                  ROST
HHHHHHHHHH     EE   ----------------------------------------    MURZIN

PPVMEFPAAHDRMVYLGLSDYFFNTAGLVYQEAGVLKMTLRDDMIPKESKFRLTTKFFGT    sequence
             EEEEEEHHHHHHHHHHHHH    EEEEE            EHHHHHH    experimental
       HHHHHHEEEE     HHHHHHHHHH  EEEEEE                        STERNBERG
 HHH   HHHHHEEEEEEEEEE     EEEEEE  EEEEEE         EEE   HHHH    JAAP
  EEEE      EEEE HHHHHHHHH EEEEHHHHHEEEE            HHHHHHHH    FINKELSTEIN
      HHHHHHHEHEEE HHHHHHHHHHHHHHHHEEEEE         HHH  EEE       MUNSON
  HHH        EEE       HHHHHHHHH   EEE            EEEEE    H    SOLOVYEV
            EEEEEEEHHHHHHHHHHHHH   EEEEEE          EEE          ROST
------------------------------------------------------------    MURZIN

FLPEVAKKFPNMKIQIHVSASTPPHLSVQPTGLTFYPAVDVQAFAVLPNSSLASLFLIGM    sequence
H HHHHHH    EEEEEEEE    EEEEE  EEEEE EEEEEEEEE     EEEEEEEEE    experimental
  HHHHHH    EEEEEE      EEE             EEEEEE        EEEEEE    STERNBERG
H HHHHHH     EEEEEEE     EEE          EE EEEEE       EEEEEEE    JAAP
HHHHHHHHH  EEEEEEE      EEEEE   EEEE  EEEEEEEEE       EEEEEE    FINKELSTEIN
  HHHHH     HHHEEEE      E           HHHHEEEEE       HHEEEEE    MUNSON
HHHHHHHH    EEEEEEE                        EE      HHHHEEEE     SOLOVYEV
              EE         EEE     EEEEEEEEEEEEE        EEEEEE    ROST
------------------------------------------------------------    MURZIN
```

```
HTTGSMEVSAESNRLVGELKLDRLLLELKHSNIGPFPVELLQDIMNYIVPILVLPRVNEK    sequence
E  EEEEEEEE  EEEEEEEE   EEEEEEE      HHH HHHHHHHHHHH HHHHHHH    experimental
    EEEEEE   HHHHH  HHHHHEEE          HHHHHHHHHEE       HHHH    STERNBERG
   HHHHHHH    EEEEE    EEEEEE        HHHHHHHHEEE EEEE   HHHH    JAAP
EE  EEEEE           HHHHHHHHHHH     HHHHHHHHHHHHEEEEEE          FINKELSTEIN
   EEEEEHH H HHH      HHHHHH        HHHHHHHHHHHEEEEEE   HHHH    MUNSON
                    HHHHHHHH         HHHHHHHHHHH                SOLOVYEV
EEEEEEEEEE    EEEEE   EEEEEEE       HHHHHHHHHHHHHHHH    HHHH    ROST
------------------------------------------------------------    MURZIN

LQKGFPLPTPARVQLYNVVLQPHQNFLLFGADVVYK                            sequence
H   EE      EEEEEEEEEEE  EEEEEEEEEE                             experimental
HH          EEEE         HHEE                                   STERNBERG
HH         EEEEEEEE       EE   EEE                              JAAP
               EEEEEE   EEEEEEEEEEE                             FINKELSTEIN
H            EEHHHHH  HHHHEEH   HEEE                            MUNSON
            EEE   EE     EEEE                                   SOLOVYEV
HHH          EEEE  E      EEE  EEEE                             ROST
------------------------------------                            MURZIN
```

Figure 54. Sequence and predictions from the CASP2 site and experimental secondary structure[332] for bactericidal permeability-increasing protein, human (456 residues), T0010, 1bpi, P17213, BPI_HUMAN. Experimental secondary structural assignments (DSSP) were taken from the CASP2 site. Key: E, β strand; H, α helix. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: JAAP, 61.8, 64.3; FINKELSTEIN, 57.6, 64.7; ROST, 56.7, 69.5; STERNBERG, 55.9, 60.3; MUNSON, 47.8, 53.9; SOLOVYEV, 43.8, 49.8; MURZIN, 43.5, 56.0, from coordinate model.

tionary-based tools. Further, the protein is big, with 456 amino acids. For whatever the reason, the S_{ov} and Q_3 scores for this target were poor even in the best prediction (61.8 and 64.3, by JAAP). Inspection of Figure 54, which collects the secondary structure predictions submitted for the bactericidal permeability-increasing protein, shows the problems in detail. In the N-terminal domain, which is the region of the protein that binds lipopolysaccharides, the predictions underestimate the lengths of the β strands that distinguish the experimental secondary structure. None of this can be ascribed to divergence in secondary structure, as the multiple alignment contains no gaps. In the second half of the prediction, a small number of strands are misassigned as helices.

5. HSP90 N-Terminal Domain (T0011)

With over 30 homologous sequences and substantial evolutionary divergence, the N-terminal domain of the heat shock protein 90 (HSP90) provides an excellent target for evolution-based modeling. As expected for such an input, the Q_3 and S_{ov}–O scores for evolution-based predictions were high. Figure 55 contains the secondary structure predictions and the experimentally assigned secondary structure, with core and edge strands assigned.

Considering issues related to scoring, the importance of distinguishing between mistakes in core and noncore assignments is illustrated here. For example, the GOLDSTEIN prediction misassigns a core strand as a helix, while the MUNSON prediction misassigns an edge strand as a helix. The two misassignments score identically, but only the first prevents assembly of a correct tertiary structural model from the predicted secondary structural elements. Further, a three-residue helix (Figure 55, line 2) is assigned to the experimental structure. Such a helix is, of course, less than a full turn, and is rarely a core element. No tool predicts it, and the tools are not deficient for not doing so. Likewise, the four residue helices at the end of the first line and at the start of line 4 are not significant, and the value of predictions that do not predict them are not diminished.

The value of the predicted models for secondary structure in this protein was illustrated by the application of the models to predict tertiary structure in the family, and the use of the tertiary structure models to solve biochemical problems identified in the literature of this family. This was done by two participants in the CASP2 project, both who make transparent predictions, the COBEGETJ team and BAZAN.

The COBEGETJ team recognized that the predicted secondary structural elements for T0011 could be mapped on the ATPase domain of gyrase (found by SCOP browsing).[333,334] The team obtained the coordinates as a personal communication from D. B. Wigley. Upon closer comparison of the predicted tertiary structure model, based on the predicted secondary structure elements and active-site assignments, the COBEGETJ team concluded T0011 might be a distant homolog of gyrase, was likely to adopt the same fold except for an inserted hairpin structure (residues 54–70) and a region (86–117) that forms a lid in the gyrase structure.[334]

The model was then used to address a biochemical question concerning HSP90 (target T0011). The literature had not established by "wet" biochemical experiments whether HSP90 bound ATP. Indeed, a report issued just as the CASP2 project was running stated that "highly purified Hsp90 does not bind ATP".[335] The prediction identified an ATP-binding site, however, and the COBEGETJ team drew the correct conclusion that the protein did indeed bind ATP.

The prediction and biochemical conclusions made by the COBEGETJ team involved human interven-

```
 edge
ASETFEFQAEITQLMSLIINTVYSNKEIFLRELISNASDALDKIRYKSLSDPKQLETEPD  sequence
  EEEE  HHHHHHHHHHHH       HHHHHHHHHHHHHHHHHHHHH   HHHH         experimental
      HHHHHHHHH         HHHHHHHHHHHHHHHHHHH                     BAZAN
    EEEEEEE HHHHHHHHHHHH  HHHHHHHHHHHHHHH  EEEEEEE    EEEE      COHEN
  HHHHHHHHHHHHHHEEE        EEEHHHH   HHHHHHEEEE                 STERNBERG
 HHHHHHHHHHHHHHHHHHHHH   HHHHHHHHHH HHHHHHHHHHHEE          E    ROST (2)
 HHHHHHHHHHHHHHHHHHHHEE  HHHHHHHHHH HHHHHHHHHHH            H    JAAP
HHHHHHHHHHHHHHHHHEH      HHHEHHHHHHHHHHHHHHHHHH            H    MUNSON
HHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHHHHHHH             E    SOLOVYEV
HHHHHHHHHHHHHHHHHHHHEE    HHHHHHHHHHHHHHHHHHHH                  GOLDSTEIN
 HHHHHHHHHHHHHHHHHHHHH  HHHHHHHHHHHHHHHHHHHHHHEEE               VALENCIA
    HHHHHHHHHHHHH HHHHH HHHHHHH   -     --------------   ---    BAKER
------------------------------------------------------------    ROSE (2)

 core       core
LFIRITPKPEQKVLEIRDSGIGMTKAELINNLGTIAKSGTKAFMEALSAGADVSMIGQFG    sequence
  EEEEEEHHH EEEEEE      HHHHHHH        HHHHHHHHHH   HHHHHHH     experimental
EEEEEEE    EEEEEEE                     HHHHHHHHHH               BAZAN
EEEEEE      EEEEEEE      HHHHHHH       HHHHHHHHH                COHEN
EEEEEE      EEEEEE      EEE            HHHHHHHH                 STERNBERG
EEEEEEE     EEEEEE      HHHHHHHHHHHHH  HHHHHHHHHH   EEEEEEE     ROST (2)
HHHEEE      EEEEE       HHHHH  HHHHH   HHHHHHHHHH    EEEEEE     JAAP
EEEEE       EEEEEE  EE EHHHHHH  HHHHHH HHHHHHHHHH   HEEE E      MUNSON
EEEEEE      EEEEE       HHHHHHHHHHHHHH HHHHHHHHHH    EEEE       SOLOVYEV
 EEEE     HHHHHHH        HHHHHHHHHHHH  HHHHHHHHH    EEEE        GOLDSTEIN
EEEEEE      EEEEE       HHHHHHHHHHHHHH HHHHHHHHHH   EEEEEEE     VALENCIA
------  -  -            HHHHHHH     HHHHHHHHHH    ----------    BAKER
------------   EEE  EEE  HHHHHHHHHHHHH    HHHHHHHHHH   -----    ROSE (2)

           core         core      core            core
VGFYSLFLVADRVQVISKSNDDEQYIWESNAGGSFTVTLDEVNERIGRGTILRLFLKDDQ    sequence
 HHHHHHH EEEEEEEEE     EEEEEE    EEEEEE         EEEEEEEE        experimental
HHHHHHHHH  EEEEEEE    EEEEEEE    EEEEEE          EEEEEEE        BAZAN
 HHHHHHHH  EEEEEE     EEEEEE      EEEEEE        EEEEEEEE        COHEN
  EEEEEE   EEEEEE        EEEE    EEEEE            EEEEEE        STERNBERG
  HHHEEEEEEEEEEEEE      EEEEE    EEEEEE           EEEEE    H    ROST (2)
  EEEEEEEE EEEEEEE    HHHEEEE    EEEEEHHH  HHH    EEEEE    H    JAAP
  EEEHEEEHEEEEEEE      HHHHHH     EEEEE           EHEEEHHH H    MUNSON
    EE EE  EEEEEE       EEEE      EEEEE           EEEEEE   H    SOLOVYEV
    HEEEEH  EEEEE       EEEEE      EEEHHHHHHH    EEEEEEE HHH    GOLDSTEIN
   EEEEEEEEEEEEEEE       EEEE     EEEEE           EEEEE    H    VALENCIA
-                  --------------      ------            HHH    BAKER
-----------------------------------------------------------     ROSE (2)

                      edge
LEYLEEKRIKEVIKRHSEFVAYPIQLVVTKEVEKEV                            sequence
HHHH HHHHHHHHHHH       EEE                                      experimental
 HHHHHHHHHHHHHHHH     EEEEEE  EEEEEE                            BAZAN
  HHHHHHHHHHHHHHHHHH   EEEEEE   EEEE                            COHEN
  HHHHHHHHHHHH         EEE                                      STERNBERG
HHHHHHHHHHHHHHHHHH    EEEEEEEE                                  ROST (2)
HHHHHHHHHHHHHHHHHH E   EEEEEEEHHH                               JAAP
HHHHHHHHHHHHHHHHHHH E  HHHHHH                                   MUNSON
HHHHHHHHHHHHHHHH        EEEE  HHHHH                             SOLOVYEV
HHHHHHHHHHHHHHH  HHE   EEEEE HHHH                               GOLDSTEIN
HHHHHHHHHHHHHHHHHH    EEEEEEE                                   VALENCIA
HHHH HHHHHHHHHH      ------- --                                 BAKER
-------------------------------------                           ROSE (2)
```

Figure 55. Sequence and predictions from the CASP2 site and experimental secondary structure for HSP-90 N-terminal domain,[337] *S. cerevisiae* (220 residues), T0011, PO2829, HS82_YEAST. Experimental secondary structural assignments, calculated with DSSP, were taken from the CASP2 site. Key: E, β strand; H, α helix. The number in parentheses (n) indicates the prediction was a weighted average of n predictions. The prediction with the highest S_{ov}–O is shown. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: COHEN, 75.6, 68.1; ROST (2), 72.4, 74.5; VALENCIA, 72.1, 71.8; BAZAN, 70.3, 71.3; SOLOVYEV, 67.9, 69.4; STERNBERG, 66.4, 68.5; JAAP, 61.5, 65.7; GOLDSTEIN, 59.6, 62.5; MUNSON, 53.6, 64.4; ROSE (2), 49.5, 47.8; BAKER, 49.3, 52.0.

tion. At noted above, several individuals in the field have criticized such procedures as being unreproducible.[65] Thus, it is interesting to note that the same conclusions concerning secondary structure, tertiary

```
LTSTERLIQLFNSWMLNHNKFYENVDEKLYRFEIFKDNLNYIDETNKKNNSYWLGLNEFA    sequence
  HHHHHHHHHHHHHH        HHHHHHHHHHHHHHHHHHHHHH                  experimental
   HHHHHHHHHHHHHHHHHH   HHHHHHHHHHHHHHHHHHHHH      EEEEE        ROST
 HHHHHHHHHHHHHHHHHH      HHHHHHHHHHHHHHHHHHHH       HHHHHHHH    STERNBERG
   HHHHHHHHHHHHHHH      HHHHHHHHHHHHHHHHHHHHH      EEEEE        JAAP
  HHHHHHHHHHHHHHHH      HHHHHHHHHHHHHHH  HHHHHHHH  EEEEE        ABAGYAN
   HHHHHHHHHHHHHH       HHHHHHHHHHHHH HHHHH        HHHEE        MUNSON
   HHHHHHHHHHHHH   HHHHHHHHHH  HHHHHHHHHHHHH           HHHHH    SOLOVYEV

DLSNDEFNEKYVGSLIDATIEQSYDEEFINEDTVN                             sequence
   HHHHHHHH                                                     experimental
   HHHHHHHHH                                                    ROST
HHHHHHHHHHHH                                                    STERNBERG
   HHHHHHHHH              EEE                                   JAAP
    HHHHHHH                                                     ABAGYAN
   HHHHHHHH              HH                                     MUNSON
        HHHHHHHHHHHH     HHHHH                                  SOLOVYEV
```

Figure 56. Sequence and predictions from the CASP2 site and experimental secondary structure for proregion of procaricain, *Carica papaya* (107 residues),[338] T0012, 1pci, EM_PL:CPPRO. Experimental secondary structural assignments, calculated with DSSP, were taken from the CASP2 site. Key: E, β strand; H, α helix. For each prediction, S_{ov}–O and Q_3 were calculated for only the nonhomolgous residues and are listed in order of descending S_{ov}–O: MUNSON, 97.2, 91.7; ABAGYAN, 97.2, 91.7; JAAP, 97.2, 88.9; STERNBERG, 92.0, 83.3; ROST, 86.1, 80.6; SOLOVYEV, 68.9, 75.0; and (for residues 1−48) ABAGYAN, 97.2, 91.7; MUNSON, 97.2, 91.7, JAAP, 97.2, 88.9; STERNBERG, 92.0, 83.3; SOLOVYEV, 92.0, 75.0; ROST, 86.1, 80.6.

structure, and biochemical behavior were derived independently by Bazan. Bazan noted that he conducted an exhaustive survey of Hsp90 homologs from the nonredundant NCBI databases using the BLAST server with Gonnet−Benner[220] and Blosum45 and 30[336] comparison matrices. These sequences were collected and aligned with ClustalW to make Gribskov-type profiles, used to screen again for more distant relatives. From both the BLAST and profile searches, the human TRAP1 and *C. elegans* ORF sequences (Genbank accession U12595 and U00036) were incorporated to the profile. Next, a hypothetical prokaryotic protein (SwissProt yd3m_herau) was found. The augmented profiles, and the MPSRCH server (DISC in Japan) located a significant, albeit faint, similarity to bacterial MutL proteins (involved in DNA mismatch repair complexes) centering around an Hsp90 conserved motif of DxGxG (aa 79−83 in target). All MutL-like sequences were separately harvested and aligned (including MLH1- and PMS1-like proteins in eukaryotes, with some quite distant homologs found as ORFs in the yeast genome) using ClustalW. BLAST/profile searches next revealed two interesting matches that had appeared as bottom-type hits with the Hsp90 profile, each with a number of bacterial sensor proteins from two-component signaling pathways to central regions that correspond to putative histidine kinase domains, and to the N-terminal segments of bacterial gyrase subunit-B sequences, also ATPase domains.[334] Both of these divergent families also preserve DxGxG motifs at approximately the same spot as Hsp90s/MutLs, about $^1/_3$ of the way into the chain; another centrally located Gly-rich motif also cemented the relationship.

Bazan then writes that the growing multiple alignments were submitted to the PHD neural network prediction server, and to the PSSP server at Baylor implementing Solovyev's SSP and NNSSP programs. The Hsp90 and MutL predictions were quite similar, with an $\alpha + \beta$ pattern of $\alpha-\alpha-\beta-\beta-\alpha-\alpha-\beta-\beta-\beta-\beta-\alpha-\beta-\beta$. The histidine kinase domains, smaller in size feature a pattern of $\alpha-\alpha-\beta-\beta-\alpha-\beta-\beta-\alpha-\beta-\beta$ (minus two strands), while the gyraseB-like sequences (clustering a kinase) feature a pattern of $\alpha-\alpha-\beta-\beta-\alpha-\beta-\beta-\alpha-\beta-\beta$ (less two strands), while the gyraseB-like sequences (clustering in prokaryotic and eukaryotic families) are HSP90/MutL-like in length, and give similar $\alpha + \beta$ patterns. Routine checks were run of representative members of the Hsp90, MutL, HisKin, and GyrB families with the threading programs 123D (Alexandrov, NCI), topits (Rost, EMBL), Pscan (Eloffson, Stockholm), and ProFIT; none of these appeared to be similar, although most of the hits were with $\alpha + \beta$, or α/β folds.

Bazan then noted that the York group has earlier solved the X-ray structure of *E. coli* gyraseB,[334] but that coordinates had not yet been deposited in the PDB. The gyrase B fold is composed of two distinct domains: an N-terminal novel ATPase structure formed by a mixed β sheet with helices packed on one side, and a C-terminal α/β fold related to domains in ribosomal proteins and EF-G. The location of the GyrB ATPase secondary structural elements correspond quite well with the PHD/DSSP-derived helices and strands.

From this template fold, Bazan deduced the likely topology of the predicted HSP90 secondary structure, as well as the strand pairing/contacts. Viewing the sheet from above (looking down at the helices lying on top of the sheet), the eight strands are in order 5-4-3-6-2-1-7-8, all antiparallel save for the 1−7 pair, which are parallel to each other. Two helices precede the first β strand, and then also form links between strands 1−2 and 6−7. The more economical histidine kinase sequences may lack the edge 5−4 hairpin—this looks to be allowed by the fold. The ATP-binding site, as mapped by the presence of the ADPNP, is on top of the sheet, protected by various loops and helices. The noted Asp-Xxx-Gly-Xxx-Gly motif was observed to lie in a loop just after strand 2; in the GyrB-ADPNP complex, the Asp73 side chain interacts with

```
MKTVTVKNLIIGEGMPKIIVSLMGRDINSVKAEALAYREATFDILEWRVDHFMDIASTQS    sequence
   EEE  EEE     EEEEEE    HHHHHHHHHHHHH    EEEEEHHH      HHH    experimental
  EEEEEEEE      EEEEEHHHHHHHHHHHHHHHHHH   HHHEEHHHHHHHH  HHH    ROST (2)
   EEEEEEE      EEEEE   HHHHHHHHHHHHHH   HHHHEEEEEHHHH    HH    STERNBERG
 EEEEEEEEE      EEEEE   HHHHHHHHHHHHHHHH   HHHEEHHHHH HHHHHH    JAAP
  EEEEEEEEE   HHHHHHHHHHHHHHHHHHHH      EEEEEEE HHHHHHHHH       FINKELSTEIN
HHHHHHHHHHHHH HHHHHHHH     EEEEEEE    HHHHHHHHHH    EEE         ABAGYAN
  EEEEEEEEE      EEEEE    HHHHHHHHHHHHHH HHHHHHHHHHHH    HHH    MUNSON (7)
   EE           EEEEEE      HHHHHHHHHHH  HHHHHHHHHHHH    HHH    SOLOVYEV (2)
                         EE EEEE    HHHHHHHHHHHHHH     EEEEE    LENGAUER
        HHHHH      EEEEE      HHHHHHHHHHH  EEEEE                MURZIN

VLTAARVIRDAMPDIPLLFTFRSAKEGGEQTITTQHYLTLNRAAIDSGLVDMIDLELFTG    sequence
HHHHHHHHHHH     EEEE  EHHH   E   HHHHHHHHHHHHHH    EEEEEHHH     experimental
HHHHHHHHHHH     EEEEE            HHHHHHHHHHHHH     EEEE         ROST (2)
HHHHHHHHHHH     EEEEHHHHHH        HHHHHHHHHHHHH  HHHHHHHH       STERNBERG
HHHHHHHHHHH    EEEEEEHHHH        HHHHHHHHHHHHH    EEEEE         JAAP
HHHHHHHH      HHHHHHH         EE    EEEE            EEEEEEE     FINKELSTEIN
HHHHHHHHHHHHHHH EEEE           HHHHHHHHHHHHH      EEEEE         ABAGYAN
HHHHHHHHHHH      HE     H       HHHHHHHHHHHHHHH HHHHHHHHH       MUNSON (7)
HHHHHHHHHHH     EEEE           EEE  HHHHHHHHHH        HHHH      SOLOVYEV (2)
HHH  HHHHHHH    EEEEEE   HHHHHHHHHHHH            HHH  HHHHHH    LENGAUER
 HHHHHHHHHH     EEEE                 HHHHHHHHHH    EEEEE        MURZIN

DADVKATVDYAHAHNVYVVMSNHDFHQTPSAEEMVSRLRKMQALGADIPKIAVMPQSKHD    sequence
HHHHHHHHHHHHH   EEEEEEEE      HHHHHHHHHHHHH    EEEEEE    HHH    experimental
 HHHHHHHHHHH   EEEEEEE        HHHHHHHHHHHHH     EEEEE    HHH    ROST (2)
  HHHHHHHHHH   EEEEEE         HHHHHHHHHHHHHHHHHHHHHEE    HHH    STERNBERG
 HHHHHHHHHHH   EEEEEE         HHHHHHHHHHHHHH     EEEE    HHH    JAAP
  EEEEEEEEE   EEEEEE          HHHHHHHHHHH         EEEE          FINKELSTEIN
 HHHHHHHHHHHH    EEEEE        HHHHHHHHHHHHHH     EEEEE          ABAGYAN
  HHHHHHHHHHH  EEEEEE         HHHHHHHHHHHHHHH  HHHHEE      H    MUNSON (7)
   HHHHHHHH     EEEEEE        HHHHHHHHHHHHHH     EEEE           SOLOVYEV (2)
HHHHH HHHHHHHHHHHHHHHHH EEEE  HHH        HHHHH   EEEE  HHHHH    LENGAUER
HHHHHHHHHHHHH   EEEEEE   E       E HHHHHHHHH   EEEEE            MURZIN

VLTLLTATLEMQQHYADRPVITMSMAKEGVISRLAGEVFGSAATFGAVKQASAPGQIAVN    sequence
HHHHHHHHHHHHHH      EEEE     HHHHH HHHH   EEE E          EHH    experimental
HHHHHHHHHHHHHH     EEEEEE     EEEEEE            E         HH    ROST (2)
HHHHHHHHHHHHH       EEHHHHHHHHHHHHHHH                     HH    STERNBERG
HHHHHHHHHHHHHHH    EEEEE     EEEE   EEE  HHHH           EEHH    JAAP
HHHHHHHHHHHHHHHH   EEEEEE    HHHHHHHHHHHHHHHHHHHHHH     EEEE    FINKELSTEIN
  HHHHHHH     EEEEEE    HHHHHHHHHHH  EEEE  HHHHHHH    EEEE      ABAGYAN
HHHHHHHHHHHHH      HEEHHH  H  EHHHH HHH    E   HH         HH    MUNSON (7)
HHHHHHHHHHHHH      EEEEE     HHHHHHHH    HHHHHHHHHHHH           SOLOVYEV (2)
HHHHHHH EEEEEEEE     HHHHHHH                      HHH HHHHHH    LENGAUER
    HHHHHHHHHH      EEE            HHHHHHHHHH     EEE HHHH      MURZIN

DLRSVLMILHNA                                                    sequence
HHHHHHHHHHH                                                     experimental
HHHHHHHHHH                                                      ROST (2)
HHHHHHHH                                                        STERNBERG
HHHHHHHHHH                                                      JAAP
 HHH EEEE                                                       FINKELSTEIN
HHHHHHHHHHHH                                                    ABAGYAN
HHHHHHHHHHHH                                                    MUNSON (7)
 HHHHHHHHH                                                      SOLOVYEV (2)
HH HHH                                                          LENGAUER
 HHHHHHHHHH                                                     MURZIN
```

Figure 57. Sequence and predictions from the CASP2 site and experimental secondary structure for 3-dehydroquinase, *Salmonella typhimurium*[339] (252 residues), T0014, P24670, AROD_SALT1. Experimental secondary structural assignments, calculated with DSSP, were taken from the CASP2 site. STRIDE assignments were not available. Key: E, β strand; H, α helix. The number in parentheses (n) indicates the prediction was a weighted average of n predictions. The prediction with the highest S_{ov}–O is shown. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: JAAP, 81.4, 77.8; ROST (2), 79.5, 79.5; SOLOVYEV (2), 79.4, 73.4; STERNBERG, 73.8, 73.8; MURZIN, 69.6, 69.0, from a coordinate model; MUNSON (6), 67.1, 65.1; ABAGYAN, 54.2, 50.8; FINKELSTEIN, 50.1, 50.8; LENGAUER, 34.8, 42.5.

```
DEIGDAAKKLGDASYAFAKEVDWNNGIFLQAPGKLQPLEALKAIDKMIVMGAAADPKLLK    sequence
 HHHHHHHHHHHHHHHHHHH     HHHH       HHHHHHHHHHHHHHHHH  HHHHH    experimental
HHHHHHHHHHHHHHHHHHHHHHHHHHHHH     HHHHHHHHHHHHHHHHHHHHHHHHHH    STERNBERG
      HHHHHHHHHHHHHHHH              HHHHHHHHHHHHHHHHHHHHH  H    MURZIN
 HHHHHHHHH  HHHHHHHHHHH   EEEE      HHHHHHHHHHHHH HHHHHHHHHH    JAAP
 HHHHHHHHH   HHHHHHHHHH   EEEE      HHHHHHHHHHHHHHHHH  HHHHH    ROST
HHHHHHHHHHHHHHHHHHHH      EEE       HHHHHHHHHHHEEE     HHHHH    SOLOVYEV
HHHHHHHHHHHHHHHHHHHHHHH   HEE       HHHHHHHHHHHHHHHH   HHHHH    MUNSON
------------------------------------------------------------    HUBBARD (2)

AAAEAHHKAIGSISGPNGVTSRADWDNVNAALGRVIASVPENMVMDVYDSVSKITDPKVP    sequence
HHHHHHHHHHH  E    E  HHHHHHHHHHHHHHHH   HHHHHHHHHHHHHH    HH    experimental
HHHHHHHHHHHH        HHHHHHHHHHHHHHHHH     HHHHHHHHH             STERNBERG
HHHHHHHHHHHH              HHHHHHHHHHHH     HHHHHHHHHHHHHHH      MURZIN
HHHHHHHHHHEEEE          HHHHHHHHHHHHH   HHHEEEHHHHHHH     HH    JAAP
HHHHHHHHHHEEEE          HHHHHHHHHHHHHE  HHHEEEHHH          H    ROST
HHHHHHHHHH              HHHHHHHHHHHHH     HHHHHHHHHHHH     H    SOLOVYEV
HHHHHHHHEEEE            HHHHHHHHHHHH    H EEEEEH           H    MUNSON
------------------------------------------------------------    HUBBARD (2)

AYMKSLVNGADAEKAYEGFLAFKDVVKKSQVTSAAGPATVPSGDKIGVAAQQLSEASYPF    sequence
HHHHH   HHHHHHHHHHHHHHHHHHHHH               HHHHHHHHHHHHHHHH    experimental
HHHHHH   HHHHHHHHHHHHHHHHHHHHHHHH          HHHHHHHHHHHHHHHHH    STERNBERG
   HHHHHHHHHHHHHHHHHHHHHHHHH                     HHHHHHHHHHH    MURZIN
HHHHHHH  HHHHHHHHHHHHHHHHHHHEEEE            HHHHHHHHHHHHHHHH    JAAP
HHHHHHH  HHHHHHHHHHHHHHHHHHHEEEE            HHHHHHHHHHHH HHH    ROST
HHHHHHH    HHHHHHHHHHHHHHHHH                 HHHHHHHHHHHHHHH    SOLOVYEV
HHHHHHH    HHHHHHHHHHHHH      E               EHHHHHHHHH  HH    MUNSON
-------------------------------------------  HHHHHHHHHHH        HUBBARD (2)

LKEIDWLSDVYMKPLPGVSAQQSLKAIDKMIVMGAQADGNALKAAAEAHHKAIGSIDATG    sequence
HHH      HHH       HHHHHHHHHHHHHHHHH  HHHHHHHHHHHHHHHH          experimental
HHHHHHHHHHHH     HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH           STERNBERG
HHHHH              HHHHHHHHHHHHHHHHHHHHH  HHHHHHHHHHHHH         MURZIN
HHHHHHHHHH         HHHHHHHHHHHHHHH    HHHHHHHHHHHHHH  EEE       JAAP
HHHHHHHHHHH        HHHHHHHHHHHHHEE    HHHHHHHHHHHHHH  EEE       ROST
HHHHHHH            HHHHHHHHHHHHH       HHHHHHHHHHHHHH           SOLOVYEV
HHHHHHHHHH          HH HHHHHHHHHHHHHHHHHHHHHHHHHHHHHEE          MUNSON
HHHHHHHHHHHH HHHHHHHHHHHHHHHHHHHHHHH     HHHHH  HHHH            HUBBARD (2)

VTSAADYAAVNAALGRVIASVPKSTVMDVYNAMAGVTDTSIPLNMFSKVNPLDANAAAKA    sequence
   HHHHHHHHHHHHHHH    HHHHHHHHHHHHHH   HHHHHHHH   HHHHHHHHHH    experimental
HHHHHHHHHHHHHHHHHHH   HHHHHHHHHHHH       HHHHHH   HHHHHHHHHH    STERNBERG
        HHHHHHHHHHHH     HHHHHHHHHHHHHHH     HHHHHHHHHHHHHHH    MURZIN
  HHHHHHHHHHHHHHHHHH     HHHHHHHHH               HHHHHHHHHHH    JAAP
 HHHHHHHHHHHHHHHHHEEE     HHHHHHHHH               HHHHHHHHHH    ROST
  HHHHHHHHHHHHHHHHH      HHHHHHHHH                   HHHHHHH    SOLOVYEV
    HHHHHHHHHHHHEEEE    EEEEEEEE          HEE       HHHHHHHH    MUNSON
HHHHHHHHHHHHHHHHHHHHHHHH   HHHHHH       HHHHHHHH   HHHHHHHHH    HUBBARD (2)

FYTFKDVVQAAQ                                                    sequence
HHHHHHHHHHH                                                     experimental
HHHHHHHHH                                                       STERNBERG
HHHHHHHHHH                                                      MURZIN
HHHHHHHHH                                                       JAAP
HHHHHHHHHHH                                                     ROST
HHHHHHHHHHHH                                                    SOLOVYEV
HHHHHHHHHHHH                                                    MUNSON
HHHHH                                                           HUBBARD (2)
```

Figure 58. Sequence and predictions from CASP2 site and experimental secondary structure for peridinin chlorophyll protein, *Amphidinium carterae* (312 residues),[340] T0016, 1ppr, PCP1_AMPCA, P80484 P51872. Experimental secondary structure from DSSP. Key: E, β strand; H, α helix. The number in parentheses (n) indicates the prediction was a weighted average of n predictions. The prediction with the highest S_{ov}–O is shown. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: SOLOVYEV, 86.4, 81.1; STERNBERG, 81.2, 84.3; JAAP, 76.3, 79.2; ROST, 75.1, 77.3; MURZIN, 72.2, 75.3, from coordinate model; MUNSON, 63.3, 71.2; HUBBARD (2), 53.4, 66.4.

the amino side group of the a loop just after strand 2; in the GyrB-ADPNP complex, the Asp73 side chain interacts with the amino side group of the adenine ring. Tyr109 H bonds to the N3 atom of the adenine ring; while HSP90 has no equivalent Tyr at that position, there is a totally conserved Lys98 residue

```
   core
RKKMGLLVMAYGTPYKEEDIERYYTHIRRGRKPEPEMLQDLKDRYEAIGGISPLAQITEQ    sequence
 EEEEEEEEE      HHHHHHHHHHH       HHHHHHHHHHHHH    HHHHHHHHH    experimental
    EEEEEE      HHHHHHHHHHHHH     HHHHHHHHHHHHHH    HHHHHHHH    ROST
   EEEEEEE       HHHHHHHHHH        HHHHHH            HHHHHHH    STERNBERG
HHHHEEEEEE        HHHHHEEHHHHH  HHHHHHHHH    EEE       HHHHH    JAAP
HH   EEEEE      HHHHHHHHHHHHH     HHHHHHHHHHHHHH     HHHHHHH    GOLDSTEIN
    EEEEEE       HHHHHHHHHHHHH  HHHHHHHH     EEE    HHHHHHHH    PREDICTPROTEIN
      HHHHH           EEEEE         HHHHHHHHHHH       HHHHHH    SHESTOPALOV (2)
HHHHEEEEEEE         EEEEEEEEE      HHHHH               HHHHH    SERVER_SSPRED
HHHHHEEEEE     HHHHHHHHH EE       HHHHHHHHHHHEE        EHHHH    SERVER_GOR
     EEEE          HHHHHEH         HHHH     HEE        HHHHH    SERVER_NNPREDICT
    EEEEEE       HHHHHHHHHHHHHHH HHHHHHHHHHHH        HHHHHHH    SERVER_NNSSP_MULT
    EEEEEE      HHHHHHHHHHHH           HHHHHHHHH      HHH HH    HUBBARD
     EEEEEEE     HHHHHHHHHHHHHHH HHHHHHHHHHHHHHH     HHHHHHH    SERVER_SSP_MULT
    EEEEEE        HHHHHHHHHHHH    HHHHH     EEEE     HHHHHHH    SERVER_DSC_MULT
    EEEEEE      HHHHHHHHHHHHHHH       HHHHHHHHHH          HH    SOLOVYEV
  HHHHHHHHHH   HHHHHHHHHHHHHH      HHHHHHHHHHHHHH   HHHHHHHH    COHEN (COBEGETJ)
   HHHHHHHHHHHHHHH    EEEEEEEEE    HHHHHHHHHHHHHHHHHHHHHHHH     SMITH
-         -- -----      -   ------           --- - - HHHHHHH    BAKER


                  edge                       core
QAHNLEQHLNEIQDEITFKAYIGLKHIEPFIEDAVAEMHKDGITEAVSIVLAPHFSTFSV    sequence
HHHHHHHHHHHH   EEEEEEEEE     EHHHHHHHHHH    EEEEEE         H    experimental
HHHHHHHHHHHH      EEEEEEEE    HHHHHHHHHHH  EEEEEEE   EEEEE      ROST
HHHHHHHHHHH       EEEEEEE     HHHHHHHHHHHHHHHEEEEE     EEE      STERNBERG
HHHHHHHHHHHHHHH EEEEEEEEEE   HHHHHHHHHH    HHHHEEE     EEEE     JAAP
HHHHHHHHHHHHHHHH HHEEE     HHHHHHHHHHHHH     EEEEEE      EEH    GOLDSTEIN
HHHHHHHHHHHH    EEEEEEEEEEE   HHHHHHHHH    EEEEEEEE  EEEEE      PREDICTPROTEIN
HHHHHHHHHHHHHHHEEEEEE         HHHHHHHH        EEEEE   EEEEEE    SHESTOPALOV (2)
HHHHHHHHHHHHHHHHHEEEEEEEEEEEE  HHHHHHHH     EEEEEE     EEEEE    SERVER_SSPRED
H  HHHH HHHHHHHHHHHHHHH     HHHHHHHHHHHHHHHHHHHHEEEE    EE      SERVER_GOR
HHHHHHHHHH  HHHHHHHHHH       HHHHHHHHH      HHHEEEE             SERVER_NNPREDICT
HHHHHHHHHHHHHHH   EEEEEEHHHHHHHHHHHHHHHH    EEEEEE      EE      SERVER_NNSSP_MULT
HHHHHHHHHHH       EEEEEEEE    HHHHHHHHHHH  EEEEEEE     EE       HUBBARD
HHHHHHHHHHHH           HHHHHHHHHHHHHHHHH  EEEEEEEEE   EEEEEE    SERVER_SSP_MULT
HHHHHHHHHHH        EEE        HHHHHHHHHH      EEEEE     EEE     SERVER_DSC_MULT
HHHHHHHHHHH        EEE         HHHHHHHHHH   EEEEE               SOLOVYEV
HHHHHHHHHHHH      EEEEEE     HHHHHHHHHHHH    EEEE               COHEN (COBEGETJ)
   EEEEEEEEE    HHHHHHHHHHHHHHHHHHHHHH    EEEEEEEE     HHHHH    SMITH
HHHHHHH      -           -- HHHHHH       -  -              -    BAKER


                 edge                                  core
QSYNKRAKEEAEKLGGLTITSVESWYDEPKFVTYWVDRVKETYASMPEDERENAMLIVSA    sequence
HHHHHHHHHHHHHH   EEEE       HHHHHHHHHHHHHHHHH  HHHHH   EEEEE    experimental
    EE            EEEEE     HHHHHHHHHHHHHHHHH      EEEEEEEE     ROST
                  EEE          EEEHHHHHHH            EEEEEE     STERNBERG
   HHHHHHHHHH     EEEEEEE     EEEHHHHHHHHHHHHH      EEEEEEE     JAAP
HHHHHHHHHHHHHH   EEEEEE       EEEEEHHHHHHHH    HHHHHHHHHEHH     GOLDSTEIN
  HHHHHHHHHHH     EEEEE    HHHHHHHHHHHHHHHHHHH      EEEEEEE     PREDICTPROTEIN
EEHHHHHHHHHHH    EEEEEEEEE    EEEEEE   HHHHHH    HHHHHEEEEEE    SHESTOPALOV (2)
E   HHHHHHHHHHH  EEEEEEEE      EEEEEEHHHHHHHH        EEEEEEE    SERVER_SSPRED
     HHHHHHHHHHH EEEEE         EEEHH HE       HHHHHHHHHHHHHH    SERVER_GOR
      HHHHHHHH   EEEE          EEEHH H HH         HHHHHHHE      SERVER_NNPREDICT
  HHHHHHHHHHH    EEEE         HHHHHHHHHHHHHHH  HHHHHH EEEEE     SERVER_NNSSP_MULT
 HHHHHHHHHHH      EEEEE     HHHHHHHHHHHHHHHHH       EEEEEEEE    HUBBARD
E  HHHHHHHHHHHH EEEEEEE    HHHHHHHHHHHHHHHH    HHHHHHHHH        SERVER_SSP_MULT
    HHHHHHHHH    EEEEE         EEEHHHH              HHEEEEE     SERVER_DSC_MULT
HHHHHHHHHHH        HHHH      HHHHHHHHHHHHHHHH       EEEEEE      SOLOVYEV
     HHHHHHHHH    EEEEEEE   HHHHHHHHHHHHH             EEEE      COHEN (COBEGETJ)
HHHHHHHHHHHHHH   EEEEEEEE    HHHHHHHHHHHHHHHHHHHH    EEEEEEE    SMITH
-- - -----------  HHHHHHH -- HHHHHHHHHHHHH      -          -    BAKER
                                             note shift in hera plot
```

```
                                  edge
HSLPEKIKEFGDPYPDQLHESAKLIAEGAGVSEYAVGWQSE                       sequence
E   HHHHHH   HHHHHHHHHHHHHHH     EEEEEE                         experimental
     EE       HHHHHHHHHHHHHHH     EEEEEE                        ROST
     EEE        HHHHHHHHHHHHH     EEEEE                         STERNBERG
     HHHH       HHHHHHHHHHH       EEEEE                         JAAP
   HHHHHH     HHHHHHHHHHHHH     HEEE                            GOLDSTEIN
    HHHH      HHHHHHHHHHHHH       EEEEE                         PREDICTPROTEIN
HHHHHHHHH        HHHHHHHHH    EEEEEEEE                          SHESTOPALOV (2)
                 HHHHHHHHHHHH   EEEEE                           SERVER_SSPRED
H HHHHHE       HHHHHHHHHHHHH    EEE                             SERVER_GOR
    HHH          HHHHHHHHHH     EEEE                            SERVER_NNPREDICT
    HHHHH     HHHHHHHHHHHHH       EEE                           SERVER_NNSSP_MULT
               HHHHHHHHHHHHHH    EEEEEE                         HUBBARD
               HHHHHHHHHHHH                                     SERVER_SSP_MULT
                HHHHHHHHHHH      EEEEEE                         SERVER_DSC_MULT
     EEE      HHHHHHHHHHHHHHH     EEEE                          SOLOVYEV
               HHHHHHHHHHHHHHH   EEEEE                          COHEN (COBEGETJ)
EE    HHHHHHHHHHHHHHHHHHHHH    EEEEEEE                          SMITH
 HHHHHHHHHH      HHH - -  --- HHH -- -                          BAKER
```

Figure 59. Sequence and predictions from the CASP2 site and experimental secondary structure for ferrochelatase, *Bacillus subtilis* (320 residues),[341] T0020, 1ak1, HEMZ_BACSU, P32396. Experimental secondary structural assignments calculated with DSSP. Key: E, β strand; H, α helix. The number in parentheses (*n*) indicates the prediction was a weighted average of *n* predictions. The prediction with the highest S_{ov}–O is shown. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: SERVER_NNSSP_MULT, 86.0, 80.1; GOLDSTEIN, 84.6, 76.0; SERVER_PRREDICTPROTEIN, 82.3, 75.8; SOLOVYEV, 81.2, 71.9; COHEN, 79.9, 73.1; HUBBARD, 78.4, 78.0; JAAP, 78.4, 66.5; ROST, 75.7, 73.8; SERVER_DSC_MULT, 75.5, 67.0; SHESTOPALOV (2), 74.4, 65.9; STERNBERG, 70.9, 67.4; SERVER_SSPRED, 69.8, 60.2; SERVER_SSP_MULT, 68.1, 66.5; SMITH (fold recognition), 58.6, 58.8; SERVER_NNPREDICT, 56.3, 61.5; SERVER_GOR, 51.4, 59.3; BAKER from coordinate model (fold recognition), 44.3, 42.0.

that could play a similar role. The phosphates in ADPNP rest against a Gly motif in GyrB (Glys114, -117, and -119); in HSP90, the equivalent Gly residues were proposed to lie at positions 118, 121, and 123.

As with the COBEGETJ team, Bazan drew from his models the conclusion that HSP90 must bind to ATP.

As an example where an *ab initio* prediction generated a secondary structural model that was sufficiently accurate to support a tertiary structural model, and that the tertiary structural model was useful for detecting long-distance homology and solving a problem concerning biological function, this prediction was especially significant. As Dunbrack *et al*. noted, the fact that two groups independently reached the same conclusions indicates that the problem was approached systematically, making it probable that similar procedures can be implemented in automated systems in the future.[130]

6. Procaricain (T0012)

With 17 homologs and a family that had undergone evolutionary divergence of 120 PAM units, procaricain was an excellent target for an evolution-based prediction. Accordingly, predictions obtained by the STERNBERG and ROST groups scored highly. Another factor contributing to the high quality of this prediction was undoubtedly the fact that the protein is entirely helical; empirically, these tools seem to work well with proteins built from a single type of secondary structural elements.

Virtually all of the secondary structure predictions identified the three core helices correctly. The SOLOVYEV group, although it achieved a relatively low three-state score, also identified three helices, with the third significantly shifted. Figure 56 collects the secondary structure predictions submitted for the CASP2 project for procaricain. This protein was also identified by the contest organizers as one that had a homologous sequence with known 3D structure. Residues 49−107 were shown to have homology to the proregion of cathepsin B (rat and human). Accordingly, the scores were calculated for the non-homologous part, the first two helices.

7. 3-Dehydroquinase (T0014)

With only six homologs and an evolutionary tree spanning ~200 PAM units, the dehydroquinase target was marginal for an evolution-based structure prediction. Given this fact, the predictions produced by the JAAP, ROST, SOLOVYEV, and MUNSON teams are quite impressive. Figure 57 collects the secondary structure predictions submitted for the CASP2 project for 3-dehydroquinase. As the coordinates for the protein are not yet in the public domain, we cannot assess the significance of the misprediction of one strand in the first line of Figure 57, and the overprediction of strands in the carboxy-terminal segment of the protein.

8. Peridinin Chlorophyll Protein (T0016)

Peridinin chlorophyll protein is an all-helical protein. The family contains four members with only 15 PAM units of sequence divergence overall. This would normally not be sufficient divergence to gain the advantage that evolution-based predictions offer over those based on a single sequence. Nevertheless, both the SOLOVYEV and STERNBERG groups gave excellent S_{ov}–O scores. Figure 58 collects the secondary structure predictions submitted for the CASP2 project for the peridinin chlorophyll protein family.

										Manual		Auto	
Pos		l	ij	k	efg	h	bcd	a	SIAPred	SB	DLG	MT	rec
001	55	p				-	AAV		i				
002	56	t			qss	n	QQQ		s				
003	57	v			ssh	a	PPP		s				
004	58	r		m	qhv	Q	GEQ		s				
005	59	s	mm	t	aav	K	NRK	m	s				
006	60	g	rk	k	vva	R	RRR	s	s				
007	61	q	qq	p	dee	S	KKK	r	S				
008	62	k	ts	a	ded	P	PPP	k	s				
009	63	R	KK	K	KKK	_	RKK	k	S				
010	64	V	TL	I	VVI	T	TTT	m	s	e	h		e
011	65	G	GG	G	GGG	G	GGG	g	i	e	h	E	e
012	66	V	IV	V	VVV	I	III	l	i	E	eh	E	E
013	67	L	LL	L	LLL	V	LLL	l	i	E	eh	E	E
014	68	L	LM	L	LLL	L	MMM	v	i	E	eh	E	E
015	69	V	AV	A	LLL	M	LLL	m	i	E	eh	E	E
016	70	N	NN	N	NNN	N	NNN	a	A	e	eh		e
017	71	L	LL	L	LLL	M	MMM	y	i	eh			
018	72	G	GG	G	GGG	G	GGG	g	i	e			
019	73	T	TT	T	GGG	G	GGG	t	s				
020	74	P	PP	P	PPP	P	PPP	p	i				
021	75	D	DD	D	EEE	S	EEE		S				
022	76	T	AA	S	___	_	___		s				
023	77	A	PP	P	___	_	___	y	.				
024	78	D	TT	T	TTT	K	TTT	k	S				
025	79	A	PP	P	LLL	V	VLL	e	i		h	H	h
026	80	P	EQ	K	DNN	E	EGG	e	S	h	H	H	h
027	81	G	AA	S	DDD	E	EED	d	S	H	H	H	H
028	82	V	VV	I	VVV	T	VVV	i	i	H	H	H	H
029	83	R	KK	S	QQQ	Y	QQH	e	S	H	H	H	H
030	84	V	RR	R	PPP	D	DDD	r	S	H	H	H	H
031	85	Y	YY	Y	FFF	F	FFF	y	i	H	H	H	H
032	86	L	LL	L	LLL	L	LLL	y		H	H	H	H
033	87	K	KA	W	YFY	Y	QQL	t	S	H	H	H	H
034	88	E	QE	Q	NNN	Q	RRR	h	S	H	H	H	H
035	89	F	FF	F	LLL	L	LLL	i	I	H	H	H	H
036	90	L	LL	L	FFF	F	FFF	r	i	H	H	H	H
037	91	S	SS	T	AAA	A	LLL	r		h	H	H	H
038	92	D	DD	D	DDD	D	DDD	g	S		h	H	h
039	93	A	RR	P	PPP	N	QRQ	r	S		he		
040	94	R	RR	R	DDD	D	DDD	k	S		he		
041	95	V	VV	V	III	L	LLL		I	e	he		e
042	96	I	VV	V	III	I	MMM		I	e	he		e
043	97	E	DD	D	RRR	_	TTT		S		he		
044	98	D	TT	L	LLL	_	LLL		s		he		
045	99	Q	SS	P	PPP	P	PPP		s		he		
046	100	G	RP	R	RRR	I	VII		s				
047	101	L	LW	C	LLP	S	___		s				
048	102	V	LL	K	FFF	A	___		s				
049	103	W	WW	W	RRQ	K	___		S				
050	104	K	WW	Y	FFF	Y	___		s				
051	105	_	__	_	___	_	___		i				
052	106	_	__	_	___	_	___		s				
053	107	_	__	_	___	_	___		s				
054	108	V	PP	P	___	_	___						
055	109	V	LL	L	LLL	_	___		i			H	
056	110	L	LL	L	QQQ	Q	QQQ		i			H	
057	111	N	RR	K	ERG	K	DNN		S			H	
058	112	G	GG	A	PPT	T	KKK		S			H	
059	113	I	VV	I	LLI	I	LLL		e			H	
060	114	I	II	I	AAA	A	GAA		s	e		H	
061	115	L	FL	L	KKK	K	PPP		S	e		H	

062	116	R	PP	P	___	_	___		s				
063	117	Q	LI	L	LLF	Y	FFF		s			H	
064	118	R	RR	R	III	I	III	p	s			H	
065	119	P	SS	S	SSS	A	AAA	e	s			H	
066	120	R	PP	K	TTV	K	KKK	p	S			H	
067	121	S	RR	R	YFV	F	RRR	e	S	H	h	H	
068	122	K	VV	I	RRR	R	RRR	m	S	H	h	H	
069	123	A	AA	A	AAA	T	TTT	l	i	H	h	H	
070	124	L	KK	K	PPP	P	PPP	q	s	H	h	H	
071	125	D	LL	N	KKK	K	KKK	d	S	H	h	H	
072	126	Y	YY	Y	SSS	I	III	L	i	H	h	H	
073	127	Q	AQ	Q	KNK	E	QQQ	K	S	H	h	H	
074	128	K	SS	A	EEE	K	EEE	D	S	H	h	H	
075	129	I	VV	I	GGG	Q	Q_Q	R	S	H	h		
076	130	W	WW	W	YYY	Y	Y_Y	Y	i	h			
077	131	N	MM	T	AAA	R	RRR	E	s				
078	132	N	ED	E	SSA	E	RRR	A	S				
079	133	E	GE	Q	III	I	III	I	s				
080	134	K	__	_	GGG	G	GGG	G	s				
081	135	N	__	_	GGG	G	GGG	G	s				
082	136	E	GG	G	GGG	G	GGG	I	s				
083	137	S	SS	S	SSS	S	SSS	S	A				
084	138	P	PP	P	PPP	P	PPP	P	.	H			
085	139	L	LL	L	LLL	I	III	L	I	H	H		H
086	140	K	ML	L	RRR	R	KKK	A	S	H	H		H
087	141	T	VV	A	KKK	K	MMI	Q	S	H	H		H
088	142	I	YY	I	III	W	WWW	I	I	H	H		H
089	143	T	SS	S	TTT	S	TTT	T	s	H	H		H
090	144	R	RR	R	DDD	E	SSS	E	S	H	H		H
091	145	S	QR	Q	EEE	Y	KKK	Q	S	H	H		H
092	146	Q	QQ	Q	QQQ	Q	QQQ	Q	A	H	H		H
093	147	S	QQ	K	AAA	A	GGG	A	s	H	H	H	H
094	148	A	QK	D	QND	T	EEE	H	S	H	H	H	H
095	149	K	AA	A	AAA	E	GGG	N	S	H	H	H	H
096	150	L	LL	L	LLI	V	MMM	L	I	H	H	H	H
097	151	A	AA	Q	KKK	C	VVV	E	s	H	H	H	H
098	152	A	QE	A	MVM	K	KKK	Q	s	H	H	H	H
099	153	A	RR	Y	AAS	I	LLL	H	s	H	H	H	H
100	154	L	LM	L	LLL	L	LLL	L	I	H	H	H	H
101	155	S	PP	D	AKQ	D	DDD	N	S	H	H	H	H
102	156	D	EE	N	ESA	K	EEE	E	S	H	H	H	H
103	157	R	MI	Q	KKK	T	LLL	I	s	H		H	
104	158	D	__	N	NNN	C	SSS	Q	s				
105	159	_	__	I	MLI	P	PPP	D	s				
106	160	H	__	D	SEA	E	HAN	E	S				
107	161	V	__	T	TAA	T	TTT	I	s				
108	162	_	__	_	___	A	AAA	T	s				
109	163	_	__	_	___	P	PPP	_	i				
110	164	_	__	_	___	H	HHH	F	s				
111	165	V	PP	Q	NDN	K	KKK	K	S	e	H		E
112	166	V	VV	V	VIV	P	YYY	A	i	E	H		E
113	167	D	AE	E	YYY	Y	YYY	Y	s	E	H		E
114	168	W	LL	I	VVV	V	III	I	I	E	H		E
115	169	A	GG	A	GGG	A	GGG	G	.	E	H		E
116	170	M	MM	M	MMM	F	FFF	L	I	E	H		E
117	171	R	SS	T	RRR	R	RRR	K	S	E	H		E
118	172	Y	YY	Y	YYY	Y	YYY	H	i	E	H		E
119	173	G	GG	G	WWW	A	VVV	I	i	e	H		
120	174	N	SS	N	YYY	K	HHH	E	S	H	H		
121	175	P	PP	P	PPP	P	PPP	P	i	H	H		
122	176	S	SN	S	FFF	L	LLL	F	s		H		
123	177	I	LL	M	TTT	T	TTT	I	I	H	H	H	H
124	178	K	EP	Q	EEE	A	EEE	E	S	H	H	H	H
125	179	S	SD	S	EEE	E	EEE	D	S	H	H	H	H

126	180	G	AA A	AAA	T AAA	A	.	H	H	H	H
127	181	I	VI V	IIV	Y III	V	I	H	H	H	H
128	182	D	DD K	QDQ	K EEE	A	S	H	H	H	H
129	183	A	EK N	QQQ	Q EEE	E	S	H	H	H	H
130	184	L	LL L	III	M MMM	M	I	H	H	H	H
131	185	I	LL L	KKK	L EEE	H	S	H	H	H	H
132	186	G	AA K	RKK	K RRR	K	S	H		H	H
133	187	G	EQ N	DDD	D DDD	D	S	H		H	
134	188	M	HG Q	GKK	G GGG	G	S		H	H	
135	189	R	VV V	III	V LLL	I			H	H	
136	190	P	DT E	TTT	K EEE	T	S		H	H	
137	191	H	HK R	RKR	K RRR	E	S		H	H	
138	192	L	IL I	LLL	A AAA	A	I	E	He	H	e
139	193	A	VV I	VVV	V VII	V	I	E	He		e
140	194	V	VV V	VVV	A AAA	S	i	E	e		e
141	195	_	LL L	LLL	F FFF	I	i		e		
142	196	P	PP P	PPP	S TTT	V	s		e		
143	197	L	LL L	LLL	Q QQQ	L	i	Eh	he		
144	198	Y	YY Y	YYY	Y YYY	A	I	Eh	he		
145	199	P	PP P	PPP	P PPP	P	i	Eh	he	H	
146	200	Q	QQ Q	QQQ	H QQQ	H	i	Eh	he	H	
147	201	Y	FY Y	YYY	F YYY	F	I	Eh	he	H	
148	202	S	SS S	SSS	S SSS	S	A	Eh	he	H	
149	203	A	CC S	III	Y CCC	T	i	Eh	he	H	
150	204	S	SS S	SSS	S SSS	F	.	Eh	he	H	
151	205	T	TT T	TTT	T TTT	S		Eh	he	H	
152	206	S	VS T	TST	T TTT	V	s	E	h	H	
153	207	A	GA G	GGG	G GGG	Q	s	E	h		
154	208	T	AA A	SSS	S SSS	S	s	E	h		
155	209	V	VV V	SSS	S SSS	Y	.	h			
156	210	C	WW F	III	I LLL	N	.				
157	211	D	DD D	RRR	N NNN	K	S	h			H
158	212	_	__ _	VVV	E AAA	_	s	H	H	H	H
159	213	_	__ _	LLL	L III	_	i	H	H	H	H
160	214	_	__ _	QQQ	W YYY	_		H	H	H	H
161	215	E	EA A	KND	R RRR	R	S	H	H	H	H
162	216	V	LV F	MIL	Q YYY	A	i	H	H	H	H
163	217	F	AA A	FVF	I YYY	K		H	H	H	H
164	218	R	RR N	RKR	K NNN	E	S	H	H	H	H
165	219	V	II A	EEK	A EEQ	E	S	H	H	H	H
166	220	L	LL L	DDD	L VVV	A	.	H	h	H	H
167	221	A	AK K	APP	D GGG	E	S	H	h	H	H
168	222	R	RG E	YYY	S RQR	K	S		h	H	h
169	223	L	KY E	LFL	E KKK	L	S		h	H	h
170	224	R	RR R	SAA	R PPP	G	S			H	h
171	225	A	SR G	SGG	S TTT	G	S			H	
172	226	Q	IL L	LLV	I MMM	L	i			H	
173	227	P	PP L	PPP	S KKK	T	S			H	
174	228	T	GS P	___	_ ___	_	s				
175	231	L	II F	VIV	W WWW	I	I		H	e	h
176	232	R	SS D	SSA	S SSS	T	s		H	e	h
177	233	V	FF F	III	V TTT	S	s(i)		H	e	h
178	234	T	II I	III	I III	V	i		H	e	h
179	235	P	RR H	KEK	D DDD	E	S		H		h
180	236	P	DD S	SSS	R RRR	S	S		H		h
181	237	Y	YY Y	WWW	W WWW	W	I		H		H
182	238	Y	AA H	YYY	P PPP	Y	(i)		H		H
183	239	E	DE I	QQQ	T TTT	D	s		H		H
184	240	D	NH D	RRR	N HHH	E	S		H		H
185	241	E	HP E	EER	E PPH	P	S	H	h		H
186	242	A	DA N	GGG	G LLL	K	S	H	h	H	H
187	243	Y	YY Y	YYY	L LLL	F	I	H	H	H	H
188	244	I	II I	IVV	I III	V	I	H	H	H	H
189	245	E	NS N	KKN	K QQQ	T	S	H	H	H	H

190	246	A	AA A	SSS	A CCC	Y	i	H	H	H	H
191	247	L	LL L	MMM	F FFF	W	I	H	H	H	H
192	248	A	AK A	AAA	S AAA	V	H	H	H	H	
193	249	V	NQ D	DDD	E DDD	D	S	H	H	H	H
194	250	S	SS S	LLL	N HHH	R	S	H	H	H	H
195	251	I	VV I	MII	I III	V	I	H	H	H	H
196	252	E	RE K	QEE	T LLL	K	S	H	H	H	H
197	253	T	AN _	AKK	K KKK	E	S	H	H	H	H
198	254	H	SS _	EEE	K EEE	T	S			H	h
199	255	L	FF _	LLL	L LLL	Y	I			H	h
200	256	A	AV V	KSQ	Q DND	A	S			H	h
201	257	T	KQ R	NVT	E HHH	S	S				h
202	258	L	HH L	FFF	F FFF	M	i				
203	259	P	GG K	ASS	P PPP	P	s				
204	260	F	EK S	NND	Q PEL	E	S				
205	261	K	PP D	PPP	P EEE	D	S				
206	262	P	DD E	QEK	V KKK	E	S				
207	263	_	__ _	___	R RRR	R	s				
208	264	_	__ _	___	D RSS	E	S				
209	265	E	__ _	EEE	K EEE	N	S				
210	266	L	LR F	VVV	V VVV	A	i			E	
211	267	I	LL L	MMM	V VVV	M	I	E	h	E	
212	268	V	LV L	III	L III	L	I	E	h	E	
213	269	A	LL F	FFF	L LLL	I	I	E	he	E	
214	270	_	__ _	FFF	F FFF	V	i	E	he		
215	271	S	SS S	SSS	S SSS	S	A		he		
216	272	F	YF Y	AAA	A AAA	A	I		he		
217	273	H	HH H	HHH	H HHH	H	A		he		
218	274	G	GG G	GGG	S SSS	S			he		
219	275	M	II I	VVV	L LLL	L	I		he		
220	276	P	PP P	PPP	P PPP	P	i		he		
221	277	K	QK L	VLV	M MMM	E	s			e	
222	278	S	RR R	STS	D SSS	K	S			e	
223	279	Y	YY Y	YYY	V VVV	I	i			e	
224	280	V	AA E	VVV	V VVV	K	I			e	
225	281	D	DQ K	EKE	N NNN	E	S				
226	282	_	__ _	NDN	_ ___	_	S				
227	283	K	EL M	AAA	T RRR	F	S				
228	284	G	GG G	GGG	G GGG	G	i				
229	285	D	DD D	DDD	D DDD	D	s				
230	286	P	DD Y	PPP	A PPP	P	s				
231	287	Y	YY Y	YYY	Y YYY	Y	i				
232	288	Q	PP R	KRQ	P PPP	P	s			H	
233	289	E	QQ E	DDK	A QQQ	D	S	H		H	
234	290	H	RR H	QQQ	E EEE	Q	s	H		H	H
235	291	C	CC C	MMM	V VVV	L	I	H		H	H
236	292	I	RE K	EEE	A GGS	H	S	H		H	H
237	293	A	TD Q	EDE	A AAA	E	S	H		H	H
238	294	T	TT T	CCC	T TTT	S	(i)	H	h	H	H
239	295	T	TS T	III	V VVV	A	s	H	h	H	H
240	296	E	RR I	CAD	Y QHQ	K	S	H	H	H	H
241	297	A	EA A	LLL	N RKK	L	S	H	H	H	H
242	298	L	LL V	III	I VVV	I	I	H	H	H	H
243	299	R	AR V	MMM	M MMM	A	s	H	H	H	H
244	300	A	SA N	QEE	Q DEE	E	S	H	H	H	H
245	301	A	AE K	EEE	K KKR	G	S	H	H	H	H
246	302	R	LI L	LLL	L LLL	A	(i)	H	H	H	H
247	303	R	GA G	KKK	K GGE	G	S	H	H	H	
248	304	_	__ _	ASA	_ ___	_		H	extra		
249	305	_	__ _	RRR	_ ___	_	s	H	turn		
250	306	_	__ _	GGG	_ ___	_	i	H			
251	307	L	ML L	ITV	_ ___	_	s	H			
252	308	D	AP T	GLL	F YYY	_	s	H			
253	309	A	PA E	NNN	K SPC	V	s	H		H	

254	310	S	EE N	EDD	N NNN	S	S	H		H	
255	311	K	KQ Q	___	P PPP	E	S			H	
256	312	L	VI W	HHH	Y YYY	Y	(i)			H	
257	313	L	MM R	TTK	R RRR	A	s			H	
258	314	L	MM M	LLL	L LLL	V	i		eh	H	
259	315	T	TT T	AAA	V VVV	G	s		eh	H	e
260	316	F	FY F	YYY	W WWW	W	I		eh	H	e
261	317	Q	QQ Q	QQQ	Q QQQ	Q	A	ac	eh	H	A
262	318	S	SS S	SSS	S SSS	S	A	si	eh	H	A
263	319	R	RR R	RRR	Q KKK	E	S		eh	H	
264	320	F	FF F	VVV	V VVV	G	i		eh	H	
265	321	G	GG G	GGG	G GGG	N	s			H	
266	322	_	__ _	___	_ ___	T					
267	323	N	RR R	PPP	P PPP	P	S				
268	324	D	EE E	VVV	K MVM	D	S			H	
269	325	E	PP E	QQQ	P PPP	P	s			H	
270	326	W	WW W	WWW	W WWW	W	i	E		H	e
271	327	L	LL L	LLL	L LLL	L	i	E		H	e
272	328	Q	MT Q	KKK	G GGG	G	s	h		H	
273	329	P	PP P	PPP	A PPP	P	.	H	h	H	
274	330	Y	YY Y	YYY	Q QQQ	D	s	H	H	H	H
275	331	T	TT T	TTT	T TTT	V	i	H	H	H	H
276	332	D	DD D	DDD	A DDD	Q	S	H	H	H	H
277	333	K	EE K	EEE	E EEE	D	S	H	H	H	H
278	334	T	TT F	VVV	I AAS	L	s	H	H	H	H
279	335	M	LL L	LLL	A III	T	i	H	H	H	H
280	336	E	KK E	VVV	E KKK	R	S	H	H	H	H
281	337	R	MS S	EED	F GGG	D	S	H	H	H	H
282	338	L	LL A	LLL	L LLL	L	I	H	H	H	H
283	339	A	GP A	GGG	G CCC	F	i		H	H	
284	346	K	ES A	QQK	_ KEE	E	S			H	
285	347	_	__ _	___	_ ___	Q			h		
286	348	E	KQ Q	KKS	P RRR	K	S		h		
287	349	G	GG N	GGG	K GGG	G	s	E	h		
288	350	V	VV I	IVV	V RRR	Y	S	E	h		
289	351	R	GK Q	KKK	D KKK	Q	S	E	h		
290	352	R	HH K	SSS	G NNN	A	S	E	h		
291	353	I	II I	LLL	L III	F	I	E	he	E	E
292	354	A	QQ A	LLL	M LLL	V	i	E	he	E	E
293	355	V	VL V	AAA	F LLL	Y	I	E	e	E	E
294	356	V	MI I	VVV	I VVV	V	I	E	e	E	E
295	357	T	CC C	PPP	P PPP	P	s	e		E	
296	358	P	PP P	VVV	I III	V	i				
297	359	G	GG G	SSS	A AAA	G	i		h		
298	360	F	FF F	FFF	F FFF	F	i	h	h		
299	361	A	AS S	VVV	T TTT	V	s	h	h		
300	362	A	AA V	SSS	S SSS	A	h	h		H	
301	363	D	DD D	EEE	D DDD	D	s	hh	H	H	
302	364	C	CC C	HHH	H HHH	H	i	hh	H	H	
303	365	L	LL L	III	I III	L	I	h	h	H	
304	366	E	EE E	EEE	E EEE	E	A	h	h	H	A
305	367	T	TT T	TTT	T TTT	V	i	h	h	H	
306	368	L	LL I	LLL	L LLL	L	I	h	h	H	
307	369	E	EE E	EEE	H YYY	Y	s	h	h	H	
308	370	E	EE E	EEE	E EEE	D	S	h	h	H	
309	371	I	II I	III	I LLL	N	h	h	H	H	
310	372	A	AK D	DDD	D DDD	D	S	h		H	
311	373	Q	EE E	MMM	L III	Y	s				
312	374	E	QQ E	EEE	_ EEE	_	s				
313	381	_	__ _	___	_ YYY	_	i				
314	382	N	NN N	YYY	_ SSS	E	s				
315	383	A	RR R	KRR	G QQQ	C	S				h
316	384	E	EE E	HEE	V VVV	K	s				H
317	385	I	VV N	LLL	I LLL	V	I	e	e	h	H

318	386	F	FF F	AAA	G AAA	V	i	e	e	h	H
319	387	K	LI L	LLL	E SQK	T	S		e	h	H
320	388	H	GH N	EEE	S EKE	D	S		e	h	H
321	389	N	AA N	SSS	E CCC	D	S			h	H
322	390	G	GG G	GGG	Y GGG	I	(i)			h	H
323	391	G	GG G	IIV	K LAV	G	s	h		h	H
324	392	E	KE Q	QEE	D EEE	A	S	h		h	H
325	393	T	KK S	NNN	K NNN	S	S	h	e	h	H
326	394	F	YF Y	WWW	F III	Y	I	h	e	h	H
327	395	S	EE Q	GGG	K RRR	Y	S	h	eh	h	H
328	396	A	YY Y	RRR	R RRR	R	S	h	eh	h	H
329	397	I	II I	VVV	C AAA	P	i	h	eh	h	H
330	398	P	PP P	PPP	E EEE	E	s		h	h	h
331	399	C	AA A	AAA	S SSS	M	i		eh	h	h
332	400	L	LL L	LLL	L LLL	P	i		eh	h	
333	401	N	NN N	NGG	N NNN	N	s		eh	h	
334	402	D	Ad V	CCL	G GGG	A	s		eh	h	
335	403	S	Td E	NTT	N NNN	K	s		e	h	
336	404	E	Pe H	SSP	Q PPP	P	S		h		
337	405	P	Eg A	SSS	T LLL	E	s		H		H
338	406	G	Hp H	FFF	F FFF	F	s		H		H
339	407	M	I I	III	I SSS	I	i		H	h	H
340	408	D	E E	SST	E KKK	D	S		H	h	H
341	409	V	M M	DDD	G AAA	A	.		H	h	H
342	410	I	M M	LLL	M LLL	L	i		H	h	H
343	411	R	A G	AAA	A AAA	A	.		H	h	H
344	412	T	N K	DDD	D DDD	T	S		H	h	H
345	413	L	L L	AAA	L LLL	V	i		H	h	H
346	414	V	V I	VVV	V VVV	V	i		H	h	H
347	415	L	a L	IVI	h hhh	L	s		h	h	H
348	416	R	a E	EEE	s sss	K	S		h	h	H
349	417	E	y K	AAS	h hhh	K	s			h	h
350	418	L	r L	LLL	l lii	L	i			h	h
351	419	q	t	PPP	q qqq		s				h
352	420	g		SSS	s sss						

Figure 60. Residue-by-residue consensus secondary structure prediction for the ferrocheletase family prepared using the transparent method. The SIA Predict records assignments of positions to the surface (S, s), interior (I, i), or near the "active site" (A, a). Automated assignments are given, with the output generated by DARWIN. Services of DARWIN are available by server to the user on the Web (URL http://cbrg.inf.ethz.ch/). Secondary structure is indicated by E (strong strand assignment), e (weak strand assignment), H (strong helix assignment), and h (weak helix assignment). Sequences, designated using single letters, are from the SwissProt database and Genbank, as below. Sequence "a" is the target sequence. The column marked "Auto" contains output from the fully automated secondary structure prediction tool (Marcel Turcotte's SAINT). The columns marked "Manual" contain assignments from semimanual analysis of the same data by two experts (Steven A. Benner and Dietlind Gerloff). Key: (a) (P32396) HEMZ_BACSU ferrochelatase (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Bacillus subtilis.* (b) (P22600) HEMZ_BOVIN ferrochelatase precursor (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase) (fragment). *Bos taurus* (bovine). (c) (P22315) HEMZ_MOUSE ferrochelatase precursor (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Mus musculus* (mouse). (d) (P22830) HEMZ_HUMAN ferrochelatase precursor (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Homo sapiens* (human). (e) (P42044) HEMZ_CUCSA ferrochelatase precursor (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Cucumis sativus* (cucumber). (f) (P42045) HEMZ_HORVU ferrochelatase precursor (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Hordeum vulgare* (barley). (g) (P42043) HEMZ_ARATH ferrochelatase, chloroplast precursor (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Arabidopsis thaliana* (mouse-ear cress). (h) (P16622) HEMZ_YEAST ferrochelatase precursor (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Saccharomyces cerevisiae* (bakers' yeast). (i) (P23871) HEMZ_ECOLI ferrochelatase (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Escherichia coli.* (j) (P43413) HEMZ_YEREN ferrochelatase (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Yersinia enterocolitica.* (k) (P43868) HEMZ_HAEIN ferrochelatase (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Haemophilus influenzae.* (l) (P28602) HEMZ_BRAJA ferrochelatase (EC 4.99.1.1) (protoheme ferro-lyase) (heme synthetase). *Bradyrhizobium japonicum.*

9. Ferrochelatase (T0020)

The ferrocheletase family contains 12 proteins with substantial evolutionary divergence, and is an excellent candidate for an evolution-based prediction. Accordingly, S_{ov}–O scores were high. Figure 59 collects the secondary structure predictions submitted for the CASP2 project for ferrocheletase. A transparent prediction (COBEGETJ) can be compared with a neural network prediction (ROST) with nearly identical Q_3 scores. Each has a serious mistake, where a helix in the experimental structure was mistaken for a strand in the model, or vice versa.

To understand the significance of this comparison, we must examine the multiple alignment in greater detail. This is reproduced in Figure 60, together with transparent predictions made by two experts (SB and

```
SLPKIGIRPVIDGRRMGVRESLEEQTMNMAKATAALLTEKLRHACGAAVECVISDTCIAG    sequence
   EEEEEEE       HHHHHHHHHHHHHHHHHHHHHHH E     E  EEE    E      experimental
              HHHHHHHHHHHHHHHHHHHHHHHHHHHH      EEEE            STERNBERG
       E          HHHHHHHHHHHHHHHHHHHHHHHHH    EEEEEEE   HHH    ROST
E    EEE         HHHHHHHHHHHHHHEHHHHHHHHHHH    EEEEEEE   HHH    JAAP
                  HHHHHHHHHHHHHHHHHHHHHHHH     EEEEEE    H      MUNSON
         EE      HHHHHHHHHHHHHHHHHHHHHHHHH     EEEEEE           SOLOVYEV
    E    EE       HHHHHHHHHHHHHHHHHHHHHHHHHH    EEEEE    HHH    GOLDSTEIN
               HHHHHHHHHHHHHHHHHHHHHHHHH    --       HHHHHHH    BAKER
                 HHHHHHHHHHHHHHHHHHHHHHHHHHHH   EEEEEE   HHH    BAZAN

MAEAAACEEKFSSQNVGLTITVTPCWCYGSETIDMDPTRPKAIWGFNGTERPGAVYLAAA    sequence
HHHHHHHHHHHHH  EEEEEEEE     HHHH        EEEEE       HHHHHHHH    experimental
HHHHHHHHHHHH     EEEEEEEEE              EEEEE        HHHHHHH    STERNBERG
HHHHHHHHHHHHH   EEEEEEE                    E       HHHHHHHHH    ROST
HHHHHHHHHHHH    EEEEEEE  EEE              EE         HHHHHHH    JAAP
HHHHHHHHHHH  EEEEEEEEEEE E     E        HEEE         HHHHHHH    MUNSON
HHHHHHHHHHHHH    EEEEE                   EEEE        HHHHHHH    SOLOVYEV
HHHHHHHHHH        EEEE                   EEE        HHHHHHHH    GOLDSTEIN
HHHHHHHHHH                  -                    HHHHHHHHHHH    BAKER
HHHHHHHHHHHH     EEEEEEEE                            HHHHHHH    BAZAN

LAAHSQKGIPAFSIYGHDVQDADDTSIPADVEEKLLRFARAGLAVASMKGKSYLSLGGVS    sequence
HHHHHH     EEE              HHHHHHHHHHHHHHHHHHHH   EEEEE        experimental
HHH       EEEE              HHHHHHHHHHHHHHHHHHEE   EEEEEEEEE    STERNBERG
HHHHHH     EEE             HHHHHHHHHHHHHHHHHHHHH   EEEEE  EE    ROST
HHHHHH     EEE    HHH      HHHHHHHHHHHHH  HHHHHH   EEEEE   E    JAAP
HHHHHH     EE               HHHHHHHHHHHHHHHHHHH    EEE    E    MUNSON
HHHH      EEEEE             HHHHHHHHHHHHH HHH       EEEE  EE    SOLOVYEV
HHHHHH    EEEEE             HHHHHHHHHHHHHHHHHHHH    EEE         GOLDSTEIN
H -              -    -   HHHHHHHHHHHHHHH                       BAKER
HHHHHH     EEEE             HHHHHHHHHHHHHHHHHH    EEEEEE        BAZAN

MGIAGSIVDHNFFESWLGMKVQAVDMTELRRRIDQKIYDEAELEMALAWADKNFRYGEDE    sequence
   HHH   HHHHHHH   EEEEE  HHHHHHHH     HHHHHHHHHHHHHH EE        experimental
EEE     HHHHHHHHH    HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH           STERNBERG
E EHHHHHHHHHHHHHH    EE  HHHHHHHHHHH   HHHHHHHHHHHHHHHH         ROST
E HHHHEHHHHHHHHHE   EEEE HHHHHHHHHHH   HHHHHHHHHHHHHHHH         JAAP
      E  HHHHHHHH    HHHHHHHHHHHH H    HHHHHHHHHHHHH            MUNSON
E        HHHHHHHH        HHHHHHHHHHHH  HHHHHHHHHHHHHHHH         SOLOVYEV
 EE   EE     HHHH  HHHHHHHHHHHHHHHHHH HHHHHHHHHHHHHHHHH   HH    GOLDSTEIN
         HHHHH      HHHHHHHHHHHHHHH    HHHHHHHHHHHH   HHHHHH    BAKER
                  HHHHHHHHHHHHHHHHHHHHH     EEEEE               BAZAN

NNKQYQRNAEQSRAVLRESLLMAMCIRDMMQGNSKLADIGRVEESLGYNAIAAGFQGQRH    sequence
  HHH    HHHHHHHHHHHHHHHHHHHHHH   HHHH   HHHH    EEEEEE         experimental
  HHHHHHH   HHHHHHHHHHHHHHHHHH     HHHHHHHHHHHHHHHHHHH          STERNBERG
    HHHHHHHHHHHHHHHHHHHHHHHHHH   HHHHHHHHHHHHH HHHHHH           ROST
    HHHHHHHHHHHHHHHHHHHH HHHHHH  HHHH   HHHHHH HHHHH            JAAP
  HH  HHHHHHHHHHHHHHHHHHE HHHH        H HHHHHHHHHHHHHHH    E    MUNSON
 HHHHHHHHHHHHHHHHHHHHHHHHHHHH       HHHHHHHHHH   EEE            SOLOVYEV
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH    HHHH  HHHHH HHHHHHHH   H     GOLDSTEIN
   -    HHHHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHH                   BAKER
        HHHHHHHHHHHHHHHHHHHHHHH                   EEEEE         BAZAN

WTDQYPNGDTAEAILNSSFDWNGVREPFVVATENDSLNGVAMLMGHQLTGTAQVFADVRT    sequence
        HHHHHHH   EE  EE     EEE   HHHHHHHHHHHHHH    EEEEEEE    experimental
         HHHHHH            EEEE     HHHHHHHHHH      EEEEEEEE    STERNBERG
        HHHHHHHHH          EEEEE    HHHHHHHHH      EEEEEHHHH    ROST
         HHHHHHHH          EEEEE    HHHHHHHH       EEEEEHHHE    JAAP
E        HHHHHE            EEEEE    HHHHEHHHHHHH  HHHEEH        MUNSON
         HHHHHHH           EEEE     HHHHHHHH       EEEEE        SOLOVYEV
          HHHHH            EEEE        HHHHH       HHEEEH       GOLDSTEIN
                                                   -            BAKER
         HHHHHHHH          EEEEE                   HHHHHHHHH    BAZAN
   -    HHHHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHH                   BAKER
        HHHHHHHHHHHHHHHHHHHHHHH                   EEEEE         BAZAN
```

```
YWSPEAIERVTGHKLDGLAEHGIIHLINSGSAALDGSCKQRDSEGNPTMKPHWEISQQEA   sequence
EE HHHHHHHH     HHHHH EEEEE      HHHH  EE     EE   HHH  HHHH   experimental
     HHHHH            EEEE                             HHHHH   STERNBERG
   HHHHHHH           EEEEEE                           HHHHHH   ROST
E  HHHHHHEE           EEEEE                           HHHHHH   JAAP
             HH HHHHHHHEEHH   HHHHH                 HHHHHHHH   MUNSON
   HHHHHHHH     HHHH  EEEEE          HHHHH              HHHH   SOLOVYEV
   HHHHHHH        HH  EEEEE     H                   HHHHHHHH   GOLDSTEIN
   HHHHHHHHH  - HHHHH          HHHHHHHH --              HHHH   BAKER
HHHHHHHHHHH           EEEEEE                         HHHHHHH   BAZAN

DACLAATEWCPAIHEYFRGGGYSSRFLTEGGVPFTMTRVNIIKGLGPVLQIAEGWSVELP   sequence
HHHHH  EEEE          EEEE E     EEEEEEEEEE   EEEEEEEEEEE       experimental
HHHHHH         EE       EEE      EEEEEEEE      HHHHH           STERNBERG
HHHHHHH    HHHHHH    HHHHHHH      EEEEEHHH      EEEE           ROST
HHHHHHH    HHHHHE    HHHHHHH      EEEEEHHH      EEEE           JAAP
HHHH H     HHHEEE       EEEE     EEEEEEEEE     HHEEHH  EH      MUNSON
HHHHH     HHHHHHHH       HHH     EEEEEE        EEEE            SOLOVYEV
HHHHHHHHH HHHHHHH       EEE       EEEEEEEE     EEEE     H      GOLDSTEIN
HHH HHHHHHHHHHHHHH  - HHHHHHHH               -                 BAKER
HHHHHHHHHHHHHHHHHH              EEEEE             HHHHHHHHHH   BAZAN

KDVHDILNKRTNSTWPTTWFAPRLTGKGPFTDVYSVMANWGANHGVLTIGHVGADFITLA   sequence
HHHHHHHHHHH     EEEEEE     HHH  HHHHHH     EEEEEE   HHHHHHHH   experimental
HHHHHHH          EEEE          EEEEEEE      EEEEEE    HHHHHH   STERNBERG
HHHHHHHHH                       EHHHHHHH     EEEE   HHHHHHHH   ROST
HHHHHHHHH                       EEEHHHH      EEEE    HHHHHHH   JAAP
HHHHHH          EEEEEH        EEEEEHHHHH      EEH    HHHHHHH   MUNSON
HHHHHHHHHH                      HHHHHHH    EEEEE       HHHHH   SOLOVYEV
HHHHHHH          EE             EEEEHH       EEEE      HHHHH   GOLDSTEIN
 HHHHHHHHHHHHH                       -      -                  BAKER
HHHHHHHHHHHHHHH              HHHHHHHHH  HHHHH        HHHHHHH   BAZAN

SMLRIPVCMHNVEETKVYRPSAWAAHGMDIEGQDYRACQNYGPLYK                 sequence
HHH   EEE    HHH    HHHHHH   HHHHHHHHHHHH                      experimental
HH   EEE                          HHHHH                        STERNBERG
HHH  EEEEE           HHH           HHHH                        ROST
HHHEE  EEE     EE     HHH          HHH                         JAAP
HH    E             HHHH          HH         E                 MUNSON
HHH   EEEEE                       HHHH                         SOLOVYEV
HH    EEE           HHHHHH      HHHHHHH    HHH                 GOLDSTEIN
          -             --              HHHHH                  BAKER
H                  HHHHHH   HHHHHHHHHHHHH                      BAZAN
```

Figure 61. Sequence and predictions from the CASP2 site and experimental secondary structure for L-fucose isomerase, *E. coli* (591 residues), T0022,[342] pdb code 1fui. Experimental secondary structural assignments, calculated with DSSP, were taken from the CASP2 site. Key: E, β strand; H, α helix. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: GOLDSTEIN, 68.1, 69.3; ROST, 67.3, 71.8; SOLOVYEV, 66.8, 71.7; JAAP, 64.9, 69.6; STERNBERG, 63.2, 69.1; MUNSON, 62.8, 68.1; BAZAN, 50.2, 63.1; BAKER, 40.7, 55.1, from coordinate model.

DLG) and by an automated version of the transparent evolution-based analysis (MT) known by the acronym SAINT (Structure Assignment with INFormative Transparency). The difficulty that the transparent prediction has in identifying the first strand arises because of a difficult alignment in this segment. The SAINT tool is fully automatic. In addition to a prediction of the secondary structure, however, it generates an output which explains why the secondary structure prediction is made. Thus, it combines the facility of an automated tool with the informative nature of a transparent prediction. The correspondence between the manual and SAINT-generated predictions was quite good; indeed, the SAINT prediction correctly identified the first strand that the manual prediction misassigned.

10. L-Fucose Isomerase (T0022)

The fucose isomerase family contains only two identifiable proteins with an evolutionary divergence of only 40 PAM units. Thus, evolution-based methods are not expected to perform well in this protein. The S_{ov}–O and Q_3 scores are low, and no prediction does well in the C-terminal half of the protein. The predictions are all remarkably good on the amino terminal end of the protein. Figure 61 collects the secondary structure predictions submitted for the CASP2 project for L-fucose isomerase.

11. Protein g3 (T0030)

Target T0030 has only four homologs, which come as two pairs of proteins, the members of each pair being essentially identical in sequence. Thus, the

```
           edge    edge   core      core            core
ETVESCLAKPHTENSFTNVWKDDKTLDRYANYEGCLWNATGVVVCTGDETQCYGTWVPIGLAIPEN sequence
  HHHHHH   EEEEE   EEE    EEEEEE  EEEEEEEEEEEE    EEEEEEEEEE       DSSP
  HHHHHH    EEEE   EEE    EEEEEE  EEEEE   EEEE    EEEEE            STRIDE
   EEE        E    E       EEEE   EEEEEEEEEEE    EEEEEEEE   EEEE   ROST (2)
   EEEE       EEEEE        EEEE         EEEEEE    EEE   EEEEE      STERNBERG
   EEE         EE EE        EEE   EEEEEE EEEE    EEEEEEEE    EEE   JAAP
   EEE        EE   E       EEEE   EEEEEEEEEEE    EEEEEEE    EEEE   PREDICTPROTEIN
HHHHHHH                              EEEEEEEE     EEEEE            SSPRED
HHHHHHHH            H   E                EEEE          EEEEE       GOR
   HHH                   H   H     EE    EEEE      E   E           NNPREDICT
                         HHHHH          EEEEE     EEE              NNSSP_MULT
HHHHHHHHH               HHHHHHHHH     EEEEEEE                      SSP_MULT
   EEEE        EEEEE       EEE          EEEEE     EEEEEEEEEEEE     DSC_MULT
HHHHHHHHHHHHH            HHHHH           EEEE                      SHESTOPALOV
  EEEE        EEEEEE     EEEEEE  EEEEEE EEEEE     EEEEEEE    EEE   HUBBARD
HHHHHH          EEE      HHHHH           EEEE     EE               GOLDSTEIN
    EEEEEEE    EEEEE           EEEEE    EEEEEE  EEEEEEEE  HHHHHHHH ABAGYAN
  EEE          EEEEE     EEEEEE         EEEEE     EE        EE     SOLOVYEV
  EEE          EEEEE     HHHHHHHHHHHHHHH EEEEE   EEEE  EE   EEE    ROSE
                        HHHHH  HHHHHHHHH                           MOULT (2)
```

Figure 62. Sequence and predictions from the CASP2 site and experimental secondary structure for domain 1 of protein g3, filamentous phage fd (66 residues),[343] T0030, 1fgp, P03661, CDAA–BPFD. Experimental secondary structural assignments, calculated with DSSP and STRIDE, from the CASP2 site. Key: E, β strand; H, α helix. A number in parentheses (n) indicates the prediction was a weighted average of n predictions. For these predictions, the prediction with the highest S_{ov}–O is shown. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: ROST (2), 75.6, 66.2; SERVER–PREDICTPROTEIN, 74.4, 65.2; SERVER–DSC–MULT, 61.4, 59.1; HUBBARD, 59.4, 62.1; JAAP, 59.3, 60.6; STERNBERG, 58.3, 59.1; SERVER–SSPRED, 54.1, 60.6; SOLOVYEV, 53.3, 48.5; SERVER–GOR, 46.6, 54.5; ROSE, 40.9, 39.4; GOLDSTEIN, 40.5, 42.4; ABAGYAN, 39.0, 37.9; SHESTOPALOV, 36.1, 45.2; SERVER–SSP–MULT, 34.4, 47.0; SERVER–NNPRREDICT, 33.9, 51.5; SERVER–NNSSP–MULT, 32.9, 47.0; MOULT (2), 7.2, 29.8, from coordinate model.

family contains effectively only two sequences, and these are 140 PAM units divergent. The family is therefore not expected to give strong evolution-based predictions. Accordingly, the S_{ov}–O and Q_3 scores are lower than those obtained from families with more members. Figure 62 collects the secondary structure predictions submitted for the CASP2 project for the protein g3. All of the predictions that identify the strands correctly misassign the first helix as a strand. The ROST prediction correctly identifies the long strands, and underpredicts the length of the shorter edge strands, all expected for a consensus model. Although the ROST group did not attempt to build a tertiary structure from this protein, we suspect that the ROST prediction would have sustained a successful modeling attempt, as would the SERVER–PREDICTPROTEIN prediction.

12. Exfoliative Toxin A (T0031)

The family of proteins containing target T0031 contains only three members. Although these are widely divergent, evolution-based predictions are expected to be poor. In fact, the S_{ov}–O and Q_3 scores are quite poor. Figure 63 collects the secondary structure predictions submitted for the CASP2 project for the exfoliative toxin A. In most of the predictions, the Q_3 is dramatically greater than the S_{ov}–O score. This reflects the large number of fragments of

```
VSAEEIKKHEEKWNKYYGVNAFNLPKELFSKVDEKDRQKYPYNTIGNVFVKGQTSATGVL    sequence
  HHHHHHHHHHHHHHH   HHH     EEE    HHH   HHHEEEEEE   EEEEEEE    experimental
        HHHHHH                              E   EE EEEEEEEEE    ROST
HHHHHHHHHHHHH            HHHHHHH              EEEEE     EEEE    STERNBERG
 HHHHHHHHHHHHHH             EEE             EEEEEEEEE EEEEEE    PREDICTPROTEIN
HHHHHHHHHHHHHHHEEEE        HHHHHHH          EEEEEEEE    EEEE    SERVER_SSPRED
HHHHHHHHHHHH      EEEE  HHHHHHHHHHH        EEEEEEEE    EEEEE    SERVER_GOR
  HHHHHHHHHHH            HHHH                  EEE       EEE    SERVER_NNPREDICT
  HHHHHHHHHHHHHH         HHHHHHHHH            EEEE      EEEE    SERVER_NNSSP_MULT
  HHHHHHHHH            HHHHHHHHH           EEEEEEEE   EEEEEE    SERVER_SSP_MULT
  HHHHHHHHHHH            HHHHHHH              EEEEE     EEEE    SERVER_DSC_MULT
HHHHHHHHHHHH    EEEEEE          HHHHHH         EEE        EE    SHESTOPALOV
H HHHHHHHHHHHHH          HHHHHHHHHHHH       EEEEEE      EEEE    GOLDSTEIN
  HHHHHHHHHHHHHH         HHHHH              EEEEEEE     EEEE    JAAP
HHHHHHH                                 HHH             EEE     ABAGYAN
  HHH    H        EE                        EEEEEEE  EEEEEEE    MUNSON (2)
  HHHHHHHHHH                                EEEEEE       EEE    SOLOVYEV
---------------------------------------    EEE    EEE  EEEEE    MURZIN
 HHHHHHH            EEEEE     HHHH      EEEEEEEEEEEEHHHEEEEE    LENGAUER
```

```
IGKNTVLTNRHIAKFANGDPSKVSFRPSINTDDNGNTETPYGEYEVKEILQEPFGAGVDL    sequence
    EEEE HHHHHHH   HHHEEEEEE EE     EE     EEEEEEE              experimental
EE  EEE    HHHHH                          HHHHHHHH         E    ROST
E       HHHHHHH                            HHHHHHHH             STERNBERG
EEEEEE   HHHHHHH      EEE                 HHHHHHHHH             PREDICTPROTEIN
EEEEEEEEEEEHHHH                             HHHHHH       HHH    SERVER_SSPRED
EEE EEEEHHHHEEE      EEEEEE               HHHHHHHHHH    HHHH    SERVER_GOR
E     HHHHHHHH                             HHHHHHHH       HH    SERVER_NNPREDICT
E   EEE  HHHHHHH     EEE                     HHHHHH             SERVER_NNSSP_MULT
E   EEEEHHHHHHHHH     EEEE                 HHHHHHHHH     HHH    SERVER_SSP_MULT
E       HHHHHHHH      EEEE                 HHHHHHHH        H    SERVER_DSC_MULT
E        HHHHH        EEE                    HHHHH              SHESTOPALOV
E    E   HHHHHHH      EEE                  HHHHHHH       HHH    GOLDSTEIN
EEEEEEEE HHHHHH      EEEE                 HHHHHHHHH        H    JAAP
 EEEE                                    EEEEE            EE    ABAGYAN
EE  EEEE   E         EEEE            EEE EEEEEE                 MUNSON (2)
E   EEEEE HHHH        EEE                 EEEE    EEEE          SOLOVYEV
  EEEEEE HHHHHH        EEE                     EEEEEEEE   EE    MURZIN
EE     EEEEE          EEEEEEEEEEEEEEE  EEEE     HHHEEEEEEEEE    LENGAUER

ALIRLKPDQNGVSLGDKISPAKIGTSNDLKDGDKLELIGYPFDHKVNQMHRSEIELTTLS    sequence
EEEEE        HHH     EE  HHH     EEEEEE          EEEEEEEE HH    experimental
EEEEEE                           EEEEEE        EEEE   EEE       ROST
 EEE                              EEEE       HHHHH   EEEEE      STERNBERG
EEEEE                             EEEEE             EEEEE       PREDICTPROTEIN
HHHHH                              EEE         HHHHHHHHHHH      SERVER_SSPRED
HHEEE     EEEE   EEEEE  E      HHHEEEEE   H  HHHH   HHHHHH      SERVER_GOR
HHE                               HEEE       H       EHHHHH     SERVER_NNPREDICT
EEEE                              EEEEE              EEEE       SERVER_NNSSP_MULT
HHHHHH                                                          SERVER_SSP_MULT
HEEE                               EEE              EEEEEHHH    SERVER_DSC_MULT
           EEE                    EEE                           SHESTOPALOV
HEHE                              EEEE        HH    EEEEHHH     GOLDSTEIN
HHHHH                             EEEEE             HHHH        JAAP
EE                 EEEE  EEEE                                   ABAGYAN
EEEEE                            EEEEEE               EEEEEE    MUNSON (2)
EEEEE                             EEEEE              EEEEE      SOLOVYEV
EEEEE                   EE       EEEEE       EEEEEEEEE          MURZIN
EEE             EE  HHHHHHHHHHHHHHHH    EEE          HHH   E    LENGAUER

RGLRYYGFTVPGNSGSGIFNSNGELVGIHSSKVSHLDREHQINYGVGIGNYVKRIINEKNE    sequence
H EEEE    HHH   EEE    EEEEEEEEEEE      EEEEEEE  HHHHHHHHHH      experimental
  EEEEEEE              EEEEEE                    HHHHHHHHH       ROST
   EEEE          EEE    EEEEE                     HHHHHH         STERNBERG
   EEEEEEE             EEEEEE                    EEEEEEEE        PREDICTPROTEIN
  EEEEEE                EEEE    HHHHHHHHHEEEEEEEEEEEEEHHHHHH     SERVER_SSPRED
     EEEEE       EE     HEEHH      HHHH         EEEEEEEE         SERVER_GOR
H  EEE E         EE     EEEEE               E     HEEHEH         SERVER_NNPREDICT
    EEEEEE       EEE    EEEEE                     HHHHHHH        SERVER_NNSSP_MULT
  EEEEEEEE                               EEEEEE   EEEEEEE        SERVER_SSP_MULT
H   EEEEEE       EEE    EEEEE                     HHHHHH         SERVER_DSC_MULT
   EEEEE                 EEE       HHHHHHEEEEEEE  HHHHHHHH       SHESTOPALOV
   EEE E         EE    EEEEE                 EE   HHHHHHHHH      GOLDSTEIN
   EEEEEEE             EEEEE            EEE      EEEEEEEE        JAAP
  EEE           EEE    EEEEEE                         EEEEE      ABAGYAN
    EEEEE               EEEE            EEEEEEEEE                MUNSON (2)
                 EEE    EEEEE           EEE   HHHHHHHHHHHH       SOLOVYEV
 EEEE           EEE    EEEEEEEE             EEEEHHHHHHHH  EE     MURZIN
EEEE HHHHHH          HHHHHHHHH   HHHHH         HHHHHHH           LENGAUER
```

Figure 63. Sequence and predictions from the CASP2 site and experimental secondary structure for exfoliative toxin A, *Staphylococcus aureus*[344] (242 residues), T0031, P09331, ETA_STAAU. Experimental secondary structural assignments (DSSP) from the CASP2 site. STRIDE assignments were not available. Key: E, β strand; H, α helix. A number in parentheses (*n*) indicates the prediction was a weighted average of *n* predictions. For these predictions, the prediction with the highest S_{ov}–O is shown. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: MURZIN, from coordinate model, 61.8, 63.9; SOLOVYEV, 56.8, 65.6; SERVER_NNSSP_MULT, 55.2, 63.5; GOLDSTEIN, 55.0, 64.3; SERVER_DSC_MULT, 53.2, 61.0; SHESTOPALOV, 53.1, 58.4; SERVER_PREDICTPROTEIN, 48.5, 63.6; STERNBERG, 48.5, 57.3; MUNSON (2), 46.8, 57.9; ROST, 45.6, 62.2; JAAP, 45.3, 58.5; SERVER_GOR, 41.0, 41.9; SERVER_NNPREDICT, 40.1, 56.0; SERVER_SSPRED, 39.4, 50.6; SERVER_SSP_MULT, 36.9, 53.5; ABAGYAN (2), 33.5, 46.6; LENGAUER, 29.6, 34.4.

```
TACTATQQTAAYKTLVSILSDASFNQCSTDSGYSMLTAKALPTTAQYKLMCASTACNTMI    sequence
 E  HHHHHHHHHHHHHHH  HHHHHHHHHH            HHHHHHHHH HHHHHHH    DSSP
    HHHHHHHHHHHHHHH  HHHHHHHHHH            HHHHHHHHH HHHHHHH    STRIDE
           EEEEEEE               EEEEE       EEEEEE    HHHHH    STERNBERG
        HHHHHHHEEEEE            EEEEEEE     HHHHHHHHHHHHHHHH    ROST
       HHHHHHHHHEEE   EE E       HHHHHHH    HHHHHHHHHHHHHHHH    JAAP
HHHHHHHHHHHHHHHHHHH              HHHHHH     HHHHHHHHHHHHHHHH    GOLDSTEIN
     HHHHHHHHHHHHH       HHHHHHHH          HHHHHHHHHHHHHHHHH    MUNSON
         HHHHHHHHHH             EEE           EEEEEE            SOLOVYEV
    HHHHHHHHHHHHHH       HHHHHHHH          HHHHHHHHHHHHHH  H    VALENCIA

         hairpin hairpin
KKIVTLNPPNCDLTVPTSGLVLNVYSYANGFSNKCSSL                          sequence
HHHHHH     EEE      EE HHHHHHHHHHHHH                            DSSP
HHHHHH     EEE      EEEHHHHHHHHHHHHHH                           STRIDE
HHEEE      EEE     EEEEEEE                                      STERNBERG
HHHHH      EEE    EEEEEEEEE                                     ROST
HHHEEE      EE     EEEEEEEE                                     JAAP
HHEEE              EEEEEEEE                                     GOLDSTEIN
HHHHHHH  EEEEEE    EEEEEEHHHHHHHHHHHHH                          MUNSON
 EEEEE      EEE    EEEEEEE                                      SOLOVYEV
HHHHHHH   EEEEEE  EEEEEE HHHHHHHHHHHHH                          VALENCIA
```

Figure 64. Sequence and predictions from the CASP2 site and experimental secondary structure for β-cryptogein, fungus *Phytophthora cryptogea* (98 residues),[345] T0032, 1beo, P15570, ELIB_PHYCR. Experimental secondary structural assignments (DSSP and STRIDE) from the CASP2 site. Key: E, β strand; H, α helix. The MUNSON and VALENCIA predictions were based on published secondary structure assignments made using NMR data. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: MUNSON, 79.3, 79.6; VALENCIA, 75.7, 77.6; JAAP, 48.7, 54.1; ROST, 44.1, 53.1; GOLDSTEIN, 40.5, 55.1; SOLOVYEV, 38.5, 40.8; STERNBERG, 32.2, 37.8; BAKER, 18.4, 35.5.

secondary structure in the experimental assignment. The three-residue helices do not represent canonical helices, which require at least four residues to complete a standard turn of an α helix. As the coordinates are not yet available, it is not clear how critical these omissions and mispredictions are. This example represents one of the worst performances for the high scoring automated nontransparent tools, with several serious mistakes.

13. β-Cryptogein (T0032)

As noted above, a paper reporting NMR experiments that assigned secondary structure to cryptogein was published before the CASP2 contest began. Both MUNSON and VALENCIA used the experimental information in making their models, and stated so. This accounts for their high Q_3 scores. The other methods performed poorly on this protein. Lesk was puzzled by the fact that automated prediction methods that did so well (at least by the Q_3 score) on many of the predictions did so poorly on cryptogein. He considered the possibility that cryptogein might be difficult to predict because it contained multiple disulfide bonds.[174] Similar problems were not encountered, however, by these tools with other disulfide-containing proteins that were targets of the CASP2 contest.

From an evolutionary perspective, it is not surprising that the predictions are generally poor. Although the cryptogein family has 11 homologs, the most divergent pair is only 35 PAM units distant. The effect is that the prediction is little better than one made with a single sequence. As noted at many points in this review, evolutionary-based methods do not work well when applied to a family of proteins that have undergone little sequence divergence. Figure 64 collects the secondary structure predictions submitted for the CASP2 project for β-cryptogein.

14. The Calponin Homology Domain (T0037)

With 18 members having an evolutionary divergence of 150 PAM, the calponin homology domain was an excellent target for evolution-based structure prediction methods. Figure 65 collects the secondary structure predictions submitted for the CASP2 project for the calponin homology domain of β-spectrin. As before, the helices containing only three or four residues are not canonical and can be ignored in modeling the four-helix bundle that is at the core of the fold. Nevertheless, they depress the S_{ov}–O scores in several of the predictions, and provide an illustration of how low S_{ov}–O scores can be misleading about the true value of a prediction. For the core elements, most of the prediction tools (except that of ROSE) perform equally well except for the final helix, which proved difficult to identify for some of the tools.

15. CBDN1 (T0038)

The CBDN1 protein is an endoglucanase that is homologous to the protein macromomycin in its central segment. The protein fold is built entirely from β strands. If the homolog with known structure is excluded, the protein family contains only two members approximately 60 PAM units divergent. Figure 66 collects the secondary structure predictions submitted for the CASP2 project for the CBDN1 protein from *Cellulomonas fimi*. Several of the predictions are very good, ignoring an extra strand and a fusion of two strands.

16. NK-Lysin (T0042)

The NK-lysin family contains 20 homologs with good evolutionary divergences, and should give good

```
KSAKDALLLWCQMKTAGYPNVNIHNFTTSWRDGMAFNALIHKHRPDLIDFDKLKKSNAHY    sequence
 HHHHHHHHHHHHH                   HHHHHHHHHH HHH  HHH     HHH    experimental
  HHHHHHHHHHHHH      EE  HHHHHHHHHHHHHHHHH                HH    ROST
   HHHHHHHHHH            HHHHHHHHHHHHHHHH                HHH    STERNBERG
  HHHHHHHHHHHHH     EEE  HHHHHHHHHHHHHHHHH   HHHH HHH     HH    PREDICTPROTEIN
  HHHHHHHHHHH       EEEEEE       HHHHHHHHH                      SERVER_SSPRED
HHHHHHHHHHHHE       EEEE          HHHHHHH     HHHHHHHHH  H      SERVER_GOR
   HHHHHHHHH        EEEE         HHHHHHHH       HH HH      H    SERVER_NNPREDICT
  HHHHHHHHHHH       EEEE         HHHHHHHHH                HH    SERVER_NNSSP_MULT
    HHHHHHHHHH      EEEEEEE      HHHHHHHHHHH                    SERVER_SSP_MULT
   HHHHHHHHHH            HHHHHHHHHHHHHHHH     HHHHHHHH   HHH    SERVER_DSC_MULT
     HHHHHHHHH      EEE   EEEEE   HHHHHHHHH    EEE       HHH    SHESTOPALOV
HHHHHHHHHHHHHHH      EE          HHHHHHHHH       HHHHHHH HHH    GOLDSTEIN
  HHHHHHHHHHHHH     EEE     HHHHHHHHHHHHH     HHHHHHH     HH    JAAP
  HHHHHHHHHHHHH      EE     HHHHHHHHHHHHHH                HH    HUBBARD
  HHHHHHHHHHH                    HHHHHHHH                 HH    SOLOVYEV
  HHHHHHHHHHHHHH      HHHHHHHHH   EEEEEEE     EEEEE             ROSE
    HHHHHHHHHHHH     HHHHHHHHHH  HHHHHHHH                HHH    COHEN
  HHHHHHHHHHHHH     EEE     HHHHHHHHHHHHHH                HH    VALENCIA
  HHHHHHHHHHHH ---      -- - HHHHHHHHHHHH - ---     ------ H    BAKER

NLQNAFNLAEQHLGLTKLLDPEDISVDHPDEKSIITYVVTYYHYFSKM                sequence
HHHHHHHHHHHHH       HHHH      HHHHHHHHHHHHHHHH                  experimental
HHHHHHHHHHHH                  HHHHHHHHHHHHHHHH                  ROST
HHHHHHHHHHHHHHHHH             HHHHHH   HHHHH                    STERNBERG
HHHHHHHHHHHHH HHHH            HHHHHHHHHHHHHHH                   PREDICTPROTEIN
HHHHHHHHHHHHH                    EEEEEEEEEEEE                   SERVER_SSPRED
     HHHHHHHHHH   E H HHHHH   HHHEEEEEEEEE                      SERVER_GOR
HHHHHHHHHHHHHHH                  EEEEEEEEEHH                    SERVER_NNPREDICT
HHHHHHHHHHHH                    EEEEEE                          SERVER_NNSSP_MULT
 HHHHHHHHHHHH                   EEEEEEE                         SERVER_SSP_MULT
HHHHHHHHHHHHHHHHHH  HHHHH     HHHHHHH  HHHHHH                   SERVER_DSC_MULT
HHHHHHHHHHHHH          EEE       EEEEEEEEEEEE                   SHESTOPALOV
HHHHHHHHHHHHH HHH       EE       EEEEEEEEEEE                    GOLDSTEIN
HHHHHHHHHHHH  HHHH               HHHHHHHHHHH                    JAAP
HHHHHHHHHHHHH    E             HHHHHHHHHHHHHHH                  HUBBARD
HHHHHHHHHHHHH                   HHHHHHHHHHHH                    SOLOVYEV
 HHHHHHHHHHHH          EEEEE     EEEEEEEEE                      ROSE
HHHHHHHHHHHHHH   EE    EEE     HHHHHHHHHHHHHHHH                 COHEN
HHHHHHHHHHHHH   EE             HHEHHHHHHHHHHHH                  VALENCIA
HHHHH HHHHH -            -----       HHHHHH - --                BAKER
```

Figure 65. Sequence and predictions from the CASP2 site and experimental secondary structure for calponin homology domain of β-spectrin, *Homo sapiens*[346] (109 residues), T0037, 1aa2. Experimental secondary structural assignments (DSSP) from the CASP2 site. Key: E, β strand; H, α helix. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: SOLOVYEV, 78.7, 82.4; SERVER–DSC–MULT, 75.0, 74.1; HUBBARD, 66.7, 78.5; VALENCIA, 66.3, 76.6; BAKER, from coordinate model, 65.9, 69.9; STERNBERG, 64.1, 70.4; JAAP, 62.7, 70.4; SERVER–NNPREDICT, 61.2, 64.8; SERVER–PREDICTPROTEIN, 61.1, 73.1; ROST, 60.5, 76.9; COHEN, 59.3, 68.5; SHESTOPALOV, 58.8, 60.4; GOLDSTEIN, 58.0, 65.7; SERVER–NNSSP–MULT, 55.5, 67.6; SERVER–SSPRED, 54.0, 64.8; SERVER–SSP–MULT, 52.6, 62.0; SERVER–GOR, 51.8, 59.3; ROSE, 44.0, 45.3.

evolution-based secondary structure predictions. It does, with S_{ov}–O scores in the 90s. Figure 67 collects the secondary structure predictions submitted for the family. Only the transparent (COBEGETJ) prediction identifies the correct helices and helix junctions throughout the protein, but several of the automated tools come close. The transparent prediction was used to predict contacts between secondary structural elements that were cited by Lesk in his review of the CASP2 project.[174]

E. Conclusions from CASP2

CASP2 confirmed and extended conclusions already evident from CASP1 and other *bona fide* predictions made independently of the CASP projects. First, evolution-based prediction tools could produce excellent secondary structural models when an adequate number of sequences having adequate evolutionary divergence was used as input. Where evolution-based methods did poorly, the poor performance could in general be traced to few homologous sequences for the target or inadequate sequence divergence among the homologs within the family. For proteins with few homologs, results for different predictions cluster around those expected for single sequence predictions (see Figure 7, Nishikawa Ooi). With some of the protein targets (for example, T0032) the scores are worse than for single targets; for others (for example, T0038) the scores are better.

```
    not core   edge             edge        core        edge
ASPIGEGTFDDGPEGWVAYGTDGPLDTSTGALCVAVPAGSAQYGVGVVLNGVAIEEGTTY    sequence
      E         EEE      E      EEEEEE      EEEEEE      EEEE    experimental
                EE            EEEEEE       EEEEEE  EEEE   EE    ROST
      EEE     EEEEEE          EEEEEE      EEEEEEEEEEEEE   EE    STERNBERG
    EE         EEEE           EEEEEE       EEEEEEE EEEE   EE    JAAP
                EEE            EEEE         EEEEE  EEE    EE    GOLDSTEIN
               EEEEE      E   EEEEEE    EEEEEEEEE  EEE    EE    SMITH
     E         EEEEE          EEEEEE       EEEEEEEEEEEE   EE    PREDICTPROTEIN
               EEEE                          EEEEEEEEEE   EE    SERVER_SSPRED
      E        EEEE           EEEEEEE    EEEEEEEEEE EEEE        SERVER_GOR
               EEE             EEEE       EEEEEEE   EEE    E    SRVER_NNPREDICT
               EEEE           EEEEEE         EEEE         EE    SERVER_NNSSP_MULT
                            EEEEEEEE          EEEE       EEE    SERVER_SSP_MULT
   EEE          EE     EEEE   EEEEEE     EEEEEEE  EEEEEE  EE    HUBBARD
               EEEEE           EEEEEE        EEEE  EEE    EE    SOLOVYEV
  E    E    EHHH  EE    EEE    EEE     E EEEE E          EE    MURZIN
   EEEEEE   EEEEEEEE   EEEEE   EE       EEEEEEEEE    EEEEEEE    LENGAUER

              edge        not core       core
TLRYTATASTDVTVRALVGQNGAPYGTVLDTSPALTSEPRQVTETFTASATYPATPAADD    sequence
EEEEEE      E EEEEEEE       EEE  E     EEEEEEEEEE               experimental
EEEEEEEE   EEEEEEEE       E    EEE      EEEEEEEEE               ROST
EEEEEEE   EEEEEEEEEE      EEEE   EEE     EEEEEEEEE              STERNBERG
EEEEEEEEEEEEEEEEEE        EEEE          EEEEEEEEEEE             JAAP
EEEEEE      EEEEEE        EEE            EEEEEE                 GOLDSTEIN
EEEEEEE    EEEEEEE     EEEEEEE   EEEE    EEEEEEEE       EE      SMITH
EEEEEEEEEEEEEEEEEE        EEEE          EEEEEEEEEEE             PREDICTPROTEIN
EEEE       EEEEEE         EEEE            EEEEEEE               SERVER_SSPRED
EEEEE    EEEEEEEEE       EEEEEE          EEEEE        E         SERVER_GOR
EEEEE      HEEEEEE        EE             E EEHE                 SRVER_NNPREDICT
EEEEE      EEEEEEE        EEE              EEEEEE               SERVER_NNSSP_MULT
EEEEE      EEEEEEE                       EEEEEEE                SERVER_SSP_MULT
EEEEEEEE    EEEEEEE       EE   EEEE   EEEEEEEEE                 HUBBARD
EEEEEEEE    EEEEEEEE       EE      EE       EEEEEEE             SOLOVYEV
EEEEEE     EEEEEEE    EEEEEEEE           HHH EEEE     EE        MURZIN
E  EEEE     EEEEEEEE   EEEEEEEEEEEEEEEE  EEEEEEEEEEEE           LENGAUER

  core         edge   not core
PEGQIAFQLGGFSADAWTLCLDDVALDSEVEL                                sequence
 EEEEEEEE      EEEEE   EE                                       experimental
  EEEEEEE       EEEEEEEEEEEEEEE                                 ROST
   EEEEEE      EEEEEEEEEEEE                                     STERNBERG
    HHHHH   EHHHHEEEE        EE                                 JAAP
    HEEH     HHHHHHHH H     HHHH                                GOLDSTEIN
    EEEEE  E   EEEEE    EEEEEEEE                                SMITH
  EEEEEE     HHHHHEEEEEE EE EE                                  PREDICTPROTEIN
     EEE         EEEEHHHHHHHHHH                                 SERVER_SSPRED
 HHHHEEE        HHHHHHHHHHHHHHHH                                SERVER_GOR
    EEEEE      HHHHH        H                                   SRVER_NNPREDICT
    EEEEE       EEEEE        EEE                                SERVER_NNSSP_MULT
 HHHHHHHHH                                                      SERVER_SSP_MULT
  EEEEEEE       EEEEEEEEEEEEEE                                  HUBBARD
   EEEE          EEEE  EEE                                      SOLOVYEV
   EEEEEEE        - EEEEEEEEE                                   MURZIN
 HHH      EEEEEEE    EEEEEEEEEE                                 LENGAUER
```

Figure 66. Sequence and predictions from the CASP2 site and experimental secondary structure for CBDN1, *Cellulomonas fimi* (152 residues),[347] T0038, 1ulo, P14090, GUNC–CELFI. Experimental secondary structural assignments (DSSP) were from the CASP2 site. Key: E, β strand; H, α helix. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: STERNBERG, 79.1, 74.3; SERVER–NNSSP–MULT, 78.9, 76.3; SOLOVYEV, 76.9, 75.0; HUBBARD, 75.0, 66.9; ROST, 68.9, 74.3; GOLDSTEIN, 68.8, 69.7; SMITH, 67.0, 67.1; SERVER–PREDICTPROTEIN, 65.4, 67.8; SERVER–GOR, 63.1, 62.5; JAAP, 62.8, 66.4; SERVER–SSPRED, 60.7, 67.1; SERVER–NNPREDICT, 58.1, 69.7; MURZIN, from coordinate model, 57.1, 60.9; SERVER–SSP–MULT, 55.3, 71.7; LENGAUER, 53.0, 50.7.

One prescription for improvement is clear from this observation: more sequences need to be collected. This will be the inevitable outcome of genome projects. As the sequence databases grow, fewer protein

```
GYFCESCRKIIQKLEDMVGPQPNEDTVTQAASQVCDKLKI     sequence
   HHHHHHHHHHHHHHH     HHHHHHHHHHHHHH        experimental
 EEEHHHHHHHHHH              HHHHHHHHHHHH     SHESTOPALOV
   HHHHHHHHHHHHHH       HHHHHHHHHHHHH        STERNBERG
     HHHHHHHHHHH        HHHHHHHHHHHHHHHH     SERVER_DSC_MULT
   HHHHHHHHHHHHHHH     HHHHHHHHHHHHH         SERVER_PREDICTPROTEIN
   EEEEEEEEHHHHHH         HHHHHHHH           SERVER_SSPRED
    HHHHHHHHHEE          EHHHHHHHHHHHHHH     SERVER_GOR
      HHHHHHHH            HHHHHHHHHHHHHH     SERVER_NNPREDICT
    HHHHHHHHHHHHHH     HHHHHHHHHHHH  HHH     SERVER_NNSSP_MULT
      HHHHHHHHHHH      HHHHHHHHHHHHHHHHH     SERVER_SSP_MULT
   HHHHHHHHHHHHHH      HHHHHHHHHHHHHHH H     JAAP
    HHHHHHHHHHHHHH     HHHHHHHHHHHHH         ROST
   HHHHHHHHHHHHHH      HHHHHHHHHHHHHHHHH     SERVER_PREDICTPROTEIN_SINGLE
  HHHHHHHHHHHHHHHH     HHHHHHHHHHHH    H     SOLOVYEV
   HHHHHHHHHHHHHHH        HHHHHHHHHHH        COBEGETJ
  HHHHHHHHHHHHHHHHH     HHHHHHHHHHHHHH       BENNER (2)
- HHHHHHHHHHHHHHHH        HHHHHHHHH          BAKER
      HHHHHHHHHHHH        EHHHHHHHHHHH       MURZIN
   HHHHHHHHHHHH         HHHHHHHHHHHHHH       EISENBERG (2)
-  HHHHHHHHHHHHH          HHHHHHHHHHH        MOULT
  HHHHHHHHHHHHHHHH       HHHHHHHHHHHH        COHEN

LRGLCKKIMRSFLRRISWDILTGKKPQAICVDIKICKE       sequence
 HHHHHHHHHH     HHHHH    HHHHHHH             experimental
HHHHHHHHHHHHH EEEEEE        EEEEEEE          SHESTOPALOV
HHHHHHHHHHHHHHHHHHHHH     HHHHHHH            STERNBERG
HHHHHHHHHHHHHHHHHHHHH    HHHHHHHHH           SERVER_DSC_MULT
HHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHH         SERVER_PREDICTPROTEIN
      EEEEEEEEEEEEEE HHHHHHHHHHHHHHHH        SERVER_SSPRED
HHHHHHEEEEEE     EE        EEEHHHHHHHH       SERVER_GOR
HHHHHHHHHHHHHHH E HH        EEE H            SERVER_NNPREDICT
HHHHHHHHHHHHHHHHHHHHH       EEEE EE          SERVER_NNSSP_MULT
HH HHHHHHHHHH               EEEEEEEE         SERVER_SSP_MULT
HHHHHHHHHHHHHHHHHHHH       HHHHHH            JAAP
HHHHHHHHHHHHHHHHHHHHH    HHHHHHH             ROST
HHHHHHHHHHHHHHHHHHHHH      EEEEEEEEE         SERVER_PREDICTPROTEIN_SINGLE
HHHHHHHHHHHHHHHHHHHHHH     HHHHHH            SOLOVYEV
HHHHHHHHHHHH   HHHHHHHH HHHHHHHHHH           COBEGETJ
HHHHHHHHHHHHH HHHHHHHHH  HHHHHHHHH           BENNER (2)
     HHHHHHHHHHHHHH         HHHHHHH --       BAKER
    HHHHHHHH      E     HHHHHHHHHHHHH        MURZIN
 HHHHHHHHHHHHHHHHHHH         HHHH            EISENBERG (2)
    HHHHHHHHHHHHHHHHH                        MOULT
   HHHHHHHHHHHHHHHHHHH      HHHHHH           COHEN
```

Figure 67. Sequence and predictions from the CASP2 site and experimental secondary structure for NK-lysin, pig (78 residues),[348] T0042, 1nkl. Experimental secondary structural assignments, calculated with DSSP, were taken from the CASP2 site. Key: E, β strand; H, α helix. A number in parentheses (n) indicates the prediction was a weighted average of n predictions. For these predictions, the prediction with the highest S_{ov}–O is shown. For each prediction, S_{ov}–O and Q_3 are listed in order of descending S_{ov}–O: BENNER (2), 92.1, 84.6; ROST, 85.7, 89.7; STERNBERG, 85.7, 87.2; EISENBERG (2), BAKER, 82.5, 87.0; SOLOVYEV, 82.1, 82.1; COHEN, 81.2, 79.5; SERVER–PREDICTPROTEIN, 81.0, 85.9; JAAP, 80.7, 83.3; MURZIN, from coordinate model, 79.8, 70.5; MOULT, 65.7, 74.0; SERVER–DSC–MULT, 65.6, 79.5; SERVER–NNSSP–MULT, 63.3, 74.4; SERVER–SSP–MULT, 60.0, 65.4; SERVER–GOR, 56.8, 55.1; SERVER–NNPREDICT, 55.1, 62.8; SERVER–PREDICTPROTEIN–SINGLE (2), 54.6, 73.1; SERVER–SSPRED, 45.2, 43.6; SHESTOPALOV, 44.3, 58.0.

families will be small (in their representation in the database), and the quality of evolution-based predictions should improve accordingly.

This observation belies efforts to rank the relative value of different evolution-based prediction methods, both transparent and nontransparent. Much of the difference observed in the different prediction methods arose from the fact that different methods were tested with different subsets of the set of target proteins accessible for *ab initio* predictions, or different input was used by different methods. In many cases, the application of classical scoring methods to targets that contained substantial noncore segments caused an underevaluation of the quality of the prediction (as was the case in CASP1).

Nevertheless, evolution-based methods continued to have difficulties assigning secondary structure near active sites and distinguishing between internal strands and internal helices. Therefore, one still cannot be certain that secondary structure models produced by evolution-based methods are free of all serious mistakes, even when adequate diversity is contained within the protein family being examined. Thus, any model needs to be inspected in detail, and full transparent predictions that call attention to possible serious mistakes (as was done, for example, with the HSP90 and protein serine/threonine phosphatase families) remain an important part of a prediction. The emergence of fully automated, transparent prediction tools (such as SAINT) should combine the informative nature of a transparent prediction with the convenience of an automated prediction.

Two further prescriptions can be made. First, future CASP projects should provide an expanded submission format that allows predictors to identify segments that might be incorrectly assigned for specific reasons. Second, the prediction community should actively discourage referees from blocking publication of predictions in manuscript form. In several cases, predictions submitted to the CASP2 were also submitted as manuscripts for publication in journals, but blocked from publication by an anonymous referee who considered the publication of *bona fide* predictions to be inappropriate. For this reason, manuscripts analyzing the structure of NK-lysin (Richard Russell, personal communication), ferrocheletase, and the S1 domain of polynucleotide nucleotidyltransferase were not published. The dialog and insight that they contained has therefore been lost, especially that which might prove helpful for improving prediction heuristics for difficult proteins and difficult types of secondary structure. Referee anonymity has made it remarkably difficult to persuade a few members of the prediction community that blocking publication never contributes to a scientific enterprise. The effort to persuade must be redoubled.

Despite the problematic serious mistake that characterizes many predictions, the models predicted in CASP2 were useful. As with CASP1, where the core tertiary structural model of phospho-β-galactosidase was successfully predicted, CASP2 yielded convincing tertiary structural models. Perhaps the most valuable of these were made for the HSP90 family, where long-distance homology was established and biological function confirmed, both by prediction. The residue−residue contacts predicted by the VALENCIA group, and the segment−segment contacts predicted by the COBEGETJ group showed clear improvement over the results observed in CASP1.

With respect to methods for scoring evolution-based predictions, CASP2 also confirmed conclusions that were established earlier. First, Q_3 and S_{ov} scores are not appropriately used in evaluating predictions, even as a cutoff to distinguish predictions worthy of closer inspection from those that are not. If the experimental structure is for a protein with large segments introduced in addition to the core segments, the Q_3 and S_{ov} scores can be arbitrarily low.

Last, one cannot help but be impressed with the improvement made by neural networks and other nontransparent tools in the past two years. We cannot say what the neural networks are considering when they make a prediction. The fact that they do poorly when few homologous sequences are used as input, however, suggests that they are identifying some feature in the divergence of sequences, similar to the transparent methods. Intriguingly, in several examples, transparent approaches and the nontransparent neural networks made mistakes in parallel, suggesting that the neural networks have “learned” some of the “rules” that scientists working transparently had deduced. Whatever the reason, neural networks are performing now quite well, as inspection of the above figures shows.

IX. Prospects for the Future

One cannot help but be impressed by the progress that the summary above represents. In the 1980s, the only method to predict a folded structure of a protein was to identify it as a homolog of a protein with a known structure, or to be assisted by experimental information (most notably circular dichroism spectra) that indicated that a protein adopted a regular class of fold (generally all helical). Today, tools are available that have permitted the construction of models of secondary structure that are useable for other purposes.

It would be a mistake to dismiss this progress as an inevitable outcome of having more sequence data. Evolution-based predictions do, of course, incorporate more information than a classical prediction. Additional information certainly cannot hurt prediction, if only by allowing “noise” to be averaged out. To the extent to which mistakes in classical predictions arise from “noise”, then averaging the predictions over several homologs should diminish mistakes. The prediction of the eight-fold $\alpha-\beta$ barrel structure for tryptophan synthase by averaging GOR predictions over a set of homologous sequences, of the annexins by a similar approach (although assisted by circular dichroism data) and the cytokine receptor superfamily are landmarks in this approach.

However, it was clear at the outset with the work of Lenstra *et al.* on ribonuclease in the 1970s (section IV.D) that the approach would not be general. The approach works best on $\alpha-\beta$ structures. It appears to overpredict them, however, suggesting that the component predictions introduce systematic error into the evolution-based prediction. An evolutionary analysis, coupled with an understanding of organic chemistry, offers explanations why.

First, evolutionary considerations about how natural selection, protein stability, and conformation showed the nature of the problem. As the products of natural selection, natural proteins have evolved to violate folding rules to engineer a desired level of instability (section I.A). As organic molecules, proteins should have local conformations that are influenced by long-range interactions. These observations suggested that classical prediction methods based on single sequences would not work, indeed *could* not work, for the general protein. These suggestions, in turn, guided work toward areas that ultimately proved to be more productive, work that focused on identifying elements of tertiary structure (in particular, surface accessibility), constrained ways for using patterns of variation and conservation as indicators of tertiary structure, and exploited manual analysis of homologous protein sequences to speed the development of insight that, in turn, speeded the development of improved prediction heuristics. The prediction of the core antiparallel β sheet of protein kinase and the secondary structure of the src homology 2 domain are landmarks in this approach. The first was especially interesting, as the prediction explicitly denied a homology model, the first example where the confidence in a secondary structure prediction was sufficient to allow such a conclusion to be drawn.

While prognostication is always difficult, an interplay between evolutionary theory, chemical principles, and massive amounts of sequence data may well be useful in analyzing problems generally in biological chemistry, including the role of biological macromolecules in differentiation and development, the design of biological pathways, and the biological chemistry of disease. If so, then this interplay in the protein structure prediction field may serve as a model for a significant part of the future development of biological chemistry.

Much remains to be done, however. Approaches that model the conformation of a target protein from the known conformation of a homologous protein are quite successful, but only to the extent that the target and reference structures are the same. To the extent that the target and reference proteins do not have the same conformation, homology modeling confronts directly the most difficult problems in contemporary physical chemistry: How to model quantitatively the interaction of molecules and molecular fragments with each other, especially in solution, especially when the solvent is water. Much more work must be directed toward understanding the underlying physical chemical issues involved in this interaction, both in proteins and in small molecules.

Long-distance searches for homologs (profiling, threading) often encounter the same physical chemical issues, as potentials and force fields must at some point be called upon to evaluate the superimposition of the target sequence upon the reference structure. Physical potentials and empirical potentials reflect two distinct underlying philosophies for evaluating reference structures identified in a threading exercise. The first confronts again directly the physical chemical problems discussed above. The second must confront the problems associated with the statistical analysis of protein structures, including the relatively small size and potential bias in the crystallographic database. Again, much more work is needed, and much is underway.[169]

Tools that extract information residue-by-residue from a set of aligned homologous sequences using physical chemical models that incorporate an understanding of molecular evolution remain incomplete. For example, the physical chemical models that underlie the approach are best applied to monomeric globular proteins that have physiological function in solution. In particular, membrane proteins have not yet come fully within the scope of these tools (but see refs 349 and 350, where steps have been taken in this direction).

Even if *ab initio* tools based on evolutionary information work at the level of the secondary structure, they do not represent a comprehensive solution to the structure prediction problem. At best, an *ab initio* secondary structure prediction will identify a homolog of the target protein in the crystallographic database. This converts the *ab initio* problem into a homology modeling problem, and the problems associated with homology modeling must then be solved.

This step is, of course, not insignificant. This approach has been successful so often in *bona fide* prediction settings, both in public “contests” and in private industry, that it is easy to imagine that it will work generally. It should not be long before a particular class of prediction problem can be declared “solved”, those in which *ab initio* predictions of secondary structure are used to identify protein homologs in the database too distant to detect by any simple threading or profile methods.

At worst, the *ab initio* problem yields a consensus model for the protein fold, one that does not apply to any individual protein in the family, but applies to the family as a whole. Here, the present task is to learn how to make *ab initio* modeling of tertiary structure from predicted secondary structural elements routine, even in the absence of homologs or analogs in the database. This is the forefront of research in this area at this time. Friesner and Gunn have outlined progress in this area, drawing the conclusion “the problem of determining tertiary structure once secondary structure is specified, although nontrivial from the point of view of both algorithms and potential functions, is tractable with current computing technology”.[40] This is good news indeed, especially as some rather simple potential functions can generate some tertiary structural models robustly in the 4−6 Å range.[40]

Even if *ab initio* tertiary structure modeling from predicted secondary structural elements becomes routine, however, the problem is not solved. Given a consensus model for tertiary structure, most users want to proceed to a model for the conformation of a specific protein in the family. This is, of course, another problem in homology modeling, with the specific protein being the target structure and the consensus model being the homolog. It therefore also confronts the central problems in physical chemistry mentioned above.

Thus, virtually all lines of progress in *ab initio* prediction merely reduce the problem to one of homology modeling, which must then confront and resolve problems in physical chemistry that are difficult to resolve. The message is clear: sooner or later the physical chemical problems alluded to above will need to be solved.

Further, a realist must point out that structure prediction has a competitor: *experimental* structure determination. During the time that modeling has made the advances outlined in this review, crystallography, electron microscopy, and NMR analysis of protein structure have also made dramatic progress. Assisted by molecular biological tools yielding proteins in large amounts, a rationalization of conditions for crystallizing proteins, new methods for phasing diffraction data, and computational advances that speed the solution of the structure, the number of crystal structures per year is about 10-fold higher today than it was a decade ago. To this is added increasing numbers of structures determined by NMR methods.

The general problem of structural biology is not unbounded. The number of families of proteins readily recognizable by sequence similarities will be less than 10 000 when the genomes of *all* organisms on the planet are sequenced.[220] The number of distinct folds may be less than 1000.[351] At some point, experimental analysis of protein structure becomes similar to the analysis of other types of chemical structure. A good analogy is the work done between 1850 and 1950 to identify all of the elements in the Periodic Table. After 1950, the elements were all known, and the research problem became obsolete.

Sooner or later (current estimates are in the year 2020), a crystal structure will be available for each of the recognizable families of proteins that have been produced by Darwinian evolution on planet earth. Barring the discovery of extraterrestrial life, this will effectively remove the need for any *ab initio* structure prediction; all protein-modeling problems will be problems in homology modeling. Ironically, *ab initio* structure prediction may help hasten the progress that will make itself obsolete as a discipline. As noted above, *ab initio* prediction tools are already able to identify proteins that most likely belong to a class of structures already represented in the crystallographic database. Thus, *ab initio* tools already available should help crystallographers and NMR spectrometrists select proteins to study that are *not* members of families of proteins already represented in the database, hastening the time when a representative experimental structure is known for all families of proteins on earth.

When this time comes, it seems certain that the protein structure prediction effort of the 1990s will not be remembered for the scores that prediction methods produced in any particular contest, but for what it contributed to our understanding of protein chemistry and molecular evolution. Hence the emphasis in this review on transparency.

Here, it is worth noting how far the attitude of the computational biochemistry community has evolved in just the past five years. The scope of this review, covering *bona fide* predictions made by transparent analysis of homologous sequences based on an understanding of molecular evolution, where the implementation of the analysis required active participation of an expert, was far from the mainstream of the field. Just three years ago, leading members in the community viewed *bona fide* prediction as fundamentally and scientifically flawed as a research method.[65] Further, those advocating transparency in a prediction method explicitly stated the premise that the “best structural modelling is done by biological chemists who understand the biochemistry of the system that they are studying and use what they know in the modelling effort”. While this was obvious to those with a background in physical organic chemistry, experts in the field found this grounds to assert that transparent methods were neither reproducible nor testable.[65,176]

Further, many computational chemists recognize that a set of scores does not allow one to learn optimally from the prediction exercise, which requires that the prediction must be examined in detail. One can detect increasingly among the community the sentiment that “black box” tools will not produce an understanding of the problem that will last after the problem itself becomes obsolete. Hence the emphasis on what went wrong, what went right, and why, in CASP1 and CASP2.

Last, and perhaps most significantly, the field is beginning to accept a role for human participation in the prediction exercise. For example, reviewing the conclusions of a workshop on structure prediction, Hubbard *et al*. conceded that “more predictions will be obtained if the central figure in the prediction process is the experimentalist working on the protein rather than the theoretician”.[203] Regardless of one’s view, this metamorphosis is noteworthy.

X. Acknowledgements

We wish to thank several groups for making experimental structural data available to us in preparing this review, including W. Bode, I. D. Campbell, E. Meyer, A. Musacchio, D. C. Rees, T. Gallagher, D. B. Wigley, and H. Yu. We also wish to thank Judy Hempel (BioSym), Rick Lee (BioSym), Professor Susan Taylor, and Professor Edgar Meyer for many helpful comments on predictions that we have made, Daniel Caraco and David Liberles for careful reading of the text, and Arthur Lesk for a preview of his manuscript.

XI. Glossary

BLAST A program to perform fast database searching combined with rigorous statistics for judging the significance of matches: http://www.ncbi.nlm.nih.gov/BLAST/.

Core The part of the protein fold that is conserved during divergent evolution.

DARWIN Data Analysis and Retrieval With Indexed Nucleotide/peptide sequences. A programming environment for organizing and analyzing large amounts of sequence data. Services from DARWIN are available on the Web at http://cbrg.inf.ethz.ch.

Define Define produces a list of the secondary structure of a protein and some relations between the secondary elements based solely on the coordinates of the α carbon atoms. The principal procedure uses difference distance matrices for evaluating the match of interatomic distances in the protein to those from idealized secondary structures.

DSC Discrimination of protein Secondary structure Class, a program to predict secondary structure:[106] http://bonsai.lif.icnet.uk/bmm/dsc/dsc_form_align.html.

DSSP Define Secondary Structure of Proteins, a program to standardize secondary structure assignment from X-ray coordinates. The hydrogen bonds and torsion angles are the main parameters that are used by the program to make these assignments: http://www.sander.embl-heidelberg.de/dssp/.

GOR The Garnier−Osguthorpe−Robson method for predicting secondary structure for a protein sequence. The method, discussed in detail in ref 105 is based on the theory of information, which has its roots in probability theory. Central to this method is the concept that residues, considered individually and as part of a sequence pattern, have a tendency to adopt certain conformations. The following are some servers that provide GOR analysis on the Internet: http://molbiol.soton.ac.uk/compute/GOR.html and http://absalpha.dcrt.nih.gov:8008/gor.html.

Hydrophobic moment Analog of the electric dipole moment. It measures the asymmetry of hydrophobicity or amphiphilicity.

Indel In̲sertion or del̲etion. An evolutionary event that either adds amino acids or subtracts amino acids from a polypeptide chain.

Markov A Markov chain is a sequence of random variables such that the future of the variable is determined by its present state (but independent of the way in which the present state arose).

NNSSP Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments:[81] http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html.

P-curve Another program to assign secondary structure from Cartesian coordinates. The assignments are made from a set of helicoidal parameters.

Parse A segment of polypeptide chain or section of a multiple sequence alignment that lies between two standard secondary structural units; α helix or *β* strand.

PHD A neural network program[208] for assigning secondary structure: http://www.embl-heidelberg.de/predictprotein/predictprotein.html.

PREDATOR A secondary structure prediction program. It takes as input a single protein sequence to be predicted and can optimally use a set of unaligned sequences as additional information to predict the query sequence. The mean prediction accuracy of PREDATOR is 68% for a single sequence and 75% for a set of related sequences. PREDATOR does not use multiple sequence alignment. Instead, it relies on careful pairwise local alignments of the sequences in the set with the query sequence to be predicted: http://www.embl-heidelberg.de/cgi/predator_serv.pl.

Q_3 A score assigned to a secondary structure prediction that involves comparing the prediction to the experimental structure. $Q_3 = Q_{ok}/Q_{total}$, where Q_{ok} is the number of correct assignments and Q_{total} is the total number of assignments.

QL The Quadratic-Logistic prediction method is based on maximum−likelihood methods: http://absalpha.dcrt.nih.gov:8008/predict.html.

QSLAVE PSLAVE/QSLAVE Alignment and searching for common protein folds using a databank of structural templates: http://www-cryst.bioc.cam.ac.uk/local_html/soft-base.html.

SIMPA SIMilarity Peptide Analysis,[132] a program to predict secondary structure based on sequence similarity between peptides (17 amino acid long) and sequences of known structure.

SOPMA Self-Optimized Prediction Method from Alignment[83] is a package to make secondary structure predictions of proteins: http://ibcp.fr/serv_pred.html.

SSPRED A three-state secondary structure prediction routine. The computer routine PreferCal was first written to determine the preference or avoidance weights for each possible pair of residue exchanges and for each of the three secondary structural states. PreferPred predicts secondary structural elements within a query sequence multiply aligned to related primary structures. Finally, PreferEval allows evaluation of the accuracy of the secondary structure predictions relative to those known from three-dimensional structural determinations: http://www.embl-heidelberg.de/cgi/sspred_mul.pl.

STRIDE Program to assign secondary structure from experimental coordinates.[88] STRIDE uses both hydrogen-bonding and main chain dihedral angles as input, parameterizes this information against secondary structures assigned by crystallographers, and optimizes the relative contributions of the two with the specific goal of producing assignments which are in closer agreement with the assignments that crystallographers made. The propensities of amino acid residues with specific ϕ and ψ angles to be part of helices and strands are also considered, so the method depends as well on the nature of the amino acids involved: http://www.embl-heidelberg.de/cgi/stride_serv.

Target protein A protein of unknown conformation, whose conformation is sought.

Threading A process that involves superimposing the sequence of a target protein on the three-dimensional structure of a possible distant homolog to see if the target sequence might fold to give the same overall conformation.

Transparent prediction method A tool for assigning secondary structure to a protein sequence that yields an assignment where the user can understand why the assignment was made.

ZPRED Computer program[21] that predicts secondary structure using physicochemical information from a set of aligned sequences and the Garnier *et al.*[105] secondary structure decision constants: http://kestrel.ludwig.ucl.ac.uk/zpred.html.•

XII. Appendix

Protein Structure Prediction Tools on the World-Wide Web

Homology Modeling (Comparative Modeling)

• Map123d: evaluation of 3D-models, Sallantin group

http://www-bio.lirmm.fr:8090/eval.html

REF: J. Gracy, L. Chiche, and J. Sallantin, Improved alignment of weakly homologous protein sequences using structural information. *Protein Eng.* **1993**, *6*, 821−829.

• MODELLER: homology modeling program by satisfaction of spatial restraints, Sali group

ftp://guitar.rockefeller.edu/pub/modeller/ (ftp site)

REF: A. Sali and T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.

• SWISS-MODEL (part of ExPasy server): automated homology modeling, Peitsch group

http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html

REF: M. C. Peitsch, ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.* **1996**, Feb, 24(1), 274−9.•

Threading (Fold Recognition)

• 123D TopLign: threading tool based on secondary structure prediction and residue−residue contact potential (part of the GMD-SCAI server), Zimmer group

http://cartan.gmd.de/123D-test.html

REF: N. N. Alexandrov, R. Nussinov, and R. M. Zimmer, Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pacific Symposium on Biocomputing '96*; Hunter, L., Klein, T. E., Eds.; World Scientific Publishing Co.: Singapore, 1996; pp 53−72.

• Gon+predss/Gon+predss+MULT: (part of the UCLA-DOE frsvr server) Fischer threading approach, considers predicted secondary structure in addition to fold recognition, Eisenberg group

http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html

REF: D. Fischer and D. Eisenberg, Fold recognition using sequence-derived predictions. *Protein Sci.* **1996**, *5*, 947−955.

• H3P2: Rice threading approach (part of the UCLA-DOE frsvr server), considers predicted secondary structure, Eisenberg group
http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html
REF: D. Rice and D. Eisenberg, A 3D−1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **1996**, submitted for publication.

• ProFit: threading based on an empirical "energy" function, code can be downloaded, Sippl group
ftp://gundi.came.sbg.ac.at/publ (ftp site)
REF: M. J. Sippl, Recognition of errors in three-dimensional structures of proteins. *Proteins* **1993**, *17*, 355−62.

• PSCAN: profilescan threading, Arne Elofsson
http://www.biokemi.su.se/~arne/pscan/
REF (most closely related): A. Elofsson, D. Fischer, D. W. Rice, S. M. LeGrand, and D. Eisenberg, A study of combined structure−sequence profiles. *Folding & Design* **1996**, *1*, 451−461.

• RDP: threading by recursive dynamic programming (part of the GMD-SCAI server), Lengauer group
http://cartan.gmd.de/cgi-bin/ToPLignLogin?/home/protal/WWW+/home/protal/WWW/fast+FastLogin.rc+FastRDP
REF: R. Thiele, R. Zimmer, and T. Lengauer, Recursive dynamic programming for adaptive sequence and structure alignment. *Intelligent Systems for Molecular Biology* **1995**, *3*, 384−92.

• THREADER: threading code can be downloaded, Thornton group
ftp://ftp.biochem.ucl.ac.uk/pub/THREADER
REF: D. T. Jones, W. R. Taylor, and J. M. Thornton, A new approach to protein fold recognition. *Nature* **1992**, *358*, 86−89.

• TOPITS (called PHD threader as part of the PredictProtein server): threading based on secondary structure prediction and solvent accessibility prediction, Burkhard Rost
http://www.embl-heidelberg.de/predictprotein/
REF: B. Rost, TOPITS: threading one-dimensional predictions into three-dimensional structures. *Ismb* **1995**, *3*, 314−321.•

Solvent Accessibility Prediction

• PHD (called PHDacc as part of the PredictProtein server): accessibility prediction (10 states in output) by a neural network
http://www.embl-heidelberg.de/predictprotein/
REF: B. Rost, PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **1996**, *266*, 525−539.

•DARWIN, prediction of surface, interior, active site, and parse positions from homologous sequences
http://cbrg.inf.ethz.ch

Ab Initio Secondary Structure Prediction

(servers accepting multiple alignments as input are marked [MULT+])

• COILS: probabilistic coiled coil prediction
http://ulrec3.unil.ch/software/COILS_form.html [MULT−]
REF: A. Lupas, M. Van Dyke, and J. Stock, Predicting Coled Coils from Protein Sequences. *Science* **1991**, *252*, 1162−1164.

• DAS: transmembrane helix prediction using low-stringengy dot plots, Eloffsson group
http://www.biokemi.su.se/~server/DAS/ [MULT−]
REF: (Web only) M. Cserzo, E. Wallin, I. Simon, G. von Heijne, and A. Elofsson, Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: application of the Dense Alignment Surface method. http://www.biokemi.su.se/~server/DAS/abstract.html.

• DPM (Double Prediction Method): secondary structure prediction by combining Chou−Fasman-type parameters and protein-folding class prediction (as part of the Protein Sequence Analysis server at IBCP), Deleage group
http://www.ibcp.fr/serv_pred.html [MULT−]
REF: G. Deleage and B. Roux, An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* **1987**, *1*, 289−294.

• DSC: secondary structure prediction by discrimination of secondary structure class, Sternberg group
http://bonsai.lif.icnet.uk/bmm/dsc/dsc_read_align.html - [MULT+]
REF: R. D. King and M. J. E. Sternberg, Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **1996**, *5*, 2298−2310.

• GOR: classic, statistical method for protein secondary structure prediction, online at SBD Southampton
http://molbiol.soton.ac.uk/compute/GOR.html [MULT−]
or at the University of Leeds
http://bmbsgi11.leeds.ac.uk/bmb5dp/gor.html [MULT−]
REF: J. Garnier, D. J. Osguthorpe, and B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **1978**, *120*, 97−120.

• Map123d: secondary structure prediction (neural network) for homology modeling, Sallantin group
http://www-bio.lirmm.fr:8090/intro.html [MULT−]
REF: J. Gracy, L. Chiche, and J. Sallantin, Learning and alignment methods applied to protein structure prediction. *Biochimie* **1993**, *75*, 353−361.

• Multicoil: two- and three-stranded coiled coil prediction by analysis of correlated residues, Kim group (program can also be downloaded), based on Paircoils program
http://ostrich.lcs.mit.edu/cgi-bin/multicoil [MULT−]
REF: E. Wolf, P. S. Kim, and B. Berger, MultiCoil: A program for predicting two- and three-stranded coiled coils. *Protein Sci.* **1997**, in press.

• MultPredict (also known as ZPRED): statistical secondary structure prediction, based on physicochemical residue properties, from AMPS (Barton) multiple sequence alignments, Sternberg group
http://kestrel.ludwig.ucl.ac.uk/zpred.html [MULT+]
REF: M. J. Zvelebil, G. J. Barton, W. R. Taylor, and M. J. Sternberg, Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **1987**, *195*, 957−961.

• NNPREDICT: secondary structure prediction by a neural network, Cohen group
http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html - [MULT−]
REF: D. G. Kneller, F. E. Cohen, and R. Langridge, Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **1990**, *214*, 171−182.

• NNSSP: secondary structure prediction by an improved nearest-neighbor method using multiple sequence information (part of the structure prediction server at the Baylor College of Medicine), Solovyev group
http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html [MULT+] (e-mail)
REF: A. A. Salamov and V. V. Solovyev, Prediction of protein secondary structure by combining nearest-neighbor

algorithms and multiple sequence alignments. *J. Mol. Biol.* **1995**, *247*, 11−15.

- Paircoils: two-stranded coiled coil prediction by analysis of correlated residues, Kim group (program can also be downloaded)

http://ostrich.lcs.mit.edu/cgi-bin/score [MULT−]

REF: B. Berger, D. B. Wilson, E. Wolf, T. Tonchev, M. Milla, and P. S. Kim, Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 8259−8263.

- PHD (called PHDsec as part of PredictProtein Server): secondary structure prediction by a profile fed neural network, Sander group

http://www.embl-heidelberg.de/predictprotein/ [MULT+]

REF: B. Rost and C. Sander, Prediction of protein structure at better than 70% accuracy. *J. Mol. Biol.* **1993**, *232*, 584−599.

REF: B. Rost, PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* **1996**, *266*, 525−539.

- PHD (called PHDhtm as part of PredictProtein Server): transmembrane helix prediction by a neural network, Sander group

http://www.embl-heidelberg.de/predictprotein/ [MULT+]

REF: B. Rost, R. Casadio, P. Fariselli, and C. Sander, Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.* **1995**, *4*, 521−533.

REF: B. Rost, PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* **1996**, *266*, 525−539.

- PREDATOR: secondary structure prediction from local sequence alignments with known structures, Argos Group

http://www.embl-heidelberg.de/argos/predator/predator_info.html [MULT+]

REF: D. Frishman and P. Argos, Incorporation of non-local interactions in protein secondary structure prediction from amino acid sequence. *Protein Eng.* **1996**, *9*, 133−42.

- PSA: probabilistic folding class, secondary and supersecondary structure prediction, Smith group

http://bmerc-www.bu.edu/psa/ [MULT−]

REF: C. M. Stultz, J. V. White, and T. F. Smith, Structural analysis based on state-space modeling. *Protein Sci.* **1993**, *2*, 305−314.

- QL: quadratic-logistic secondary structure prediction, Munson group

http://absalpha.dcrt.nih.gov:8008/predict.html [MULT−]

REF: P. J. Munson, V. Di Francesco, and R. Porrelli, Protein secondary structure prediction using periodic-Quadratic-Logistic Models: theoretical and practical Issues. 27th Annual Hawaii International Conference on System Science 5:375−384, IEEE, Los Alamitos, CA, 1994.

- SAPS: statistical analysis of protein sequences [MULT−]

http://ulrec3.unil.ch/software/SAPS_form.html

REF: V. Brendel, P. Bucher, I. Nourbakhsh, B. E. Blaisdell, and S. Karlin, Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2002−2006.

- SOPMA (as part of the Protein Sequence Analysis server at IBCP): self-optimized secondary structure prediction method, Deleage group

http://www.ibcp.fr/serv_pred.html [MULT−]

REF: C. Geourjon and G. Deleage, SOPM: a self-optimized method for protein secondary structure prediction. *Protein Eng.* **1994**, *7*, 157−64.

REF: C. Geourjon and G. Deleage, SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS* **1995**, *11*, 681−684.

- SOSUI: secondary structure prediction for membrane proteins, Mitaku group, Tokyo University of Agriculture and Technology

http://www.tuat.ac.jp/~mitaku/adv_sosui/ [MULT−]

REF: n/a (March 1997).

- SSCP: secondary structure content prediction from sequence, Argos group

http://www.embl-heidelberg.de/argos/sscp/sscp_info.html [MULT−]

REF: F. Eisenhaber, F. Imperiale, P. Argos, and Frommel C., Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods. *Proteins* **1996**, *25*, 157−68.

REF: F. Eisenhaber, F. Frommel, and P. Argos, Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins* **1996**, *25*, 169−79.

- SSP: segment-oriented secondary structure prediction using linear discriminant analysis (part of the structure prediction server at the Baylor College of Medicine), Solovyev group

http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html [MULT+] (e-mail)

REF: V. V. Solovyev and A. A. Salamov, Predicting alpha-helix and beta-strand segments of globular proteins. *CABIOS* **1994**, *10*, 661−669.

- SSPAL: secondary structure prediction for single sequences (NO multiple sequence information required) by a nearest neighbor method looking for local sequence alignments with known structures (part of the structure prediction server at the Baylor College of Medicine), Solovyev group

http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html [MULT−]

REF: A. A. Salamov and V. V. Solovyev, Protein secondary structure prediction using local alignments. *J. Mol. Biol.* **1997**, *268*, 31−36.

- SSPRED: secondary structure prediction based on residue exchange weight matrixes in different secondary structures, Argos group

http://www.embl-heidelberg.de/cgi/sspred_mul.pl - [MULT+]

REF: P. K. Mehta, J. Heringa, and P. Argos, A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci.* **1995**, *4*, 2517−2525.

- TMAP: prediction of transmembrane segments using multiple sequence alignments, Argos group

http://www.embl-heidelberg.de/tmap/tmap_info.html [MULT+]

REF: B. Persson and P. Argos, Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.* **1994**, *237*, 182−192.

Tertiary Structure Prediction

- PHD (called PHDtopology as part of the PredictProtein server): topology (IN or OUT) prediction for transmembrane helices by a neural network, Sander group

http://www.embl-heidelberg.de/predictprotein/

REF: B. Rost, P. Fariselli, and R. Casadio, Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* **1996**, *5*, 1704−1718.

REF: B. Rost, PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* **1996**, *266*, 525−539.

• TM pred: prediction of transmembrane secondary structure and orientation, Stoffel group

http://ulrec3.unil.ch/software/TMPRED_form.html

REF: K. Hofmann and W. Stoffel, TMbase - A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler* **1993**, *347*, 166.•

Evaluation of Secondary Structure Prediction

• EvalSec (part of the PredictProtein server): calculation of evaluation indices for secondary structure predictions, Sander group

http://www.embl-heidelberg.de/predictprotein/

REF: B. Rost, C. Sander, and R. Schneider, Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* **1994**, *235*, 13−26.

Joint Servers Allowing Submission to Different Methods Simultaneously

Threading (Fold Recognition)

• UCLA-DOE frsvr

Gon+predss+MULT	(D. Fischer and D. Eisenberg, UCLA)
H3P2	(D. Rice and D. Eisenberg, UCLA)
TOPITS	(B. Rost, EMBL)
123D	(N. N. Alexandrov, R. Nussinov, and R. M. Zimmer, Amgen/GMD)
PSCAN	(A. Elofsson, D. Fischer, D. W. Rice, S. M. Legrand, and D. Eisenberg, Stockholm U./UCLA)

http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html

Ab Initio Secondary Structure Prediction

• IBCP server

DPM	(G. Deleage and B. Roux, CNRS)
PHDsec	(B. Rost and C. Sander, EMBL)
SOPMA	(C. Geourjon and G. Deleage, IBCP-CNRS)
+ statistical methods	

http://www.ibcp.fr/serv_pred.html

• BCM server

SSP	(V. V. Solovyev and A. A. Salamov, BCM)
NNSSP	(A. A. Salamov and V. V. Solovyev, BCM)
SSPAL	(A. A. Salamov and V. V. Solovyev, BCM)

http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html [MULT+]

XIII. References

(1) Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M.; et al. *Science* **1995**, *269*, 496−512.

(2) Fraser, C. M.; Gocayne, J. D.; White, O.; Adams, M. D.; Clayton, R. A.; Fleischmann, R. D.; Bult, C. J.; Kerlavage, A. R.; Sutton, G.; Kelley, J. M.; et al. *Science* **1995**, *270*, 397−403.

(3) Bult, C. J.; White, O.; Olsen, G. J.; Zhou, L.; Fleischmann, R. D.; Sutton, G. G.; Blake, J. A.; FitzGerald, L. M.; Clayton, R. A.; Gocayne, J. D.; Kerlavage, A. R.; Dougherty, B. A.; Tomb, J. F.; Adams, M. D.; Reich, C. I.; Overbeek, R.; Kirkness, E. F.; Weinstock, K. G.; Merrick, J. M.; Glodek, A.; Scott, J. L.; Geoghagen, N. S. M.; Weidman, J. F.; Fuhrmann, J. L.; Venter, J. C.; et al. *Science* **1996**, *273*, 1058−73.

(4) Williams, N. *Science* **1996**, *272*, 481.

(5) Sulston, J.; Du, Z.; Thomas, K.; Wilson, R.; Hillier, L.; Staden, R.; Halloran, N.; Green, P.; Thierry-Mieg, J.; Qiu, L.; et al. *Nature* **1992**, *356*, 37−41.

(6) Ramachandran, G. N.; Sasisekharan, V. *Adv. Protein Chem.* **1968**, *23*, 283−438.

(7) Saunders, M.; Houk, K. N.; Wu, W.-D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. *J. Am. Chem. Soc.* **1990**, *112*, 1419−1420.

(8) Vasquez, M.; Nementhy, G.; Scheraga, H. A. *Chem. Rev.* **1994**, *94*, 2183−2239.

(9) Evans, J. S.; Mathiowetz, A. M.; Chan, S. I.; Goddard, W. A. *Protein Sci.* **1995**, *4*, 1203−1216.

(10) Levitt, M. *J. Mol. Biol.* **1992**, *226*, 507−33.

(11) Schiffer, C. A.; Caldwell, J. W.; Stroud, R. M.; Kollman, P. A. *Protein Sci.* **1992**, *1*, 396−400.

(12) Park, B.; Levitt, M. *J. Mol. Biol.* **1996**, *258*, 367−92.

(13) Hao, M. H.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4984−9.

(14) Fraternali, F.; Van Gunsteren, W. F. *J. Mol. Biol.* **1996**, *256*, 939−48.

(15) Benner, S. A. *Adv. Enzyme Regul.* **1989**, *28*, 219−36.

(16) Pascarella, S.; Argos, P. *Protein Eng.* **1994**, *7*, 185−93.

(17) Wako, H.; Blundell, T. L. *J. Mol. Biol.* **1994**, *238*, 693−708.

(18) Wako, H.; Blundell, T. L. *J. Mol. Biol.* **1994**, *238*, 682−92.

(19) Rost, B.; Sander, C. *Proteins* **1994**, *19*, 55−72.

(20) Taylor, W. R. *Protein Eng.* **1993**, *6*, 593−604.

(21) Zvelebil, M. J.; Barton, G. J.; Taylor, W. R.; Sternberg, M. J. *J. Mol. Biol.* **1987**, *195*, 957−61.

(22) Rossman, M. G.; Liljas, A.; Branden, C. I.; Banaszak, L. J. *Enzymes* **1975**, *11*, 61.

(23) Chothia, C.; Lesk, A. M. *EMBO J.* **1986**, *5*, 823−6.

(24) Sternberg, M. J.; Cohen, F. E. *Int. J. Biol. Macromol.* **1982**, *4*, 137−144.

(25) Maxfield, F. R.; Scheraga, H. A. *Biochemistry* **1979**, *18*, 697−704.

(26) Lenstra, J. A.; Hofsteenge, J.; Beintema, J. J. *J. Mol. Biol.* **1977**, *109*, 185−93.

(27) Crawford, I. P.; Niermann, T.; Kirschner, K. *Proteins* **1987**, *2*, 118−29.

(28) Bowie, J. U.; Luethy, R.; Eisenberg, D. *Science* **1991**, *253*, 164−70.

(29) Shortle, D. *Nat. Struct. Biol.* **1995**, *2*, 91−3.

(30) Gray, P. M. D.; Kemp, G. J. L.; Rawlings, C. J.; Brown, N. P.; Sander, C.; Thornton, J. M.; Orengo, C. M.; Wodak, S. J.; Richelle, J. *Trends Biochem. Sci.* **1996**, *21*, 251−256.

(31) Feng, D. F.; Johnson, M. S.; Doolittle, R. F. *J. Mol. Evol.* **1984**, *21*, 112−25.

(32) Smith, T. F.; Waterman, M. S.; Fitch, W. M. *J. Mol. Evol.* **1981**, *18*, 38−46.

(33) Taubes, G. *Science* **1996**, *273*, 588−590.

(34) Woese, C. R. *Microbiol. Rev.* **1987**, *51*, 221−271.

(35) Benner, S. A.; Ellington, A. D. *Bioorg. Chem. Front.* **1990**, *1*, 1−70.

(36) Adey, N. B.; Tollefsbol, T. O.; Sparks, A. B.; Edgell, M. H.; Hutchison, C. A. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 1569−73.

(37) Malcolm, B. A.; Wilson, K. P.; Matthews, B. W.; Kirsch, J. F.; Wilson, A. C. *Nature* **1990**, *345*, 86−9.

(38) Stackhouse, J.; Presnell, S. R.; McGeehan, G. M.; Nambiar, K. P.; Benner, S. A. *FEBS Lett.* **1990**, *262*, 104−6.

(39) Jermann, T. M.; Opitz, J. G.; Stackhouse, J.; Benner, S. A. *Nature* **1995**, *374*, 57−9.

(40) Friesner, R. A.; Gunn, J. R. *Annu. Rev. Biophys. Biomol. Struct.* **1996**, *25*, 315−342.

(41) Pedersen, J. T.; Moult, J. *Curr. Opin. Struct. Biol.* **1996**, *6*, 227−31.

(42) Eisenhaber, F.; Persson, B.; Argos, P. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 1−94.

(43) Bohm, G. *Biophys. Chem.* **1996**, *59*, 1−32.

(44) Rost, B.; Sander, C. *Annu. Rev. Biophys. Biomol. Struct.* **1996**, *25*, 113−136.

(45) Benner, S. A.; Gerloff, D. L. *FEBS Lett.* **1993**, *325*, 29−33.

(46) Benner, S. A.; Gerloff, D. L.; Jenny, T. F. *Science* **1994**, *265*, 1642−4.

(47) Barton, G. J. *Curr. Opin. Struct. Biol.* **1995**, *5*, 372−6.

(48) Lattman, E. E. *Proteins* **1995**, *23*, R1.

(49) Moult, J. *Curr. Opin. Biotechnol.* **1996**, *7*, 422−7.

(50) Genuine. *Websters New International Dictionary*, 3rd ed.; Simon and Schuster: New York, 1981; definition 3.

(51) Toulmin, S. E. *Foresight and understanding; an enquiry into the aims of science*; Harper & Row: New York, 1963.

(52) Hunt, T.; Purton, M. *Trends Biochem. Sci.* **1992**, *17*.

(53) Schulz, G. E.; Schirmer, R. H. *Principles of Protein Structure*; Springer-Verlag: New York, 1979.

(54) Schulz, G. E.; Barry, C. D.; Friedman, J.; Chou, P. Y.; Fasman, G. D.; Finkelstein, A. V.; Lim, V. I.; Ptitsyn, O. B.; Kabat, E. A.; Wu, T. T.; Levitt, M.; Robson, B.; Nagano, K. *Nature* **1974**, *250*, 140−2.

(55) Matthews, B. W. *Biochim. Biophys. Acta* **1975**, *405*, 442−51.

(56) Kabsch, W.; Sander, C. *FEBS Lett.* **1983**, *155*, 179−82.
(57) Rees, D. C. In *Current Research in Protein Chemistry*; Villafranca, J., Ed.; Academic Press: New York, 1990.
(58) Sippl, M. J.; Flöckner, H. *Structure* **1996**, *4*, 15−19.
(59) Thornton, J. M.; Flores, T. P.; Jones, D. T.; Swindells, M. B. *Nature* **1991**, *354*, 105−6.
(60) Russell, R. B.; Sternberg, M. J. *Curr. Biol.* **1995**, *5*, 488−90.
(61) Lesk, A. M.; Boswell, D. R. *Bioessays* **1992**, *14*, 407−10.
(62) Defay, T.; Cohen, F. E. *Proteins* **1995**, *23*, 431−45.
(63) Fasman, G. D. *Prediction of Protein Structure and the Principles of Protein Conformation*; Plenum: New York, 1989.
(64) Garnier, J.; Robson, B. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum: New York, 1989.
(65) Robson, B.; Garnier, J. *Nature* **1993**, *361*, 506.
(66) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−637.
(67) Schiffer, M.; Edmundson, A. B. *Biophys. J.* **1967**, *7*, 121−35.
(68) Colloc'h, N.; Etchebest, C.; Thoreau, E.; Henrissat, B.; Mornon, J. P. *Protein Eng.* **1993**, *6*, 377−82.
(69) Sklenar, H.; Etchebest, C.; Lavery, R. *Proteins* **1989**, *6*, 46−60.
(70) Richards, F. M.; Kundrot, C. E. *Proteins* **1988**, *3*, 71−84.
(71) Booker, G. W.; Gout, I.; Downing, A. K.; Driscoll, P. C.; Boyd, J.; Waterfield, M. D.; Campbell, I. D. *Cell* **1993**, *73*, 813−22.
(72) Koyama, S.; Yu, H.; Dalgarno, D. C.; Shin, T. B.; Zydowsky, L. D.; Schreiber, S. L. *Cell* **1993**, *72*, 945−52.
(73) Gerloff, D. L.; Jenny, T. F.; Knecht, L. J.; Gonnet, G. H.; Benner, S. A. *FEBS Lett.* **1993**, *318*, 118−24.
(74) Russell, R. B.; Barton, G. J. *J. Mol. Biol.* **1993**, *234*, 951−7.
(75) Rost, B.; Sander, C.; Schneider, R. *J. Mol. Biol.* **1994**, *235*, 13−26.
(76) Yu, H.; Rosen, M. K.; Shin, T. B.; Seidel-Dugan, C.; Brugge, J. S.; Schreiber, S. L. *Science* **1992**, *258*, 1665−8.
(77) Summers, N. L.; Carlson, W. D.; Karplus, M. *J. Mol. Biol.* **1987**, *196*, 175−98.
(78) Jenny, T. F.; Benner, S. A. *Biochem. Biophys. Res. Commun.* **1994**, *200*, 149−55.
(79) Benner, S. A.; Gerloff, D.; Chelvanayagam, G. *Proteins* **1995**, *23*, 446−53.
(80) Wiesmann, C.; Beste, G.; Hengstenberg, W.; Schulz, G. E. *Structure* **1995**, *3*, 961−8.
(81) Salamov, A. A.; Solovyev, V. V. *J. Mol. Biol.* **1995**, *247*, 11−5.
(82) Mehta, P. K.; Heringa, J.; Argos, P. *Protein Sci.* **1995**, *4*, 2517−25.
(83) Geourjon, C.; Deléage, G. *CABIOS* **1995**, *11*, 681−684.
(84) Chandonia, J. M.; Karplus, M. *Protein Sci.* **1996**, *5*, 768−74.
(85) Musacchio, A.; Noble, M.; Pauptit, R.; Wierenga, R.; Saraste, M. *Nature* **1992**, *359*, 851−5.
(86) Koyama, S.; Yu, H.; Dalgarno, D. C.; Shin, T. B.; Zydowsky, L. D.; Schreiber, S. L. *FEBS Lett.* **1993**, *324*, 93−8.
(87) Kohda, D.; Hatanaka, H.; Odaka, M.; Mandiyan, V.; Ullrich, A.; Schlessinger, J.; Inagaki, F. *Cell* **1993**, *72*, 953−60.
(88) Frishman, D.; Argos, P. *Proteins* **1995**, *23*, 566−79.
(89) Benner, S. A.; Cohen, M. A.; Gonnet, G. H. *J. Mol. Biol.* **1993**, *229*, 1065−82.
(90) Gerloff, D. L.; Jenny, T. F.; Knecht, L. J.; Benner, S. A. *Biochem. Biophys. Res. Commun.* **1993**, *194*, 560−5.
(91) Benner, S. A.; Gerloff, D. *Adv. Enzyme Regul.* **1991**, *31*, 121−81.
(92) Benner, S. A.; Cohen, M. A.; Gonnet, G. H.; Berkowitz, D. B.; Johnsson, K. In *The RNA World*; Gesteland, R., Atkins, J., Eds.; Cold Spring Harbor: New York, 1993.
(93) Fischer, D.; Eisenberg, D. *Protein Sci.* **1996**, *5*, 947−955.
(94) Hopp, T. P.; Woods, K. R. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 3824−8.
(95) Hopp, T. P. *Pept. Res.* **1993**, *6*, 183−90.
(96) Jenny, T. F.; Gerloff, D. L.; Cohen, M. A.; Benner, S. A. *Proteins* **1995**, *21*, 1−10.
(97) Scheraga, H. A. *J. Am. Chem. Soc.* **1960**, *82*, 3847−3852.
(98) Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H. *Proc. Natl. Acad. Sci. U.S.A.* **1961**, *47*, 1309−1314.
(99) Hartl, D. U. *Nature* **1996**, *381*, 571−9.
(100) Baker, D.; Agard, D. A. *Biochemistry* **1994**, *33*, 7505−9.
(101) Dodge, R. W.; Laity, J. H.; Rothwarf, D. M.; Shimotakahara, S.; Scheraga, H. A. *J. Protein Chem.* **1994**, *13*, 409−21.
(102) Guzzo, A. V. *Biophys. J.* **1965**, *5*, 809−822.
(103) Burgess, A. W.; Scheraga, H. A. *J. Theor. Biol.* **1975**, *53*, 403−20.
(104) Chou, P. Y.; Fasman, G. D. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1978**, *47*, 45−148.
(105) Garnier, J.; Osguthorpe, D. J.; Robson, B. *J. Mol. Biol.* **1978**, *120*, 97−120.
(106) King, R. D.; Sternberg, M. J. E. *Protein Sci.* **1996**, *5*, 2298−2310.
(107) Ellis, L. B.; Milius, R. P. *Comput. Appl. Biosci.* **1994**, *10*, 341−8.
(108) Jones, D. T.; Moody, C. M.; Uppenbrink, J.; Viles, J. H.; Doyle, P. M.; Harris, C. J.; Pearl, L. H.; Sadler, P. J.; Thornton, J. M. *Proteins* **1996**, *24*, 502−513.
(109) Kabsch, W.; Sander, C. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 1075−8.
(110) Argos, P. *J. Mol. Biol.* **1987**, *197*, 331−48.
(111) Cohen, B. I.; Presnell, S. R.; Cohen, F. E. *Protein Sci.* **1993**, *2*, 2134−45.
(112) Rooman, M. J.; Wodak, S. J. *Biochemistry* **1992**, *31*, 10239−49.
(113) Rooman, M. J.; Kocher, J. P.; Wodak, S. J. *Biochemistry* **1992**, *31*, 10226−38.
(114) Orengo, C. A.; Jones, D. T.; Thornton, J. M. *Nature* **1994**, *372*, 631−4.
(115) Niermann, T.; Kirschner, K. *Protein Eng.* **1991**, *4*, 359−70.
(116) Benner, S. A.; Cohen, M. A.; Gerloff, D. *Nature* **1992**, *359*, 781.
(117) Fauchere, J. L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. *Int. J. Pept. Protein Res.* **1988**, *32*, 269−78.
(118) Rose, G. D. *Nature* **1978**, *272*, 586−90.
(119) Luque, I.; Mayorga, O. L.; Freire, E. *Biochemistry* **1996**, *35*, 13681−13688.
(120) Lim, V. I. *J. Mol. Biol.* **1974**, *88*, 873−94.
(121) Eisenberg, D.; Wesson, M.; Wilcox, W. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G., Ed.; Plenum: New York, 1989.
(122) Matthews, B. W.; Nicholson, H.; Becktel, W. J. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6663−7.
(123) Alber, T. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G., Ed.; Plenum: New York, 1989.
(124) McCammon, J. A.; Wong, C. F.; Lybrand, T. P. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G., Ed.; Plenum: New York, 1989.
(125) Mackay, D. H. J.; Cross, A. J.; Hagler, A. T. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G., Ed.; Plenum: New York, 1989.
(126) Bohm, G.; Jaenicke, R. *Protein Sci.* **1992**, *1*, 1269−78.
(127) Gibson, T. J.; Postma, J. P.; Brown, R. S.; Argos, P. *Protein Eng.* **1988**, *2*, 209−18.
(128) Kolinski, A.; Skolnick, J. *Proteins* **1994**, *18*, 353−66.
(129) Srinivasan, R.; Rose, G. D. *Proteins* **1995**, *22*, 81−99.
(130) Dunbrack, R. L.; Gerloff, D. L.; Bower, M.; Chen, X. W.; Lichtarge, O.; Cohen, F. E. *Folding Des.* **1997**, *2*, R27−R42.
(131) Doolittle, R. F. *Protein Sci.* **1992**, *1*, 1563−77.
(132) Levin, J.; Garnier, J. *Biochim. Biophys. Acta* **1988**, *955*, 1177−1192.
(133) Donnelly, D.; Overington, J. P.; Blundell, T. L. *Protein Eng.* **1994**, *7*, 645−53.
(134) Nishikawa, K.; Ooi, T. *Biochim. Biophys. Acta* **1986**, *871*, 45−54.
(135) Benner, S. A.; Ellington, A. D. *CRC Crit. Rev. Biochem.* **1988**, *23*, 369−426.
(136) Sali, A. *Curr. Opin. Biotechnol.* **1995**, *6*, 437−51.
(137) May, A. C. W.; Blundell, T. L. *Curr. Opin. Biotechnol.* **1995**, *5*, 355−360.
(138) Brown, W. J.; North, A. C. T.; Phillips, D. C.; Brew, K.; Vanaman, T. C.; Hill, R. L. *J. Mol. Biol.* **1969**, *42*, 65−86.
(139) Rossmann, M. G.; Argos, P. *J. Mol. Biol.* **1976**, *105*, 75−95.
(140) Greer, J. *J. Mol. Biol.* **1981**, *153*, 1027−42.
(141) Blundell, T. L. *Food Chem. Toxicol.* **1995**, *33*, 979−85.
(142) Johnson, M. S.; Srinivasan, N.; Sowdhamini, R.; Blundell, T. L. *Crit. Rev. Biochem. Mol. Biol.* **1994**, *29*, 1−68.
(143) Crippen, G. M. *Proteins* **1996**, *26*, 167−171.
(144) Schiffer, C. A.; Caldwell, J. W.; Kollman, P. A.; Stroud, R. M. *Proteins* **1990**, *8*, 30−43.
(145) Ponder, J. W.; Richards, F. M. *J. Mol. Biol.* **1987**, *193*, 775−91.
(146) Laughton, C. A. *J. Mol. Biol.* **1994**, *235*, 1088−97.
(147) Harrison, R. W.; Chatterjee, D.; Weber, I. T. *Proteins* **1995**, *23*, 463−71.
(148) Moult, J.; Pedersen, J. T.; Judson, R.; Fidelis, K. *Proteins* **1995**, *23*, ii−v.
(149) Doolittle, R. F. *Of urfs and orfs: A primer on how to analyze derived amino acid sequences*; University Science Books: Mill Valley, 1986.
(150) Benner, S. A.; Cohen, M. A.; Gonnet, G. H. *Protein Eng.* **1994**, *7*, 1323−32.
(151) Vogt, G.; Etzold, T.; Argos, P. *J. Mol. Biol.* **1995**, *249*, 816−31.
(152) Argos, P. *Curr. Opin. Biotechnol.* **1995**, *5*, 361−371.
(153) Bowie, J. U.; Eisenberg, D. *Curr. Opin. Struct. Biol.* **1993**, *3*, 437−444.
(154) Bryant, S. H.; Altschul, S. F. *Curr. Opin. Struct. Biol.* **1995**, *5*, 236−44.
(155) Gribskov, M.; McLachlan, A. D.; Eisenberg, D. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 4355−8.
(156) Gribskov, M.; Luethy, R.; Eisenberg, D. *Meth. Enzymol.* **1990**, *183*, 146−59.
(157) Overington, J.; Donnelly, D.; Johnson, M. S.; Sali, A.; Blundell, T. L. *Protein Sci.* **1992**, *1*, 216−26.
(158) Bryant, S. H.; Lawrence, C. E. *Proteins* **1993**, *16*, 92−112.
(159) Jones, D. T.; Taylor, W. R.; Thornton, J. M. *Nature* **1992**, *358*, 86−9.
(160) Sippl, M. J. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 473−501.
(161) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534−552.
(162) Eddy, S. R. *Curr. Opin. Struct. Biol.* **1996**, *6*, 361−365.
(163) Miller, R. T.; Jones, D. T.; Thornton, J. M. *FASEB J.* **1996**, *10*, 171−8.

(164) Westhead, D. R.; Collura, V. P.; Eldridge, M. D.; Firth, M. A.; Li, J.; Murray, C. W. *Protein Eng.* **1995**, *8*, 1197−1204.
(165) Bryant, S. H. *Proteins* **1996**, *26*, 172−185.
(166) Madej, T.; Gibrat, J. F.; Bryant, S. H. *Proteins* **1995**, *23*, 356−69.
(167) Jones, D. T.; Miller, R. T.; Thornton, J. M. *Proteins* **1995**, *23*, 387−97.
(168) Lemer, C. M.; Rooman, M. J.; Wodak, S. J. *Proteins* **1995**, *23*, 337−55.
(169) Defay, T. R.; Cohen, F. E. *J. Mol. Biol.* **1996**, *262*, 314−323.
(170) Jones, D. T.; Thornton, J. M. *Curr. Opin. Struct. Biol.* **1996**, *6*, 210−6.
(171) Madej, T.; Boguski, M. S.; Bryant, S. H. *FEBS Lett.* **1995**, *373*, 13−8.
(172) Baumann, H.; Morella, K. K.; White, D. W.; Dembski, M.; Bailon, P. S.; Kim, H.; Lai, C. F.; Tartaglia, L. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8374−8.
(173) Zhang, F. M.; Basinski, M. B.; Beals, J. M.; Briggs, S. L.; Churgay, L. M.; Clawson, D. K.; DiMarchi, R. D.; Furman, T. C.; Hale, J. E.; et al. *Nature* **1997**, *387*, 206−209.
(174) Lesk, A. M. *Proteins Struct. Funct. Genet.* **1997**, *30*, 1−16.
(175) Wodak, S. J.; Rooman, M. J. *Curr. Opin. Struct. Biol.* **1993**, *3*, 247−259.
(176) Rost, B.; Schneider, R.; Sander, C. *Trends Biochem. Sci.* **1993**, *18*, 120−3.
(177) Burgess, A. W.; Scheraga, H. A. *J. Theor. Biol.* **1975**, *53*, 403−420.
(178) Levin, J. M.; Pascarella, S.; Argos, P.; Garnier, J. *Protein Eng.* **1993**, *6*, 849−54.
(179) Di Francesco, V.; Garnier, J.; Munson, P. J. *Protein Sci.* **1996**, *5*, 106−13.
(180) DeGrado, W. F.; Wasserman, Z. R.; Chowdhry, V. *Nature* **1982**, *300*, 379−81.
(181) Bewley, T. A.; Levine, H. L.; Wetzel, R. *Int. J. Pept. Protein Res.* **1982**, *20*, 93−6.
(182) Senda, T.; Shimazu, T.; Matsuda, S.; Kawano, G.; Shimizu, H.; Nakamura, K. T.; Mitsui, Y. *EMBO J.* **1992**, *11*, 3193−201.
(183) Murgolo, N. J.; Windsor, W. T.; Hruza, A.; Reichert, P.; Tsarbopoulos, A.; Baldwin, S.; Huang, E.; Pramanik, B.; Ealick, S.; Trotta, P. P. *Proteins* **1993**, *17*, 62−74.
(184) Mowbray, S. L.; Foster, D. L.; Koshland, D. E., Jr. *J. Biol. Chem.* **1985**, *260*, 11711−8.
(185) Milburn, M. V.; Prive, G. G.; Milligan, D. L.; Scott, W. G.; Yeh, J.; Jancarik, J.; Koshland, D. E., Jr.; Kim, S. H. *Science* **1991**, *254*, 1342−7.
(186) Moe, G. R.; Koshland, J. D. E. In *Microbial Energy Transduction, Genetics, Structure and Function of Membrane Proteins*; Youvan, D. C., Daldal, F., Eds.; Cold Spring Harbor Press: New York, 1986.
(187) Taylor, W. R.; Geisow, M. J. *Protein Eng.* **1987**, *1*, 183−7.
(188) Barton, G. J.; Newman, R. H.; Freemont, P. S.; Crumpton, M. J. *Eur. J. Biochem.* **1991**, *198*, 749−760.
(189) Pearl, L. H.; Taylor, W. R. *Nature* **1987**, *329*, 351−4.
(190) Bazan, J. F.; Fletterick, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 7872−6.
(191) Hyde, C. C.; Ahmed, S. A.; Padlan, E. A.; Miles, E. W.; Davies, D. R. *J. Biol. Chem.* **1988**, *263*, 17857−17871.
(192) Kyte, J.; Doolittle, R. F. *J. Mol. Biol.* **1982**, *157*, 105−32.
(193) Karplus, P. A.; Schulz, G. E. *Naturwissenschaften* **1985**, *72*, 212−213.
(194) Farber, G. K.; Petsko, G. A. *Trends Biochem. Sci.* **1990**, *15*, 228−34.
(195) Niermann, T.; Kirschner, K. *Meth. Enzymol.* **1991**, *202*, 45−59.
(196) Hurle, M. R.; Matthews, C. R.; Cohen, F. E.; Kuntz, I. D.; Toumadje, A.; Johnson, J., W. C. *Proteins: Struct., Funct., Genet.* **1987**, *2*, 210−224.
(197) Niermann, T.; Kirschner, K. *Protein Eng.* **1995**, *8*, 535−42.
(198) Tesmer, J. G.; Klem, T. J.; Deras, M. L.; Davisson, V. J.; Smith, J. L. *Nature Struct. Biol.* **1996**, *3*, 74−86.
(199) Chen, A.; Kroon, P. A.; Poulter, C. D. *Protein Sci.* **1994**, *3*, 600−7.
(200) Tarshis, L. C.; Yan, M.; Poulter, C. D.; Sacchettini, J. C. *Biochemistry* **1994**, *33*, 10871−7.
(201) Bazan, J. F. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 6934−8.
(202) de Vos, A. M.; Ultsch, M.; Kossiakoff, A. A. *Science* **1992**, *255*, 306−12.
(203) Hubbard, T.; Park, J. *Trends Biochem. Sci.* **1996**, *21*, 279−281.
(204) Bazan, J. F. *Proteins* **1996**, *24*, 1−17.
(205) Qian, N.; Sejnowski, T. J. *J. Mol. Biol.* **1988**, *202*, 865−84.
(206) Holley, L. H.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 152−6.
(207) Hirst, J. D.; Sternberg, M. J. *Biochemistry* **1992**, *31*, 7211−8.
(208) Rost, B.; Sander, C. *J. Mol. Biol.* **1993**, *232*, 584−99.
(209) Salzberg, S.; Cost, S. *J. Mol. Biol.* **1992**, *227*, 371−4.
(210) Benner, S. A. *J. Mol. Recog.* **1995**, *8*, 9−28.
(211) Rost, B.; Sander, C. *Nature* **1992**, *360*, 540.
(212) Gomis-Ruth, F. X.; Kress, L. F.; Bode, W. *EMBO J.* **1993**, *12*, 4151−7.
(213) Zhang, D.; Botos, I.; Gomis-Ruth, F. X.; Doll, R.; Blood, C.; Njoroge, F. G.; Fox, J. W.; Bode, W.; Meyer, E. F. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 8447−51.
(214) Bode, W.; Kress, L. F.; Meyer, E. F.; Gomis-Ruth, F. X. *Braz. J. Med. Biol. Res.* **1994**, *27*, 2049−68.
(215) Gomis-Ruth, F. X.; Kress, L. F.; Kellermann, J.; Mayr, I.; Lee, X.; Huber, R.; Bode, W. *J. Mol. Biol.* **1994**, *239*, 513−44.
(216) Hubbard, T. J.; Park, J. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 398−402.
(217) Jabri, E.; Carr, M. B.; Hausinger, R. P.; Karplus, P. A. *Science* **1995**, *268*, 998−1004.
(218) Rost, B.; Sander, C.; Schneider, R. *Comput. Appl. Biosci.* **1994**, *10*, 53−60.
(219) Hodgkin, E. E.; Gillman, I. C.; Gilbert, R. J. *Protein Sci.* **1994**, *3*, 984−6.
(220) Gonnet, G. H.; Cohen, M. A.; Benner, S. A. *Science* **1992**, *256*, 1443−5.
(221) Zuckerkandl, E. *Sci. Am.* **1965**, *212*, 110−118.
(222) *Molecular Evolution, Computer Analysis of Protein and Nucleic Acid Sequences*; Doolittle, R. F., Ed.; Academic Press: New York, 1990.
(223) King, J. L.; Jukes, T. H. *Science* **1969**, *164*, 788−98.
(224) Kimura, M. In *Molecular Evolution, Protein Polymorphism, and the Neutral Theory*; Kimura, M., Ed.; Springer-Verlag: Berlin, 1982; pp 3-56.
(225) Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. In *Atlas of Protein Sequence and Structure*; Dayhoff, M. O., Ed.; National Biomedical Research Foundation: Washington, DC, 1978; Vol. 5.
(226) Jones, D. T.; Taylor, W. R.; Thornton, J. M. *Comput. Appl. Biosci.* **1992**, *8*, 275−82.
(227) Perutz, M. F.; Lehmann, H. *Nature* **1968**, *219*, 902−9.
(228) Go, M.; Miyazawa, S. *Int. J. Pept. Protein Res.* **1980**, *15*, 211−24.
(229) Lim, W. A.; Sauer, R. T. *Nature* **1989**, *339*, 31−6.
(230) Hubbard, T. J.; Blundell, T. L. *Protein Eng.* **1987**, *1*, 159−71.
(231) Patthy, L. *Acta Biochim. Biophys. Hung.* **1989**, *24*, 3−13.
(232) Overington, J. P.; Johnson, M. S.; Sali, A.; Blundell, T. L. *Proc. R. Soc. London B.* **1990**, *241*, 132−145.
(233) Bowie, J. U.; Reidhaar-Olson, J. F.; Lim, W. A.; Sauer, R. T. *Science* **1990**, *247*, 1306−10.
(234) Benner, S. A.; Badcoe, I.; Cohen, M. A.; Gerloff, D. L. *J. Mol. Biol.* **1994**, *235*, 926−58.
(235) Cohen, M. A.; Benner, S. A.; Gonnet, G. H. *Biochem. Biophys. Res. Commun.* **1994**, *199*, 489−496.
(236) Cohen, F. E.; Abarbanel, R. M.; Kuntz, I. D.; Fletterick, R. J. *Biochemistry* **1983**, *22*, 4894−904.
(237) Pascarella, S.; Argos, P. *J. Mol. Biol.* **1992**, *224*, 461−71.
(238) Needleman, S. B.; Wunsch, C. D. *J. Mol. Biol.* **1970**, *48*, 443−53.
(239) Smith, T. F.; Waterman, M. S. *J. Mol. Biol.* **1981**, *147*, 195−7.
(240) Flory, P. A. *Principles of Polymer Chemistry*; Cornell Univ. Press: Ithaca, New York, 1953.
(241) Brant, D. A.; Flory, P. A. *J. Am. Chem. Soc.* **1965**, *87*, 2788−2791.
(242) Cohen, F. E.; Abarbanel, R. M.; Kuntz, I. D.; Fletterick, R. J. *Biochemistry* **1986**, *25*, 266−75.
(243) Brown, R. S.; Argos, P. *Nature* **1986**, *324*, 215.
(244) Kimura, M. *Molecular Evolution, Protein Polymporphism and the Neutral Theory*; Springer-Verlag: Berlin, 1982; pp 3−56.
(245) Benner, S. A. *Curr. Opin. Struct. Biol.* **1992**, *2*, 402−412.
(246) McClure, M. A.; Vasi, T. K.; Fitch, W. M. *Mol. Biol. Evol.* **1994**, *11*, 571−92.
(247) Knighton, D. R.; Zheng, J. H.; Ten Eyck, L. F.; Ashford, V. A.; Xuong, N. H.; Taylor, S. S.; Sowadski, J. M. *Science* **1991**, *253*, 407−14.
(248) Benner, S. A.; Jenny, T. F.; Cohen, M. A.; Gonnet, G. H. *Adv. Enzyme Regul.* **1994**, *34*, 269−353.
(249) Riddihough, G. *Nat. Struct. Biol.* **1994**, *1*, 265−266.
(250) Wentrup, C. *Reactive Molecules*; Wiley: New York, 1984.
(251) Tauer, A.; Benner, S. A. *Proc. Nat. Acad. Sci. U.S.A.* **1997**, *94*, 53−58.
(252) Fry, D. C.; Kuby, S. A.; Mildvan, A. S. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 907−11.
(253) Shoji, S.; Ericsson, L. H.; Walsh, K. A.; Fischer, E. H.; Titani, K. *Biochemistry* **1983**, *22*, 3702−9.
(254) Taylor, S. S.; Buechler, J. A.; Slice, L. W.; Knighton, D. K.; Durgerian, S.; Ringheim, G. E.; Neitzel, J. J.; Yonemoto, W. M.; Sowadski, J. M.; Dospmann, W. *Cold Spring Harbor Symp. Quant. Biol.* **1988**, *53*, 121−30.
(255) Taylor, S. S. *J. Biol. Chem.* **1989**, *264*, 8443−6.
(256) Sternberg, M. J. E.; Taylor, W. R. *FEBS Lett.* **1984**, *175*, 387−92.
(257) Bork, P. *Current Opin. Struct. Biol.* **1992**, *2*, 413−421.
(258) Gonnet, G. H.; Benner, S. A. “Computational biochemistry research at ETH,” E. T. H. Department Informatik, 1991.
(259) Kim, J.; Rees, D. C. *Nature* **1992**, *360*, 553−560.
(260) Benner, S. A.; Cohen, M. A.; Gerloff, D. *J. Mol. Biol.* **1993**, *229*, 295−305.
(261) Noble, M. E. M.; Musacchio, A.; Saraste, M.; Courtneidge, S. A.; Wierenga, R. K. *EMBO J.* **1993**, *12*, 2617−2624.

(262) Biou, V.; Gibrat, J. F.; Levin, J. M.; Robson, B.; Garnier, J. *Protein Eng.* **1988**, *2*, 185−91.

(263) Musacchio, A.; Gibson, T.; Lehto, V. P.; Saraste, M. *FEBS Lett.* **1992**, *307*, 55−61.

(264) Panayotou, G.; Bax, B.; Gout, I.; Federwisch, M.; Wroblowski, B.; Dhand, R.; Fry, M. J.; Blundell, T. L.; Wollmer, A.; Waterfield, M. D. *EMBO J.* **1992**, *11*, 4261−72.

(265) Russell, R. B.; Breed, J.; Barton, G. J. *FEBS Lett.* **1992**, *304*, 15−20.

(266) Waksman, G.; Kominos, D.; Robertson, S. C.; Pant, N.; Baltimore, D.; Birge, R. B.; Cowburn, D.; Hanafusa, H.; Mayer, B. J.; Overduin, M.; et al. *Nature* **1992**, *358*, 646−53.

(267) Musacchio, A.; Gibson, T.; Rice, P.; Thompson, J.; Saraste, M. *Trends Biochem. Sci.* **1993**, *18*, 343−8.

(268) Jenny, T. F.; Benner, S. A. *Proteins* **1994**, *20*, 1−3.

(269) Haslam, R. J.; Koide, H. B.; Hemmings, B. A. *Nature* **1993**, *363*, 309−10.

(270) Mayer, B. J.; Ren, R.; Clark, K. L.; Baltimore, D. *Cell* **1993**, *73*, 629−30.

(271) Yoon, H. S.; Hajduk, P. J.; Petros, A. M.; Olejniczak, E. T.; Meadows, R. P.; Fesik, S. W. *Nature* **1994**, *369*, 672−5.

(272) Macias, M. J.; Musacchio, A.; Ponstingl, H.; Nilges, M.; Saraste, M.; Oschkinat, H. *Nature* **1994**, *369*, 675−7.

(273) Gerloff, D. L.; Cohen, F. E. *Proteins* **1996**, *24*, 18−34.

(274) Jeffrey, P. D.; Russo, A. A.; Polyak, K.; Gibbs, E.; Hurwitz, J.; Massague, J.; Pavletich, N. P. *Nature* **1995**, *376*, 313−20.

(275) Gibson, T. J.; Thompson, J. D.; Blocker, A.; Kouzarides, T. *Nucl. Acids Res.* **1994**, *22*, 946−52.

(276) Lees, E. M.; Harlow, E. *Mol. Cell. Biol.* **1993**, *13*, 1194−201.

(277) Bazan, J. F. *Science* **1992**, *257*, 410−3.

(278) Roach, P. L.; Clifton, I. J.; Fulop, V.; Harlos, K.; Barton, G. J.; Hajdu, J.; Andersson, I.; Schofield, C. J.; Baldwin, J. E. *Nature* **1995**, *375*, 700−4.

(279) Yee, V. C.; Pedersen, L. C.; Le Trong, I.; Bishop, P. D.; Stenkamp, R. E.; Teller, D. C. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 7296−300.

(280) Takahashi, N.; Takahashi, Y.; Putnam, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 8019−23.

(281) Livingstone, C. D.; Barton, G. J. *Int. J. Pept. Protein Res.* **1994**, *44*, 239−44.

(282) Edwards, Y. J.; Perkins, S. J. *FEBS Lett.* **1995**, *358*, 283−6.

(283) Johnson, M. S.; Overington, J. P.; Blundell, T. L. *J. Mol. Biol.* **1993**, *231*, 735−52.

(284) Lee, J. O.; Rieu, P.; Arnaout, M. A.; Liddington, R. *Cell* **1995**, *80*, 631−8.

(285) Barford, D.; Flint, A. J.; Tonks, N. K. *Science* **1994**, *263*, 1397−404.

(286) Barton, G. J.; Cohen, P. T.; Barford, D. *Eur. J. Biochem.* **1994**, *220*, 225−37.

(287) Griffith, J. P.; Kim, J. L.; Kim, E. E.; Sintchak, M. D.; Thomson, J. A.; Fitzgibbon, M. J.; Fleming, M. A.; Caron, P. R.; Hsiao, K.; Navia, M. A. *Cell* **1995**, *82*, 507−22.

(288) Barford, D.; Jia, Z.; Tonks, N. K. *Nature Struct. Biol.* **1995**, *2*, 1043−1053.

(289) Lupas, A.; Koster, A. J.; Walz, J.; Baumeister, W. *FEBS Lett.* **1994**, *354*, 45−9.

(290) Cohen, B. I.; Presnell, S. R.; Cohen, F. E. *Meth. Enzymol.* **1991**, *202*, 252−68.

(291) Loewe, J.; Stock, D.; Jap, B.; Zwickl, P.; Baumeister, W.; Huber, R. *Science* **1995**, *268*, 533−9.

(292) Gerloff, D. L.; Benner, S. A. *Proteins* **1995**, *21*, 273−81.

(293) Leng, B.; Buchanan, B. G.; Nicholas, H. B. *J. Comp. Biol.* **1994**, *1*, 25−38.

(294) Rost, B.; Sander, C. *Proteins* **1995**, *23*, 295−300.

(295) Munson, P. J.; Di Francesco, V.; Porrelli, R. *27th Anual Hawaii International Conference on Systems Science* **1994**, *5*, 375−384.

(296) Harris, G. W.; Jenkins, J. A.; Connerton, I.; Pickersgill, R. W. *Acta Crystallogr. D.* **1996**, *52*, 393−401.

(297) Gerloff, D. L.; Chelvanayagam, G.; Benner, S. A. *Proteins* **1995**, *22*, 299−310.

(298) Sutton, R. B.; Davletov, B. A.; Berghuis, A. M.; Sudhof, T. C.; Sprang, S. R. *Cell* **1995**, *80*, 929−38.

(299) Floeckner, H.; Braxenthaler, M.; Lackner, P.; Jaritz, M.; Ortner, M.; Sippl, M. J. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 376−86.

(300) Woolfson, D. N.; Evans, P. A.; Hutchinson, E. G.; Thornton, J. M. *Protein Eng.* **1993**, *6*, 461−70.

(301) Bycroft, M.; Proctor, M.; Freund, S. M.; St Johnston, D. *FEBS Lett.* **1995**, *362*, 333−6.

(302) Zanotti, G.; Berni, R.; Monaco, H. L. *J. Biol. Chem.* **1993**, *268*, 10728−38.

(303) Petratos, K.; Banner, D. W.; Beppu, T.; Wilson, K. S.; Tsernoglou, D. *FEBS Lett.* **1987**, *218*, 209−14.

(304) Davies, C.; White, S. W.; Ramakrishnan, V. *Structure* **1996**, *4*, 55−66.

(305) Gallagher, D. T.; Gilliland, G. L.; Wang, L.; Bryan, P. *Structure* **1995**, *3*, 907−914.

(306) Pai, K. S.; Bussiere, D. E.; Wang, F.; Hutchison, C. A., III; White, S. W.; Bastia, D. *EMBO J.* **1996**, *15*, 3164−3173.

(307) Pellequer, J. L.; Westhof, E.; Van Regenmortel, M. H. *Immunol. Lett.* **1993**, *36*, 83−99.

(308) Weinhold, E. G.; Glasfeld, A.; Ellington, A. D.; Benner, S. A. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 8420−4.

(309) Bairoch, A. *Nucleic Acids Res.* **1991**, *19*, 2241−5.

(310) Taylor, W. R. *Comput. Chem.* **1993**, *17*, 117.

(311) Russell, R. B.; Copley, R. R.; Barton, G. J. *J. Mol. Biol.* **1996**, *259*, 349−65.

(312) Monge, A.; Friesner, R. A.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5027−9.

(313) Neher, E. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 98−102.

(314) Taylor, W. R.; Hatrick, K. *Protein Eng.* **1994**, *7*, 341−8.

(315) Shindyalov, I. N.; Kolchanov, N. A.; Sander, C. *Protein Eng.* **1994**, *7*, 349−58.

(316) Gobel, U.; Sander, C.; Schneider, R.; Valencia, A. *Proteins* **1994**, *18*, 309−17.

(317) Cohen, F. E.; Sternberg, M. J.; Taylor, W. R. *Nature* **1980**, *285*, 378−82.

(318) Valencia, A.; Hubbard, T. J.; Muga, A.; Banuelos, S.; Llorca, O.; Carrascosa, J. L.; Valpuesta, J. M. *Proteins* **1995**, *22*, 199−209.

(319) Hunt, J. F.; Weaver, A. J.; Landry, S. J.; Gierasch, L.; Deisenfoger, J. *Nature* **1996**, *379*, 37−45.

(320) Lesk, A. *J. Mol. Graphics* **1995**, *13*, 159−164.

(321) Bouaziz, S. V., C.; Huet, J C.; Pernollet, J C.; Guittet, E. *Biochemistry* **1994**, *33*, 8188−8197.

(322) Hutchinson, E. G.; Thornton, J. M. *Proteins* **1990**, *8*, 203−212.

(323) Solovyev, V. V.; Salamov, A. A. *Comput. Appl. Biosci.* **1994**, *10*, 661−9.

(324) Gallagher, T. Personal communication, 1997.

(325) Subramanian, A. R. *Prog. Nucleic Acid Res. Mol. Biol.* **1981**, *28*, 101−142.

(326) Giorginis, S.; Subramanian, A. R. *J. Mol. Biol.* **1983**, *141*, 393−408.

(327) Régnier, P.; Grunberg-Manago, M.; Portier, C. *J. Biol. Chem.* **1987**, *262*, 63−68.

(328) Gribskov, M. *Gene* **1992**, *119*, 107−111.

(329) Bycroft, M.; Hubbard, T. J.; Proctor, M.; Freund, S. M.; Murzin, A. G. *Cell* **1997**, *88*, 235−242.

(330) Gerloff, D. L.; Cohen, F. E.; Benner, S. A. *Proteins: Struct., Funct., Genet.* **1997**, *27*, 279−289.

(331) Yee, V.; Teller, D. C. *Structure* **1997**, *5*, 125−138.

(332) Beamer, L. J.; Carroll, S. F.; Eisenberg, D. *Science* **1997**, *276*, 1861−1864.

(333) Gerloff, D. L. C.; Fred, E.; Korostensky, C.; Turcotte, M.; Gonnet, G. H.; Benner, S. A *Proteins: Struct., Funct., Genet.* **1997**, *27*, 450−458.

(334) Wigley, D. B.; Davies, G. J.; Dodson, E. J.; Maxwell, A.; Dodson, G. *Nature* **1991**, *351*, 624−629.

(335) Jakob, U.; Scheibel, T.; Bose, S.; Reinstein, J.; Buchner, J. *J. Biol. Chem.* **1996**, *271*, 10035−10041.

(336) Henikoff, S.; Henikoff, J. G. *Proteins* **1993**, *17*, 49−61.

(337) Prodromou, C.; Roe, S. M.; Piper, P. W.; Pearl, L. H. *Nature Struct. Biol.* **1997**, *4* (6), 477−82.

(338) Groves, M. R.; Taylor, M. A. J.; Scott, M.; Cummings, N. J.; Pickersgill, R. W.; Jenkins, J. A. *Structure* **1996**, *4*, 1193−1203.

(339) Shrive, A. K.; Polikarpov, I.; Krell, T.; Coulson, A.; Hawkins, A.; Sawyer, L. *Nat. Struct. Biol.* **1997**, submitted.

(340) Hofmann, E.; Wrench, P. M.; Sharples, F. P.; Hiller, R. G.; Welte, W.; Diederichs, K. *Science* **1996**, *272*, 1788−1791.

(341) Al-Karadaghi, S.; Hansson, M.; Nikonov, S.; Jonsson, B.; Hederstedt, L. *EMBO J.* **1997**, submitted.

(342) Seemann, J. E.; Schulz, G. E. *J. Mol. Biol.* **1997**, *273*, 256−268.

(343) Holliger, P.; Riechmann, L. *Structure* **1997**, *5*, 265−275.

(344) Vath, G. M.; Earhart, C. A.; Rago, J. V.; Kim, M. H.; Bohach, G. A.; Schlievert, P. M.; Ohlendort, D. H. *Biochemistry* **1997**, *36*, 1559−1566.

(345) Boissy, G.; de La Fortelle, E.; Kahn, R.; Huet, J. C.; Bricogne, G.; Pernollet, J. C.; Brunie, S. *Structure* **1996**, *4*, 1429−1439.

(346) Carugo, K. D.; Banuellos, S.; Saraste, M. *Nat. Struct. Biol.* **1997**, *4*, 175−179.

(347) Johnson, P. E.; Joshi, M. D.; Tomme, P.; Kilburn, D. G.; McIntosh, L. P. *Biochemistry* **1996**, *35*, 14381−14394.

(348) Liepinsh, E.; Andersson, M.; Ruysschaert, J. M.; Otting, G. *Nat. Struct. Biol.* **1997**, *4*, 793−795.

(349) Taylor, W. R.; Jones, D. T.; Green, N. M. *Proteins* **1994**, *18*, 281−94.

(350) Rees, D. C.; DeAntonio, L.; Eisenberg, D. *Science* **1989**, *245*, 510−3.

(351) Chothia, C. *Nature* **1992**, *357*, 543−544.

CR940469A