Pergamon

Advan. Enzyme Regul., Vol. 38, pp. 155–180, 1998

Printed in Great Britain
0065-2571/98/$19.00 + 0.00

PII: S0065-2571(97)00019-8

POST-GENOMIC SCIENCE: CONVERTING PRIMARY STRUCTURE INTO PHYSIOLOGICAL FUNCTION

STEVEN A. BENNER, NATHALIE TRABESINGER and DAVID SCHREIBER

Departments of Chemistry, Anatomy and Cell Biology, University of Florida, Gainesville, FL 32611, USA

INTRODUCTION

The 1990's is the decade of the genome. At its midpoint, complete sequences were available for the genomes of several eubacteria, an archaebacterium, and a eukaryote (yeast). Before the decade is out, a half dozen additional bacterial genomes will be added to this list, together with much of the genomes of *C. elegans* and *H. sapiens*. To these will be added sequences from dozens of other organisms, collected with varying degrees of systematic effort.

As these data have accumulated, it has become increasingly apparent that new methods are needed to exploit the information that they (must) contain. Chemistry has always been driven by the discovery of new natural products, elucidation of their structures, and exploration of their behaviors. The genome database provides an enormous new collection of natural product structures to study. These display every behavior of interest to chemists: conformation, supramolecular organization, combinatorial assembly, photochemistry, and catalysis are just a few. Organic chemistry *should* be revolutionized by genomic data. But how?

A similar question can be framed for the biomedical sciences. Pharmaceutical and genome corporations worldwide are collecting and cataloging sequence data, comparing the expressed genetic inventory of diseased and normal tissues, and attempting to correlate genomic data with physiological function. It seems that the treatment of human disease should be revolutionized by genomic sequence data. But it is not obvious precisely how this will happen.

Last, genomic projects present both problems and opportunities for the science of exobiology, defined as the study of the origin, evolution and distribution of life (including life on earth) within the context of cosmic evolution. Throughout much of its history, exobiology has been viewed as

a science without a subject matter. With the recent advances in planetary science, including landing on Mars and close inspection of the moons of Jupiter, questions central to the biochemistry of life important to exobiological research have become more relevant. In particular, it is important now to distinguish features of terrestrial life that reflect unique solutions to problems presented by life from those that do not. The first are likely to be mirrored in life that originated independently on other planets, the second are not. Unique solutions are likely to arise from constraints imposed by fundamental chemical reactivity (which is assumed to be universal) and Darwinian processes that drive organisms to optimize chemical behavior, also assumed to be universal. Phrased this way, exobiological research seeks to identify and distinguish chemical features of terrestrial life that reflect selection, neutral drift, and origins. Genome projects, it turns out, help address these issues as well.

One consequence of genomic sequences from a variety of organisms is the ready availability of the *evolutionary histories* of families of proteins represented within the genome sequence databases. Sequence data will be organized as a set of approximately 10,000 independently evolving protein sequence "modules". For each of these, an evolutionary history can be built that will consist of a multiple alignment of the sequences of the proteins in the module themselves (as well as their encoding DNA sequences), an evolutionary tree, and a reconstructed ancestral DNA and protein sequence for each branch point in the tree. Given a detailed model of biomolecular evolution, these histories can be used to connect sequence, structure, chemical reactivity, and biological function.

Over the past decade, we have used *Advances in Enzyme Regulation* and its associated conference organized in the Indiana University School of Medicine by Dr. George Weber to lay out tools that exploit evolutionary analysis of sequence data to solve problems in biological chemistry. These have included methods to identify functional regions of protein structure (1), predict the conformation of proteins from a family of homologous sequences (2), analyze evolutionary covariation at residues distant in the polypeptide chain (2), and use protein structure prediction to detect long distance homologs (2). This article lays out "post-genomic" tools that exploit evolutionary histories directly to take the next step: to extract information concerning protein structure, behavior, and function from a detailed understanding of how protein sequences divergently evolve under functional constraints. In a post genomic world, with volumes of sequence data from an unlimited number of organisms, these tools will be used widely to learn from sequence data about living systems, their chemistry and their diseases.

The Conventional Evolutionary Paradigm

Since the mid 1970's, it has been known that homologous proteins—those related by common ancestry—have analogous conformations (folds) (3, 4), at least in their core domains. From this observation has emerged many tools, including profile analysis, homology modeling, and threading (5, 6), that help biological chemists model the conformation of a protein from the conformation of one of its homologs.

If homologous proteins have analogous conformations, it might be reasoned that homologous proteins might be analogous in other ways as well. Biomolecular behavior depends in part on conformation. Thus, two homologous proteins with analogous conformations might have analogous behaviors. The reasoning can be carried further. As biomolecular behavior is important for biomolecular function, perhaps two homologous proteins will also have analogous functions.

This train of logic has been viewed as a starting point for analyzing genomic data. Genome sequencing projects generate sequences of proteins without any supporting experimental data describing the protein itself, data that accompany most sequences generated in classical biochemical research. While applications for genomic sequences are many, virtually all require that a structure, behavior, or function be assigned to proteins known only as sequences. Chemical theory is insufficient to allow us to assign structure, behavior, or function directly to a sequence. But bioinformatic theory is adequate to use the genomic sequence to identify homologous proteins in the database. Some of these homologs may have known structures, behaviors, or physiological functions. If homology implies structural analogy, and (from there) behavioral analogy, and (from there) functional analogy, then the task of assigning structure, behavior, or function to the genomic data should become trivial whenever a homolog with a known structure, behavior, or function can be identified.

Many tools for analyzing genomic sequence data are based on this logic, which we shall call the *conventional evolutionary paradigm*. The conventional evolutionary paradigm is implemented throughout bioinformatics research in various simple ways. Consider the following recipe, used in a variety of commercial software packages to analyze a new sequence for a new protein collected in a genome project:

(a) The new sequence is first recorded.
(b) The sequence is then used as a probe in a BLAST search done against the existing sequence database.
(c) Proteins in the database that have sequences similar to the new sequence (by some scoring criterion) are recorded. These are putative homologs of the new sequence.

(d) The "functions" of the putative homologs are read from their documentation in the database.
(e) The function of the new sequence is presumed to be the same as that of the homologs.

This approach has many well known limitations. First, a BLAST search need not find an analogous sequence in the database whose function is already known. In many cases, a BLAST search fails to find *any* sequence in the database with a statistically significant sequence similarity. If no putative homolog can be identified, then it follows that no homolog with known structure, behavior or function can be identified.

In other cases, a BLAST search identifies one or more putative homologs, but the homologs have not been studied sufficiently to assign a structure, behavior, or function. Again, the paradigm fails to resolve the problem.

More frequently, the BLAST search identifies *many* possible homologs in the database, all with statistically marginal sequence similarities. The approach frequently presents the user with the documentation from several of these, and leaves the user to decide which (if any) of the putative functions are correct. As the literature shows, weak sequence similarities reflecting distant homologies have been both profoundly informative and profoundly deceptive (2). Very frequently, the functions of these putative homologs are widely different, making it difficult to decide which, if any, function should be imputed to the probe sequence.

These problems are all well recognized as limitations of the conventional evolutionary paradigm. Less well recognized, however, are problems with the elements of the logic upon which the conventional evolutionary paradigm is based. Proteins with analogous folds need not have analogous behaviors. Proteins with analogous behaviors need not have analogous functions. As has been reviewed in detail elsewhere (7–10), evolution is a potent process for recruiting old protein folds to perform new functions, and many cases are known where the conventional evolutionary paradigm would generate (indeed, has generated) incorrect conclusions. For example, fumarase (functioning in the citric acid cycle), adenylosuccinate lyase (functioning in nucleotide biosynthesis) and aspartate ammonia lyase (functioning in amino acid metabolism) are all identified as homologs by a BLAST search. Yet their behaviors are analogous only at the level of organic reaction mechanism, and there only at the most abstract level. The conventional evolutionary paradigm fails to suggest correct conclusions in this case. The recent work of Babbitt, Kenyon, Gerlt, and Petsko has added other fascinating examples of how microorganisms can recruit a common fold to catalyze a variety of reactions from the racemization of mandelic acid to the opening/closing of a lactone (11).

In proteins involved in "advanced" functions (for example, developmental biology) in more complex organisms, difficulties with the "homology-implies-analogous-structure/behavior/function" assumption underlying the conventional evolutionary paradigm become confounding. For example, protein serine kinases and protein tyrosine kinases are clearly identifiable homologs at the level of sequence similarity. The chemist would say that both classes of protein operate using analogous reaction mechanisms, differing only in the source of the oxygen nucleophile in the phosphoryl transfer reaction. The biologist would note that the physiological function of the two classes of proteins are greatly different, however. For any biomedical application, the biologist would be correct. The physiologically relevant differences in behavior, central to the understanding of biological function (phosphorylation on tyrosine versus phosphorylation on serine) cannot be inferred for one family from the other using the conventional evolutionary paradigm.

The deeper the chemistry of developmental biology is probed in metazoa (multicellular animals), the more it becomes apparent that function in the Darwinian sense can change with very little change in sequence (see also the summary in references 7–10). For example, a variety of src homology 2 (SH2) domains all bind peptide sequences containing phosphotyrosine residues. The binding specificities of the different SH2 domains are different, however, for the peptide sequences surrounding the phosphotyrosine residues. It is these specificities that determine which protein binds to each individual SH2 domain. The physiologically relevant function of each SH2 domain centers on this pairwise interaction. Thus, any statement of function for any particular SH2 domain must at least identify its phosphotyrosine-containing partner. To assign function at this level, the conventional evolutionary paradigm has little to say.

Post-genomic Science: Modeling Molecular Evolution

These considerations suggest that if it is to be used to analyze genomic data, evolutionary theory must be applied at a more sophisticated level than the level exploited by the conventional evolutionary paradigm. Much of this sophistication is available at the present time. Let us examine briefly how biomolecular evolution is modeled as a starting point to designing tools for interpreting genomic data.

The Markov model. Virtually all of contemporary genomic science is based in some sense on an analysis where sequence data are treated simply as strings of characters. This analysis is based on a specific model of sequence evolution at the level of proteins. Overwhelmingly, these models are in some sense "Markovian". They assume, implicitly or otherwise, that variation in a protein sequence occurs independently at each position

in the sequence, that future mutations occur independently of past mutations, and that a single substitution matrix adequately describes the relative likelihood of each amino acid being replaced by any other amino acid during an episode of molecular evolution. Further, simplifying assumptions are made when scoring gaps (indels) in a pairwise alignment.

The primary virtue of this model is its simplicity. Yet the model is certainly false. Proteins are not linear strings of independent characters. Rather, they are folded structures that have arisen by divergent evolution under constraints imposed by natural selection seeking adaptive function. Because real proteins fold in three dimensions, mutations at different positions in the protein sequence are not independent (12). Nor are future mutations independent of past mutations (12). Insertions and deletions display complex patterns that reflect functional constraints on protein evolution (13).

Here, the second virtue of Markovian models, their exactness, becomes useful. Markovian models generate specific expectations. These can be explicitly violated by the behavior of real proteins during divergent evolution, and the extent of the violation can frequently be quantitated. The differences between the expected and actual evolution of proteins contain clues concerning how proteins fold, function, and create new function. Let us examine some tools to extract these clues.

Non-Markovian protein evolution as a post-genomic tool for structure prediction. Tools that extract information from the non-Markovian behavior of proteins undergoing divergent evolution have been exceptionally valuable for predicting the conformation of proteins from an evolutionary history (14). These tools have made major strides towards solving a problem that as recently as a decade ago was considered to be unsolvable. The approach is described in detail previously, including on these pages (1, 2), and will not be described here. Rather, we shall assume that reasonably accurate (if not perfect) models of protein secondary structure can be predicted for a family of homologous protein sequences. We will then ask how these predictions might be exploited to solve one of the problems noted above with the conventional evolutionary paradigm: the need to have methods more powerful than simple sequence analysis to detect long distance homologs.

Structure prediction as a tool for identifying long distance homologs. The core of a protein fold is conserved during divergent evolution long after the sequences within the family have diverged so much that homology is no longer evident by sequence analysis alone. In the mid 1970's, Rossman and his coworkers suggested that analogous folds in two proteins might indicate that the proteins are homologs, even when the sequences of the

two proteins bear no statistically significant similarities (3, 4). This observation prompted many groups to develop methods for building models from protein sequences starting from marginal sequence similarities (5, 6). These are known as profile or threading approaches.

In practical application, the primary difficulty with these approaches is that they generate too many "hits", or matchings of a probe sequence against a sequence database. The analyses frequently return many proteins that are possible homologs. As has been shown in many joint prediction projects (such as the Critical Assessment of Structure Prediction, or CASP projects (15)), these hits are sometimes correct and informative. Equally likely, however, the putative homologs identified by a threading or profile analysis are not true homologs, or are misaligned with true homologs using the analysis. Especially needed, therefore, are tools to confirm or (especially) deny the possibility that a target identified from a sub-statistical sequence similarity is a homolog.

In the early 1990's, the first example was presented where a structure prediction was used to critically evaluate suggestions, based on sub-statistical sequence similarities, that two proteins were homologous. The case, discussed on these pages, concerned the protein kinases (2). Protein kinase contains the sequence motif Gly-Xxx-Gly-Xxx-Xxx-Gly (where Xxx is any amino acid) preceded by a strand and followed by a helix (16). A similar motif was found in adenylate kinase, where a crystal structure was known. Therefore, following the conventional evolutionary paradigm, several groups proposed that the two structures were homologs. This proposal implied in turn that protein kinase would adopt the same fold as adenylate kinase. Several groups built models based on this inference (17–20).

Contrasting with this view was a model built for the secondary and tertiary structure of the protein kinase family based on the evolutionary history of the protein. In this model, the Gly-Xxx-Gly-Xxx-Xxx-Gly motif was predicted to be flanked by beta strand both before *and* afterwards. Thus, the predicted secondary structural model of protein kinase was not congruent with the experimental structure of adenylate kinase. Accordingly, the prediction noted that the two folds could not be the same. From this, it was inferred that the two proteins could not be homologs.

This alternative model based on post-genomic methods was later shown to be correct by a subsequently determined crystal structure (21). This was perhaps the first time that a predicted structure and a motif analysis had been used to infer the *absence* of homology between two families catalyzing analogous chemical reactions. The tool used in the protein kinase prediction proved to be an example of a more general tool to confirm or deny long distance homology between the two protein families.

Using this tool, core secondary structural elements of the protein families are aligned sequentially. In this process, the secondary structural elements are considered to be congruent when every core element from one family finds a core element in the other of the same type (helix or strand), in the same order, where gaps matched against non-core elements (where a non-core element in one family is not aligned against any element in the other) are allowed in any number, and a core element in one protein may be missing in the other, but may not be aligned with a core element in the other of a different type (i.e., helix against strand).

This approach is especially powerful when coupled with motif analysis, as was done for protein kinase. Congruent secondary structural elements can indicate whether a motif is a significant indicator of homology or not. Thus, flanking a motif, a secondary structural model might have one of four forms: helix-motif-helix, helix-motif-strand, strand-motif-helix, and strand-motif-strand. The secondary structural alignments are congruent if and only if the secondary structural elements flanking the motifs correspond between the two proteins. Homology is not denied if and only if the secondary structures are congruent. This method is preferably applied when each family contains proteins that are at least 120 point accepted mutations per 100 amino acids (PAM units) divergent.

Important to this tool is a distinction between core and non-core secondary structural elements. The failure of core secondary structural elements to correspond between two protein families is a clear contradiction of homology. But non-core elements need not be conserved over long periods of divergent evolution, and failure to find a correspondence between non-core elements from one protein family in another does not exclude homology. Several ways of defining "core" exist in this context. Some involve crystal structure data; others do not.

When an experimental structure is known for a protein (for example, by crystallography or n.m.r.), core elements are conveniently defined geometrically; a core element is one where a substantial fraction is buried in the protein fold unexposed to solvent water. Most useful is the application of this concept to beta strands in a beta sheet. A core strand is one that forms backbone hydrogen bonding interactions with two other strands on both of its edges. Thus, 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. Core strands are highly conserved during divergent evolution. If they are lost, the sheet in which they participate cannot be conserved.

A more general definition of a core secondary structural unit focuses on the evolutionary stability of the secondary structural unit. For the purpose of detecting long distance homologs, a core secondary structural element, predicted or otherwise, is one that cannot be lost during divergent evol-

ution without damaging the integrity of the protein fold. This is based on notions of continuity in protein evolution, most fundamentally on the assumption that a protein that has one "topology" of protein fold (e.g., an eight fold alpha-beta barrel) cannot by continuous evolutionary processes be converted into a protein with another (e.g., an immunoglobulin fold). It is clear that divergence of biological function can add or subtract peripheral secondary structural elements to create or remove contact elements, expand or eliminate binding sites, or to modify the performance of the protein in other fashions.

In this case, non-core segments of a protein family are recognized from a family of sequences, preferably between 100 and 150 PAM units divergent for the most divergent pairs. If a segment (including a segment containing a helix or a strand) is deleted in a protein family built from members all sharing significant sequence similarity, it cannot be essential for the integrity of the fold in the family. In applying this tool, one must be concerned about database mistakes; a pair of sequence that is "deleted" because the scientist providing the entry into the database neglected to collect it, or neglected to enter it, is not a deletion from the purpose of detecting non-core segments.

A third method for identifying a core segment of a protein sequence is applicable to any alignment containing three or more sequences. 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 of 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.

A final method relates to the reconstructed ancestral sequence of the protein. It has long been appreciated (22) that when the sequences of two or more homologous proteins are available, it is possible to construct a probabilistic model for the sequence of the ancestral protein. The part of the ancestral sequence that is reconstructed with high probability is the "core" of the protein. These reconstructions are done by well-known maximum likelihood tools (for example, as implemented within Darwin, available via the web at the address cbrg.inf.ethz.ch, see also (23)). A core is defined from the ancestral sequences as a segment of the multiple alignment where the average probability of the most frequent amino acid at that position is greater than one standard deviation above the average probability of all of the reconstructed positions in the multiple alignment. This is a tree-weighted measure of the divergence in the family as a whole, and correlates with core regions defined in the other ways, as the region of the ancestral sequence that is reconstructed with high probability is also the one that has not suffered insertions and deletions, and the one

that has seen relatively little sequence divergence. These segments also correlate with core segments defined geometrically.

Structure predictions made using post-genomic tools have now been applied in several cases to make statements about long distant homology and, from there, catalytic behavior. One of the most dramatic was for the heat shock protein 90 (HSP 90) family. The predicted secondary structural elements were assembled to yield a tertiary fold that resembled closely the fold determined in the N-terminal fragment of DNA gyrase B (the ATPase fragment) (24). This analysis was published before an experimental structure of HSP 90 was known (25), and after an experimental study had suggested that HSP 90 did not have catalytic activity as an ATPase. The prediction thus generated a statement about catalytic behavior (and, presumably, physiological function) in addition to secondary structure.

Post-genomic prediction tools were also applied to the family of ribonucleotide reductases (26). Here, proteins with highly divergent sequences were shown to belong to one universal family using a combination of protein structural analysis and gene sequencing. The tools in this case confirmed a speculation by Stubbe and coworkers based on a mechanistic analysis that all ribonucleotide reductases are related by common ancestry (27).

Comparing predicted secondary structure models for a protein family is now a proven technique with an impressive track record to detect or deny long distance homologs based on sequence data alone. As genome projects are completed; and evolutionary histories for their constituent protein families articulated, secondary structural models for those families should become routinely available. Thus, these tools should solve the first problem with practical implementation of the conventional evolutionary paradigm—the difficulty of detecting homologs using standard alignment tools.

Recruitment Of Function

Predicted secondary structures of a protein family offer an approach to solve the general problem of finding distant homologs in a database. They do not, however, address other problems associated with the conventional evolutionary paradigm for assigning function to sequences—those that arise when the assumptions underlying the conventional evolutionary paradigm are invalid. This is especially the case when one protein family gives rise to proteins with different function. This possibility can never be excluded *a priori*. In the case of the heat shock protein 90, for example, the post-genomic prediction tool showed that the fold of SHP 90 was analogous to the fold of gyrase. It suggested (under the conventional evolutionary paradigm) that the behavior and function of HSP 90 was analogous to the behavior and function of gyrase. Some of these sugges-

tions may in fact be true. But they need not be. The gyrase fold could be recruited in HSP 90 to perform many other (indeed, *any* other) function.

The premise for the post-genomic sciences, however, is that the evolutionary histories of protein families will become generally available as a result of massive genome sequencing projects. Non-Markovian behavior is not only expected to arise because of folded structure. In addition, patterns of sequence evolution are expected to violate the Markovian model differently if the protein in question is undergoing an episode where function is being changed.

On these pages in 1988, the power of this approach was illustrated to identify the active site residues of mammalian alcohol dehydrogenase (E.C. 1.1.1.1). Mammalian alcohol dehydrogenases have undergone a rapid episode of sequence evolution in and around the active site as substrate specificity has divergently evolved to handle xenobiotic substances in the liver. In contrast, over a comparable span of evolutionary distance, the active site of yeast alcohol dehydrogenase has changed very little, corresponding to an apparently constant role of the enzyme to act on the ethanol-acetaldehyde redox couple. Indeed, by identifying positions in mammalian dehydrogenases where amino acid variation was observed over a span of evolution where the same residues were conserved in the yeast dehydrogenases provided a clear map of the active site of the protein.

A particularly clever use of this approach has recently been described by Lichtarge *et al.* These workers described an *evolutionary trace method* that defined functionally significant residues as those that are conserved within a family (28). They then used this approach to identify patches on the surface of proteins that contribute to functionality.

The approaches work because the scientist understands something about the function of a protein family. Thus, the active site of alcohol dehydrogenase could be identified because the scientist knew that substrate specificity is conserved in one branch of a protein family (yeast alcohol dehydrogenases) but not in another (liver alcohol dehydrogenases). The evolutionary trace method works only when the function being traced is conserved within the family. Neither approach is applicable when the evolutionary status of the protein function (conserved or not conserved) is not known. In general, this status will not be known to the post-genomic scientists examining only genomic data. To apply these post-genomic tools, therefore, the post-genomic scientist needs a tool for learning whether function is changing within the family in an episode of protein sequence evolution—one based on an analysis of the sequence data alone.

One tool is based on the fact that the genetic code is degenerate. More than one triplet codon encodes the same amino acid. Therefore, a mutation in a gene can be either silent (not changing the encoded amino

acid) or expressed (changing the encoded amino acid). Especially in multicellular organisms, and most particularly in multicellular animals (metazoa), silent changes are not under selective pressure. In contrast, expressed changes at the DNA level, by changing the structure of the protein that the gene encodes, change the property of the protein. This frequently places these changes under selective behavior.

The outcome of different selection pressures on silent and expressed mutations is non-Markovian behavior in the evolution of DNA sequences. Consider three cases. If the first, we examine an episode of protein sequence evolution during a period of evolutionary history where, at the outset of the period, the behavior of a protein whose sequence has been perfectly optimized for a specific biological function, and where that function remains constant for the protein throughout the period being examined. During this period, changes in the DNA sequence that lead to a change in the sequence of the encoded protein (expressed changes) will diminish the survival value of the protein (7) and therefore will be removed by natural selection. Silent changes will not be removed by natural selection, but will accumulate at an approximately clock-like rate, as silent changes are approximately neutral, especially in higher organisms. Thus, the ratio of expressed to silent changes will be low during a period of evolution of a protein family where the ancestor and its descendants share a common function.

A second case concerns a period of evolution where a protein is acquiring a new derived function. Its amino acid sequence at the beginning of this episode will be optimized for the ancestral function, rather than the derived function. To be optimally suited for the derived function, therefore, amino acids must be changed in the protein. Thus, changes in the gene that are expressed will have a chance of improving the behavior of the protein *vis a vis* its new biological function, and these will be selected for. The ratio of expressed to silent substitutions at the DNA level will be high, and a high expressed/silent ratio will reveal a period of evolution of a protein family where the function of the ancestor is changing.

In a third case, consider the evolution of a gene encoding a protein that has no function (a pseudogene, for example), neutrally drifting without functional constraints. In this case, the expressed/silent ratio will reflect random introduction of point mutations. Given the genetic code and a typical distribution of amino acid codons within the gene, a ratio of expressed to silent changes will be approximately 2.5 during the period of evolution of a protein family where the ancestor and its descendants have no function.

Stewart and Messier introduced this approach through a clever and systematic analysis of conservation and variation within the lysozyme family (29). Episodes of rapid sequence evolution were identified by high

expressed/silent ratios in specific branches of the evolutionary trees. These were correlated with specific episodes in the evolution of the protein itself and the physiology of the organisms that contained the protein.

This approach can be illustrated in a biomedically interesting family of proteins by examining the protein leptin, a protein whose mutation in mice is evidently correlated with obesity, and was previously known as the "obesity gene protein". The protein has attracted substantial interest in the pharmaceutical industry, especially after a human gene encoding a leptin homolog was isolated. According to the conventional evolutionary paradigm, because it is a homolog of the mouse leptin, the human leptin must also play a role in obesity, and might be an appropriate target for pharmaceutical companies seeking human pharmaceuticals to combat this common condition in the first world.

DNA and protein sequences were retrieved for the genes encoding leptins. A multiple alignment for the protein sequences was constructed for the DNA sequences and the protein sequences. Congruent trees for both the DNA and protein sequences were then constructed, and sequences at the nodes of the tree reconstructed using MacClade (30) and the known relationship between the organisms from which these sequences were derived. For the DNA sequences, the biologically most plausible tree proved to be the most parsimonious tree as well. The most parsimonious tree for the protein sequences proved *not* to be the most plausible tree (by one change) from a biological perspective. The DNA tree was taken to be definitive because of its consistency with the biological (cladistic) data.

A secondary structure prediction was made for the protein family. The evolutionary divergence of the sequences available for the leptin family is small—only 21 PAM units (point accepted mutations per 100 amino acids)—and predictions were biased to favor surface assignments (31). Thus, positions holding conserved KREND were assigned as surface residues, conserved H and Q were assigned to the surface as well, while positions holding conserved CST were assigned as uncertain.

Five separate secondary structural elements were identified. The results are summarized in Fig. 1. A disulfide bond was presumed to connect positions 96 and 146. These secondary structural elements can be accommodated by only a small number of overall folds. Interestingly, the pattern of secondary structure in this prediction is consistent with an overall fold that resembles that seen in cytokines such as colony stimulating factor (32) and human growth hormone (33).

To decide whether evolutionary function may have changed under selected pressure during the divergent evolution of the protein family, silent and expressed mutations were assigned to individual branches on the evolutionary tree. For each branch of the tree, the sum of the number

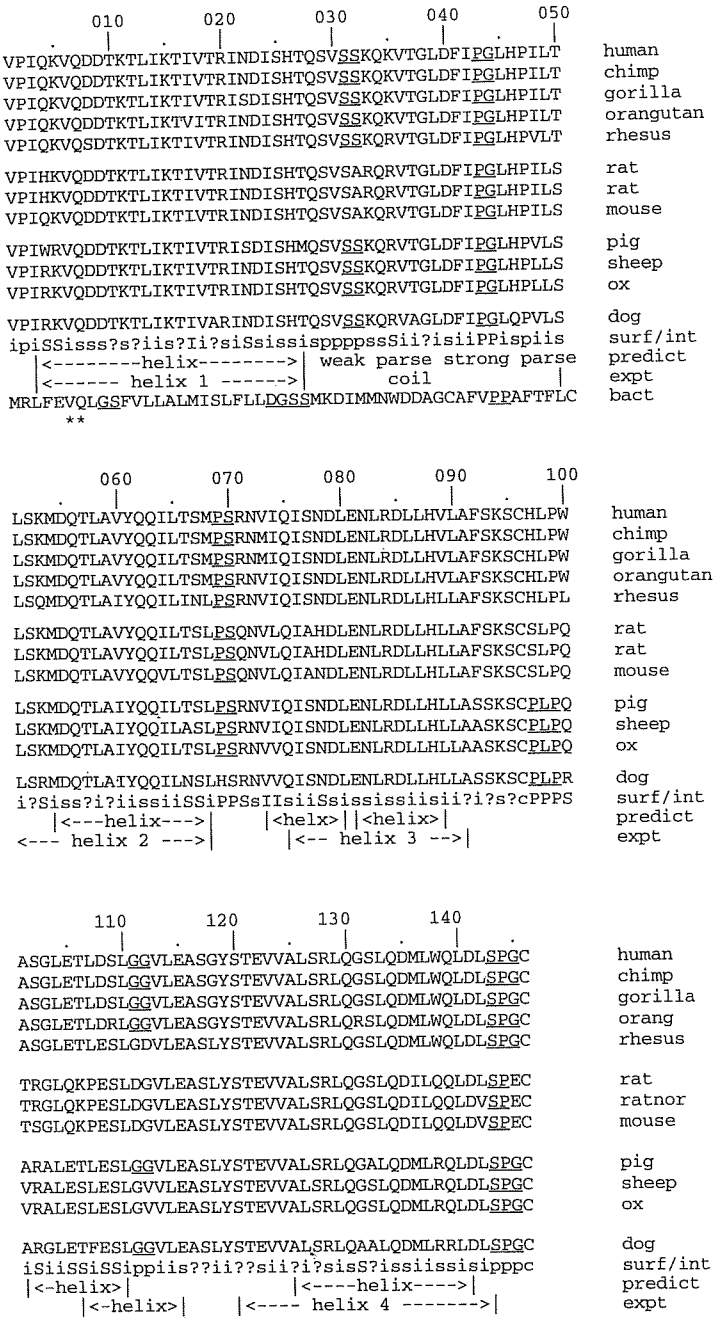

FIG. 1. Predicted surface, interior, and secondary structure assignments for leptin. S and s, I and i indicate strong and weak surface and interior assignments respectively. P and p indicate strong and weak parses respectively. A "?" indicates that no assignment is made. A "c" indicates that the position is involved in a disulfide bond. Secondary structure was assigned using the method of Benner and Gerloff (2) where positions denoted "?" were permitted to fall in either the surface or interior arc of the helix. Underlined residues are part of parsing strings.

of silent and expressed changes were tabulated, and the ratio of expressed to silent changes calculated. These are shown in Fig. 2.

The branches on the evolutionary tree leading to the primate leptins from their ancestors at the time that rodents and primates diverged had an extremely high ratio of expressed to silent changes. From this analysis, it was concluded that the biological function of leptins has changed significantly in the primates relative to the function of the leptin in the common ancestor of primates and rodents. This conclusion has several implications of importance, not the least being for pharmaceutical companies asked whether they should explore leptins as a pharmaceutical target. At the very least, it suggests that the mouse is not a good pharmacological model for compounds to be tested for their ability to combat obesity in humans. The post-genomic analysis suggests that a primate model must

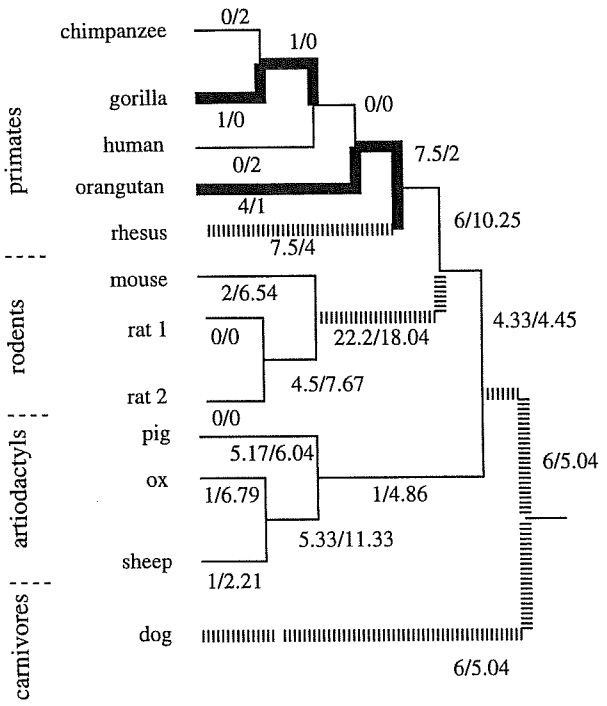

FIG. 2. Evolutionary tree showing the evolutionary history of the leptins. Heavy lines show branches with expressed/silent ratios higher than 2. Hatched lines show branches with expressed/silent ratios from 1 to 2. Thin, solid lines show branches with expressed/silent ratios less than 1. Numbers on the lines indicate the ratio of expressed/silent changes for that branch. The branch lengths do not correspond to either geological time or evolutionary distance.

be used to test those compounds, with implications for the cost of developing an anti-obesity drug based on the leptin protein.

Intriguingly, a tree can also be built for the leptin receptor (Fig. 3). Here, the evolutionary history is not so complete. In particular, fewer primate sequences are available for the leptin receptor than for leptin itself. Thus, the reconstructed ancestral sequences are less precise with the leptin receptor family, and the assignment of expressed and silent mutations to the tree are less certain. Nevertheless, it appears that the leptin receptor has undergone an episode of rapid sequence evolution in the primate half of the family as well. The example illustrates how much sequence data is needed (much) to build reliable models of this nature, as the ambiguity in the assignment of ancestral sequences makes it possible that the receptor was evolving rapidly not only in the lineage leading to primates but also in the lineage leading to mouse.

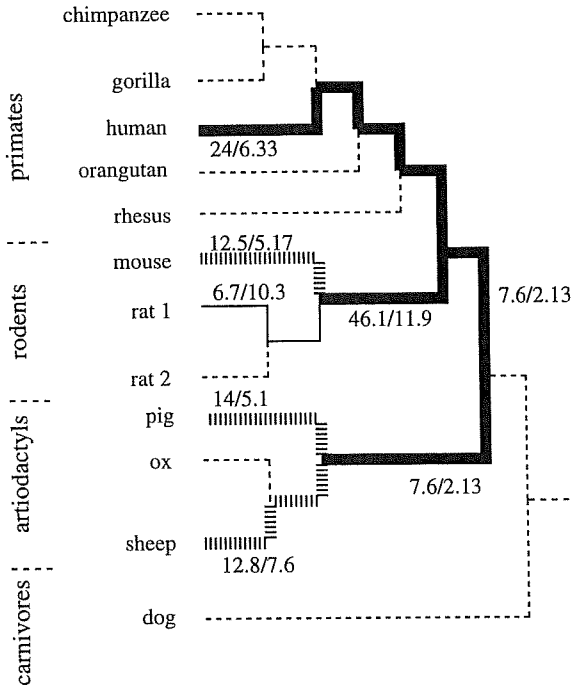

FIG. 3. Evolutionary tree showing the evolutionary history of the extracellular domain of the leptin receptors. Heavy lines show branches with expressed/silent ratios higher than 3. Hatched lines show branches with expressed/silent ratios from 2 to 3. Thin, solid lines show branches with expressed/silent ratios less than 2. Notice the greater overall divergence in the leptin receptor than in leptin itself. Dotted lines show branches with indeterminate ratios. Numbers on the lines indicate the ratio of expressed/silent changes for that branch. The branch lengths do not correspond to either geological time or evolutionary distance.

Nevertheless, the approximate correlation between the episode of rapid sequence evolution in the leptin family and in the leptin receptor family suggests a tool that might become useful in the advanced stages of post-genomic science when evolutionary histories are very well articulated. Here, it might be possible to detect ligand-receptor relationships between protein families in the database by a correspondence between their episodes of rapid sequence evolution. Thus, ligand families should evolve rapidly (in a non-Markovian fashion) at the same time in geological history as their receptors evolve. It will be interesting to identify more sequences for primate leptin receptors to see if a more complete evolutionary history allows us to see more clearly the co-evolution of the leptin receptor and leptin itself.

Correlating the Paleontological Record with Episodes of Sequence Evolution

As discussed above, detailed analyses of evolutionary histories frequently can provide a solution to the most general problem of the conventional evolutionary paradigm, the difficulty in routinely identifying a homolog of a target sequence with known function within the database. By analysis of non-Markovian evolutionary behavior at the level of the protein, a model of secondary structure can be predicted. This prediction can be used in turn to detect long distance homologs in some cases and exclude the possibility of distant homology in others. This increases the likelihood that a homolog will be found with a known structure, behavior, or function for a new protein sequence. If one is found, then the logic associated with the conventional evolutionary paradigm can be applied to generate a hypothesis concerning the behavior or function of the protein.

The value of this post-genomic tool to assign behavior and structure to a target sequence problem is expected to grow over the near term, as the ratio of sequences supported by experimental studies to those not supported increases with the conclusion of genome projects, and as more sequences increase the detail of the evolutionary histories that can be extracted from the database directly, and therefore the quality of the predicted secondary structural model.

At the next level, analysis of non-Markovian behavior at the level of the gene can alert the biological chemist that the logic associated with the conventional evolutionary paradigm might not apply in individual cases. In particular, if an episode of rapid sequence evolution intervenes in the evolutionary tree between the sequence of interest and the sequence with the known behavior and function, the biological chemist is alerted to the possibility that the function of the protein might have changed. This alert is useful even with close homologs, as illustrated in the example with leptin.

But what if the evolutionary tree contains *no* protein with a sequence with assigned function, even one with low sequence similarity? Even with more limited evolutionary histories, post-genomic tools that analyze non-Markovian evolution at the level of the codon can be useful. By identifying the organisms that provide the sequences at the "leaves" of the evolutionary tree, it is frequently possible to correlate branches in the evolutionary tree with episodes in geological history, as determined from the fossil record. Especially in multicellular animals (metazoa), the fossil record can provide approximate dates for the emergence of new physiological function. In this case, it is possible to ask whether an episode of rapid sequence evolution in a protein family (in particular, an episode with a high expressed/silent ratio) occurred at the same time as a new physiological function emerged on earth. If so, a first level of hypothesis about physiological function can be proposed, even if no behavior or function of any kind is known for any of the modern proteins.

Perhaps the most transparent analysis of this type concerns proteins that underwent massive radiative divergences in metazoa approximately 600 million years ago. This is the time of the *Cambrian explosion*, an episode in terrestrial history that marks the massive radiative divergence of multicellular animals, including chordates. Protein families undergoing rapid evolution at this time (for example, of protein tyrosine kinases and src homology 2 domains) are almost certainly involved in the basic processes by which multicellular animals develop from a single fertilized egg.

This type of analysis might be applied in the family of ribonuclease (RNase) A (E.C. 2.7.7.16), a well known family of digestive proteins found in ruminants. The protein underwent rapid sequence evolution approximately 45 million years ago, a time where ruminant digestion emerged in mammals (34). Thus, the rapid molecular evolution evident in the reconstructed evolutionary history of this protein suggests that the protein is important for ruminant digestive function.

Identification of in vitro *Behaviors that Contribute to Physiological Function*

In vitro experiments in biological chemistry extract data on proteins and nucleic acids (for example) that are removed from their native environment, often in pure or purified states. While isolation and purification of molecules and molecular aggregates from biological systems is an essential part of contemporary biological research, the fact that the data are obtained in a non-native environment raises questions concerning their physiological relevance. Properties of biological systems determined *in vitro* need not correspond to those *in vivo*, and properties determined *in vitro* need have no biological relevance *in vivo*.

To date, there has been no simple way to say whether or not biological behaviors are important physiologically to a host organism. Even in those cases where a relatively strong case can be made for physiological relevance (for example, for enzymes that catalyze steps in primary metabolism), it has proven to be difficult to decide whether individual properties of that enzyme (k_{cat}, K_m, kinetic order, stereospecificity, etc.) have physiological relevance. Especially difficult, however, is to ascertain which behaviors measures *in vitro* play roles in "higher" function in metazoa, including digestion, development, regulation, reproduction, and complex behavior.

Analysis of non-Markovian behavior, as described above, permits the biological chemist to identify episodes in the history of a protein family where new function is emerging. This suggests a general method to determine whether a behavior measured *in vitro* is important to the evolution of new physiological function. We may take the following steps:

(a) Prepare in the laboratory proteins that have the reconstructed sequences corresponding to the ancestral proteins before, during, and after the evolution of new biological function (34), as revealed by an episode of high expressed to silent ratio of substitution in a protein. This high ratio compels the conclusion that the protein itself serves a physiological role, one that is changing during the period of rapid non-Markovian sequence evolution.

(b) Measure in the laboratory the behavior in question in ancestral proteins before, during, and after the evolution of new biological function, as revealed by an episode of high expressed to silent ratio of substitution. Those behaviors that increase during this episode are deduced to be important for physiological function. Those that do not are not.

An example of this method was applied to the bovine seminal ribonuclease (RNase) family. Bovine seminal RNase diverged from bovine pancreatic RNase approximately 35 million years ago. Seminal RNase represents approximately 2% of the total protein in bovine seminal plasma. It displays antispermatogenic activity (35), immunosuppressive activity (36–38), and cytostatic activity against many transformed cell lines (39, 40). Each of these biological activities is essentially absent from pancreatic RNase. Further, seminal RNase binds to anionic glycolipids, binds and melts duplex DNA, hydrolyzes duplex RNA, has a dimeric quaternary structure, and binds to spermatozoa.

Each of these behaviors is measured *in vitro*, as is the case for a wide range of biological phenomenology recorded in the literature. The behaviors are difficult to interpret. Some, any, or all of the behaviors might serve an adaptive role. It is possible that none of these behaviors serve

adaptive roles. Indeed, it is conceivable that the protein has no adaptive role at all. This makes it difficult to make even the simplest research decisions, as the only *in vitro* properties of a protein that are interesting to study are those that have a physiological function.

To resolve these issues using the post-genomic method outlined above, genes for seminal and pancreatic RNases were obtained from a variety of organisms closely related to *Bos taurus*, using cloning procedures well known in the art. These were then sequenced, and a maximum parsimony tree was constructed using MacClade. From this tree were calculated the sequences of RNases that were intermediates in the evolution of the seminal RNase, using the maximum parsimony method and checked using maximum likelihood tools implemented in Darwin (23).

Next, the ratio of expressed to silent substitutions was calculated along each branch of the evolutionary tree. A very high ratio of expressed to silent substitutions was observed in the evolutionary period following the divergence of cape buffalo (41) from the lineage leading to ox, until the divergence of water buffalo and ox. This is indicative of an episode of adaptive evolution, where the protein acquires a new physiological function. Further work indicated that the seminal RNase gene was not expressed in the period of evolution since the divergence of the seminal RNase family and the divergence of cape buffalo.

Last, protein engineering methods were used to prepare the seminal RNase that existed at the beginning of the episode of rapid sequence evolution. Its properties were then examined experimentally. It was discovered that the ability of the protein to bind to anionic glycolipids was roughly the same before and after this episode of rapid evolution. So too was its sensitivity to inhibition by placental RNase inhibitor. Thus, both of these properties are not likely to be under selective pressure.

In contrast, the immunosuppressivity of the ancestral RNase (IC_{50} ca. 8 micrograms/mL) was greater than that of pancreatic RNase (IC_{50} ca. 100 micrograms/mL) (J. Sleasman, M. Rojas, personal communication). But following the period of rapid sequence evolution characteristic of a protein evolving to serve a new physiological function, the immunosuppressivity became still greater (IC_{50} ca. 2 micrograms/mL). Thus, one concludes that immunosuppressivity as measured *in vitro* is a selected trait of the protein, or is closely structurally coupled to a trait that is selected.

Likewise, the ability of the seminal RNase protein to bind and melt duplex DNA, and to hydrolyze duplex RNA, also underwent rapid increases between the time of divergence of cape buffalo from modern ox. Thus, it too is either a selected trait of the protein, or is closely structurally coupled to a trait that is selected. In contrast, dimeric structure did not emerge during this period. Dimeric structure, therefore, is presumably not as important to the new selected function of the protein, although it

may be a trait that was initially useful in the selection of the system for further optimization during the period of rapid evolution.

Structure Prediction and a Rapidly Searchable Database

The overarching problem with genomic sequence databases is their sheer size. This makes them tedious to search, and nearly impossible to subject to exhaustive self-matching (42). Predicted structures can be connected with reconstructed evolutionary histories to resolve this problem, to build a rapidly searchable database. Consider the following steps:

(a) A multiple alignment, an evolutionary tree, and ancestral sequences at nodes in the tree are constructed for a set of homologous proteins.
(b) A corresponding multiple alignment is constructed by methods well known in the art for the DNA sequences that encode the proteins in the protein family. This multiple alignment is constructed in parallel with the protein alignment. In regions of gaps or ambiguities, the amino acid sequence alignment is adjusted to give the alignment with the most parsimonious DNA tree.
(c) Mutations in the DNA sequences are then assigned to each branch of the DNA evolutionary tree. These may be fractional mutations to reflect ambiguities in the sequences at the nodes of the tree. When ambiguities are encountered, alternatives are weighted equally. Mutations along each branch are then assigned as being "silent", meaning that they do not have an impact on the encoded protein sequence, and "expressed", meaning that they do not have an impact on the encoded protein sequence. Fractional assignments are made in the case of ambiguities in the reconstructed sequences at nodes in a tree.
(d) A prediction is then made for each protein family. A secondary structure is predicted for the family, and this predicted secondary structure is aligned with the ancestral sequence at the root of the tree. If the root of the tree is unassigned, the predicted secondary structure is aligned with the ancestral sequence calculated for an arbitrary point near the center of gravity of the tree.
(e) An ancestral sequence is then reconstructed at nodes on the tree, and at a point on the tree as near as possible to its *root*—the point on the tree representing the oldest (geologically) sequence.

Steps (a) through (e) provide a method to organize the protein sequence database in a rapidly searchable form. The ancestral sequences and the predicted secondary structures associated with the families defined by steps (a) through (e) are surrogates for the sequences and structures of the individual proteins that are members of the family. The reconstructed ancestral sequence represents in a single sequence all of the sequences of

the descendent proteins. The predicted secondary structure associated with the ancestral sequence represents in a single structural model all of the core secondary structural elements of the descendent proteins. Thus, the ancestral sequences can replace the descendent sequences, and the corresponding core secondary structural models can replace the secondary structures of the descendent proteins.

This makes it possible to define two surrogate databases, one for the sequences, the other for secondary structures. The first surrogate database is the database that collects from each of the families of proteins in the databases a single ancestral sequence, at the point in the tree that most accurately approximates the root of the tree. If the root cannot be determined, the ancestral sequence chosen for the surrogate sequence database is near the center of mass of the tree. The second surrogate database is a database of the corresponding secondary structural elements. The surrogate databases are much smaller than the complete databases that contain the actual sequences or actual structures for each protein in the family, as each ancestral sequence represents many descendent proteins. Further, because there is a limited number of protein families on the planet, there is a limit to the size of the surrogate databases. Based on our work with partial sequence databases (42), we expect there to be fewer than 10,000 families as defined by steps (a) through (e).

Searching the surrogate databases for homologs of a probe sequence proceeds in two steps. In the first, the probe sequence (or structure) is matched against the database of surrogate sequences (or structures). As there will be on the order of 10,000 families of proteins as defined by steps (a) through (e) after all the genomes are sequenced for all of the organisms on earth, there will be only on the order of 10,000 surrogate sequences to search. Thus, this search will be far more rapid than with the complete databases. A probe protein sequence (or DNA sequence in translated form) can be exhaustively matched (42) against this surrogate database (that is, every subsequence of the probe sequence will be matched against every subsequence in the ancestral proteins) more rapidly than it could be matched against the complete database.

Should the search yield a significant match, the probe sequence is identified as a member of one of the families already defined. The probe sequence is then matched with the members of this family to determine where it fits within the evolutionary tree defined by the family. The multiple alignment, evolutionary tree, predicted secondary structure and reconstructed ancestral sequences may be different once the new probe sequence is incorporated into the family. If so, the different multiple alignment, evolutionary tree, and predicted secondary structure are recorded, and the modified reconstructed ancestral sequence and structure are incorporated into their respective surrogate databases for future use.

The advantage of this data structure over those presently used is apparent. As presently organized, sequence and structure databases treat each entry as a distinct sequence. Each new sequence that is determined increases the size of the database that must be searched. The database will grow roughly linearly with the number of organismal genomes whose sequences are completed, and become increasingly more expensive to search.

The surrogate database will not grow linearly. Most of the sequence families are already represented in the existing database. Addition of more sequences will therefore, in most cases, simply refine the ancestral sequences and associated structures. In any case, the total number of sequences and structures in their respective databases will not grow past ca. 10000—the estimate for the total number of sequence families that will be identifiable after the genomes of all organisms on earth are sequenced. If a dramatically new class of organism is identified, this estimate may grow, but not exponentially (as is the growth of the present database).

Conclusions

The evolutionary histories of protein families are now becoming routinely available through genome sequencing projects. These histories comprise a multiple alignment for their protein sequences and the corresponding DNA sequences, an evolutionary tree showing the pedigree of these sequences, and reconstructed ancestral sequences for each node in the tree. In a post-genomic world having genomic sequences from an unlimited number of organisms, these histories will be used to connect structure, chemical reactivity, and physiological function to these families.

This paper describes several "post-genomic" tools that exploit these evolutionary histories. Predictions of secondary structure can be made via an analysis of non-Markovian evolution at the level of the protein, and used to confirm or deny long distance homology between two protein families. Analysis of non-Markovian behavior at the level of the gene can identify proteins within a family that have new functions. Coupling these analyses with the paleontological record can suggest hypotheses for physiological function in families where no function is suggested by any experimental work for any of its members. Coupling these analyses with experimental work reconstructing ancestral proteins can identify specific *in vitro* properties of the protein that are important for its physiological role. Last, evolution-based data structures can be used to organize large sequence databases in a fashion that allows them to be searched with extreme efficiency. Together, these post-genomic tools can join with classical methods whose value is now becoming widely recognized (43).

ACKNOWLEDGEMENTS

We are indebted to the National Aeronautics and Space Administration for support.

REFERENCES

1. S. A. BENNER, Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure, *Advan. Enzym. Regul.* **28,** 219–236 (1989).
2. S. A. BENNER and D. L. GERLOFF, Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure. The catalytic domain of protein kinases, *Advan. Enzyme Regul.* **31,** 121–181 (1991).
3. M. G. ROSSMAN and P. ARGOS, Exploring structural homology of proteins, *J. Mol. Biol.* **105,** 75–95 (1976).
4. C. CHOTHIA and A. M. LESK, The relation between the divergence of sequence and structure in proteins, *EMBO J.* **5,** 823–826 (1986).
5. M. GRIBSKOV, A. D. MCLACHLAN and D. EISENBERG, Profile analysis: Detection of distantly related proteins, *Proc. Nat. Acad. Sci.* **84,** 4355–4358 (1987).
6. A. C. W. MAY and T. L. BLUNDELL, Automated comparative modelling of protein structures, *Curr. Opin. Biotech.* **5,** 355–360 (1995).
7. S. A. BENNER and A. D. ELLINGTON, Interpreting the behavior of enzymes. Purpose of pedigree? *CRC Crit. Rev. Biochem.* **23,** 369–426 (1988).
8. S. A. BENNER, A. GLASFELD and J. A. PICCIRILLI, Stereospecificity in enzymology. Its place in evolution, *Topics in Stereochem.* **19,** 127–207 (1989).
9. S. A. BENNER, Enzyme kinetics and molecular evolution, *Chem. Rev.* **89,** 789–806 (19??).
10. S. A. BENNER and A. D. ELLINGTON, Evolution and structural theory. The frontier between chemistry and biochemistry, *Bioorganic Chemistry Frontiers* **1,** 1–70 (1990).
11. P. C. BABBITT, G. T. MRACHKO, M. S. HASSON, G. W. HUISMAN, R. KOLTER, D. RINGE, G. A. PETSKO and G. L. KENYON, Functionally diverse enzyme superfamily that abstracts the alpha-protons of carboxylic acids, *Science* **267,** 1159–1161 (1995).
12. M. A. COHEN, S. A. BENNER and G. H. GONNET, Analysis of mutation during divergent evolution. The 400 by 400 dipeptide mutation matrix, *Biochem. Biophys. Res. Comm.* **199,** 489–496 (1994).
13. S. A. BENNER, M. A. COHEN and G. H. GONNET, Empirical and structural models for insertions and deletions in the divergent evolution of proteins, *J. Mol. Biol.* **229,** 1065–1082 (1993).
14. S. A. BENNER, G. CANNAROZZI, G. CHELVANAYAGAM and M. TURCOTTE, *Bona fide* predictions of protein secondary structure using transparent analyses of multiple sequence alignments, *Chem. Rev.* **97,** 2725–2843 (1997).
15. J. MOULT, The current state-of-the-art in protein-structure prediction, *Current Opin. Biotech.* **7,** 422–427 (1996).
16. M. J. E. STERNBERG and W. R. TAYLOR, Modeling the ATP binding site of oncogene products, the epidermal growth-factor receptor and related proteins, *FEBS Lett.* **175,** 387–392 (1984).
17. W. R. TAYLOR, Identification of protein sequence homology by consensus template alignment, *J. Mol. Biol.* **188,** 233–258 (1986).
18. W. R. TAYLOR and J. M. THORNTON, Recognition of super-secondary structure in proteins, *J. Mol. Biol.* **173,** 487–514 (1984).
19. R. K. WIERENGA, P. TERPSTRA and W. G. J. HOL, Prediction of the occurrence of the ADP-binding beta-alpha-beta fold in proteins using an amino acid sequence fingerprint, *J. Mol. Biol.* **187,** 101–107 (1986).
20. S. SHOJI, D. C. PARMELEE, R. D. WADE, S. KUMAR, L. H. ERICSSON, K. A. WALSH, H. NEURATH, H. L. LONG, J. G. DEMAILLE, E. H. FISCHER and K. TITANI, Complete amino acid sequence of the catalytic subunit of bovine cardiac muscle cyclic AMP-dependent protein kinase, *Proc. Nat. Acad. Sci.* **78,** 848–851 (1981).

21. D. R. KNIGHTON, J. ZHENG, L. TEN EYCK, F. V. A. ASHFORD, N. H. XUONG, S. S. TAYLOR and J. M. SOWADSKI, Crystal structure of the catalytic subunit of cyclic adenosine-monophosphate dependent protein-kinase, *Science* **253,** 407–414 (1991).
22. E. ZUCKERKANDL and L. PAULING, Evolutionary divergence and convergence in proteins, pp. 97–166 in *Evolving Genes and Proteins*, (V. BRYSON and H. J. VOGEL, eds.). Academic Press, New York (1965).
23. G. H. GONNET and S. A. BENNER, *Computational Biochemistry Research at ETH*, Technical Report 154, Departement Informatik (March 1991).
24. D. B. WIGLEY, G. J. DAVIES, E. J. DODSON, A. MAXWELL and G. DODSON, Crystal structure of an N-terminal fragment of the DNA gyrase B protein, *Nature* **351,** 624–629 (1991).
25. D. L. GERLOFF, F. E. COHEN, C. KOROSTENSKY, M. TURCOTTE, G. H. GONNET and S. A. BENNER, A predicted consensus structure for the N-terminal fragment of the heat shock protein HSP90 family, *Proteins: Struct. Funct. Genet.* **27,** 450–458 (1997).
26. A. TAUER and S. A. BENNER, The B12-dependent ribonucleotide reductase from the archaebacterium *Thermoplasma acidophila*. An evolutionary conundrum, *Proc. Natl. Acad. Sci.* **94,** 53–58 (1997).
27. S. S. MAO, T. P. HOLLER, G. X. YU, J. M. BOLLINGER JR, S. BOOKER, M. I. JOHNSTON and J. STUBBE, A model for the role of multiple cysteine residues involved in ribonucleotide reduction: Amazing and still confusing, *Biochemistry* **31,** 9733–9743 (1992).
28. O. LICHTARGE, H. R. BOURNE and F. E. COHEN, An evolutionary trace analysis defines binding surfaces common to protein families, *J. Mol. Biol.* **257,** 342–358 (1996).
29. W. MESSIER and C-B. STEWART, Episodic adaptive evolution of primate lysozymes, *Nature* **385,** 151–154 (1996).
30. W. P. MADDISON and D. R. MADDISON, *MacClade. Analysis of Phylogeny and Character Evolution*, Sinauer Associates, Sunderland MA (1992).
31. S. A. BENNER, I. BADCOE, M. A. COHEN and D. L. GERLOFF, *Bona fide* prediction of aspects of protein conformation. Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences, *J. Mol. Biol.* **235,** 926–958 (1994).
32. C. P. HILL, T. D. OSSLUND and D. EISENBERG, The structure of granulocyte colony stimulating factor and its relationship to other growth factors, *Proc. Nat. Acad. Sci.* **90,** 5176–5181 (1993).
33. A. M. DE VOS, M. ULTSCH and A. A. KOSSIAKOFF, Human growth-hormone and extracellular domain of its receptor. Crystal-structure of the complex, *Science* **255,** 306–312 (1992).
34. T. M. JERMANN, J. G. OPITZ, J. STACKHOUSE and S. A. BENNER, Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily, *Nature* **374,** 57–59 (1995).
35. J. DOSTAL and J. MATOUSEK, Isolation and some chemical properties of aspermatogenic substance from bull seminal vesicle fluid, *J. Reprod. Fertil.* **33,** 263–274 (1973).
36. J. SOUCEK and J. MATOUSEK, Inhibitory effect of bovine seminal ribonuclease on activated lymphocytes and lympoblastoid cell lines in vitro, *Folia Biol. Praha* **27,** 334–345 (1981).
37. J. SOUCEK, A. HRUBÁ, E. PALUSKA, V. CHUDOMEL, J. DOSTÁL and J. MATOUSEK, Immunosuppressive effects of bovine seminal fluid fractions with ribonuclease activity, *Folia biologica (Praha)* **29,** 250–261 (1983).
38. J. SOUCEK, V. CHUDOMEL, I. POTMESILOVA and J. T. NOVAK, Effect of ribonucleases on cell, mediated lympholysis reaction and on GM, CFC colonies in bone marrow culture, *Nat. Immun. Cell Growth Regul.* **5,** 250–258 (1986).
39. J. MATOUSEK, The effect of bovine seminal ribonuclease on cells of Crocker tumor in mice, *Experientia* **29,** 858–859 (1973).

40. S. VESCIA, D. TRAMONTANO, G. AUGUSTI-TOCCO and G. D'ALESSIO, In vitro studies on selective inhibition of tumor cell growth by seminal ribonuclease, *Cancer Res.* **40,** 3740–3744 (1980).
41. N. TRABESINGER-RÜF, T. M. JERMANN, T. R. ZANKEL, B. DURRANT, G. FRANK and S. A. BENNER, Pseudogenes in ribonuclease evolution. A source of new biomacromolecular function? *FEBS Lett.* **382,** 319–322 (1996).
42. G. H. GONNET, M. A. COHEN and S. A. BENNER, Exhaustive matching of the entire protein sequence database, *Science* **256,** 1443–1445 (1992).
43. J. HUELSENBECK and B. RANNALA, Phytogenetic methods come of age: Testing hypotheses in an evolutionary context, *Science* **276,** 227–232 (1997).