PATTERNS OF DIVERGENCE IN HOMOLOGOUS PROTEINS AS INDICATORS OF TERTIARY AND QUATERNARY STRUCTURE

STEVEN A. BENNER

Laboratory for Organic Chemistry,
Swiss Federal Institute of Technology, CH-8092 Zurich, Switzerland

INTRODUCTION

Conformational analysis in peptide chemistry presents two historically significant problems: (a) Can polypeptide sequences be routinely designed to fold in solution to yield a predicted tertiary structure? (b) Can the tertiary structure of a natural polypeptide sequence be predicted from sequence data? With the first problem rapidly approaching solution (1) it is appropriate to focus on the second.

We use the term "conformational analysis" to remind the reader that the "protein folding problem" is a special case of a topic common in organic chemistry: how constitution determines conformation. In view of this fact, it might be surprising that the field is dominated by crystallographers, molecular biologists, physical chemists, and computational chemists; few organic chemists can be found doing research on the "protein folding problem".

In part, the absence of organic chemists reflects their accurate appreciation of the magnitude of the problem. Virtually every organic chemist deals with organic reactivity where conformation plays a central role. However, organic molecules routinely refuse to behave as predicted, and a poor understanding of their conformation in solution is often the reason why.

There is no theory that allows the chemist to predict the conformation of any organic molecule in solution. This is true even for molecules much smaller than normal proteins. Computational methods for making these predictions are improving; optimistically (and perhaps naively) given enough computer time, such methods might ultimately be satisfactory for predicting the conformation in solution of all molecules, including proteins. However, at present, the best computational programs do respectable jobs only in solvents similar to a vacuum, and there only with small molecules. Conformational analysis in a solvent that interacts strongly with the solute (water being a good example) is far more difficult. For organic chemists it makes sense to say, "Let's solve the conformation problem with small molecules first; then we can worry about proteins."

This prelude is intended to convey one view of chemical reality to those interested in protein folding. The commonly stated goals: a "code" for protein folding (analogous to the genetic code); patterns in sequences that indicate with reliability a specific secondary structure; a "meta-language" that permits the analysis of protein folded forms without recourse to analysis at the level of atoms, bonds, and orbitals; or a distributable computer program that will allow the novice to extract information useful for predicting tertiary structure from sequence data, all may someday be attained. However, realistically, such goals have not been achieved in any branch of organic chemistry in the past.

Realism need not be discouraging. Individuals *can* acquire the ability to predict the behavior of molecules to a degree sufficient to manipulate them for practical ends. This ability in an organic chemist comes from training and experience, creating an "intuition" (a single word meaning "training and experience") about organic reactivity. Further, an organic chemist is trained to analyze from many theoretical perspectives a problem in reactivity concerning a single molecule. First, he considers steric aspects, then perhaps electronic aspects, then perhaps acid–base properties. With each pass, he deepens his understanding of the molecule.

Finally, there is one opportunity for the conformational analysis of biological macromolecules that is not available in normal organic molecules. Proteins are the products of evolutionary processes, and are often present in the natural world in a variety of forms. With recent advances in purification and sequencing, it is now far easier to collect constitutional information on macromolecules than it is to determine their conformation by crystallography. Our understanding of how proteins evolve is also growing rapidly. This understanding together with sequence data are already providing much information useful for tackling the protein folding problem.

We develop here themes that incorporate both considerations. These themes constitute a partial solution to the protein folding problem. The reader should not expect this discussion to be uncomplicated, or our approach to be simple; after all, nothing in organic chemistry is simple or uncomplicated. However, by examinations of many proteins, from many theoretical points of view, and with an eye towards evolutionary processes, progress can be made.

WHAT WE CAN LEARN FROM CHEMISTRY

Those who synthesize peptides routinely experience solubility problems. Many synthetic peptides are intractable because they precipitate. A peptide that precipitates is a peptide that prefers interactions with other peptide units over interactions with solvent. Of course, interactions with other peptide units (in preference to interactions with solvent) are simply folding interactions. The general insolubility of peptides suggests that there is abundant opportunity for peptides to fold with immense conformational stability. The fact that most proteins from thermostable organisms are quite conformationally stable has long been recognized as another manifestation of this same principle.

This is one reason why progress in the design of peptides that fold in aqueous solution is faster than in the prediction of conformation of natural peptides from their sequences. Indeed, the most important problem in design is to obtain a sufficiently soluble peptide for n.m.r. analysis to prove conformation. However, in predicting the conformation of natural polypeptides whose conformational stability is usually rather low, this fact creates a problem. It indicates that there will be few constitutional features (i.e., sequences) that are *essential* for conformational stability. Given an excess of opportunities for conformationally stabilizing interactions, some can be absent in some proteins, and others in other proteins, and still have proteins that fold.

This point is important, as much effort has recently been devoted to identifying structural features that are necessary for the formation of particular secondary structures. For example, a recent hypothesis is that side chains capable of forming hydrogen bonds to amide groups at the end of helices are a *necessary* condition for helix formation (2). While such efforts should be encouraged given the possibility that they might lead to discoveries, they are unlikely to be successful in the manner envisioned. Amphiphilicity, or hydrogen bonding patterns, or electrostatic interactions, or polarity in the environment, each appear *sufficient* to induce helix formation in model peptides in solution. A helix that has all of these structures is very stable indeed. Thus, it is unlikely that one of these factors will be found to be a necessary condition for helix formation in natural proteins.

WHAT WE CAN LEARN FROM BIOLOGY

In the past five years, our understanding of the role of natural selection in determining the properties of proteins has grown enormously. This subject is reviewed at length elsewhere (3–6). In most cases we are now able to distinguish between macromolecular behavior that is adaptive (i.e., influences the survival of a host organism) from behavior that is neutral. This distinction is critical, as adaptive and neutral traits in proteins behave differently during divergent evolution. Neutral traits drift randomly, and structural aspects of a protein that determine neutral behaviors alone are unconstrained from drifting. However, structures that influence adaptive traits (function) do not drift; the function is said to "constrain" drift. However, protein sequences can diverge for adaptive reasons. Different proteins in different environments need different structures to be optimally suited to assist efforts of the host organism to survive.

An understanding of evolution in proteins can assist the study of folding in several ways. First, tertiary structure diverges far slower than primary structure. This has been shown best in the elegant investigations of Chothia and Lesk, where 75% divergence in sequence leads to less than 2 Å rms divergence in backbone positions (7). This means that in predicting tertiary structures at this resolution, it is relevant to examine the sequences of many homologous proteins, even if their sequences have diverged substantially. Thornton, Blundell and their colleagues have made use of this fact to extrapolate the structures of proteins when the tertiary structure of one of its homologs is already known by crystallography (8).

However, an understanding of divergence in function can, on occasion, help build a picture of a protein with unknown tertiary structure. For example, if, in a series of homologous proteins, adaptation requires different catalytic behaviors in different proteins, structural differences are expected to reflect this. As is illustrated below, this permits the deduction of a piece of information concerning which parts of the primary structure come together to form the active site.

However, the most important outcome of studies on adaptation and drift in proteins is a full appreciation that conformational instability in proteins is an evolutionarily desirable behavior. Conformational instability in a protein is believed to be important for protein turnover, and instability appears to be engineered into proteins to give them a desirable lifetime *in vivo*. Much evidence supports this statement. In many cases, inducible proteins (those which are designed to be recycled more frequently under physiological conditions) often have lower thermal stability than constitutive proteins (9). Further, it is quite easy to create a mutant protein that is more stable by introducing a point mutation into a protein. This argues strongly that natural selection does not maximize stability (6).

The implications of this are disappointing to those attempting to predict tertiary structure from primary structure. Opportunities for conformational stabilization are abundant (*vide supra*), meaning that proteins do not *need* to use them all to form a stable tertiary structure. But also it is clear that natural selection does not *want* to form a very stable tertiary structure. Enzymes that have evolved to have low conformational stability will use only a small subset of the stabilizing interactions possible. Indeed, selective forces might produce proteins that introduce destabilizing interactions, simply to achieve the desired level of instability in the protein.

Thus, even if rules can be deduced that connect primary and tertiary structure, biological systems will have evolved to violate them a certain fraction of the time to achieve proteins with the desired level of instability. This implies that the sequence of a single protein will contain pitfalls, elements in the sequence that will deceive the chemist setting out to predict conformation, even one who thoroughly understands the rules that connect constitution and conformation.

WHAT WE CAN LEARN FROM ALIGNMENTS OF HOMOLOGOUS SEQUENCES

The simplest approach to solving this last problem is to examine alignments of many sequences of homologous proteins. As mentioned above, each protein in such a collection will have approximately the same tertiary structure. While no single structural feature is necessary for forming a particular secondary structure, and while every rule can be violated in a protein's search for instability, there does not appear to be selective pressure to determine *which particular* destabilizing interactions will be incorporated. Thus, destabilizing interactions are expected to drift during divergent evolution. An "average" over many sequences might therefore filter out the destabilizing "noise".

This approach has been suggested in its general form by many others seeking to improve the success of statistical methods for secondary structure prediction. For example, Chou-Fasman parameters successfully predict secondary structure only about 60–70% of the time. However, by "averaging" these parameters over a set of homologous structures, the hope has been to improve structural predictions. This hope has been realized in at least one case. When averaging of statistical parameters was combined with other considerations, Kirschner and his co-workers successfully predicted the tertiary structure of tryptophan synthetase (10).

However, much more information can be obtained from an alignment of sequences if we use a different type of analysis (11). Information about tertiary structure is contained in the *pattern* of sequence divergence in homologous proteins. The sequence divergence in a protein is a combination of adaptive variation and neutral drift. Because the impact on behavior of these two types of primary structural variation is intended to be different, their tertiary structural implications are different.

Neutral variation must (by definition) have no impact on any selectable behavior of a protein. Evidence discussed elsewhere (3–6) shows that most behaviors are selectable, even very subtle ones. Thus, neutral variation can only occur in limited regions of a protein. In normal catalytic proteins (although not in binding proteins), neutral variation is generally found on the surface of the protein. This statement must be qualified. As the sequences of two proteins diverge, neutral variation need no longer occur simply on the surface. Compensating changes are conceivable (and in fact are observed) where several changes together have no impact on behavior, but where the changes individually do. Thus, as the sequences of two proteins drift apart, neutral variation moves from the surface into the protein (Fig. 1).

In contrast, adaptive variation must perturb the behavior of a protein. In the case of an enzyme, adaptive variation alters the kinetic, physical, or other catalytic properties to suit a particular substrate or a particular environment. Thus, adaptive variation can occur essentially anywhere in a protein; often it is seen at or near the active site.

This suggests an approach for detecting surface residues in a set of homologous proteins with unknown tertiary structure, provided that adaptive variation can be distinguished from neutral variation. We illustrate this procedure with alcohol dehydrogenases (EC 1.1.1.1). These proteins form an especially challenging prediction problem for several reasons. First, the alcohol dehydrogenase structure is "irregular" (for example, in comparison with beta-barrels). Second, there has been a considerable amount of adaptive variation superimposed on neutral drift. In glycolytic enzymes (e.g., triose phosphate isomers) prediction is easier, as the exact substrate has been conserved, suggesting that a higher proportion of the structural variation that is observed is neutral.

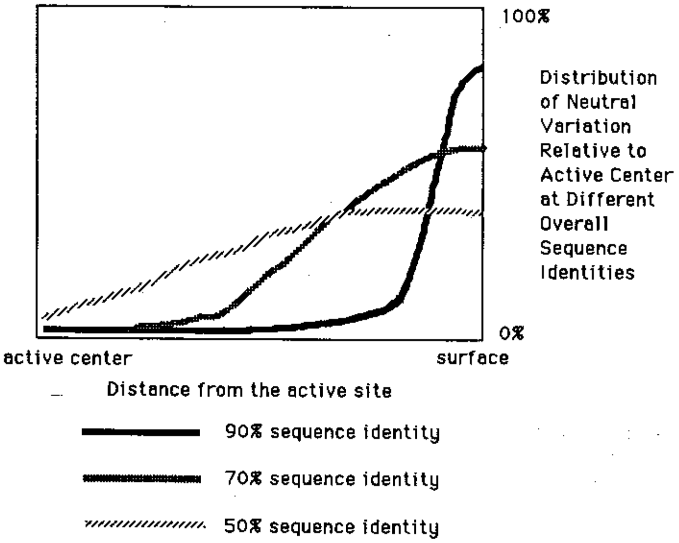

FIG. 1. A generalized picture of the positions of neutral variation in a protein, relative to the active center, in proteins with different overall sequence identity. As sequences diverge, neutral variation can move in towards the active center, due to the presence of increasing possibility for compensating mutations.

THE ALIGNMENT

To illustrate how information about tertiary structure can be extracted from the patterns of sequence divergence in a set of homologous proteins, we consider an alignment of 17 alcohol dehydrogenase sequences (the "master alignment"). Eight are from mammals, four from plants, and five from fungi. A table of the pairwise sequence identities is shown in Table 1.

Subgroups of this alignment are chosen for analysis. A group can be characterized by a "minimum pairwise identity" (MPI) value, which is simply the percentage identity shared by the two proteins in the group that are the least similar. For example, the MPI for the total alignment of 17 proteins is 21%. Alternatively, a group can be characterized by the percentage of residues that are "absolutely positively conserved" (APC) throughout the alignment. Clearly, the APC value is always less than or equal to the MPI value. The lower the values are for the group, the greater the sequence divergence within the group.

TABLE 1. PAIRWISE IDENTITIES IN ALIGNMENT OF ALCOHOL DEHYDROGENASES

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	98*	88	87	88	64	85	82	53	53	48	53	24	24	24	24	25
2		88	87	88	64	85	83	52	53	48	52	25	24	24	24	25
3			94	93	63	84	82	52	52	48	52	26	26	26	25	26
4				95	63	84	83	52	52	49	52	26	26	27	26	27
5					62	85	83	51	52	48	51	26	26	26	26	26
6						60	60	51	52	50	52	24	21	21	22	24
7							90	51	51	48	51	26	26	26	26	26
8								51	51	48	51	26	26	26	26	25
9									87	83	81	23	22	21	22	23
10										83	81	24	23	23	24	24
11											87	24	23	23	24	25
12												23	22	22	24	24
13													58	59	58	55
14														93	80	56
15															81	56
16																53

*Percent sequence identity.

Key:

1	Horse liver E.
2	Horse liver S.
3	Human class 1, alpha chain.
4	Human class 1, beta chain.
5	Human class 1, gamma chain.
6	Human class 2.
7	Mouse.
8	Rat.
9	Maize chain 1.
10	Maize chain 2.
11	Pea.
12	*Arabidopsis*.
13	*Aspergillus nidulans*.
14	*Saccharomyces cerivisiae* 1.
15	*Saccharomyces cerivisiae* 2.
16	*Saccharomyces cerivisiae* 3.
17	*Schizosaccharomyces pombe*.

THE COMPUTER PROGRAMS

The operations that are described below are best executed by computer. Programs were run on an IBM-AT, on a hand-entered alignment. We emphasize that the computer programs used here are only tools to assist the chemist in manipulating data; they do not predict tertiary structure directly.

DETECTION OF SURFACE RESIDUES

As mentioned above, neutral variation is expected to occur most often at positions where the side chain protrudes into solvent. Thus, surface residues are likely to display high variability; conversely, one might like to develop a procedure that identifies positions in an alignment that are observed to be highly variable, and then assign those positions to the surface of a protein with unknown structure.

However, variability can occur inside the protein and still be "neutral" in terms of its impact on selectable behavior in the protein, once proteins have diverged sufficiently so that compensating changes can be incorporated into the structure (*vide supra*). Thus, only in proteins that are otherwise highly similar in sequence will neutral variation be confined primarily to the surface.

Of course, variability even in highly homologous proteins cannot itself infallibly predict surface residues. Variability observed at least in some positions in an alignment is undoubtedly adaptive, and adaptive variation can occur anywhere in the folded structure. Therefore, any algorithm that assigns residues to the surface of a protein, based on their variability in proteins with high sequence homology, must first "filter" the variability to remove the part that is adaptive.

To solve the first problem, we look for variability within subgroups of the master alignment where the MPI of each subgroup is greater than 85%. Thus, any variability that we detect is variability between two proteins that are otherwise highly similar. This implies (the implication to be tested by examination of actual data) that the neutral variation that is detected will be on the surface of the protein. This number is not absolute. Lower values will identify more residues as being on the surface, but with a greater risk that the identification will be in error. Higher values will provide more secure identification of surface residues, but at the expense of overlooking a higher fraction of the residues actually on the surface.

To filter out adaptive variation, we assign significance to variability only if it appears at a particular position in more than one subgroup. The rationale for this "filter" is simple. Variability at position X in subgroup 1 may be adaptive; the two proteins in subgroup 1 might perform slightly different functions in slightly different environments, and variation at position X might be needed to create two proteins optimized for two different functions. However, it is unlikely that the same position will be altered adaptively in subgroup 2. Thus, by assuming that the variation at position X is neutral only if variation at this position is observed in another subgroup should filter out variation that is adaptive (at the expense of filtering out some variation that truly identifies surface residues).

Five subgroups from the master alignment were selected: Subgroup 1 (mammalian Adh No. 1, proteins, 1, 2, 3, 4 and 5) has an MPI of 87%. Subgroup 2 (mammalian Adh No. 2, proteins 7 and 8) has an MPI of 90%. Subgroup 3 (plant Adh No. 1, proteins 9 and 10) has an MPI of 87%. Subgroup 4 (plant Adh No. 2, proteins 11 and 12) has an MPI of 87%. Subgroup 5 (yeast Adh, proteins 14 and 15) has an MPI of 93%. Thus, the sequences of the proteins within each subgroup are quite similar.

The computer then searched for positions in the alignment where variability was observed in two or more of these subgroups. Fifty-three residues were identified, corresponding to 14% of the total sequence. These were divided into 8 classes, depending on the nature of the variability (Table 2).

TABLE 2. POSITIONS IN THE ALIGNMENT FOR ALCOHOL DEHYDROGENASE SHOWING VARIABILITY IN MORE THAN ONE SUBGROUP, WHERE EACH SUBGROUP HAS A MINIMUM PAIRWISE IDENTITY OF GREATER THAN 85%

Class A: Hydrophilic hypervariable residues					
118	MMNNN	MQ	TV	TT	GG
124	QQQQQ	RL	AG	NH	—
133	RRRRS	KR	NS	KN	—
138	YHHHH	HH	YF	HY	—
185	KKNNK	KQ	NN	NN	NS
277	TTAAA	SS	QS	SQ	AA
327	SSCGS	SA	DD	DD	AA
Class B: Hydrophobic hypervariable residues					
141	LLLLV	IL	VV	VL	TT
208	IIIVV	IV	AM	AA	VV
Class C: Hydrophilic variable residues					
33	AAAAA	AA	AA	AK	KP
56	TTTNN	TG	KK	KK	DD
84	RRKKK	KK	AA	KQ	KK
99	KKKKK	EK	EE	ED	AA
101	RSRRR	RR	AA	PR	EE
115	DSDDD	DN	RR	RR	DD
120	RRQRR	RK	RR	RR	—
156	SSAAA	AA	CC	CQ	QQ

TABLE 2.—*Cont'd*

191	QQPPP	PP	KK	PK	AA
227	DDDDD	DD	SA	SK	CG
233	KKKKK	KK	RK	KK	RT
247	KKKKK	ST	ND	DD	KK
300	NNNNN	NS	EQ	AA	CC
310	SSTTT	LL	NS	NN	KK
Class D: Amphiphilic variable					
17	ENLVL	LP	AA	AA	SS
34	HHHTH	HH	MM	GH	NH
117	SSSGG	LT	NN	NN	SS
190	TTTTT	TT	PA	KK	MR
231	KKKKK	KK	EQ	LQ	LL
297	DDDAD	MS	KK	KK	GG
303	MMMII	MM	VT	TT	DD
307	LLLLL	LS	NN	NN	QH
373	TTMTT	TT	RR	KT	DD
Class E: Hydrophobic variable					
38	IIIII	II	VI	LI	II
65	AALLL	LL	FL	FG	GG
76	IIVVV	VI	VV	VV	MM
110	FLYYY	FL	MM	MM	NN
123	MMLLL	LL	II	LI	—
172	IIIII	II	LL	LV	VI
224	IIIII	II	LI	LF	GG
235	VVLLL	LL	FF	FF	IL
319	FFLYF	FF	FF	YF	VV
328	VVVIV	VV	LL	LI	DD
Class F: Reflexive hydrophobic variable					
184	VVVVV	VV	IL	IL	KK
272	LLLLL	LL	IV	IV	EE
Class G: Reflexive hydrophilic variable					
10	KKKKK	KR	KR	R	KK
135	KKKKK	KK	KQ	QK	—
259	NNDDD	DD	NN	ND	NN
Class H: Hydrophobicity split					
213	AAAAA	AT	IL	II	AA
255	TTKKK	QQ	AI	AA	KK
265	SSSSS	SS	SS	AS	VI
343	DDDDD	DE	EE	EE	IV
363	RRHHR	RR	AL	LL	EE
367	SSSSS	SS	GS	SS	VA

The one letter code for amino acids, and numbering of the horse liver enzyme are used. For the purpose of classification in this table, G, P, and A are treated as either hydrophobic or hydrophilic, with the assignment chosen to agree with the polarity of the other amino acids in the subgroup. C, D, E, H, K, N, Q, R, S, and T are considered hydrophilic. F, I, L, M, V, W, and Y are considered hydrophobic. The first column corresponds to the mammalian Adh No. 1 subgroup, the second to mammalian Adh No. 2 subgroup, the third to plant Adh No. 1 subgroup, the fourth to plant Adh No. 2 subgroup, and the fifth to yeast Adh subgroup.

A. Hydrophilic hypervariable residues; those where more than 2 subgroups showed variability, and where variation involved polar amino acid side chains. Eight positions were identified.

B. Hydrophobic hypervariable residues; those where more than 2 subgroups showed variability, and where variation involved only non-polar amino acid side chains. Two positions were identified.

C. Hydrophilic variable residues; where polar substitution was observed in 2 subgroups. Fourteen positions were identified.

D. Amphiphilic variable; where variability is observed in 2 subgroups, where at least in one case polar for non-polar substitution is observed. Nine positions are identified.

E. Hydrophobic variable; where variability is observed in 2 subgroups, but where hydrophobic residues are substituted always for other hydrophobic residues. Ten positions are identified.

F. Reflexive hydrophobic variable; where variability is observed in 2 subgroups, but where the two amino acids in one subgroup showing variation are the same as the two in the other, and hydrophobic residues are substituted always for other hydrophobic residues. Two positions are identified.

G. Reflexive hydrophilic variable; where variability is observed in 2 subgroups, but where the two amino acids in one subgroup are the same as the two in the other, and hydrophilic residues are sometimes substituted for other hydrophobic residues. Three positions are identified.

H. Hydrophobicity "splits"; where variability in both classes conserves the polarity of the residue within each class. Six positions are identified.

RESULTS

Four classes of variation are expected to be strong indicators that the residues occupying the variable position lie on the surface of the protein. In order of decreasing reliability, these are class A (hydrophilic hypervariable residues), class C (hydrophilic variable residues), class D (amphiphilic variable residues), and class G (reflexive hydrophilic variables). In the first case, it is unlikely that variation in all three classes is adaptive; further, the facility with which polar groups are substituted suggests that such substitution could not be neutral if it is in the interior of the protein. Thus, this variation is expected to be on the surface. Similar arguments apply to classes C and D, although less strongly since variation is observed in only two groups. Finally, reflexive variation suggests somewhat greater constraints on variation.

Inspection of the crystal structure for Adh (using Frodo software on an Evans and Sutherland graphics display apparatus interacting with a Vax computer) permits two important conclusions. First, the residues identified by this procedure are indeed all on the surface. Second, the method is reliable for members of these classes, even in the cases where variability is observed in only two subgroups. The side chains of only two residues identified by the algorithm, Asp 115 and Ser 196 are partially buried. The rest are fully exposed.

In other classes, the polarity of the side chain displays considerable conservation, even though the amino acid undergoes variation. For example, in class H, the polarity of the side chain is conserved within the variable subgroups, even though the polarity is different between different subgroups. This suggests constraints on drift at these positions. Nevertheless, 4 of the 6 residues are on the surface; however, the side chains of the residues at two of these positions (Ser 265 and Ser 367) are buried.

In classes B, E, and F, hydrophobicity is conserved within the variable groups, suggesting still more stringent constraints on neutral drift. Residues at these positions are not likely to reside on the surface, but rather in interior positions where some variability in structure is possible. Examples of the latter type are (a) on the inside of secondary structural units which are on the surface, (b) at subunit contact sites, and (c) at domain contact sites. Indeed, the algorithm is a good (but not perfect) indicator of such residues: residues at positions 76, 141, 184, 235, and 272 lie internally but near the surface, and residues at positions 110, 172, 224, and 319 lie at interfaces between domains.

DISCUSSION

The algorithm presented here is a reliable predictor of surface residues, at least in alcohol dehydrogenases. At the very least, such an algorithm provides a fully independent way to test tertiary structure predictions for proteins with unknown structures. However, more information clearly can be extracted from the patterns of sequence divergence used in this algorithm. Some consideration should be given as to how this information might be extracted.

Two perspectives are possible. We might simply use evolutionary information to help us analyze the structure of alcohol dehydrogenases, together with crystallographic, kinetic, and other information, to gain insight into the reactivity of the molecule itself. Such an analysis allows the investigator to improve his success rate with site-directed mutagenesis. For example, positions identified in class D are those where some proteins direct a hydrophobic side chain into solvent. Interactions between solvent and the hydrophobic side chain are expected to destabilize the folded form. This implies that more stable proteins could be obtained by site-directed mutagenesis that replaces the hydrophobic residue at this position with a hydrophilic residue. Thus, a yeast Adh with Met 190 replaced by an Arg should be a more stable protein.

Similarly, class C identifies some proteins where Pro is substituted by another amino acid. This implies that, at this point in the tertiary structure, the side chain can accommodate the constraints imposed by a Pro. Locking the chain in that conformation should stabilize the protein. Thus, a yeast Adh with Lys 33 replaced by a Pro should also be more stable.

However, from a second perspective, we might try to develop algorithms similar to the one presented here to make predictions about details of the tertiary structure for proteins that lack crystal structures. Detailed considerations of the divergence in function and behavior can help. Mammalian alcohol dehydrogenases display a wide range of substrate specificity, and quite low specific activity with ethanol as a substrate. This variation is probably adaptive. However, regardless of whether the variation is adaptive or neutral, it is clear that structural variation near the active site is needed to account for it. In contrast, there is relatively little catalytic variability in the homologous fungal Adh's. All appear to act on ethanol and acetaldehyde as sole substrates; all are quite active catalysts with this substrate.

Together, these facts suggest that positions displaying variability in mammalian Adh's, but which are highly conserved in fungal Adh's, are likely to identify residues at or near the active site.

There is, unfortunately, a complication. Mammalian enzymes are dimers, while yeast alcohol dehydrogenases, and perhaps all fungal alcohol dehydrogenases, are tetramers. Thus, residues involved in the tetrameric contact in the fungal enzyme will also be conserved more highly than the corresponding residues in the mammalian dimer (where these residues protrude into solvent). This complication is not necessarily bad. Much of the work with site-directed mutagenesis in this system is being done on the yeast enzyme. Although some guidance for this work comes from the crystal structure of the dimeric enzyme from horse liver, the sequences are 70% different. Any insight that the patterns of evolutionary divergence can give us about the different quaternary structures in the yeast and horse enzymes would help us guide these studies.

Nevertheless, the complication remains. Here again, we have a choice of subgroups to compare. To make this comparison reasonable, the subgroup of Adh's that is searched for conservation (in this case, fungal Adh's) should have a lower MPI than the group of Adh's that is searched for variability (mammalian, in this case). If the reverse is true, many residues will be identified that are identical in the first subgroup Adh's simply because residues in this group are generally more highly conserved. Subgroups can be chosen with differing levels of stringency. We have chosen 5 mammalian Adh's (proteins 1–5, MPI=87%) and 5 fungal Adh's (proteins 13–17, MPI=53%).

This algorithm using these subgroups identifies 24 amino acids that are variable in the mammalian enzymes but conserved in the fungal enzymes (Table 3). Their distribution is striking. Four clusters are evident (residues 47–57, 45%; 108–118, 36%; 318–328, 27%; 341–348, 37%). Of course, some of the variation in the mammalian subgroup could be the result of neutral drift, undesirable (in this case), since we are seeking adaptive variation in the mammalian series. Thus, the data are best filtered by removing from the list those residues already identified as highly variable (above) and therefore presumably able to undergo facile neutral drift. Alternatively, the data can be filtered by excluding positions that display variability in plant Adh's (proteins 9–12, MPI=81%).

TABLE 3. RESIDUES CONSERVED IN FUNGAL ALCOHOL DEHYDROGENASES (MPI=53%) BUT VARIABLE IN MAMMALIAN ALCOHOL DEHYDROGENASES (MPI=87%)

Position	Fungal Adh	Mammalian Adh		
47	H	RRGRR		Confirmed plant
48	T	SSTTS		Confirmed plant
50	L	DDDDE		Confirmed plant
56	D	TTTNN	Highly variable	Confirmed plant
57	L	LLMLL		Confirmed plant
65	G	AALLL	Highly variable	
84	K	RRKKK	Highly variable	
93	W	FFAFF		Confirmed plant
101	E	RSRRR	Highly variable	
108	E	GGSSS		Confirmed plant
116	L	LLVLL		Confirmed plant
117	S	SSSGG	Highly variable	Confirmed plant
118	G	MMNNN	Highly variable	
141	T	LLLLV	Highly variable	
143	D	TTITV	Variable	Confirmed plant
207	A	VVAAV		Confirmed plant
258	T	SSTTT		Confirmed plant
283	R	QQHHH		Confirmed plant
318	I	IIIVI		Confirmed plant
326	R	DDEEE		Confirmed plant
328	D	VVVIV	Highly variable	
341	G	AASSS		Confirmed plant
344	K	PPAAA		Confirmed plant
348	K	HHHHN		Confirmed plant

The one letter code for amino acids, and numbering of the horse liver enzyme are used. Residues in the region of mammalian Adh deleted in yeast (residues 120–139) are ignored. Positions designated "Highly variable" are identified by the algorithm discussed in the text in Table 2. Positions identified "Confirmed plant" indicate positions where residue is absolutely conserved in the 4 plant Adh's.

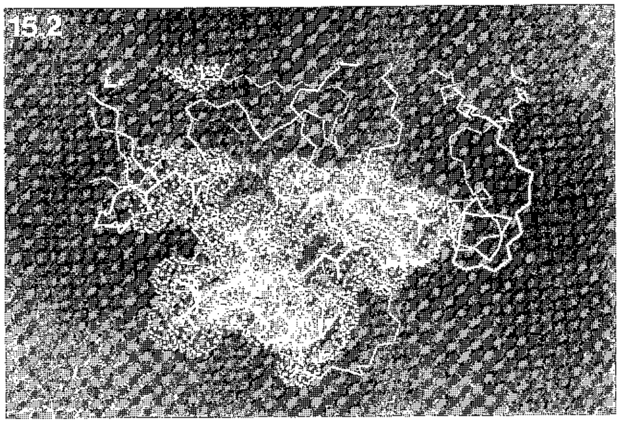

FIG. 2. View of the wall of the active site cleft at the position where the substrate binds. Residues that are variable in mammalian Adh's, but conserved in fungal Adh's, are shown as dotted spheres. These variations in structure correspond to variation in the substrate specificities of different mammalian Adh's. Fungal Adh's show much less variation in substrate specificity.

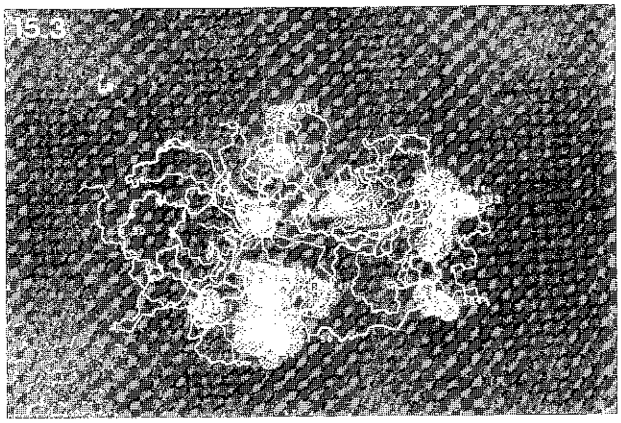

FIG. 3. A view of alcohol dehydrogenase, with residues that are variable in mammalian Adh's, but conserved in fungal Adh's, shown as dotted spheres. The indicated residues at the right and on the top represent the positions of the presumed quaternary contacts in fungal Adh. Note in particular the absence of such residues on the left side of the structure (near amino acid 30, marked). This suggests that the subunit contact sites are *not* in this region of the protein.

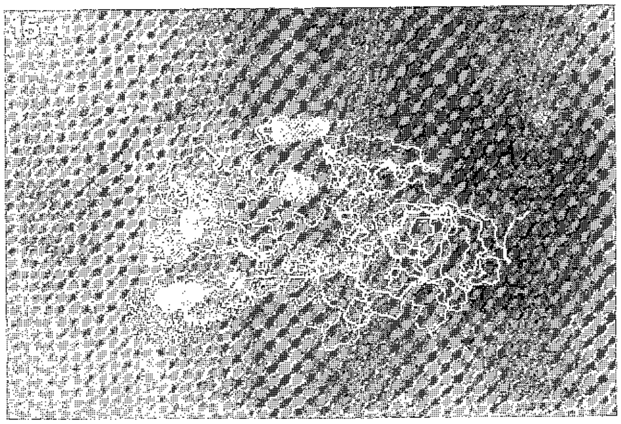

FIG. 4. A view of alcohol dehydrogenase as in Figure 3, but with the known subunit contact sites in dimeric mammalian Adh shown as large dotted circles.

TABLE 4. POSITIONS VARIABLE IN MAMMALIAN ALCOHOL DEHYDROGENASE BUT CONSERVED IN FUNGAL ALCOHOL DEHYDROGENASES, AFTER REMOVAL OF HIGHLY VARIABLE POSITIONS AND POSITIONS DISPLAYING VARIABILITY IN PLANTS

Position	Fungal Adh	Mammalian Adh	
47	H	RRGRR	Active site
48	T	SSTTS	Active site
50	L	DDDDE	Active site
57	L	LLMLL	Active site
93	W	FFAFF	Active site
108	E	GGSS	Active site
116	L	LLVLL	Near active site
207	A	VVAAV	Near active site
258	T	SSTTT	
283	R	QQHHH	
318	I	IIIVI	Active site
326	R	DDEEE	Active site
341	G	AASSS	
344	K	PPAAA	
348	K	HHHHN	

The one letter code for amino acids, and numbering of the horse liver enzyme are used.

Sixteen residues remain (Table 4). If we did not know the tertiary structure of any alcohol dehydrogenase, we might build a model that brings the polypeptide chain together at these residues. The result would not be very much in error. However, two significant errors would be made, due to the fact that some of these residues reflect different quaternary structures in mammalian and fungal Adh's.

In fact, 10 of the residues identified by the algorithm are at or near the active site. Indeed, both walls of the substrate binding cavity are identified by the algorithm; Figure 2 shows residues on one of the walls.

The remaining 5 positions are noteworthy in their location (Fig. 3). Two (positions 258 and 283) lie on the surface of the protein above where mammalian Adh makes a dimer contact. The remaining 3 lie together on the surface at the top of the protein. The obvious explanation why these residues are conserved in the fungal tetramers but not in the mammalian dimers is that these residues are involved in the (as yet structurally undefined) tetrameric contacts of the yeast enzyme.

Thus, the algorithm permits us to make a prediction about the position of the subunit contact sites in yeast alcohol dehydrogenase. This prediction awaits test by crystallography or mutagenesis.

However, a final word must be said before leaving the discussion of quaternary structure. Thirteen residues are identified by crystallography as being involved in the dimer contact in horse liver alcohol dehydrogenase

TABLE 5. POSITIONS INVOLVED IN SUBUNIT CONTACTS IN DIMERS ARE HIGHLY CONSERVED IN MAMMALIAN AND PLANT ALCOHOL DEHYDROGENASES, BUT VARIABLE IN FUNGAL ALCOHOL DEHYDROGENASES

Position	Mammalian residues	Plant residues	Fungal residues
275	M	M	FIIIY
291	IIIIIFVI	L	ALLLT
294	V	V	LHLLH
296	PPPPPAPP	HHSS	A
301	L	F	KCSFG
303	MMMIIIMM	VTTT	PDDED
305	P	P	F
306	MMMMMEMM	M	TNNSW
308	L	F	VVVVT
309	LLLLLILL	L	V
312	R	RKRR	deletion
314	WWWWWIWW	L	deletion
316	g	g	

The one letter code for amino acids, and numbering of the horse liver enzyme are used. A single letter in a subgroup indicates that the indicated amino acid is found at the indicated position in all proteins in the subgroup.

(Fig. 4) (Table 5). Seven of these are absolutely conserved in mammalian enzymes (MPI=62%). Ten are absolutely conserved in plant Adh's (MPI=81%). However, in fungal Adh (all presumably tetramers), only one of the positions that shows absolute conservation in mammalian Adh is conserved. Indeed, in fungal enzymes (MPI=53%), 5 of the positions undergo polar or amphiphilic variability; an additional position has been deleted. This is evolutionary behavior characteristic of residues occupying positions on the surface!

Thus, rapid divergence is seen in fungal Adh's at positions that in mammalian Adh are involved in quaternary interactions, and are highly conserved because of this. Conversely, fungal Adh's display two sets of residues on the surface that are much more highly conserved than residues at corresponding positions in mammalian enzymes. This suggests that the quaternary interactions in the two groups of proteins are quite different; the fungal tetramers cannot be simply considered to be dimers of the dimers found in mammalian enzymes. This statement is quite significant for those trying to engineer the behavior of yeast Adh using the crystal structure of mammalian Adh as a guide.

CONCLUSIONS

We have discussed here a fundamentally new approach for extracting conformational information from an alignment of homologous proteins.

A reliable algorithm is developed for identifying surface residues in a protein. An algorithm is also developed for identifying active site residues; this algorithm can be applied in cases where functional divergence occurs in one subgroup of homologous proteins but not in others. Finally, we have used these algorithms to make a prediction regarding the quaternary structure of alcohol dehydrogenase from yeast.

Clearly, much additional information remains to be extracted from these alignments. Some of this information has been summarized previously (11). For those who are disappointed with the complexity of the analysis presented here, we can offer only two consolations. First, chemical and evolutionary considerations, mentioned at the beginning of this article, make it unlikely that any simpler approach will be productive. A set of homologous proteins, together with considerations of evolution, function, and structure, seem to be necessary to provide sufficient information to make predictions.

Second, in organic chemistry, nothing is ever simple.

SUMMARY

A new approach for extracting conformational information from an alignment of homologous proteins is presented. This approach extracts information from the pattern of sequence divergence in proteins, and considers evolutionary issues, such as functional adaptation and neutral drift, in assigning roles in tertiary structure to residues at specific positions in the alignment. A reliable algorithm is developed for identifying surface residues in a protein. An algorithm is also developed for identifying active site residues; this algorithm can be applied in cases where functional divergence occurs in one subgroup of homologous proteins but not in others. Finally, these algorithms are used to make predictions regarding the quaternary structure of alcohol dehydrogenase from yeast.

ACKNOWLEDGEMENTS

I am indebted to the Swiss Federal Institute of Technology, the Swiss National Science Foundation, and Sandoz for generous support for my laboratory during the period that this work was executed.

REFERENCES

1. K. JOHNNSON, Synthese von polypeptiden mitdefinierter sekundaerstruktur und deren strukturuntersuchung in loesung, *Diplomarbeit*, E.T.H. Zurich (1988), V. L. RATH and R. J. FLETTERICK, Protein structure and design 1987, *Cell* **49**, 583–586 (1987).
2. L. G. PRESTA and G. D. ROSE, Helix signals in proteins, *Science* **240**, 1632–1641 (1988).
3. S. A. BENNER and A. D. ELLINGTON, Interpreting the behavior of enzymes: Purpose or pedigree? *CRC Crit. Rev. Biochem.*, in press (1988).

4. S. A. BENNER, Stereospecificity in enzymatic reactions: Its place in evolution, *Topics in Stereochem.*, in press (1988).
5. S. A. BENNER, Enzyme kinetics and biochemical adaptation, *Chem. Rev.*, in press (1988).
6. S. A. BENNER, Reconstructing the evolution of proteins, pp. 115–175 in *Redesigning the Molecules of Life*, (S. A. BENNER, ed.), Springer-Verlag, Heidelberg, (1988).
7. C. CHOTHIA and A. M. LESK, The relationship between the divergence of sequence and structure in proteins, *EMBO J.* **5**, 823–826 (1986).
8. T. L. BLUNDELL, B. L. SIBANDA, M. J. E. STERNBERG and J. M. THORNTON, Knowledge-based prediction of protein structures and the design of novel molecules *Nature* **326**, 347–352 (1987).
9. S. CHESNE, N. MONNIER and J. PELMONT, The two aspartate aminotransferases of *Escherichia coli* K12, *Biochimie* **60**, 403–407 (1978).
10. I. P. CRAWFORD, T. NIERMANN and K. KIRSCHNER, Prediction of secondary structure by evolutionary comparison: Application to the alpha subunit of tryptophan synthase, *Proteins* **2**, 118–129 (1988).
11. S. A. BENNER, Evolution, physical organic chemistry, and understanding enzymes, pp. 14–23 in *Enzymes—Tools and Targets*, Proceedings of the 6th International Conference on Clinical Enzymology, (D. M. GOLDBERG, D. W. MOSS and F. W. SCHMIDT, eds.), Karger, Basel (1988).