Pergamon

0065-2571/94/$26.00

0065-2571(93)E0011-C

PREDICTING THE CONFORMATION OF PROTEINS FROM SEQUENCES. PROGRESS AND FUTURE PROGRESS

STEVEN A. BENNER*, THOMAS F. JENNY†*, MARK A. COHEN* and GASTON H. GONNET†

*Institute for Organic Chemistry, E.T.H., CH-8092 Zürich Switzerland
†Institute for Scientific Computation, E.T.H., CH-8092 Zürich Switzerland

INTRODUCTION

Two complementary challenges define the frontier of structural biology in proteins: design and prediction. The design challenge will be met when biological chemists routinely design polypeptides that fold and catalyze reactions (1). The prediction challenge will be met when biological chemists regularly predict the conformation of polypeptide sequences that evolutionary processes have created to fold and catalyze reactions (2).

Improvements in methods for synthesizing and purifying polypeptides have enabled steady progress towards meeting the first challenge (3–7). In one case, the solution structure of a designed enzyme has been proven by multidimensional N.M.R. and its catalytic mechanism explored by physical organic methods (6, 8). The simpler goal, designing a polypeptide that folds, has been approached in still more laboratories (9–13), and the conformation of an additional designed protein has now been established by N.M.R. (14).

In contrast, progress towards solution of the protein structure prediction problem has been slow. A typical view was presented last year by the editors of *Trends in Biochemical Sciences* in a celebratory commentary for its 200th edition (15):

> When we asked ourselves and our friends what truly great advances were made since 1984, we got a strikingly large amount of foot shuffling and head scratching. Some of the old problems remain stubbornly unyielding, in particular the protein folding problem. The ability to predict folding patterns from sequences is still more a matter for soothsayers than scientists, despite lavish support from optimistic protein and drug designers.

As readers of *Advances in Enzyme Regulation* have known for some time (16), this view is mistaken. For many years in several laboratories, the conformations of proteins have been predicted, these predictions have

been published before crystallographic data became available (17–24), and subsequently determined crystallographic and N.M.R. structures (25–32) have shown these predictions to have been remarkably accurate. These predictions include both secondary structures and, in the best cases, supersecondary and tertiary structural models, and join others made with the help of spectroscopic information that provided conformational clues. Three key examples of the latter are the bacterial aspartate receptor (33, 34), interleukin II (35, 36), and interleukin IV (37). In some cases, predictions may have offered a better view of conformation in a protein than the reported crystal structure (38–40).

As part of the emergence of a new paradigm in structure prediction, we at the E.T.H. in Zurich have been developing a method for extracting conformational information from a set of aligned homologous protein sequences. In *Advances in Enzyme Regulation* three years ago, a model for the conformation of the protein kinase superfamily, an enzyme central to the regulation of metabolic processes in higher organisms, was built using the E.T.H. method (19). The prediction was also transmitted to the crystallographers involved in solving the structure of a member of this superfamily, for the cAMP-dependent protein kinase from mouse. The crystal structure was published one year later (26).

Thus, it was possible to review the prediction in light of subsequently determined structural data. This was done first by the crystallographers themselves, who found the prediction to have been "remarkably accurate" (26). Later, Thornton pointed out that although the residue-by-residue score achieved in the prediction was not much better than that obtained using classical methods, "the possibility of extending [the] prediction to a tertiary fold is much better with the Benner–Gerloff method" (41). Lesk and Boswell referred to the protein kinase prediction as a "spectacular achievement" that "will come to be recognized as a major breakthrough". It is, however, only a single test case (42).

Today, the protein kinase prediction is only one of several bona fide predictions (those made and published before crystallographic analysis is available) that have been made in Zurich. Several of these can now be evaluated in light of subsequently determined crystal structures. Very recently, the first example has been published where classical structure predictions (e.g. those based on Chou-Fasman (43) or GOR (44) methodology), a neoclassical prediction derived from the neural network developed by Sander and his group in Heidelberg (45), and predictions based on the new paradigm have been tested in parallel in a bona fide structure prediction contest (24).

The E.T.H. method is based on a rather detailed model of divergent evolution of function (46) and structure (47) in proteins. A detailed discussion of the method has appeared elsewhere (16, 19, 66). However, no comprehensive overview has been presented illustrating the importance of the evolutionary model to the method, or the advantages of prediction methods based solidly upon evolutionary foundations over those based on more classical approaches. Thus, it is timely to present this aspect of the problem here.

This article presents a discussion of the evolutionary model underlying the E.T.H. prediction method to illustrate how the model suggests tools for extracting conformational information in a protein family from a set of aligned homologous protein sequences. We then review briefly the early structure predictions made using this method. A review follows of the controversy that has emerged as several from the classical field of protein structure prediction have criticized the new method. We review the philosophical disagreements as well as more recent predictions that show these criticisms to be misdirected.

BACKGROUND

The Zurich laboratory began as outsiders in the field of structure prediction in 1986. At that time, much creative work within the field had produced enormous strides towards an understanding of protein conformation. Nevertheless, the field had also developed a research paradigm that, in several of its details, appeared to be obstructing further progress, at least from the perspective of the outsider (48, 49). In particular:

(1) The field had come to focus on one particular goal: to obtain a residue-by-residue assignment of secondary structure from sequence data. Tertiary structure was considered to be too difficult a challenge at the present.

(2) Prediction tools were being sought in the form of a fully automated computer program. Automation was presumed to guarantee objectivity, testability and reproducibility, while human involvement in the prediction process was believed to prevent these.

(3) The "fair and proper" method for evaluating an individual prediction was believed to be a "three state per-residue score" (50). To calculate this score, crystallographic data are first used to assign each residue in the protein to one of three conformational states (an alpha helix, a beta strand, or neither). The prediction method is then called upon to assign each residue in the protein to one of these three states. A per-residue score is then calculated by dividing the number of residues correctly assigned by the total number of residues. The score is then reported for the entire protein as a unit. Different types of errors (for example, misassigning an alpha helix as a coil versus misassigning an alpha helix as a beta strand) are weighted equally in this score. Evidently, many in the field had come to view such scores as informative (44, 48–50).

(4) Prediction of protein conformation was viewed as a statistical problem (48). Thus, prediction methods were being evaluated based on the average three-state per residue scores obtained from a statistically large sample of proteins. In contrast, evaluations of a prediction method based on examination of individual proteins or protein families was viewed as fundamentally flawed (48).

(5) Finally, the accepted procedure for testing prediction heuristics was evidently to apply them blind to a set of proteins whose structures were already known. Few were making bona fide predictions, that is, predictions announced before crystallographic (or, with increasing frequency, N.M.R.) structures became available (but see Refs 17–24, 33–37, 51).

Several of these rules of procedure were adopted following the experience in the previous decade. In the early 1970s, several methods for predicting the structure of proteins had been proposed (2). These methods had, however, been tested on only a small number of protein structures, as only a small number of structures were available at the time. Yet the methods may have been viewed with excessive enthusiasm, as demonstrated in the mid 1970s by a prediction contest (52). In this contest, different methods were challenged to predict the conformation of a structure that was known, but not released until after the predictions were announced. The methods had all done rather poorly (50). The conclusion had evidently been drawn that a more rigorous research paradigm was necessary to ensure objectivity, and that the five rules listed above were useful to this end.

As organic chemists, we appreciated the desire for objectivity. However, chemistry as a field had already had by 1986 a rich tradition in conformational analysis. From this tradition could be gleaned a general view of which sorts of approaches were likely to be productive, and which were not, in efforts to develop an understanding of conformation within a class of molecules. In particular, as discussed in *Advances in Enzyme Regulation* in 1989 (16), chemical experience suggested that the rules of procedure listed above would obstruct progress towards a solution to the structure prediction problem:

(1) Distinctions between local conformation (e.g., secondary structure) and global conformation (e.g., tertiary structure) had proven to be arbitrary in chemistry, even in small molecules. Information emerging from studies of protein structure pointed towards the same conclusion. For example, a pentapeptide of defined sequence was known by 1986 to be unreliable as an indicator of a unique secondary structure (53), suggesting that tertiary structural interactions were more important than local sequence in determining secondary structure. For protein prediction, this implied that aspects of tertiary structure must be predicted before secondary structure could be predicted.

(2) It is extremely difficult to obtain an objective description of conformation, even in small molecules, and doubly so when that description must be computer-based. For protein structures, this implies that three-state scores were likely to be uninformative about the value of a secondary structure prediction. Information emerging from studies of protein structure reinforced this view. The three most commonly used automated tools for assigning secondary structure to crystallographic data all concur for only 65% of the residues even when presented the same experimental data (54). The inaccuracy in assigning per residue secondary structure makes it impossible to attain three state secondary structure scores much higher than 75%, even if the secondary structure prediction is in fact "perfect". Further, it is clear that different types of assignment error have different impact on the value of a secondary structure prediction as a starting point for building a tertiary structural model. Assigning (for example) five consecutive residues to a helical conformation when the residues in fact form a core beta strand is likely to irrevocably frustrate any effort to build a tertiary structural model from a set of predicted secondary structural elements. Assigning five residues to a helical conformation in a region that a crystallographer records as a coil is not. Yet in a three-state score, the two misassignments are counted with equal weight.

(3) It has rarely been productive in chemistry to treat molecular behavior as a statistical average over many different molecules. In developing conformational theory in chemistry as it applies to small molecules, the conformation of individual molecules was first analyzed individually to develop an understanding of underlying principles in a single case. This understanding was then tested by applying it to another individual case, and then to another. Gradually, expertise was built for a class of molecules. For proteins, this implies that averaging three-state scores for individual proteins, themselves uninformative as averages over the entire length of a protein, to a statistically large sample of proteins will lose essentially all information that might be useful in an attempt to understand why the prediction method was successful (when successful) or unsuccessful (when it was not).

(4) Automation disconnects the chemist from the chemical details that are essential to developing this understanding. Once a conformational problem is solved, automation is certainly appropriate and often very desirable to save effort. However, an approach that involves interaction by a chemist is always superior to a fully automated approach before a problem is solved, and there was little doubt that the structure prediction problem in proteins had not been solved in 1986.

(5) Nor does automation guarantee objectivity. Experience in chemistry has provided many cases that illustrate that knowledge of fact can bias the output of highly parameterized computer programs, even those that are fully automated and tested "blind", even when the best efforts

are made to exclude such bias (55). Knowledge of reality can influence the output by indicating to the scientist when to stop parameterizing the program. The selection of which of several differing predictions made by a highly parameterized program to publish may be influenced by knowledge of reality; few theoretical results that differ widely from experimental reality make it into print. This means that while "blind" tests are useful in developing prediction tools, the tools must ultimately be tested and established by making and publishing bona fide predictions.

In Zurich, work in the structure prediction field therefore began with a rejection of the five aspects of the classical paradigm listed above. Indeed, the new paradigm in structure prediction, as outlined in our first effort in *Advances in Enzyme Regulation* (16), is largely defined by the rejection of these aspects of the classical paradigm. We were aware at the time that this approach was likely to attract criticism from adherents to the classical paradigm, who would regard the new approach as non-objective. We have not been disappointed (48, 49). Indeed, extraordinary effort must be made to ensure objectivity in the new paradigm as well, much of which is discussed below. However, the advantage of rejecting the elements of the classical paradigm listed above, in particular, reliance on three state scores calculated for a statistically large set of proteins, is that significant progress can be made.

Another aspect of the new paradigm was a recognition that natural enzymes have evolved to violate folding "rules" to engineer within themselves a degree of conformational instability (16, 46). This implied that natural polypeptide sequences have evolved to deceive the structure predictor, that prediction of conformation of a single sequence was unlikely to be direct even were these folding rules firmly in hand, and heuristics must be designed to accommodate the fact that selective pressures yielded "deceptive" polypeptide sequences.

Complementing this bad news was some good news. Nature often presents a set of sequences from homologous proteins diverging from a common ancestor under functional constraints; there are few parallels for this within small molecule chemistry. Thus, proteins come in families with similar functions and similar conformations. In 1986, Chothia and Lesk quantitated the relationship between sequence similarity and conformational similarity (56). Thus, if conformational information contained in homologous sequences could be extracted, analyzing the conformation of proteins could prove to be easier than analyzing the conformation of smaller organic molecules.

The potential application of homologous sequences in protein structure prediction was recognized at an early stage even among classical workers. Lenstra (57) *et al.* explored the possibility of obtaining improved secondary structure predictions by averaging predictions made by a classical method (e.g., Chou-Fasman or GOR) obtained from a family of homologous

sequences. Maxfield and Scheraga (58) and Garnier *et al.* (44) both suggested that such averaging might be productive in improving the scores obtained by their methods. Nishikawa and Ooi (51) first systematically applied averaging of classical predictions over homologous sequences to bona fide prediction, also constructing joint predictions by combining predictions made by several classical methods. Sternberg, Barton, and others have explored this approach in greater depth (59, 60). Kirschner and his coworkers (17) used such averaging in their bona fide prediction of the conformation of tryptophan synthase.

Nevertheless, averaging classical predictions over a set of aligned homologous protein sequences has proven disappointing as a way for improving scores, as already recognized by Lenstra *et al.* in their early work (57). In most families of proteins, improvements are marginal (51, 57–60), indicating that the predictions are not randomly disposed around reality such that the error behaves as "noise." The approach has therefore been shown to be unsatisfactory as a general solution to the prediction problem, at least in its simplest form. Despite this disappointment, some authoritative classical workers continue to emphasize this as *the* way in which evolutionary information might be exploited to improve structure predictions (49).

THE E.T.H. METHOD— EVOLUTIONARY MODELS

A model for the divergence of protein sequences evolving under functional constraints is necessary to begin to exploit the rich structural information that can be found in a set of aligned homologous protein sequences. These models can be developed both theoretically and empirically. We review three specific cases where a broadly based empirical investigation of divergent evolution can guide the formulation of prediction heuristics. We use a segment of the protein kinase alignment shown in Figure 1 in this discussion. Further we draw on the data provided by the 1.7 million aligned pairs of homologous protein sequences derived from the exhaustive matching of the protein sequence database (47). These data provide an enormous amount of information that allows us to systematize the prediction heuristics that were, for the most part, developed intuitively (16, 19). In this discussion, each aligned pair of sequences is characterized by an evolutionary distance measured in PAM units, the number of *P*oint *A*ccepted *M*utations per 100 residues (61).

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

FIG. 1. Segment of the protein kinase alignment.

The Probability of a Gap in an Alignment is Proportional to its (Length)$^{-1.7}$

It has long been appreciated that the probability of a gap in an alignment decreases with its length. The exhaustive matching (47) permitted two quantitative statements to be made about the gap-length distribution (62). First, it was largely independent of PAM distance, implying that gaps arose from single insertions and deletions occurring at a single point in time, not through multiple indels at the same position occurring consecutively in time. Second, the probability of an insertion or deletion was inversely proportional to its length raised to the 1.7 power. This empirical law applies with remarkable accuracy for indels ranging in length from a single amino acid up to those of over 60 amino acids.

Three hypotheses were suggested to account for these empirical facts (62):

(a) Inserted or deleted segments must be introduced or removed from points in the folded structure that are near in space, so that the insertion or deletion event does not create major alterations in the global conformation of the protein.

(b) Segments of polypeptide that are inserted or deleted from a typical protein form random coils in the folded structure.

(c) The same laws governing conformation of free coils govern coils in a polypeptide chain (63).

From the statistical mechanics of polymer chains (63), it has long been known that the mean radius of an unidimensional polymer chain in a random coil conformation is proportional to the length of the chain raised to the one-half power. Thus, the volume occupied by the coil is proportional to the length of the chain raised to the three-halves power. As the probability that the two ends of a polymer lie together in space is inversely proportional to that volume, one expects (if the three hypotheses are correct) that the probability of an indel should be inversely proportional to the three-halves power of its length.

This calculation applies, of course, only to an idealized unidimensional polymer. Real polypeptide chains are not unidimensional, and have an excluded volume that raises the power of the expression relating the length of the polymer to the volume that it occupies. Brant and Flory experimentally determined this with model polypeptides (64). With the excluded volume taken into consideration, the expression relating the volume occupied by a polypeptide chain in a random coil conformation to its length has a higher exponent, corresponding to the 1.7 power observed in the empirical gap-length dependence.

Several of the hypotheses are controversial, judging by the informal literature (mostly the electronic mail and journal and grant referee comments). Nevertheless, the gap length distribution can be used as a tool in structure prediction. If gaps are assumed to arise from insertions

or deletions that insert or extract randomly coiled chains, segments of a protein sequence that are matched against a deletion can be assigned the conformation of a random coil. While this conformational assignment itself is not of great importance, the assignment implies that the sequence preceding the gap and the sequence following the gap do not form one continuous secondary structure element.

We were not, of course, the first to suggest this approach for dividing (or parsing) a protein sequence. Indeed, Kirschner and his coworkers used gaps to parse the multiple alignment for tryptophan synthases, and extract a pattern of secondary structure that allowed them to identify the 8-fold alpha–beta barrel conformation adopted by this protein (17). However, the evolutionary model places this procedure on a more solid footing. Further, it provides a context for interpreting cases where the generalization does not apply.

We can use this rule to assign secondary structural information to the segment of the protein kinase alignment that we are considering. The segment matched against the one residue insert is assigned a "coil", and the secondary structure broken at this point (Fig. 2).

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

FIG. 2. Assignment of a break in the secondary structure from gaps in a multiple alignment.

Mutations Do Not Occur Independently at Consecutive Positions in a Protein

When building an alignment of homologous protein sequences, it is common to score matches and mismatches using an empirically derived "log-odds" scoring matrix developed by Dayhoff and her coworkers in the 1970s (61). This matrix makes several assumptions, one of which is that mutation at position i occurs independently of mutation at position $(i + 1)$. Even the most naive view of protein structure recognizes that this is a severe approximation of reality. However, with the small database available to Dayhoff when she constructed her scoring matrices, it was impossible to use a more sophisticated model.

With the results of the exhaustive matching (47), it became possible to examine this assumption directly. A 400 × 400 dipeptide mutation matrix was built that showed the probability of each dipeptide being replaced by every other dipeptide divided by the probability that the exchange would occur if adjacent residues mutated independently (65). The matrix showed that if residue i is conserved, then residue $i + 1$ is more likely to

TABLE 1. PROBABILITY THAT RESIDUE ($i + 1$) IS CONSERVED IF RESIDUE i IS CONSERVED

Amino Acid		log[prob (X = Z)] - log[prob ($X \neq$ Z)]
Pro-X	Pro-Z	−12.5
Gly-X	Gly-Z	−3.9
Glu-X	Glu-Z	−2.1
Lys-X	Lys-Z	0.0
Asp-X	Asp-Z	0.6
Ser-X	Ser-Z	1.2
Leu-X	Leu-Z	1.5
Ala-X	Ala-Z	1.5
Asn-X	Asn-Z	3.8
Arg-X	Arg-Z	4.8
Gln-X	Gln-Z	5.0
Thr-X	Thr-Z	5.4
Phe-X	Phe-Z	5.7
Ile-X	Ile-Z	7.1
Tyr-X	Tyr-Z	8.0
Val-X	Val-Z	8.3
Cys-X	Cys-Z	8.5
Trp-X	Trp-Z	10.5
His-X	His-Z	16.3
Met-X	Met-Z	16.8

be conserved, while if residue i is variable, then residue $i + 1$ is more likely to be variable (Table 1). This is not unexpected on structural grounds. If residue i is conserved, it is more likely to lie inside the folded structure. This implies that the adjacent residue is also more likely to be inside the folded structure, and also more likely to be conserved.

Table 1 shows exceptions to this rule, however. If residue i is a conserved proline or glycine, then it is more likely that the adjacent residues will be *variable*. Why this might be the case is conjectural. However, three hypotheses could be proposed (65):

(a) A conserved Pro or Gly defines a critical turn or break in a regular secondary structural element.

(b) Turns normally occur on the surface.

(c) Surface residues are generally more variable than residues not on the surface.

Again, certain of these hypotheses are controversial. Nevertheless, if they are accepted, then they can be used in predicting the conformation of proteins. When a Pro or Gly is conserved and adjacent to a polar variable residue, then a turn can be assigned at this point in the sequence. In our sample segment of the protein kinase alignment, a conserved G can be found adjacent to a variable residue; a turn is therefore assigned to this segment (Fig. 3).


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

FIG. 3. Assigning turns using conserved Gly and Pro adjacent to variable positions as parsing indicators.

The segment in the middle of the sample alignment in Figure 3 is now "parsed". Neither the secondary structure of the segment lying between the coil and the turn, nor the structures before the coil or after the turn, are known. However, each segment must have a secondary structure independent of what lies before and after. Thus, through models based on a detailed empirical understanding of the divergent evolution of protein sequences, heuristics can be developed that localize the conformation problem, at least to a degree.

Patterns of Amino Acid Substitution Depend on Evolutionary Distance

Another assumption made when using a Dayhoff scoring matrix (61) to align homologous protein sequences is that subsequent mutations during the evolutionary history of a protein occur independently of earlier mutations, that is, that the pattern of mutation at long PAM distances can be obtained simply by repeating the pattern of substitution at short distances multiple times. An empirical study of point mutation in the aligned sequence pairs obtained from the exhaustive matching showed that this was not the case (66). Specifically, terms in the Dayhoff matrix (normalized to PAM 250 for comparison) are different depending on the evolutionary distance separating the proteins used to derive it. Two sets of these terms are shown in Table 2.

At low PAM distances, pairing Trp with Arg yields a relatively high score, while pairing Trp with Phe yields a relatively low score. This is remarkable on chemical grounds, as Trp and Arg are as different in chemical properties

TABLE 2. DEPENDENCE OF THE MATRIX ELEMENT SCORING TRP-ARG AND TRP-PHE MISMATCHES AS A FUNCTION OF EVOLUTIONARY DISTANCE

Average PAM of data set	Trp Arg cross term	Trp Phe cross term
5.5	1.5	−3.9
10.2	0.5	−0.9
42.5	−1.3	1.3
86.5	−1.8	3.0

as possible within the 20 proteinogenic amino acids, while Trp and Phe are as similar as possible. Again, explanations why are conjectural, but possible. The codon for Trp is UGG. Of the 9 possible point mutations in this codon, two lead to Arg codons, while none lead to Phe codons. This observation implies the hypothesis that at low divergence, the code influences strongly not only the distribution of point mutations (a trivial hypothesis), but also the distribution of accepted point mutations (a surprising hypothesis).

Again, this view, although controversial, is useful in structure prediction. Code-driven variation (e.g., Trp–Arg) is consistent with continued biological function only when the amino acid lies on the surface of the folded protein. As code driven variation is easily recognized in similar proteins, it can be assigned to the surface. In particular, the sample segment of the protein kinase alignment has a Trp–Arg code-driven match. This position is therefore assigned to the surface of the folded protein (Fig. 4).

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

FIG. 4. Assignment of a surface residue using code-driven variation.

Note that an assignment of a residue to the surface of a protein provides tertiary structural information, information likely to be essential for assigning secondary structure, even if parses are available which localize secondary structure elements. It is also important to note how a detailed, empirically-based understanding of divergent evolution helps develop tools for extracting secondary and tertiary information from a set of aligned homologous sequences. In no case does this analysis follow the prescription (49) that evolutionary information is useful because it permits averaging of classical predictions to remove noise.

THE E.T.H. METHOD— OBTAINING TERTIARY STRUCTURAL INFORMATION

Clearly, considerably more tertiary structural information must be generated before secondary structure can be assigned. To illustrate how the Zurich method uses both an evolutionary model and sequence information to extract tertiary structural information useful for making secondary structural predictions, we consider heuristics for identifying positions in a multiple alignment that hold amino acids whose side chains protrude on the surface of the folded structure.

Identifying Surface Residues

In principle, amino acids with side chains lying outside the folded structure should be more easily changed without disrupting function than those lying inside the folded structure. Given this assumption, one might define a *naive heuristic* for assigning positions in an alignment to the surface. If there are multiple amino acid replacements at a position, and if some of the amino acids are "surface indicating" (Lys, Arg, Glu, Asn, or Asp, KREND), the heuristic predicts that the side chain at that position protrudes into solvent (16, 19, 67).

The results obtained using this heuristic are disappointing (Table 3). Two hypotheses may explain why. First, the significance of variation depends on its *context*. Variation is more significant between two proteins that are overall similar in sequence, and less significant in proteins that are overall less similar in sequence. Thus, variation within a family of proteins can be productively discussed only in the context of a quantitative measure of evolutionary divergence within the family. Two methods have been used in Zurich to quantitate evolutionary divergence. Both involve subdividing the proteins in the alignment into subfamilies or "subgroups" representing subbranches of the evolutionary tree; the first uses minimum pairwise identity (MPI) (16, 19) to quantitate the extent of evolutionary divergence within a subfamily of proteins, the second the more rigorous PAM distances (number of accepted point mutations per 100 amino acids) (67).

TABLE 3. ACCURACY OF SURFACE ASSIGNMENTS USING THE NAIVE AND REVISED SURFACE HEURISTICS

	Naive (%)	Revised (%)
Aspartate aminotransferase	82	93
Alcohol dehydrogenase	69	91
Lactate dehydrogenase	78	86
Myoglobin	85	94
Plastocyanin	91	100
Superoxide dismutase	81	98
Average	81	93

Second, the naive heuristic fails to distinguish between neutral and adaptive variation. Neutral variation involves amino acid substitutions that do not influence the behavior of a protein sufficiently to influence survival. Adaptive variation is the opposite. It involves amino acid substitutions that *do* change the behavior of a protein sufficiently to influence survival, and accumulates during the divergent evolution of a protein family whose

members are evolving to fulfill different physiological roles in different environments.

Neutral and adaptive variation convey essentially opposite messages regarding the structural environment of a particular amino acid. Neutral variation is on the surface, where it has a minimal impact on the behavior of a protein. But adaptive variation is *supposed* to change behavior. It can lie anywhere in the folded structure, even in the active site. Unfortunately, both types of variation look the same to the naive heuristic. To identify surface positions from variation, a heuristic must therefore distinguish between neutral and adaptive variation.

An evolutionary hypothesis can be used to interpret sequence variation among homologous functional proteins. A particularly useful hypothesis is the "hypothesis of independent variation". It holds that position n in a multiple alignment is not likely *both* to adapt two different proteins in one branch of an evolutionary tree to two different environments in that branch *and* to adapt two different proteins in another branch to two different environments as well (Fig. 5). This hypothesis is not "proven"; it would be difficult to prove in any general way. However, it suggests a way that the naive surface heuristic might be revised to improve its performance. In the revised heuristic, a position is assigned to the surface only if it is variable in at least *two* subfamilies within the multiple alignment. According to the hypothesis of independent variation, this should filter out adaptive variation, and leave behind neutral variation likely to indicate surface position. Consistent with this view, the predictive success is higher (Table 3). A detailed study of this and related heuristic is presented elsewhere (67).

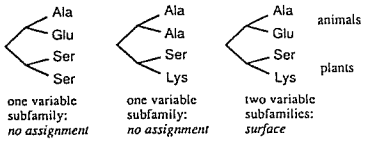


FIG. 5. An illustration of concurrent variation as a tool for assigning surface positions in a multiple alignment.

Although the heuristic is most conveniently applied by computer, it is simple enough to apply by hand to a segment of an alignment for protein kinase, where several sequences have been added to the two considered to this point (Fig. 6). For each subfamily, at each position, if the position contains more than one type of amino acid (is variable) and at least one


```
                coil                        turn
      s   s  s     ss        s   s  s   s
44 DLYTYLSRRLNP--LGRPQIAAVSRQLLSAVDYIHRQGIIHRDIK
46 DLFDFITERGA---LQEDLARGFFWQVLEAVRHCHNCGVLHRDIK
   s     sssss    ss   s    s  s  s   sss
04 EMFSHLRRIGR---FSEPHARFYAAQIVLTFEYLHSLDLIYRDLK
09 ELWTILRDRGS---FEDSTTRFYTACVVEAFAYLHSKGIIYRDLK
11 DLMYHIQQVGK---FKEPQAVFYAAEISIGLFFLHKRGIIYRDLK
    ss  sss        ss ss  s   s  s     s
20 ELFEDIVAREY---YSEADASHCIQQILEAVLHCHQMGVVHRDLK
23 ELFDYLTEKVT---LTEKETRKIMRALLEVICTLHKLNIVHRDLK
24 ELFERIVDEDYH--LTEVDTMVFVRQICDGILFMHKMRVLHLDLK
25 ELFERIIDEDF---LTERECIKYMRQISEGVEYIHKQGIVHLDLK
   s  s  sssss    s ss ss  s   s  s   sss s
28 ELFDYIVQRDK---MSEQEARRFFQQIISAVEYCHRHKIVHRDLK
29 ELFHYIRKHGP---LSEREAAHYLSQILDAVAHCHRFRIRHRDLK
30 QLLHYIIQHGS---IREHQARKFARGIASALIYLHANNIVHRDLK
    ss  sss ss   s ss ss  s   s  s     ss
33 DLKRYMEGIPKDQPLGADIVKKFMMQLCKGIAYCHSHRILHRDLK
35 DLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCHSRRVLHRDLK
36 DLRTYLDKAPPPG-LPAETIKDLMRQFLRGLDFLHANCIVHRDLK
37 DLEVVIKDKSIL--FTPADIKAWMLMTLRGVYHCHRNFILHRDLK
    s     s s
48 SLYKHLHVQETK--FQMFQLIDIARQTAQGMDYLHAKNIIHRDMK
49 SLYHHLHVADTR--FDMVQLIDVARQTAQGMDYLHAKNIIHRDLK
     s   s  s     ss       s   s  s   s
   s     sssss    ss   s    s  s  s   sss
      ss  sss        ss ss  s   s  s     s
   s  s  sssss    s ss ss  s   s  s   sss s
     ss  sss ss   s ss ss  s   s  s    ss
      s     s s

        assignments for all families

   s  s  sssss    s ss ss  s   s  s   sss
44 DLYTYLSRRLNP--LGRPQIAAVSRQLLSAVDYIHRQGIIHRDIK
46 DLFDFITERGA---LQEDLARGFFWQVLEAVRHCHNCGVLHRDIK
   1    1    1    1    1    1    1    1    1
   1    2    2    3    3    4    4    5    5
   5    0    5    0    5    0    5    0    5
```

FIG. 6. Using concurrent variation to assign surface positions in the multiple alignment for protein kinase.

"surface indicating" amino acid (KREND), we write an "s". Collecting these into a common prediction for the entire alignment involves writing "S" whenever more than one subfamily is assigned an "s".

Interior Heuristics

Corresponding heuristics can be developed for assigning positions in a multiple alignment to the inside of the folded structure. These heuristics often rely on conservation (as opposed to variation) of interior-indicating (as opposed to surface-indicating) residues (FAMILYVW). However, the most reliable indicators of interior position are those where interior-indicating residues are coupled with variability. These are reviewed in detail elsewhere (67). For the purpose of the example in Figure 7, a

```
                coil                          turn
     ii ii          i       ii    ii  i ii     ii   i
44 DLYTYLSRRLNP--LGRPQIAAVSRQLLSAVDYIHRQGIIHRDIK
46 DLFDFITERGA---LQEDLARGFFWQVLEAVRHCHNCGVLHRDIK
     ii  i         i       ii i i   i ii    iii  i
04 EMFSHLRRIGR---FSEPHARFYAAQIVLTFEYLHSLDLIYRDLK
09 ELWTILRDRGS---FEDSTTRFYTACVVEAFAYLHSKGIIYRDLK
11 DLMYHIQQVGK---FKEPQAVFYAAEISIGLFFLHKRGIIYRDLK
     ii  i         i         i  i  ii        ii   i
20 ELFEDIVAREY---YSEADASHCIQQILEAVLHCHQMGVVHRDLK
23 ELFDYLTEKVT---LTEKETRKIMRALLEVICTLHKLNIVHRDLK
24 ELFERIVDEDYH--LTEVDTMVFVRQICDGILFMHKMRVLHLDLK
25 ELFERIIDEDF---LTERECIKYMRQISEGVEYIHKQGIVHLDLK
     ii ii         i    i  ii  ii ii        i    i
28 ELFDYIVQRDK---MSEQEARRFFQQIISAVEYCHRHKIVHRDLK
29 ELFHYIRKHGP---LSEREAAHYLSQILDAVAHCHRFRFRHRDLK
30 QLLHYIIQHGS---IREHQARKFARGIASALIYLHANNIVHRDLK
     i  ii         i    i  ii       i        ii   i
33 DLKRYMEGIPKDQPLGADIVKKFMMQLCKGIAYCHSHRILHRDLK
35 DLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCHSRRVLHRDLK
36 DLRTYLDKAPPPG-LPAETIKDLMRQFLRGLDFLHANCIVHRDLK
37 DLEVVIKDKSIL--FTPADIKAWMLMTLRGVYHCHRNFILHRDLK
     ii  i i       i ii ii ii    i  i ii i  ii   i
48 SLYKHLHVQETK--FQMFQLIDIARQTAQGMDYLHAKNIIHRDMK
49 SLYHHLHVADTR--FDMVQLIDVARQTAQGMDYLHAKNIIHRDLK
assignments for all families
    II iI        I i  I  II  iI II II    IiI  I
Combine surface and interior assignments.
Question: What is the secondary structure between the coil and
the turn?
                coil                          turn
44 DLYTYLSRRLNP--LGRPQIAAVSRQLLSAVDYIHRQGIIHRDIK
46 DLFDFITERGA---LQEDLARGFFWQVLEAVRHCHNCGVLHRDIK
   SIISiIssSSSS  ISisSIssIIssiISIIsII sSSIiI  I
   1    1    1    1    1    1    1    1    1
   1    2    2    3    3    4    4    5    5
   5    0    5    0    5    0    5    0    5
```

FIG. 7. Using concurrent variation to assign interior positions in the multiple alignment for protein kinase interior heuristics. Interior heuristics assigning positions to the inside of the fold are based on patterns of conservation and variation in hydrophobic residues. For each subfamily, at each position, if every amino acid is "interior indicating" (Phe, Ala, Met, Ile, Leu, Tyr, Val, or Trp, Pro, Gly, FAMILYVWPG), an "i" is written.

subfamily contributes to an interior assignment (denoted by an "i") when all of the proteins have an interior indicating amino acid (FAMILYVW) at the designated position.

THE E.T.H. METHOD — PREDICTING SECONDARY STRUCTURE

Of the heuristics discussed above, only the parsing heuristics directly provide secondary structural information. The surface, interior, and active site heuristics (the last discussed in detail elsewhere) (16, 19) do not. Rather they orient a side chain relative to the global conformation of a protein, and therefore provide a type of tertiary structural information. This information could conceivably be used to model tertiary structure directly. For example,

the protein could be represented by two spheres, one describing the surface of the protein, the other describing the active site. Positions assigned to the surface could be placed on the surface sphere, positions assigned to the active site could be placed on the active site sphere, the assigned residues moved on the spheres to satisfy constraints imposed by the connectivity of the polypeptide chain while contracting the surface sphere until it is packed by the interior residues to an appropriate density. It is important to note that in this model, both the alpha helix and the beta strand are defined by relative side chain orientations, not by hydrogen bonding. This definition is different from that often used when assigning secondary structure from crystallographic data.

In Zurich, a different approach has been followed, where surface, interior, active site, and parse assignments are first used to create an unrefined secondary structure prediction. Here, alpha helices and beta strands are assigned to parsed segments of the alignment based on 3.6 residue and 2 residue periodicities in the surface and interior assignments. In our test example taken from protein kinase, this approach can be followed directly. In the region between the coil and the turn, the pattern of surface and interior assignments clearly displays a 3.6 residue periodicity, indicating an alpha helical structure (Fig. 8).

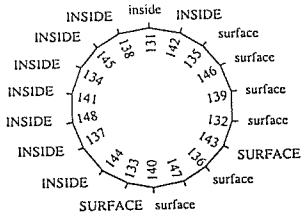


FIG. 8. Assigning a helix to the parsed segment of the protein kinase multiple alignment.

In the bona fide prediction of the secondary structure of protein kinase published on these pages three years ago (19), a helix was in fact assigned to these positions using this rationale. The helix was found in the subsequently determined crystal structure (26), illustrating the power of the E.T.H. approach, at least for identifying surface helices. Indeed, overall the secondary structure prediction for the protein kinase superfamily was remarkably accurate (26) (but see also Ref. (49)). In particular, it was adequate to support an effort to model the tertiary

structure of the first domain of the protein family, where an antiparallel beta sheet in the first domain was successfully identified. The power of the E.T.H. method was especially noteworthy when compared with bona fide predictions made by classical methods (68–74). Standard Chou-Fasman (2) and GOR (44) predictions gave rather poor bona fide predictions for this protein. Further, prediction methods based on motifs were defeated by the Gly–X–Gly–X–X–Gly–$(X)_5$K sequence found in protein kinases, which was interpreted as evidence that the protein folded to form a Rossmann fold found in many other nucleotide binding enzymes. Protein kinase turned out not to contain a Rossmann fold (32).

Using patterns in surface and interior assignments alone to assign secondary structures has clear limitations when applied to secondary structural units that lie entirely within the folded structure. In early predictions, strings of consecutive interior positions were assigned canonically as beta strands. Internal helices were assigned (e.g., in alcohol dehydrogenases) because they passed near the active site, and active site residues appeared with a distinctive 3.6 residue periodicity (16). A method of this type is clearly vulnerable to error when presented with an internal helix that does not pass near the active site.

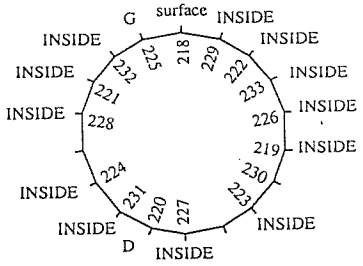


FIG. 9. The internal helix missed in the protein kinase prediction.

This error was in fact made in the protein kinase prediction (Fig. 9), where a helix that lay entirely within the folded structure was assigned to be a rather long internal beta strand. Had the classical paradigm been followed, this error would have been simply lost in a three-state score aggregated over the entire sequence, itself buried within the three-state scores for a large number of other protein families. However, following the new paradigm, the error was analyzed, and effort was devoted to

developing improved heuristics for identifying internal helices. These were then applied to subsequent predictions. For example, an internal helix was predicted for the hemorrhagic metalloprotease (collagenase) family from snake venom (Fig. 10) using these heuristics (24). The prediction is discussed further below.

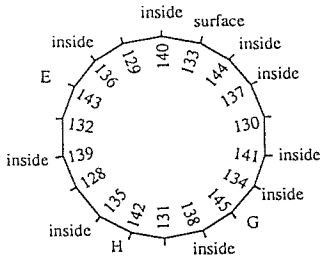


FIG. 10. An internal helix predicted for the collagenase family.

THE E.T.H. METHOD — CONSENSUS MODELS AND THEIR LIMITATIONS

A second shortcoming was demonstrated by the protein kinase prediction. When building a single structural model for a family of proteins, one assumes that the conformations of all of the proteins in the family are the same (56). This is, of course, only true as an approximation. To the extent that it is not true, the structural model is an "average" or "consensus" prediction, applying exactly to no individual protein in the family. It can, however, serve as the starting point for modeling the structure of every protein in the family (75).

Within a superfamily, structure can diverge. This creates problems, not least because classical automated scoring methods rarely take divergence into account. For example, Rost *et al.* (49) used the crystal structure from mouse cAMP-dependent protein kinase to evaluate the consensus prediction for the entire kinase family (19). The cAMP-dependent kinase contained an additional short helix (alignment positions 052–057, see Ref. [19]) that was matched against a gap in other members of the protein kinase superfamily. Because of this gap, the consensus model did not assign any structure to this region, although the text of the prediction paper stated that "segment 52–57 could adopt a standard secondary structure" in the cAMP-dependent protein kinases (19). Rost *et al.* evidently did not read this text, as they counted the prediction incorrectly in this region.

Related problems arise when the multiple alignment used to build an unrefined prediction is bad. The technology for constructing multiple alignments remains imperfect, and experts routinely rearrange computer-generated multiple alignments by hand to meet an aesthetic that is often difficult to explain. Nevertheless, the rearranged multiple alignments are often more useful in structure prediction than the original multiple alignment. Thus, in a refinement process, the multiple alignment is adjusted. This adjustment is guided by secondary structure predictions for subfamilies of proteins. During this process, the secondary structure itself can be reassigned. For example, evidence for rotation of a helix during divergent evolution is a strong confirmation of a helix prediction.

Because a consensus model for a protein family does not apply exactly to any single family member, it must be evaluated by examining more than one experimental structure, especially if the family has undergone substantial divergence in sequence. As will be noted below, one of the primary difficulties that scientists trained in the classical paradigm have had with the new paradigm has been their failure to recognize the implications of divergence in conformation when evaluating consensus predictions (see Ref. (49) and below). An experimental structure from each significant subfamily (joined by a bridge at a PAM distance greater than 120) is ideal, of course, but many structures are available for a single protein family only in special cases. Nevertheless, even a single additional structure can distinguish between predictions that are "wrong" because conformation within the family of proteins has diverged, and predictions that are in fact wrong.

THE E.T.H. METHOD— REFINING A SECONDARY STRUCTURE AND MODELLING TERTIARY STRUCTURE

These and other limitations inherent to the building of a consensus prediction for secondary structure were recognized early in the development of the E.T.H. method (16), and a refinement procedure was introduced to address them. Refinement involves several steps:

(a) Separate secondary structure predictions for subfamilies of the multiple alignment are constructed to assess the possibility of divergence in secondary structure.

(b) The multiple alignment is adjusted and the consensus secondary structure prediction revised accordingly.

(c) A preliminary supersecondary structure model is assembled, with revision of the secondary structure if necessary.

(d) A covariation analysis is performed within the context of the supersecondary structural model, and the secondary and supersecondary structures are modified as appropriate. In the covariation analysis, residues

distant in the primary structure are sought that undergo compensatory variation that might suggest that they lie close in space. Significant covariation cannot generally be distinguished from general variation when taking the protein as a whole or when examining both distantly and closely related members of a superfamily. Rather, covariation is used to confirm specific hypotheses that juxtapose specific secondary structural units, and even then is useful only in protein pairs that have diverged from 30–80 PAM units. At longer PAM distances, pairwise compensatory mutation does not appear to be the rule. At shorter PAM distances, variation is localized on the surface. A good example of the use of covariation within the context of a supersecondary structural model can be found in the bona fide prediction for protein kinase, where covariation at alignment positions 87 and 108 was used to assign an antiparallel orientation of predicted beta strands 3 and 4 (19).

A refined secondary structure prediction is the starting point for building a tertiary structural model. The goal of the modeling is first to orient secondary structural units with respect to active site residues, disulfide bonds, and whatever remote interactions might be detected by covariation analysis. The orientation also may exploit packing rules (the packing of a beta–alpha–beta unit being the most reliable). In some cases, patterns in the secondary structure assignment suggest a particular type of fold (e.g., an 8-fold alpha beta barrel).

Preparing a consensus tertiary structural model can prompt further refinement of the consensus secondary structure prediction. Secondary structure predictions for individual subfamilies of the protein family are used to construct preliminary models for supersecondary structural units. These are then analyzed using covariation analysis to identify positions distant in the primary structure but near in the tertiary structure (19). The predicted supersecondary structural units are then oriented with respect to an active site and (when available) disulfide bonds. This may lead to alterations in the secondary structure assignments. Further, understanding divergence in function and environment in the family of proteins influences refinement; as we have noted elsewhere, the best analyses of conformation are done by those who understand function and reactivity of the family of the macromolecule in question, and use what they know in building models (16, 19). In this respect again, conformational analysis in proteins is similar to that in organic chemistry generally.

Several predictions from Zurich have not exploited this refinement procedure, in particular those made in response to challenges from outside Zurich where inadequate time has been available. To the crystallographer, the challenge is often an afterthought, made after the structure is solved, the manuscript is accepted for publication, and the proofs are in the mail. To the predictor, this often means that only a few days are available after

a challenge is received to assemble a prediction and write the paper. This effectively excludes refinement.

RESULTS — THE SH3 DOMAIN PREDICTION AND THE CRITICS

While the protein kinase prediction was remarkably accurate (26), a single prediction is not sufficient to test a method. Indeed, in the new paradigm, a prediction is simply a tool to develop the insight into protein structure needed to improve prediction heuristics. These in turn become interesting only when they are tested in another bona fide prediction exercise. Therefore, the Zurich group and other groups (76) have produced a steady stream of bona fide predictions. To date, bona fide predictions made using the E.T.H. or related methods have included: (a) the quaternary structure of yeast alcohol dehydrogenase (16), (b) the secondary and supersecondary structure of protein kinase (19), (c) the secondary structure of the SH2 domain (20), (d) the secondary structure of the SH3 domain (22), (e) the secondary structure of a fragment of histidinol phosphate aminotransferase (77), (f) the secondary structure of the MoFe nitrogenase protein (23), (g) the secondary and supersecondary structure of hemorrhagic metalloprotease (24), (h) the secondary, supersecondary, and tertiary structure of venom allergen, and (i) the secondary structure of isopenicillin N synthase. Experimental structures are now available for protein kinase (26), the SH2 domain (78), the SH3 domain (28), the MoFe nitrogenase protein (30), and hemorrhagic metalloprotease. In addition, models have been assembled for other protein families, including the fumarase-aspartase superfamily and the protein phosphatase superfamily, where experimental structures do not appear to be imminent.

The review by Thornton *et al.* (41) of the protein kinase prediction made in Zurich encouraged several groups to contact the Zurich group with prediction challenges. In particular, the September of 1992, Dr A. Musacchio challenged the Zurich group to predict the conformation of the SH3 domain. The structure had been solved in the laboratory of M. Saraste in Heidelberg, and was scheduled to be published in October. A manuscript describing an unrefined prediction for the SH3 domain superfamily was prepared in a week and rushed through the reviewing process at the *Journal of Molecular Biology*, where it was accepted on the day that the crystal structure appeared. Simultaneously, a letter from Zurich appeared in *Nature* summarizing the prediction (79).

The publication of the prediction of the secondary structure for the SH3 domain caused the first direct engagement between the classical and new paradigms. This engagement came in the form of a Scientific Correspondence in *Nature* contributed by Sander and his coworkers, who offered their services as a "jury" to render a verdict on the SH3 domain

prediction (80). The "good news" they concluded was that the secondary prediction made in Zurich was "correct." The "bad news", they concluded, was that the tertiary structure prediction made in Zurich was "wrong."

The Heidelberg jury's deliberations were hampered by the fact that none of its members had actually read the prediction paper, which had not yet appeared in print. In this paper, the Zurich group explicitly noted that it had not predicted a tertiary structure for the protein (nor is a tertiary structure prediction sensible with only an unrefined secondary structure prediction). Nor had the jury evidently reviewed the structures of homologs for the SH3 domain that were then beginning to emerge (29, 81). Therefore, the jury did not have the information from homologous structures needed to evaluate a consensus prediction in a superfamily of proteins that had undergone substantial sequence divergence. One point that can be made from Ref. (79) is that it is important to hear the evidence before rendering a verdict.

Another point is how rapidly misinformation presented as Scientific Correspondence in *Nature* is propagated throughout the scientific world. Scarcely two months later, Geoffrey Barton cited the jury's verdict in asserting that "the prediction of Benner *et al.* was in the event disappointing in some respects." (82). Barry Robson and Jean Garner also cited the jury's verdict in asserting that "in attempting to predict the structure of SH3, the dice have fallen badly for Benner *et al.*" (48). These authors also had not read either the prediction paper or the papers reporting the structures of homologs for the SH3 domain. We are certain that Barton appreciated the importance of examining multiple homologous structures in evaluating a prediction, as he had done so when he evaluated his own prediction for the SH2 domain (82). The question of the accuracy of the SH3 domain prediction was central to the controversy that ensued, and we shall return to it momentarily.

The most important point to extract from the jury's verdict relates, however, to how predictions and prediction methods must be evaluated. In the "jury" paper (80), Rost and Sander asserted that a neural network developed in Heidelberg would have done "significantly better" in predicting secondary structure for the SH3 domain than the E.T.H. method did. This conclusion was based on a per-residue score obtained in a "blind" prediction applying the neural network to the sequence of the c-spec SH3 domain, whose structure had been solved in Heidelberg and was known to the jury. The three state score calculated by Rost and Sander was 56% for the E.T.H. prediction and "reached 70%" for the Heidelberg method.

The use of the per residue score as the deciding criterion to compare the two predictions identified Rost and Sander as adherents to the classical paradigm. Other classical characteristics of their neural network

was that it predicted secondary structure without explicitly considering tertiary structure, it was fully automated, and it had been tested only on proteins with known crystal structures. Other aspects of the jury's verdict suggested, however, that the adherence was not dogmatic. For example, Rost and Sander (79) also scored the SH3 domain prediction by counting the number of secondary structural *elements* correctly assigned. This is, of course, a much better indicator of the value of a secondary structure prediction. Using per-segment scores, the Zurich and Heidelberg predictions were equally satisfactory; in both cases, four out of the five (80%) of the predicted secondary structural units were viewed by the jury to have been correct.

Further, the Heidelberg neural network used as input a set of aligned protein sequences, rather than a single sequence, for making a prediction. In using the network, the user submits a single "guide sequence" to the Heidelberg server. The server then finds in the database all sequences that are clearly homologous, builds a multiple alignment, and makes a secondary structure prediction. This is reported together with the multiple alignment and numerical values that indicate, residue-by-residue, the probability that the secondary structure assignment is correct. Thus, the network presumably produced a consensus model for the proteins being aligned, rather than a model for any particular protein in the multiple alignment, just as did the method in Zurich.

If the neural network in fact produced a consensus prediction, the 70% per residue score reported for the SH3 domain is problematic: it is too good for a consensus prediction made for a family of proteins that has undergone substantial sequence divergence. Within the SH3 domain superfamily, sequence similarities between the most divergent proteins (15–20%) were barely adequate to indicate common ancestry (Fig. 11) (22). The high sequence divergence implies substantial divergence in secondary structure. The secondary structures of different SH3 domains from different branches of the evolutionary tree correspond only ca. 70% of the positions (Fig. 11). Considering the subjective nature by which secondary structure is assigned to experimental structures in the first place (54), the claimed per-residue accuracy was higher than it could possibly be for a consensus prediction evaluated by one member of the SH3 domain superfamily, in this case the c-spec structure.

We considered three possible explanations for this phenomenon. First, it was possible that the neural network did not in fact build a consensus model for the entire SH3 domain family, but rather only for the c-spec sequence and close homologs. If this were the case, the network should produce different predictions, but with similarly high scores, when proteins from other branches of the SH3 domain evolutionary tree were used as guide sequences.

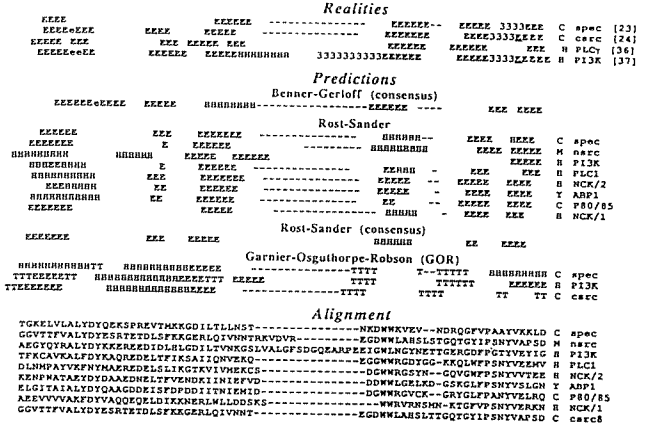


FIG. 11. Multiple alignment and predictions for the SH3 domain. For the SH3 domain superfamily (a) the experimental secondary structures assigned for four homologous SH3 domains by three different research groups, (b) the consensus prediction made for the SH3 domain superfamily in Zurich, (c) the individual predictions made for a sampling of the SH3 domains by the Heidelberg neural network, and (d) individual predictions made for a sample of SH3 domains by the classical GOR method.

To test this possibility, sequences of proteins from other subfamilies of the SH3 domain family were submitted to the Heidelberg neural network via a server. The results are shown in Figure 11. The neural network indeed yielded different predictions depending on which homolog was entered as the guide sequence. Further, different guide sequences found different subsets of the SH3 domain superfamily. Thus, different guide sequences led to different consensus predictions for different protein sequences; the different predictions could be explained by the difference in the multiple alignment retrieved by the server. However, the predictions for each subfamily did not match the experimental structure in the subfamily. In all but perhaps one case, the neural network prediction was worse (judging by a per-residue three-state score) than the prediction that had been published in *Nature* for the c-spec domain, the experimental structure structure known to the jury at the time that their made the neural network prediction (83).

Figure 11 presents the following information for the SH3 domain superfamily (a) the experimental secondary structures assigned for four homologous SH3 domains by three different research groups, (b) the consensus prediction made for the SH3 domain superfamily in Zurich, (c) the individual predictions made for a sampling of the SH3 domains by the Heidelberg neural network, and (d) individual predictions made for a sample of SH3 domains by the classical GOR method. Several points are evident. First, far from being disappointing, the Zurich prediction was as good as could have been expected given the divergence in secondary structure within the superfamily as a whole. Second, the GOR method does extremely poorly with this superfamily, and averaging the GOR predictions over the homologous proteins could not improve its poor performance. Finally, the prediction from the neural network, although not as bad as those obtained by the GOR method, are poor, especially if the fortuitiously good prediction for the c-spec domain is excluded. The reason why the c-spec domain was predicted so well remains unknown.

RESULTS— THE CLASSICAL CRITIQUE OF THE CONTEMPORARY PARADIGM

Just two months after the Heidelberg jury had rendered its verdict, *Nature* published another Scientific Correspondence, this time by Barry Robson and Jean Garnier, dons of the classical structure prediction field (48). Their letter evidently had been prepared not as a response to the SH3 domain prediction, but rather to Thornton's earlier review of the protein kinase prediction (41). "Benner *et al.* have fallen foul of the classic mistake of forgetting that one swallow does not make a summer," wrote Robson and Garnier. "By seeking to incorporate intuition, insight, and expertise interactively, Benner *et al.* do not satisfy [the] criteria" of "formal correctness, ease of reproduction and various methods of objective testing." "Reproducibility is the cornerstone of science, and although that may appear hampering to the creative, it does have very considerable benefits."

Much of the Robson–Garnier letter was merely a restatement of the points of the classical paradigm that had been rejected (16) in the formulation of the new paradigm (16, 19). Robson and Garnier again asserted that the "fair and proper" way of evaluating a secondary structure prediction was when "a residue is either right or wrong in its assignment to one and only one of three states." They again insisted that prediction was "fundamentally statistical" in nature, and that all methods must be evaluated by examining a statistically large sample of proteins. They also asserted that it was possible to test a method by examining known structures only, ignoring mechanisms by which knowledge of correct structures might influence

even fully automated prediction tools. They gave no evidence of being aware of the arguments that these elements of the classical paradigm were making the field of structure prediction "stubbornly unyielding." Except for its tongue-in-cheek style, the letter could be viewed as an example of obdurate dogmatism.

RESULTS— THE IMPORTANCE OF BONA FIDE PREDICTIONS

The Heidelberg group also provided a critique of the protein kinase prediction, this time on the pages of *Trends in Biochemical Sciences*, again asserting that their neural network would have predicted the conformation of the family "significantly better" than the Zurich group did (49). Particularly prominent in their critique was the suggestion that the positive evaluations of the protein kinase prediction made by others (41) in the field were "exaggerated."

Much of Rost *et al.* (41) is also a reassertion of the very elements of the classical paradigm that are rejected by the new paradigm. The three-state per residue score is again introduced as the criteria for measuring the quality of a prediction. Methods are evaluated by averaging three-state scores over a statistically large sample of proteins. The neural network was tested on proteins whose structures were already known, and evolutionary information is considered useful only to the extent to which it enables averaging of the outputs of classical prediction heuristics. Finally the role of human involvement in development of prediction tools is profoundly misunderstood.

Most notable about the *TIBS* paper (49), however, is that the Heidelberg group again claimed a remarkably high three-state per residue score for their neural network, 76% when applied to the protein kinase family. This score is still higher than the score claimed for the SH3 domain family. Again, it is surprisingly high for a consensus prediction for this family, where fewer than 15% of the residues are conserved in all members. Again, the high score might be explained in part by the presumption that the sequences used in the Heidelberg prediction were close homologs of the kinase (the cAMP-dependent protein kinase) whose crystal structure had been solved; in contrast, the Zurich prediction was a consensus model for a far larger family of proteins (see above).

However, the *TIBS* article from the Heidelberg group (49) may turn out to have its greatest significance by providing an example of how bias arising from knowledge of the experimental structure influences the *post-hoc* prediction, even when they are made blind by a fully automated computer program. As noted above, the most serious error made in the protein kinase prediction in Zurich was the misassignment of an internal helix as a beta strand (Fig. 9). In the *post-hoc* prediction made by the

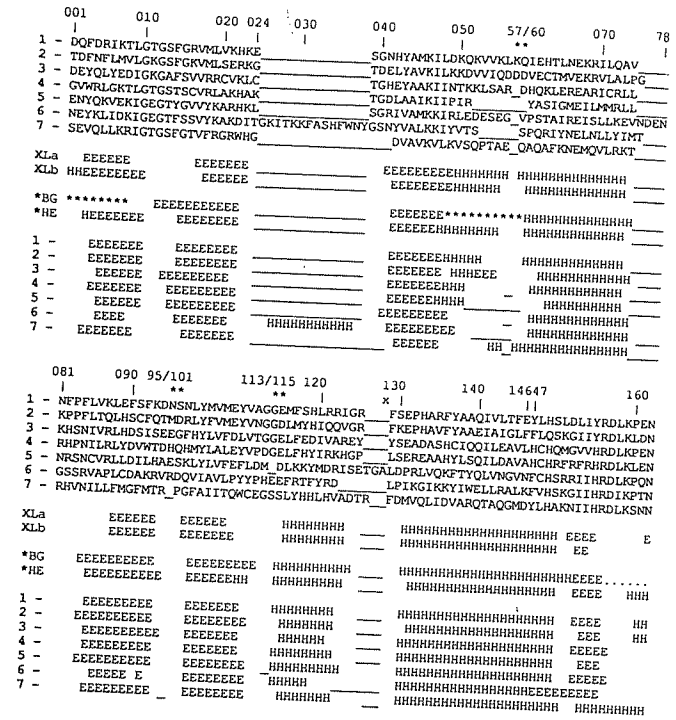


FIG. 13. A comparison of protein kinase predictions made in Zurich and by the Heidelberg neural network. Lines 1–7. Representative sequences taken from the multiple alignment for the protein kinase superfamily. X1a and X1b. Two secondary structure assignments taken from the experimental crystal structure. BG. The Benner–Gerloff prediction made using the E.T.H. method [19]. The published [49] prediction made by the Heidelberg neural network. Secondary structure prediction made by the neural network via server, for the seven protein kinase sequences. Special attention is drawn to (a) positions 225–240, the region of the internal helix missed in the Zurich prediction, correctly assigned in the published prediction made by the neural network, but incorrectly assigned in unpublished predictions made by the network, and (b) positions 048–055, helical in the cAMP-dependent protein kinases (sequence 1), but matched against gap in other kinases.

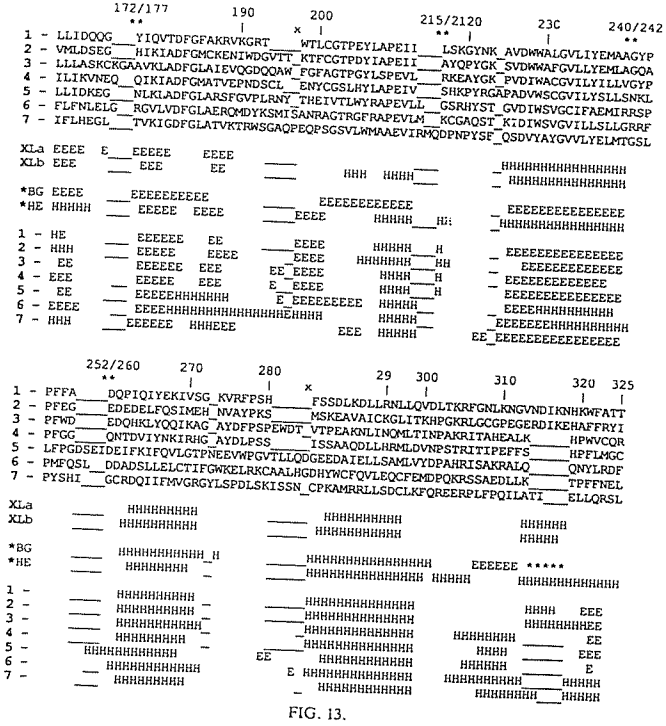


FIG. 13.

Heidelberg neural network and published in *TIBS* (49), this internal helix was correctly identified. This result was, of course, extremely interesting to the Zurich group, as it suggested that a neural network could correctly identify a type of secondary structure (internal helices) that the E.T.H. method detects only with difficulty. This would suggest that combining the Zurich and Heidelberg methods might yield better predictions than either method individually.

Therefore, several sequences of homologous protein kinases (including the sequence of the cAMP-dependent protein kinase from mouse used in Ref. (49)) were submitted to the server. The relevant output is shown in

Figure 13. Remarkably, the only output from the neural network where this internal helix was correctly identified was the one published by the Heidelberg group in their *TIBS* article (49). In every other case, using either the sequence from the cAMP-dependent protein kinase or homologous sequences, the neural network assigned a beta strand to this region, or a mixture of alpha helix and beta strand. In other words, the correct helical assignment for this region reported by the Heidelberg group could not be reproduced.

It is difficult to explain this fact other than to assume that the Heidelberg group selected (presumably inadvertently) the prediction that yielded the highest per residue score from many secondary structure predictions produced by the server. The Heidelberg neural network, as with many other neural networks, behaves somewhat chaotically, with output often sensitive to minor changes in input (83). This example underscores two important points made by the new paradigm: automation and blind testing do not guarantee objectivity, and bona fide predictions are essential as a part of evaluating any new prediction method.

RESULTS— HOW MANY SWALLOWS MAKETH A SUMMER?

We were intrigued by the literary reference made by Robson and Garnier that we had run foul of the classic mistake of forgetting that "one swallow does not make a summer." Fortunately, one of us had organized the Oxford English Dictionary using the same data structure used in the exhaustive matching of the protein sequence database (47). From this we learned that the first citation ("It is not one swallow that bryngeth in somer") in the English language comes from 1552, and is based on a much older anonymous Greek proverb. In its modern interpretation, the proverb implies that one successful prediction does not prove a method.

We do not, of course, know of anyone (including anyone in Zurich) who believes that one (protein kinase) or two (the SH3 domain) remarkably accurate predictions means that the structure prediction problem is solved. Therefore, we have continued to publish bona fide predictions. Two of these, one unrefined and the other refined, can now be evaluated by subsequently reported crystal structures. The unrefined prediction was made for the MoFe nitrogenase protein (23), responding to a challenge from Prof. D. Rees at the California Institute of Technology (30). The refined prediction made for the metal-dependent collagenases from snake venom (24), responding to a challenge from Prof. E. Meyer from Texas A&M. Each demonstrates the state of current technology in Zurich, as well as illustrating problems in evaluating consensus predictions.

For the MoFe nitrogenase, the unrefined prediction (23) was evaluated by grouping the assigned units in 7 categories: "correct" (a predicted secondary structure unit that would not adversely affect an effort to build a tertiary structure model), "possibly correct" (a predicted secondary structure unit whose effect on a tertiary structure model depends on context), "wrong" (a helix assigned as a strand, tabulated as an incorrect strand assignment, or a strand assigned as a helix, an incorrect helix assignment), "missed significant" (a helix or strand not identified in a region that does not contain a gap, and where the missed unit is important to a tertiary structural model), "missed insignificant" (a helix or strand not identified in a region that does not contain a gap, but where the missed unit does not appear important to building a tertiary structure), "gapped" (a helix or strand not identified because of the canonical treatment of gaps (19)), and "overpredicted" (a helix or strand assigned to a region left unassigned by the experimentalists). These numbers for the MoFe nitrogenase protein are collected in Table 4. Note that these are preliminary assignments; precise assignments can be made only in the context of an effort to assemble a tertiary structure model, which necessarily follows refinement.

TABLE 4. SECONDARY STRUCTURE OF THE MoFe NITROGENASE PROTEIN: COMPARISON OF THE PREDICTION AND THE CRYSTAL STRUCTURE

	Alpha Helices	Beta Strands
Correct	10	7
Possibly correct	0	2
Wrong	0	3
Missed significant	3	4
Missed insignificant	3	0
Gapped	2	1
Overpredicted	0	2

An overview of the MoFe nitrogenase prediction is given in Ref. (23). The prediction shows both the strengths and the limitations of an unrefined prediction made using the E.T.H. method. For example, each of the ten predicted helices corresponded to a helix found in the experimental structure. The prediction of long surface helices with nearly perfect accuracy is a characteristic of unrefined predictions made by the E.T.H. method.

The errors are of two types. First, errors arise because of a poor alignment in a region of significant sequence divergence. The multiple alignment of the MoFe nitrogenase homologs included some rather distant sequences, and regions of the alignment were rather poor. Several additional helices could have been detected (together with a reorientation of individual helices with

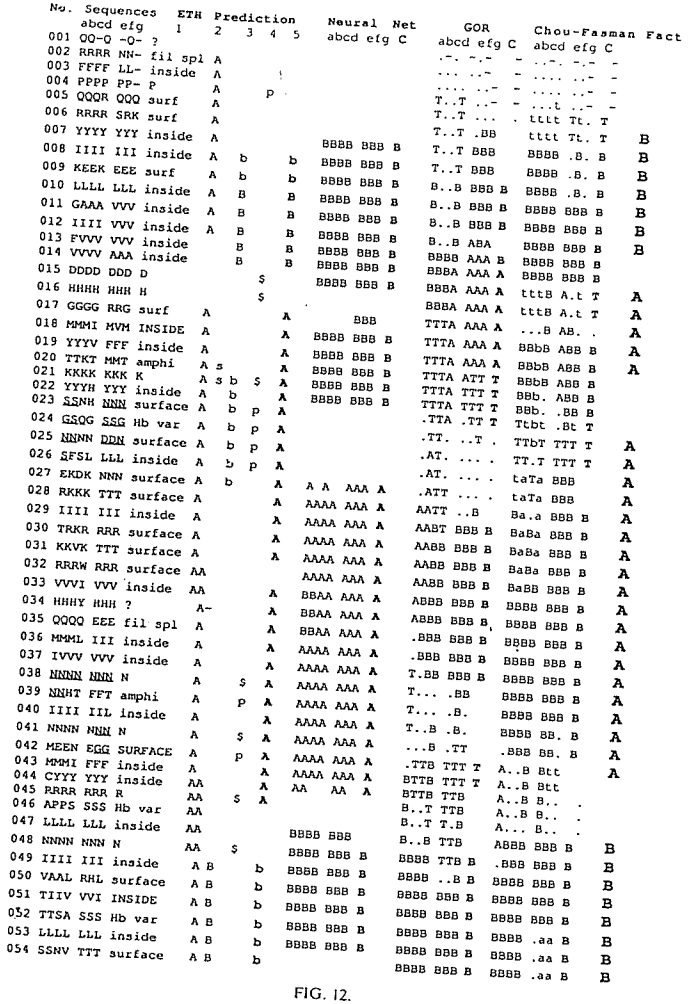


FIG. 12.

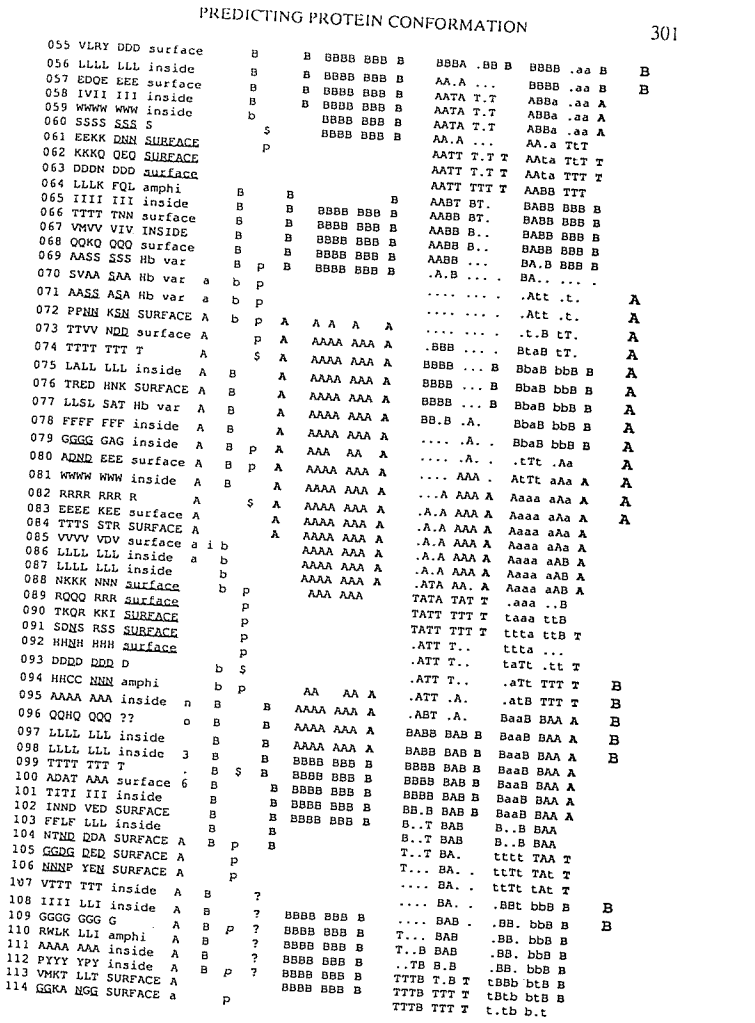


FIG. 12 (cont'd).

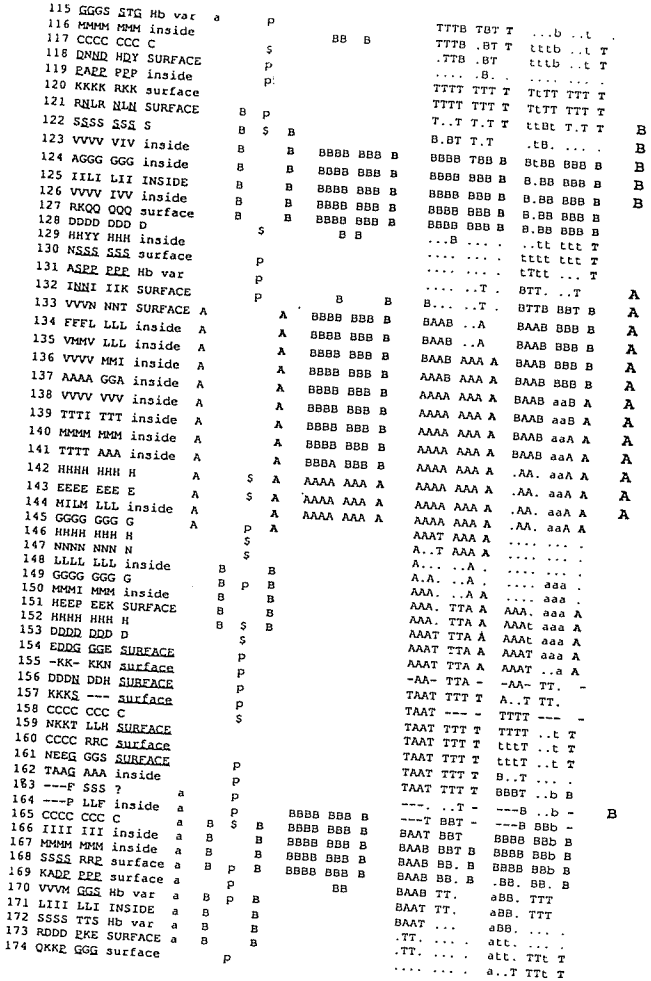


FIG. 12 (cont'd).

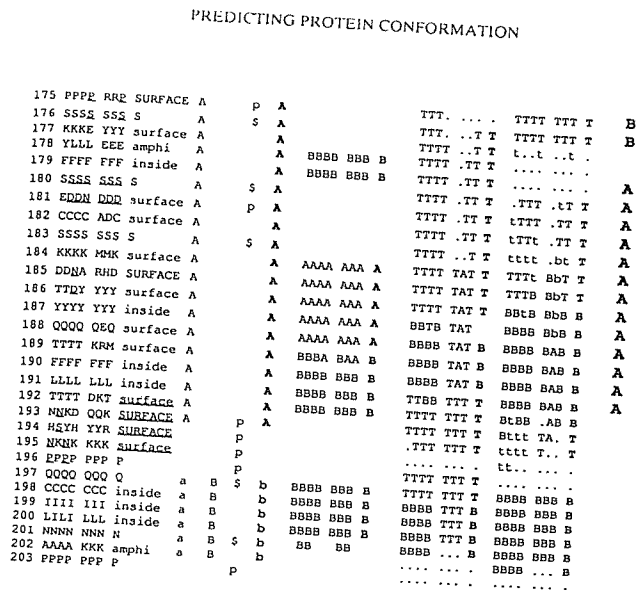


FIG. 12. Comparison of bona fide secondary structure predictions made by various methods for seven hemorrhagic metalloproteases, or collagenases, with a secondary structure assignment derived from a subsequently determined crystal structure. Alignment position numbers and sequences shown vertically, one letter code used, underlining indicates parsing strings. Accession numbers (SwissProt): (a) P15503; (b) P14530; (c) P20165; (d) P20164; (e) P20897; (f) P15167; (g) P22796, followed by secondary structure predictions. "A", "a", "B", "b", "T" and "t" indicate strong and weak alpha helix, beta strand, and turn assignments, respectively. A "?" indicates that an assignment is uncertain. *E.T.H. prediction* [1–5]. Column 1 shows surface/interior assignments made using the E.T.H. method (underlining indicates consecutive surface assignments indicative of breaks in secondary structure); "fil spl": hydrophilic split; "amphi": amphiphilic position separating two branches in tree; "Hb var": hydrogen bonding variable. Column 2 shows possible alpha helix assignments made using the E.T.H. method. Trailing character "s" indicates that amphiphilicity is preserved only if the position is assigned to the surface, while trailing character "i" indicates that amphiphilicity is preserved only if the position is assigned to the inside. Unassigned is a coil. Column 3 shows possible beta strand assignments made using the E.T.H. method. Column 4 shows parses made using the E.T.H. method ("*p*" indicates a strong parse, "p" a weak parse) and active site assignments ($). Column 5 shows consensus secondary structure assignments made using the E.T.H. method. *Neural net prediction*. Provided by Rost and Sander. Consensus prediction under C. *GOR Prediction*: Consensus prediction under C. *Chou-Fasman prediction*: Consensus prediction under C.

respect to the surface of the protein during divergent evolution) had time for a refinement been available. Second, errors in unrefined predictions are frequent near the active site of a protein, where the functional constraints on divergent evolution arising from the need to maintain catalytic activity dominate over those that indicate secondary structure.

The refined prediction for the family of collagenases (hemorrhagic metalloproteases) was built from only 7 sequences spanning only 74 PAM units. This alignment itself is paltry; earlier predictions made by the E.T.H. method has been built from multiple alignments containing more sequences with higher overall sequence divergence. Three other predictions were made, one from the Heidelberg neural network kindly provided by Rost and Sander, a "consensus Chou–Fasman" prediction where a standard Chou–Fasman heuristic was assigned to each sequence in the multiple alignment individually, and the results averaged over the alignment, and a "consensus GOR" assignment similar to that performed by Kirschner and his coworkers for tryptophan synthase, where a version of the standard GOR heuristic was assigned to each sequence individually and a consensus prediction obtained by averaging. Thus, for the first time, four popular methods for predicting secondary structure were placed head-to-head in a bona fide prediction.

The predictions were evaluated using structural information communicated privately by Edgar Meyer a short time ago (Fig. 12). Data evaluating the relative merits of the different predictions are presented in Table 5. By every method of evaluation, the E.T.H. method performed the best. In particular, it was best in the classical three-state per-residue score, surpassing all classical methods. It is important to note that the classical predictions were, in all cases, consensus predictions made for the entire alignment.

However, the three state per residue scores, by aggregating errors of different types, obscure the important differences in the four predictions.

TABLE 5. SCORING THE SECONDARY STRUCTURE PREDICTIONS FOR THE HEMORRHAGIC METALLOPROTEASES (COLLAGENASES)

	Three state residue score	Total assignments α or β	Correct assignments α or β	Incorrect assignments α or β
ETH	70.4%	131	91	2
Heidelberg	66.5%	119	70	20
GOR	56.2	94	40	33
Chou–Fasman	48.8	122	37	51

To use a secondary structure prediction as the starting point for assembling a tertiary structure model, the predictions of alpha helices and beta strands are important. Further, serious errors are those that mistake alpha helices for beta strands and beta strands for alpha helices, not those that mistake an alpha helix for a coil or a beta strand for a coil. Of the 131 predictions made by the E.T.H. method, *only two mistake an alpha helix for a beta strand or vice versa*. Some 16% of the neural network's predictions are incorrect, and the consensus GOR and consensus Chou–Fasman predictions perform still more poorly. Nor does the superiority of the E.T.H. method arise because it makes fewer predictions of alpha helix and beta strand; indeed the E.T.H. method makes the largest number of assignments (alpha or beta) overall (the neural network makes 10% fewer assignments, while the consensus GOR method makes almost 30% fewer assignments). Such statistics underlie Thornton's comment that "although the residue by residue comparison looks very similar, the possibility of extending prediction to a tertiary fold is much better with the Benner–Gerloff method" (41).

CONCLUSIONS: PROGRESS AT LAST

These results add additional evidence of the power of the E.T.H. method for producing reliable secondary structure predictions from a set of aligned homologous protein sequences. Nevertheless, we continue to avoid making claims that our *method* works better than other methods in the *general* case. The history of chemistry shows how difficult it is to evaluate competing approaches for solving conformational problems. Often, alternative models survive in parallel for years before one becomes favored. We expected this to be the case in structure prediction.

The E.T.H. method as it presently stands contains the following components:

(a) All homologous sequences are found in the database (automated)
(b) An evolutionary tree is constructed (automated)
(c) A multiple alignment is constructed (automated)
(d) Parses are identified to subdivide the sequence (automated)
(e) Surface and interior positions are assigned (automated)
(f) Secondary structure is assigned (computer assisted)
(g) Active sites are identified (computer assisted)
(h) Supersecondary structure is assigned (manual)
(i) Covariation analysis identifies amino acids distant in the sequence that are near in space (computer assisted, in the context of a supersecondary structural model)
(j) A tertiary structural model is built (manual with computer assistance).

Many of these outputs are available to the public free of charge via electronic mail using a server at the address: cbrg@inf.ethz.ch.

What is clear is that, with the new paradigm, the structure prediction problem is no longer stubbornly unyielding, the province of soothsayers needing lavish support from optimistic protein and drug designers. Summarizing its key elements, tertiary structural information is generated before secondary structure is predicted. Prediction tools are based on detailed evolutionary models, making them more powerful than those available simply by averaging classical approaches over a set of aligned homologous sequences. The acquisition of predictive power has been developed by predicting structures for many individual cases, with human involvement in making the predictions and understanding their successes and failures. Objectivity in this process is guaranteed by making bona fide predictions, and challenges from those determining experimental structures are still being sought and the method is sustaining serious efforts to model tertiary structure.

APPENDIX: A CONSENSUS SECONDARY STRUCTURAL MODEL FOR PROTEINS HOMOLOGOUS TO THE ISOPENICILLIN N SYNTHASES

INTRODUCTION

The development of structure prediction methods at the E.T.H. has been facilitated by the publication of worked examples, where a structural model for a family of proteins has been built step-by-step, with the details of the steps explained in some detail (16, 19). These discussions are often highly qualitative, and therefore are in some respects the antithesis of the automated programs sought under the classical paradigm as a first goal, and sought by all paradigms once the underlying scientific issues are resolved. Nevertheless, they offer an insight into the state of thinking about structure prediction from an evolutionary perspective. This is especially true if the protein family examined presents an extremely difficult prediction challenge.

This is the case for the superfamily of proteins homologous to the isopenicillin N synthases. We are indebted to Prof. Ian Scott (Texas A & M) for calling this superfamily to our attention as a potential prediction target. The family has proven to be extremely difficult to model for the reasons outlined below. It has forced us to examine many issues in structure prediction, including those relating to extreme divergence of sequence, domain exchange during divergent evolution, and the impact of liganded cofactor on analyses of secondary structure. The appearance of a crystal structure for one superfamily member in the near future will allow us to determine how well we addressed these issues.

The common feature of the superfamily of proteins is that they all catalyze reactions with the aid of a non-heme iron cofactor. Otherwise, the superfamily demonstrates the problematic nature of the Enzyme Commission Catalog number as a systematic way of identifying enzymes. The superfamily includes flavanone hydroxylases (84) (EC 1.14.11.9), a number of 2-oxoglutarate-dependent (EC 1.14.11) oxidases such as hyoscyamine 6β-hydroxylase (85), deacetoxycephalosporin C synthase (expandase) (86, 87, 88), and a protein encoded by the A2 locus of the anthocyanin biosynthesis pathway in various plants (89). These are represented by the tree in Figure 14.

The divergence in catalytic function in this superfamily presents several interesting challenges for the structure predictor. The cofactors required by the different enzymes can be different. In particular, 2-ketoglutarate is evidently required by all proteins in the family except the isopenicillin N synthases. In other cases, it is not clear what cofactors are involved. For example, ascorbate is required for many (and perhaps all) of these enzymes.

This makes identification of active site positions difficult. These are normally identified as functionalized amino acids conserved across the entire protein family (16, 19, 90). In this superfamily, however, adaptive variation is expected to alter many active site residues.

Further, the subfamilies within the superfamilies are all rather small. Several contain protein sequences that have diverged among themselves only modestly. Thus, few if any of the subfamilies could obviously sustain a reliable secondary structure prediction by themselves. Thus, a consensus structural model must be built as far as possible for the entire superfamily, despite the fact that it is almost certain that secondary and tertiary structure have diverged considerably within the superfamily. Thus, as was the case for the SH3 domain (21, 22, 28, 29, 80), the consensus model that is derived by this process is best used as the starting point for homology modelling for individual proteins within the superfamily rather than a definitive structural model for any specific protein within the superfamily.

This prediction exercise therefore involves development of prediction methods. First, we have relied heavily upon parallel analysis of both the superfamily and the individual subfamilies. This is the first time that this has been done, and the outcome is uncertain.

Next, this prediction relies heavily (for the first time) on a set of automated heuristics for assigning surface and interior residues, parses and secondary structure. The first two of these are available to the public via a server accessible by electronic mail at the address cbrg@inf.ethz.ch. Throughout this discussion, parallel predictions are also made by hand. Thus, the prediction offers the opportunity to compare man and machine directly in a bona fide structure prediction exercise.

Finally, given the special challenges of this superfamily, we have relied heavily on parsing heuristics that identify breaks in secondary structure

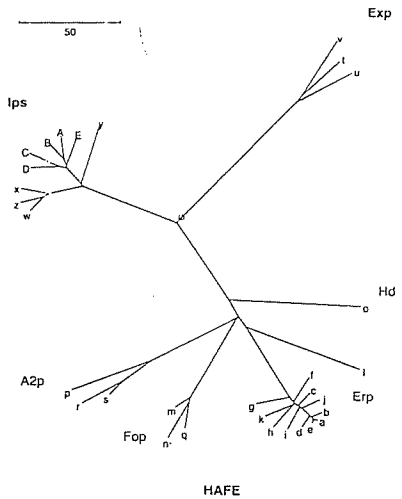


FIG. 14. Evolutionary tree for IPNS superfamily. Letters indicate the proteins below:

The HAFE family

Subfamily Erp

a — (SwissProt P05116) ethylene biosynthesis protein pTom13, tomato, *Lycopersicon esculentum*.

b — (mipsx M90294) ethylene forming enzyme, *Petunia hybrida* (cultivar pink flash).

c — (mipsx M97961) ACC oxidase homologue, kiwi fruit, *Actinidia deliciosa*.

d — (SwissProt P07920) ethylene related protein (GTOMA), tomato, *Lycopersicon esculentum*.

e — (SwissProt P24157) ethylene forming enzyme (EFE), tomato, *Lycopersicon esculentum*.

f — (SwissProt P19464) ripening related protein, avacado, *Persea americana*.

g — (EMBL Z11750) ethylene forming enzyme, *Brassica juncea* (India mustard).

h — (EMBL X66719) ethylene forming enzyme, *Arabidopsis thaliana*.

i — (mipsx M81794) ripening related protein, apple, *Malus sylvestris*.

j — (mipsx M98357) ACC oxidase, pea, *Pisum sativum* (strain Alaska).

k — (mipsx M62380) senescence related protein, carnation, *Dianthus caryophyllus*.

l — (SwissProt P10967) ethylene responsive protein E8, tomato, *Lycopersicon esculentum*.

Subfamily Fop

m — (EMBL X60512) flavanone 3-beta-hydroxylase, *Petunia hybrida*.

n — (SwissProt P28038) flavanone 3-dioxygenase, *Hordeum vulgare* (barley).

q — (Cathie Martin, pers. commun.) inc gene product, *A. majus*.

Subfamily Hd

o — (SwissProt P24397) hyoscyamine 6-dioxygenase, *Hyoscyamus niger*.

Subfamily A2p

p — (EMBO J. 9:3051–3057 (1990)) A2 gene product, maize, *Zea mays*.

(continued next page)

by searching for particular strings of amino acids in individual sequences. This is necessary because gap parses, although frequent throughout the superfamily, are generally poorly anchored. It is therefore difficult to know exactly where to place these gaps. As noted at specific instances throughout the prediction, alternative placement of parses leads to alternative secondary structure predictions in some cases.

Thus, this prediction exercise achieves the goal of being "risky". Of the various families that we have examined as a result of correspondence with many colleagues around the world, this has been the most challenging.

A key problem in this analysis is keeping track of the sequence numbers. We have used the following convention. A number without a prefix designates a position in the master alignment shown in Figure 15. Many alignment position numbers correspond to a specific residue number in the target protein, the isopenicillin N synthase whose structure is presumably soon to emerge. These residues are designated by numerals carrying the prefix "t" (e.g., t034). Separate multiple alignments were constructed for three families of proteins, the expandase family (referred to as the Exp family, with alignment position numbers preceded by the prefix "h"), the isopenicillin N synthase family (referred to as the Ips family, with alignment position numbers preceded by the prefix "g"), and a family of protein (referred to as the HAFE family, with alignment position numbers preceded by the prefix "f") that contains four subfamilies of proteins: (a) the ethylene biosynthesis related protein family (Erp) which contains 11 enzymes (Ebs) and an ethylene responsive unit, (b) a subfamily containing flavanone 3-beta-hydroxylase and flavanone 3-dioxygenases (Fop), (c) a subfamily containing a single sequence for a hyoscyamine 6-dioxygenase (Hd), and

FIG. 14 (cont'd).

r — (Andy Prescott Anton Gerats, pers. commun.) petunia A2 homologue, *Petunia hybrida*.

s — (Cathie Martin, pers. commun.) anthocyanidin synthase, *Candi candi*.

The Exp family

t — (SwissProt P18548) expandase (DAOCS), *Streptomyces clavuligerus*.

u — (SwissProt P11935) expandase (DAOCS/DACS multienzyme complex), *Cephalosporium acremonium*.

v — (J. Bact. 173:398–400 (1991)) DACS, *Streptomyces sp.*

The Ips family

w — (Swiss Prot P05326) IPNS, *Aspergillus nidulans*.

x — (Swiss Prot P05189) IPNS, *Cephalosporium acremonium*.

y — (Swiss Prot P16020) IPNS, *Flavobacterium sp.* (Strain SC 12.154).

z — (Swiss Prot P08703) IPNS, *Penicillium chrysogenum*.

A — (Swiss Prot P10621) IPNS, *Streptomyces clavuligerus*.

B — (Swiss Prot P18286) IPNS, *Streptomyces jumonjinensis*.

C — (Swiss Prot P12438) IPNS, *Streptomyces lipmanii*.

D — (Ann. Rev. Microbiol. 46:462–495 (1992)) IPNS, *Streptomyces griseus*.

E — (SwissProt P27744) IPNS, *Nocardia lactamdurans*.

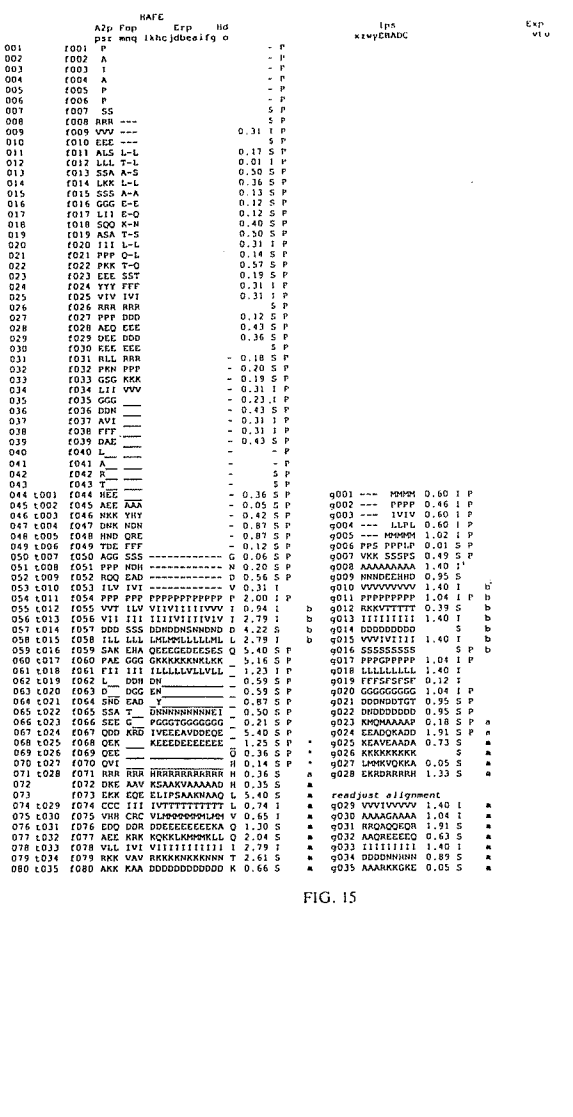


FIG. 15

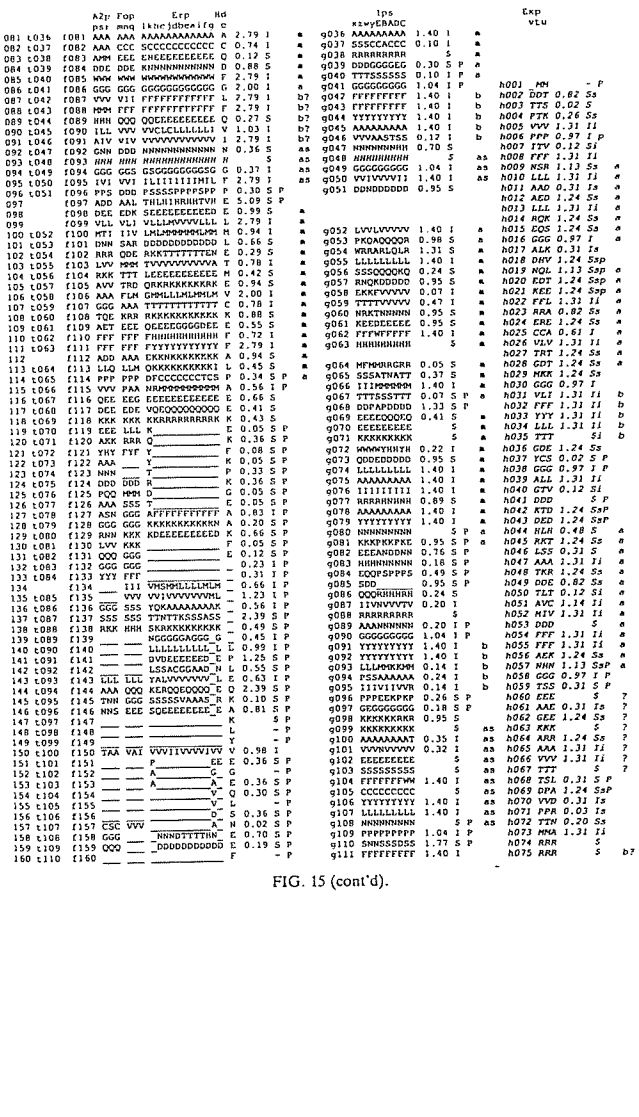


FIG. 15 (cont'd).

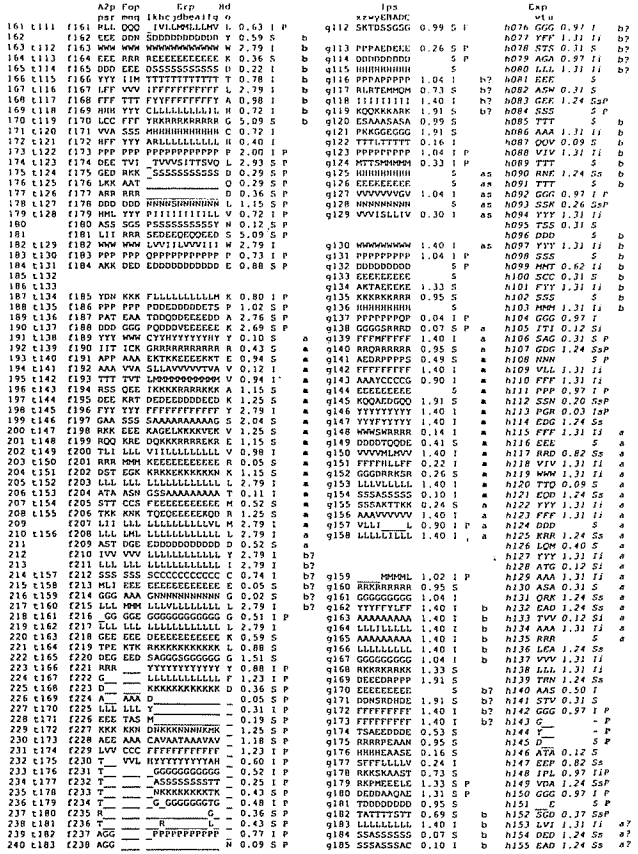


FIG. 15 (cont'd).

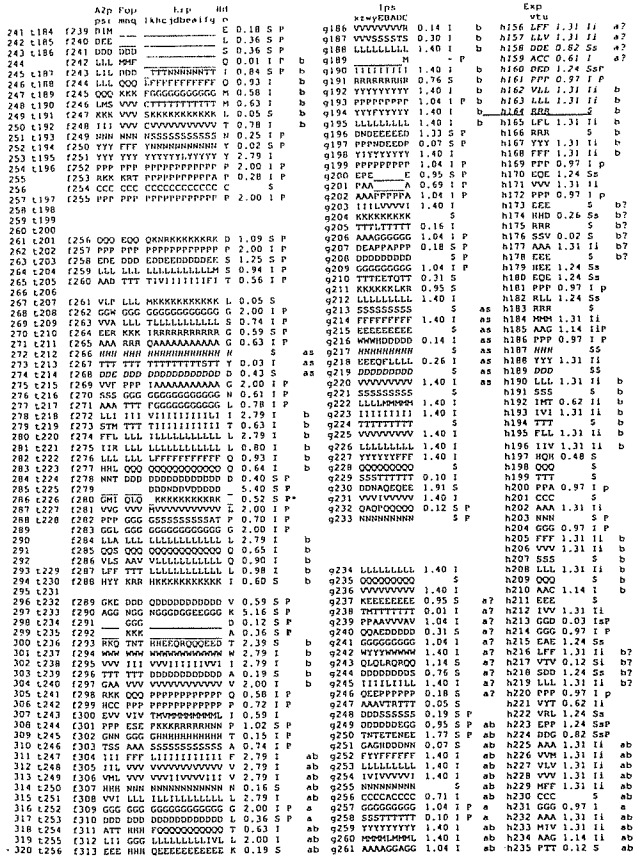


FIG. 15 (cont'd).

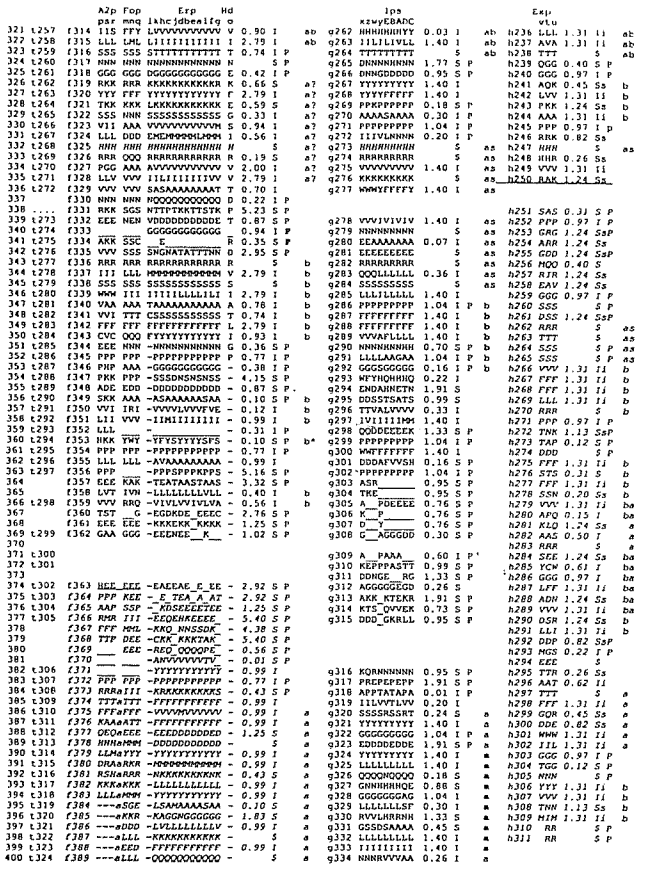


FIG. 15 (cont'd).

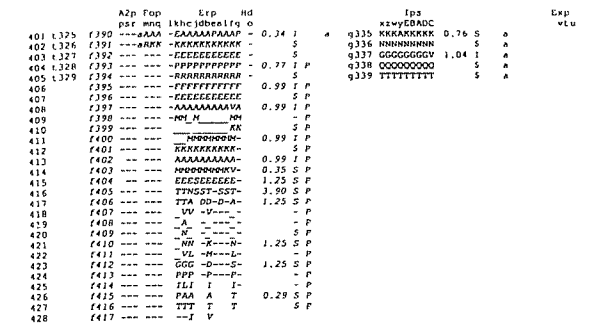


FIG. 15. Multiple alignment of the superfamily containing isopenicillin N synthases (the Ips family), the expandases (the Exp family), and the HAFE family composed of four subfamilies of proteins: (a) the ethylene biosynthesis related protein subfamily (Erp) which contains 11 enzymes (Ebs) and an ethylene responsive unit, (b) a subfamily containing flavanone 3-beta-hydroxylase and flavanone 3-dioxygenases (Fop), (c) a subfamily containing a single sequence for a hyoscyamine 6-dioxygenase (Hd), and (d) a subfamily containing a series of proteins homologous to anthocyanidin synthase (A2p). Boldface predictions indicate the core of secondary structural elements. *Indicates positions where the adjacent secondary structural element could be extended if the multiple alignment were readjusted. Underscoring (_____) indicates a deletion. A dash (–) indicates a segment containing amino acids where the sequence similarity is too low for DARWIN to make an adequate multiple alignment. A dollar sign ($) indicates a functionalized residue conserved across a subalignment. When the subalignment has sufficient evolutionary width, this is an indicator of an active site assignment. I indicates an interior position, assigned by automatic program with a numerical value corresponding roughly to the accuracy of the assignment. S indicates a surface position, assigned by automatic program, with a numerical value corresponding roughly to the accuracy of the assignment. These numerical values are not scaled and have no precise theoretical interpretation. For the purpose of constructing Schiffer-Edmundson helical wheels, the values are presented to one decimal place. Five hundredths are rounded up. In the Exp family, second interior and surface assignments in lower case letters are made by hand. Italicized regions of the alignment are segments that do not align with the Ips family. Italicized single residues indicate active site ligands to iron.

a — amphiphilic alpha helix, b — beta strand, as — active site region, ab — internal alpha helix preferred over a beta strand. The target sequence is B.

(d) a subfamily containing a series of proteins homologous to anthocyanidin synthase (A2p).

A multiple alignment for each family was produced using the DARWIN operating system (47), and the three multiple alignments joined manually to form the master alignment.

Abbreviations used throughout are: APC, all proteins conserved, indicating a residue that is conserved at the indicated position in all proteins; CM1, conserved in all proteins except for one.

PARSING

Identifying breaks in secondary structure has been a key step in most bona fide secondary structure predictions, both made in Zurich and elsewhere. For proteins divergently evolving under a consistent set of functional constraints, several relatively simple parsing heuristics can be applied (16, 19). With the IPNS superfamily, the functional constraints are not consistent. Much of the multiple alignment is poor, and it is clear that simple analysis of indels will not provide reliable parsing. A number of alternative parsing heuristics have therefore been explored in proteins with known structure and in several bona fide predictions. These are used collectively here.

The simplest approach to analyzing parsing begins by examining subfamilies of the alignment that have divergently evolved under a consistent set of functional constraints. These subfamilies have, of course, separately less divergence overall. The three families in the superfamily were considered separately. The strongest parsing elements are considered below. Weaker possible parsing elements are discussed as secondary structure assignments are presented.

The Expandase Family (the Exp Family)

This family contains only three sequences (t, u and v), with an overall PAM distance of 64 PAM units. The three proteins are essentially equally divergent from a common ancestor (the pairwise PAM distances are 58, 64, and 64). The multiple alignment is therefore highly reliable. However, it is extremely difficult to make a reliable prediction from so few proteins with so little overall sequence divergence. There are only two indels, both in the region h142–151 of the subfamily alignment. The parse established by these indels is confirmed by a PDGG tetrapeptide parse in the segment aligned with the gap. Prior to this gap, the expandases do not have sufficient sequence similarity with the rest of the superfamily for DARWIN to recognize them as homologs. Thus, it is appropriate to regard the h142–151 segment not only as a break between secondary structural units, but also presumably a joining of two independently evolving domains.

To obtain additional parses, a variety of weaker "secondary" parsing heuristics were used. These find parses at subfamily alignment positions h169 (APC P), h172 (APC P), h181 (APC P), h185–186 (GP dipeptide parse), h213–214 (GG dipeptide parse), h220 (APC P), h223–224 (PG dipeptide parse), h239–240 (GG dipeptide parse), h245 (APC P), h251–253 (SPG tripeptide parse), h259–261 (GSS tripeptide parse), h271–274 (PN and PD dipeptide parses, and PxPD tripeptide parse), h292–293 (PS dipeptide parse), and h303–305 (GGN tripeptide parse). Due to the small amount of evolutionary divergence within this subfamily, the APC P parses are especially unreliable.

At the end of the sequence, DARWIN ceases to align the Exp family with the other proteins. When attempting to align the Exp family with both the HAFE and Ips families, DARWIN terminates the alignment at h271. When attempting to align the Exp family with the Ips family alone, DARWIN terminates the alignment at positions h250. In both cases, the alignment is terminated at a position where the Exp family has a parse. It is worth noting that the expandases require a 2-oxoglutarate cofactor, in common with most other subfamilies of the superfamily, but not in common with the IPN synthases.

The Isopenicillin N Synthases (the Ips Family)

The isopenicillin N synthase subfamily contains 9 sequences (w, x, y, z, A, B, C, D and E) with a maximum PAM distance of ca. 60. Thus, although the subfamily contains more proteins than the expandase subfamily, the evolutionary breadth is less than that seen with the expandases. Indeed, the overall alignment is less than that used in the bona fide prediction of the secondary structure of the collagenase (hemorrhagic metalloprotease) family (see above). The subfamily is of special importance, however, as it contains the protein B (the IPN synthase from *Streptomyces jumonjinensis*) whose crystal structure presumably will be solved.

In the IPN synthase subfamily, gap parses are found in the subfamily alignment positions g085, g157–159, g189, g200–201, and g303 to ca. g313. The third corresponds to the point in the superfamily where DARWIN recognizes the expandases as homologs. The last is 10 positions after the point where the expandases cease to be alignable by DARWIN. Within the IPN synthase subfamily, secondary parses are found at positions g011 (APC P), g016–017 (SP dipeptide), g020–024 (concatenated dipeptide and tripeptide parses, weak), g039–041 (GSG tripeptide), g067–068 (SP dipeptide), g080–084 (PDNP tetrapeptide), g089–090 (NG dipeptide), g096–097 (PG dipeptide), g108–110 (NPS tripeptide), g112–114 (SPD tripeptide), g123–124 (PS dipeptide), g131–132 (PD dipeptide), g137–138 (PG dipeptide), g179–181 (PDD tripeptide), g193 (APC P), g196–199 (DPT P tripeptide parse), g206–208 (GPD dipeptide), g232–233 (PN dipeptide), g248–250 (DDN tripeptide), g269–272 (PSP, APC P, and PS parses), g286 (APC P), g298–299 (DP dipeptide), g301–303 (DPS tripeptide), and g316–318 (NPP tripeptide).

The Bottom of the Tree (the HAFE Family)

We can repeat this process now with the bottom of the tree (Fig. 14), which includes four subfamilies of proteins, designated Hd, A2p, Fop, and Erp. The family is termed the HAFE family overall, from the first initial of the names of the constituent subfamilies. Overall, ca. 160 PAM units of

sequence divergence have taken place within this family. This family would form a nice target for structure prediction, if its subfamilies did not perform quite different catalytic roles.

Here as before, some of these families contain additional protein units not found elsewhere in the superfamily, presumably reflecting different catalytic roles. The Fop and A2p protein subfamilies contain two modules that proceed the superfamily bulk. The first is from positions f008–022, and is missing in the flavone 3-dioxygenase (Fop subfamily). The second is found in

TABLE 6. PARSES IN THE HAFE FAMILY*

Alignment	Target	comments
A. Parses in the Erp Subfamily		
062–064	**t019–021**	confirmed in A2p
119–126	**t070–077**	
139–146	**t089–096**	single del, confirmed by del in A2p
151–161	**t101–111**	confirmed by del A2p, Hd
173–174	**t123–124**	PNPPS pentapeptide parse
225–227	**t169–171**	single insert, confirmed by indel A2p
232–244	**t176–187**	NGP tripeptide parse, indel in A2p
253–261	t196–202	PPCPxP parse
285	**t225**	confirmed in A2p
344	**t275**	confirmed in A2p
354		ethylene responsive protein drops out; NPGS parse
370–383	**t297–305**	A2p and Fop subfamilies drop out at 377/303
411–413	**after 329**	ethylene forming enz *B. jun.* (sequence 1) drops out at 413
B. Parses in the A2p Subfamily		
040–043	**t < 0**	not confirmed
062–063	**t019–020**	confirmed in Erp
217	**t161**	largely conserved G in Erp and Fop
223–225	**t167–169**	not confirmed; may need adjustment
231–237	**t175–181**	Sequence p contains a remarkable TTTTT
355–357	t286–288	PPP tripeptide parse
C. Parses in the Fop Subfamily		
023		sequence n begins
065–066	**t022–023**	PG in Erp
370	**t297**	indel in Erp
D. Parses in the Hd Subfamily		
156–157	t106–107	Dipeptide SN
179–180	gap	Dipeptide NS
254–255	t196–gap	Dipeptide PP
257–264	t197–204	Pentapeptide PDPSS
268–271	t208–211	Tetrapeptide GSGG
275–276	t215–216	Dipeptide GN
329–330	t265–266	Dipeptide GS
335–336	t271–272	Dipeptide DP
351–352	t285–286	Dipeptide GP

*Deletion parses are in bold.

all proteins in the Fop and A2p subfamilies, and includes positions f023–034, extending to position f039 in the A2p subfamily. The PP dipeptide parse at positions f021–022 breaks the two segments reliably. The PG dipeptide parse f032–033 terminates the second segment. There is no reason to assign these anything other than coils.

Likewise, DARWIN fails to find a significant alignment following f363 within this family. The A2p and Fop subfamilies remain aligned 19 residues further, with each subfamily carrying a C-terminal tail of 20–40 residues that aligns with nothing else. The conformation of these tails is unassigned. However, the segment where the A2p and Fop subfamilies align can be assigned a helical structure (see below). This corresponds to a helix tentatively assigned to the unaligned segment in the Erp subfamily. This suggests that these two regions are homologous even though homology cannot be detected by standard sequence comparison.

To parse the HAFE family, we identify parsing units in separate subfamilies, confirmed in other subfamilies, and combined into a parsing scheme in the family, and then added to parsing schemes from another part of the superfamily. We start with the ethylene biosynthesis proteins. The parses are summarized in Table 6.

Some segments of the multiple alignment of the HAFE family are highly gapped, making parses difficult to identify even thought the HAFE family has far narrower overall divergence than the superfamily as a whole. An excellent example of this is found at alignment positions f120–160. This segment is discussed in the next section in greater detail.

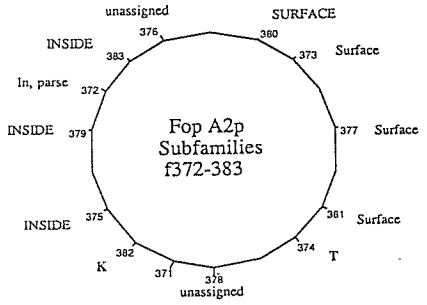
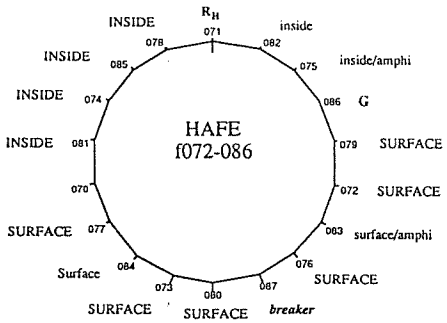
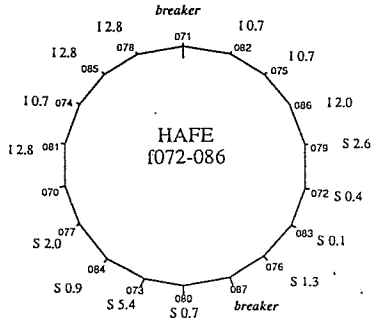


FIG. 15.





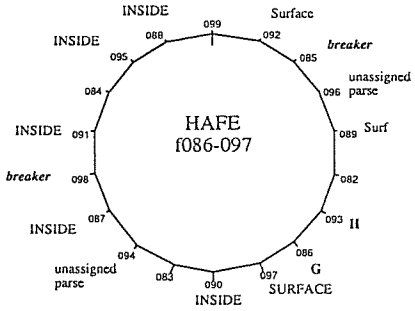
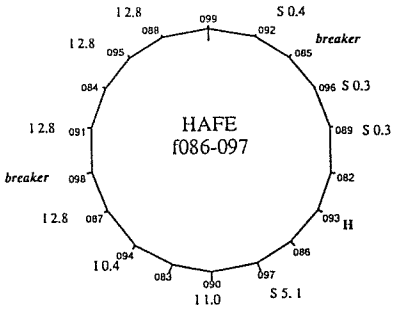
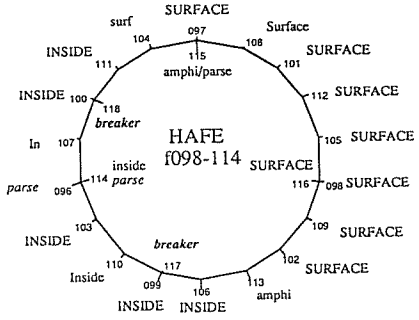


FIG. 16 (cont'd).





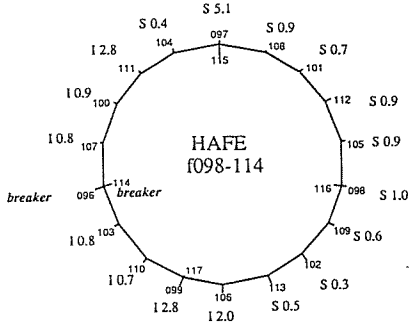
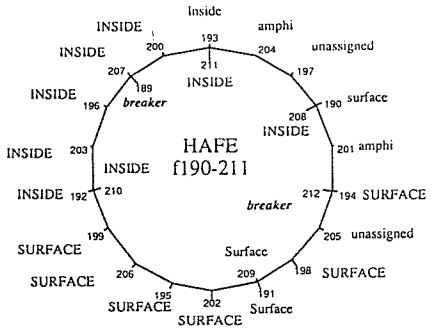
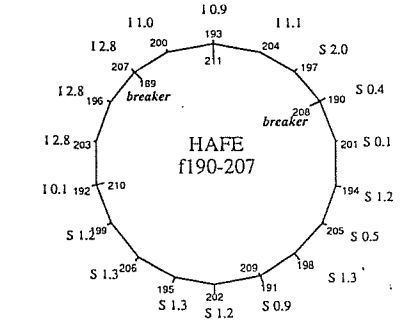


FIG. 16 (cont'd).





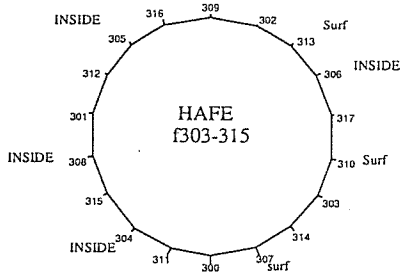
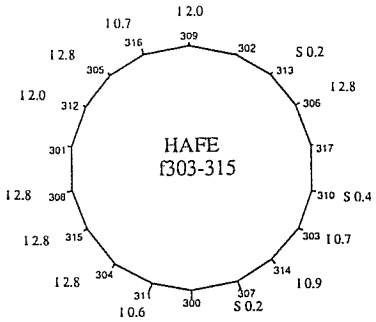
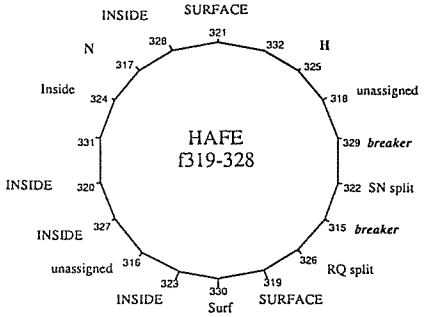


FIG. 16 (cont'd).





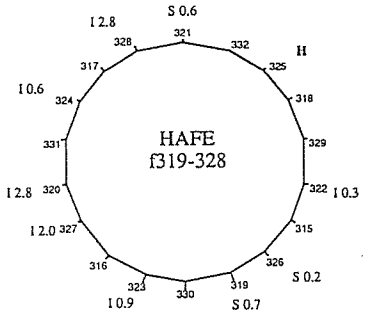
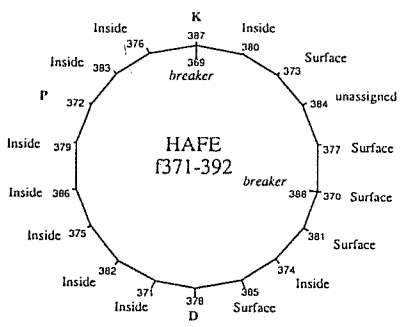
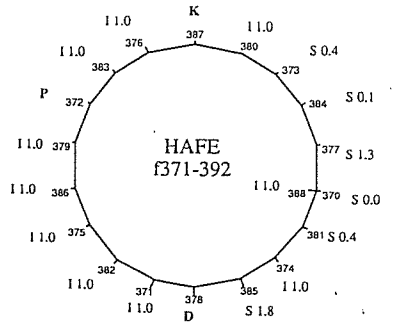


FIG. 16 (cont'd).





SURFACE

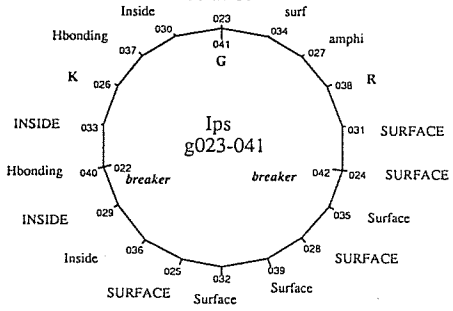
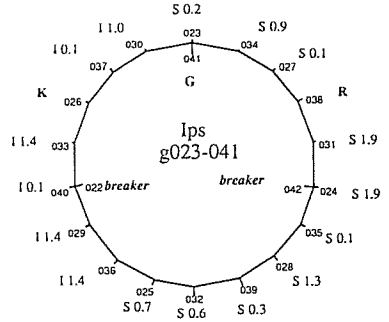
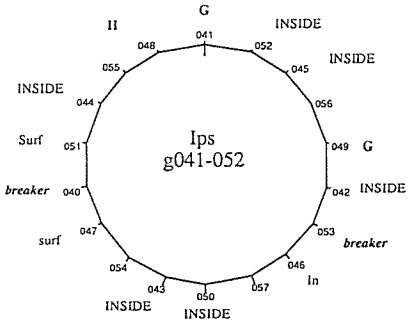


FIG. 16 (cont'd).





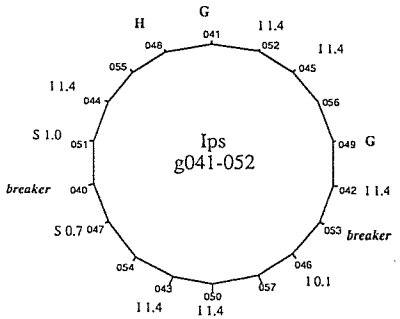


FIG. 16 (cont'd).

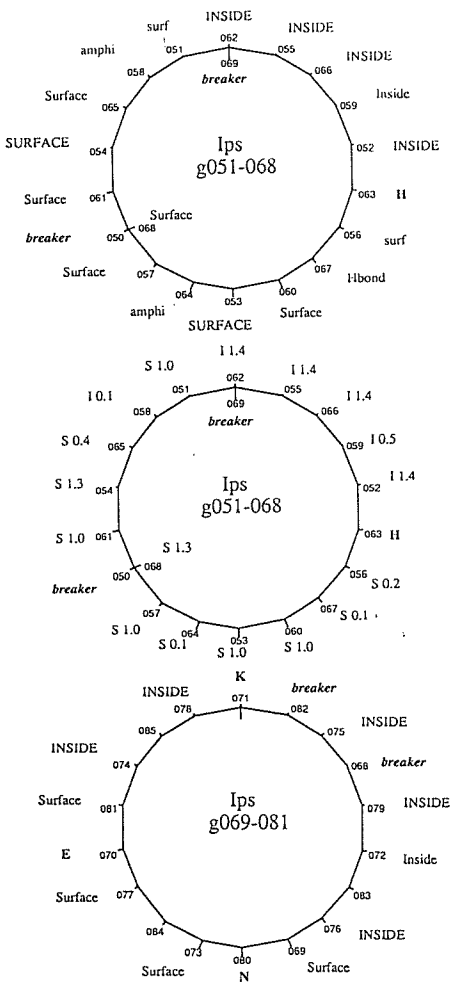


FIG. 16 (cont'd).

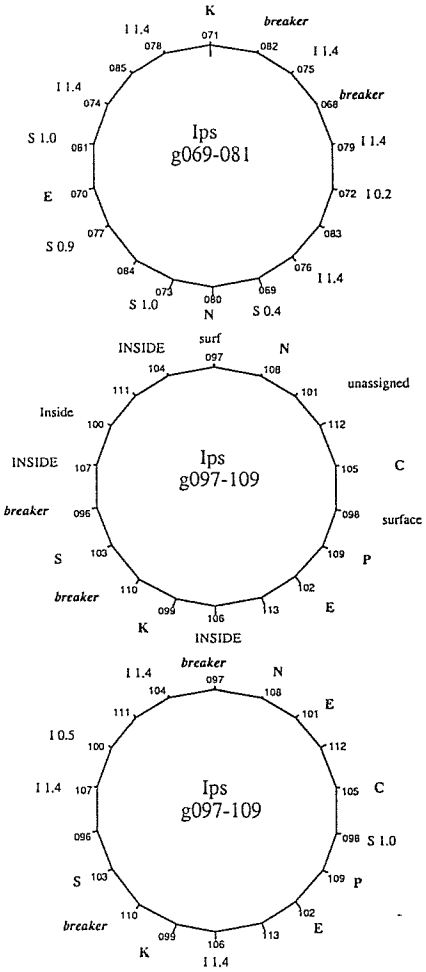


FIG. 16 (cont'd).

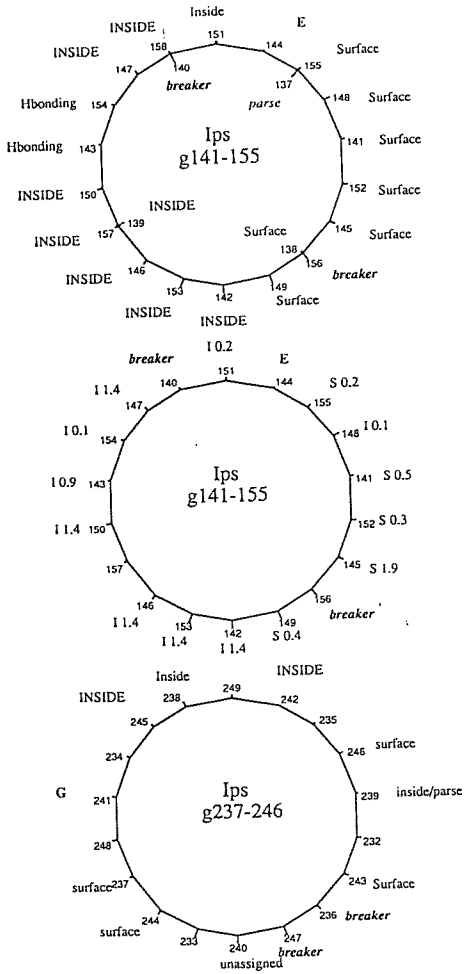


FIG. 16 (cont'd).

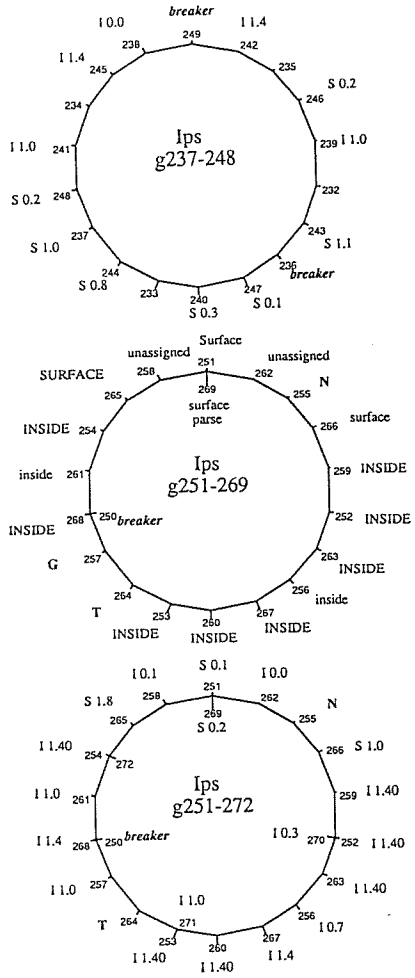


FIG. 16 (cont'd).

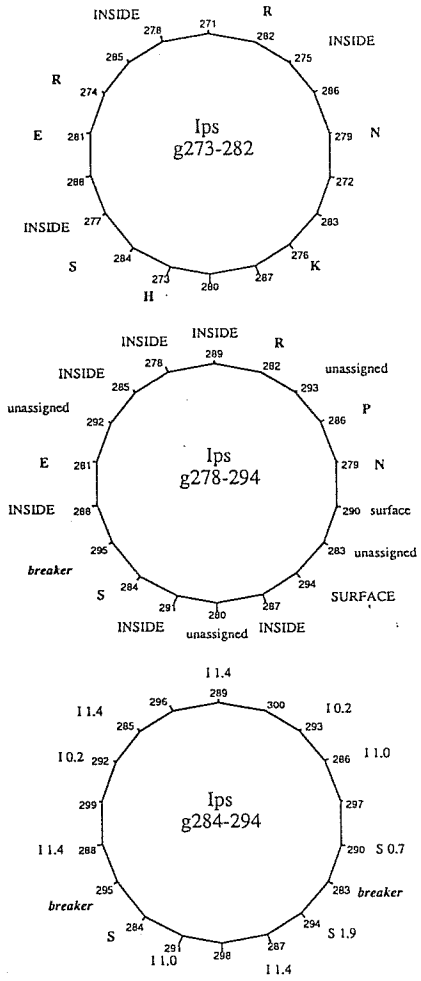


FIG. 16 (cont'd).

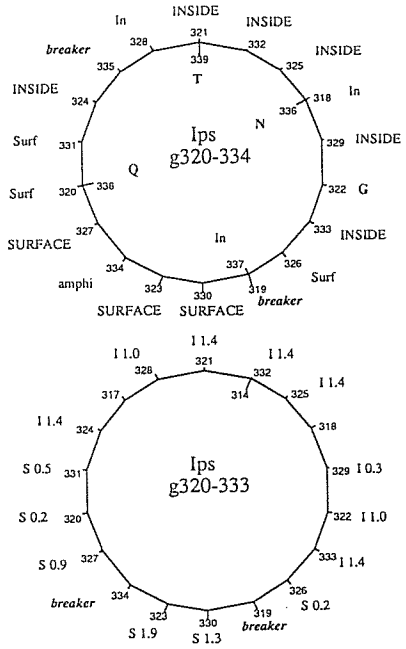


FIG. 16. Helical wheels used to assign secondary structure in the superfamily aligned in Figure 15. Pairs of wheels have surface and interior assignments made by hand (top) and made by computer (bottom). Positions that break 3.6 residue amphiphilicty are marked "*breaker*". Bold upper case letters indicate conserved amino acids (one letter code).

SECONDARY STRUCTURE ASSIGNMENT

Secondary structural assignments were then generated for each of the three families. As before (16, 19, 22–24, 76), this was done first by assigning surface, interior, and active site positions in the alignment from patterns of conservation and variation within the homologous proteins. These assignments are collected in the master alignment. Where the surface (S) and interior (I) assignments are generated by an automated heuristic, they are associated with a numerical value. Larger values generally indicate a more reliable surface or interior assignment, while smaller values generally

indicate a less reliable surface or interior prediction; the values are not scaled in any way, however, and have no direct theoretical meaning.

Patterns in the surface and interior assignments are then used in the first effort to assign secondary structures separately for each family (or, in some cases, subfamily). These are presented in the master alignment. In the final round, efforts are made to reconcile secondary structure predictions between the subfamilies. These are collected in Table 7.

The Expandase Family (the Exp family)

Because there are only three sequences in the Exp family, and because the overall divergence of sequence is quite low, surface and interior assignments are more inaccurate. Perhaps indicative of this is the fact that the automated assignments and the hand-derived assignments (Fig. 15) often do not correspond. This implies that the secondary structure assignments for the Exp family are rather unreliable, especially when compared with those made for other families within the superfamily. Thus, these secondary structure predictions must themselves be viewed casually. They are useful only in comparison with those made for other family members.

Parse h006 (APC P)
Segment Exp h007–017
A helix is assigned in positions h009–016
Parse h018–021 (4 consecutive surfaces)
Segment Exp h022–036
A helix is assigned in positions h019–028; a beta strand is assigned in positions h031–035.
Parse h037–h043 (SG and DDD)
Segment Exp h044–057
A helix is assigned in positions h044–057.
Parse h057–059 (NGS)
Segment Exp h060–067
The assignment of secondary structure to this segment is uncertain.
Parse h068–069
Segment Exp h070–h082
A beta is tentatively assigned to this segment.
Parse h083–084 (GS)
Segment Exp h085–091
A beta is tentatively assigned to this segment.
Parse h092–093 (GS)
Segment Exp h094–105
A beta is tentatively assigned to this segment.
Parse h106–108 (GGN)
Segment Exp h109–h110

Hydrophobic anchor for coil
Parse h111–113 (PSP)
Segment Exp h114–141
A helix is assigned in this segment.
Parse h142–152 (gap GD)
Segment Exp h153–159
A short helix might be weakly assigned in this position.
Parse h160–161 (DP),
Segment Exp h162–171
A beta strand is tentatively assigned to this segment.
Parse h172 (APC P),
Segment Exp h173–180 (258–265, t198–205)
A beta strand is weakly assigned to this segment. The alignment of the Exp family with the rest of the superfamily begins at this point.
Parse h181 (APC P)
Segment Exp h182–184 (267–269, t207–209)
This segment is most probably a coil.
Parse h185–186 (GP dipeptide parse)
Segment Exp h187–212 (272–297, t212–233)
Two possible assignments apply to this segment. First, the dipeptide parses at positions h203–204 (NG) and the scattered prolines at position h200 can be viewed as definitive. Then, beta strands would be assigned to positions h190–196 and h205–210. Alternatively, an internal helix can extend throughout the entire segment. The parses at positions h200–204 make the former option stronger.
Parse h213–214 (GG dipeptide parse)
Segment Exp h215–219 (300–304, t236–240)
A beta strand is assigned to this segment.
Parse h220 (APC P),
Segment Exp h221–222 (306–307, t242–243)
This segment is most probably a coil.
Parse h223–224 (PG dipeptide parse)
Segment Exp h225–238 (310–323, t246–259)
This segment contains either an internal alpha helix or two consecutive beta strands.
Parse h239–240 (GG dipeptide parse)
Segment Exp h241–244 (326–329, t262–265)
A beta strand is assigned to this segment.
Parse h245 (APC P),
Segment Exp h246–250 (331–335, t267–271)
This segment includes several active site positions, and is assigned "as". The alignment of the Exp family with the rest of the superfamily ceases at this point.

Parse h251–253 (SPG tripeptide parse)
Segment Exp h254–258
This segment is most probably a coil.
Parse h259–261 (GSS tripeptide parse)
Segment Exp h262–270
Active site/beta. A region where the expandases obtain catalytic groups for catalytic function unique to the Exp family.
Parse h271–274 (PN and PD dipeptide parses, and PxPD tripeptide parse)
Segment Exp h275–291
Both an alpha helix or two beta strands are possible in this region.
Parse h292–293 (PS dipeptide parse)
Segment Exp h294–302
A helix is possible in this segment.
Parse h303–305 (GGN tripeptide parse)
Segment Exp h305–309
A beta strand is assigned to this segment.

The Isopenicillin N Synthases (the Ips Family)

The secondary structure predictions within the Ips family are expected to be more reliable than with the Exp family. Even with the low overall sequence divergence (ca. 64 PAM) and only 9 family members, however, some secondary structural elements can be assigned with high reliability.

Parse g001–007 (044–050, t001–007)
Segment Ips g008–015 (051–058, t008–015)
This segment is divided by an APC P parse at position g011. However, parses of this type are not particularly strong given the low overall divergence of the Ips family. Here, a beta strand is a canonical assignment from positions g012–015, with one position extensions on each end possible. As an alternative assignment, an alpha helix for positions g008–019 (12 residues) was considered. The wheel is not particularly convincing; parses occur at positions g011, g016 and g017. All of these are strengthened by parses in the corresponding segments in the Erp subfamily. If all of the parses are accepted, a helix cannot be assigned. If the g011 parse is ignored, the helix length is only 8 positions long, and the 3.6 residue periodicity not statistically significant. Therefore, the beta assignment is made.

Parse g016–017 (SP dipeptide)
Segment Ips g018–019 (061–062, t018–019)
These two residues are both assigned to the inside. This is assigned as a coil, especially in light of the assignment in the Erp subfamily (see below). A two residue beta strand cannot be excluded, however, and remains a candidate to be considered in light of secondary structure assignments made for other families.

Parse g020–024 (overlapping tripeptide DPD GDD parses)
Segment Ips g025–038 (068–083; t025–038)
An alpha helix can be assigned to positions g023–041 (19 residues), both using surface and interior assignments made by hand and those made by the automated computer tools. To avoid all parses, the helix would be assigned to positions g025–038 (14 positions). These define the minimum and maximum lengths of the helix.

Parse g039–041 (GSG tripeptide)
Segment Ips g042–066 (087–115; t042–066)
This segment is quite long, and almost certainly contains more than one standard secondary structural unit. Internal parsing units are possible at positions g049, g051, and g053. The situation is complicated by the identification of g048 as an active site position (see below). Constraints on divergence imposed by catalytic demands in the active site often obscure patterns that might indicate specific secondary structures. Canonically, the assignments are for a beta strand (g042–046), an active site segment with undefined secondary structure (g047–050), and an alpha helix (g051–068). Excluding parsing elements, a minimal helix covers positions g054–066 (13 residues). Notably, the helix is followed by another predicted helix; however, the amphiphilic pattern is not the same in the two helices, and therefore they are assigned separately.

Parse g067–068 (SP dipeptide)
Segment Ips g069–079 (118–128; t069–079)
A helical wheel with good 3.6 residue periodicity can be assigned to the segment g069–081. Excluding parsing elements makes a minimal helix covering positions g069–079.

Parse g080–085 (NPDNP pentapeptide, gap)
Segment Ips g086–088 (136–138; t086–088)
This segment is assigned as a coil.
Parse g089–090 (NG dipeptide)
Segment Ips g091–095 (140–144; t090–094)
This segment of inside positions is assigned as a beta strand.
Parse g096–097 (PG dipeptide)
Segment Ips g098–107 (147–156; t097–106)
Six of the 10 positions in this segment are APC, including an APC K, and APC E, and an APC S. When the following parsed region is included, a string of 5 consecutive APC residues are found. This string is one of two APC pentapeptides in the Ips family, the longest APC strings in the family. The implication is that this segment is near the active site. However, the segment is not highly conserved in other branches of the evolutionary tree. This implies that the this segment performs a function in the Ips family not found in other members of the superfamily, most likely substrate binding. Photoaffinity labelling studies (91) support the

notion that this segment (in particular, Cys-g105, corresponding to t104) is near the substrate binding site.

Canonically, this segment is assigned as a beta strand. An attempt to construct a helical wheel shows no convincing 3.6 residue periodicity. However, placing this segment near the active site complicates assignment of a secondary structure, as conservation of residues for catalytic function often obscure patterns that might indicate a particular secondary structure. Thus, an alternative assignment is an active site coil.

Parse g108–110 (NPS tripeptide)

Segment Ips g111 (160; t110)

This internal residue stands alone, and presumably serves as a hydrophobic anchor for a coil in the flanking regions.

Parse g112–114 (SPD tripeptide)

Segment Ips g115–122 (165–172; t114–121)

Canonically, this segment is assigned as a coil. There is no 3.6 residue periodicity that would indicate even a single turn of a helix. There is, however the possibility of beta strand periodicity (g116–119, 4 residues) presence of a weak dipeptide parse (SG) at positions g120–121 strengthens the coil assignment.

Parse g123–124 (PS dipeptide)

Segment Ips g125–130 (175–182; t124–129)

This segment also contains a large number of APC positions (7 out of 9 a positions g125–133), 4 of them functionalized, again indicating a position near the active site. Again, the conserved segments are not found in other branches of the evolutionary tree. This implies that the role of this segment in the active site of the isopenicillin N synthases is not shared in the other branches of the evolutionary tree.

Parse g131–132 (PD dipeptide)

Segment Ips g133–136 (185–188; t132–135)

This segment is assigned as a surface coil.

Parse g137–138 (PG dipeptide)

Segment Ips g139–156 (191–208; t138–155)

A good alpha helix can be assigned to positions g141–155. In the automated assignment, position g148 breaks the amphiphilicity; however, this contains an example of a Trp–Arg codon-driven substitution, a surface indicator (see above) not yet incorporated into the automated surface prediction heuristics. Interestingly, the helix can be extended by a single turn towards the amino terminus in some members of this family. Further, rearranging the multiple alignment can yield either an extra turn of a helix at the carboxyl terminal end. Thus the helix is assigned maximally to positions g138–158, minimally to positions g143–154, and preferred to positions g141–155.

Parse g157–159 (209–214; t156–157). The multiple alignment can be readjusted in this region to move the deletion parse.

Segment Ips g160–178 (215–233; t158–176)

This is a long segment that may contain more than one standard secondary structural element. Two beta strands are possible in this region. The first is canonical at positions g162–166. The second, in positions g170–173, is defined only by two interior positions, and is very weak.

There is little likelihood that the first part of this segment forms an internal alpha helix. Efforts to construct a helix in this region identify a possible helical segment from beta strand g173–182. The statistics of the model are poor, with only two interior positions of significance. Thus, a coil assignment for positions g174–181 is preferred.

Parse g179–180 (PDD tripeptide)

Segment Ips g181–188 (236–243; t181–186)

This segment includes a string of 6 interior assignments broken only by a very weak surface assignment at position g184. A beta strand is canonically assigned in this region, with its extent possibly altered by altering the multiple alignment to move the gap, presently at position g189.

Parse g189 (gap)

Segment Ips g190–192 (245–247; t187–189)

A beta strand is canonically assigned in this region, possibly joined to the beta strand preceding or the beta strand following, depending on the adjustment of the multiple alignment (see below).

Parse g193 (APC P)

Segment Ips g194–195 (249–250; t191–192)

A beta strand is canonically assigned in this region, possibly joined to the beta preceding strand, depending on the adjustment of the multiple alignment and the confirmation of the parse at position g193 in other branches of the evolutionary tree. Following this point, the Exp family first becomes alignable by DARWIN.

Parse g196–197 (DP dipeptide)

Segment Ips g198 (253; t195)

Canonically assigned as an anchor for a surface coil.

Parse g199–201

Segment Ips g202–205 (257–260; t197–200)

A coil is preferred in this region, although a beta strand cannot be excluded.

Parse 206–209 (GPDG tetrapeptide)

Segment Ips g209–231 (264–286; t204–226)

There are several regions in the Ips family where secondary structure is extremely difficult to assign. This is the first. The most striking aspect of this segment is the alternating functionalized APC residues at positions g213 (Ser), g215 (Glu), g217 (His), g219 (Asp) and g221 (Ser). Positions g217

and g219 are assigned as ligands to the ferrous ion in the active site. This strongly suggests an extended beta-like structure in this region.

The competing assignment is an internal helix in this segment. This is all but ruled out if both position g217 and position g219 contribute ligands to a ferrous ion. It is possible to organize APC positions g221, g224, and g228 on one face of an alpha helix. However, there is no convincing 3.6 residue periodicity, at least in the first part of this segment.

Parse g232–233 (PN dipeptide)

Segment Ips g234–247 (293–306; t229–242)

An amphiphilic alpha helix is possible between positions g237–246 (10 residues, using hand surface-interior assignments), or g237–248 (12 residues, using automated surface-interior assignments). The helix contains possible parses at positions g239 and g241, and one unassigned position (g240). Thus, the assignment is not strong.

Parse g248–250 (DDN tripeptide)

Segment Ips g251–268 (310–327; t246–263)

This is the second difficult segment to assign. There are two different tactics. First, one can accept the GS dipeptide parse at positions g257–258 and the NG dipeptide parse at positions g265–266 as definitive. This makes a beta assignments canonical for positions g251–256, positions g259–264, and positions 267–268. However, as neither of these dipeptide parses is particularly strong, an internal helix must be considered. Especially important signals favoring an internal helix are (a) the stretch of 13 interior positions, long enough to traverse a globular domain of the size expected for the Ips family, (b) the internal APC G at position g257, often seen in internal helices, (c) the juxtaposition on the helical wheel of surface positions g251 and g265 at the ends of the putative helix, and (d) the placement of the two weakest interior residues (positions g258 and g262) on the "surface" arc of the helix. Thus, the internal helix assignment is preferred, to be evaluated by comparison with secondary structure predictions in other branches of the evolutionary tree.

Parse g269–272 (PSP and PS overlapping parses)

Segment Ips g273–285 (332–346; t268–280)

This segment includes an active site string at positions g273–276. Significantly, following these active site residues, DARWIN is no longer able to align the Exp family. The APC positions following g276 presumably contribute functionality to the active site that is unique to the HAFE and Ips families. This is a region that exemplifies the difficulties of identifying secondary structure clearly in regions where residues are highly conserved, evidently under functional constraints associated with an active site. A helix wheel does not show a 3.6 residue periodicity in the APC functionalized residues. Therefore this segment is assigned simply as an active site (as) region.

Parse g286 (APC P)

Segment Ips g287–289 (348–350; t282–284)

This segment is canonically assigned as a beta strand, perhaps including parts of the adjacent parses. An alpha helix might be assigned for positions g278–294 based on interior and surface assignments made by hand. The automated surface and interior assignments allow a possible helix from positions g284–294. However, these helices must be largely internal, and the wheels are not statistically convincing.

Parse g290–292

Segment Ips g293–297 (356–358; t290–292)

There is no evidence to assign this short segment any standard secondary structure.

Parse g298–299 (DP dipeptide)

Segment Ips g300 (361; t295)

There is no evidence to assign this short segment any standard secondary structure.

Parse g301–318 (various overlapping parses)

Segment Ips g319–321 (385–387; t309–311)

There is no evidence to assign this short segment any standard secondary structure.

Parse g322–323 (GD dipeptide)

Segment Ips g324–339 (390–405; t314–329)

An alpha helix with good 3.6 residue periodicity can be assigned to positions g320–333.

Excluding the parses at the beginning gives a minimal helix covering positions g324–333. Including the APC residues at the end gives a maximal helix covering positions g320–339.

The Bottom of the Tree (the HAFE family)

An entirely independent secondary structure prediction can be made for the Erp, A2p, Fop and Hd subfamilies that together form the HAFE family. The individual parsed segments are discussed below:

Parse f052 (end of sequence with GND, DGP tripeptides)

Segment HAFE f053–058 (053–058; t010–015)

A beta strand in this region may be broken by the APC parse at position f054. This parse is strengthened by the corresponding APC P in the Ips family.

Parse f059–070 (gap) possibly realignable

Segment HAFE f071–095 (071–095; t028–050)

Clean 3.6 residue amphiphilicity indicating an alpha helix is observed from positions f072–086. Adjusting the multiple alignment to eliminate the break at position f070 does not allow the amphiphilic pattern to be

extended to earlier positions, and the interior assignment that breaks the amphiphilic pattern at the end of the helix (f087) is solid. The end of the helical amphiphilicity is also marked by an APC G, a weak parse. The subsegment that follows (f087–095) is difficult to assign, as it is approaching an active site residue, APC H (f093). A 3.6 residue periodicity (see helical wheels) makes a helix possible, if this helix were to include the active site histidine, it might be as long as 12 residues long. However, a beta strand from positions f087–091 is preferred, in part because of the weakness of the surface assignment at position f089. Positions f092–095 are assigned as an active site region.

Parse f096–097 (PD dipeptide)

Segment HAFE f098–113 (098–113; t052–064)

A convincing pattern of 3.6 residue amphiphilicity indicating an alpha helix extends across this entire segment.

Parse f114–115 (PP dipeptide)

Segment HAFE f116–118 (116–118; t067–069)

The consecutive surface residues in this segment indicate a coil.

Parse f119–161 (119–160; t070–110)

This segment must be realigned and analyzed again before a secondary assignment can be made with any reliability.

Segment HAFE f162–172 (162–172; t112–122)

There is a string of 5 consecutive interior residues (f165–169) that are canonically assigned as a beta strand. This is extended by several residues on each end that display alternating patterns of amphiphilicity.

Parse f173–181 (PDG, PNPPS, and gap)

Segment HAFE f182 (182; t129)

This single interior position is assigned as an anchor for a coil.

Parse f183–188 (PD, NP, and PPD strings)

Segment HAFE f189–215 (191–217; t138–160)

A good helix is possible at positions f190–211. A beta strand is possible at positions f210–215 because of three consecutive interior assignments (f210–212) followed by two very weak surface assignments. No clear parse separates the putative helix f190–211 from the putative strand f210–215. The placement of a gap at position f216 is critical to the assignment. This gap is poorly anchored; it may in fact be a sequencing error in one of the A2p subfamily members. Were this gap not at this position, a beta strand would be assigned more strongly to this region. The 3.6 residue periodicity of the previous alpha helix is not extendable into this region.

Parse f216 (Gap)

Segment HAFE f218–220 (220–222; t163–165)

As parsed, this segment is assigned a coil structure.

Parse f221–243 (gap)

Possibly realignable

Segment HAFE f244–248 (246–250; t188–192)

This string of interior positions is canonically assigned a beta conformation. In the A2p and Fop subfamilies, this strand might extend two amino acids towards the amino terminus.

Parse f249–265 (SN, PP, PNPD, PDPSS, DP, and GSGG strings)

With occasional anchors.

Segment HAFE f266–268 (272–274; t212–214)

Active site (see below).

Parse f269–271 (PG and GG dipeptide, overlapping)

Segment HAFE f272–278 (278–284; t218–224)

This is canonically assigned as a core beta strand following an active site segment. The alternative assignment disregards the GG dipeptide parse (f270–271) and constructs an internal alpha helix that includes the active site residues f266 and f268. There is no basis for favoring this alternative, however.

Parse f278–283 (DD and PG dipeptide and gap)

Segment HAFE f284–288 (290–294, t226–231)

This segment of consecutive interior residues is canonically assigned as a beta strand.

Parse f289–293 (DG dipeptide, gap)

Segment HAFE f293–297 (300–304; t236–240)

This segment contains 3 strongly assigned interior positions. The segment can be lengthened by a rearrangement of the multiple alignment, but the tripeptide parse (DNG) in the Fop subfamily (f289–291) would make assignment of a helix difficult. The segment is canonically assigned as a beta strand (f294–297).

Parse f298–299 (PP dipeptide)

Segment HAFE f300 (307; t243)

This single interior position is assigned as an anchor for a coil.

Parse 301–302 (PG dipeptide)

Segment HAFE f303–315 (310–322; t246–258)

The secondary structure for this segment and the segment following are the most difficult in the HAFE family to make. There are two strategies to making an assignment. First, the weak GD dipeptide parse (positions f309–310) can be taken as definitive. In this case, two beta strands are assigned at positions f303–308 and f311–315. Alternatively, the entire segment can be assigned as an internal helix (f303–315, 13 residues). A weak pattern of 3.6 residue periodicity in the assignments strengthens the helix assignment.

In any case, it is certain that this is a core segment inside the globular structure. This implies that if a clear assignment cannot be obtained by

examining the other families within this superfamily, tertiary structural modelling must be attempted with both secondary structure assignments.

Parse f316–318 (SNG tripeptide)

Segment HAFE f319–329 (326–336; t262–272)

This segment includes an active site His (f325, t268). A helix can be assigned to positions f319–328, delivering the His side chain to the surface arc. In the automated output, the amphiphilicity is broken by a SN split (f322) assigned to the inside but appearing on the surface arc. Following this segment, the Exp family ceases to align significantly.

Parse f330–335 (NSNSS pentapeptide; gap)

Segment HAFE f336–343 (343–350, t277–284)

A canonical beta strand, supported by the APC R (f336, t277) and APC S (f338, t279). Following this segment, the Hd subfamily and one protein in the Erp subfamily cease to align significantly.

Parse f344–348 (NPGSDS hexapeptide)

Segment HAFE f350–351 (357–358; t291–292)

This segment would canonically be assigned as a hydrophobic anchor for a coil. However, if the multiple alignment can be adjusted, a short beta strand is assigned. It is worth noting that the Hd subfamily ceases to be aligned by DARWIN starting at this point, implying the end of the common fold within the HAFE family.

Parse f352–357 (gap, overlapping SP, PS dipeptide)

Segment HAFE f358–359 (365–366; t---)

This segment is assigned as a hydrophobic anchor for the flanking coil regions.

Parse f360–370 (gap)

Possibly realignable.

Segment HAFE f371–392 (382–403; t306–327)

Following position f363, separate predictions must be made for the A2p+Fop subfamilies and the Ebs subfamily (subfamily Erp minus sequence l), as DARWIN ceases to find a statistically significant sequence similarity between the two. Of course, this requires structure predictions based on 6 and 11 sequences respectively. For the Ebs subfamily, a helix can be proposed for positions f375–390. The prediction for the Fop and A2p subfamilies also shows a good amphiphilic helix. Therefore, a helix can be assigned in this region. Further, the similar secondary structures assigned to the non-alignable segments of the A2p, Fop, and Ebs subfamilies suggests that these regions are indeed related by common ancestry, and are unalignable due to accumulation of an extraordinary number of point mutations.

THE CONSENSUS SECONDARY STRUCTURE

Comparison of the secondary structure predictions made for the three families show regions with good correspondence and regions with poor correspondence. In this comparison, regions in the Exp family should be ignored before alignment position 250 and after alignment position 335, as these segments of the Exp sequences are not significantly similar to the other sequences in the superfamily, and may indeed not be related by common ancestry.

TABLE 7. CONSENSUS SECONDARY STRUCTURE PREDICTION FOR THE SUPERFAMILY*

Unit	preferred	minimum	maximum	comments
Beta 1	055–059	055–058	053–059	alternating periodicity
Alpha 1	071–083	074–083	066–085	
Beta 2	087–091	087–090	086–092	
Active site 1	092–096	093–094	091–096	
Alpha 2	100–115	102–113	099–116	
Alpha 3	**118–130**	**118–128**	**118–130**	**adjust HAFE alignment**
Beta 3	**140–144**	**140–142**	**139–144**	**adjust HAFE alignment**
Active site 2	**148–157**	**missing**	**148–158**	**possibly beta**
Beta 4	163–169	166–169	162–170	
Active site 3	**175–179**	**175–179**	**175–179**	**possibly beta**
Alpha 4	192–207	193–206	190–213	
Beta 5	217–222	218–221	216–222	**shifted in HAFE family**
Beta 6	**237–242**	**coil**	**236–243**	**adjust HAFE alignment**
Beta 7	246–250	244–248	244–251	
Active site 4	**266–275**	**266–275**	**266–275**	**possibly beta**
Beta 8	**278–283**	**279–282**	**277–284**	
Beta 9	**290–294**	**coil**	**288–294**	
Beta 10	**300–304**	**coil**	**300–305**	
Alpha/2 beta	**310–323**	**310–322**	**308–335**	**internal helix preferred**
Active site 5	**326–335**	**326–335**	**326–335**	
Beta 11	347–351	347–350	343–353	
Beta 12	**356–358**	**coil**	**356–360**	
Alpha 5	388–401	390–399	386–402	

*Problematic assignments are in bold.

A summary of the consensus predictions is shown in Table 7. The strongly assigned secondary structures are those that are consistently assigned in all three families (or two families where the Exp family does not align). The other regions, where the three families do not have consistent secondary structure assignments, are problematic, and are discussed below.

118–130 coil alpha unaligned

The Ips family shows a satisfactory 3.6 residue periodicity from positions 118–130, provided that the conserved lysine at position 120 is permitted on the interior arc of the helical wheel. The alignment of the HAFE family is, however, poor, with multiple gaps that suggest that the alignment needs to be adjusted. The alignment can be adjusted to yield a helical pattern between positions 117 and 138 (a total of 10 residues total). Because the Ips family contains the target sequence, the consensus prediction favors the assignment for this family. However, considerable structure divergence is possible.

140–144 coil beta unaligned

The Ips family shows a canonical beta strand in this region. Again, however, the alignment of the HAFE family is poor in this region. Readjustment of the multiple alignment does not reveal an obvious strand in the HAFE family; a possible strand is from positions 140–143. Because the Ips family contains the target sequence, the consensus prediction favors the assignment for this family. However, considerable structure divergence is possible.

150–156 parse active site unaligned

As discussed above, this region is most likely a substrate binding region in the Ips family that is not found in the HAFE family with a possible beta-like structure.

175–179 parse active site unaligned

This segment is assigned as an active site in the Ips series. The segment from the HAFE family that is aligned, is clearly a coil. It contains gaps (poorly anchored) and parsing strings throughout. The differences might be explained by a poor alignment; consistent with this explanation, there are few alignment anchors. The strongest anchor is the assignment of helices in both families of protein starting with position ca. 190. There is no solid anchor on the amino terminus; the computer alignment is anchoring the segment on an APC P at position 123 and a CM1 N at position 178. These need not be homologous residues, however.

Comparison of this segment with other segments in the Ips family suggests that this string is near the active site. Therefore, an active site coil is a satisfactory alternative assignment for this region. Overall, this suggests a functional role in the Ips family that is not present in the HAFE family.

212–243 beta?/parse/parse beta/beta?/beta unaligned

These three uncertain portions are in another segment showing major gapping in the HAFE family. They are anchored on the amino end by a helix ending at ca. 213, and on the carboxyl end by a strand beginning at ca. 245. The first of the three beta assignments (217–222) is strong in the Ips family. Although not overlapping in the master multiple alignment, a beta strand in the beginning of this segment in the HAFE family (212–217) might correspond to this beta strand in the Ips family.

The second beta strand is defined only by two interior positions in the Ips family (227–228), and is therefore weak. This assignment receives no support from the HAFE family, where there are multiple gaps. Therefore, this beta strand is abandoned in the consensus secondary structure assignment.

The third involves a string of 6 interior assignments in the Ips family broken only by a very weak surface assignment at position 239. Again, this segment is heavily gapped in the HAFE family. One A2p subfamily member has residues throughout, including a remarkable pentapeptide

containing threonine as the sole residue type. Even within the HAFE family the gaps are not well anchored. Nevertheless, a second beta strand, not entirely convincing, might be squeezed out by a rearrangement of the multiple alignment. The beta would include alignment positions 224–233 (7 residues). Thus, it seems clear that the secondary structure in this region has undergone massive divergence in the superfamily, where functional constraints on divergence in this region have been greater in the Ips family than in the HAFE family. Thus, the beta assignment is retained in the consensus prediction.

266–275 active site active site unassigned

The active site segment in the Ips family can form an extended beta-like structure based on the unusual alternation in conserved functionalized residues. This may overinterpret this pattern in a family with only a few sequences and little overall sequence divergence. Given the differences in the reactions catalyzed by proteins in the HAFE and Ips families, the conformations of the polypeptide chain in this region should be different and difficult to model in a consensus model.

278–283 beta uncertain beta

Canonical beta assignments are made in the HAFE and Exp families. A beta assignment is not contradicted by the sequences in the Ips family. Therefore, a beta assignment is made in the consensus prediction. Again, so near the active site in a superfamily with such different catalytic functions, it is unlikely that the conformation in all three families is the same.

290–307 beta/beta gap/alpha? beta/beta?

The multiple alignment is clearly incorrect, an error that is not attributable to DARWIN but rather to the manual assembly of the master multiple alignment from alignments made by DARWIN for each of the three families. Once the alignment is readjusted, a prediction of two beta strands (290–294, then 301–304) is most plausible.

310–322 internal alpha or beta

This is the most difficult segment to assign. Following this segment, the conformations of proteins in the HAFE and Ips families do not appear to be identical. For example, in alignment positions 328–331, the Ips family has parses that are absent in the HAFE family. This segment in the Exp family shows alternating periodicity, admittedly with a very small multiple alignment.

In the first part of the segment, however, correspondence between the HAFE and Ips families is convincing. The three interior residues at positions 311–313 are found in all three families, as is the G at position 316. The ITN tripeptide at positions 322–324 is found in both the HAFE and Ips families. Thus, it appears that in these regions at least, secondary structure is conserved.

The segment from positions 310–323 is clearly a core structure. It is

either a pair of beta strands or an internal alpha helix. When modelling a tertiary structure, both assignments should be examined. Deciding which assignment to favor rests on the strength of the parse at position 316. An APC G is a parsing indicator, but not a strong one; indeed, APC Gs are found in internal helices, where packing constraints prevents substitution by other amino acids. There is no obvious covariation between residues at position i and $i + 2$, $i + 3$, and $i + 4$ that might be used to favor one of the two possible assignments.

326–335 alpha active site unaligned

The helical assignment in the HAFE family is unconvincing. The preferred assignment is simply an active site segment.

356–358 beta coil unaligned

The short beta assignment is accepted in the Ips structure, following adjustment of the multiple alignment.

365–366 beta parse unaligned

In this region, divergence is so significant that attempts to obtain a consensus secondary structure are not likely to be useful.

ACTIVE SITE ASSIGNMENTS

Given the enormous range of reactions catalyzed by the IPNS superfamily, many of the heuristics traditionally used to identify active site residues (16, 19, 90) fail to do so. For example, substrate binding residues almost certainly have not been conserved within this superfamily. Further, even in the best cases, a detailed understanding of the mechanism, reactivity, and divergent function of the members of a protein family can offer substantial insight in assigning secondary structure. The best structure predictions are made by those who understand these aspects of a protein system and use their understanding when modelling conformation.

The best studied enzyme from a mechanistic standpoint is isopenicillin N synthase itself (92, 93). The enzyme contains a single high spin non-heme iron (II) center in its active site (94). However, unlike with oxidative ring cleaving enzymes, where atoms from the dioxygen substrate end up in the product, the dioxygen substrate in isopenicillin N synthase is completely reduced to 2 equivalents of water. EPR, Mössbauer, electronic and NMR studies with iron, copper, and cobalt-containing enzyme suggest that the iron is bound by three histidine imidazole ligands, leaving three coordination sites free for the peptide substrate, oxygen, and solvent. NOE studies suggest that one of these sites is filled with a carboxylate, probably from aspartate (95). This model has been supported by very recent Fe K-edge X-ray adsorption studies (96). The alternative hypothesis that the enzyme might also contribute a thiol ligand to the iron led to several experiments where site-directed mutagenesis was used to replace cysteine residues in the protein (97, 98). Data from these studies makes a convincing case that the enzyme does not contribute an essential Cys residue to the reactive center.

The presence of three active site His ligands and the absence of a critical cysteine residue could have been deduced directly from the sequence analysis even in the absence of experimental data. In the regions where an alignment can be established by DARWIN, there are exactly 3 histidines that are absolutely conserved, at positions 093 (t048), position 272 (t212), and position 332 (t268). Given the conservation in a superfamily of proteins that shares little except a reactive center iron, these would be assigned ligands. Position 343 (t277) containing an Arg, and position 345 (t279) containing a Ser are also APC within the DARWIN alignment. Each of these are potential additional active site residues.

Interestingly, there is no APC D in the alignment. The only position that plausibly might contribute an aspartate as a ligand is the CM1 D found at position 274 (t214). This is two positions away from one of the of putative His ligands. However, in the A2 homolog from *Petunia hybrida*, the Asp is replaced by Glu. This could, of course, represent an error in the sequence; the sequence was derived in this alignment by private communication and was entered by hand. It is also conceivable that the carboxylate of Glu plays the same role in the single protein as the carboxylate in Asp does in all of the other proteins. Time did not allow resolution of this issue.

Importantly, the expandases do not align in the region containing His 093; they join the multiple alignment only at position 249 (t191). There is no APC H within the region of the expandases that precede the point where the expandases join the multiple alignment, and no APC H elsewhere in the expandases either. If the expandases were naively aligned with the rest of the superfamily, His 093 would lie at approximately position h010 in the expandases, and there is no obvious ligand that might be contributed from this segment. This implies that the ligands to iron have changed in the expandases. Identifying additional ligands is difficult within the Exp subfamily, as it contains only three proteins and the divergence is less than 70 PAM units. Thus, the most simple heuristic for assigning active sites, simply looking for conserved functionality, is not particularly useful.

However, an alternative heuristic searches for conserved functionality embedded within a conserved string. Within the expandase subfamily, the longest conserved strings containing conserved functionality are at positions h262–271 (RTSSVFFLRRP, 10 positions, containing the APC R at position 343 and APC S at position 345 conserved throughout the superfamily), h201–209 (CANGFVSLQ, 9 positions), h186–191 (PHYDLS, 6 positions, with the APC H and APC D), and h032–035 (HYLT, 4 positions). Shorter strings with conserved functionality include the tripeptides DFF (h053–055), AVT (h065–067), RRG (h073–075), STA

(h 083–085), DYS (h096–098), SMG (h102–104), YFD (h122–124), CGA (h 230–232), and NYV (h305–309).

Potential additional ligands to iron might be sought within these strings, although it is worth noting that the presumed second His ligand (position 332) is not embedded in a conserved string. The second longest string (h201–209) is matched in the multiple alignment against significant parsing elements, and contains a cysteine (h201) that might act as a ligand to iron. It is worth noting that the substrate for the expandases lacks a free thiol group that can act as a ligand.

SUMMARY

A new paradigm for predicting the secondary and tertiary structure of functional proteins from sequence data has emerged from detailed models of how natural selection, conservation, and neutral drift, the three fundamental factors in molecular evolution, leave their mark upon protein sequences. Structural information is extracted from a set of aligned homologous sequences via an analysis of patterns of conservation and variation between proteins with quantitatively defined evolutionary relationships. Tertiary structural information is obtained prior to the assignment of secondary structure, where it plays an important role. Throughout, structural predictions are made with the active involvement of a biochemist whose expertise and insight is critical both for making the prediction and in analyzing its successful and unsuccessful parts. Secondary structure predictions are evaluated based on their ability to sustain an effort to model tertiary structure. Several predictions made using the new paradigm can now be compared with those made under the classical paradigm, including a neural network. The results obtained from the new paradigm are clearly superior to those obtained with the classical paradigm, at least within the protein families that were examined.

ACKNOWLEDGEMENT

We are indebted to the Swiss National Science Foundation and Sandoz AG for partial support of this work. We are also indebted to Dr Cathie Martin (John Innes Inst., Norwich, U.K.) and Andy Prescott (John Innes Inst.) and Anton Gerats (Univ. Gent, Belgium) for communicating unpublished sequence data.

REFERENCES

1. D. L. OXENDER, *Protein Engineering*, Liss, New York (1987).
2. G. FASMAN, editor, *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum, New York (1989).

3. P. K. CHAKRAVARTY, K. B. MATHUR and M. M. DHAR, The synthesis of a decapeptide with glycosidase activity, *Experientia* **29**, 786–788 (1973).
4. B. GUTTE, M. DAEUMIGEN and E. WITTSCHIEBER, Design, synthesis and characterization of a 34-residue polypeptide that interacts with nucleic acids, *Nature* **281**, 650–655 (1979).
5. R. K. ALLEMANN, *Evolutionary Guidance as a Tool in Organic Chemistry*, Dissertation, E. T. H. No. 8804 (1989).
6. K. JOHNSSON, R. K. ALLEMANN and S. A. BENNER, Designed enzymes: new peptides that fold in aqueous solution and catalyze reactions, 166–187 in *Molecular Mechanisms in Bioorganic Processes* (C. BLEASDALE and B. T. GOLDING, eds.), Cambridge, Royal Society of Chemistry (1990).
7. K. W. HAHN, W. A. KLIS and J. STEWART, Design and synthesis of a peptide having chymotrypsin-like esterase activity, *Science* **248**, 1544–1546 (1990).
8. K. JOHNSSON, R. K. ALLEMANN, H. WIDMER and S. A. BENNER, Synthesis, structure, and activity of artificial, rationally designed catalytic polypeptides, *Nature*, **365**, 530–532 (1993).
9. E. T. KAISER and F. J. KEZDY, Amphiphilic secondary structure: Design of peptide hormones, *Science* **223**, 249–255 (1984).
10. D. EISENBERG, W. WILCOX, S. M. ESHITA, P. M. PRYCIAK, S. P. HO and W. F. DEGRADO, The design, synthesis, and crystallization of an alpha-helical peptide, *Proteins* **1**, 16–22 (1986).
11. M. H. HECHT, J. S. RICHARDSON, D. C. RICHARDSON and R. C. OGDEN, *De novo* design, expression, and characterization of Felix: A four-helix bundle protein of native-like sequence, *Science* **249**, 884–891 (1990).
12. K. GORAJ, A. RENARD and J. A. MARTIAL, Synthesis, purification and initial structural characterization of octarellin, a *de novo* peptide modelled on the α/β-barrel, *Protein Engineering* **3**, 259–266 (1990).
13. S. PADMANABHAM, S. MARQUSEE, T. RIDGEWAY, T. M. LAUE and R. L. BALDWIN, Relative helix-forming tendencies of nonpolar amino acids, *Nature* **344**, 268–270 (1990).
14. J. J. OSTERHOUT, Jr, T. HANDEL, G. NA, A. TOUMADJE, R. C. LONG, P. J. CONNOLLY, J. C. HOCH, W. C. JOHNSON, Jr, D. LIVE and W. F. DEGRADO, Characterization of the structural properties of α1b, a peptide designed to form a four-helix bundle, *J. Am. Chem. Soc.* **114**, 331–337 (1992).
15. T. HUNT and M. PURTON, 200 issues of TIBS, *Trends Biochem. Sci.* **17**, 273 (1992).
16. S. A. BENNER, Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure, *Advan. Enzyme Regul.* **31**, 121–181 (1989).
17. I. P. CRAWFORD, T. NIERMANN and K. KIRSCHNER, Prediction of secondary structure by evolutionary comparison: application to the a subunit of tryptophan synthase, *Proteins* **2**, 118–129 (1987).
18. J. F. BAZAN, Structural design and molecular evolution of a cytokine receptor superfamily, *Proc. Natl. Acad. Sci. USA* **87**, 6934–6938 (1990).
19. S. A. BENNER and D. 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).
20. R. B. RUSSELL, J. BREED and G. J. BARTON, Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains, *FEBS Lett.* **304**, 15–20 (1992).
21. A. MUSACCHIO, T. GIBSON, V.-P. LEHTO and M. SARASTE, SH3- An abundant protein domain in search of a function, *FEBS Lett.* **307**, 55–61 (1992).
22. S. A. BENNER, M. A. COHEN and D. GERLOFF, A predicted secondary structure for the src homology domain 3, *J. Mol. Biol.* **229**, 295–305 (1993).
23. D. L. GERLOFF, T. F. JENNY, L. J. KNECHT, G. H. GONNET and S. A. BENNER, The nitrogenase MoFe protein: A secondary structure prediction, *FEBS Lett.* **318**, 118–124 (1993).

of variation and conservation in homologous protein sequences. *J. Mol. Biol.* in press (1993).
68. S. SHOJI, D. C. PARMELEE, R. D. WADE, S. KUMAR, L. H. ERICSSON, K. A. WALSH, H. NEURATH, G. 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. Natl. Acad. Sci.* **78**, 848–851 (1981).
69. H. N. BRAMSON, E. T. KAISER and A. S. MILDVAN, Mechanistic studies of cAMP-dependent protein kinase action, *CRC Crit. Rev. Biochem.* **15**, 93–124 (1984).
70. S. SHOJI, L. H. ERICSSON, K. A. WALSH, E. H. FISCHER and K. TITANI, Amino-acid-sequence of the catalytic subunit of bovine type-II adenosine cyclic 3',5'-phosphate dependent protein-kinase, *Biochemistry* **22**, 3702–3709 (1983).
71. S. S. TAYLOR, J. A. BUECHLER, L. W. SLICE, D. K. KNIGHTON, S. DURGERIAN, G. E. RINGHEIM, J. J. NEITZEL, W. M. YONEMOTO, J. M. SOWADSKI and W. DOSPMANN, cAMP-dependent protein-kinase: A framework for a diverse family of enzymes, *Cold Spring Harbor Symp. Quant. Biol.* **53**, 121–130 (1988).
72. M. J. E. STERNBERG and W. R. TAYLOR, Modelling the ATP-binding site of oncogene products, the epidermal growth factor receptor and related proteins, *FEBS Lett.* **175**, 387–392 (1984).
73. D. C. FRY, S. A. KUBY and A. S. MILDVAN, ATP-binding site of adenylate kinase: mechanistic implications of its homology with ras-encoded p21, F1-ATPase, and other nucleotide-binding proteins, *Proc. Natl. Acad. Sci. USA* **83**, 907–911 (1986).
74. S. S. TAYLOR, cAMP-dependent protein kinase: Model for an enzyme family, *J. Biol. Chem.* **264**, 8443–8446 (1989).
75. 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).
76. R. B. RUSSELL, J. BREED and G. J. BARTON, Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains, *FEBS Lett.* **304**, 15–20 (1992).
77. S. A. BENNER, M. A. COHEN, G. H. GONNET, D. B. BERKOWITZ and K. JOHNSSON, Reading the palimpsest. Contemporary biochemical data and the RNA world, in *The RNA World* (R. GESTELAND and J. ATKINS, eds.), Cold String Harbor Press, 27–70 (1993).
78. R. HUBER, J. ROMISCH and E. P. PAQUES, The crystal and molecular-structure of human annexin-v, an anticoagulant protein that binds to calcium and membranes, *EMBO J.* **9**, 3867–3874 (1990).
79. S. A. BENNER, M. A. COHEN and D. GERLOFF, Correct structure prediction? *Nature* **359**, 781 (1992).
80. B. ROST and C. SANDER, Jury returns on structure prediction, *Nature* **360**, 540 (1992).
81. S. KOYAMA, H. T. YU, D. C. DALGARNO, T. B. SHIN and L. D. ZYDOWSKY, Structure of the P13K SH3 domain and analysis of the SH3 family, *Cell* **72**, 945–952 (1993).
82. G. J. BARTON and R. B. RUSSELL, Protein-structure prediction, *Nature* **361**, 505–506 (1993).
83. D. L. GERLOFF and S. A. BENNER, Predicting the conformation of proteins: Man versus machine, *FEBS Letters* **325**, 29–33 (1993).
84. L. BRITSCH, B. RUHNAU-BRICH and G. FORKMANN, Molecular cloning, sequence analysis, and *in vitro* expression of flavanone 3b-hydroxylase from *Petunia hybrida*, *J. Biol. Chem.* **267**, 5380–5387 (1992).
85. T. HASHIMOTO and Y. YAMADA, Purification and characterization of hyoscyamine 6b-hydroxylase from root cultures of *Hyoscyamus niger L.*, *Eur. J. Biochem.* **164**, 277–285 (1987).
86. J. KUPKA, Y.-Q. SHEN, S. WOLF and A. L. DEMAIN, Partial purification and properties of the α-ketoglutarate-linked ring-expansion enzyme of β-lactam biosynthesis of *Cephalosporium acremonium*, *FEMS Microbiol. Lett.* **16**, 1–6 (1983).
87. J. E. DOTZLAF and W.-K. YEH, Copurification and characterization of deacetoxycephalosporin C synthase/hydroxylase from *Cephalosporium acremonium*, *J. Bacteriol.* **169**, 1611–1618.
88. S. KOVACEVIC and J. R. MILLER, Cloning and sequencing of the beta-lactam hydroxylase gene (cefF) from *Streptomyces clavuligerus*. Gene duplication may have led to separate hydroxylase and expandase activities in the actinomycetes, *J. Bact.* **173**, 398–400 (1991).
89. A. MENSSEN, S. HOEHMANN, W. MARTIN, P. S. SCHNABLE, P. A. PETERSON, H. SAEDLER and A. GIERL, The En/Spm transposable element of *Zea mays* contains splice sites at the termini generating a novel intron from a dSpm element in the A2 gene, *EMBO J.* **9**, 3051–3057 (1990).
90. M. J. J. M. ZVELEBIL and M. J. E. STERNBERG, Analysis and prediction of the location of catalytic residues in enzymes, *Prot. Eng.* **2**, 127–138 (1988).
91. J. E. BALDWIN, J. B. COATES, M. G. MOLONEY, A. J. PRATT and A. C. WILLIS, Photoaffinity labelling of isopenicillin N synthetase, *Biochem. J.* **266**, 561–567 (1990).
92. J. E. BALDWIN, G. P. LYNCH and C. J. SCHOFIELD, Isopenicillin N synthase: a new mode of reactivity, *Tetrahedron* **48**, 9085–9100 (1992).
93. J. E. BALDWIN and E. ABRAHAM, The biosynthesis of penicillins and cephalosporins, *Natural Prod. Rep.* **5**, 129–145 (1988).
94. V. J. CHEN, A. M. ORVILLE, M. R. HARPEL, C. A. FROLIK, K. K. SURERUS, E. MÜNCK and J. D. LIPSCOMB, Spectroscopic studies with isopenicillin N synthase, *J. Biol. Chem.* **264**, 21677–21681 (1989).
95. L.-J. MING, L. QUE, JR, A. KRIAUCIUNAS, C. A. FROLIK and V. J. CHEN, NMR studies of the active site of isopenicillin-N synthase. A nonheme iron (II) enzyme, *Biochemistry* **30**, 1653–1659 (1991).
96. C. R. RANDALL, Y. ZANG, A. E. TRUE, L. QUE, JR, J. M. CHARNOCK, C. D. GARNER, Y. FUJISHIMA, C. J. SCHOFIELD and J. E. BALDWIN, X-ray absorption studies of the ferrous active site of isopenicillin N synthase and related model compounds, *Biochemistry* **32**, 6664–6673 (1993).
97. S. M. SAMSON, J. L. CHAPMAN, R. BELAGAJE, S. W. QUEENER and T. D. INGOLIA, Analysis of the role of cysteine residues in isopenicillin N synthetase activity by site-directed mutagenesis, *Proc. Natl. Acad. Sci.* **84**, 5705–5709 (1987).
98. A. M. ORVILLE, V. J. CHEN, M. R. HARPEL and B. G. FOX, Thiolate ligation of the active site Fe^{2+} of isopenicillin N synthase derives from substrate rather than endogenous cysteine. Spectroscopic studies of site-specific Cys->Ser mutated enzymes, *Biochemistry* **31**, 4602–4612 (1992).