PROTEINS: Structure, Function, and Genetics 23:431–445 (1995)

Evaluation of Current Techniques for Ab Initio Protein Structure Prediction

Tom Defay[1] and Fred E. Cohen[2]

[1]Graduate Group in Biophysics and [2]Departments of Pharmaceutical Chemistry, Biochemistry and Biophysics, Medicine, and Cellular and Molecular Pharmacology, University of California, San Francisco, California 94131-0450

ABSTRACT **The results of a protein structure prediction contest are reviewed. Twelve different groups entered predictions on 14 proteins of known sequence whose structures had been determined but not yet disseminated to the scientific community. Thus, these represent true tests of the current state of structure prediction methodologies. From this work, it is clear that accurate tertiary structure prediction is not yet possible. However, protein fold and motif prediction are possible when the motif is recognizably similar to another known structure. Internal symmetry and the information inherent in an aligned family of homologous sequences facilitate predictive efforts. Novel folds remain a major challenge for prediction efforts.** © 1995 Wiley-Liss, Inc.

Key words: protein structure prediction, secondary structure, evaluation

INTRODUCTION

In December, in Asilomar, California, a "Meeting on Critical Assessment of Techniques for Protein Structure Prediction" was held to determine the status of current methods for predicting the three-dimensional structure of proteins. Thirty-five different laboratories attempted, in a blinded fashion, to predict some aspect of the structure of 33 different proteins. The structures of these proteins were contemporaneously determined by NMR spectroscopy or X-ray crystallography, but were unavailable to the predictors prior to the submission of their predictions. Thus, these represent true or bona fide predictions in the spirit of the work of Schulz and collaborators on adenylate kinase,[1] or Curtis et al. on interleukin-4,[2] and not the "retrodictions" of structure that have been called into question by Benner and co-workers.[3]

The structure predictions fell into three categories: comparative modeling, threading, and ab initio structure prediction. Comparative modeling was defined as structure prediction when the structure of an homologous protein was known.[4–7] Threading predictions were computational attempts to align the sequence of a protein of unknown structure (that lacks clear similarity to another sequence of a protein of known structure) with the side chain environmental preferences dictated by a known protein structure.[8–11] Ab initio structure predictions attempt to solve the folding problem; given a protein sequence that is unrelated to any protein of known structure, what is its secondary and tertiary structure? While a great deal of effort has been devoted to this problem, many issues at the secondary structure level and most concerning tertiary structure remain unresolved.[12–22] Three different laboratories were chosen to evaluate the structure predictions; we evaluated the ab initio predictions.

METHOD

Categories of Predictions

The ab initio predictions were divided into four categories: *class, secondary structure, fold,* and *structure. Class* prediction was the simplest level of prediction. Predictors evaluated which of the following protein classes the protein most resembled: all α-helix, all β-sheet, α/β, or α + β.[23–27] The *secondary structure* category included predictions for each residue of the protein to be in one of three backbone conformations compatible with secondary structure: α-helix, β-sheet, or loop. Predictions of *fold* described the overall fold or shape of the protein, including many of the common folding motifs originally characterized by J. Richardson[28] and expanded upon recently by a number of groups.[29–31] Predictions in this category included secondary structure predictions. The final category, *structure,* was reserved for predictions of the three-dimensional coordinates of the protein. Predictions in this category naturally included specifications for the other three categories. The investigators and the categories in which they made predictions are shown in Table I. A short synopsis of their methods is given in Table II. Additional information about some of the prediction methodologies can be found in other articles in this issue of the Journal.

Received March 23, 1995; revision accepted July 10, 1995.

Address reprint requests to Dr. Fred E. Cohen, Department of Cellular and Molecular Pharmacology, University of California, San Francisco, 513 Parnassus Ave., San Francisco, CA 94131-0450.

TABLE I. Predictors and Categories

Predictor	Structure	Fold	Secondary structure	Class
Benner		X	X	
Covell	X			
Garnier			X	
Hubbard		X	X	
Lee	X			
Livingston			X	
Marshall	X			
Mekler		X	X	
Moult	X			
Munson			X	
Osguthorpe	X			
Rose				X
Rose and Sander			X	
Sander	X	X		

Evaluation of Predictions

The success or failure of class prediction was decided by visually assigning a class to each protein and comparing to the predicted class.[24]

Secondary structure predictions were evaluated by comparing the predicted with the experimentally determined secondary structure. The percentage correct score in a three-state system (Q_3: α-helix, β-sheet, or loop) was used.[32] The secondary structure of the experimentally determined structure was calculated with the program DSSP[33] which assigns secondary structure by examining hydrogen bonding patterns in the context of backbone dihedral angle preferences. The secondary structure predictions were further evaluated by subcategorizing the incorrect predictions into three categories: OVER, UNDER, and WRONG. OVER was defined as predicting an α-helix or β-sheet when the protein formed an aperiodic or loop structure in reality. UNDER was defined as predicting a loop conformation when the residue adopted an α-helical or β-sheet geometry. WRONG was defined as predicting an α-helix when the amino acid was in a β-sheet or vice versa. While one can imagine molecular dynamics simulations or other optimization methods correcting UNDER and OVER prediction, WRONG predictions are likely to be extremely difficult to recover from. For each protein, the secondary structure prediction methods were compared to the GOR[17] method as an historical standard.

The overall fold of the protein was evaluated qualitatively, from a visual comparison of the experimentally determined structure with the predicted description.

The precise structure of the protein was evaluated by the root mean square deviation (rms) between equivalent α-carbons in the predicted and experimental structures. This calculated value was compared to the rms value expected for random compact structures.[34]

The structure of the protein was also evaluated using a recently developed method that minimizes the area of a "soap film" that would join the predicted and experimentally determined polypeptide backbone. Benchmarks for this type of comparison have been developed to help to assess if any of the predicted models have captured features of the chain topology and fold.[35]

RESULTS AND DISCUSSION

All the predictors taking part in this contest should be congratulated. Many of the structure predictions were completed under less than ideal conditions. Some prediction methods, which typically require 6 months to apply in their entirety, were made in one. Some of the prediction strategies remain in a developmental phase and so these predictions should be regarded as work in progress. For this reason, we are stressing the promising results from the meeting, while still noting all of the results.

The results of the *structure* predictions are shown in Table III. The *fold* predictions are shown in Table IV; the *secondary structure* predictions are shown in Figure 1. The *class* predictions are shown in Table V.

Overview

The main issue in this section of the conference was whether or not it is presently possible to predict ab initio the tertiary structure of proteins. Two different approaches were used to predict tertiary structures. The first was Primary → Secondary → Tertiary, which involved predicting the secondary structure of the protein from the amino acid sequence(s) (Primary → Secondary) and then assembling a tertiary structure from the secondary structure elements (Secondary → Tertiary). The second method involved going straight from the sequence(s) to the tertiary structure. These two techniques met with varying success.

Primary → Secondary → Tertiary

Primary → secondary

Q_3 can be a misleading measure of prediction accuracy. The first step in predicting the structure of a protein following the Primary → Secondary → Tertiary approach is predicting the secondary structure. Secondary structure predictions are traditionally evaluated using a three state percentage correct score, Q_3. This approach was used to evaluate each prediction for the meeting. For 6-phospho-β-D-galactosidase, Benner and Sander correctly predicted this protein to be an α/β barrel. However, their secondary structure prediction accuracy differed considerably, 67% for Benner compared to 75.4% for Sander. Benner's level of accuracy was more similar to that of the GOR algorithm,[17] which correctly predicted 62% of the secondary structure of this protein (see

Fig. 2). With these modest per-residue prediction scores, how was Benner able to predict the correct fold? The answer lies in the analysis of the secondary structure prediction shown in Figure 1a. This figure demonstrates that Benner, Rost and Sander, Munson, and Garnier-COMBINE all had an exceptionally small number of "WRONG" predictions. From the viewpoint of fold prediction, the correct assignment of amino acids comprising the structural core of the protein is more important than the conformational assignments for amino acids that form the end of secondary structure elements and loop regions. For example, the structure of 6-phospho-β-D-galactosidase has a large excursion from the standard α/β barrel fold (see Fig. 3). This structure was missed by both Rost and Sander and Benner, and counts in the UNDER category. In addition, the exact beginnings and ends of secondary structure are less important for fold prediction. Errors in these places result in UNDER or OVER prediction. Figure 4 shows the parts of 6-phospho-β-D-galactosidase predicted correctly and as WRONG for both the Benner prediction and standard GOR predictions. It can be seen how a correct fold prediction was possible for Benner, but would have been unlikely if a fold prediction was made from the GOR secondary structure prediction.

Benefits of multiple sequence alignments. The main difference between the GOR algorithm and the secondary structure prediction methods demonstrated in this meeting is the exploitation of the structural information implicit in multiple sequence alignments. Since each sequence in the alignment codes for approximately the same structure, the secondary structure elements for each of these proteins should colocalize. This redundancy of information allows the central portion of most secondary structure regions to be assigned correctly (very few WRONG assignments). The ends of secondary structure regions vary between the sequences, and as expected OVER and UNDER prediction remains common in methods based on aligned sequence information.

It is thus not surprising that for xylanase, a more regular α/β barrel than 6-phospho-β-D-galactosidase, the secondary structure prediction accuracy for the multiple sequence methods improved (Fig. 1b), while the GOR method maintained the traditional level of accuracy.

As another example, each of T. Hubbard's predictions was analyzed, and compared to the GOR score for the same proteins. The average percent correct secondary structure per protein was 68.6% for Hubbard and 58.3% for GOR. However, the difference was much more pronounced with respect to the WRONG predictions. Hubbard had only 2.3% WRONG predictions while the GOR method produced WRONG predictions for 10.2% of the residues. The total errors for the OVER and UNDER predictions were 29.0% for Hubbard and 31.5% for GOR—almost identical.

When the multiple sequence alignment is limited in sequence number or covers a narrow phylogeny, the quality of the secondary structure prediction suffered. For example, the replication terminator protein had only two homologous sequences in the sequence data banks. Table VI shows the number of sequences present in a family of aligned homologous sequences for each protein. Figure 1h demonstrates the poor results. This affected the overall fold predictions as shown in Table IV; neither of the fold predictions was correct.

Synaptotagmin I C2 presented a different alignment problem. Table VI shows that synaptotagmin I C2 had 40 sequences in its multiple sequence alignment. However, the C terminal end of the protein showed a large amount of sequence divergence. Both Hubbard and Benner correctly predicted the first six strands of synaptotagmin I C2, but the C terminal strand was mispredicted to be a helix by both labs.

Prokaryotic ribosomal protein L14 demonstrated that even with the highly unusual structure shown in Fig. 5, accurate secondary structure assignments can be made with a sufficiently large family of aligned sequences, 25 sequences in this case. Figure 1j shows the secondary structure prediction.

Is human intervention superior to totally automated approaches? A point which has been argued in the literature is whether or not human intervention is superior to totally automatic methods in secondary structure prediction.[36,37] For a group of proteins, Hubbard and Rost and Sander both used the same secondary structure prediction algorithm: PHD.[20] However, Hubbard aligned the sequences automatically and optimized them by hand; Rost and Sander's alignment method was totally automated. Figure 6 indicates that Hubbard's hand alignment improved the ability of PHD to accurately predict secondary structure.

Another example of man vs. machine is shown by the predictions made by Benner and co-workers. Although much of Benner's technique is now automated, there is still a large human element in the structure predictions that their lab performed. Figure 7a and b demonstrates that their method has a level of effectiveness for secondary structure prediction similar to many of the automated methods.

What have we learned about secondary structure prediction? The Mystery protein demonstrates the correspondence between the information we use to predict secondary structure and that used to design proteins. Mystery was a designed TIM barrel called RORO. It was designed by the first EMBO protein design course, improved by Chris Sander and Gert Vriend,[38] and produced by Steve Emery.[39] The extremely accurate secondary structure predictions shown in Figure 1c indicate that the rules used to design this TIM barrel strongly resemble the rules

TABLE II. Synopsis of Methods

Investigator	Abbreviation	Method
The ETH Prediction Group: D. Gerloff, G. Chelvanayagam, and S.A. Benner	Benner	The prediction method applies automated heuristics to assign surface, interior, active site (tertiary structural information), and parsing residues by analysis of patterns of conservation and variation among homologous protein sequences in light of evolutionary models that interpret amino acid substitutions as the consequence of neutral variation subjected to functional constraints together with adaptive variation that alters the properties of homologous proteins to make them optimally suited to different environments. Secondary structural elements are assigned from patterns in the tertiary structural information.[59]
B.K. Lee and N. Kurochkina	Lee	For a polypeptide chain, a biased Monte Carlo search was applied for the dihedral angles of the main chain phi and psi and side chain dihedral angels chi. Conformational space was reduced into a small number of allowed regions in Ramachandran phi and psi map.[60] Weighted sum of hydrophobic energy based on pairwise surface area sum[61] and hydrogen bond energy calculated as electrostatic Coulomb sum was used to estimate the energy of the structure.
S.G. Galaktionov and G.R. Marshall	Marshall	The secondary structure of the protein is predicted using a consensus of three methods implemented in SYBYL 5.5. Next an algorithm was used to predict coordination number vectors for the amino acid residues.[62] Then the residue–residue contact matrix was predicted using an iterative procedure to improve heuristic gain function. Finally, the spatial structure was reconstructed.[63]
J.T. Pedersen and J. Moult	Moult	A torsion space representation of a protein is used with an all atom force field[64] together with a genetic algorithm[65] and a Monte-Carlo algorithm[66] to predict the structure of small proteins.
D.J. Osguthorpe	Osguthorpe	A simplified model of protein structure with potentials developed to reproduce the physical behavior of atoms rather than protein statistics derived from the database. The potentials are being continuously improved to reproduce protein-like structures.
T. Hubbard and J. Park	Hubbard	Automatic alignment of sequences using the PHD server[20] followed by addition of more sequences and hand alignment. These alignments were then submitted to the PHD neural network in Heidelberg. Fold prediction was aided by a strand pairing algorithm.[67]
B. Rost and C. Sander	Rost and Sander	Secondary structure was predicted for all proteins using the neural network method that uses sequence profiles as input.[20]
L. Holm, B. Rost, P. Bork, and C. Sander	Sander	The secondary structure elements were then assembled into three-dimensional structures.
G. Livingston and H.B. Nicholas	Livingston	Case based learning approach. Various 22 amino acid segments are compared to the protein to be predicted; if the sequence matching score exceeds a threshold, the structure of the 22 amino acid segment is used as evidence to predict the secondary structure.[68,69]
J. Garnier and J.M. Levin	Garnier-SIMPA	SIMPA (SIMilarity Peptide Analysis) program is based on sequence similarity between a stretch of amino acids (17 amino acid long) of the test sequence and the sequences in a data base of protein structures. Q_3 of 86% when a homologous protein structure is present, otherwise 63–65%.[70] When homologous sequences are known, it can be associated with the CONSENSUS program[41] to yield an accuracy of 68–69%.

(*continued*)

TABLE II. Synopsis of Methods (*Continued*)

Investigator	Abbreviation	Method
J. Garnier and V. DiFrancesco	Garnier-COMBINE	The COMBINE method is an expert system amalgamation of three secondary structure prediction algorithms: GOR III, SIMPA, and Bit Pattern.[71] It can be associated with multiple sequence alignments (CONSENSUS) to yield an accuracy of 69–70%.[72]
P. J. Munson and V. DiFrancesco	Munson	Two different multiple sequence methods: QL (quadratic logistic), Profile-QL. The QL method is a calibrated logistic model for a three state prediction using the maximum likelihood principle.[73] The profile method combines this method with multiple sequence alignment information.[72] The expected accuracy for Profile-QL is 67–69% measured in two separate cross-validated tests.
R.G. Idlis and L.B. Mekler	Mekler	Prediction of specific contacts between amino acid residues of the protein molecule being in the intermediate conformation, the so called "molten globule." These contacts are supposed to be determined by the specific binding of amino acid residues encoded by a codon and its anticodon. The folding of an amino acid sequence into the "molten globule" is a step-by-step cotranslational process of the formation and reorganization of these code bonds. An additional stereochemical code is supposed to determine the first-order phase transition that underlies protein activity. It is supposed that the two conformations of a protein molecule have a similar topology of the backbones by the entirely different systems of hydrogen bonds and van der Waals interatomic contacts.[74]
D. Covell	Covell	Simulated annealing methods are applied to a simple cubic lattice α-carbon model of a protein. Each amino acid occupies only one lattice site. Several simulations of greater than 100,000 steps are carried out to determine the consensus configuration of the protein.[75,76]
R. Srinivasan and G. Rose	Rose	Not specified at the time of submission.

used to predict its structure. These are not, however, the rules nature uses for folding proteins; RORO showed approximately the correct helical content but almost no beta sheet by CD and NMR spectroscopy.[40]

Difficult protein substructures. The following are some examples of specific errors that occurred in secondary structure prediction.

- Completely exposed helices are consistently difficult to predict due to the lack of the classical hydrophilic, hydrophobic repeating pattern of the more common partially buried helices. For instance, most groups missed one of the exposed helices in ribosomal protein L14.
- Completely buried strands and helices are also difficult to predict due to their lack of a repeating hydrophobic/hydrophilic pattern.
- As previously mentioned, excursions of secondary structure not present in the entire family of aligned sequences are difficult to predict.
- Finally, the ends of secondary structure units are still frequently misassigned. Perhaps work on capping structures will serve to address this problem.[41–44]

Secondary → tertiary

Approaches to secondary structure assembly. One approach to assembling the structure of a protein is to attempt to combine the secondary structure units of the protein in every plausible way, and evaluate which assembly is most likely to be correct.[15,45–47] Benner attempted to assemble synaptotagmin I C2 by this combinatorial approach. Unfortunately, they mispredicted a strand for a helix. Even so, the correct overall fold was present in their list of plausible folds. This fold was rejected in the evaluation stage (Table IV).

The task of assembling secondary structure units is simplified when the secondary structure exhibits a pattern seen before. The overall fold of 6-phospho-β-D-galactosidase was determined largely because the repeated α-β pattern was familiar to the investigators (Table IV). Sander was also able to propose a structure with coordinates (Table III). Even when the secondary structure exhibits a familiar pattern, mistakes are possible. Sander rejected an α/β barrel in favor of an alternative α/β structure with one or more sheets rather than a closed barrel (Table IV).

Assembly of the secondary structure units into a

TABLE III. Evaluation of Protein Structure Prediction

Protein	Length (residues)	Random rms (Å)*	Predictor	rms (Å)	Soap bubble†	Energy
Membrane binding domain for the C2 domain of human coagulation factor VIII	22	10.3	Moult	4.4	0.15	−46.1
				8.8	0.29	−45.1
				9.1	0.35	−41.2
			Lee	4.4	0.14	−236
				7.7	0.23	−212
Subtilisin propiece	71	12.6	Marshall	11.4	0.35	
Subtilisin propiece segment	16	10.0	Moult	10.2	0.43	
Domain 3 of staufen	68	12.4	Marshall	13.7	0.36	
			Osguthorpe	19.6	0.33	
				21.3	0.31	
				12.9	0.35	
Chymotrypsin/elastase inhibitor-1	63	12.2	Covell	7.3	0.32	
6-Phospho-β-D-galactosidase	454	30.4	Sander	‡	0.26	

*The standard deviation associated with the average random rms scores is ± 1.4 Å.[34]
†A soap bubble value of <0.35 is somewhat accurate, <0.3 is adequate, <0.25 is good, <0.2 is very good.[35]
‡The rms could not be determined due to a reversal in chain tracing.

fold is also aided by local folding motifs such as the leucine zipper.[48,49] This motif is formed when two helices pack against one another with leucines at the interface (see Fig. 8). The fingerprint of this motif is a leucine repeated every seventh amino acid. Hubbard was able to recognize this motif and correctly predict a leucine zipper in chorismate mutase. Since only one leucine zipper helix was shown, Hubbard correctly predicted that chorismate mutase was "an all helical dimer with a coiled coil along the N-terminal helix." Another example of the leucine zipper motif was seen in the replication terminator protein. In this case again the presence of a dimer was identified.

When the protein to be predicted has an unusual fold, it is more difficult to assemble the secondary structure units. Two of the proteins that several labs made fold predictions for had an unusual fold: domain 3 of staufen and subtilisin propiece. They both had a β-sheet sandwiched against a pair of helices. This can be contrasted with the more common motif of helices covering both sides of a β-sheet.[46]

Marshall and Mekler's groups both predicted domain 3 of staufen to have the two helices on opposite sides of the sheet. Hubbard's group correctly predicted that the helices would lie on the same side of the sheet. The predicted folds with coordinates are shown in Figure 9. The rms and soap bubble values are shown in Table III.

Another unusual motif was the subtilisin propiece, although its structure is surprisingly similar to that of domain 3 of staufen. As shown in Table III, Marshall predicted the overall fold with an rms error of 11.4 Å. Moult predicted the conformation of residues 7–22 with an rms error of 10.2 Å. The unusual folds seem to have affected the secondary structure predictions as well. As seen in Figure 1f and g, the secondary structure predictions are frequently in error. It is possible that common folds have improved secondary structure prediction accuracy due to the presence of similar structures in the databases used to derive prediction parameters.

The high symmetry of the TIM barrels clearly aided prediction of the overall fold of proteins in this study as well as in previous efforts.[16] This effect was also seen with biphenyl-2,3-diol 1,2-dioxygenase whose overall fold was predicted by Hubbard to be "two symmetrical regions, each split into two E-H-E-E-E-E regions," where E is an extended β-strand and H is an α-helix. In reality each region is E-H-E-E-E. The secondary structure prediction is shown in Figure 1d. Gene duplication and other evolutionary mechanisms frequently lead to proteins with substantial internal symmetry. Clearly, this can be put to advantage in prediction efforts. When recognized, symmetry elements can serve to multiply the amount of homologous sequence information and frequently extends the phylogenetic separation between structurally related elements. This was observed with the internal 2-fold identified by Hubbard in biphenyl-2,3-diol 1,2-dioxygenase, as well as the implicit 4-fold in four-helix bundles or 8-fold symmetry in α/β barrels.

Are we really just threading? Threading matches a protein sequence with known protein structures. The two correct α/β barrel predictions were essentially matching the patterns of secondary structure of the unknown protein to that of known proteins.

TABLE IV. Evaluation of Protein Fold Prediction

Protein	Actual fold	Predictor	Prediction
6-Phospho-β-D-galactosidase	α/β barrel	Benner	α/β barrel
		Sander	α/β barrel
Xylanase	α/β barrel	Sander	α/β structure with one or more β-sheets rather than a closed barrel
Biphenyl-2,3-diol 1,2-dioxygenase	Two symmetrical regions split into two regions of E-H-E-E-E	Hubbard	Two symmetrical regions split into two regions of E-H-E-E-E-E
Membrane binding domain for the C2 domain of human coagulation factor VIII	α-Helix with a twist on the end	Moult (3 predictions)	1. α-Helix with a twist on the end
			2. Disordered β-structure
			3. Short helix packed against a strand/coil
		Lee (2 predictions)	1. α-Helix with a twist on the end
			2. α-Helix with a β-strand pair at the end
Chorismate mutase	All α-helical dimer with a coiled coil along the N-terminal helix	Hubbard	All α-helical dimer with a coiled coil along the N-terminal helix
Domain 3 of staufen	Two α-helices packed against the same face of a three stranded β-sheet	Hubbard	Two α-helices packed against the same face of a two stranded β-sheet
		Mekler	Two α-helices packed against opposite sites of a two-stranded β-sheet
		Osguthorpe (3 predictions)	1. Disordered β structure
			2. N-terminal α-helix and disordered coil
			3. Compact disordered coil
		Marshall	Two α-helices packed against opposite sides of a two stranded β-sheet
Chymotrypsin/elastase inhibitor-1	Coiled structure with five disulfide bonds	Covell	Coiled structure with five disulfide bonds
Replication terminator protein	α+β leucine zipper dimer	Hubbard	All α-helical protein making a leucine zipper dimer
		Mekler	α + β dimer differing in placement of the secondary structure regions, resulting in different overall fold
Synaptotagmin I C2	β-Sandwich	Benner	Pleckstrin like seven β-strands plus one α-helix
		Hubbard	Pleckstrin like seven β-strands plus one α-helix
Subtilisin propiece	Three strand β-sheet packed against two helices	Marshall	One stranded β-sheet with helices on either side
Subtilisin propiece segment	Extended	Moult	β-Hairpin

Benner's misprediction of synaptotagmin C2 was partly due to mispredicted secondary structure. The secondary structure pattern they predicted matched the pattern of the pleckstrin family of folds. Hubbard's fold predictions for biphenyl-2,3-diol 1,2-dioxygenase and synaptotagmin C2 were accomplished with a combination of secondary structure prediction and threading. For these reasons, it can be argued that these approaches to fold prediction should be classified under the "threading" category of structure prediction. Ab initio fold prediction would then be limited to the strict combinatorial approaches to tertiary structure or methods that did not employ secondary structure units as intermediates. This

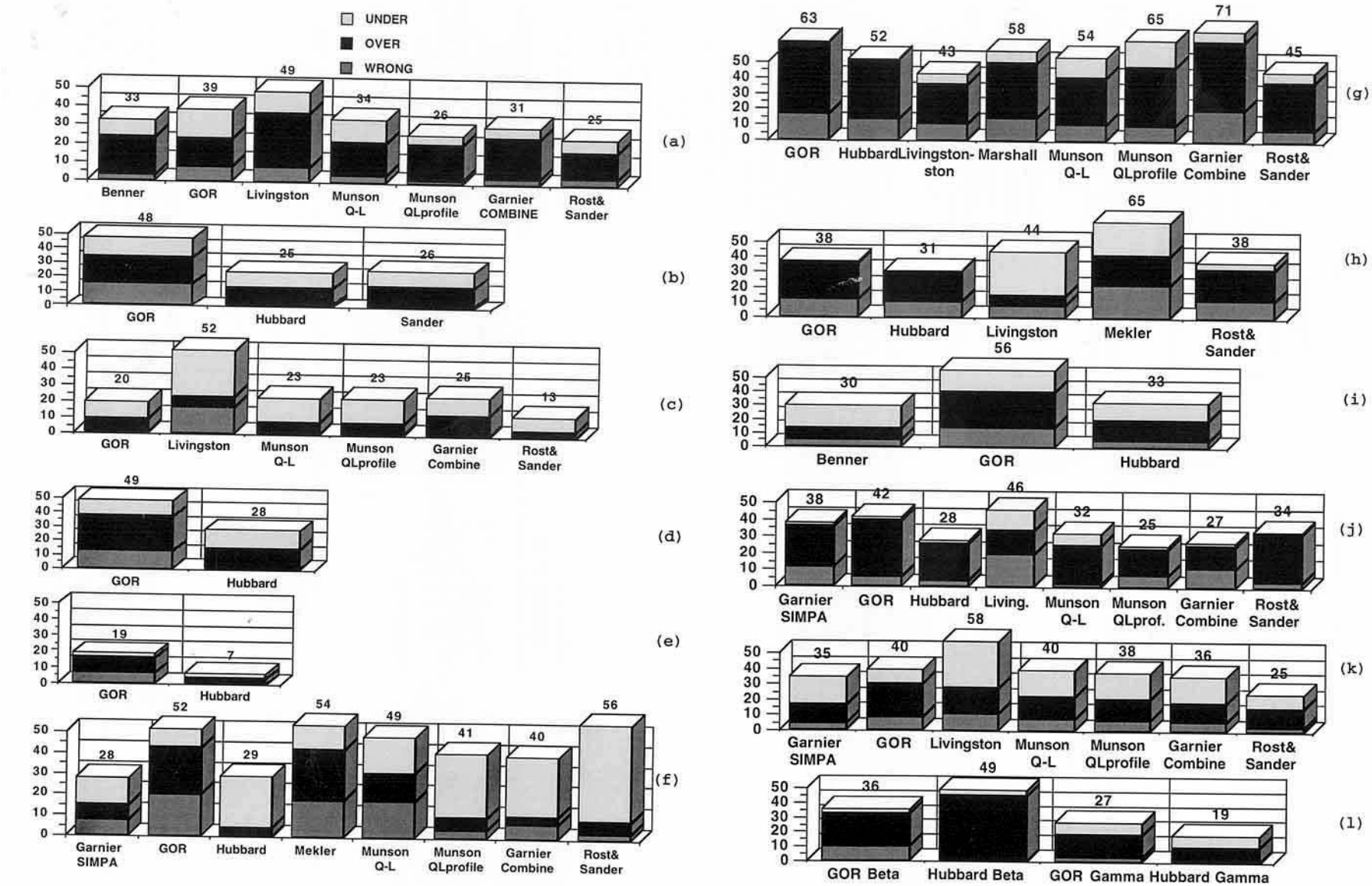


Fig. 1. Secondary structure predictions for proteins broken down into three categories: OVER, UNDER, WRONG (see text for definitions). The percentage score above the histogram is the total percentage incorrect predictions. **(a)** 6-Phospho-β-D-galactosidase, **(b)** xylanase, **(c)** mystery, **(d)** biphenyl-2,3-diol 1,2-dioxygenase, **(e)** chorismate mutase, **(f)** domain 3 of staufen, **(g)** subtilisin propiece, **(h)** replication terminator protein, **(i)** synaptotagmin I C2, **(j)** prokaryotic ribosomal protein L14, **(k)** pyruvate phosphate, **(l)** *Klebsiella aerogenes* urease: beta and gamma subunits.

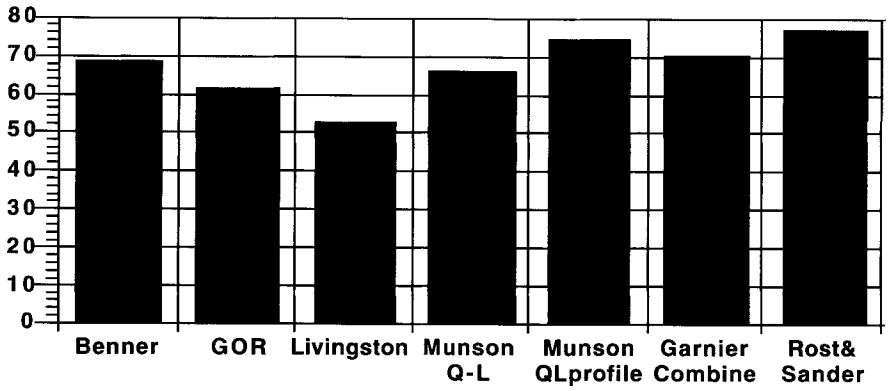


Fig. 2. Percent correct secondary structure of 6-phospho-β-D-galactosidase in a three state system (α-helix, β-sheet, or other). The secondary structure of the experimentally determined structure was calculated by DSSP.[33]

TABLE V. Evaluation of Protein Class Prediction

Protein	Class	Predictor	Prediction
Chorismate mutase	All α-helical	Rose	All α-helical
Synaptotagmin I C2	All β-sheet	Rose	All β-sheet

may prove to be an arbitrary distinction. As ab initio methods improve, threading algorithms will exploit these features to their own advantage. If there are only a limited number of protein folds, and X-ray crystallographers and NMR spectroscopists continue to solve a wide array of new structures, threading algorithms are likely to limit the need for true ab initio approaches.

Primary → Tertiary

This category of methods includes those that do not use the calculation of secondary structure as an intermediate for structure determination. These include the work of Mekler, Marshall, Lee, Moult, and Covell. These methods have the advantage over Primary → Secondary → Tertiary methods in that they are not simply threading. They are clearly applicable to the prediction of novel structures and folds. Moreover, the work of Chan and Dill[50] would suggest that secondary and tertiary structure are inextricably tied. This has led to the notion that tertiary structure determines secondary structures. While this extreme point of view is unlikely to be true in all cases,[51,52] it is clear that the structure of some local sequences is largely influenced by their tertiary context.[23,53]

Contact matrices

Mekler and Marshall both used methods that involved generating the tertiary structure from a predicted set of interresidue contacts. Unfortunately, as with the Primary → Secondary → Tertiary methods,

TABLE VI. Number of Aligned Sequences for Predicted Proteins*

Protein	Sequences*
6-Phospho-β-D-galactosidase	16
Mystery	2
Xylanase	13
Biphenyl-2,3-diol 1,2-dioxygenase	16
Membrane binding domain for the C2 domain of human coagulation factor VIII	1
Pyruvate phosphate dikinase	6
Chorismate mutase	7
Domain 3 of staufen	8
Klebsiella aerogenes urease β	9
Klebsiella aerogenes urease γ	11
Chymotrypsin/elastase inhibitor-1	1
Replication terminator protein	3
Synaptotagmin I C2	40
Prokaryotic ribosomal protein L14	25
Subtilisin propiece	12

*Each of the sequences in the alignment was considered to be homologous to the probe sequence when the sequence identity was ≥ 30%.

Mekler and Marshall were unable to correctly predict the structure of domain 3 of staufen. Marshall incorrectly predicted the structure of subtilisin propiece, and Mekler incorrectly predicted the fold of the replication terminator protein. These results are not surprising in light of the predicted contact matrices of Mekler. For domain 3 of staufen, Mekler

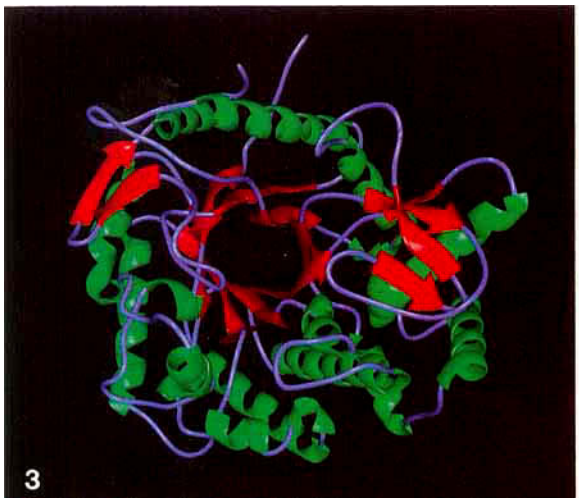


Fig. 3. 6-Phospho-β-D-galactosidase. Picture generated by midas Ribbonjr.[58] β-Strands are in red, α-helices are in green, and the rest of the chain is in purple. Secondary structure calculated by DSSP.[33]

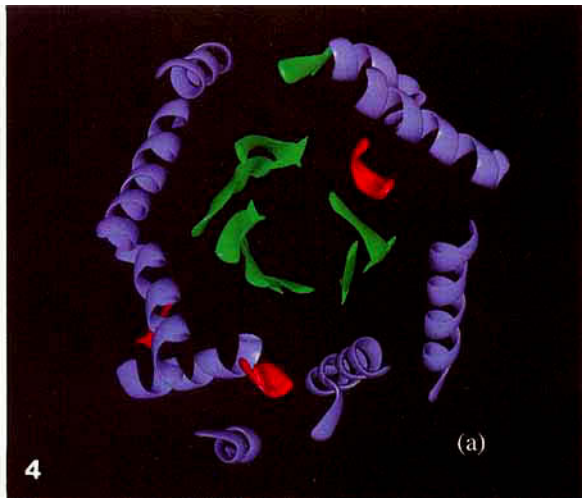


Fig. 4. 6-Phospho-β-D-galactosidase. Correct α-helix predictions are shown in purple, correct β-strand predictions are shown in green, and WRONG (α-helix predicted for β-sheet region or β-sheet region predicted for α-helix region) predictions in red. The prediction in **(a)** was done by Benner. The prediction in **(b)** was done with the 1977 version of the GOR algorithm.

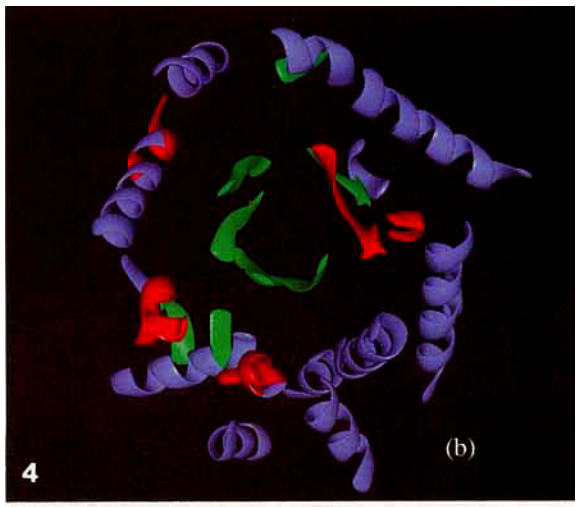
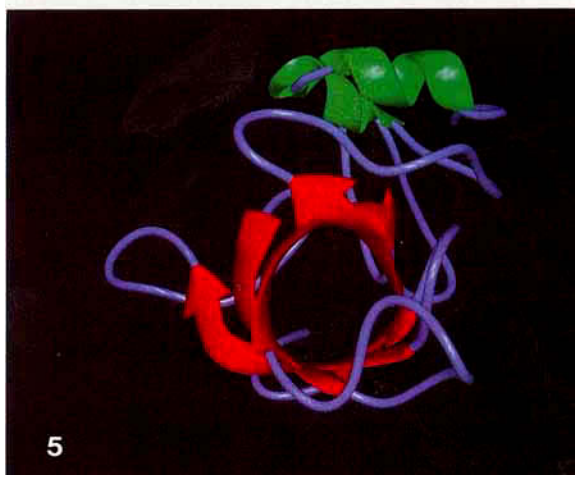


Fig. 5. Prokaryotic ribosomal protein L14. Picture generated by midas Ribbonjr.[58] β-Strands are in red, α-helices are in green, and the rest of the chain is in purple. Secondary structure calculated by DSSP.[33]



correctly predicted none of the seven long range (greater than five residue separation) contacts. For the replication terminator protein, Mekler predicted none of the 23 long range contacts correctly.

Semi-exhaustive methods

The membrane-binding domain for the C2 domain of human coagulation factor VIII indicates that structures of very small proteins may be easier to predict. This peptide is only 22 amino acids in length. The NMR structure and two lowest energy predicted structures are shown in Figure 10. The structures from both Moult's group and Lee's group are qualitatively quite accurate: a helix followed by an N-terminal twist. As shown in Table III, each predicted structure deviated from the NMR structure by 4.4 Å rms, and had low soap bubble values. Moult's group made two other predictions and Lee's group one other. These other predictions were higher in energy and correspondingly less accurate. However, Lee's high energy prediction was convincing enough that it was chosen by that group to be their preferred prediction.

One partially successful example of ab initio protein folding was the effort of Covell on chymotrypsin/elastase inhibitor-1. His predicted structure for this 65 residue protein was 7.3 Å rms from the actual structure. At the simplest level of comparison,

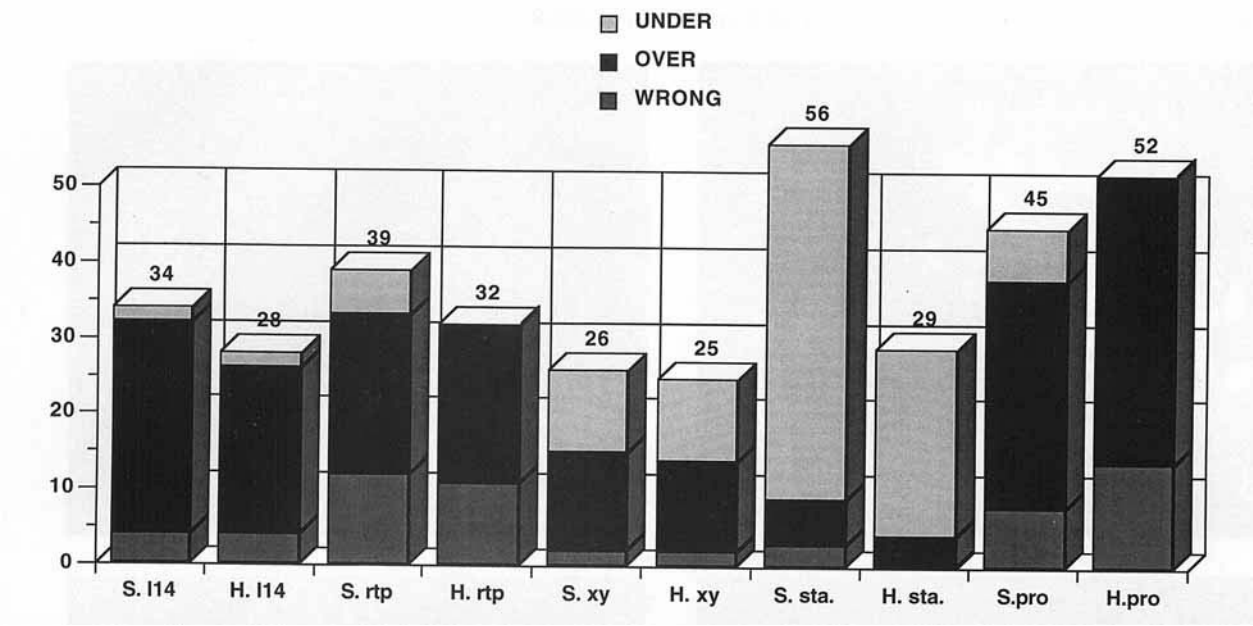


Fig. 6. Comparison of automated sequence alignment versus automated plus hand alignment. Secondary structure predictions were broken down into three categories OVER, UNDER, WRONG. The percentage score above the histogram is the total percentage incorrect predictions. Rost and Sander and Hubbard are abbreviated S and H. Proteins evaluated include ribosomal protein L14 (L14), replication terminator protein (rtp), xylanase (Xy), domain 3 of staufen (Sta), and subtilisin propiece (Pro). The secondary structure of each of these proteins was predicted by both investigators. Hubbard used automated alignments supplemented by hand alignment, Rost and Sander used automated alignments. They both submitted the alignments to Rost and Sander's PHD server.[20]

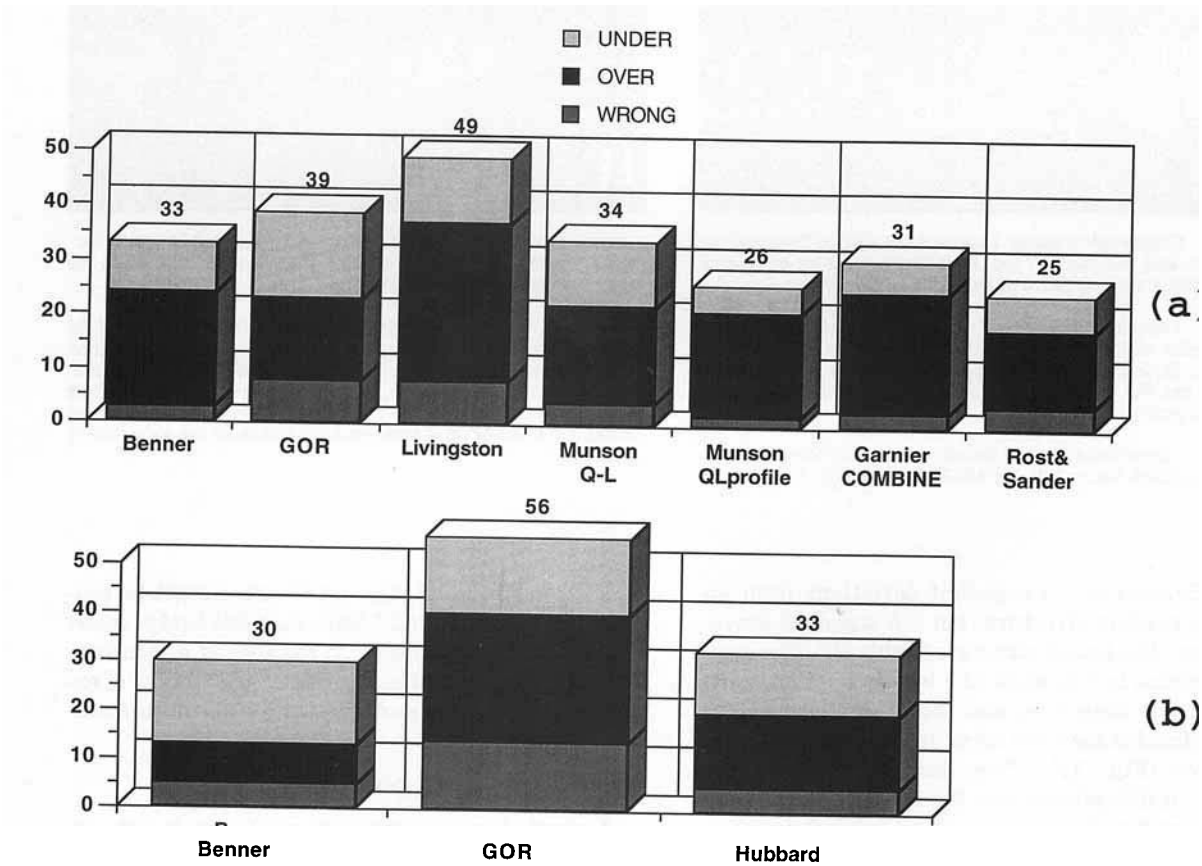
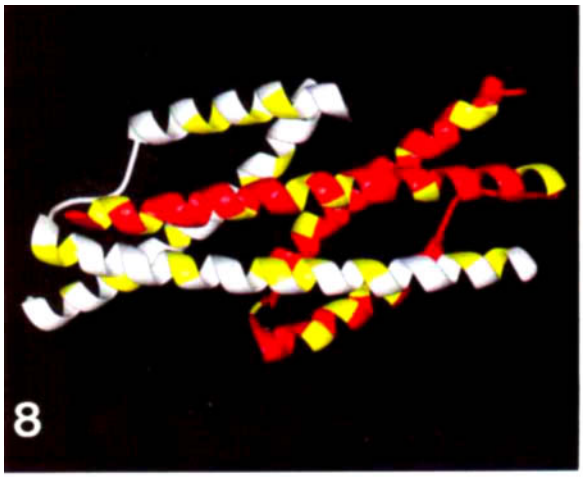
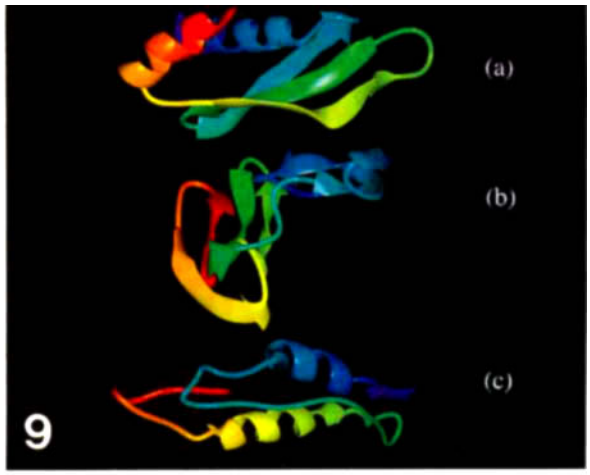
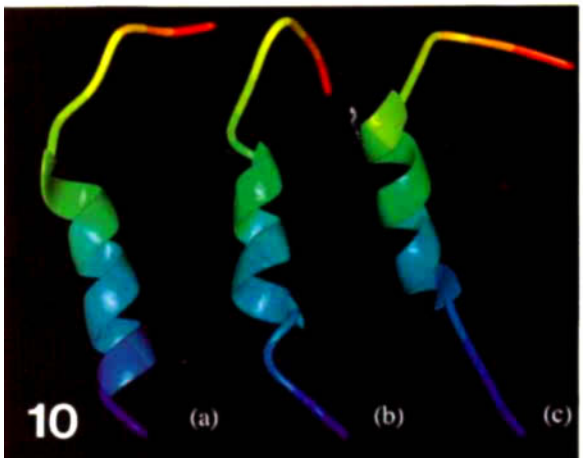


Fig. 7. Comparisons of the Benner prediction of **(a)** 6-phospho-β-D-galactosidase with several other investigators and of **(b)** synaptotagmin I C2 with Hubbard. Secondary structure predictions were broken down into three categories OVER, UNDER, WRONG. The percentage score above the histogram is the total percentage incorrect predictions. This is a comparison between human guided structure prediction (Benner) and automated approaches (Hubbard and others).







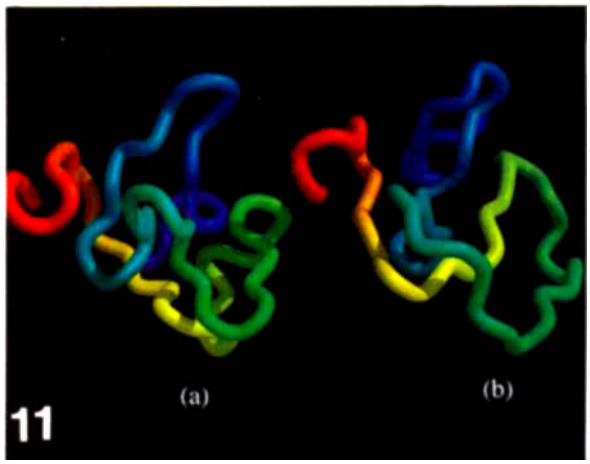


Fig. 8. Chorismate mutase. Leucines are yellow. One subunit is in white and the other in red. Secondary structure evaluated using DSSP. Picture generated by Midas Ribbonjr.[58]

Fig. 9. Domain 3 of staufen. **(a)** Experimentally determined X-ray crystal structure. **(b)** Osguthorpe prediction. **(c)** Marshall prediction. Structures generated by Midas Ribbonjr.[58] The N-terminus is red, the C-terminus is blue, and the middle of the sequence is green.

Fig. 10. Membrane binding domain for the C2 domain of human coagulation factor VIII. **(a)** NMR structure. **(b)** Lee prediction.**(c)** Moult prediction. Pictures generated with MIDAS Ribbonjr.[58] The N-terminus is red, the C-terminus is blue, and the middle of the sequence is green.

Fig. 11. Chymotrypsin/elastase inhibitor **(a)** predicted and **(b)** experimental. Picture generated with Midas Neon.[58] α-Carbon representation with exaggerated carbon radius used to emphasize the topological features. The tube was colored in a rainbow pattern corresponding to the amino acid number. The N-terminus is red, the C-terminus is blue, and the middle of the sequence is green.

this difference is ~3 standard deviations from an average random structure, but ~5 standard deviations from the actual structure (Table III). The soap bubble value is indicative of a loosely similar structure. Visual inspection also shows some structural similarities between the experimental and predicted structure (Fig. 11). Five disulfide bonds were present in this protein and the exact pairings were made known to the investigators. Covell did not use these pairing during the prediction phase of his work, but only later as a check on the accuracy of his prediction. Since he did not have knowledge of these pairing before hand, this prediction cannot be considered entirely "blind." Still, disulfide bridge information is often available in advance of a structure determination and thus provides a useful type of experimental structure constraint for ab initio methods.[2,54]

CONCLUSION

Accurate tertiary structure prediction is not possible today. Overall fold prediction is possible and has been demonstrated. We can predict the overall fold of a protein when that protein has a recogniz-

able motif. Examples are the leucine zipper seen in chorismate mutase and replication terminator protein, and the α/β barrels xylanase and 6-phospho-β-D-galactosidase. We are also aided by a large degree of symmetry present in the amino acid sequence, which translates to symmetry in the three-dimensional structure. We can predict the approximate structure of extremely small proteins, such as the membrane-binding domain for the C2 domain of human coagulation factor VIII. For these tiny proteins, it is possible to pursue extensive conformational searches to predict a protein's tertiary structure. Unfortunately, the predicted structures are still 4.4 Å rms from the experimental structures. There is hope that these methods may be extended to somewhat larger proteins as shown by Covell's prediction of chymotrypsin/elastase inhibitor-1, but in this example, the structural resemblance is tenuous.

We still have difficulty with proteins that have unusual folding motifs, such as domain 3 of staufen and the subtilisin propiece. In addition, most of the recent advances made in structure predictions have been due to the exploitation of multiple sequence alignments. When the quality of these alignments was poor as was seen with the replication terminator protein and synaptotagmin I C2, prediction accuracy suffered. Given the current level of prediction accuracy, we recommend the use of as much experimental information as possible in structure prediction and/or subsequent validation.[54–56]

To improve tertiary structure prediction, multiple sequence information may have to be included in the Primary → Tertiary methods. Already, this information could improve the distance matrix approaches by using sequence variability information across an aligned family to narrow the range of coordination numbers for contact map approaches. Multiple sequence alignments could also be adapted to calculate more specific contact potentials. We expect that the next generation of tertiary structure prediction strategies will exploit multiple sequence information.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institutes of Health (GM39900). Molecular graphics images were produced using the MidasPlus program from the Computer Graphics Laboratory, University of California, San Francisco (supported by NIH RR-01081).

REFERENCES

1. Schulz, G.E., Barry, C.D., Friedman, J., Chou, P.Y., Fasman, G.D., Finkelstein, A.V., Lim, V.I., Ptitsyn, O.B., Kabat, E.A., Wu, T.T., Levitt, M., Robson, B., Nagano, K. Comparison of predicted and experimentally determined secondary structure of adenyl kinase. Nature (London) 250:140–142, 1974.
2. Curtis, B.M., Presnell, S.R., Srinivasan, S., Sassenfeld, H., Klinke, R., Jeffrey, E., Cosman, D., March, C.J., Cohen, F.E. Experimental and theoretic studies of the 3-dimensional structure of human interleukin-4. Proteins Struct. Funct. Genet. 11:111–119, 1991.
3. Benner, S.A., Cohen, M.A., Gerloff, D. Correct structure prediction? Nature (London) 359:781, 1992.
4. Greer, J. Comparative modeling methods: Application to the family of the mammalian serine proteases. Proteins 7:317–334, 1990.
5. Ring, C.S., Cohen, F.E. Modeling protein structures—construction and their applications. FASEB J. 7:783–790, 1993.
6. Sali, A., Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234(3):779–815, 1993.
7. Summers, N.L., Karplus, M. Modeling of globular proteins. A distance-based data search procedure for the construction of insertion/deletion regions and Pro–non-Pro mutations. J. Mol. Biol. 216:991–1016, 1990.
8. Bowie, J.U., Luthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–169, 1991.
9. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through the folding motif. Proteins Struct. Funct. Genet. 16:92–112, 1993.
10. Godzik, A., Skolnick, J. Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. Proc. Natl. Acad. Sci. U.S.A. 89:12098–12102, 1992.
11. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature (London) 358:86–89, 1992.
12. Benner, S.A., Gerloff, D. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: A prediction of the structure of the catalytic domain of protein kinases. Adv. Enzyme Reg. 31:121–181, 1991.
13. Chou, P.Y., Fasman, G.D. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. 47:45–148, 1978.
14. Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J. Turn prediction in proteins using a pattern-matching approach. Biochemistry 25:266–275, 1986.
15. Cohen, F.E., Richmond, T.J., Richards, F.M. Protein folding: Evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. J. Mol. Biol. 132:275–288, 1979.
16. Crawford, I.P., Niermann, T., Kirschner, K. Prediction of secondary structure by evolutionary comparison: Application to the alpha subunit of tryptophan synthase. Proteins 2:118–129, 1987.
17. Garnier, J.R., Osguthorpe, D.J., Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
18. King, R.D., Sternberg, M.J. Machine learning approach for the prediction of protein secondary structure. J. Mol. Biol. 216:441–457, 1990.
19. Levitt, M., Warshel, A. Computer simulation of protein folding. Nature (London) 254:694–698, 1975.
20. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
21. Russell, R.B., Breed, J., Barton, G.J. Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. FEBS Lett. 304:15–20, 1992.
22. Smith, R.F., Smith, T.F. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc. Natl. Acad. Sci. U.S.A. 87:118–122, 1990.
23. Cohen, B.I., Presnell, S.R., Cohen, F.E. Origins of structural diversity within sequentially identical hexapeptides. Protein Sci. 2:2134–2145, 1993.
24. Levitt, M., Chothia, C. Structural patterns in globular proteins. Nature (London) 261:552–558, 1976.
25. Muskal, S.M., Kim, S.-H. Predicting protein secondary structure content. J. Mol. Biol. 225:713–727, 1992.

26. Nishikawa, K., Kubota, Y., Ooi, T. Classification of proteins into groups based on amino acid composition and *other characters*. I. Angular distribution. J. Biochem. 94: 981–995, 1983.
27. Sheridan, R.P., Dixon, J.S., Venkataraghavan, R. Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. Biopolymers 24: 1995–2023, 1985.
28. Richardson, J.S. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–339, 1981.
29. Harris, N.L., Presnell, S.R., Cohen, F.E. Four helix bundle diversity in globular proteins. J. Mol. Biol. 236:1356–1368, 1994.
30. Orengo, C.A., Flores, T.P., Taylor, W.R., Thornton, J.M. Identification and classification of protein fold families. Prot. Eng. 6:485–500, 1993.
31. Chothia, C., Murzin, A.G. New folds for all-beta proteins. Structure 1:217–222, 1993.
32. Schulz, G.E., Schirmer, R.H. "Evaluation of Prediction Methods. Principles of Protein Structure." New York: Springer-Verlag, 1979:122–128.
33. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
34. Cohen, F.E., Sternberg, M.J.E. On the prediction of protein structure: The significance of the root-mean-square deviation. J. Mol. Biol. 138:321–333, 1980.
35. Falicov, A., Cohen, F.E. Novel metric for structural comparison of proteins. J. Mol. Biol., in press.
36. Benner, S.A., Gerloff, D. Predicting the conformation of proteins. Man versus machine. FEBS Lett. 325:29–33, 1993.
37. Robson, B., Garnier, J.R. Protein structure prediction. Nature (London) 361:506, 1993.
38. Sander, C., Vriend, G. Results of class on protein design. 1992.
39. Emery, S.C., Fritz, H.J. Gene synthesis, expression and purification, 1994.
40. Schmid, F.X. Spectra of RORO, 1994.
41. Levin, J.M., Pascarella, S., Argos, P., Garnier, J.R. Quantification of secondary structure prediction improvement using multiple alignments. Prot. Eng. 6:849–854, 1993.
42. Richardson, J.S., Richardson, D.C. Amino acid preferences for specific locations at the ends of alpha helices. Science 240:1648–1652, 1988.
43. Harper, E.T., Rose, G.D. Helix stop signals in proteins and peptides: The capping box. Biochemistry 32:7605–7609, 1993.
44. Zhou, H.X., Lyu, P., Wemmer, D.E., Kallenbach, N.R. Alpha helix capping in synthetic model peptides by reciprocal side chain-main chain interactions: evidence for an N terminal "capping box". Proteins 18:1–7, 1994.
45. Cohen, F.E., Sternberg, M.J.E., Taylor, W.R. Analysis and prediction of protein β-sheet structures by a combinatorial approach. Nature (London) 285:378–382, 1980.
46. Cohen, F.E., Sternberg, M.J.E., Taylor, W.R. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. J. Mol. Biol. 156:821–862, 1982.
47. Ptitsyn, O.B., Rashin, A.A. A model of myoglobin self-organization. Biophys. Chem. 3:1–20, 1975.
48. Landschulz, W.H., Johnson, P.F., McKnight, S.L. The leucine zipper: A hypothetical structure common to a new class of DNA binding proteins. Science 240:1759–1764, 1988.
49. O'Shea, E.K., Rutkowski, R., Kim, P.S. Evidence that the leucine zipper is a coiled coil. Science 243:538–542, 1989.
50. Chan, H.S., Dill, K.A. Origins of structure in globular proteins. Proc. Natl. Acad. Sci. U.S.A. 74:4130–4134, 1990.
51. Kim, P.S., Bierzynski, A., Baldwin, R.L. A competing salt-bridge suppresses helix formation by the isolated C-peptide carboxylate of ribonuclease A. J. Mol. Biol. 162:187–199, 1982.
52. Gregoret, L.M., Cohen, F.E. Effect of packing density on chain conformation. J. Mol. Biol. 219:109–122, 1991.
53. Kabsch, W., Sander, C. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. U.S.A. 81:1075–1078, 1984.
54. Cohen, F.E., Kosen, P.A., Kuntz, I.D., Epstein, L.B., Ciardelli, T.L., Smith, K.A. Structure-activity studies of interleukin-2. Science 234:349–352, 1986.
55. Cohen, F.E., Sternberg, M.J.E. On the use of chemically derived distance constraints in the prediction of protein structure with myoglobin as an example. J. Mol. Biol. 137: 9–22, 1980.
56. Jin, L., Cohen, F.E., Wells, J.A. Structure from function: screening structural models with functional data. Proc. Natl. Acad. Sci. U.S.A. 91:113–117, 1994.
57. Richards, F.M., Kundrot, C.E. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 3:71–84, 1988.
58. Ferrin, T.E., Huang, C.C., Jarvis, L.E., Langridge, R. The MIDAS display system. J. Mol. Graphics 6:13–37, 1988.
59. Benner, S.A., Badcoe, I., Cohen, M.A., Gerloff, D.L. Bona fide prediction of aspects of protein conformation. J. Mol. Biol. 235:926–958, 1994.
60. Kang, H.S., Kurochkina, N., Lee, B.K. Estimation and use of protein backbone angle probabilities. J. Mol. Biol. 229: 448–460, 1993.
61. Kurochkina, N., Lee, B.K. Hydrophobic potential by pairwise surface area sum. Prot. Eng., in press.
62. Rodionov, M.A., Galaktionov, S.G. Analysis of the three-dimensional structure of proteins in terms of residue-residue contact matrices. II. Coordination numbers. Mol. Biol. 26:777–783, 1992.
63. Galaktionov, S.G., Marshall, G.R. Properties of intraglobular contacts in proteins: An approach to prediction of tertiary structure. In: "Proceedings of the 27th Hawaiian International Conference on Systems Sciences, Biotechnology Computing." Los Alamitos: IEEE Computer Society Press 1994: 5:326–335.
64. Avbelj, F., Moult, J. Role of electrostatic screening in determining protein main chain conformational preferences. Biochemistry 34:755–764, 1995.
65. Pedersen, J.T., Moult, J. Genetic algorithms in protein folding: An efficient, full atom representation torsion space algorithm for the minimization of global energy functions. Document in preparation.
66. Avbelj, F., Moult, J. Determination of the conformation of folding initiation sites in proteins by computer simulation. Proteins 23:129–141, 1995.
67. Hubbard, T.J. Use of b-strand interaction pseudo-potentials in protein structure prediction and modelling. In: "Proceedings of the Biotechnology Computing Track, Protein Structure Prediction MiniTrack of the 27th HICSS." Los Alamitos: IEEE Computer Society Press, 1994:336–354.
68. Leng, B., Buchanan, B.G., Nicholas, H.B. Protein secondary structure prediction using two-level case-based reasoning. J. Compt. Biol. 1:25–38, 1994.
69. Leng, B. A knowledge-based approach for predicting the internal structure of objects with two-level case-based reasoning. Ph.D. thesis, University of Pittsburgh, 1994.
70. Levin, J.M., Garnier, J.R. Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. Biochim. Biophys. Acta 955:283–285, 1988.
71. Biou, V., Gibrat, J.F., Levin, J.M., Robson, B., Garnier, J.R. Secondary structure prediction: Combination of three different methods. Protein Eng. 2:185–191, 1988.
72. DiFrancesco, V., Munson, P.J., Garnier, J.R. Use of multiple alignments in protein secondary structure prediction. In: "28th Hawaii International Conference on System Sciences." Los Alamitos: IEEE Computer Society Press, 1995: 5:285–291.
73. Munson, P.J., DiFrancesco, V., Porrelli, R. Protein secondary structure prediction using periodic-quadratic-logistic models: Theoretical and practical issues. In: "27th Annual Hawaii International Conference on System Sciences." Los Alamitos: IEEE Computer Society Press, 1994: 5:285–291.
74. Mekler, L.B., Idlis, R.G. The general stereochemical genetic code—the way to 21st-century biotechnology and universal medicine—already today. Priroda (Nature) the

monthly scientific journal of the Russian Academy of Science (in Russian; translation into English is available from the authors by request) 5:22–25, 1993.
75. Covell, D.G. Low resolution models of polypeptide chain collapse. J. Mol. Biol. 235:1032–1043, 1994.
76. Covell, D.G. Folding protein alpha-carbon chains into compact forms by Monte Carlo methods. Proteins 14:192–204, 1992.