Critical Reviews in Biochemistry and Molecular Biology, 30(1):1–94 (1995)

Protein Structure Prediction: Recognition of Primary, Secondary, and Tertiary Structural Features from Amino Acid Sequence

Frank Eisenhaber,[1,2] *Bengt Persson,*[2,3] *and Patrick Argos*[2]

[1]Institut für Biochemie der Charité, Medizinische Fakultät, Humboldt-Universität zu Berlin, Hessische Str. 3–4, D–10098 Berlin-Mitte, Fed. Rep. Germany; [2]European Molecular Biology Laboratory, Meyerhofstr. 1, Postfach 10.2209, D–69012 Heidelberg, Fed. Rep. Germany; [3]Dept. of Medical Biochemistry and Biophysics, Karolinska Institutet, S-17177 Stockholm, Sweden

Table of Contents

Address for correspondence: Mail European Molecular Biology Laboratory, Meyerhofstr.1, Postfach 10.2209, D-69012 Heidelberg, Fed. Rep. Germany. Tel. +49-6221-387-275 (or 452); Fax +49-6221-387-517 (or 306); E-mail Eisenhaber@EMBL-Heidelberg.DE, bengt.persson@mbb.ki.se, Argos@EMBL-Heidelberg.DE.

1040-9238/95/$.50

ABSTRACT: This review attempts a critical stock-taking of the current state of the science aimed at predicting structural features of proteins from their amino acid sequences. At the primary structure level, methods are considered for detection of remotely related sequences and for recognizing amino acid patterns to predict posttranslational modifications and binding sites. The techniques involving secondary structural features include prediction of secondary structure, membrane-spanning regions, and secondary structural class. At the tertiary structural level, methods for threading a sequence into a mainchain fold, homology modeling and assigning sequences to protein families with similar folds are discussed. A literature analysis suggests that, to date, threading techniques are not able to show their superiority over sequence pattern recognition methods. Recent progress in the state of *ab initio* structure calculation is reviewed in detail.

The analysis shows that many structural features can be predicted from the amino acid sequence much better than just a few years ago and with attendant utility in experimental research. Best prediction can be achieved for new protein sequences that can be assigned to well-studied protein families. For single sequences without homologues, the folding problem has not yet been solved.

KEY WORDS: protein structure prediction, sequence motif recognition; secondary structure prediction, prediction of transmembrane regions, prediction of secondary structural class, conformational energy, conformational search; reduced protein representation, threading, remotely related amino acid sequences, homology modeling, side chain placement.

I. INTRODUCTION

There is a Holy Grail in the protein structure wing of theoretical molecular biology, and it is ever sought. The chalice contains a secret and with its knowledge, the prediction of a protein's tertiary structure from only its amino acid sequence will be possible. The three-dimensional architecture of a protein represents the ultimate in molecular information, and from it springs many significant scientific results: the understanding of protein folding mechanisms, enzymatic catalysis, molecular structural stablity, and scaffolding and interactions among protein subunits, receptors, and ligands, substrates, and active-site pockets, and the like; the ability to design and engineer proteins through synthesis and mutation; comprehension of the evolutionary development of life; and the creation of drugs and utilization of protein processes to confront human disease and suffering. Alas, the Holy Grail is not at hand and theoretical molecular biology must settle for lesser goals as well as the quest, which is described and referenced in this review.

The treatise is divided into three parts for the separate consideration of prediction techniques designed for increasing complexity of structural organization: the primary, secondary, and tertiary structures. At the protein primary structural level, techniques are described to elicit subsequence consensus patterns, motifs, binding and modification sites, and profiles from multiple sequence alignments and to use them for recognition of even more distant protein familial constituents.

Certain secondary structural aspects can be predicted from the sequence alone, albeit not with 100% accuracy. These include membrane spanning subsequence regions; secondary structural segements such as helices, strands, or loops in soluble proteins, and the structural class characterized by the content and sequence successive order of secondary structural types.

On the tertiary-structural level, important present-day pursuits are in abundance. Procedures have been developed for the comparison and equivalencing of tertiary structures, thereby eliciting essential relationships, conservations, and structural and functional principles. It is possible to predict by computation the three-dimensional structure of sequences with reasonably strong homology to those for which the structure is experimentally available. Even mainchain topologies have been predicted for small proteins, largely helical, from the sequence alone. The methods range from complex and computer-intensive energy calculations to simplified models involving a few dominant folding forces. New techniques to test if a sequence can fit one of the known backbone folds have also lately appeared in great number, especially because there will be probably less than 10,000 mainchain folds possible for hundreds of thousands of sequences. Such methods may partly close the large and growing gap between the number of amino acid sequences and tertiary structures (Figure 1).

Although this treatise attempts to be comprehensive and critical in the research areas previously listed, it will, like the methods it describes, not be 100% successful in its thoroughness.

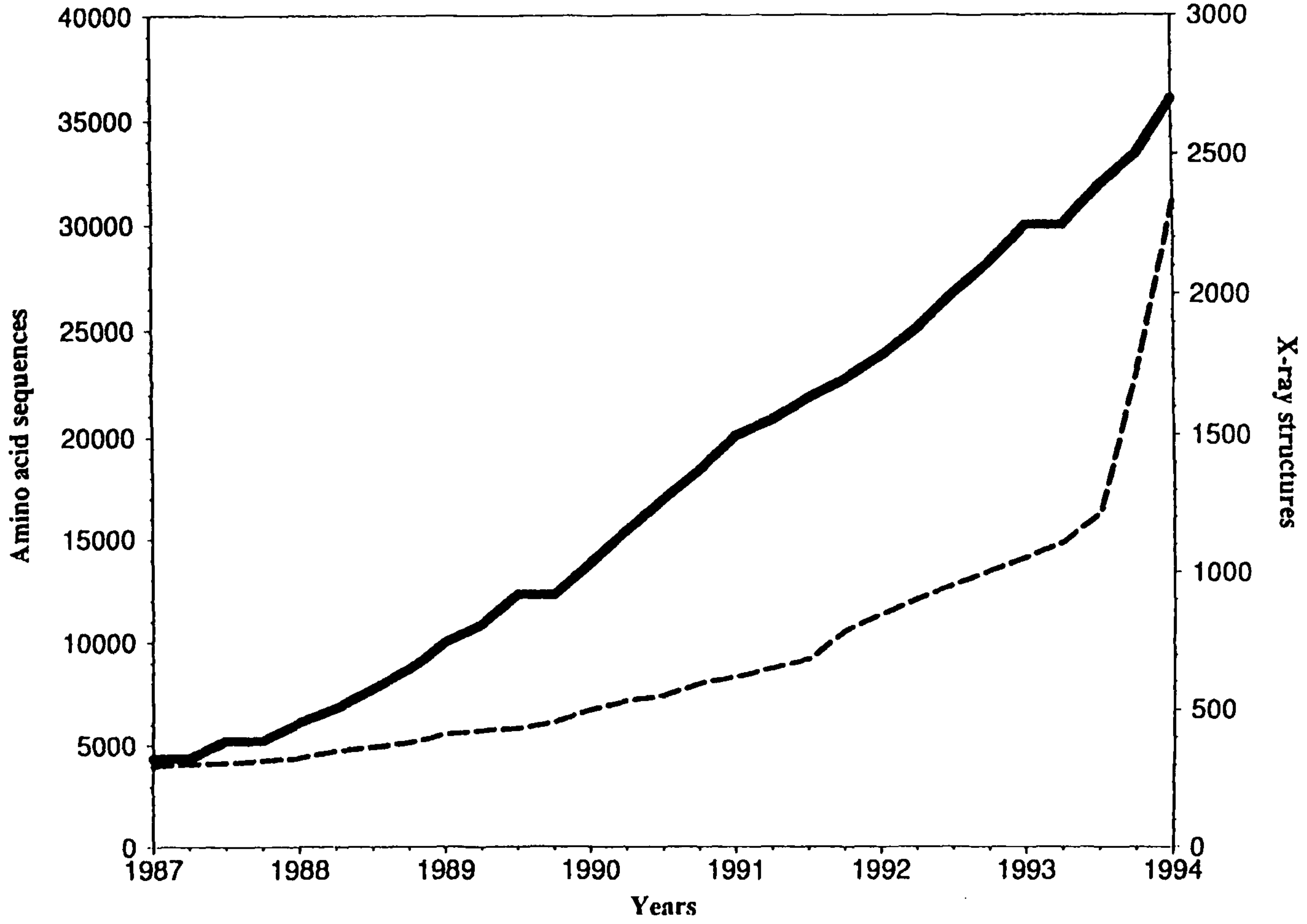


FIGURE 1. The gap between the number of protein sequences and structures. The graph shows the accumulation of protein amino acid sequences (bold line) in SWISS-PROT (Bairoch and Boeckmann, 1993) and of protein tertiary structures (dashed line) in the Brookhaven Protein Databank (PDB; [Bernstein *et al.*, 1977; Abola *et al.*, 1987]) since 1987. The data were provided by Karen Smith, and the graph was drawn by Dimitrij Frishman. The gap between structures and sequences is even more impressive if closely related homologous structures (e.g., mutants with only a few changed residues) are removed. After such a cleaning of the PDB, only a few hundred unique tertiary structures remain.

II. MOTIF RECOGNITION IN PRIMARY STRUCTURE

A. Prediction of Posttranslational Modifications

Many proteins are posttranslationally modified, necessary for a wide variety of reasons, for example, protection against proteolysis, directions of transport, genetic regulations, membrane anchoring, and regulation of degradation. The modifications can be N-terminal (e.g., acetylation, myristoylation, and pyroglutaminylation); C-terminal (e.g., amidation, isoprenylation, and farnesylation), or affect the side chains (e.g., glycosylation, phosphorylation, and hydroxylation).

Several of these modifications are guided by signals in the amino acid sequence. Knowledge about the determinative amino acid sequence patterns can therefore be utilized in the prediction of structural modifications. Such sequence patterns have been described for acetylation (Persson *et al.*, 1985); myristoylation (Hancock *et al.*, 1989; Resh, 1994); palmitoylation (Hancock *et al.*, 1989; Resh, 1994); phosphorylation (Kemp and Pearson, 1990); and glycosylation (Lis and Sharon, 1993). Further types of modification are treated in a recent review by Han and Martinage (1992).

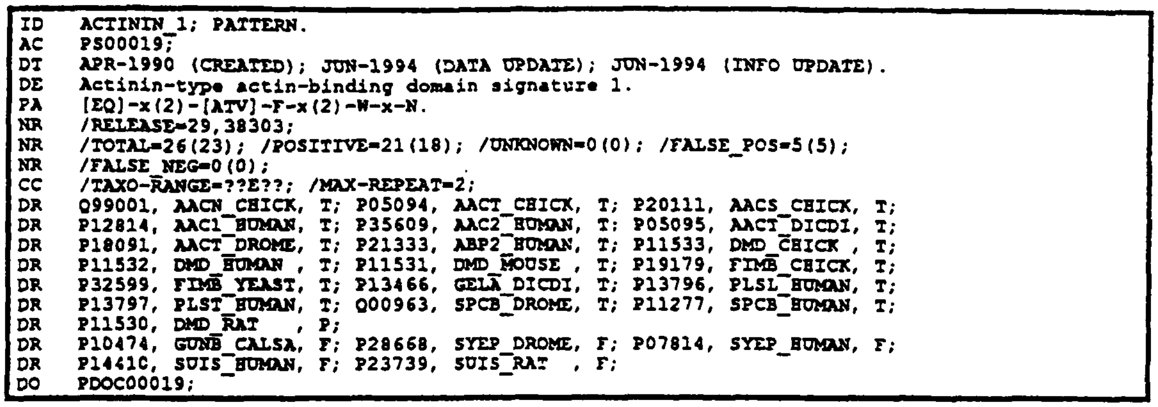
Some of the these characteristic sequence patterns are found in the sequence motif databases (*vide infra*), while others are not included there due to that additional factors (e.g., the secondary structure) are also determining the modificiation. Still further patterns have not yet been detected but remain to be elucidated.

B. Databases of Motifs and Consensus Patterns

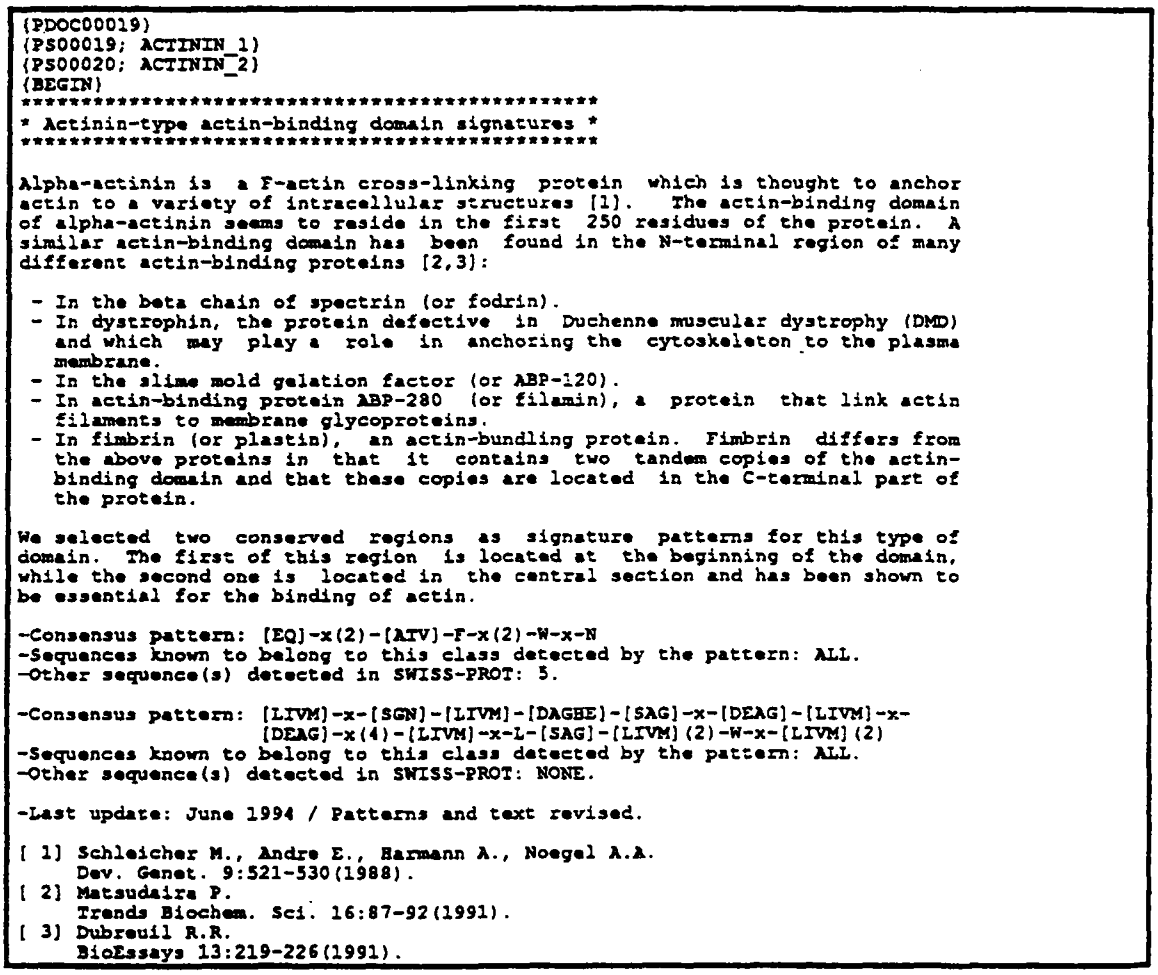
The use of protein sequence patterns to delineate protein function is widely popular. Similarly, a protein can be related to a family of sequences, thereby aiding with prediction of its secondary and tertiary structures. Several databases of patterns are available, of which three are described subsequently.

1. PROSITE

Since 1988, Amos Bairoch has compiled a database, PROSITE, of amino acid sequence sites and patterns found in proteins. The database consists of over 600 different patterns with short descriptions and key references (Bairoch, 1993); for example, the tyrosine kinase phosphorylation site, the protein kinases signature, the zinc-containing alcohol dehydrogenases pattern, or the dihydrofolate reductase signature. For each pattern, a list of proteins (SWISS-PROT identifiers) as well as false-positives containing the pattern is provided (Figure 2a). There is also an accompanying brief documentation, as exemplified in Figure 2b. The false-positives are known not to exhibit the corresponding function. In general, the patterns chosen are those with high specificity to keep the number of unrelated sequences displaying the same pattern low. The database has cross-references to the SWISS-PROT protein sequence database. PROSITE is available via anonymous ftp (Bairoch, 1993).



```
ID   ACTININ_1; PATTERN.
AC   PS00019;
DT   APR-1990 (CREATED); JUN-1994 (DATA UPDATE); JUN-1994 (INFO UPDATE).
DE   Actinin-type actin-binding domain signature 1.
PA   [EQ]-x(2)-[ATV]-F-x(2)-W-x-N.
NR   /RELEASE=29,38303;
NR   /TOTAL=26(23); /POSITIVE=21(18); /UNKNOWN=0(0); /FALSE_POS=5(5);
NR   /FALSE_NEG=0(0);
CC   /TAXO-RANGE=??E??; /MAX-REPEAT=2;
DR   Q99001, AACN_CHICK, T; P05094, AACT_CHICK, T; P20111, AACS_CHICK, T;
DR   P12814, AAC1_HUMAN, T; P35609, AAC2_HUMAN, T; P05095, AACT_DICDI, T;
DR   P18091, AACT_DROME, T; P21333, ABP2_HUMAN, T; P11533, DMD_CHICK , T;
DR   P11532, DMD_HUMAN , T; P11531, DMD_MOUSE , T; P19179, FIMB_CHICK, T;
DR   P32599, FIMB_YEAST, T; P13466, GELA_DICDI, T; P13796, PLSL_HUMAN, T;
DR   P13797, PLST_HUMAN, T; Q00963, SPCB_DROME, T; P11277, SPCB_HUMAN, T;
DR   P11530, DMD_RAT   , P;
DR   P10474, GUNB_CALSA, F; P28668, SYEP_DROME, F; P07814, SYEP_HUMAN, F;
DR   P1441C, SUIS_HUMAN, F; P23739, SUIS_RAT  , F;
DO   PDOC00019;
```



```
{PDOC00019}
{PS00019; ACTININ_1}
{PS00020; ACTININ_2}
{BEGIN}
**************************************************
* Actinin-type actin-binding domain signatures *
**************************************************

Alpha-actinin is  a F-actin cross-linking  protein  which is thought to anchor
actin to a variety of intracellular structures [1].   The actin-binding domain
of alpha-actinin seems to reside in the first  250 residues of the protein.  A
similar actin-binding domain has  been  found in the N-terminal region of many
different actin-binding proteins [2,3]:

 - In the beta chain of spectrin (or fodrin).
 - In dystrophin, the protein defective  in  Duchenne muscular dystrophy (DMD)
   and which  may  play a  role  in  anchoring the  cytoskeleton to the plasma
   membrane.
 - In the slime mold gelation factor (or ABP-120).
 - In actin-binding protein ABP-280  (or filamin), a  protein  that link actin
   filaments to membrane glycoproteins.
 - In fimbrin (or plastin),  an actin-bundling protein.  Fimbrin  differs from
   the above proteins in  that  it  contains  two  tandem copies of the actin-
   binding domain and that these copies are located  in the C-terminal part of
   the protein.

We selected  two  conserved  regions  as  signature  patterns for this type of
domain.  The first of this region  is located at  the beginning of the domain,
while the second one is  located in  the central section and has been shown to
be essential for the binding of actin.

-Consensus pattern: [EQ]-x(2)-[ATV]-F-x(2)-W-x-N
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in SWISS-PROT: 5.

-Consensus pattern: [LIVM]-x-[SGN]-[LIVM]-[DAGHE]-[SAG]-x-[DEAG]-[LIVM]-x-
                    [DEAG]-x(4)-[LIVM]-x-L-[SAG]-[LIVM](2)-W-x-[LIVM](2)
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in SWISS-PROT: NONE.

-Last update: June 1994 / Patterns and text revised.

[ 1] Schleicher M., Andre E., Harmann A., Noegel A.A.
     Dev. Genet. 9:521-530(1988).
[ 2] Matsudaira P.
     Trends Biochem. Sci. 16:87-92(1991).
[ 3] Dubreuil R.R.
     BioEssays 13:219-226(1991).
```

FIGURE 2. Sequence pattern databases. An example entry of the PROSITE (a) and PROSITEDOC (b) databanks presenting the ACTININ_1 pattern. Two-letter identifiers at the beginning of each line of the PROSITE entry facilitate computer access: ID, entry name; AC, accession number; DT, update information; DO, reference to PROSITEDOC; DE, full name of the pattern; PA; pattern description (x denotes any amino acid); NR, numerical result of SWISS-PROT database search; DR, identifiers of proteins containing the pattern (true (T), potential (P), or false (F) hits); CC, taxonomic range.

2. PRINTS

Another databank, called PRINTS, of amino acid sequence patterns has been collected by Attwood and Beck (1994). This database contains sets of motifs excised from conserved regions of multiple familial sequence alignments, thus reflecting the protein's "fingerprint". This is advantageous, because amino acid patterns typical of a certain protein type might be distributed along the protein chain, but nearly always the distinctive subpatterns are located in the same order along the primary structure for all members. Therefore, a search for related sequences can be more sensitive than one over a single-pattern database such as PROSITE. However, the shorter and fewer the patterns, the more distant are the familial members detected. The version of April 1994 contains about 200 entries. The entries are cross-referenced to the PROSITE database, which makes a combined use of the different pattern collections more facile for the user.

3. ProDom

Sonnhammer and Kahn (1994) have undertaken an automated apporach to detect and collect amino acid sequence patterns. They have developed a method to systematically search the protein sequence databases and automatically identify protein domains from sequence comparisons. These domains are subsequently clustered into families, from which the consensus sequence can be deduced from the multiple aligned primary structures. In the clustering phase, the automated routine does not allow insertion of gaps in the alignments to simplify the matching procedure and to avoid unreliable alignments resulting from a given selection of gap penalties. The consensus subsequences can be used as motif or pattern in database searches to find further family members. The applicability of the method is exemplified by showing two previously unrecognized domain arrangements detected with the new procedure. The database is called ProDom and has 5765 entries (edition 21.0) for protein families. It is available via anonymous ftp and regular updates are planned (Sonnhammer and Kahn, 1994).

C. Antigenic Motifs

An important step in the biochemical characterization of a protein is the detection of antigenic sites responsible for specific antibody binding. Prediction algorithms attempt to locate antigenic sites indirectly as hydrophilic and malleable loops at the protein surface.

Several algorithms for prediction of these sites have been developed. One of the first methods was that of Hopp and Woods (1981), who calculated averaged hydrophilicity values in segments along the protein chain. This method is still one of the most popular and has a high reliablity for prediction of antigenic determinants (Hopp, 1993). In general, the highest peak is in the vicinity of one of the antigenic sites.

An alternative set of hydrophilicity values has been used by Parker *et al.* (1986), resulting in a higher success rate than the Hopp and Woods method. Jameson and Wolf (1988) combined linearly in a weighted manner hydrophilicity, surface accessibility (Janin *et al.*, 1978), flexibility (Karplus and Schultz, 1985), and secondary structure prediction for turn (Chou and Fasman, 1978; Garnier *et al.*, 1978), producing a commonly used measure named the antigenic index.

Neural network techniques (cf. Section III.B.4 and IV.D) have been used to predict the surface exposure of amino acids (Holbrook *et al.*, 1990; Rost and Sander, 1994b). This is of interest in the estimation of antigenic sites and other molecular interactions. The method is able to predict correctly the surface exposure for 72% of the residues using a two-state evaluation model (accessible or not accessible). More accurate predictions can be achieved by making a consensus prediction of exposure for a homologous family, but the residue accessibility is not as well conserved as, for example, the secondary structure (Flores *et al.*, 1993).

Antigenic epitopes are also known as mutation hotspots. Frömmel (1988) has quantified the relationship between antigenicity and mutability and showed that it is possible to successfully predict antigenic sites of proteins using the averaged local mutability of amino acids.

D. Profile Methods

One strategy to predict the structure of a protein is to find a related protein for which the tertiary structure is known. A sensitive approach to solve this problem is to use the sequence profile method. The profile consists of a scoring table in which each position is assigned 20 scores reflecting the probability for each amino acid residue to be in that position as deduced from alignments of a set of known familial sequences (Gribskov *et al.*, 1987; Gribskov *et al.*, 1990).

A few years ago, an extension of the profile method was described (Bowie *et al.*, 1991) by defining a profile of tertiary structure environments for each amino acid residue in a protein sequence, using information from known X-ray structures. Tertiary structure profiles were constructed by first assigning each amino acid residue to one of 18 environmental classes, according to the area of the side-chain that is buried by other atoms, the fraction of the side-chain area that is covered by polar atoms or water, and the local secondary structure. Second, the 20 scores for that environmental class is assigned to that position of the profile. These residue scores for the different environmental classes have been calculated from well-refined tertiary structures. Subsequently, the profiles can be utilized to search for related proteins having similar fold. By aligning the sequences with the tertiary profile, it is possible to extract those that might adopt a similar tertiary structure. This has been tested for different protein families (i.e., globins, cAMP-receptor-like proteins, periplasmic binding proteins, and actins [Bowie *et al.*, 1991]). The tertiary structure profile method arrives at better results than the traditional profile method, which only considers the primary structure.

A similar approach is taken by Overington *et al.* (1992), who have calculated environment-specific amino acid substitution tables according to amino acid type, accessibility, side-chain interactions, and main-chain conformation. The substitution tables were generated by accumulating substitutions observed in homologous structures. They show that using these tables in template-based database searches gives more reliable results than using only sequence-based techniques.

Recently, the sensitivity of the sequence profile method has been improved (Lüthy *et al.*, 1994; Henikoff and Henikoff, 1994). This is achieved by inclusion of sequence weights and amino acid substitution weight tables in the calculations. The sequence weights reflect the distance of the sequences calculated according to Sibbald and Argos (1990) and inclusion of these avoids the bias toward many closely related sequences. Furthermore, three different amino acid substitution tables were tested, of which the most sensitive was found to be BLOSUM45 (Henikoff and Henikoff, 1992). The total increased sensitivity is exemplified by application to four different protein families.

Another recent improvement of profile searches by weighting of sequences and the use of alternative substition matrices has been described (Thompson *et al.*, 1994). In addition, all alignment positions with the density of gaps higher than a threshold were excluded from the profile. By these means, the sensitivity was improved as demonstrated by application of the technique to four different protein families.

III. PREDICTION OF SECONDARY STRUCTURAL FEATURES

A. Introduction

An intermediate goal along the way to predict the three-dimensional structure is the assignment of secondary structure elements to segments of the amino acid sequence. It was found early on that different amino acid residue types have different preponderance for particular secondary structural environments. This established the basis for the so-called statistical methods (treated in Section III.B.1), normally having a success rate of about 50 to 60%. Improvements have been made by utilization of information in multiple sequence alignments (Section III.B.2) or consideration of secondary structural motifs (Section III.B.3). The methodology of neural networks has also been adopted for this task (Section III.B.4). The success rate is now at best about 70%, which is still too low for allowing any detailed conclusions regarding the tertiary structure to be drawn but nevertheless of importance as a rough guideline. The statistics for evaluating the success rate are naturally dependent on which judgment method is utilized and this is discussed further in Section III.B.5.

Prediction of secondary structural features has also been adapted to the problem of finding membrane-spanning segments in proteins. Here the task seems somewhat easier, because these segments are supposed to be hydrophobic, and the methods consequently directed toward identification of these regions. However, the number of known three-dimensional structures for membrane proteins is still small, and thus there probably exist several not yet discovered types of architecture not considered by present prediction methods. A wide variety of prediction techniques are available, some of which are described in Section III.C.

Finally, prediction schemes have also been developed to assign the structural class of a protein from the amino acid sequence alone. This task is treated in Section III.D.

B. Prediction of Secondary Structure

1. Statistical Methods for Single Sequences

Several secondary structure prediction methods are based on the principle that different amino acid residue types have different likelihoods to occur in various secondary structural conformations. The exemplary techniques of Chou and Fasman and Garnier *et al.* are briefly described here.

Propensity values representing the statistical occurrence of the 20 common amino acid residues in three structural states (α-helix, β-sheet, and coil) have been calculated (Chou and Fasman, 1974b). The preferences for each amino acid type were defined as the ratio of the fractional occurrence of the residue in secondary structure elements of a given type to the fractional occurrence in all structures. These values were computed from the occurrences of the secondary structural elements in 15 X-ray crystallography determined structures (Chou and Fasman, 1974b). A procedure for secondary structure predictions was developed that is based on the sequential successive appearance of large propensity values (Chou and Fasman, 1974a). For α helices and β strands, respectively, each of the amino acids was categorized as formers, breakers, or indifferent. These properties were used to identify potential α- and β-forming sites, which were then extended along the protein chain as long as the average propensity values calculated over a window of 5 or 6 residues were above a threshold value. Turns were predicted by considering the probabilities of different amino acids at the various positions in a β-turn.

The method was reported to predict successfully 80% of the helical and 86% of the β-sheet residues in the 19 proteins evaluated (Chou and Fasman, 1974a). The combined accuracy for α, β, and coil was 77%. However, these values are biased since most of the proteins predicted (15 out of 19) were included in the set from which the propensity parameters were derived. The values are also dependent on which scoring scheme is used (cf. Section III.B.5). For a set of 62 proteins, the Chou-Fasman algorithm is reported to achieve only about 50% overall prediction success using a three-state accuracy measure (Kabsch and Sander, 1983a).

The GOR III method (Garnier *et al.*,1978; Gibrat *et al.*, 1987) is a representative of methods based not only on single residue propensities but also on statistically significant pairwise residue interactions. The preference (information content) I of a residue with sequence number j and amino acid type R_j for a secondary structure type $Z \in \{\text{helix, sheet, coil}\}$ is approximated as

$$I\left(S_j = Z;\ R_{j-8},\ldots,R_{j+8}\right) = \sum_{m=-8}^{8} I\left(S_j Z;\ R_{j+m}\middle|R_j\right)$$

in a sequence environment of eight residues on either side of a central one. The information I carried by the amino acid pair $(R_{j+m}|R_j)$ on the occurrence of the event Z (adoption of a specific secondary structural state) is defined as

$$I\left(S_j = Z;\ R_{j+m}\middle|R_j\right) = \log\left[P\left(Z\middle|\left(R_{j+m}\middle|R_j\right)\right)\middle/P(Z)\right]$$

where P denotes the conditional probability. The enormous amount of parameters (3 structural states × 20 amino acid types × 20 amino acid types × 17 sequence positions) is estimated from a set of 68 nonredundant protein crystallographic structures. The prediction accuracy achieved was about 63% (Gibrat *et al.*, 1987; Garnier and Levin, 1991). A further improvement of 2.5 to 6.5% (Biou *et al.*, 1988) was attained by combining the GORIII method with two other prediction schemes: one based on hydrophobicity patterns that are often observed in regular secondary structures (bit pattern method), and the other using structural similarity between short, sequentially homologous peptides (Levin and Garnier, 1988). Both GORIII alone and COMBI are still widely in use and often as reference methods. As was shown by Gibrat *et al.*, (1991), the predictive power of methods relying on only sequentially local structure information is limited by about 65%. A further increase in prediction accuracy requires the consideration of long-range (tertiary) interactions.

The knowledge that a residue might not adopt a secondary structural state may be as useful as a corresponding positive statement. The impact of a combination of several statistical prediction methods (Chou-Fasman, GOR) together with hydrophobicity information and helix capping rules was shown to be significant for negative predictions (Arnold *et al.*, 1992).

It seems that all methods of this type, which predict the secondary structure based on the information in only a single sequence (more accurately based on the local sequence environment of the residue under consideration), do not achieve higher three-state accuracy than about 60%. In order to get significantly beyond this level, more information is required, for instance, knowledge derived from related sequences.

2. Predictions with Multiple Alignments

The extension of the algorithms to base the prediction on multiple sequence alignments of related sequences has been shown to improve accuracy. This approach has been implemented in several ways. One is to make the predictions for each sequence separately and thereafter calculate the consensus prediction, which is the strategy of Levin *et al.* (1993). Multiple alignments make it possible to distinguish characteristic patterns, which is used in the method of Donnelly *et al.* (1994). There are also methods based on neural networks (see Section III.B.4).

Levin *et al.* (1993) showed that the use of multiple aligned sequences can improve secondary structure prediction accuracy by approximately 8%. The quality of the alignments naturally affects the results such that the sequences should have at least 25% pairwise residue identity to assure improvement in prediction accuracy. The study used two statistical secondary structure prediction methods (Gibrat *et al.*, 1987; Levin and Garnier, 1988), which, when applied to single sequences, gave on the average an accuracy of 61.5%, defined as number of residues correctly predicted in the three states (as observed crystallographically) divided by the total number of residues in the protein. With multiple sequence alignments, obtained by spatial superposition of the known tertiary structures, a predictive accuracy of 69.5% was achieved (Levin *et al.*, 1993), while alignments from an automatic alignment procedure gave 68.5%. For multiple sequences, the prediction taken was that in the majority for the two prediction techniques and over the residues at a given alignment position.

Wako and Blundell (1994b), who used a three-step method for secondary structure predictions from multiple alignments, achieved 69% prediction accuracy for three-state (α, β, coil). The first step involves evaluation of mean propensities and amino acid substitution patterns for four conformational states: α, β, buried coil, and exposed coil. Capping rules are also applied to define secondary structure boundaries. The second step concerns detection of patterns characteristic for α-helix or β-strand. In the third step, the alignment is considered, so if one conformational state is present at an alignment site in more than a given fraction of the proteins, this state is reassigned to all residues at that site. The authors applied the method to 13 protein families, representing different folding types, and found the prediction accuracy to range between 60 and 79% with an average of 69%. Though this figure is similar to that achieved by neural networks (see Section III.B.4), the method of Wako and Blundell (1994a) has the advantage that it is clear to follow and can be inspected during all stages of the prediction procedure.

Geourjon and Deléage (1994) have described a self-optimized method for secondary structure prediction based on multiple alignments. It too achieves a three-state accuracy of 69%. First, a limited database is created of proteins with known secondary structures and similar to the query protein sequence. Second, the secondary structures are predicted for the proteins in the database, whereupon the results are evaluated and the prediction parameters are adjusted. This step is repeated until the predictions do not improve anymore. Third, these parameters are applied to the protein to be predicted. The method was evaluated using 239 proteins with known tertiary structure and with less than 50% pairwise residue identity. It was thus found that the three-state prediction accuracy on average was 69%.

An improved method for prediction of α-helices has been described by Donnelly *et al.* (1994). It also relies on multiple sequence alignments. They used the characteristic periodicities of solvent accessible/inaccessible residues to predict α-helices through a Fourier transform procedure. The predictions were subsequently refined by examining for sequence patterns that identify the N- and C-termini of helices. The authors report that this combined method correctly predicts 79% of the α-helix residues, while the number of false-positives only amounts to 12%. Inasmuch as the technique is dependent on sequence variations in a family of related proteins, the authors claim that ideally the proteins should have 30 to 65% pairwise residue identity. The method also predicts the internal face of each helix, and these results can be used to postulate three-dimensional arrangements of α-helices.

3. Secondary Structure Motifs

Amino acid sequence motifs predictive for secondary structure have been investigated by Rooman and Wodak (1988). They have systematically searched the protein tertiary structure database for sequence patterns involving two or three residues within a segment of maximally 8 residues. It was observed that several patterns have predictive power, even though identical short peptides can adopt different secondary structure formations (Kabsch and Sander, 1984; Argos, 1987). It was also shown that the number of predictive two-residue patterns is limited to about 70. The small size of the present structure database limited results for patterns consisting of three and more residues (Rooman and Wodak, 1988). A similar approach was used by Shestopalov (1990). He identified some di-, tri-, and pentapeptides that have predictive power for the secondary structural states.

Presnell and co-workers (1992) have developed sequence patterns to predict helices in all-helix proteins, extending the methods of Cohen and co-workers (1986) who identified the location of turns between helices. Three different sets of patterns were developed to recognize the amino-terminal, core, and carboxy-terminal portions of an α-helix. These patterns were subsequently used to identify putative turn segments (Cohen *et al.*, 1986). The strong α-helix signals are first evaluated, while the weak signals are considered in a subsequent analysis. Using a residue scoring scheme, this method has a 71% success rate, which is comparable with the accuracy of the GOR algorithm (Garnier *et al.*, 1978). However, if only considering the core segments, the algorithm achieves 78% compared with 75% for the GOR method (Presnell *et al.*, 1992).

4. Neural Networks

A neural network mimics the architecture of neurons in the human brain. It consists of a number of simple, connected computational units that take signals from other units such that, if the sum of the signals is above a threshold, a signal is passed onto further units. The methodological advantage with a neural network is that it is suitable to detect subtle patterns and correlations in data. Each node has an attached weight for its signaling power. The weights are determined by training on a set data. In fact, parameters in a complex, nonlinear recognition function are fitted to given input and output conditions (Figure 3). Subsequently, the network can be used, for instance, in predictions of secondary structures of proteins. Several different applications of neural networks to the study of proteins and nucleic acids have been reviewed recently (Hirst and Sternberg, 1992).

Qian and Sejnowski (1988) presented a neural network method for prediction of secondary structures in single protein sequences trained with set of 106 tertiary structures. They achieved a success rate of 64.3% for a three-state model (α, β, coil). This accuracy, being reproducibly obtained by other researchers also (e.g., Holley and Karplus [1989]), is substantially better than the prediction accuracy from statistical approaches (Chou and Fasman, 1974a; Lim, 1974; Garnier *et al.*, 1978) in the 50 to 56% range (Kabsch and Sander, 1983a). The test set contained only proteins not homologous to those in the training set.

A variant of the neural network, the so-called hybrid system, has been developed by Zhang and co-workers (1992). Three subsystems are used: a neural network module, a statistical module, and a memory-based reasoning module. The output from these three modules are fed into a "combiner", which makes the final predictions. The three subsystems are individually

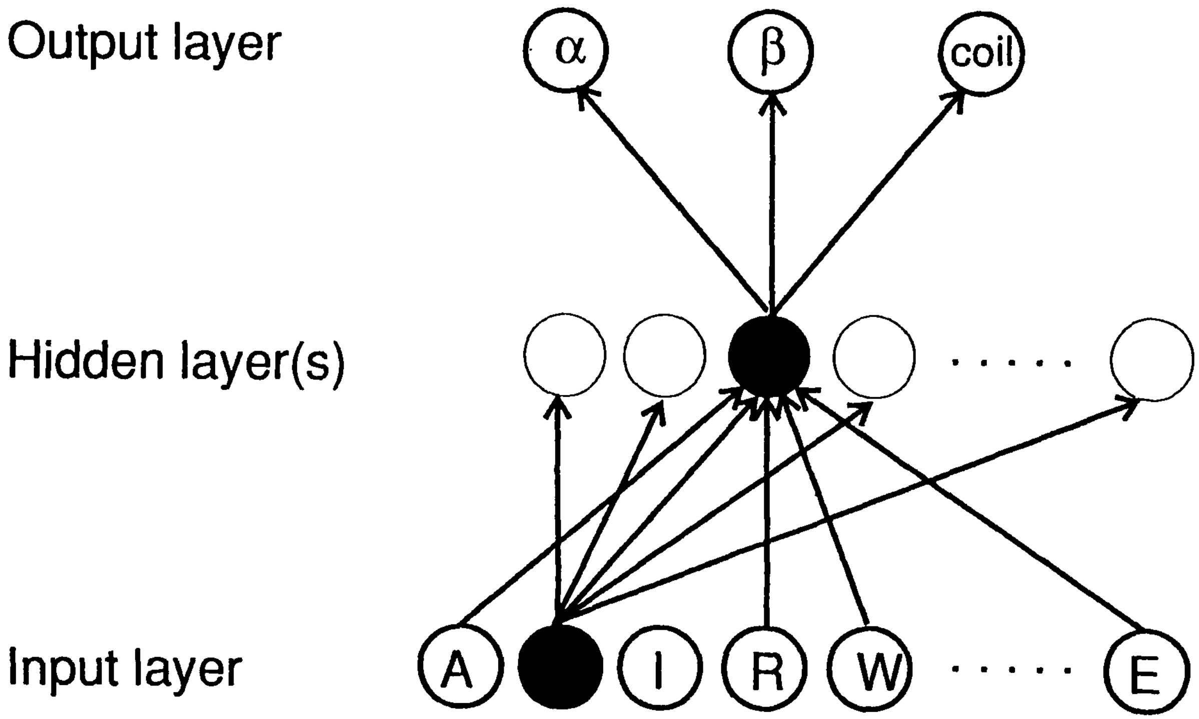


FIGURE 3. Neural network technique. The basic units of a neural network are nodes grouped into layers. Each node j of the layer i accepts input $S_{j'}^{(i-1)}$ from the nodes j′ of the previous layer i–1 and transmits a signal $S_j^{(1)}$ to the nodes j″ of the next layer i + 1 such that

$$S_{j'}^{(i)} = f\left(\sum_{j'} w(j',j) S_{j'}^{(i-1)}\right)$$

The weights w(j′,j) are adjustable parameters fitted to given pairs of input and output values for the whole network. The function f is usually chosen as a sigmoidal step function, for example, *Tangens hyperbolicus*. In principle, the neural network can be considered as a complex nonlinear function of multidimensional input vectors. The figure shows a hypothetical neural network with input, hidden, and output layer. For clarity, only the connection to one hidden node and from one input node (both units indicated in grey) are shown in full detail.

trained on test sets. The method was checked over 107 protein structures, divided into 8 different sets to allow some cross-validation. A prediction accuracy of 66.4% was achieved for the three-state model (α, β, and coil). It was also observed that for 20% of the residues all three subsystems gave the same incorrect predictions, indicating a possible limit of accuracy for secondary structure predictions due to non-local interactions. Similar techniques have been applied for the calculation of distance tables that can be used for secondary structure prediction (Salzberg and Cost, 1992).

Muggleton *et al.* (1992; 1993) have used neural networks in the prediction of all-helix domains. With a learning set of 12 nonhomologous proteins, they achieve 81% accuracy (three-state model) for four different proteins. McGregor and coworkers (1989) used the technique to predict β-turns, giving an overall improvement of more than 5 percentage points over the Chou-Fasman algorithm. However, for the type I and II turns, the Chou-Fasman method is more successful — 79 and 86% vs. 69 and 81%, respectively. An approach considering both α, β and turn simultaneously would increase the success rate, where, for instance, a weak turn prediction would be discarded if the same segment has a strong prediction for an α-helix (McGregor *et al.*, 1989).

Rost and Sander (1994a) have described a prediction method based on neural networks and multiple sequence alignments. This network system has an accuracy of 71.6%, albeit the cross-validation test on 126 proteins is limited because the training of neural networks is computationally very expensive. The success rate is similar to that of Levin *et al.* (1993), which also depends on multiple alignment information. The method of Rost and Sander (1994a) calculates a position-specific reliability index, giving hints regarding which of the predicted regions can be considered with more confidence. It has been shown by Yi and Lander (1993) that for a subset of residues the secondary structural state can be predicted with very high probability (86 for 28% of the residues).

Neural networks (Bohr *et al.*, 1990) have also been applied to tertiary structure predictions by homology (see Sections IV.C and IV.D), but so far that technique seems not to be as successful as the traditional methods.

5. Evaluation of Secondary Structure Prediction Methods

When comparing the results of different secondary structure prediction methods, the method used for evaluation of the results is significant. The most commonly used measure for secondary structure prediction accuracy is the three-state residue-by-residue score giving the percentage of correct predictions, Q, defined as

$$Q = \frac{(p(\alpha) + p(\beta) + p(\text{coil}))}{N}$$

where p is the prediction accuracy (number of residues predicted correctly) for a given secondary structural type and N is the total number of residues. However, this measure is insensitive to the distribution of prediction in terms of secondary structural type, and, more importantly, it does not take overprediction into account. A more meaningful scale is Matthews' correlation coefficient (Kneller *et al.*, 1990). It is defined as

$$Q(s) = \frac{p(s)n(s) - u(s)o(s)}{\sqrt{(n(s)+u(s))(n(s)+o(s))(p(s)+u(s))(p(s)+o(s))}}$$

for a particular secondary structural type s where p(s) is the number of residues properly predicted, n(s) is the number of residues correctly identified as not being in secondary structural state s, u(s) is the number underpredicted, and o(s) the number overpredicted. The coefficient ranges from 1.0 (perfect prediction) to –1.0 (completely anticorrelated).

When comparing the location of the secondary structure elements in homologous proteins with known three-dimensional structure, the three-state (helix, strand, coil) residue-by-residue score does not achieve even 100% (Jenny and Benner, 1994). Naturally, the more deviant the structures, the further from 100% are the scores. The "expected" maximum value could be calculated and should be taken into account when judging secondary structure prediction methods (Russel and Barton, 1993). A further problem is that there is no universally accepted way to assign an experimental secondary structure to a set of crystallographic data (Colloc'h *et al.*, 1993; Jenny and Benner, 1994), even though many researchers use the DSSP algorithm for achieving some type of "standard-of-truth" (Kabsch and Sander, 1983b).

The "reliability" of these types of automatic secondary structure assigment methods was investigated by Colloc'h and coworkers (1993). They compared three approaches, DSSP (Kabsch and Sander, 1983b), P-Curve (Sklenar *et al.*, 1989), and Define (Richards and Kundrot, 1988). They found that only in 63% of the residues the assignments coincide. However, in most of the cases of non-agreement, one or two of the methods deviated by not assigning α or β to a structure identified as α or β by the other techniques. This can be explained by the particularities of each method. The DSSP approach considers hydrogen bond patterns, while the P-Curve algorithm finds regularities along the helicoidal axis and the Define technique measures distances between Cα atoms. Therefore, when evalutating predictions, the "standard-of-truth" might vary depending on which property was used for the secondary structure assignment.

However, it may sometimes be sufficient to predict the approximate locations of helices, strands, and loops (Russel and Barton, 1993). It may also be more important to predict the correct number of secondary structure elements rather than the exact borders of these elements. Recently, a new method to evaluate secondary structure predictions was suggested by Rost *et al.* (1994). They define a measure of segment overlap, weighing down minor deviations at the ends of secondary structure segments. This measure gives values around 90% for homologous protein pairs, while random protein pairs only achieve around 37%.

Several researchers (Qian and Sejnowski, 1988; Salzberg and Cost, 1992) have experimented with balancing the learning set of structural examples for deriving the prediction rule (even portions of sequence pieces exhibiting coil, helix, and strand). The differences in prediction accuracies for coil, helix, and strand (about 70%, 65%, and 40% in methods using single sequences only) have usually been attributed to the uneven influence of non-local interactions stabilizing these structures. Rost and Sander (1994c) reported that this might not be the case. If the learning set of structural examples for a network is balanced, then the prediction accuracy for all three structural states is about 60% using only single sequence information.

The lengths of the predicted structural elements are often too short compared with the real helices and strands. Several methods have been published to smooth the string of residue secondary structure propensities obtained in the prediction process. Rost and Sander (1994c) apply a second level network for analyzing the propensities predicted by the first-level networks for a window of 13 residues. They achieve almost a doubling of the predicted helix length compared with the original (single-level) secondary structure prediction. Various analytical smoothing techniques have been reviewed by Zimmermann (1994). It must be emphasized that these regularization methods do not affect the overall prediction accuracy but redistribute the structural states making the prediction more "protein-like".

6. Conclusions

Because secondary structure prediction methods are based on sequence-specific information in the close vicinity of the residue to be predicted, only local interactions are considered. The secondary structure is also dependent on distant interactions, and thus these methods cannot be expected to achieve 100% accuracy. Most of the statistical methods have a prediction accuracy of 50 to 65% when based on single sequences. However, inclusion of mulitple alignments, reflecting the "allowed" variations in the sequences (and indirectly some distant interactions), improves the accuracy to around 70%. Interestingly, it seems that it is the multiple alignment per se that is important for this improvement, because regardless of which method is used, all prediction schemes arrive at a similar degree of accuracy.

Further improvement of secondary structure prediction can only be achieved by influencing the local decision (secondary structural state of a given residue) with more information. Possible candidates are environmental states (Yi and Lander, 1993), sequence length (number of residues, domains), amino acid (Kneller *et al.*, 1990; Rost and Sander, 1994c), and dipeptide composition (see Sections III.D 3 and IV.D). Tertiary interactions are very important for the length of a secondary structural element. For example, it was demonstrated (Hayward and Collins, 1992) that neural networks may predict longer helices than observed in the crystal structure. The authors concluded that a potentially longer helix is disrupted by global constraints.

Extension of secondary structure predictions methods to other structural states than only helix, strand, and coil is another possible research task. A likely candidate is the extended structure of PPII type. This conformation frequently occurs in globular proteins (Adzhubei and Sternberg, 1993) in the form of stretches a few residues long. The PPII conformation is tightly related to the unfolding of proteins (Moore and Fasman, 1993).

C. Prediction of Membrane-Spanning Regions

1. Introduction

Membrane proteins fulfill several important tasks in the cells, for example, hormone receptors, transport proteins, photosynthetic reaction center. Because they are hydrophobically bound to the membrane, they are difficult to purify and to crystallize. Therefore, there are still only few three-dimensional structures available. Due to this and to the medical importance of these molecules, many methods have been developed to predict the membrane-spanning parts of proteins. It is believed that the membrane-spanning parts in general consist of helices perpendicular to the membrane plane. The structures of bacteriorhodopsin (Henderson *et al.*, 1990) and reaction center (Deisenhofer and Michel, 1989) support this view. However, there also exists another type of membrane proteins, the porins (Weiss and Schulz, 1992), which have a build-up consisting of 16 β-sheets arranged as a barrel, giving rise to a big central hole. Recent three-dimensional structures also show further variants. The helices can be tilted against the perpendicular plane, like the case in light harvesting complex (Kühlbrandt *et al.*, 1994). Presence of helices parallel to the membrane plane has also been shown (Kühlbrandt *et al.*, 1994; Picot *et al.*, 1994). There are also indications that the membrane-spanning segments can consist of single β-strands, thus making it possible to spawn the membrane with fewer residues than in the case of an α-helix (Hucho *et al.*, 1994). All this implies that prediction of membrane-spanning regions might be a more difficult task than has hitherto been anticipated.

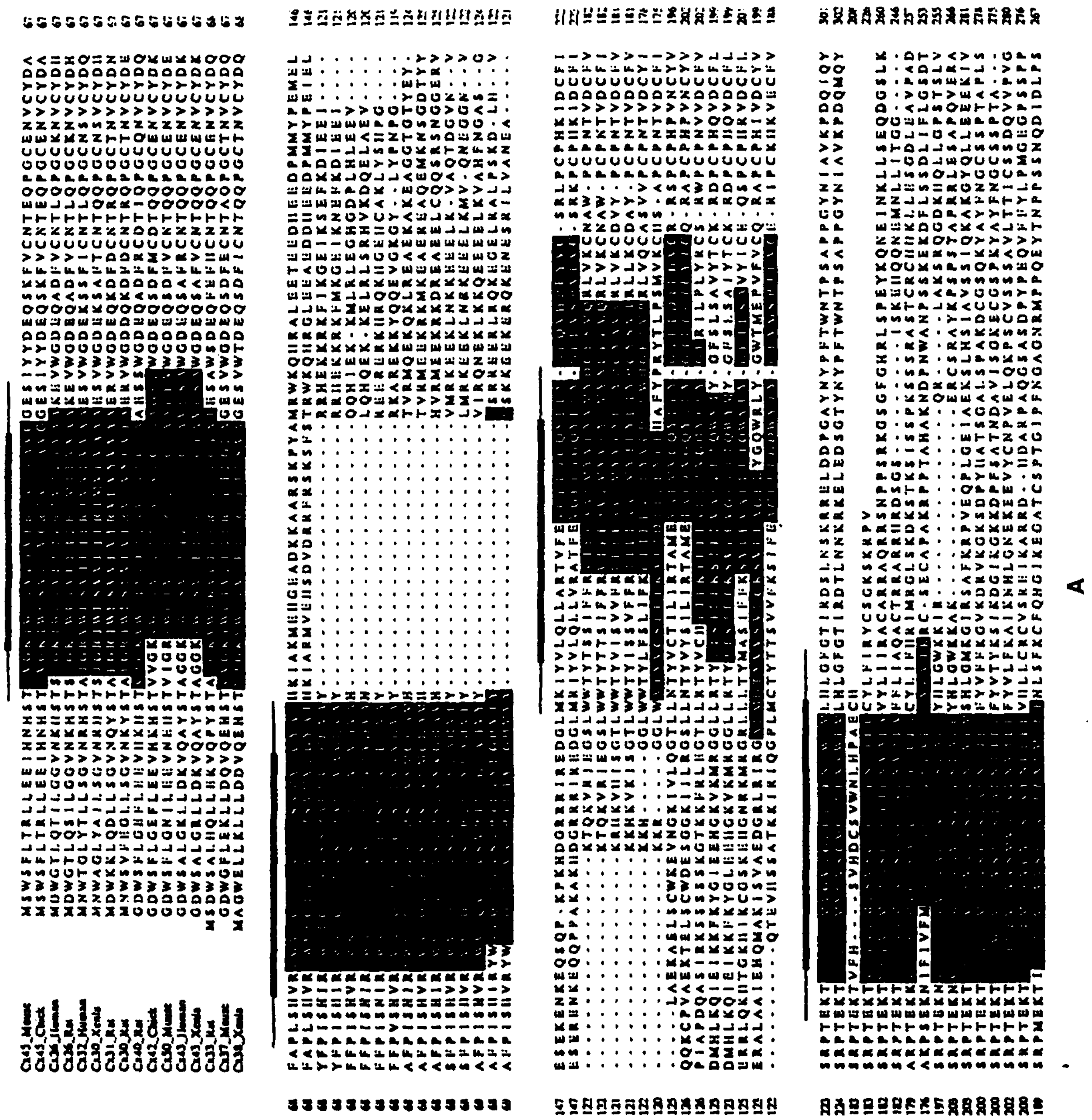
Most available methods are designed to identify hydrophobic stretches of around 20 amino acid residues, corresponding to an α-helix traversing the membrane. This might still be a common structural theme (or the most common theme) in membrane proteins, because total compositions of putative transmembrane segments, as found in the SWISS-PROT database, in general, display a segment of 21 consecutive hydrophobic residues, corresponding to the hydrophobic part of the membrane (Persson and Argos, 1994).

2. Hydrophobicity Methods

All methods for prediction of membrane-spanning segments have in common that they identify hydrophobic stretches in the sequence (Figure 4b). This is often graphically presented as a profile of hydrophobicity against the amino acid position along the sequence. One of the most widely used methods for prediction of membrane-spanning segments was invented by Kyte and Doolittle (1982), where mean residue hydrophobicity values were calculated for consecutive 19-residue sequence spans. Segments with hydrophobicity above a certain threshold were predicted to be membrane-spanning. A similar approach was adopted by Rao and Argos (1986), who also considered residues that break the transmembrane helices to improve reliability.

Different prediction methods were reviewed and evaluated by Degli Esposti and co-workers (1990). They examined the correlations among the various amino acid hydrophobicity scales used and compared the accuracy of the various prediction approaches. They also calculated a new set of parameters derived from seven different hydrophobicity scales.

Von Heijne (1986; 1992) used a trapezoidal sliding window in his hydrophobic analysis of the sequence together with a consideration of positively charged residues interior to the membrane. Several studies have also been effected regarding helix-helix interactions in membrane proteins (Lemmon and Engelmann, 1992).


Cx45_Mouse
Cx45_Chick
Cx36_Human
Cx36_Rat
Cx32_Human
Cx30_Xenla
Cx31_Rat
Cx30_Rat
Cx40_Rat
Cx42_Chick
Cx50_Mouse
Cx43_Human
Cx43_Xenla
Cx33_Rat
Cx37_Mouse
Cx38_Xenla
A

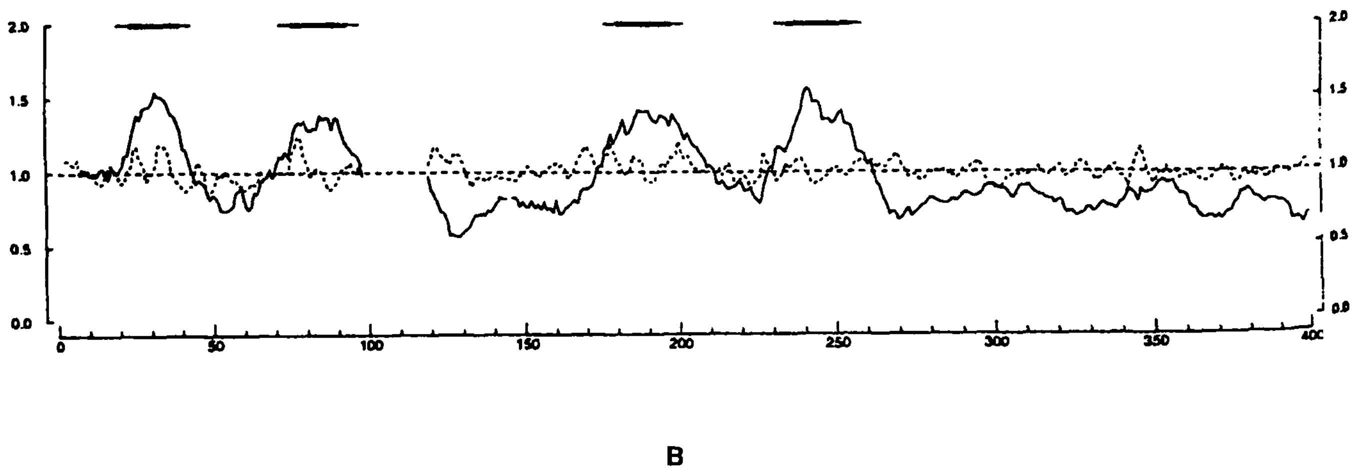


FIGURE 4. Prediction of transmembrane regions utilizing multiple alignments. (**A**) Multiple alignment of connexins. Residues shown in white against a black background indicate putative transmembrane regions according to SWISS-PROT annotations; the data are not consistent for all members of the protein family. Bars above the alignment indicate the predicted transmembrane segments utilizing the method discussed in Persson and Argos (1994). Residue numbers are indicated at the beginning and at the end of each line for each sequence. Gaps are shown by dashes. The figure exhibits the first 304 alignment positions. The protein codes correspond to SWISS-PROT entries. (**B**) Plots of mean propensity values (for the hydrophobic region as continuous line and for the flanking four-residue extensions at the both termini as dashed line) are shown against the alignment position j for the sequences in **A**. The four predicted segments are indicated by horizontal bars, where the thick lines correspond to the central parts of the transmembrane segments and the thin lines to the flanking parts (Persson and Argos, 1994).

3. Prediction of Membrane Protein Topology

Methods have been developed to predict how the protein is inserted into the membrane, that is, which parts of the protein face the inside and which face outward. The most successful of these prediction schemes is that so-called "positive-inside rule" of von Heijne (1986; 1992). He showed that the loops between membrane-spanning segments have a preponderance for positively charged residues on the inside (normally facing the cytosol) of the membrane. This was shown first for prokaryotic (von Heijne, 1986) and later also to be valid for several eukaryotic proteins (Sipos and von Heijne, 1993).

4. Utilization of Multiple Sequence Alignments

Similarly to secondary structure prediction, improvements can be made by inclusion of the extended information in a multiple sequence alignment of related proteins (Figure 4). This has been utilized in a recent algorithm for prediction of transmembrane (Persson and Argos, 1994). This technique has higher accuracy in predicting transmembrane segments than previous methods based on individual sequences. Present primary structural databases are large and expanding at such a rate that homologous sequences are often found making this approach often applicable.

D. Prediction of Secondary Structural Class from Amino Acid Sequence

1. Introduction

Early in 1976, when only about 40 crystallographic structures of proteins were known, Levitt and Chothia (1976) studied the succession of secondary structural elements along the amino acid sequence. Intuitively, they grouped the proteins into four structural classes (or folding types):

- All-α proteins having only α-helix secondary structural elements (more than 60% of the residues adopt helical conformation; no residues in β-strands)
- All-β proteins consisting mainly of (often antiparallel) β-strands;
- α + β proteins having independent clusters of α-helices and (often antiparallel) β-strands in the sequence
- α/β proteins with mixed (often alternating) segments of α-helix and (mostly parallel) β-strands.

Many more protein structures are known today, and, for an increasing number, it is not easy to classify them in accordance with the definitions of Levitt and Chothia (1976). For example, both the acylphosphatase, PDB entry 1APS (Pastore *et al.*, 1992), and the B chain of the regulatory domain of the aspartate carbamoyltransferase, PDB entry 8ATC (Stevens *et al.*, 1990), have a two-layered structure consisting of an antiparallel β-sheet and two parallel α-helices (Figure 5). The existence of an antiparallel sheet is characteristic for α+ β structures. However, a more detailed investigation of the structures reveals a high degree of secondary structural alternation βαββαβ and a doublet of the βαβ motif, both observations pointing to class α/β.

Nevertheless, the concept of structural class, based on the secondary structural content of the protein and the directionality of β-strands, is very useful from the experimental as well as the theoretical point of view.

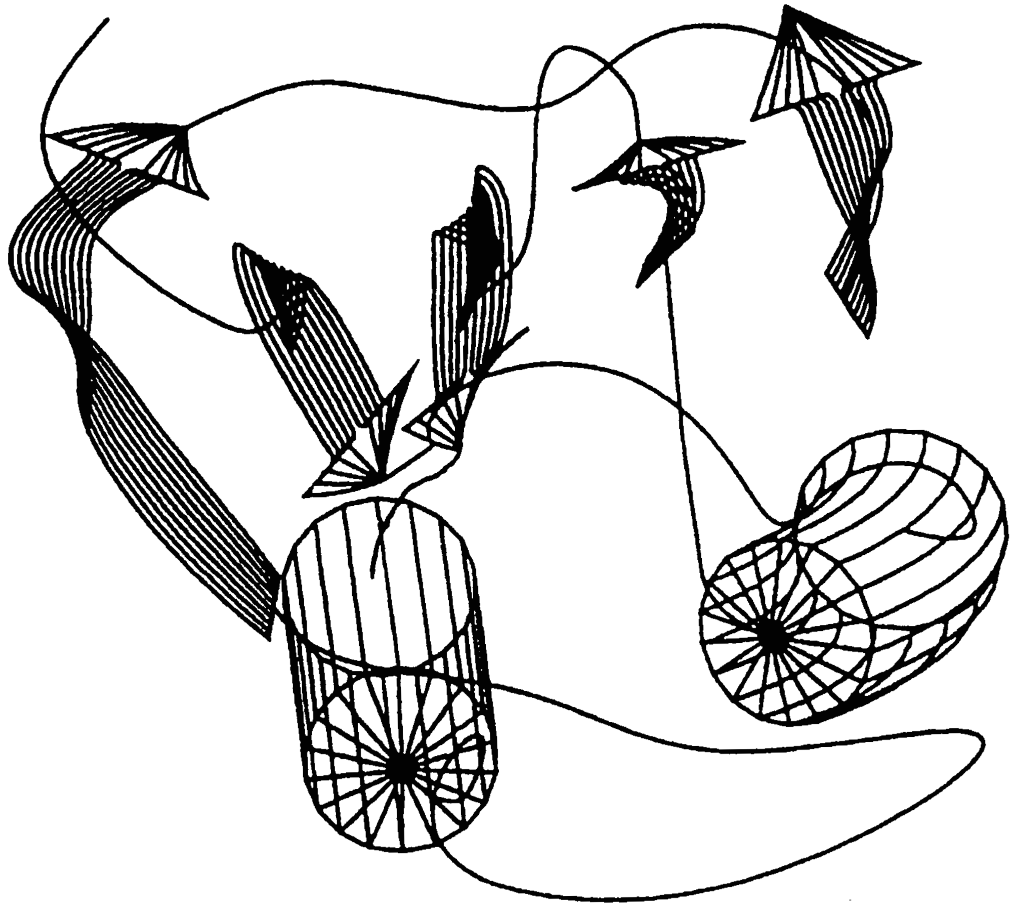


FIGURE 5. Packing in acylphosphatase. The sequence and packing of secondary structural elements in acylphosphatase (PDB entry 1APS, Pastore *et al.* [1992]) are shown. The complete sequence of secondary structures in this protein is βαββαββ. Cylinders denote α-helices, and flat arrows represent β-strands. The plot was produced with the package WHATIF (Gerrit Vriend) with a graphical option developed by David Thomas. The βαβ elements usually observed in α/β proteins form an antiparallel β-sheet characteristic of α+ β domains.

The folding type of a protein can be directly determined by relatively simple spectroscopic methods. With a sufficient quantity of the protein available, circular dichroism (CD) spectroscopy in the UV absorption range can be used to obtain reliable measures of secondary structure content, especially for α-helices, but also for parallel and antiparallel β-strands (Johnson, Jr. 1990; Perczel *et al.,* 1991; Sreerama and Woody, 1994). Similar information can be obtained from IR Raman spectroscopy (Bussian and Sander, 1989).

Secondary structural class restrictions have a high impact for secondary and tertiary structure prediction. The accuracy of secondary structure prediction from the amino acid sequence with methods designed for all α proteins is larger than 80% compared with maximally ~70% in the general case (Kneller *et al.,* 1990; Muggleton *et al.,* 1992; Muggleton *et al.,* 1993). The effect of knowledge of structural class alone in improving secondary structure prediction is comparable with the use of the extra information contained in multiple alignments of homologous sequences (Levin *et al.,* 1993; Rost and Sander, 1993). Class-dependent turn prediction is about 90% accurate (Cohen *et al.,* 1986; Cohen *et al.,* 1991). The secondary structural class is related to various properties of a protein such as its location in extra- or intracellular compartments, biological function (being an enzyme or not), or the existence of disulfide bonds (Nishikawa and Ooi, 1982; Nishikawa *et al.,* 1983a; Nishikawa *et al.,* 1983b).

Historically, the finding of Nishikawa *et al.* (Nishikawa and Ooi, 1982; Nishikawa *et al.,* 1983a) that structural classes of proteins correlate strongly with amino acid composition marked the onset of algorithm developments aimed at predicting the structural class of a protein from the amino acid sequence alone. The basic characteristics of the various methods are

1. The selection of a learning set of protein structures and their classification into folding types (Section III.D.2)
2. The proposal of a criterion taken from the primary structure to recognize the structural class of a protein not contained in the learning set (Section III.D.3)
3. A check of the prediction accuracy resulting from the selected classification criterion (Section III.D.4).

In the following, the efficiency of different approaches described in literature is reviewed.

2. Folding Type Classification of a Learning Set of Protein Structures

The learning set of structures generally consists of proteins with non-homologous sequences. The secondary structure assignments given in the Brookhaven Protein Databank entries (Bernstein *et al.,* 1977; Abola *et al.,* 1987) by crystallographers might be subjective; therefore, a computer algorithm such as DSSP (Kabsch and Sander, 1983b), DEFINE (Richards and Kundrot, 1988), or P-CURVES (Sklenar *et al.,* 1989) are mostly used for objectification. These methods have been critically reviewed (Colloc'h *et al.,* 1993).

As the set of available protein structures has enlarged, it has become difficult to define quantitatively the borderlines between the structural classes (Richardson, 1981; Nakashima *et al.,* 1986) as increas-

TABLE 1
Definition of Secondary Structural Classes of Proteins[a]

Class	Definition		Refs.
	α-content	β-content	
All-α	>15%	<10%	Nishikawa and Ooi, 1982
All-β	<10%	>15%	
Mixed	>15%	>10%	
Irregular	<15%	<15%	
All-α[c]	$\alpha > \beta$	$\alpha < \beta$	Sheridan *et al.*, 1985
all-β[c]	$\alpha < \beta$	$\alpha > \beta$	
Parallel[c]	With parallel β-sheet		
Irregular	With >4.5% Cys		
All-α	>40%	<5%	Klein and DeLisi, 1986
all-β	<10%	>30%	
Mixed[b]	≥15%	≥15%	
Irregular		$\alpha + \beta$ <20%	
All-α[d]	≥30%	≥0.15 · ($\alpha + \beta$)	Kneller *et al.*, 1990
All-β[d]	≤10%		
Mixed[b,d]	>15%	>5%	
Irregular	All remaining proteins		

[a] The definitions are given in terms of secondary structural content in percent.

[b] Mixed domains are classified as α/β or $\alpha + \beta$ depending on the existence of parallel β-sheet or a "sufficient" alternation of secondary structural types (Kneller *et al.*, 1990).

[c] A smooth transition between α- and β-structure is considered. Mixed domains always contain parallel sheets.

[d] The authors define all regular proteins as having a minimal length.

ingly more structures populate transition regions. A somewhat improved resolution is achieved if, in the case of large proteins, only their individual folding domains (sometimes within a single polypeptide chain) are considered (Sheridan *et al.*, 1985). It is clear that sufficiently large proteins will probably contain both α-helices, and β-sheets and, consequently, the sample would be biased toward mixed proteins.

The quantitative criteria utilized to classify folding types in terms of secondary structural content seem to a certain extent arbitrary (see Table 1). The criteria of Nishikawa *et al.* (1983b) are most often used. Many authors merge the two mixed classes with α- and β-structure into one. Sometimes, a fifth class ζ of irregular proteins is added that is comprised mainly of small proteins (<100 residues) with a low content of both α-helix and β-sheet and with many disulfide bridges and/or metal ion-mediated links. Further research in this area should apply cluster analysis to justify the presently accepted classification. It would also be useful to study the criterion of alternation frequency of α- and β-structure along the sequence to distinguish between α + β and α/β folding types instead of reliance on the existence of parallel β-sheet.

3. Criteria for Folding Type Recognition from Amino Acid Sequence

In addition to differences in amino acid composition, it might be expected that certain patterns of hydrophobic amino acids could have predictive power to distinguish structural classes. Studies of such patterns include the occurrences of hydrophobic and hydrophilic stretches of 4 or more amino acids (Sheridan *et al.*, 1985), the periodicity of hydrophobic residues in amphiphatic α- and β-structure (Klein and DeLisi, 1986; Klein, 1986; Metfessel *et al.*, 1993), and the hydrophobic moment vector for a subsequence window over which helical or extended structure is assumed (Sheridan *et al.*, 1985; Klein and DeLisi, 1986). In only two reports is the effect of inclusion of pattern search into the prediction algorithm investigated (Klein, 1986; Kneller *et al.*, 1990), both concluding that simple hydrophobic patterns add almost nothing in prediction accuracy compared with the criterion of amino acid composition alone.

Busetta and Barans (1982; 1984) classify proteins into structural classes with a function based on amino acid secondary structural propensities. Interestingly, different scales for parallel and antiparallel sheets have been used. The prediction accuracy of their method is low, suffering from the same deficiencies as propensity-based secondary structure predictions (see Section III.B).

Having for each protein a vector of attributes such as the fraction of each amino acid type or the frequency of pattern occurrence, the structural class can be represented by a cloud of points in a multidimensional attribute space. A new protein sequence could be assigned to a folding type according to the shortest distance from its attribute vector to the center of the cloud. Early approaches follow this immediate idea with Euclidean (Nakashima *et al.*, 1986) or Minkowski's (Chou, 1989) distance definitions and achieve a prediction accuracy in the range of 70 to 80%. Two authors published three papers only to show that instead of Euclidean distance, the maximal component of the attribute vector along the vectors pointing to the cloud centers (Zhang and Chou, 1992b) the smallest angle between two of these vectors (Chou and Zhang, 1993), or the correlation coefficient for such a pair of vectors (Chou and Zhang, 1992) may be used. It is not surprising that the prediction accuracy does not change significantly because of the mathematical equivalence of all four approaches. In another paper by the same authors, it was demonstrated that, assuming a normal distribution of points around the center of the structural class (more generally a multidimensional spherical density distribution), Euclidean metrics resulted in better predictions than Minkowski´s distance definition (Zhang and Chou, 1992a); indeed, the latter is more suitable for polyhedral geometry. The cloud can be even better approximated by a multidimensional ellipsoid that takes into account the relative size and shape of the point distribution. In this technique, the distance to the new sequence attribute point is scaled in accordance with each of the main axes of the hyperellipsoid (Mao *et al.*, 1994). In an earlier work (Genfa *et al.*, 1992), weighting coefficients for the fraction of each amino acid type were computed by a linear programming procedure. This approach is essentially equivalent to the hyperellipsoidal scaling of the distance (Mao *et al.*, 1994), if the ellipsoid main axes are fixed to the coordinate axes of the attribute space. The similar effectiveness of both methods implies that correlations between the fractions of different amino acid types do not greatly influence the structural class prediction. The distribution of the cloud in the attribute space can also be taken into account by defining a potential of attraction to each of its points (Sheridan *et al.*, 1985).

In contrast to the relatively simple analytical methods just described, more complicated approaches such as statistical discriminant analysis (Klein and DeLisi, 1986; Klein, 1986) or different types of neural networks (Metfessel *et al.*, 1993) have been applied to the structural class prediction problem. Also, the secondary structural content (Muskal and Kim, 1992; Pancoska *et al.*, 1992; Rost and Sander, 1994a) was estimated with a neural network using amino acid composition data as input. Nonetheless, the prediction accuracy for the structural class achieved with these techniques is not better (and sometimes even lower) than that of the analytical methods. In the case of neural networks, the computational costs for training the network is considerable. It is also scientifically disappointing that no physical explanation for the possible prediction success is afforded by the latter technique.

The dipeptide composition, rather than the amino acid composition, was used as input to neural networks for prediction of structural class combined with folding motif prediction (out of 45 variants) in a hierarchical way (Reczko *et al.*, 1994; Reczko and Bohr, 1994). The prediction accuracy was reported to be enhanced compared with pure structural class prediction. This effect might be explained by the fact that, in the latter method, the query sequence is compared with several specific fold patterns memorized in the neural net compared with direct approaches in which the neural nets store only a single, less detailed structural class pattern.

4. Accuracy of Structural Class Prediction

The prediction accuracies reported by different authors vary from 40 to 100% and are difficult to compare. The prediction success depends on the size and selection of members in the learning set and of the test set of structures. Klein (1986) notes that his method "generally" assigns the structural class correctly with a probability of 83%, while, for a set of 27 unseen examples, only 63% of the predictions were free from error. Both the learning set and the test set should contain as many structurally unrelated protein domains as possible to be sure that the sparsely populated regions of the attribute space are comprehensively checked. The structure selections may also be biased by the fact that certain groups of proteins do not readily crystallize. Given the results of different researchers, a modern algorithm for structural class prediction has a prediction accuracy of more than 80% for α- and β-type proteins and of more than 70% for α+ β- and α/β-classes (Genfa *et al.*, 1992; Metfessel *et al.*, 1993; Mao *et al.*, 1994). Currently, the cloud approximation by a multidimensional ellipsoid (Mao *et al.*, 1994) is the method of choice given its high prediction accuracy of about 96% for a 132 protein set (Nakashima *et al.*, 1986) and algorithmic simplicity.

Structural class prediction techniques have been applied to a set of 1490 human protein sequences (Chou and Zhang, 1993). The authors predict mixed proteins, especially α/β proteins, to predominate in the set of human proteins. This result may be biased not only by the fact that mostly enzymes have been sequenced but also by the aforementioned note that if large proteins are not divided into domains the sample is expected to overrepresent mixed structures.

It must be emphasized that the distributions of the five folding classes (all α, all β, α + β, α/β, and ζ) are not well separated in amino acid composition space, but show a considerable overlap. Although additional independent parameters other than sequence composition could increase discrimination, the overlap may indicate that transitions between structural classes are not abrupt and that the folding type is determined only by a more complete set of sequence properties and environmental conditions, at least for some sequences. In conclusion, the classification in accordance with the traditional five structural classes might also not be optimal. Several authors (Holm *et al.*, 1992; Orengo *et al.*, 1993) who attempted to achieve complete family classifications for presently known sequences had serious difficulties in distinguishing the two mixed structural classes from each other and from the other two classes containing mainly one type of secondary structure. A possibly complete enumeration of packing variants for secondary structural elements has been published by Efimov (Efimov, 1994).

IV. PREDICTION OF TERTIARY STRUCTURE

A. Computation of Tertiary Structure with Molecular-Mechanical Methods Based on Fundamental Physical Principles

1. The Hypothesis on the Native Structure as a Minimum of Free Energy

The fundamental physical approach to the protein folding problem (predicting the tertiary fold from the amino acid sequence alone) is based on the assumption that the native protein structure corresponds to a system at thermodynamic equilibrium with a minimum of free energy. *In vitro* renaturation experiments strongly support this view (Anfinsen, 1973; Creighton, 1992) because they imply that the complete information necessary for protein folding is contained in the amino acid sequence. Therefore, it would be sufficient to compute an ensemble of conformations representative for the state of lowest free energy. The conformational invariants of this ensemble (e.g., the densely packed protein core) are the characteristics of the native structure. In a more simplified approach, a unique conformation with the lowest sum of intramolecular potential energy, conformational entropy term, and solvation-free energy is considered to represent the native state (Sali *et al.,* 1994).

Recently, the important role of biological factors such as the peptidyl-prolyl-isomerase, disulfide isomerase, and molecular chaperonins in controlling the kinetics of protein folding and subunit assembly has been discovered (Gething and Sambrook, 1992; Hartl, 1994). Being similar to enzymes in their mode of action, these proteins reduce energetic barriers of conformational transitions and protect folding intermediates from aggregation. Although proteins that assist folding might be considered as a source of external information, at the moment, this experimental evidence does not force rejection of the hypothesis in which the native state is a free energy minimum of the system "protein-solvent environment".

It was demonstrated that the computational problem of finding the lowest energy conformation of a polypeptide chain from an energy function containing pairwise terms and possibly other expressions is NP-complete (Ngo and Marks, 1992; Unger and Moult, 1994; Fraenkel, 1993). The contention of Levinthal (1968) that proteins search only a tiny fraction of the conformational space and move into the lowest kinetically accessible free energy minimum appears much more likely in this context. First, experimental evidence in support of this view has been provided recently. The α-lytic protease was shown to exist in two forms: an inactive, metastable intermediate and an active native structure. Both conformations are separated by a barrier with activation energy of about 27 kcal/mol. A catalyst that is normally covalently attached to the protein is necessary to complete folding of the intermediate "molten globule", a less compact state compared with the native conformation (Baker *et al.,* 1992a; Baker *et al.,* 1992b). Metabolic energy was found to be necessary for accurate folding, for correct disulfide bond formation and for maintaining influenza hemagglutinin in its oligomerization-competent state (Braakman *et al.,* 1992). In the case of serpins, a metastable kinetically trapped five-stranded β-sheet conformation was found that only slowly rearranges to the native six-stranded form (Mottonen *et al.,* 1992).

To rescue the idea that native structures reach free energy minima, evolutionary arguments (Abagyan, 1993; Unger and Moult, 1994) are taken into consideration. In accordance with the "weak thermodynamic" hypothesis of Unger and Moult (1994), an originally functional structure of a protein corresponding to a local minimum will drift toward the global free energy minimum due to the combined effect of random mutations and the constant selective pressure of evolution. Hiding the native structure behind a large energy barrier may also be a sophisticated variant of enzyme activity regulation. A recently published opinion states that only those sequences fold into unique conformations, fold fast and fold via the all-or-none transition (with the release of substantial latent heat) that have a pronounced energy minimum that is sufficiently distinguished from all other conformational states in the energy spectrum (Sali *et al.,* 1994). The energy landscape for folding sequences might have guiding forces ("the folding funnel") that avoid the Levinthal paradox (Leopold *et al.*, 1992; Gulukota and Wolynes, 1994).

In conclusion, the search for low-energy conformations of polypeptide chains is still a promising plan for prediction of 3-D structure and function. The attempt to calculate the native structure from an extended polypeptide conformation is often called *ab initio* folding. Two prerequisites for this approach are necessary:

- An energy function for discriminating the native fold from other conformations
- A procedure for efficient searching of the conformational space

Both aspects are considered separately in the following sections. Although 3 decades of enormous scientific efforts have been concentrated on the protein folding problem, a complete solution to compute the structure from sequence is not yet in sight.

2. The Molecular-Mechanical Energy Function

Conformational analysis in organic chemistry has a quantum-chemical direction and a molecular-mechanical branch.

TABLE 2
Components of the Energy in the System "Protein–Solvent Environment"

Intramolecular energy (*in vacuo*)
- **Covalent terms**
  - Closure and disruption of chemical bonds (e.g., disulfide bonds)
  - Strain in chemical structure
    - Bond stretching
    - Valence angle bending
    - Improper (plane) torsions
  - Proper (dihedral) torsions
- **Nonvalent terms**
  - Pairwise terms
    - van der Waals interactions
    - Hydrogen bond energy
    - Coulomb electrostatics of atomic charges and dipoles
  - Conformational entropy term
    - Main chain conformational entropy
    - Side chain conformational entropy

Solvation energy
- **Long-range part**
  - Volume-related term
    - Polarization of various dielectric media due to protein charges
- **Short-range part**
  - Surface-related term
    - Cavity formation
    - Solute-solvent dispersion interactions
    - Solvent structure changes near the surface of the solute

Inasmuch as typical problems in protein structure involve many hundreds or even thousands of atoms, only molecular-mechanical methods with further simplifications are applicable. At the same time, one has to be aware of the severe approximations necessary for macromolecular simulations that make some calculated structural details unreliable (Brooks *et al.*, 1983).

To date, the molecular-mechanical energy function is considered to be a sum of several terms (see Table 2). The *in vacuo* energy components have been modeled in various atomic force fields such as ECEPP (Nemethy *et al.*, 1983; Nemethy *et al.*, 1992), GROMOS (van Gunsteren and Berendsen, 1977; van Gunsteren and Berendsen, 1990), CHARMM (Brooks *et al.*, 1983), AMBER (Weiner *et al.*, 1986), DISCOVER (Dauber-Osguthorpe *et al.*, 1988) and the rigid geometry model of Robson and Platt (1986). The force fields are consistent only as a whole such that individual terms cannot be interchanged as the interactions parameters are fitted to describe implicitly also cross-terms of the covalent energy and many-body effects. The applicability and the limitations of these force fields have been discussed repeatedly (Roterman *et al.*, 1989a; Roterman *et al.*, 1989b; Kollman and Dill, 1991; Gibson and Scheraga, 1991; Veenstra *et al.*, 1992; Aleman and Oroczo, 1992). The polarization energy, hydrogen bonds, and the entropic terms of the free energy are especially poorly modeled.

The calculation of variations in covalent energy and of the van der Waals energy is relatively easy and accurate. These components of energy can be relatively reliably determined in contrast to the other contributions that are related to electrostatic energy or entropy. Therefore, many studies estimate only the former terms. It is controversial as to whether anharmonic and bond/angle cross terms should be added to Hooke's law in the formulas for bond stretching and valence angle bending (Oroczo and Luque, 1993; Maple *et al.*, 1994) in macromolecular simulations.

Quantum mechanical results imply Buckingham's potential as the functional form for van der Waals interactions, namely,

$$E(\bar{r}) = Ae^{-B\bar{r}} - \frac{C}{\bar{r}^6} - \frac{D}{\bar{r}^8}$$

where $\bar{r}$ is the distance between two atoms and A, B, C, and D are fitted parameters. Due to computational efficiency, the Lenard-Jones 12–6 potential is used in which the exponential term is substituted with $A'/\bar{r}^{12}$ and the final term describing dipole-quadrupole interactions is omitted. The stiffness of the Lenard-Jones potential over small distances is often a source of computational artefacts and problems, therefore, polynoms in $\bar{r}$ approximating the potential function over the necessary range of distances but having a finite value at zero distance are sometimes preferred (Gibson *et al.*, 1993).

The non-electrostatic component of a hydrogen bond can be estimated with the Lippincott-Schroeder potential functions (Schroll *et al.*, 1991). In the packages ECEPP, AMBER, CHARMM, etc., 12–10 potentials are in use because of computational efficiency. Lavery *et al.* (1986) have proposed only linear hydrogen bonds be described with the 12-10 potential, while the 12-6 potential be used for non-hydrogen bonding interaction between acceptors and donors. The smooth transition between both regimes is achieved with switching function containing the cosinus of the angle α describing the linearity of the hydrogen bond

$$E_{HB} = \cos(\alpha)E_{12-10} + \left[1-\cos(\alpha)\right]E_{12-6}$$

The solvation energy of the protein consists of short- and long-range components (Still *et al.,* 1990; Cramer and Truhlar, 1992), namely, the polarization term and the energy contribution of cavity formation, solute-solvent dispersion interaction, and solvent structure changes resulting from the introduction of the solute into the solvent (Table 2). Experimental evidence from saturated hydrocarbons shows that the latter energy component is linearly related to the solvent-accessible surface area.

As reviewed by Eisenhaber *et al.* (1995), the fast and accurate calculation of the solvent-accessible surface area has been a serious stumbling block for a long time. Presently, two very efficient and accurate algorithms ASC (Eisenhaber and Argos, 1993) and NSC (Eisenhaber *et al.,* 1995) for analytical and numerical surface calculation, respectively, represent a significant advance. The program ASC implements intersection circle tests that reduce the dimensionality of the computation task. A built-in mechanism based on integer comparisons removes singularities resulting from multiple intersections of spheres. NSC, a variant of the dot sphere technique of Shrake and Rupley (1973) with spatially ordered dots, allows almost interactive surface calculation for proteins with about 1000 non-hydrogen atoms, even on a single processor workstation.

As first approximation, the energy weight for the surface-related part of the solvation energy does not depend on atom type and is in the range of 10 to 30 cal/(mole · Å^2) (Reynolds *et al.,* 1974; Eisenberg and McLachlan, 1986; Still *et al.,* 1990). For energy minimization, the analytical derivatives of the analytically calculated surface are available in Descartes' (Richmond, 1984; von Freyberg and Braun, 1993) and internal coordinates (ASC). For Monte Carlo searches, the faster numerical surface computation technique NSC is preferable (Eisenhaber *et al.,* 1994).

The electrostatic energy of the protein in homogeneous medium can be easily calculated with Coulomb's law; however, in practice, it is much more difficult. The protein charges are embedded into the low dielectric protein interior, which itself is polarizable and which is surrounded by highly polarizable solvent (water) and, maybe, salt ions. Solvation also influences the electrostatic component of intramolecular hydrogen bond interactions. Therefore, the hydrogen bond energies appear very much overestimated in force fields currently used. An example has been given by Nagy *et al.* (1994).

In principle, thè solution of the Poisson-Boltzmann equation for a very complex shape geometry (Harvey, 1989; Davis and McCammon, 1990; Juffer *et al.,* 1991) is required to calculate the protein electrostatic energy and the polarization part of the solvation energy. Even with recent improvements in the speed and accuracy of potential calculations (Oberoi and Allewell, 1993; Holst and Saied, 1993; Holst *et al.,* 1994a; Holst *et al.,* 1994b), the direct numerical solution of this differential equation is far too computer-time demanding for conformational analysis as well as the discretization of the solvent in form of individual solvent molecules (Levitt and Sharon, 1988) or dipoles (Warshel and Åqvist, 1991; Lee *et al.,* 1993). In the latter cases, the parametrization of water molecules (geometry, charges, dipoles, and polarizabilities) poses additional problems.

The Poisson-Boltzmann equation can be solved analytically for spherical geometries. The spherical image method is a computationally fast approximation to this solution. Many methods have been developed relying on the assumption that the protein geometry does not differ much from the symmetry of a sphere (Tanford and Kirkwood, 1957; Friedman, 1975; Matthew, 1985; Karshikov *et al.*, 1989; Schaefer and Frömmel, 1990; Wallqvist, 1993). Only the MIMEL technique of Abagyan and Totrov (1994) appears to be applicable for direct use in macromolecular simulations. Of course, the spherical approximations break down in the case of large conformational changes, partly unfolded states, and rod-like proteins.

Approaches relying on distance-dependent dielectric permittivities, albeit their lack of physical justification, are still more reasonable than the simple Coulomb's formula and are in use widely. The various functional forms have been reviewed by Mehler and Solmajer (1991). Unfortunately, all dielectric models cannot estimate the self-energy of charges. This is not significant if the energy function is applied to structure refinement. For large conformational changes, the self-energy is very important. The FIESTA approach to solve the Poisson-Boltzmann equation in an approximate form by the placement of virtual dipoles and charges at the coordinates of real and virtual atoms (i.e., into the molecular volume; as a result, the remaining volume may be not explicitly considered in the calculation) looks promising because it combines physical justification with fast calculation (Sklenar *et al.*, 1990). The authors also demonstrated that the electrostatic energy cannot simply be represented as a sum of pairwise atom function because terms involving the interaction of two protein charges with volume elements of third atoms are not negligible. For example, there is always a repulsive force between a low dielectric volume (e.g., a methyl group) and any charge. This force decays in a first approximation as $1/\bar{r}^4$. The virtual source technique as described by Sklenar *et al.* (1990) uses quite crude approximations for calculating volume overlap and field strength in the case of nearby atoms (e.g., in hydrogen bonds). Recently, new ways of solving this problem have been reported (Vorobyev *et al.*, 1992; Davis, 1994). The virtual source technique has been applied successfully to calculate transfer energies of nucleic acid bases (Sklenar *et al.*, 1990), the azurin pK_a shifts and the field around superoxide dismutase (Davis, 1994).

The strong dependency of the electrostatic computations on atom radii and sources (atom charges, dipoles, etc.) is of serious concern. Perhaps, an optimization of these parameters together with the other force field parameters could solve the problem, although the independence of the radii and the sources (atom charges and dipoles) on conformation has to be justified.

The difficulties in calculating the polarization term of solvation stimulated attempts to combine both components of the solvation energy E into a single function involving atomic accessibilities (Eisenberg and McLachlan, 1986; Ooi *et al.*, 1987; Vila *et al.*, 1991; Wesson and Eisenberg, 1992):

$$E = \sum_i \sigma_i A_i ,$$

where the products of atomic solvent-accessible surface A_i and atom-specific surface energy weight σ_i is summed over all atoms i. This approach may appear convincing because most charged groups are exposed anyway and their self-energy might be computable with their accessibility. The attempt to use such a surface-based function together with an *in vacuo* force field and to fit appropriately the surface energy weights has led to contradictory results (Schiffer *et al.*, 1992; von Freyberg *et al.*, 1993; Schiffer *et al.*, 1993). Not surprisingly, von Freyberg *et al.* (1993) had the best structure refinement results for BPTI and the α-amylase inhibitor tendamistat (compared with the native structure) if the ECEPP energy was considered in combination with the total or apolar surface without differentiation of energy weights with respect to atom types. This result can only be interpreted in the way that the polarization part of the solvation energy (Table 2) cannot be described by a surface energy function.

To date, the entropic contributions to the free energy from the protein molecule itself (probabilities of different conformations of the molecule or its substructural elements) are rarely considered. Yet the enthalpic and entropic contributions on burying atomic groups are of similar magnitude at room temperature. For example, the exposure of one methylene group to solvent costs about 1 kcal/mol as reported by different authors (Reynolds *et al.*, 1974; Tunon *et al.*, 1992). At the same time, the entropic gain from the flexibility of one aliphatic bond is -Rln3 = –0.66 kcal/mol (Finkelstein and Janin, 1989; Abagyan, 1993), where R is the gas constant. The entropy change of the main chain is considerable for large conformational variations (Karplus *et al.*, 1987), especially if disulfide bonds are closed or removed (Zhang *et al.*, 1994).

A major effect on the energetics of protein folding is the loss of conformational entropy of side chains with respect to the unfolded state. Estimates for the change in entropy have been obtained from statistics of side chain conformations in the Brookhaven Protein Databank (Pickett and Sternberg, 1993; Blaber *et al.*, 1993; Abagyan and Totrov, 1994; Blaber *et al.*, 1994), from the amino acid fusion energies (Sternberg and Chickos, 1994), from Monte Carlo simulations of conformational ensembles (Creamer and Rose, 1992; Creamer and Rose, 1994), and from energy maps (Lee *et al.*, 1994). The latter authors calculate the side chain entropy as the sum of two terms, one representing the rotamer distribution and the other estimating harmonic oscillations inside a given energy well. The different entropy scales are generally in reasonable agreement except for a few errors in accounting for symmetrical conformations (Lee *et al.*, 1994). Further improvement of the side chain entropy scales may require differentiation in (repetitive) main chain conformations (Schrauber *et al.*, 1993; Lee *et al.*, 1994). A first attempt to incorporate the side chain entropy as a function of the accessibility of the tip atom was described by Abagyan and Totrov (1994).

Because the energy function must be evaluated many times for searching the conformational space of proteins (for rough orientation, 10^6–10^9 times), the computational cost of one evaluation must be reduced to a minimum. Fast processors and massive parallelization are only useful if the procedure of energy calculation itself is tuned for optimal performance in every detail. There are efficient hierarchical algorithms, tree codes, and fast summation techniques for pairwise interactions developed in astronomy, fluid dynamics, and other areas of physics (Greengard, 1994). Neighbor list computation (Yip and Elber, 1989) may be improved by clustering near atoms into chemical groups (atoms connected via a small number of chemical bonds) or spatial groups (atoms in one grid cell of a space lattice). Numerical surface calculations were drastically sped up by tabulating spatially ordered points of dot spheres and by neighbor list calculations with the help of a cubic grid (Eisenhaber *et al.*, 1995). Much remains to be done to achieve further increase in efficiency for molecular energy calculations.

3. Methods for Searching the Conformation Space

The conformational space, even for a small protein, is of very high dimensionality such that the determination of low-energy conformations presents a big challenge. The complex form of the potential function containing both repulsive and attractive terms centered at many atomic positions is directly responsible for the roughness of the energy hypersurface and the existence of multiple local minima. The number of local energy minima Ω is believed to increase exponentially with the number of residues N as

$$\Omega = b^N$$

where b is in the range of 10 (Head-Gordon *et al.*, 1991). The following approaches to the search problem have been applied:

1. Complete enumerations of conformations
2. Build-up techniques from building blocks
3. Deterministic methods of global optimization
4. Stochastic searches (Monte Carlo methods, genetic algorithms)
5. Integration of Newton's equations of motion (molecular dynamics)

Complete enumerations are only applicable in a special context restricting the task to the consideration of only a small subspace of conformations. For example, a few side chains have to be altered in an otherwise homologous protein structure or a backbone has to be reconstructed from α-carbon positions. Efficient enumeration strategies for these purposes are implemented in the package CONGEN (Bruccoleri, 1993). A drastic reduction of the conformation space is also achieved in lattice models that allows checking of all available conformations for medium-sized polypeptide chains (Finkelstein and Reva, 1991). Lattice techniques will be discussed in more detail below (Section IV.A.4).

Build-up techniques or fragment-based approaches represent the global conformation of a protein as a combination of low-energy conformations of oligopeptide segments. Non-local interactions are assumed to be dominated by local ones and to introduce only little strain in local substructures. These approaches are very attractive as the difficulties involving the exponential growth of conformational space with the number of degrees of freedom are bypassed (Vasquez and Scheraga, 1985; Vajda and DeLisi, 1990; Sippl, 1990; Simon *et al.*, 1991). For example, Sippl (1990) attempts to enumerate all low-energy conformations for pentapeptides that could be combined at a later stage in a stochastic way to form polypeptide conformations. The physical reality is more complex. It is known that oligopeptides with an identical sequence can have many different conformations in proteins ranging from extended to helical states (Kabsch and Sander, 1984; Argos, 1987; Zhong and Johnson, Jr. 1992; Cohen *et al.*, 1993) depending on the structural context. Even taking into account an average energy width of only 2 kcal/mol per residue, which is not much, the enormous number of all configurations of a pentapeptide differing from the optimal state by less 10 kcal/mol has to be considered. The inaccuracy of the energy function can further increase the number of oligopeptide conformations because false positively evaluated conformations are not rejected.

Deterministic methods for global optimization achieve a flattening of the potential energy hypersurface so that roughness is reduced and many local minima disappear. The basic tenet is the averaging of the potential over a suitably wide range to achieve a smoother potential function. Protocols based on Schroedinger's (Samorjai, 1991; Olszewsky *et al.*, 1992) and the diffusion equations (Piela *et al.*, 1989; Kostrowicki and Scheraga, 1992; Shalloway, 1992) have been developed. Both approaches have bottlenecks and it is not obvious how to remove them. In the first case, the problem of the basis set arises, but it may be softened by a mean field approximation. In the second approach, the relationship between the global minimum of the smoothed function and of the real energy cannot be resolved easily. Also, the treatment of complicated energy terms as solvation is currently impossible with these methods. Deterministic global optimization has been applied successfully to simple molecules such as terminally blocked alanine and Met-enkephalin (Kostrowicki and Scheraga, 1992). The computation time was two orders of magnitude smaller than for special Monte Carlo search techniques. The regulation of the degree of hypersurface smoothing by a neural network has been applied successfully to the folding of mellitin (Head-Gordon *et al.*, 1991; Head-Gordon and Stillinger, 1993).

Monte Carlo searches are a powerful stochastic tool for searching low-energy conformations because large jumps in the conformational space are possible and the search can be continued outside the convergence intervals of local minima. A Markov chain of conformations is generated that represent a walk through the conformational space. The matrix of transition probabilities can be adjusted to sample conformations of special interest. For example, a newly generated conformation $i + 1$ can be accepted as a continuation of a Markov chain following the conformation i if its energy is smaller than the previous energy or if a randomly chosen number in the interval [0,1] is smaller than the Boltzmann probability of transition exp(–ΔE/RT), where ΔE is the energy difference of conformations $i + 1$ and i and RT is the product of gas constant and absolute temperature (Metropolis *et al.*, 1953). In this case, the ensemble of conformations represented by a large Markov chain at constant temperature is equivalent to the trajectory of a long molecular dynamics run without successive time-ordered conformations (ergodic theorem). The ensemble can be used to calculate the partition function Z in regions of conformational space visited with high probability by

$$Z = \int \cdots \int \exp\left(\frac{-E(\bar{r})}{RT}\right) d\{\bar{r}\}$$

(E is the conformational energy, $\bar{r}$ is the set of coordinate vectors, and RT is the product of gas constant and absolute temperature) and to estimate directly free energies (Paine and Scheraga, 1987).

To perform classic ensemble-oriented Monte Carlo simulation in other than Cartesian coordinates (e.g., in internal coordinates), an even search in all coordinate dimensions has to be ensured. In the present example, it is not obvious how to appropriately weigh different rotations and translations against each other.

In simple searches for low-energy conformations, modifications of the corollary Monte Carlo algorithm are more efficient and many procedural restrictions (e.g., the demand of even sampling of each coordinate direction and the fixation of the transition probabilities to Boltzmann values) can be bypassed. After each random alteration of the preceding conformation *i*, the resulting conformation *i* + 1 can be subjected to energy minimization to move to the closest local minimum (Li and Scheraga, 1987). This method is especially efficient if analytical gradients of the energy function are available and, therefore, the computational costs of finding the local minimum are relatively low. The MCM method combining Monte Carlo search and energy minimization can be considered as a real step forward as this technique allowed, for the first time, to locate reproducibly the low-energy conformations of enkephalin over fewer than a million energy evaluations. Other alternatives are Monte Carlo procedures starting in a high-dimensional coordinate space that is gradually pro-jected to three dimensions (Purisima and Scheraga, 1987), with successively lowering of an initially high-temperature and/or progressive decrease in Monte Carlo step size (Wilson and Cui, 1990; Okamoto, 1994). In the electrostatically driven Monte Carlo procedure, a fast independent energy minimization of the peptide dipole interactions (in the form of iterative orientation optimizations of individual dipoles in the field of all charges) is included in each step (Ripoll *et al.*, 1991). The influence of the Monte Carlo protocol on the efficiency of finding all low-energy conformations for a small peptide has been studied in detail (Nayeem *et al.*, 1991; Abagyan and Argos, 1992). Because the probability of steric clash with Monte Carlo steps in internal coordinates is quite large, biased jumps into regions of sterically allowed combinations of dihedral angles can drastically improve performance (Lambert and Scheraga, 1989; Bascle *et al.*, 1993; Abagyan and Totrov, 1994). Other attempts have been based on the covariance matrix of atomic movements (Vanderbilt and Louie, 1984; Noguti and Go, 1985; Shin and Jhon, 1991) that can reveal collective motions for particular subsets of atoms. Bouzida *et al.* (1992) have exploited relations between acceptance rate and step size to determine the amplitude of the next move. However, these methods make assumptions regarding the functional form of the energy hypersurface that render them appropriate to local rather than to global conformational sampling.

Genetic algorithms are in some way similar to Monte Carlo algorithms. Instead of the trajectory of a single conformation, the evolution of a population of conformations in the conformational space is monitored. In addition to mutations that resemble Monte Carlo steps, exchanges of structural features (crossing-over) are possible to produce new conformations of the subsequent generation. Therefore, once optimized substructures may be distributed throughout the population and accumulated in individual representatives. This advantage has to trade with higher computational expense due to the need of maintaining in storage a whole bunch of conformations and of calculating the energy for each representative. Therefore, genetic algorithms are usually applied in context with reduced representations of proteins and simplified energy functions (Dandekar and Argos, 1992; Unger and Moult, 1993; Sun, 1993; Dandekar and Argos, 1994). Genetic algorithms involve many parameters such as population size, number of generations, rate of crossing over and mutation, and the possibility of combination with energy minimization. The optimal choices influence critically the computational burden (McGarrah and Judson, 1993).

Integrating Newton's equations of motion (molecular dynamics, MD) in either Cartesian (van Gunsteren and Berendsen, 1990) or internal (Mazur *et al.*, 1991; Dorofeev and Mazur, 1993) coordinates would be an ideal solution as the problem of the native structure could be solved together with the question of the folding pathway. The interaction forces between atoms and, maybe, additional stochastic forces to simulate thermal noise, friction, and the like result in a move for each atom per time step, typically in the order of a few femtoseconds (fs). The possible time period that can be feasibly explored is determined by the time step size and the number of time steps. The upper limit of the time step size is set by 1/15 of the highest-frequency vibration modes of atomic movement (Bouzida *et al.*, 1992) defined by the stiffness of the energy function (Harrison, 1993) to achieve a reasonable integration accuracy. The number of time steps is also limited by the accuracy of the numerical integration procedure applied (Venneri and Hoover, 1987; Pastor, 1991; Norman *et al.*, 1994; Okunbor and Skeel, 1994) because the numerical error of integrating Newton's equations with any initial conditions accumulates with time. Both the time step size and the number of time steps are, therefore, limited by the available computer resources for accurate integration. In addition to these restrictions, the simulation will increasingly depart from the behavior of the real system in a long-term run because the functional form of the potential used does not accurately describe the real energy. Therefore, MD simulations of the assembly of native structures that take typically 1 ms to 1 s are currently not feasible.

Many attempts have aimed at enlarging the time step size. The coordinates can be classified into soft (dihedral angles) and hard (valence bonds and angles) degrees of freedom depending on the potential energy change resulting from a small coordinate shift. Fixing the hard degrees of freedom removes some high-frequency oscillations and allows to increase the integration time step to 10 to 20 fs (Dorofeev and Mazur, 1993; Rudnicki *et al.*, 1994). In Cartesian coordinate representation, the coordinate constraints resulting from fixing bond lengths and angles result in so-called holonomic conditions that increase the computational costs of integrating the equations of motions. Only for special cases is their analytical resolution possible (Miyamoto and Kollman, 1992). In internal c-oordinates, the constraining is easy because the possibility to do it analytically for any combination of fixed internal variables is an inherent property of this representation of the molecular geometry (Mazur *et al.*, 1991; Dorofeev and Mazur, 1993). Combination of molecular dynamics with normal mode analysis is a generalization of the technique of fixing hard degrees of freedoms (Zhang and Schlick, 1993; Dauber-Osguthorpe and Osguthorpe, 1993).

For simple model systems, the time step could be increased up to 50 fs. The high-frequency motions may be removed also in an overdamped Langevin dynamics mode (Gronbech-Jensen and Doniach, 1994). The limits for speed-up of molecular dynamics simulations by increasing the time step have been studied empirically by Scully and Hermans (1993), and they advise to use multiple time steps for covalent, short-range and long-range, forces (2, 4, and 16 fs, respectively).

Molecular dynamics are often used as refinement tools at the last stage of model building to relax conformational strain, for example, in homology modeling. Another application is the generation of low-energy conformational ensembles that comply with X-ray crystallographic stucture factors or NMR distance restraints (Brünger and Karplus, 1991; Gros and van Gunsteren, 1993; Schmitz *et al.*, 1993). The MD algorithm has also been used for estimating free energies of conformational transitions, docking events, and the like (Beveridge and DiCapua, 1989; Mark *et al.*, 1991; Mezei, 1993). Such simulations are extremely computing intensive (CPU weeks on supercomputers) and the accuracy of the results is only half-quantitative at best. Pearlman (1994) estimated that the MD trajectory sampling should exceed 700 ps to achieve free energy convergence for the simulation of a small organic system.

Classic molecular dynamics are a mostly local search method selecting from a few local minima in a small region of the conformational space — the deepest one. For example, Elofsson and Nilsson (1993) conclude that 10 MD runs, each of 100 ps with different starting conformations, visit more regions of the conformational space than a single 1000 ps calculation. The application of MD to molecular conformations separated by energy barriers of 3 kcal/mol or more is problematic because of slow rates of convergence (Guarnieri and Still, 1994). Several modifications have been proposed to make the search strategy more global among which are relaxation from four dimensions (van Schaik *et al.*, 1993a), molecular dynamics at constant or decreasing potential energy levels, the so-called PEACS algorithm (van Schaik *et al.*, 1993b; Byrne *et al.*, 1994), temperature changes (van Schaik *et al.*, 1993b), atom mass weighting (Mao and Friedman, 1990), or combinations with Monte Carlo jumps (Byrne *et al.*, 1994; Guarnieri and Still, 1994). At the moment, the equivalent Monte Carlo algorithms, especially Monte Carlo with energy minimization and simulated annealing, appear computationally cheaper and more efficient than the molecular dynamics variant in finding low-energy conformations.

In conclusion, the multiminima problem can be considered as solved only for special applications, for oligopeptides, or for local refinement tasks. The complete sampling of low energy conformations of medium-sized proteins is not yet in sight.

4. Hierarchy of Molecular-Mechanical Models for Protein Folding

Inasmuch as the folding simulation with full physical detail is extremely computationally expensive, a variety of techniques and methods have been developed to simplify the task. The modifications concern three areas:

1. The energy function
2. The geometric representation of the molecule
3. The search space

Without leaving the level of atomic detail, the following approximations have been used.

- The number of pairwise interactions can be reduced by the introduction of cutoffs, or the neighbor lists may be not updated for each conformational change. Switching functions can be applied to scale such interactions gracefully to zero. These approximations are applicable to van der Waals interactions and to Coulomb electrostatic interactions but should be used with care (Abagyan and Argos, 1992; Schreiber and Steinhauser, 1992; Steinbach and Brooks, 1994; Abagyan and Totrov, 1994).
- Long-range interactions of distant atoms may be approximated by a mean field approximation (Barnes and Hut, 1989; Beglov and Lipanov, 1991; Ding *et al.*, 1992). For example, different levels of monopole charge detailization can be used for different distance ranges: atomic charges for short distances and sum charges of atomic groups for large distances (Abagyan and Totrov, 1994). A mean field approximation can compensate the error introduced by a short cut-off for detailed interaction as in the case of solvent polarization (Schreiber and Steinhauser, 1992). A very specific application of the mean-field approach is the multi-copy sampling, for example, the average field of all available side chain conformations (Roitberg and Elber, 1991; Zheng and Kyle, 1994). Periodic or stochastic (random force) boundary conditions are often applied on explicit solvent. They combine atomic detail at the boundary with average influence from distant parts of the volume.
- Non-polar hydrogens have been treated as united atoms together with the corresponding carbons. Sometimes, this approach is used for all hydrogens (Pletnev *et al.*, 1974; Dunfield *et al.*, 1978; Brooks *et al.*, 1983). This approximation is more severe than one may think because, factually, the protein surface is almost exclusively formed by hydrogens (Richardson *et al.*, 1992).
- Hard degrees of freedom (valence bond lengths and valence angles) can be fixed, and conformations are searched only in the dihedral angle space. The consequences of this assumption have not been thorougly discussed in the literature. In a normal mode analysis (Kitao *et al.*, 1994), it was established that the main effect of fluctuations in bond lengths and valence angles is to allow the dihedral angles to become more flexible. The r.m.s. deviation of atomic fluctuations was 13% larger in Cartesian coordinate space than in dihedral angle space. It is also commonly accepted that the flexibility of ring structures demands variable valence angles.
- The introduction of rigid bodies is another tool to reduce dimensionality (Mazur and Abagyan, 1989). For example, the dihedral angles of the main chain inside helices may be frozen and only their packing can be analyzed (Abagyan *et al.*, 1994b). As a result, atomic interaction inside the rigid body does not change in conformational transitions. The application of rigid bodies in molecular dynamics has stimulated the development of special numerical techniques (Evans and Murad, 1977; Kneller and Geiger, 1989; Mazur *et al.*, 1991; Fincham, 1993). The helix motions in deoxymyoglobin contribute to 86% of the overall dynamics as reported by a rigid body MD study (Furois-Corbin *et al.*, 1993). The authors emphasized the importance of retaining the flexibility of side chains. The idea of predicting tertiary structure as packing of secondary structural

elements (established in advance by secondary structure prediction) is an application of the rigid body technique.

- The discretization of certain degrees of freedom reduces the search space, for example, the introduction of rotamers for side chain conformations (see Section IV.C) or the use of a limited set of conformations for small strained rings such as proline. The discretization of conformational states is in some way contradictory to the other efforts to represent the system in atomic detail, and errors caused by the impossibility to represent the geometry with less than 1.5 Å absolute accuracy should be avoided.
- Symmetry operations can be used to reduce the number of degrees of freedom if the geometry of the system allows it.

The knowledge of experimental restraints, for example, experimental distance or accessibility information, can be used for reducing the search space by including them as penalty functions in the energy function.

At the next level of simplification, single atoms disappear and atomic groups, residues, or building blocks are approximated by points at which spherical potential functions are centered. The reduction in the number of effective atoms and in the number of degrees of freedom make the energy hypersurface much smoother and removes many local minima. Many tricks used in simulations with atomic detail are also applicable here, such as cutoffs, mean field approximations, fixing degrees of freedoms, and specifying rigid bodies. Reduced representation modeling methods may be useful to study low-resolution structural details (similar to X-ray crystallographic structures with a resolution level 2.5 to 3 Å at best).

Levitt (1976) considered the backbone as a chain of α-carbons connected by virtual bonds. Two points are rigidly fixed to each virtual bond and represent N and O atoms of the peptide group. Side chains are modeled by a centroid anchored by a virtual bond at the α-carbons. Using physical considerations and experimental data for small compounds, he managed to compile a complete force field, including a solvation function. Minimization of the X-ray structure of bovine pancreatic trypsin inhibitor (BPTI) resulted in an r.m.s.d. of about 3 Å. Models obtained from an open starting conformation differed from the native conformation with an r.m.s.d. of about 7 Å if some helices were fixed as rigid bodies. Free simulations starting from fully extended structures resulted in largely deviant conformations with an r.m.s.d. clearly above 10 Å. This value is similar to that of randomly generated compact structures (Cohen and Sternberg, 1980).

Followers of Levitt maintained the general methodology. In some studies, the peptide group is modeled with less detail (Liwo *et al.*, 1993; Herczyk and Hubbard, 1993; Sun, 1993; Wallqvist and Ullner, 1994), other models have a 2 to 3 point representation for longer side chains (Wilson and Doniach, 1989; Herczyk and Hubbard, 1993; Wallqvist and Ullner, 1994) or do not consider the side chain at all (Gerber, 1992; Aszodi and Taylor, 1994). In addition to energy functions based only on physical considerations and physicochemical data on organic compounds, knowledge-based potentials of mean force derived from macromolecular structures have been used to discriminate native folds. The quality of the potentials of mean force are discussed in detail in the following Section (IV.B). The accuracy of the potential functions is not very good and many false-positive conformations may not be rejected. For example, Monge *et al.* (1994) achieved a four-helix bundle packing of the rigid helices of myohemerythrin only if a penalty on the gyration radius (explicite demand on compactness!) was added to the knowledge-based potential. Similarly, many successful modeling studies with reduced protein representation have applied potential functions containing in addition to true energetic terms other penalties biasing toward the formation of secondary structures, specific interactions, or known tertiary packing features.

Therefore, another, more promising perspective for application of simplified protein models is their use in generating structures from low-resolution experimental information (Ycas, 1990; Smith-Brown *et al.,* 1993; Saitoh *et al.,* 1993; Herczyk and Hubbard, 1993; Bohr *et al.,* 1993) such as a limited number of distance restraints (NMR, chemical linkage experiments) and accessibility information from spectroscopy or accessibility prediction (Wako and Blundell, 1994a; Rost and Sander, 1994b).

The discretization of the conformational space in the form of allowing only points on some lattice as pseudoatom positions (the so-called lattice models) is a consequent continuation of the idea of a reduced polypeptide chain representation at a lower resolution level. Many interaction types in lattices can be precalculated and tabulated. Accordingly, lattice conformation searches are more than 100 times faster than traditional molecular dynamics. To avoid bias from the lattice, its coordination number should be not too low as helical and extended structures may be differently well described by lattice points. Godzik *et al.* (1993) have reviewed the properties of many lattice types applied in protein folding studies and recommend the use of diamond, hybrid, or ultra lattices as a good compromise between precise representation (accuracy of 1 to 2 Å) and computational costs.

Although the computational task of folding a simplified chain is significantly reduced compared with an equivalent atomic detail calculation, sophisticated global search techniques such as complete enumeration, energy minimization, Monte Carlo or genetic algorithms are necessary to locate the low-energy conformations. The examples in the literature are oriented on reproducing X-ray crystallographic structures of small proteins such as bovine pancreatic trypsin inhibitor (Levitt, 1976; Hinds and Levitt, 1992), avian pancreatic polypeptide (Liwo *et al.,* 1993; Sun, 1993), melittin (Sun, 1993), crambin (Herczyk and Hubbard, 1993; Dandekar and Argos, 1994), and four-helix bundles (Monge *et al.,* 1994; Dandekar and Argos, 1994). A typical r.m.s.d. for backbone atoms is in the range 2 to 12 Å depending on the degree of how many experimental data are penalized in the potential function. As a rule, some conformations qualitatively different from the native one exhibit lower energies; thus, the force field is suboptimal.

Vieth *et al.* (1994) reported the reconstruction of the GNC4 leucine zipper with a very low r.m.s.d. of about 1 Å in a two-step approach. At the beginning, an approximate model was calculated with a reduced protein representation, a knowledge-based potential (their definition of hydrogen bond cooperativity favors secondary structure formation) and a Monte Carlo search in a lattice space. After completion, the r.m.s.d. was in the range of 3 Å, a surprisingly high accuracy. At the second step, the best low-resolution conformations were used for constructing polypeptide chains with atomic detail that were subsequently refined with CHARMM. The applicability of this technique to other cases, especially to sequences without known structure, has not been investigated.

Pure *ab initio* (i.e., without experimental distance restraints, explicit demands for secondary structures and compactness, etc.) modeling methods using a simplified protein representation and relying only on a true energy function still need to prove their ability to predict native conformations for sequences with unknown structure because energy functions involving complete atomic detail already have difficulties in recognizing the native conformation as the state of lowest energy.

Further, low-resolution models may not be well suited for extension to higher resolution (Hinds and Levitt, 1994). If the errors in the model coordinates are fairly significant, then a relatively large window of conformations around any particular low-resolution model still needs to be searched. If, in addition, the low-resolution models do not cluster well, such a window has to be explored around each of these models.

B. Threading Amino Acid Sequences through Structural Motifs

1. Introduction

The many unsuccessful attempts to predict protein tertiary structures from the amino acid sequence alone based only on fundamental physical principles (because computationally it is too expensive) have stimulated the development of more simplified approaches relying on experimental experience from observed protein structures. The inclusion of experimental data in the form of penalty functions involving, for instance, distance restraints, in the overall energy or the fixation of some degrees of freedom such as bond lengths and valence angles is one possible line of thought. The "threading" techniques described in this section are united by the far-reaching assumption that the sequences under study might accept one of the protein folds already studied by X-ray crystallography or NMR. The structure prediction problem is thus greatly simplified as the allowable conformational space is reduced to about 200 unique protein topologies presently known. The primary goal of a "threading" method is to establish relationships between amino acid sequences and folding patterns, that is, to select the most probable fold for a given sequence or to recognize suitable sequences that might fold into a given structure. This approach has been stimulated by three observations:

1. The number of different folds appears to be limited (see Section IV.B.2).
2. Distant relationships between sequences may be found by alignment to property profiles (see Section IV.D).
3. Empirical potential functions for estimating solvation can distinguish incorrect folds (see Section IV.B.3).

As introduced by Bryant and Lawrence (1993), "threading" a sequence through a fold means a specific alignment between the amino acids of the sequence under consideration and the residue positions of the folding motif. The known structure establishes a set of possible amino acid positions in the three-dimensional space (the tertiary template). The query sequence is made similar to the structure by placing its amino acids into their aligned positions. The recognition of sequence and structure is mediated by a suitable score or potential function for the evaluation of each alignment. The methods described in the literature differ in details of

- The derivation of the score function (see Section IV.B.3)
- The alignment procedure for a single sequence with a single structure (Section IV.B.4).

The following two subsections consider each point individually. Then a discussion of two search protocols will ensue, namely,

- a single sequence vs. multiple structures (Section IV.B.5.a) and
- a single fold vs. multiple sequences (Section IV.B.5.b).

During the last 3 years, the number of publications in this field has exploded. A few reviews describing some aspects of the material (Fetrow and Bryant, 1993; Wodak and Rooman, 1993; Argos, 1994; Lathrop, 1994) can be recommended.

2. How Many Folds Do Exist?

Analyses of known three-dimensional protein structures and amino acid sequences revealed that proteins are clustered into families whose members have evolved from a common ancestor, share a characteristic fold, and, sometimes, have a similar function (Pascarella and Argos, 1992b; Holm *et al.*, 1992; Yee and Dill, 1993; Orengo *et al.*, 1993; Holm and Sander, 1994a; Rufino and Blundell, 1994; Lessel and Schomburg, 1994).

The number of folds can be estimated from different genetic and structural data. Bacteria are likely to have explored most or all of the topologies. The genome of *E. coli* contains about 4000 different genes, the proteins of which are known to share folds in certain cases. The number of different exons throughout the phyla is estimated to be less than 10,000 (Dorit *et al.*, 1990; Doolittle, 1991; Patthy, 1991; Dorit *et al.*, 1991). Linear extrapolation from current genome projects yields a number of about 1000 different folds (Chothia, 1992).

With the assumption that the number of different proteins belonging to each protein family is normally distributed, the uneven representation of folds in the Brookhaven Protein Data Bank provides help in estimating the number of protein families. Using this method, Alexandrov and Go (1994) found an estimated 6700 different folds.

The search for ancient conserved regions (ACR) predating the coelomate radiation in gene sequences revealed that about 40% of the currently known protein sequences can be classified into less than 1000 groups (Green *et al.*, 1993; Claverie, 1993). At the same time, these latter authors emphasize that there is a large number of evolutionary new sequences without similarity to ACR, which may imply continuous creativity of nature in designing folds. Attempts to prove the limitedness of the available number of main chain topologies for amino acid sequences as a result of physical constraints such as compactness, hydrophobic effect, and uniqueness appear not completely convincing (Lau and Dill, 1990; Sali *et al.*, 1994) because the underlying assumptions are restrictive.

The number of families of related protèin structures critically depends on the value of the homology threshold applied in the protein structure comparison routines (see Section IV.C.2). With sequence identity below 50%, many structural parameters such as r.m.s.d. between equivalent Cα atoms, residue accessibility, side chain torsion angles, and side chain-side chain contacts are increasingly less conserved, although the overall fold and the secondary structural elements may not change (Hilbert *et al.*, 1993; Flores *et al.*, 1993; Rost *et al.*, 1994; Rost and Sander, 1994; Chelvanayagam *et al.*, 1994; Russel and Barton, 1994). For pairs of distantly related proteins (residue identity ~20%), the regions with the same general fold comprise less than half of each molecule and the r.m.s. deviation between equivalent main chain atoms is 1.8 to 3.2 Å (Chothia and Lesk, 1986; Russel and Barton, 1994; Aronson *et al.*, 1994). The relative shift of equivalent secondary structural units may be as large as 7 Å with rotations up to 30° (Lesk and Chothia, 1980; Chelvanayagam *et al.*, 1994). As was pointed out by Russel and Barton (1994), proteins can adopt very similar folds by using almost completely different stabilizing interactions and, therefore, may have little in common (from a physical point of view) apart from a scaffold of common core secondary structures.

3. Reduced Protein Representation and Knowledge-Based Potentials

Empirical potential functions estimating solvation have shown their ability to distinguish some incorrect folds. Thought of as a test for molecular mechanics calculations, the misfolded protein conformations of Novotny *et al.* (1984; 1988) prompted the development of a score of functions for the assessment of the quality of tertiary structure determination. Both the composition of the solvent-exposed surface (Eisenberg and McLachlan, 1986; Novotny *et al.*, 1988; Baumann *et al.*, 1989; Vila *et al.*, 1991; Holm and Sander, 1992b; Luthardt and Frömmel, 1994; Koehl and Delarue, 1994a) as well as the frequency of contacts between hydrophobic groups (Gregoret and Cohen, 1990; Colovos and Yeates, 1993; Vriend and Sander, 1993), may serve as discrimination criteria.

Effective knowledge-based potential functions with a one- or two-point residue description represent an intermediate between detailed atomic force fields and residue-specific secondary structure propensities. The former are very computer time consuming in evaluation of many conformations (alignments), especially when a correct description of the electrostatic part of the solvation term is attempted. The latter are insensitive to significant structural details. Some of the potential functions described in this section can be applied to not only the threading problem but also to *ab initio* structure calculations with reduced residue representation.

The approaches of different researchers share two main features:

1. *A priori* assumptions are made about the functional form of the potential, which is based on physical considerations (solvation effect, packing considerations, etc.). At the same time, the choice of the potential function reflects restrictions of various alignment techniques and some personal preferences.
2. A basis set of known protein tertiary structures and, occasionally, also of

> multiple alignments is used for statistical surveys or for optimization (learning) procedures to derive parameters associated with the different terms of the potential. In contrast to these two main alternatives, probabilities of a residue occupying a given position on a fold have also been calculated in a direct way (Abagyan *et al.*, 1994a) by relaxing the one-residue mutated structure in a force field with full atomic detail and using the resulting surface energy for constructing a 3-D profile. Ad hoc values for interaction potentials have been used only in early studies.

Ideally, the potential is finally tested on a large set of structures that have not been used for parameter derivation.

In optimization and learning procedures, the potential is constructed such that minimum values result for sequences in their native conformation, while computer-generated incorrectly folded alternatives yield larger values for the energy. The parameters are subject to constraints imposed on the value of the potential for a training set of known structures. However, the substantial computational burden to optimize a large number of parameters simultaneously restricts the variety of interaction classes that can be introduced. As a consequence, one has to be satisfied with coarse distance intervals (even to the extent of contact yes/no) and with only few residue categories instead of 20 amino acid types. These restrictions lower the sensitivity of the potentials. Only a few authors have explored complex optimization techniques for potential derivation such as multidimensional minimization (Crippen and Snow, 1990; Crippen, 1991; Maiorov and Crippen, 1992) or associative memory Hamiltonians (Goldstein *et al.*, 1992a; Goldstein *et al.*, 1992b; Madej and Mossing, 1993). The transferability of the potential derived in this way to other systems can be questioned and is very limited (Crippen and Snow, 1990; Snow, 1993).

In statistical approaches, the probability P of observing a parameter of interest, for example, the frequency of the C_α atoms of two *specific* residues being within some distance range and sequence separation, is first calculated from the learning set of structures, then normalized by a value P_o (in the present example, the expected probability of two such C_α atoms being in some distance range and sequence separation) to correct for sample bias, possibly smoothed to correct for sparse data in the learning set, and translated into a score G expressed as

$$G = -kT \ln (P/P_o) + \ln (Z/Z_o)$$

or some proportional value (Sippl, 1993a) where kT is the product of Boltzmann's constant and the absolute temperature, Z and Z_o are the partition functions of the specific and reference systems, respectively. Occasionally, the scores for different arguments are tabulated. Other authors use the experimental information to fit parameters in the analytical form of the potentials. Inasmuch as it is derived from a hypothetical Boltzmann statistics of molecular properties without estimation of the partition functions for the sequence studied (Z) and for the reference state (Z_o), knowledge-based potentials represent only relative free energies and, strictly speaking, can only be used to evaluate different conformations of one and the same sequence. Also, the approximation of independent probabilities describing the structural state of individual residues or residue pairs is a source of concern because the cooperativity between structural states of interacting residues is insufficiently taken into account. It should be noted that, for statistically derived potentials, it is not *a priori* known whether the native fold is the best conformation for a given sequence from the basis set.

Following earlier work (Tanaka and Scheraga, 1976; Crippen and Viswanadhan, 1985; Miyazawa and Jernigan, 1985), many researchers use a term summing pairwise residue-residue interactions derived from the frequency of certain residue-residue distances or contacts. In this case, the assumed Boltzmann distribution of interaction has not been proven yet. The normalization is effected essentially in two ways: by the observed interaction frequencies of *all* residue-residue types in the basis set or by interaction frequencies of the *same* residue-residue type in random sequences (with the same amino acid composition) in the structures of the same basis set (Bryant and Lawrence, 1993; Goldstein *et al.*, 1992a; Goldstein *et al.*, 1992b). If, in the latter case, all possible random sequences are considered, the frequencies should obviously converge to those in the first normalization scheme because of the mathematical equivalence of both procedures. Thus, it is not surprising that the method of normalization does not affect the potentials, and the first method can be chosen because of its simplicity.

Another aspect of the normalization is really important (Figure 6). Polar residues that are usually located closer to the surface of the protein have fewer residues at short distance contacts than would core residues (Walther, 1994). Only if the weight in the short distance range is changed, for example, by restriction to residues with similar accessibility (Wodak and Rooman, 1993) or by generating an ensemble of random sequence permutations with Boltzmann-distributed solvation energy (Walther, 1994), is the effect of specific residue-residue interaction details such as the coulombic attraction of oppositely charged ionic pairs reflected by the potential. The distance statistics show that the Debye-Hückel law $e^{-\kappa r}/r$ is a good description for the interaction between charged groups (Casari and Beyer, 1994) inside the protein (with $\kappa \approx 3.5$ Å similar to the Debye length in physiological salt solution). Therefore, from the physical point of view, residue-residue potentials, as usually normalized, describe mainly the hydrophobic effect (Casari and Sippl, 1992; Bryant and Lawrence, 1993) in addition to specific interactions between residues.

The residue-pair potentials have been sampled in the form of simple yes/no contacts (Godzik and Skolnick, 1992; Godzik *et al.*, 1992; Maiorov and Crippen, 1992; Hinds and Levitt, 1992) or in histograms with up to ~1 Å wide residue-distance bins. It is not fully clear to which extent simple contact numbers may memorize the size of the original residue. Interactions beyond 10 Å are compiled only by some authors (Wilson and Doniach, 1989; Sippl, 1990; Hendlich *et al.*, 1990; Sippl and Weitckus, 1992; Casari and Sippl, 1992; Sippl, 1993b), while many other reports ignore these contacts as non-residue-specific bulk effects (Jones *et al.*, 1992; Bryant and Lawrence, 1993; Wilmanns and Eisenberg, 1993). The average size of a globular protein seems to set an upper distance limit (e.g., 30 Å) for sampling statistically reliable residue-residue interactions (Figure 6). Nevertheless, some researchers continue beyond this value (Casari and Sippl, 1992). The partitioning of the interactions into ranges of

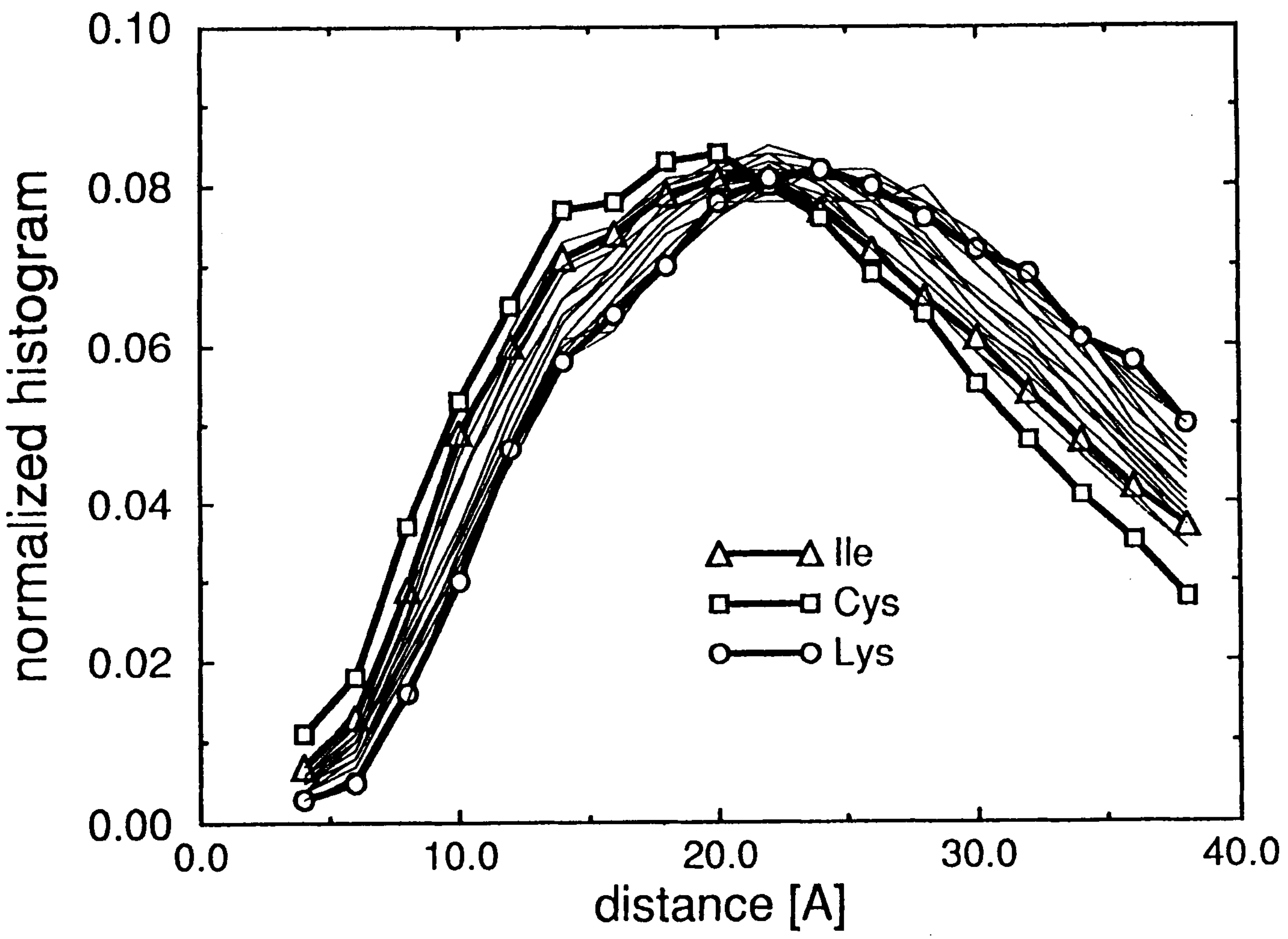


FIGURE 6. Histogram of pairwise amino acid distances. The observed frequency of distances between a residue of given amino acid type and any other residue in a large set of globular proteins is shown (courtesy of D. Walther). The maxima differ considerably for various amino acid types and cover a range from 18 to 25 Å. The shifts correlate with hydrophobicity, ILE and LYS being extreme examples. CYS is overrepresented at low distances because it is frequently involved in disulfide bridges, which are less polar than single cysteines. The reduction of distance frequency above 25 Å is due to the limited size of globular proteins.

residue separation along the amino acid sequence distinguishes between local effects (helical or extended secondary structure) and tertiary contributions (Sippl and Weitckus, 1992; Casari and Sippl, 1992; Jones *et al.*, 1992). This stratification appears to be important for the potential quality (Kocher *et al.*, 1994). The independent sampling of hydrogen-bonded and nonvalent contacts has also been proposed (Nishikawa and Matsuo, 1993).

The size of the present protein structure database makes it difficult to sample enough residue-residue interactions at all distance ranges; therefore, smoothing must be applied to the raw data. Kocher *et al.* (1994) who compared the efficiency of numerous potential functions strongly suggest that it is better to sacrifice specificity of multiparametric potentials in favor of statistical reliability of more simple functions. It was proposed to enlarge the data set of residue-residue interactions by considering together with each 3-D structure of the Brookhaven Protein Data Bank all known homologous amino acid sequences (Bauer and Beyer, 1994). This approach may not improve the potential as even homologous protein structures may differ considerably in relative positions of equivalent atoms, not only in the loops but also in more structurally conserved regions. The distance error being in the range of up to 5 Å for C_α atoms of proteins with about 25% residue identity (Chelvanayagam *et al.*, 1994) has unpredictable consequences for the potential parameters. It is difficult to substantiate the statistical reliability of residue-triplet interaction potentials with the present dataset restrictions. Nevertheless, they have been published (Godzik *et al.*, 1992; Godzik and Skolnick, 1992), and it has been reported that they constitute about 40% of the overall energy. Conversely, Crippen and Snow (1990) demonstrated that there is no three-body effect at the resolution level of a single-point residue representation.

It should be taken into account that the single-point approximation of an amino acid residue is very crude, ignoring the spatial anisotropy of a given residue. As a consequence of the spherical approximation, the resolution limit of knowledge-based pairwise potentials is expected to be half of the residue size or about 3 to 5 Å. Approaching a C_α atom from the direction where the side chain is located will result in a steric clash sooner than from the opposite side. Knowledge-based pairwise potentials cannot distinguish both cases. Therefore, it is not surprising that potentials centered at the geometric midpoint of the side chain. perform much better than C_β- or surface-based potentials that are in turn more efficient than C_α- potentials (Kocher *et al.*, 1994). This view is also confirmed by the absence of correlation of C_α distances with distances between side chain centers (Karlin *et al.*, 1994). At the same time, such functions cannot be easily used in *ab initio* calculations because the residue center is not fixed with respect to the backbone. A more detailed description of a residue with two or three pseudoatoms (one backbone point and one or maybe two points for the side chain) was used as a basis to compile statistics of distances in protein structures and to derive a potential that takes the asymmetry of an amino acid residue into account (Wilson and Doniach, 1989; Sun, 1993; Wallqvist and Ullner, 1994). The statistical reliability of this function is expected to be comparable with that of the usual one-point-per-residue potential.

For use in profile search and dynamic programming sequence alignment algorithms, it is necessary to express the potential in the form of single residue propensities without non-local effects that are introduced, for example, by pairwise residue-residue potentials. The environment-specific amino acid substitution tables of Blundell, Johnson, and their co-workers (Johnson *et al.*, 1993; Johnson and Overington, 1993; Topham *et al.*, 1993) contain such a potential in implicit form. Explicit functions have been proposed on the basis of hydrophobicity terms (total buried surface, fraction of polar surface, surface energy of a one-residue mutated and conformationally relaxed protein structure, number of neighboring protein atoms) and are sometimes combined with terms describing local structure (secondary structure class, backbone dihedral φ-ψ angle-dependent functions) in discrete or continuous analytical form (Bowie *et al.*, 1991; Lüthy *et al.*, 1992; Wilmanns and Eisenberg, 1993; Zhang and Eisenberg, 1994; Ouzounis *et al.*, 1993; Abagyan *et al.*, 1994a; Sippl, 1993a). Pairwise residue-residue potentials cannot be directly included into the residue propensities because of their non-local nature. With additional assumptions, pairwise residue-residue potentials may be used, for example, at the expense of calculating the contact energy with the original sequence (Wilmanns and Eisenberg, 1993) or generic bulk peptide (Ouzounis *et al.*, 1993) and not with the query sequence to be threaded. Of course, the sensitivity of propensity-type potential function is limited because it does not see the amino acid residues of the query sequence in the updated environment.

It has been attempted to enhance the discriminative power of the score function by combining several potential functions with pairwise potentials (Gracy *et al.*, 1993; Nishikawa and Matsuo, 1993; Kocher *et al.*, 1994). Possible candidates include single residue propensities (as discussed above) that monitor local backbone structure (Rooman *et al.*, 1991; Rooman *et al.*, 1992; Kang *et al.*, 1993; Kocher *et al.*, 1994; Kamimura and Takahashi, 1994) and residue solvation (Toma, 1994; Nishikawa and Matsuo, 1993; Kocher *et al.*, 1994). Each measure must be properly normalized to exclude memory effects related to the original sequence, especially under conditions of limited sequence update. It was shown that the residue solvent-accessible surface area introduces such bias to the original amino acid residue (memorization of residue size), whereas the relative residue accessibility does not (Kocher *et al.*, 1994). Best discrimination of native structures was achieved by applying a potential combining the pairwise term based at the side chain centers with a local backbone torsion angle function, although the tests were performed for only a limited set of protein structures. Single-residue solvation terms do not appear to increase recognition any more because this effect has already been implicitely described by the pairwise potential normalized in the usual way (Kocher *et al.*, 1994).

In conclusion, the proposed knowledge-based potential or score functions are quite crude and do not unambiguously distinguish between native conformation and incorrect alternatives because they do not correctly describe the free energy of an appropriate system but only some part of it. As was pointed out by Abagyan (1993), the hope that the number of false-positive conformations selected by an incomplete potential function might be small is unfounded given the experience of calculating low energy conformations of peptides.

4. Alignment between Sequence and Structure

In this section, we review procedural details of the aligment ("threading") of amino acid types of the query sequence with the spatial residue positions of a given fold. Unfortunately, the computational complexity of the threading problem is comparable to that of the direct folding problem (Ngo and Marks, 1992; Unger and Moult, 1994; Fraenkel, 1993; Lathrop, 1994) because of the significance of non-local interactions between residues. The search of an alignment between an amino acid sequence and the spatial residue positions of a fold belongs to the class of NP-hard problems (Lathrop, 1994). Consequently, any threading algorithm has to accept (at least) one of the following drawbacks:

1. It does not calculate residue-residue interactions from the query sequence being threaded (no or incomplete sequence update)
2. It does not admit variable-length gaps into the alignment
3. It cannot guarantee finding the optimal threading in some cases
4. It may require an exponential amount of computer time for some alignments

Originally, threading algorithms were designed to be directly compatible with dynamic programming sequence alignment procedures and profile analysis programs. Therefore, the potential function is left insensitive to the non-local interaction effect resulting from placing new amino acids into all positions of the template. Although this approximation leads to a much simpler search for optimal threading, the algorithm loses sensitivity because of the bias toward the original sequence. The following implementations have been described.

1. Only the local environment is taken into account ignoring contacts altogether (Bowie *et al.*, 1991; Lüthy *et al.*, 1992; Zhang and Eisenberg, 1994; Johnson *et al.*, 1993).
2. The non-local contacts are assumed with generic bulk peptide (Ouzounis *et al.*, 1993).
3. The interaction preferences are always evaluated with the residues from the original sequence of the structure instead of the sequence being threaded (Wilmanns and Eisenberg, 1993; Sippl, 1993b; Abagyan *et al.*, 1994a).
4. First, all suboptimal alignments are compiled with a dynamic programming routine ignoring non-local contacts and then all residues are substituted to select the favored alignment with a pairwise potential (Gracy *et al.*, 1993).

These techniques are especially useful if the main concern is not to get the spatial alignment but to recognize sequences that may have similar folds. The low computational costs make database screenings possible.

Other algorithms do not allow gaps in the alignment (Hendlich *et al.*, 1990; Sippl and Weitckus, 1992; Crippen, 1991; Maiorov and Crippen, 1992; Bryant and Lawrence, 1993). The sequence, the length of which is fixed by a window, is directly aligned with the structure (residue i at position i). In this case, the attention can be completely concentrated on the development of a good recognition function. The predictive power of such technique is limited because sequence and structure may be grossly out of register if variable-length gaps are not allowed.

Some algorithms using modified dynamic programming procedures do not guarantee that an optimal threading will be found; instead, suboptimal solutions are calculated in an amount of time increasing polynomially with the number of degrees of freedom. One modification consists in iterative steps of substituting the residues of the motif as, for example, in the "frozen neighbor" approximation of Godzik *et al.* (1992). Jones *et al.* (1992) applies a modified comparison routine with secondary levels of dynamic programming (Taylor and Orengo, 1989; Orengo and Taylor, 1990) to fix the neighbors for the first level. Other algorithms, among them stochastic search techniques, also may fail to find the optimum (Finkelstein and Reva, 1991; Goldstein *et al.*, 1992b; Nishikawa and Matsuo, 1993; Gracy *et al.*, 1993). In their gapped algorithm, Nishikawa and Matsuo (1993) combine the possibility of missing the optimal solution with a limited sequence update by allowing only residue substitutions in the homologous core regions, accepting the original sequence in the loops. All algorithms of this class favor the loop size in the original sequence by the value of the gap penalty. At the same time, for practical applications, it may not be so critical to find the optimal alignment given the limited discriminative power of the potential functions used, although Jones *et al.* (1992) found cases where accuracy suffered due to nonoptimal threading.

Due to the NP-completeness, algorithms that guarantee finding the optimal threading will use exponential time in some cases such as the procedure of Bryant and Lawrence (Bryant and Lawrence, 1993) for gapped alignments. These authors introduced lower and upper gap lengths based on observations in alignments of homologous sequences to reduce the search space. This limitation may miss distant homologues with very long loop deletions up to 140 residues (Starzyk *et al.*, 1987). The experience of solving NP-hard problems can be used to design effective algorithms, for example, knapsack packing techniques (Lathrop and Smith, 1994). Because of the exponential size of the search space, massive parallelization together with efficient coding in hardware-near languages are required for implementations of such techniques. With highly discriminative energy functions, only these methods will find native structures because each suboptimal solution may represent a misfolded conformation.

5. Application of Threading Techniques

a. Sequence Recognizes Structure

The method for deriving knowledge-based potentials (no estimate of the partition function) indicates that threading techniques are well suited for selecting the best possible structure among those known for a given amino acid sequence. The new sequence is aligned with each tertiary template, and the structure with the highest score is the probable fold. The significance of preference (the stability) of the selected conformation can be estimated by the relationship $\Delta E/\delta E$ (Z-score), where ΔE is the difference between the energy (score) of the probable fold and the average energy (score) of the ensemble of rejected alternatives considered as reference state and δE is the width of the energy (score) distribution of the uncorrectly folded structures (Goldstein *et al.*, 1992b; Kocher *et al.*, 1994). The sequence-recognizes-structure protocol mimics many aspects of protein folding.

Although almost every author claims that his or her potential recognizes the native structure for a given sequence with nearly 100% accuracy; however, the results should be viewed less optimistically (Kocher *et al.*, 1994). As a rule, the number of structural alternatives considered is small compared with the genuinely accessible conformational space. Therefore, the algorithm may recognize a conformation as well folded when it is known that the protein is a subunit stabilized only by quaternary interactions (Kocher *et al.*, 1994). Nevertheless, the threading score can be used as a hint to single out some oligomeric proteins (Eisenberg *et al.*, 1992) or badly refined regions in crystallographic structures (Lüthy *et al.*, 1992; Sippl, 1993b) and to model loops (Topham *et al.*, 1993). Wako and Blundell (1994a; 1994b) predict solvent accessibilities and secondary structures from the sequence with environment-dependent amino acid substitution tables.

Godzik and Skolnick (Godzik and Skolnick, 1992) have attempted to locate a structural motif using a supersecondary structure as a template (e.g., βαβ units) and achieved successful alignments in a number of cases. However, there is little hope that such approaches can be generalized. Numerous small polypeptides score very poorly due to being part of a larger assembly such as the crystallographic packing in the case of crambin or the oligomeric state of avian pancreatic polypeptide (Eisenberg *et al.*, 1992).

b. Structure Recognizes Sequence

The usual folding prediction starts with a primary structure (amino acid sequence) and calculates its tertiary structure and topology. The inverse technique involves starting with a structural template (set of spatial residue or atom positions) and testing if a sequence (designed or observed) can fit it. This approach has some touch of engineering because it is similar to the fine tuning of a sequence for a given structure and function as does natural evolution. The concept of inverse folding has been proposed by Drexler (1981) and Pabo (1983). An early attempt in this direction was performed by Pondor and Richards (1987), the success of which was limited because of the assumption of a fixed backbone. Homology modeling and protein *de novo* design are important fields of application for threading methods assigning sequences to tertiary templates.

The practice of the structure-recognizes-sequence protocol involves an additional approximation. Different systems (sequences) are compared in their probability to accept the same state (fold). The direct (absolute) evaluation of the partition function formally required in this context is very difficult. The scanning of each sequence against the set of all tertiary templates used as reference state would be a sensible but a computer time-consuming alternative. The sequences could then be compared by their ranking of the Z-score (see Section IV.B.5.a) calculated with respect to this reference state. However, one could argue that the number of states may well be too small and the conformations might not be very relevant for the sequence under consideration. In the simplest way, the threadings of sequences onto the same fold have been sorted in accordance with their energies (Ouzounis *et al.*, 1993). Also, significance tests known from usual sequence and profile alignments can be applied.

Sets of existing sequences with the same length (Bowie *et al.*, 1991; Lüthy *et al.*, 1992; Eisenberg *et al.*, 1992; Wilmanns and Eisenberg, 1993; Abagyan *et al.*, 1994a) and of random sequences with identical length and composition (Bryant and Lawrence, 1993) have been used as reference states for calculating Z-scores.

Ouzounis *et al.* (1993) have reported that their threading algorithm recognizes the native sequence among the top 10 threadings of a given structure for 84% of 64 tests among a set of 640 sequences. Although many higher scoring sequences were close homologues to the native sequence, the authors note that extensive searches of structural templates against sequence databases will uncover the defects in the specificity of the potential functions more easily than in the reverse sequence vs. structure protocol because the number of sequences is much higher (about 25,000 in their database) than the number of folds (100 to 200).

The possibility of providing suggestions on remote homologies that cannot be seen with conventional sequence alignment could make the threading methods valuable. The relatedness of C-phycocyanin and myoglobin (Jones *et al.*, 1992) and of actin and hsc70 (Bowie *et al.*, 1991) were found with these methods, although these similarities had already been known from X-ray crystallography and NMR. However, the simple sequence pattern search technique MAKPAT/PROPAT was even more successful in finding related members in the family, which include actin, hsc70, many prokaryotic cell cycle proteins, hexokinase, and other sugar kinases (Bork *et al.*, 1992).

The recognition of $(\alpha\beta)_8$ barrels is only marginal (Abagyan *et al.*, 1994a) but good enough to allow the hypothesis that three enzymes of the aromatic amino acid biosynthesis pathway may be structurally related (Wilmanns and Eisenberg, 1993). Pickett *et al.* (1992) show that multiple sequence alignments based on comparison of $(\alpha\beta)_8$-barrel structures can compete with threading.

Matsuo and Nishikawa (1994) threaded 11,706 sequences of the NBRF-PIR database into 101 structural templates derived from protein crystal structures. Among the 1489 pairs with high score, 67 with low sequence homology were retained, each being a fruitful directive for further biochemical research. Such screenings will hopefully produce more useful outcome when many more protein sequences from current genome projects are available.

Threading techniques have also been used to design new sequences for a given fold as possible candidates in biochemical experiments. Genetic algorithms (Jones, 1994) or simulated annealing (Hellinga and Richards, 1994) have been used to guide the walk through the sequence space. Another technique involves network minimization with random jumps in the sequence space of hydrophobic amino acids for designing core motifs (Kono and Doi, 1994).

One reason for the inefficiency of threading may consist in the different composition of the protein structure (PDB) and sequence databases. While the former contains almost exclusively water-soluble globular proteins, the latter was estimated to have about 25% proteins with transmembrane domains. Fifteen percent of the residues in SWISSPROT are in so-called regions with strong compositional bias when there are no or only single examples in the PDB (Wootton, 1994). Therefore, threading in the current fashion is applicable only to a small fraction of the sequence database.

In conclusion, the inverse folding methods have seen only a few real cases of efficient application. Their effectivity in recognizing new distantly related homologues is low. Standard multiple sequence alignment methods or profile analyses are computationally cheaper and are often able to do the same job. Threading methods will probably fail if the evolutionary divergence has removed most of the sequence similarity, if parts of the backbone have significantly moved, and if secondary structural elements are inserted or deleted despite preservation of a similar basic fold pattern. Such relative ineffectiveness is unexpected because the consideration of additional structural information should favor threading compared with simple sequence pattern search. The crudeness of the potential function and the NP-completeness of the alignment procedure appear responsible for this result. Another, important reason for the failure of threading methods is the low conservation of residue-residue interactions, residue accessibility, and secondary structure in proteins with low residue identity but identical fold (Russel and Barton, 1994). Therefore, similar backbone geometries may result from almost completely different intramolecular interactions.

C. Modeling of Tertiary Structure by Homology

1. Introduction

The first attempt of modeling a 3-D structure of a homologous protein (α-lactalbumin on the basis of the X-ray structure of chicken egg-white lysozyme) was done by Browne *et al.* (1969). The history of homology modeling is in detail described by Johnson *et al.* (1994). Greer (1981) pioneered the technique of predicting 3-D structural details for a protein sequence known to belong to a protein family having one or more representatives with resolved tertiary structure. He applied his approach to mammalian serine proteases, the classic example of the so-called modeling by homology. The principal steps of homology modeling involve the following.

1. Structurally conserved regions (SCR, the tertiary template) are determined on the basis of 3-D structural comparisons and/or multiple sequence alignments within the protein family.
2. The tertiary template (set of spatial positions of residues) must be aligned with the amino acid sequence that is a putative member of the same family. This step represents in the simplest case a multiple alignment with other sequences of the family or a profile analysis. In more complicated situations, threading techniques as reviewed in the previous Section (IV.B) may be the method of choice.
3. Given the 3-D–1-D alignment, the new backbone of the protein being modeled

is constructed. Special techniques are typically necessary for constructing the loop regions that are usually the structurally variable regions (SVR).

4. The conformations of the side chains anchored at the new backbone are determined.
5. The structure proposal is finally subjected to several cycles of refinement (with molecular-mechanical methods, see Section IV.A) and the application of verification criteria (based on stereochemical knowledge and structure database statistics, see Section IV.B).

The simple procedures in the early years requiring repeated human intervention have been gradually replaced by more sophisticated and generally automated techniques. In fact, there is not so much to predict. Inasmuch as the fold is known, only local structural details need be tuned to comply with energetic and/or database criteria. The following sections review approaches applicable to solving tasks 1, 3, and 4.

A conceptionally different but poorly investigated approach to homology modeling is based on distance geometry. In this perspective, the tertiary template restrictions are translated into distance restraints that are used as input for distance geometry programs (Havel and Snow, 1991; Taylor, 1993; Sali and Blundell, 1993). As reviewed by Johnson *et al.* (1994), the restraint-based modeling is at present inferior to the traditional approach and is not discussed here.

Site-directed mutagenesis aimed at changing physical and chemical properties of proteins (e.g., engineering enhanced thermostability) is a specific application for homology modeling methods because, as a rule, only a few amino acids are changed. The conformational space to be searched therefore, is, not very large and enumeration techniques can be applied. With a well-refined, closely homologous structure as starting point, the model can achieve accuracies in the range of 1 Å (Vriend and Eijsink, 1993; De Fillipis *et al.*, 1994).

2. Structure Comparison and Definition of Structurally Conserved Regions

Protein tertiary structure comparison is necessary to elucidate topologically equivalent regions, to determine structural differences in space, and to find insertions/deletions in one structure relative to others. The classic measure of distance between two structures is the r.m.s.d. of the distance between equivalent C_α atoms after spatial superposition. In such a metric space, structures may be classified into families (Pascarella and Argos, 1992b; Holm *et al.*, 1992; Yee and Dill, 1993; Orengo *et al.*, 1993; Holm and Sander, 1994a; Rufino and Blundell, 1994; Lessel and Schomburg, 1994). To date, superposing 3-D protein structures provides the most sensitive technique for recognizing very distant relationships between amino acid sequences with low residue identities (approaching 2%). Holm and Sander (1994b) have recently compiled examples of unexpected topological similarities verified with structure superposition.

Early comparison techniques involve rigid body superpositions (McLachlan, 1972; Kabsch, 1976; McLachlan, 1982) with the assignment of initial equivalences determined by visual inspection that can be iteratively updated (Rossmann and Argos, 1977) and with systematic variations of the relative orientation of the two structures (Rossmann and Argos, 1976). For largely divergent structures, window comparison techniques relying on successive oligopeptide superpositions along the sequence have been applied (Remington and Matthews, 1980; Karpen *et al.*, 1989) to locate similar substructures. An efficient algorithm for multiple rigid-body superposition has been described (Diamond, 1992).

To handle rigorously variable-length gaps in the alignment, structure alignment methods based on dynamic programming have been developed (Taylor and Orengo, 1989; Zuker and Somorjai, 1989; Sali and Blundell, 1990). The sensitivity of the technique has been enhanced by including hydrogen bonding, solvent exposure, torsional angles, secondary structural assignments, and the like (Taylor and Orengo, 1989; Sali and Blundell, 1990; Orengo *et al.*, 1992; Flores *et al.*, 1993) in addition to α-carbon distances. Dynamic programming cannot directly incorporate non-local effects. Therefore, some authors use multiple levels of dynamic programming (Taylor and Orengo, 1989; Orengo and Taylor, 1990), self-consistency tests for suboptimal alignments (Luo *et al.*, 1993), or stochastic optimization via genetic algorithms (May and Johnson, 1994). For the latter technique, an initial alignment is no longer needed, which is a problem in many methods. For example, Rose and Eisenmenger (1994) use geometric criteria applied to splines through C_α atoms for an automatic initial guess. The performance of such techniques is poor if the initial alignment proves to be difficult as in the case of NAD-binding proteins (Rose and Eisenmenger, 1994). Techniques for multiple structural alignment allowing gaps have also been developed, for example, the STAMP program (Russel and Barton, 1992).

Two-dimensional distance plots with the residue numbers of two structures at both axes showing the pairwise C_α distances are very useful representations of protein 3-D-structural information because of the independence of the sequential order of structural blocks. Apart from pairwise C_α distances, other characteristics such as hydrogen bonds (Factor and Mehler, 1991) or main chain dihedral angle matches (Karpen *et al.*, 1989) can be stored in distance plots. Parts of these plots can be compared to search for similar substructures or patterns (Barton and Sternberg, 1988; Richards and Kundrot, 1988). In two approaches (Vriend and Sander, 1991; Holm and Sander, 1993), the algorithm tries to assemble similar fragments into larger substructures with iterative 3-D-structural superpositions or a Monte Carlo strategy, respectively, and to delineate the largest common substructure. This is especially interesting for automatically searching structure databases to compile lists of structurally homologous domains. Two other efficient comparison techniques for extraction of conserved structural motifs but also independent of the sequential order of substructural elements have been published: one is based on iterative superposition of hexapeptides (Alexandrov *et al.*, 1992; Alexandrov and Go, 1994), and the other uses geometric hashing techniques (Bachar *et al.*, 1993).

A completely different approach consists in defining overall topological equivalence of secondary structural blocks rather than detailed atomic correspondence. Characteristic patterns of secondary structure are much more robust to structural changes than individual amino acid positions, and many mutations often do not destroy the overall main chain topology. Simplified representations of building blocks in the form of vectors, for example, along the axes of secondary structural elements, can be compared (Murthy, 1984; Richards and Kundrot, 1988; Abagyan and Maiorov, 1988; Abagyan and Maiorov, 1989). In graph theoretic approaches, protein structural elements and their relations are coded in the form of nodes and edges. The algorithms attempt to delineate common subgraphs (Mitchell *et al.*, 1989; Subbarao and Haneef, 1991; Grindley *et al.*, 1993; Mezei, 1994). The structures may also be described as strings of symbols denoting structural blocks that can be aligned by the dynamic programming technique (Matsuo and Kanehisa, 1993).

Structurally conserved regions may be defined not only from comparisons of a family of related structures but also from a multiple sequence alignment and profile analysis within a protein family. If the structure of one family member is known, the highly conserved sequence stretches indicate which part of the structure may be considered a SCR.

3. Backbone and Loop Modeling

The result of the sequence-structure alignment is a set of spatial positions of centers of amino acid residues (usually the C_α atoms) of the sequence under study. The backbone of the structurally conserved regions is usually copied by rigid body superposition from the tertiary template or a suitable fragment that is part of a homologous structure (Johnson *et al.*, 1994). If only the α-carbon positions of the template are known, there is a variety of algorithms able to reconstruct the backbone from the C_α-trace with fragment databases (Claessens *et al.*, 1989; Levitt, 1992) and also without direct database reliance, by exploiting geometric properties of peptide groups (Purisima and Scheraga, 1984; Luo *et al.*, 1992; Rey and Skolnick, 1992; Mandal and Linthicum, 1993) or their interactions (Payne, 1993; Liwo *et al.*, 1993), molecular dynamics based techniques (Correa, 1990; van Gelder *et al.*, 1994), and complete enumerations with energy minimization (Bassolino-Klimas and Bruccoleri, 1992). The accuracy expressed as r.m.s.d. between equivalent atoms of the backbone atoms is ~1 Å.

Loop regions between secondary structure elements accommodate most of the residue replacements, insertions, and deletions (Sibanda and Thornton, 1991; Pascarella and Argos, 1992a) and, consequently, constitute most of the structurally variable regions (SVR). The conformation of SVRs is defined only by the spatial location of the anchoring atomic groups at the termini of the corresponding SCR and by the interaction energy with the remainder of the protein and with the solvent. Ngo and Marks (1992) demonstrated that this modeling task is an NP-complete problem. As a result, grid search techniques (Moult and James, 1986) exhaust computer resources even for medium-sized loops.

Loop modeling is the most speculative part of homology modeling and, therefore, least automated. Categorizations of loop families may eventually lead to modeling rules or to a set of backbone motifs, especially for short turns (Sibanda and Thornton, 1991; Ring *et al.*, 1992). Typical flexibility of loops has generally not been observed in those involved in protein function and the active site. For example, key residues dictate one of the few possible main chain conformations in hypervariable loops in immunoglobulins (Tramontano *et al.*, 1989; Tramontano *et al.*, 1990). The canonical structures in hypervariable loops of immunoglobulins are surprisingly conserved throughout the vertebrate classes (Barré *et al.*, 1994).

Several conformation-generating techniques have been applied for loop construction. As proposed by Greer (1981), database fragments may be used to fill the gaps between the SCR (Jones and Thirup, 1986; Levitt, 1992; Wendoloski and Salemme, 1992; Topham *et al.*, 1993), which subsequently are relaxed by energy refinement. Other techniques rely on conformational search and energy minimization (Palmer and Scheraga, 1991; Palmer and Scheraga, 1992; Bruccoleri, 1993; Collura *et al.*, 1993; Vasmatzis *et al.*, 1994; Hubbard *et al.*, 1994). The loop closure is achieved with an analytical routine as used in the conformational analysis of cyclic organic structures or nucleic acids (for small loops) or with a penalty function on distances to atoms at the termini of SCR (for medium-size or long loops). Zheng *et al.*, (1993a; 1993b) close the backbone gap in the initial loop model by rescaling the bond lengths in the loop segments with their subsequent relaxation during energy minimizations.

Usually, a good accuracy of loop modeling is achieved only in examples for closely related proteins. If the protein environment of the loop is known in atomic detail, packing restraints will help to select the correct conformations. Monte Carlo search techniques in internal coordinates have demonstrated their predictive power in such cases (Palmer and Scheraga, 1992; Collura *et al.*, 1993; Vasmatzis *et al.*, 1994). For example, the conformation of an eight-residue loop in a triose phosphate isomerase mutant was computed successfully with a biased Monte Carlo conformational search in the space of internal coordinates (Borchert *et al.*, 1993). The resulting r.m.s.d. to the crystal structure was 0.6 Å.

Problems in homology modeling arise if the sequence of the template is evolutionary distant to the query sequence. In contrast to gross structural features that are generally well predicted, certain regions may be misaligned by a few residues. As a consequence, incorrect facing of α-helices or misaligned β-strands together with wrong side chain packing may be obtained. Optimization and energy refinement with molecular mechanical methods are not able to remove these defects but can introduce new distortions because it would be necessary to reorient complete secondary structural elements with disruption of hydrogen bonds. An r.m.s.d. for the C_α positions below 2 Å can be reliably achieved only for closely related structures.

4. Side Chain Placement

The methods for placing side chains at a fixed backbone can be classified into three groups with respect to the degree of reliance on databases.

1. Side chain dihedral angles are copied from corresponding residues in the original structure (Summers and Karplus, 1989; Schiffer *et al.*, 1990), from matching protein segments (Levitt, 1992) or from similar local residue environments (Laughton, 1994a) found in the protein structure database. This latter approach is also used in the package COMPOSER by Sutcliffe *et al.* (1987). The next step is an energy refinement with different conformational space search methods

(Summers and Karplus, 1989; Levitt, 1992; Laughton, 1994b).

2. A rotamer library is compiled from a set of high-resolution protein structures. The assignment of side chain conformation is translated into a discrete combinatorial problem. The solution has been attempted with optimization techniques such as genetic algorithms, simulated annealing, heuristic sparse-matrix driven searches (Tuffery *et al.*, 1991; Tuffery *et al.*, 1993; Holm and Sander, 1992b), energy minimization (Wilson *et al.*, 1993; Kono and Doi, 1994), and iterative searches in a self-consistent mean field (Koehl and Delarue, 1994b). The "dead-end-elimination" theorem (Desmet *et al.*, 1992) can reduce the number of rotamer combinations in a complete enumeration (Ponder and Richards, 1987). This approach was later corrected (Lasters and Desmet, 1993) and enhanced (Goldstein, 1994). A final energy minimization relaxes rotamer-rotamer clashes. Dunbrack and Karplus (1993; 1994) use a backbone-dependent rotamer library and achieve 59 to 86% correct prediction of rotamer class for various amino acid types in contrast to 46 to 76% for a backbone-independent library.
3. The third group consists of brute-force approaches that rely completely on Monte Carlo search (Lee and Subbiah, 1991), molecular dynamics protocols (David, 1993; David, 1994; van Gelder *et al.*, 1994), or enumeration techniques (Bassolino-Klimas and Bruccoleri, 1992). These methods are usually very computer time consuming and are not applicable to large proteins.

It has been demonstrated that the available conformation space for side chains is essentially restricted due to interactions with (repetitive) backbone segments (Schrauber *et al.*, 1993; Dunbrack and Karplus, 1993; Dunbrack and Karplus, 1994). Similar prediction accuracy compared with the methods listed above was achieved in a very simple manner by first minimizing for each residue individually the internal torsion energy and the non-valent and hydrogen-bond interactions with only the backbone (GAP model — all amino acid side chains are reduced to C_β atoms except glycines and prolines) and then starting a global minimization with each side chain in its local optimum (Eisenmenger *et al.*, 1993). Of all rotameric states, 47 to 60% could be correctly determined depending on the protein and amino acid type. The interactions of the side chains with the backbone are much more significant than those with other side chains.

Eisenmenger *et al.* (1993) conclude that there is hardly any jig-zaw puzzle in side chain packing by comparing the GAP-prediction with the ALL-model calculations where the remaining side chains are included in their correct conformation during local minimization. This conclusion looks especially convincing for crambin and the avian pancreatic polypeptide (see Tables 4 and 5 in Eisenmenger *et al.* [1993]). A more critical evaluation of the results reveals that a significant combinatorial effect for side chain packing remains because, for larger proteins such as uteroglobin and plastocyanin having a well-recognizable core, the number of correctly predicted rotamers increases by up to 20%, and the r.m.s.d. for core residues relative to their observed conformation in the X-ray structure drops from 0.8 to 1.2 Å to 0.4 to 1.0 Å (Eisenmenger *et al.*, 1993) after changing from the GAP-model to the ALL-model. A similar result may be expected for small proteins as avian pancreatic polypeptide and crambin if the crystallographic environment is taken into account because their core motif involves subunit interactions or crystal packing. We conclude from the results of Eisenmenger *et al.* (1993) that the combinatorial effect probably involves about 20% of the side chain conformations.

Surface residues are less restricted by main chain or side chain contacts. The prediction of surface residue conformations can be improved by considering explicit solvation terms during optimization (Schiffer *et al.*, 1990; Wilson *et al.*, 1993).

Side chain placement methods are often classified in two groups: continuum and discrete search techniques for the χ-angle conformation space. Discrete conformational states for side chains ("rotamers") can be defined in various ways, for example, as

- Intraresidue local minima of potential energy
- As spot of high density in χ-angle plots from statistics in known protein structures
- Simply as points of a grid covering the complete χ-angle space

Schrauber *et al.* (1993) demonstrated that about 20% of the side chains cannot be well modeled with a traditional rotamer library based on the first two principles only. This drawback of the discrete approach is removed if rotamers are constructed with a grid sufficiently fine to allow accurate approximation for all side chains, albeit many more rotameric states have to be considered.

The normal test of performance quality for a given algorithm consists of comparing the predicted side chain conformation with that in the protein crystal structure. Unfortunately, the criteria to assess the performance of side chain conformation prediction techniques and also the protein examples vary among researchers, making comparison difficult.

R.m.s. deviations between equivalent side chain atoms are in wide use despite their low information value. Small side chains (with one χ-angle) can occupy different energetic wells and still have low r.m.s.d. values. The same may occur, but more infrequently, with very large ones. Additionally, the squared deviation for the whole protein is averaged over many atoms, and, consequently, depends on the relative number of small and large residues. It appears reasonable to calculate the r.m.s.d. separately for each residue type (Dunbrack and Karplus, 1993) or to choose the maximal deviation between equivalent atoms (Schrauber *et al.*, 1993) as a relevant measure. Typical side chain atom r.m.s.d. values are 1.3 to 2.1 Å if the native sequence is built onto the correct backbone (0.7 to 1.7 Å for only the core residues) and 1.6 to 2.7 Å if the backbone was reconstructed by homology (Wilson *et al.*, 1993; Eisenmenger *et al.*, 1993). Therefore, the error in the location of certain functional groups at the tip of the side chain may be considerable.

A side chain dihedral angle is considered to be well predicted if it differs less than 30° (Laughton, 1994b; Laughton, 1994a), less than 40° (Summers and Karplus, 1989; Eisenmenger *et al.*, 1993; Dunbrack and Karplus, 1994; Koehl and Delarue, 1994), or if it belongs to the same rotameric class (Wilson *et al.*, 1993; Swindells and Thornton, 1993), all being very weak criteria. Typically, about 85% of all rotameric states in the core and nearly 50% of the exposed side chains are predicted correctly. Schrauber *et al.* (1993) applied energetic and geometric criteria to assess the deviation of a side chain conformation from a rotamer. A variance of more than 20° in at least one χ-angle implies an internal strain of the side chain of no less than 1 to 2 kcal/mol, and the maximal distance between equivalent side chain atoms exceeds 0.5 Å for small amino acids, 0.7 Å for nonaromatic and 1.6 Å for aromatic amino acids. However, a similarly good accuracy is expected in modeling studies for drug design. The criterion ±20° received another physical justification recently. It was shown that a harmonic approximation of the energy well corresponding to a rotamer works well up to an angular deviation of 20° (Lee *et al.*, 1994).

Additionally, the crystal structure that is the standard-of-truth may be not reliable. The side chain conformations in interleukin-I crystals (2 Å resolution) that have been analyzed independently by three groups differ considerably. For example, the same χ_1 energy well was assigned to 95% of the core residues and to 61% of the surface residues. Out of the 40 buried residues that show χ_1 angle agreement, only 75% of those possessing a χ_2 angle also agree (Swindells and Thornton, 1993). Side chains are generally well resolved at only a better resolution (~1.8 Å or less). It has been shown that the frequency of strained side chain conformations in protein crystal structures decreases up to a resolution of 1.8 Å (Schrauber *et al.*, 1993). However, with further improvement of resolution, this number actually increases for many amino acid types (Table 3). This result seriously questions the traditional rotamer approach because large deviations from rotameric states occur systematically (5 to 30% of all side chains depending on amino acid type) and cannot be traced to errors in the structures analyzed (Table 3). Often, the nonrotameric state is caused by tertiary side chain-side chain interactions (Figure 7); therefore, the number of strained side chains (about 20%) is another estimate for the extent of the combinatorial effect in side chain packing. The expected number of strained side chains under the assumption of independent conformations can be estimated with the Boltzmann distribution and is only about 4% for a 2 kcal/mol strain at room temperature. Apparently, the relaxation of some side chains is sacrificed for the global fold to be achieved (Schrauber *et al.*, 1993).

Future improvements in side chain predictions appear to be tightly connected with the improvement of molecular-mechanical energy functions because the energetic costs for straining a side chain and the energetic differences between varous rotameric states are not very large. More sophisticated techniques for searching the side chain conformational space will show their advantage in comparison with existing algorithms only in conjunction with a more discriminative energy function. The next generation of side chain placing algorithms may profit from both classes of continuum and discrete search techniques with successive steps of testing combinatorial packing within a rotamer approach and of exploring the local energy minimum with a continuum method.

D. Recognition of Tertiary Structural Features with Sequence-Analytical Methods

The sequence-analytical methods described in this section have been used to give hints as to whether a query amino acid sequence may belong to a certain structural

TABLE 3
Percentage of Side-Chain Conformations Belonging to Rotamer Classes

Amino acid	Number of rotamers	2.0		1.89		1.59		$\Delta\chi$=	Number of rotamers for the optimal grid[c]
		N	R[a]	N	R[a]	N	R[a]	±10°[b]	
PHE	4	441	72.8	310	75.8	97	75.3	33.3	21
	5		81.9		84.8		81.4	37.3	
TYR	4	442	72.6	332	74.1	134	73.9	33.0	21
	4		75.8		76.2		74.6	36.7	
TRP	6	181	74.0	121	76.9	32	81.3	34.8	23
	6		78.5		80.2		78.1	41.4	
HIS	6	278	45.7	198	50.0	76	57.9	12.2	42
	6		68.0		71.7		69.7	22.6	
LEU	4	905	74.7	638	79.6	186	78.0	44.0	39
	9		84.1		86.5		87.1	49.3	
ILE	5	611	78.6	431	81.7	154	83.1	43.5	42
	7		83.1		86.5		88.3	50.2	
MET	3	224	58.9	147	61.9	56	62.5	22.3	29
	5		77.2		78.9		82.1	30.9	
ARG	7	409	62.6	297	64.0	91	65.9	26.7	43
	5		65.8		66.3		70.3	31.2	
LYS	7	723	55.2	504	57.1	191	66.0	26.0	48
	7		60.7		63.3		69.1	29.6	
GLU	7	594	64.8	403	68.0	133	69.2	30.3	51
	8		69.4		72.2		73.7	33.6	
GLN	7	416	67.5	311	68.2	78	65.4	30.5	45
	7		73.6		74.3		73.1	33.5	
VAL	3	901	91.3	669	93.1	226	94.7	71.5	9
	4		94.0		95.4		96.5	74.1	
THR	3	720	92.4	546	94.7	203	92.1	72.6	9
	3		92.6		95.1		92.6	76.1	
SER	3	936	83.4	685	83.9	259	86.5	56.6	9
	3		85.3		86.0		90.0	59.3	
CYS	3	288	91.0	223	92.4	73	91.8	65.6	7
	3		92.4		93.7		93.2	65.2	

TABLE 3
Percentage of Side-Chain Conformations Belonging to Rotamer Classes

Amino acid	Number of rotamers	2.0		1.89		1.59		$\Delta\chi$=	Number of rotamers for the optimal grid[c]
		N	R[a]	N	R[a]	N	R[a]	±10°[b]	
ASP	3	657	46.1	462	50.2	137	48.2	19.8	34
ASN	6	550	37.5	435	40.7	151	38.4	13.8	54

[a] Rotamericity (in accordance with the criterion $\Delta\chi_1 = \pm 20°$ and $\Delta\chi_2 = \pm 20°$) of the amino acids types. For each amino acid two values are given. The upper row corresponds to the rotamers in the library of Ponder and Richards (1987), while the lower row refers that of Schrauber *et al.* (1993). The rotamers in both sets correspond to local minima of the intraresidue potential energy and/or to peaks of distribution density in the c-angle space. It must be emphasized that the number of rotamer observations for both rotameric sets were derived from the same dataset corresponding to the tertiary structures listed in Table 1 of Schrauber *et al.* (1993). The number of rotamers in each library is listed. The three resolution thresholds for the crystal structure datasets are 2.0 Å, 1.89 Å, and 1.59 Å. Additionally, in the last two cases the R-factors were less than 0.19. "N" is the number of residues in the dataset. "R" is the percentage of them with torsions within ±20° of the rotamers. For ASP and ASN rotamericities, only the rotamers from the Ponder and Richards library were taken. The very low rotamericity values illustrate the wide distribution of their side-chain conformations. The rotamericities are lower at the 1.59 Å resolution limit compared with the values at 1.89 Å resolution for the amino acids PHE, TYR, TRP, HIS, GLN, THR, CYS, and ASN, that is, mainly for amino acids with bulky side chains for which the side torsion angles can be reliably determined using crystallographic methods.

[b] The rotamericity values for the criterion $\Delta\chi_{1,2} = \pm 10°$ (instead of $\Delta\chi_{1,2} = \pm 20°$) at the 2 Å resolution level are listed. As a rule, the rotamericity is halved for all amino acids with at least two torsion angles.

[c] Schrauber *et al.* (1993) estimated the number of rotamers to cover all sample points in the 2 Å resolution protein structures. A square grid with a spacing of 40° (in accordance to the criterion $\Delta\chi_1 = \pm 20°$ and $\Delta\chi_2 = \pm 20°$) was moved in the χ_1/χ_2-plot. The number of squares with at least one sample point in them was determined. The last column in this table contains the minimal number of occupied squares for the optimal grid. For VAL, THR, SER, and CYS the maximally possible number of rotamers for the grid described would be 9; for PHE, TYR, and ASP it is equal to 45; and for the remaining amino acids 81 rotamers cover the whole χ_1/χ_2-space. The data in the table shows that the residues are spread over about one half of the χ_1/χ_2-plot.

family of proteins. At the beginning, the known sequences and structures of a given protein family (folding class) are analyzed and multiply aligned. Then conserved sequence properties are extracted. The hypothesis that the query amino acid sequence (or a significant domain of it) may be related to a structural family is put forward if the sequence properties of the folding class are found in the new sequence.

From the logical point of view, the discriminative power of this approach should be limited because it is not *a priori* clear whether a given sequence property is

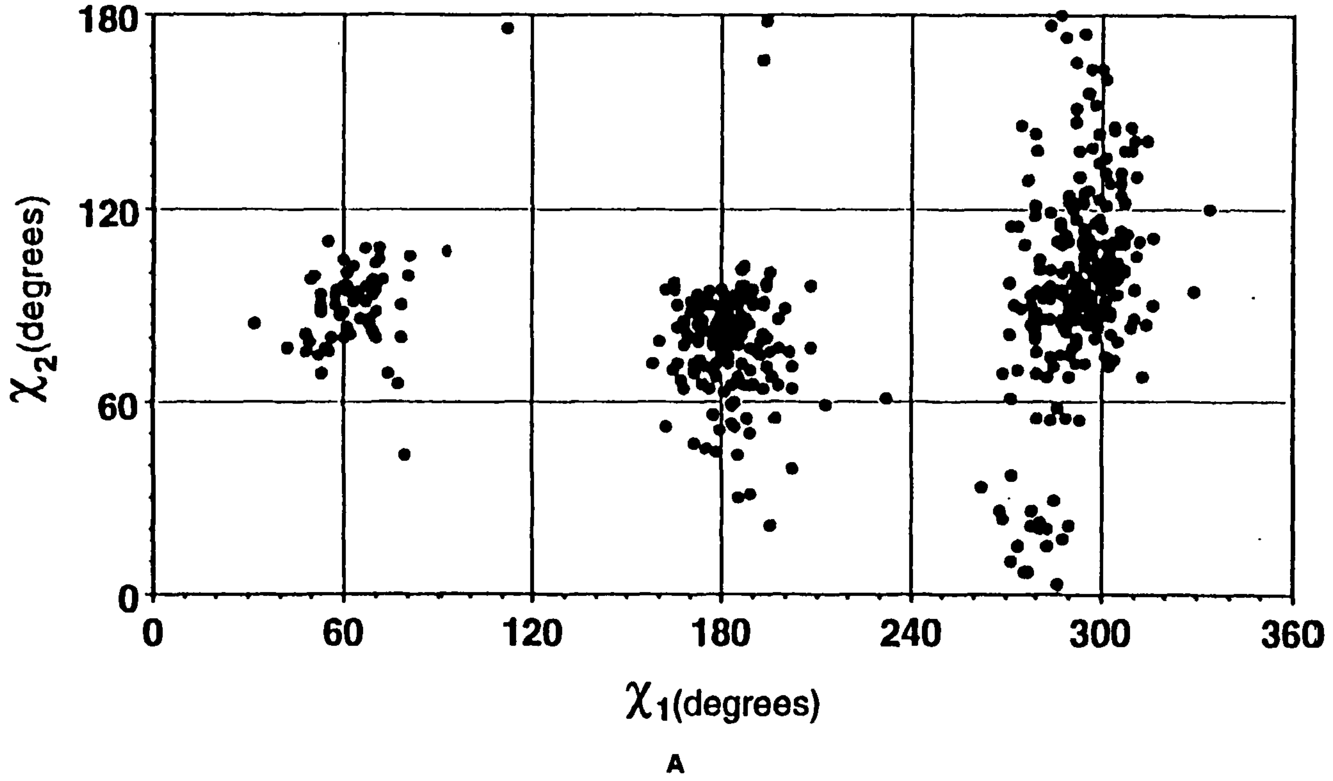


FIGURE 7. χ_1/χ_2-Plot for phenylalanine. Because of the symmetry, the range for χ_2 is reduced to the interval 0° to 180°. If we follow the standard nomenclature for χ-angles (range -180° to 180°), the cluster at χ_1 - 180° becomes divided. Therefore, χ_1 is reperesented in the range of 0° to 360°. As a result, the conformations g^-, t, and g^+ are associated with angles near 60°, 180°, and 300°, correspondingly. The figures are taken with permission from Schrauber *et al.* (1993) where the protein data set is described in detail. Three plots show the distribution for (A) all phenylalanine residues, (B) only for phenylalanine residues in regular a-helices, and (C) only for phenylalanine residues in β-sheets. The conformation space for all phenylalanine side chains is to a first approximation well described by three rotamers (-60°, 90°), (60°, 90°), and (180°, 90°). For β-sheets, this is obviously also the case. In the case of regular a-helices, the *trans*-conformation is strongly preferred in the χ_1-angle with χ_2 near 90°; yet in about one third of all phenylalanine residues, this conformation is not allowed because of atomic collisions, almost exclusively due to tertiary contacts with other (mostly aromatic) side chains. The two *gauche* conformations for χ_1 are suppressed because of clashes with the preceding turn of the helix. Inasmuch as the collision is more severe in g^+ than in g^-, the latter variant is adopted while the χ_2-angles are nearly uniformly distributed from 0° to 180°. The side chain attempts to find any acceptable conformation in the g^--region for the given tertiary packing.

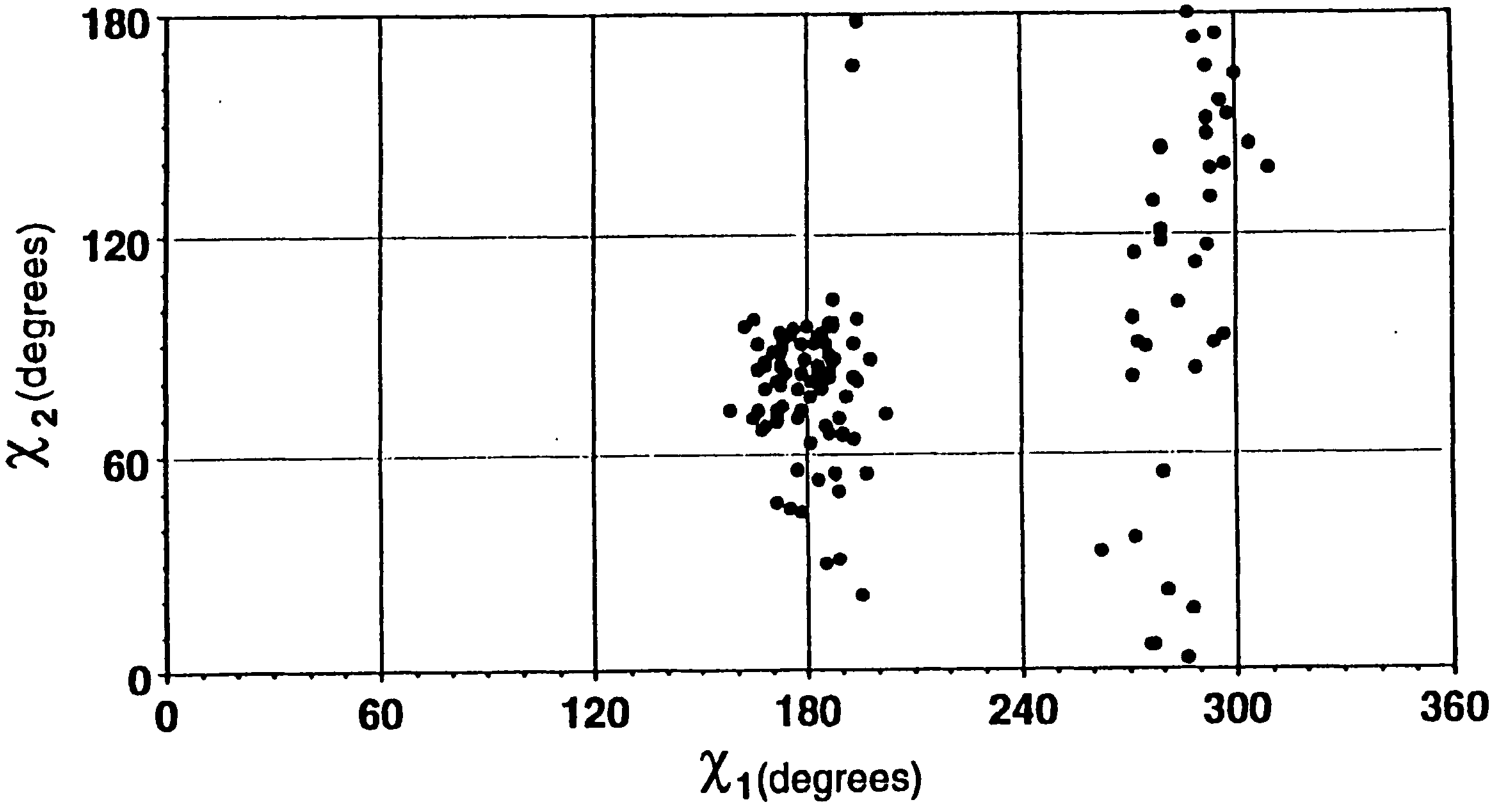


FIGURE 7B

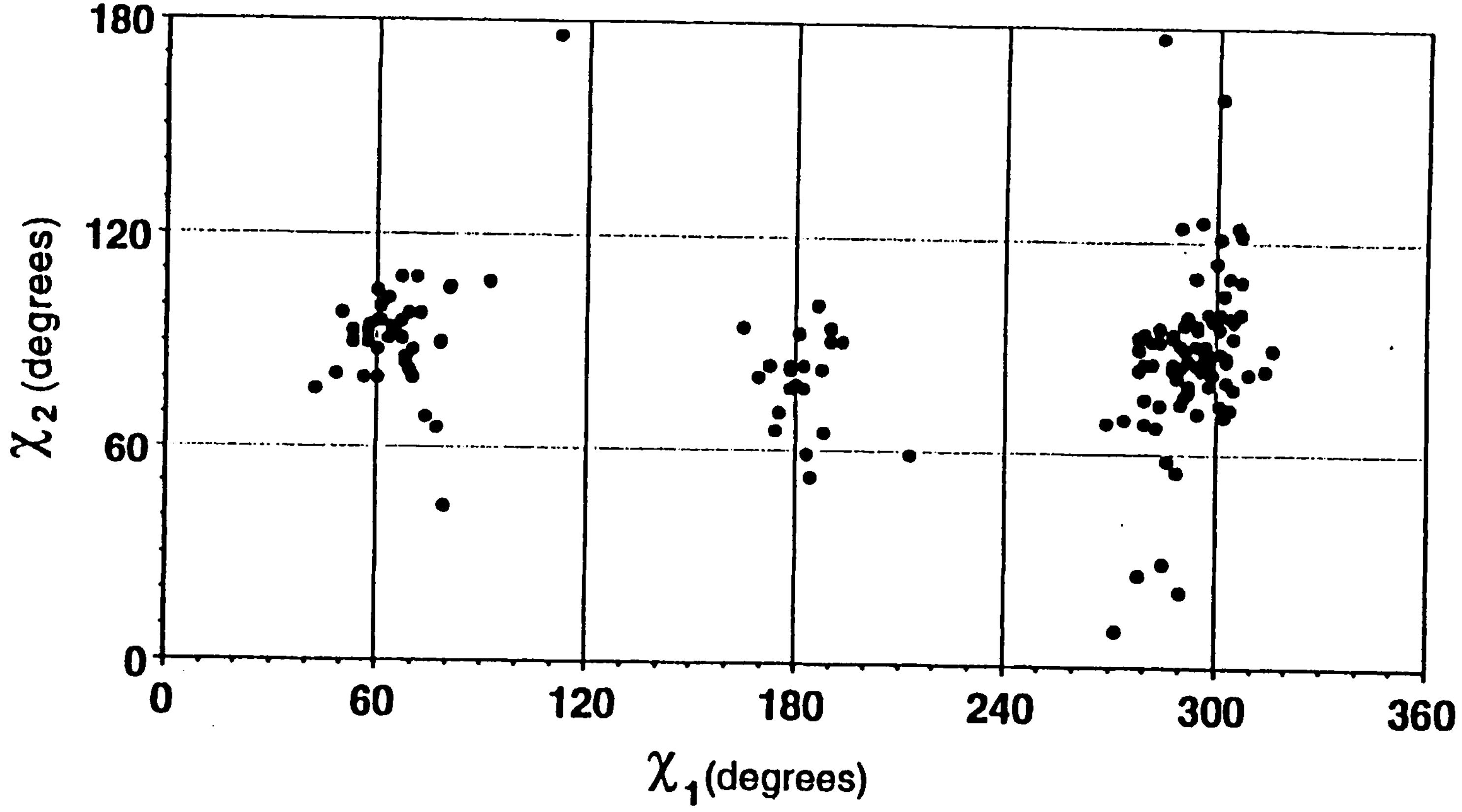


FIGURE 7C

characteristic for only the family under consideration. The reversibility of an implication is not obvious. For example, we cannot say that everything with a handle is a door and we may open it. However, in the practice of sequence analysis, this way of thinking is unexpectedly often successful. The structure and function of an unknown protein can be tested experimentally if it is supposed to be related to a protein family with well-studied members. It is at least worth investigating it further.

Various techniques have been applied for the purpose of predicting features of the tertiary structure of proteins from sequence. We suppose that the following list is by no means complete:

- Global and local sequence alignment, sometimes combined with secondary structure prediction
- Alignment with property profiles
- Search for patterns of amino acid types
- Search for patterns of physical properties of amino acids
- Discrimination by amino acid composition or dipeptide frequencies
- Generalization of sequence properties by neural networks.

Some of the techniques have been reviewed extensively in previous sections. The discussion here is restricted to some important aspects in the context of predicting tertiary structural motifs.

Sequence alignment techniques are used frequently to investigate possible structural similarity of proteins. Sequence identity above a threshold of 25 to 30% implies structural similarity with a high probability if the alignment length exceeds a threshold of about 70 residues (Doolittle, 1981; Chothia and Lesk, 1986; Sander and Schneider, 1991). The margin of 50% sequence identity was established as threshold below which many structural parameters such as r.m.s.d. between C_α atoms, residue accessibility, side chain torsion angles, and contact numbers are increasingly less conserved, although the overall fold and the secondary structural elements may not change (Hilbert *et al.*, 1993; Flores *et al.*, 1993; Rost *et al.*, 1994; Rost and Sander, 1994b; Chelvanayagam *et al.*, 1994).

With lower sequence similarity, typical sequence alignment methods can become unreliable. At the same time, sequences belonging to a protein structural family show sequence identities down to 3% (Pascarella and Argos, 1992b), which is clearly below the level of random expectation. The globin family, a remarkable example of topology conservation, displays sequence pairs with identities as low as 12% (Pascarella and Argos, 1992b). Pickett *et al.* (1992) showed that sequence similarity between $(\alpha\beta)_8$ barrel proteins is only marginal. As was demonstrated by Bell *et al.* (1993), the sensitivity of sequence alignment techniques can be enhanced with secondary structure prediction and interactive intervention taking into account biochemical and structural considerations. Through the use of automatic tools removing redundant matches and biased composition from the list of protein similarities obtained after database searches, the efficiency of using local alignment procedures such as BLAST for searching distantly related sequences may be improved (Sonnhammer and Durbin, 1994).

Distant relationships between sequences may be found by alignment to property profiles. Profile analysis (Gribskov *et al.*, 1987; Gribskov *et al.*, 1990; Bowie *et al.*, 1990; Lüthy *et al.*, 1991; Lüthy *et al.*, 1994; Thompson *et al.*, 1994; Henikoff and Henikoff, 1994) is a method to detect proteins distantly related to a known family; it compares a query sequence with a linear profile containing implicit information from multiple sequence alignments and/or structural results such as the secondary structure assignment. As a result, both the gap penalty and the amino acid preferences are position dependent in the alignment. At present, profile analysis is the standard method for detecting distant relationships between amino acid sequences and for large database searches. A recent example is the detection of the "pleckstrin homology" domain in more than 70 sequences since 1993 (Musacchio *et al.*, 1993; Gibson *et al.*, 1994). The tertiary structures of a few members of this family are known at present.

The recognition of one or several patterns of amino acid types can be considered as a hint pointing to some protein family (see Section II.B). A large compilation of single sequence patterns is contained in the database PROSITE (Bairoch, 1993). Another database PRINTS with "fingerprints" comprising sets of motifs that characterize folds better than single patterns has become available recently (Attwood and Beck, 1994). Neural networks have been trained with direct input of conserved sequence pieces in multiple alignments (Frishman and Argos, 1992). The trained nets have been subsequently applied for database searches to detect other sequences containing the same conserved regions. The method performed better than profile search for integrases, DNA-polymerases, and immunoglobulins.

While the search for amino acid type patterns is similar to local alignment, the pattern search technique of Rohde and Bork (1993) resembles features of profile analyses. From a multiple alignment of conserved sequence segments characterizing a structural family or a structural feature, patterns of conserved physical properties of amino acids are derived. The algorithm allows the inclusion of gaps. The MAKPAT/PROPAT technique of Rohde and Bork (1993) has shown its efficiency in the detection of remote homologies and in the localization of structural elements. The ATPase domain of actin and hsc70 was found in prokaryotic cell cycle proteins and sugar kinases (Bork *et al.*, 1992). Shuffled domains were observed in various mosaic (mainly extracellular) proteins (Bork, 1992). A helix capping motif with glycine has been identified for 501 out of 575 existing cases in a set of selected proteins from the Brookhaven Protein Data Bank (Preissner and Bork, 1991; Bork and Preissner, 1991).

Busetta (1988) considers the protein fold a chain of secondary structural elements. The ability of the query amino acid sequence to fit to this constraint is measured by a function of amino acid secondary structural propensities. The method has been applied to discuss model topologies for phosphoribosyl transferases. The sensitivity of the technique is comparable with secondary structure prediction methods using residue propensities (see Section III.B).

Amino acid composition was shown to discriminate strongly between protein structural classes (folding types). Some authors suggest to use the same criterion for prediction of folding class (protein structural family). For example, Dubchak *et al.* (1993) use the amino acid composition as input for neural networks trained to recognize four-helix bundles, parallel $(\alpha\beta)_8$ barrels, nucleotide binding folds, and immunoglobulin folds. For proteins not contained in the training set, their prediction accuracy was above 80%.

Dipeptide frequencies have been utilized to predict GTP-binding domains as contained in the translocation elongation factors (Solovyev and Makarova, 1993) and for clustering a sequence database into protein families (van Heel, 1992). The matrix of size 20 × 20 containing dipeptide frequencies in a query sequence was used as input for neural networks for checking the relatedness to 45 folding classes and 4 folding types (Reczko *et al.*, 1994; Reczko and Bohr, 1994). The prediction accuracy was reported larger than 80%. Every amino acid sequence of a sequence database was assigned to one of these classes in the DEF database (Reczko and Bohr, 1994). Based on known structures belonging to the recognized protein family, distance matrices are predicted for the query sequence, which may be used to construct an approximate low-resolution 3-D structure.

The amino acid accessibility is an important tertiary property the knowledge of which can reduce the conformational space to be searched. Ycas (Ycas, 1990) has estimated the distance of an amino acid residue from the protein midpoint from its hydrophobicity. Neural network approaches to the prediction of amino acid accessibility have been described (Holbrook *et al.*, 1990; Rost and Sander, 1994b). Another method is based on environment-specific amino acid substitution tables (Wako and Blundell, 1994a). Because of the lower conservation of accessibility within protein families, the improvement in prediction accuracy from the use of multiply aligned sequences is not as large as in the case of secondary structure prediction and the correlation between the predicted and observed accessibilities is only in the range of 0.36 to 0.77 for different sets of sequences (Rost and Sander, 1994b).

V. CONCLUSION

It is clear from this review that the task of predicting a protein's atomic resolution tertiary structure from sequence information is both daunting and yet essential, especially given the vast gap between the number of known primary and tertiary structures. Nonetheless, as also shown in this treatise, considerable progress is in evidence and the goal will likely be achieved in many small but significant steps over time, especially as no new chemical principles will probably be needed beyond those already known.

Already on the horizon are methods to predict the backbone fold to within 2 to 5 Å r.m.s. deviation from the sequence of small, mostly helical proteins. Techniques to place substituted side chains for half of the residues in normal sized proteins are presently available with accuracies bordering 1.5 Å r.m.s. deviation.

The ever-increasing number of experimentally determined primary and tertiary structures bodes well to crack the sequence-to-structure riddle; more information in the form of multiple sequence and structure alignments can only help with resulting knowledge of the constraints and a variety of residue types possible at the various main chain sites. It can be expected that computational techniques will appear to optimize use of this evermore abundant information. Evidence is also available that most, if not all, energetic potentials will be formulated with accuracy in the not-so-distant future; the present day nemeses of electrostatics and entropic estimations for the sidegroups as well as solvent are increasingly lessened. Such progress provides hope that the sequence/folding relationship can ultimately be expressed in an *ab initio* rather than heuristic fashion, representing profound comprehension of protein stability and functional mechanisms.

ACKNOWLEDGMENTS

The authors thank Peer Bork, Jaap Heringa, Simon Hubbard, Burkhard Rost, Gerrit Vriend, and Dirk Walther for reading and/or discussing various portions of this manuscript. The authors acknowledge financial support from the German Bundesministerium für Forschung und Technologie (grant FG5-1075 to P. A.), from the Fund "Wisssenschaftler-Integrationsprogramm" (grant 020386/B to F. E.) jointly administered by the East-German Länder and the government of the Federal Republic, and from the Swedish Medical Research Council (project no. 13F–10248). F. E. is grateful to Cornelius Frömmel for constant support and encouragement.

REFERENCES

Abagyan, R., Frishman, D. I., and Argos, P., Recognition of distantly related proteins through energy calculations. *Proteins*, **19**, 132–140 (1994a).

Abagyan, R., Totrov, M., and Kuznetsov, M., ICM — a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comp. Chem.*, **15**, 488–506 (1994b).

Abagyan, R. and Totrov, M., Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, **235**, 983–1002 (1994).

Abagyan, R. A., Towards protein folding by global energy optimization. *FEBS Lett.*, **325**, 17–22 (1993).

Abagyan, R. A. and Argos, P., Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.*, **225**, 519–532 (1992).

Abagyan, R. A. and Maiorov, V. N., A simple quantitative representation of polypeptide chain folds: comparison of protein tertiary structures. *J. Biomol. Struct. Dyn.*, **5**, 1267–1279 (1988).

Abagyan, R. A. and Maiorov, V. N., An automatic search for similar spacial arrangements of alfa-helices and beta-strands in globular proteins. *J. Biomol. Struct. Dyn.*, **6**, 1045–1059 (1989).

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., and Weng, J., Protein data bank. Allen, F. H., Bergerhoff, G., and Sievers, R., Eds., *Crystallographic databases — information content, software systems, scientific applications*, pp. 107–132. Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography (1987).

Adzhubei, A. A. and Sternberg, M. J. E., Left-handed polyproline II helices commonly occur in globular proteins. *J. Mol. Biol.*, **229**, 472–493 (1993).

Aleman, C. and Oroczo, M., On the suitability of semiempirical calculations as sources of force field parameters. *J. Comp.-Aid. Mol. Des.*, **6**, 3311–3348 (1992).

Alexandrov, N. N., Takahashi, K., and Go, N., Common spacial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.*, **225,** 5–9 (1992).

Alexandrov, N. N. and Go, N., Biological meaning, statistical significance, and classification of local spatial similarities in non-homologous proteins. *Prot. Sci.*, **3,** 866–875 (1994).

Anfinsen, C. B., Principles that govern the folding of protein chains. *Science,* **181,** 223–230 (1973).

Argos, P., Analysis of sequence similar pentapeptides in unrelated protein tertiary structures. *J. Mol. Biol.*, **197,** 331–348 (1987).

Argos, P., Sensitive methods for determining the relatedness of proteins with limited sequence homology. *Curr. Opin. Biotech.*, **5,** 361–371 (1994).

Arnold, G. E., Dunker, K., Johns, S. J., and Douthart, R. J., Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure. *Proteins,* **12,** 382–399 (1992).

Aronson, H.-E. G., Royer, W. E., Jr., and Hendrickson, W. A., Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. *Prot. Sci.*, **3,** 1706–1711 (1994).

Aszodi, A. and Taylor, W. R., Folding polypeptide α-carbon backbones by distance geometry methods. *Biopolymers,* **34,** 489–505 (1994).

Attwood, T. K. and Beck, M. E., PRINTS — a protein motif fingerprint database. *Prot. Eng.*, **7,** 841–848 (1994).

Bachar, O., Fischer, D., Nussinov, R., and Wolfson, H., A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Prot. Eng.*, **6,** 279–288 (1993).

Bairoch, A., The Prosite dictionary of sites and patterns in proteins, its current status. *Nucl. Acid Res.*, **21,** 3097–3103 (1993).

Bairoch, A. and Boeckmann, B., The SWISS-PROT protein sequence data bank, recent developments. *Nucl. Acid Res.*, **21,** 3093–3096 (1993).

Baker, D., Sohl, J. L., and Agard, D. A., A protein-folding reaction under kinetic control. *Nature,* **356,** 263–265 (1992a).

Baker, D., Sohl, J. L., and Agard, D. A., Protease Pro region required for folding is a potent inhibitor of the mature enzyme. *Proteins,* **12,** 339–344 (1992b).

Barnes, E. J. and Hut, P., Error analysis of a tree code. *Astrophys. J. Suppl. Ser.*, **70,** 389–417 (1989).

Barré, S., Greenberg, A. S., Flajnik, M. K., and Chothia, C., Structural conservation of hypervariable regions in immunoglobins evolution. *Nature Struct. Biol.*, **1,** 915–920 (1994).

Barton, G. J. and Sternberg, M. J. E., LOPAL and SCAMP: techniques for the comparison and display of protein structures. *J. Mol. Graph.*, **6,** 190–196 (1988).

Bascle, J., Garel, T., Orland, H., and Velikson, B., Biasing a Monte Carlo chain growth method with Ramachandran's plot: application to twenty-L-alanine. *Biopolymers,* **33,** 1843–1849 (1993).

Bassolino-Klimas, D. and Bruccoleri, R. E., Application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates. *Proteins,* **14,** 465–474 (1992).

Bauer, A. and Beyer, A., An improved pair potential to recognize native protein folds. *Proteins,* **18,** 254–261 (1994).

Baumann, G., Frömmel, C., and Sander, C., Polarity as criterion in protein design. *Prot. Eng.*, **2,** 329–324 (1989).

Beglov, D. B. and Lipanov, A. A., Charge grouping approaches to calculation of electrostatic forces in molecular dynamics of proteins. *J. Biomol. Struct. Dyn.*, **9,** 205–214 (1991).

Bell, L. H., Coggins, J. R., and Milner-White, E. J., Mix'n'Match: an improved multiple sequence alignment procedure for distantly related proteins using secondary structure predictions, designed to be independent of the choice of gap penalty and scoring matrix. *Prot. Eng.*, **7**, 683–690 (1993).

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M., Protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542 (1977).

Beveridge, D. L. and DiCapua, F. M., Free energy via molecular simulation. *Ann. Rev. Biophys. Biophys. Chem.*, **18**, 431–492 (1989).

Biou, V., Gibrat, J. F., Lervin, J. M., Robson, B., and Garnier, J., Secondary structure prediction: combination of three different methods. *Prot. Eng.*, **2**, 185–191 (1988).

Blaber, M., Zhang, X.-J., and Matthews, B. W., Structural basis of amino acid alpha helix propensities. *Science*, **269**, 1637–1640 (1993).

Blaber, M., Zhang, X.-J., Lindstrom, J. L., Pepiot, S. D., Baase, W. A., and Matthews, B. W., Determination of alpha-helix propensities within the context of a folded protein: sites 44 and 131 in bacteriophage T4 lysozyme. *J. Mol. Biol.*, **235**, 600–624 (1994).

Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B., and Petersen, S. B., A novel approach to the prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Lett.*, **261**, 43–46 (1990).

Bohr, J., Bohr, H., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B., and Petersen, S. B., Protein structures from distance inequalities. *J. Mol. Biol.*, **231**, 861–869 (1993).

Borchert, T. V., Abagyan, R., Rahda Kishan, K. V., Zeelen, J. P., and Wierenga, R. K., The crystal structure of an engineered triosephosphate isomerase, monoTIM: the correct modeling of an eight-residue loop. *Structure* **1**, 205–213 (1993).

Bork, P., Mobile modules and motifs. *Curr. Opin. Struct. Biol.*, **2**, 413–421 (1992).

Bork, P., Sander, C., and Valencia, A., An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 7290–7294 (1992).

Bork, P. and Preissner, R., On α-helices terminated by glycine. II. Recognition by sequence patterns. *Biophys. Biochem. Res. Comm.*, **180**, 666–672 (1991).

Bouzida, D., Kumar, S., and Swendson, R. H., Efficient Monte Carlo methods for computer simulation of biological molecules. *Phys. Rev. A*, **45**, 8894–8901 (1992).

Bowie, J. U., Clarke, N. D., Pabo, C. O., and Sauer, R. T., Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins*, **7**, 257–264 (1990).

Bowie, J. U., Lüthy, R., and Eisenberg, D., A method to identify protein sequences that fold into a known three-dimensional stricture. *Science*, **253**, 164–170 (1991).

Braakman, I., Helenius, J., and Helenius, A., Role of ATP and disulfide bonds during protein folding in the endoplasmic reticulum. *Nature*, **356**, 260–262 (1992).

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M., CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, **4**, 187–217 (1983).

Brown, W. J., North, A. C. T., Phillips, D. C., Brew, K., Vanaman, T. C., and Hill, R. L., A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme, *J. Mol. Biol.*, **42**, 65–86 (1969).

Bruccoleri, R. E., Application of systematic conformational search to protein modeling. *Mol. Simul.*, **10**, 151–174 (1993).

Brünger, A. T. and Karplus, M., Molecular dynamics simulation with experimental restraints. *Acc. Chem. Res.*, **24**, 54–61 (1991).

Bryant, S. H. and Lawrence, C. E., An empirical energy function for threading protein sequence through the folding motif. *Proteins*, **16**, 92–112 (1993).

Busetta, B., The use of folding patterns in the search of protein structural similarities: a three-dimensional model of phosphoribosyl transferases. *Biochim. Biophys. Acta.*, **957**, 21–33 (1988).

Busetta, B. and Barrans, Y., The prediction of protein topologies. *Biochim. Biophys. Acta.*, **709**, 73–83 (1982).

Busetta, B. and Barrans, Y., The prediction of protein domains. *Biochim. Biophys. Acta.*, **790**, 117–124 (1984).

Bussian, B. M. and Sander, C., How to determine protein secondary structure in solution by Raman spectroscopy: practical guide and test case DNase I. *Biochemistry*, **28**, 4271–4277 (1989).

Byrne, D., Li, J., Platt, E., Robson, B., and Weiner, P. K., Novel algorithms for searching conformational space. *J. Comp.-Aid. Mol. Des.*, **8**, 67–82 (1994).

Casari, G. and Beyer, A. *personal communication* (1994).

Casari, G. and Sippl, M. J., Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins are able to identify native folds. *J. Mol. Biol.*, **224**, 725–732 (1992).

Chelvanayagam, G., Roy, G., and Argos, P., Easy adaptation of protein structure to sequence. *Prot. Eng.*, **7**, 173–184 (1994).

Chothia, C., One thousand families for the molecular biologist. *Nature*, **357**, 543–544 (1992).

Chothia, C. and Lesk, A. M., The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826 (1986).

Chou, K.-C. and Zhang, C.-T., A correlation-coefficient method to predicting protein-structural classes from amino acid composition. *Eur. J. Biochem.*, **207**, 429–433 (1992).

Chou, K.-C. and Zhang, C.-T., A new approach to prediction protein folding types. *J. Prot. Chem.*, **12**, 169–178 (1993).

Chou, P. Y., Prediction of protein structural classes from amino acid composition. Fasman, G. D., Ed., *Prediction of protein structure*, pp. 549–586. Plenum Press, New York (1989).

Chou, P. Y. and Fasman, G., Prediction of protein conformation. *Biochemistry*, **13**, 222–245 (1974a).

Chou, P. Y. and Fasman, G., Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211–222 (1974b).

Chou, P. Y. and Fasman, G., Prediction of secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.*, **47**, 145–147 (1978).

Chou, P. Y. and Zhang, C.-T., A joint prediction of the folding types of 1490 human proteins from their genetic codons. *J. Theor. Biol.*, **161**, 251–262 (1993).

Claessens, M., Van Cutsem, E., Lasters, I., and Wodak, S. J., Modeling the polypeptide backbone with 'spare parts' from known protein structures. *Prot. Eng.*, **2**, 335–345 (1989).

Claverie, J.-M., Database of ancient sequences. *Nature*, **364**, 19–20 (1993).

Cohen, B. I., Presnell, S. R., and Cohen, F. E., Pattern-based approaches to protein structure prediction. Langone, J. J., Ed., *Methods in Enzymology*, Vol. 202, pp. 252–268, Academic Press, San Diego (1991).

Cohen, B. I., Presnell, S. R., and Cohen, F. E., Origins of structural diversity within sequentially identical hexapeptides. *Prot. Sci.*, **2**, 2134–2145 (1993).

Cohen, F. E., Abarbanel, I. D., Kuntz, I. D., and Fletterick, R. J., Turn prediction in proteins using a pattern matching approach. *Biochemistry*, **25**, 266–275 (1986).

Cohen, F. E. and 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).

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., and Mornon, J.-P., Comparison of three algorithms for the assignment of secondary structure in proteins: the advantage of a consensus assignment. *Prot. Eng.*, **6**, 377–382 (1993).

Collura, V., Higo, J., and Garnier, J., Modeling of proteins loops by simulated annealing. *Prot. Sci.*, **2**, 1502–1510 (1993).

Colovos, C. and Yeates, T. O., Verification of protein structures: patterns of nonbonded atomic interactions. *Prot. Sci.*, **2**, 1511–1519 (1993).

Correa, P. E., The building of protein structures from α-carbon coordinates. *Proteins*, **7**, 366–377 (1990).

Cramer, C. J. and Truhlar, D. G., An SCF solvation model for the hydrophobic effect and absolute free energies of aqueous solvation. *Science*, **256**, 213–217 (1992).

Creamer, T. P. and Rose, G. D., Side-chain entropy opposes α-helix formation but rationalizes experimenbtally determined helix-forming propensities. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 5937–5941 (1992).

Creamer, T. P. and Rose, G. D., α-helix-forming propensities in peptides and proteins. *Proteins*, **19**, 85–97 (1994).

Creighton, T. E., *Protein Folding*, Freeman, New York (1992).

Crippen, G. M., Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, **30**, 4232–4236 (1991).

Crippen, G. M. and Snow, M., A 1.8-Å resolution potential function for protein folding. *Biopolymers*, **29**, 1479–1489 (1990).

Crippen, G. M. and Viswanadhan, V. N., Sidechain and backbone potential function for conformational analysis of proteins. *Int. J. Peptide Prot. Res.*, **25**, 487–509 (1985).

Dandekar, T. and Argos, P., Potential of genetic algorithms in protein folding and protein engineering simulations. *Prot. Eng.*, **5**, 637–645 (1992).

Dandekar, T. and Argos, P., Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.*, **236**, 844–861 (1994).

Dauber-Osguthorpe, P., Roberts, V. A., Osguthorpe, D. J., Wolff, J., Genest, M., and Hagler, A. T., Structure and energetics of ligand binding to proteins: *Escherichia coli* dihydrofolate reductase-trimethoprim, a drug-receptor system. *Proteins*, **4**, 31–47 (1988).

Dauber-Osguthorpe, P., and Osguthorpe, D. J., Partitioning the motion in molecular dynamics simulations into characteristic modes of motion. *J. Comp. Chem.*, **14**, 1259–1271 (1993).

David, C. W., Sprouting side chain conformations in X-PLOR simulations of peptides. *J. Comp. Chem.*, **14**, 715–717 (1993).

David, C. W., Hydrating peptides using a sprouting technique. *J. Comp. Chem.*, **15**, 23–27 (1994).

Davis, M. E., The inducible multipole solvation model: a new model for solvation effects on solute electrostatics. *J. Chem. Phys.*, **100**, 5149–5159 (1994).

Davis, M. E. and McCammon, J. A., Electrostatics in biomolecular structure and dynamics. *Chem. Rev.*, **90**, 509–521 (1990).

De Fillipis, V., Sander, C., and Vriend, G., Predicting local structural changes that result from point mutations. *Prot. Eng.*, **7**, 1203–1208 (1994).

Degli Eposti, M., Crimi, M., and Venturoli, G., A critical evaluation of the hydropathy profile of membrane proteins. *Eur. J. Biochem.*, **190**, 207–219 (1990).

Deisenhofer, J. and Michel, H., The photosynthetic reaction centre from the purple bacterium *Rhodopseudomonas viridis*. *EMBO J.*, **8**, 2149–2170 (1989).

Desmet, J., De Maeyer, M., Hazes, B., and Lasters, I., The dead-end elimination theorem and its use in side-chain positioning. *Nature*, **356**, 539–542(1992).

Diamond, R., On the multiple simultaneous superposition of molecular structures by rigid body superposition. *Prot. Sci.*, **1,** 1279–1287 (1992).

Ding, H.-Q., Karasawa, N., and Goddard, W. A., III, Atomic level simulations on a million particles: the cell multipole method for Coulomb and London interactions. *J. Chem. Phys.*, **97,** 4309–4315 (1992).

Donelly, D., Overington, J. P., and Blundell, T. L., The prediction and orientation of α-helices from sequence alignments: the combined use of environment-dependent substitution tables, Fourier transform methods and helix capping rules. *Prot. Eng.*, **7,** 645–653 (1994).

Doolittle, R. F., Similar amino acid sequences: chance or common ancestry. *Science,* **214,** 149–159 (1981).

Doolittle, R. F., Counting and discounting the universe of exons. *Science,* **253,** 677–679 (1991).

Dorit, R. L., Schoenbach, L., and Gilbert, W., How big is the universe of exons? *Science,* **250,** 1377–1381 (1990).

Dorit, R. L., Schoenbach, L., and Gilbert, W., Reply to R. F. Doolittle and L. Patthy. *Science,* **253,** 679–680 (1991).

Dorofeev, V. E., and Mazur, A. K., Investigation of conformational equilibrium of polypeptides by internal coordinate stochastic dynamics. Met^5-enkephalin. *J. Biomol. Struct. Dyn.*, **10,** 143–167 (1993).

Drexler, K. E., Molecular engineering: an approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U.S.A.*, **78,** 5275–5278 (1981).

Dubchak, I., Holbrook, S. R., and Kim, S.-H., Prediction of protein folding class from amino acid composition. *Proteins,* **16,** 79–91 (1993).

Dunbrack, R. L. and Karplus, M., Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.*, **230,** 543–574 (1993).

Dunbrack, R. L. and Karplus, M., Conformational analysis of the backbone-dependent rotamer preferences of proteins sidechains. *Nature Struct. Biol.*, **1,** 334–340 (1994).

Dunfield, L. G., Burgess, A. W., and Scheraga, H. A. Energy parameters in polypeptides. VIII. Empirical potential energy algorithm for the conformational analysis of large molecules. *J. Phys. Chem.*, **82,** 2609–2616 (1978).

Efimov, A. V., Favoured structural motifs in globular proteins. *Structure,* **2,** 999–1002 (1994).

Eisenberg, D., Bowie, J. U., Lüthy, R., and Choe, S., Three-dimensional profiles for analysing protein sequence-structure relationships. *Farad. Disc.*, **93,** 25–34 (1992).

Eisenberg, D. and McLachlan, A. D., Solvation energy in protein folding and binding. *Nature,* **319,** 199–203 (1986).

Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., and Scharf, M., The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and for generating dot surfaces of molecular assemblies. *J. Comp. Chem.*, **16,** 273–284 (1995).

Eisenhaber, F. and Argos, P., Improved strategy in analytic surface calculation for molecular systems: handling of singularities and computational efficiency. *J. Comp. Chem.*, **14,** 1272–1280 (1993).

Eisenmenger, F., Argos, P., and Abagyan, R. A., A method to configure protein side-chains from the mainchain trace in homology modeling. *J. Mol. Biol.*, **231,** 849–860 (1993).

Elofsson, A. and Nilsson, L., How consistent are molecular dynamics simulations? *J. Mol. Biol.*, **233,** 766–780 (1993).

Evans, D. J. and Murad, S., Singularity free algorithm for molecular dynamics simulation of rigid polyatomics. *Mol. Phys.*, **34,** 327–331 (1977).

Factor, A. D. and Mehler, E. L., Graphical representation of hydrogen bonding patterns in proteins. *Prot. Eng.*, **4,** 421–425 (1991).

Fetrow, J. S. and Bryant, S. H., New programs for protein tertiary structure prediction. *Biotechnology*, **11**, 479–484 (1993).

Fincham, D., Leapfrog rotational algorithms for linear molecules. *Mol. Simul.*, **11**, 79–89 (1993).

Finkelstein, A. V. and Janin, J., The price of lost freedom: entropy of bimolecular complex formation. *Prot. Eng.*, **3**, 1–3 (1989).

Finkelstein, A. V. and Reva, B., A search for the most stable folds of protein chains. *Nature*, **351**, 497–499 (1991).

Flores, T. P., Orengo, C. A., Moss, D. S., and Thornton, J. M., Comparison of conformational characteristics in structurally similar protein pairs. *Prot. Sci.*, **2**, 1811–1826 (1993).

Fraenkel, A. S., Complexity of protein folding. *Bull. Math. Biol.*, **55**, 1199–1210 (1993).

Friedman, H. L., Image approximation to the reaction field. *Mol. Phys.*, **29**, 1533–1543 (1975).

Frishman, D. I. and Argos, P., Recognition of distantly related protein sequences using conserved motifs and neural networks. *J. Mol. Biol.*, **228**, 951–962 (1992).

Frömmel, C., Use of averaged mutation rate in pieces of protein sequences to predict the location of antigenic determinations. *J. Theor. Biol.*, **132**, 171–177 (1988).

Furois-Corbin, S., Smith, J. C., and Kneller, G. R., Picosecond timescale rigid-helix and side-chain motions in deoxymyoglobin. *Proteins*, **16**, 141–154 (1993).

Garnier, J., Osguthorpe, D. J. and 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) .

Garnier, J. and Levin, J. M., The protein structure code: what is its present status? *Comput. Appl. Biosci.*, **7**, 133–142 (1991).

Genfa, Z., Xinhua, X., and Zhang, C.-T., A weighting method for prediction of protein structural class from amino acid composition. *Eur. J. Biochem.*, **210**, 747–749 (1992).

Geourjon, C. and Deléage, G., SOPM: a self-optimized method for protein secondary structure prediction. *Prot. Eng.*, **7**, 157–164 (1994).

Gerber, P., Peptide mechanics: a force field for peptides and proteins working with entire residues as smallest units. *Biopolymers*, **32**, 1003–1017 (1992).

Gething, M. J. and Sambrook, J., Protein folding in the cell. *Nature*, **355**, 33–45 (1992).

Gibrat, J.-F., Garnier, J. and Robson, B., Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.*, **198**, 425–443 (1987).

Gibrat, J.-F., Robson, B., and Garnier, J., Influence of the local amino acid sequence upon the zones of the torsional angles phi and psi adopted by residues in proteins. *Biochemistry*, **30**, 1578–1586 (1991).

Gibson, K. D. and Scheraga, H. A., Decisions in force field development. Reply to Kollman and Dill. *J. Biomol. Struct. Dyn.*, **8**, 1109–1111 (1991).

Gibson, T. J., Thompson, J. D., and Abagyan, R. A., Proposed structure for the DNA-binding domain of the Helix-Loop-Helix family of eucaryotic gene regulatory proteins. *Prot. Eng.*, **6**, 41–50 (1993).

Gibson, T. J., Hyvönen, M., Musacchio, A., Saraste, M., and Birney, E., PH domain: the first anniversary. *Trends Biochem. Sci.*, **19**, 349–353 (1994).

Godzik, A., Kolinski, A., and Skolnick, J., Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.*, **227**, 227–238 (1992).

Godzik, A., Kolinski, A., and Skolnick, J., Lattice representations of globular proteins: How good are they? *J. Comp. Chem.*, **14**, 1194–1202 (1993).

Godzik, A. and 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).

Goldstein, R., Luthey-Schulten, Z. A., and Wolynes, P. G., Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. U.S.A.*, **89,** 4918–4922 (1992a).

Goldstein, R., Luthey-Schulten, Z. A., and Wolynes, P. G., Protein tertiary structure recognition using optimized hamiltonians with local interactions. *Proc. Natl. Acad. Sci. U.S.A.*, **89,** 9029–9033 (1992b).

Goldstein, R. F., Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.*, **66,** 1335–1340 (1994).

Gracy, J., Chiche, L., and Sallantin, J., Improved alignment of weakly homologous protein sequences using structural information. *Prot. Eng.*, **6,** 821–829 (1993).

Green, P., Lipman, D., Hillier, L., Waterstone, R., States, D., and Claverie, J.-M., Ancient conserved regions in new gene sequences and in the protein databases. *Science*, **259,** 1711–1715 (1993).

Greengard, L., Fast algorithms for classical physics. *Science*, **265,** 909–914 (1994).

Greer, J., Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.*, **153,** 1027–1042 (1981).

Gregoret, L. M. and Cohen, F. E., Novel method for the rapid evaluation of packing in protein structures. *J. Mol. Biol.*, **211,** 959–974 (1990).

Gribskov, M., McLachlan, A. D., and Eisenberg, D., Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **84,** 4355–4358 (1987).

Gribskov, M., Lüthy, R., and Eisenberg, D., Profile analysis. *Methods Enzymol.*, **183,** 146–159 (1990).

Grindley, H. M., Artymiuk, P. J., Rice, D. W., and Willett, P., Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.*, **229,** 707–721 (1993).

Gronbech-Jensen, N. and Doniach, S., Long-time overdamped Langevin dynamics of molecular chains. *J. Comp. Chem.*, **16,** 997–1012 (1994).

Gros, P. and van Gunsteren, W. F., Crystallographic refinement and structure-factor time-averaging by molecular dynamics in the absence of a physical force field. *Mol. Simul.*, **10,** 377–395 (1993).

Guarnieri, F. and Still, W. C., A rapidly convergent simulation method: mixed Monte Carlo/stochastic dynamics. *J. Comp. Chem.*, **15,** 1302–1310 (1994).

Gulukota, K. and Wolynes, P. G., Statistical mechanics of kinetic proofreading in protein folding *in vivo*. *Proc. Natl. Acad. Sci. U.S.A.*, **91,** 9292–9296 (1994).

Han, K.-K. and Martinage, A., Possible relationship between coding recognition of amino acid sequence motif or residue(s) and post-translational chemical modification of proteins. *Int. J. Biochem.*, **24,** 1349–1363 (1992).

Hancock, J. F., Magee, A. I., Childs, J. E., and Marshall, C. J., All *ras* proteins are polyisoprenylated but only some are palmitoylated. *Cell*, **57,** 1167–1177 (1989).

Harrison, R. W., Stiffness and energy conservation in molecular dynamics: an improved integrator. *J. Comp. Chem.*, **14,** 1112–1122 (1993).

Hartl, F. U., Secrets of a double-doughnut. *Nature*, **371,** 557–559 (1994).

Harvey, S. C., Treatment of electrostatic effects in macromolecular modeling. *Proteins*, **5,** 78–92 (1989).

Havel, T. and Snow, M. E., A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol.*, **217,** 1–7 (1991).

Hayward, S. and Collins, J. F., Limits on α-helix prediction with neural network models. *Proteins*, 14, 372–381 (1992).

Head-Gordon, T., Stillinger, F. H., and Arrecis, J., A strategy for finding classes of minima on a hypersurface: implications for approaches to the protein folding problem. *Proc. Natl. Acad. Sci. U.S.A.*, **88,** 11076–11080 (1991).

Head-Gordon, T. and Stillinger, F. H., Predicting polypeptide and protein structures from amino acid sequence: Antlion method applied to melittin. *Biopolymers,* **33,** 293–303 (1993).

Hellinga, H. W. and Richards, F. M., Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **91,** 5803–5807 (1994).

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E., and Downing, K. H., Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.*, **213,** 899–929 (1990).

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M. J., Identification of native protein folds among a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J. *Mol. Biol.*, **216,** 167–180 (1990).

Henikoff, S. and Henikoff, J. G., Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, **89,** 10915–10919 (1992).

Henikoff, S. and Henikoff, J. G., Position-based sequence weights. *J. Mol. Biol.*, 243, 574–578 (1994).

Herczyk, P. and Hubbard, R. E., A reduced representation of proteins for use in restraint satisfaction calculations. *Proteins,* **17,** 310–324 (1993).

Hilbert, M., Böhm, G., and Jaenicke, R., Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins,* **17,** 138–151 (1993).

Hinds, D. A. and Levitt, M., A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **89,** 2536–2540 (1992).

Hinds, D. A. and Levitt, M., Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, **243,** 668–682 (1994).

Hirst, J. D. and Sternberg, M. J. E., Prediction of structural and functional features of protein and nucleic acid sequences by neural networks. *Biochemistry,* **31,** 7211–7218 (1992).

Holbrook, S. R., Muskal, S. M., and Kim, S.-H., Predicting surface exposure of amino acids from protein sequence. *Prot. Eng.*, **3,** 659–665 (1990).

Holley, L. H. and Karplus, M., Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U.S.A.*, **86,** 152–156 (1989).

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., and Vriend, G., A database of protein structure families with common folding motifs. *Prot. Sci.*, **1,** 1691–1698 (1992).

Holm, L. and Sander, C., Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins,* **14,** 213–223 (1992a).

Holm, L. and Sander, C., Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.*, **225,** 93–105 (1992b).

Holm, L. and Sander, C., Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233,** 123–138 (1993).

Holm, L. and Sander, C., The FSSP database of structurally aligned protein fold families. *Nucl. Ac. Res.*, **22,** 3600–3609 (1994a).

Holm, L. and Sander, C., Searching protein structure databases has come of age. *Proteins,* **19,** 165–173 (1994b).

Holst, M., Kozack, R. E., Saied, F., and Subramaniam, S., Treatment of electrostatics effects in proteins: multigrid-based Newton iterative method for solution of the full nonlinear Poisson-Boltzmann equation. *Proteins*, **18**, 231–245 (1994a).

Holst, M., Kozack, R. E., Saied, F., and Subramaniam, S., Protein electrostatics: rapid multigrid-based Newton algorithm for solution of the full nonlinear Poisson-Boltzmann equation. *J. Biomol. Struct. Dyn.*, **11**, 1437–1445 (1994b).

Holst, M. and Saied, F., Multigrid solution of the Poisson-Boltzmann equation. *J. Comp. Chem.*, **14**, 105–113 (1993).

Hopp, T. P., Retrospective: 12 years of Antigenic determinant predictions, and more. *Peptide Res.*, **6**, 183–190 (1993).

Hopp, T. P. and Woods, K. R., Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.*, **78**, 3824–3828 (1981).

Hubbard, S. J., Eisenmenger, F., and Thornton, J. M., Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Prot. Sci.*, **3**, 757–768 (1994).

Hucho, F., Görne-Tschelnokow, U., and Strecker, A., β-Structure in the membrane-spanning part of the nicotinic acetylcholine receptor (or how helical are transmembrane helices?). *Trends Biochem. Sci.*, **19**, 383–387 (1994).

Jameson and Wolf, H., The antigenic index: a novel algorithm for predicting antigenic determinants. *Comput. Appl. Biosci.*, **4**, 181–186 (1988).

Janin, J., Wodak, S. J., Levitt, M., and Maigret, M., Conformation of amino acid side-chains in proteins. *J. Mol. Biol.*, **125**, 357–386 (1978).

Jenny, T. F. and Benner, S. A., Evaluating predictions of secondary structure in proteins. *Biophys. Biochem. Res. Comm.*, **200**, 149–155 (1994).

Johnson, M. S., Overington, J. P., and Blundell, T. L., Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.*, **231**, 735–752(1993).

Johnson, M. S., Srinivasan, N., Sowdhamini, R., and Blundell, T. L., Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.*, **29**, 1–68 (1994).

Johnson, M. S. and Overington, J. P., A structural basis for sequence comparison. An evaluation of scoring methodologies. *J. Mol. Biol.*, **233**, 716–738 (1993).

Johnson, W. C., Jr., Protein secondary structure and circular dichroism: a practical guide. *Proteins*, **7**, 205–214 (1990).

Jones, D. T., Taylor, W. R., and Thornton, J. M., A new approach to protein fold recognition. *Nature*, **358**, 86–89 (1992).

Jones, D. T., De novo protein design using pairwise potentials and a genetic algorithm. *Prot. Sci.*, **3**, 567–574 (1994).

Jones, T. A. and Thirup, S., Using known substructures in protein model building and crystallography. *EMBO J.*, **5**, 819–822 (1986).

Juffer, A. H., Botta, E. F. F., van Keulen, B. A. M., van der Ploeg, A., and Berendsen, H. J. C., The electric potential of a macromolecule in a solvent: a fundamental approach. *J. Comp. Phys.*, **97**, 144–171 (1991).

Kabsch, W., A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, **A32**, 922–923 (1976).

Kabsch, W. and Sander, C., How good are predictions of protein secondary structure. *FEBS Lett.*, **155**, 179–182 (1983a).

Kabsch, W. and Sander, C., Dictionary of protein secondary structures: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637 (1983b).

Kabsch, W. and 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).

Kamimura, M. and Takahashi, Y., phi-psi conformational pattern clustering of protein amino acid residues using the potential function method. *Comput. Appl. Biosci.*, **10**, 163–169 (1994).

Kang, H.-S., Kurochkina, N. A., and Lee, B., Estimation and use of protein backbone angle probabilities. *J. Mol. Biol.*, **229**, 448–460 (1993).

Karlin, S., Zuker, M., and Brocchieri, L., Measuring residue associations in protein structures. Possible implications for protein folding. *J. Mol. Biol.*, **239**, 227–248 (1994).

Karpen, M. E., de Haseth, P. L., and Neet, K. E., Comparing short protein substructures by a method based on backbone torsion angles. *Proteins*, **6**, 155–167 (1989).

Karplus, M., Ichiye, T., and Pettitt, B. M., Configurational entropy of native proteins. *Biophys. J.*, **52**, 1083–1085 (1987).

Karplus, M. and Schultz, G. E., Prediction of chain flexibility in proteins. *Naturwissenschaften*, **72**, 212–213 (1985).

Karshikov, A. D., Engh, R., Bode, W., and Atanasov, B. P., Electrostatic interactions in proteins: calculations of the electrostatic term of free energy and the electrostatic potential field. *Eur. Biophys. J.*, **17**, 287–297 (1989).

Kemp, B. E. and Pearson, R. B., Protein kinase recognition sequence motifs. *Trends Biochem. Sci.*, **15**, 342–346 (1990).

Kitao, A., Hayward, S., and Go, N., Comparison of normal mode analyses on a small globular protein in dihedral angle space and Cartesian coordinate space. *Biophys. Chem.*, **52**, 107–114 (1994).

Klein, P., Prediction of protein structural class by discriminant analysis. *Biochim. Biophys. Acta.*, 874, 205–215 (1986).

Klein, P. and DeLisi, C., Prediction of protein structural class from the amino acid sequence. *Biopolymers*, **25**, 1659–1672 (1986).

Kneller, D. G., Cohen, F. E., and Langridge, R., Improvements in protein secondary structure prediction by enhanced neural networks. *J. Mol. Biol.*, **214**, 171–182 (1990).

Kneller, G. R. and Geiger, A., A method to calculate the g-coefficients of the molecular pair correlation function from molecular dynamics simulations. *Mol. Simul.*, **3**, 283–300 (1989).

Kocher, J.-P. A., Rooman, M. J. and Wodak, S. J., Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J. Mol. *Biol.*, 235, 1598–1613 (1994).

Koehl, P. and Delarue, M., Polar and nonpolar atomic environments in the protein core: implications for folding and binding. *Proteins*, **20**, 264–278 (1994a).

Koehl, P. and Delarue, M., Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.*, **239**, 249–275 (1994b).

Kollman, P. A. and Dill, K. A., Decisions in force field development: an alternative to those described by Roterman *et al.* (*J. Biomol. Struct. Dyn.*, 7,421 (1989)). *J. Biomol. Struct. Dyn.*, **8**, 1103–1107 (1991).

Kono, H. and Doi, J., Energy minimization method using automata network for sequence and side-chain conformation prediction. *Proteins*, **19**, 244–255 (1994).

Kostrowicki, J. and Scheraga, H. A., Application of the diffusion equation for global optimization to oligopeptides. *J. Phys. Chem.*, **96**, 7442–7449 (1992).

Kühlbrandt, W., Wang, D. N., and Fujiyoshi, Y., Atomic model of plant light-harvesting complex by electron crystallography. *Nature*, **367**, 614–621 (1994).

Kyte, J. and Doolittle, R. F., A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.*, **157**, 105–132 (1982).

Lambert, M. H. and Scheraga, H. A., Pattern recognition in the prediction of protein structure. II. Chain conformation from a probability-directed search procedure. *J. Comp. Chem.*, **10**, 798–816 (1989).

Lasters, I. and Desmet, J., The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead end elimination theorem. *Prot. Eng.*, **6,** 717–722 (1993).

Lathrop, R. H., The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Prot. Eng.*, **7,** 1059–1068 (1994).

Lathrop, R. H. and Smith, T. F., A branch-and-bound algorithm for optimal protein threading with pairwise (contact potential) amino acid interactions. *Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences*, pp. 365–374. IEEE Computer Society Press, Los Alamos (1994).

Lau, K. T. and Dill, K. A., Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. U.S.A.*, **87,** 638–642 (1990).

Laughton, C. A., Prediction of protein side-chain conformations from local three-dimensional homology relationships. *J. Mol. Biol.*, **235,** 1088–1097 (1994a).

Laughton, C. A., A study of simulated annealing protocols for use with molecular dynamics in protein structure prediction. *Prot. Eng.*, **7,** 235–241 (1994b).

Lavery, R., Sklenar, H., Zakrzewska, K., and Pullman, B., The flexibility of nucleic acids. (II). The calculation of internal energy and applications to mononucleotide repeat DNA. *J. Biomol. Struct. Dyn.*, **3,** 989–1014 (1986).

Lee, C. and Subbiah, S., Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.*, **217,** 373–388 (1991).

Lee, F. S., Chu, Z. T., and Warshel, A., Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIX programs. *J. Comp. Chem.*, **14,** 161–185 (1993).

Lee, K. H., Xie, D., Freire, E., and Amzel, L. M., Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation. *Proteins*, **20,** 68–84 (1994).

Lemmon, M. A. and Engelmann, D. M., Helix-helix interactions inside lipid bilayers. *Curr. Opin. Struct. Biol.*, **2,** 511–518 (1992).

Leopold, P. E., Montal, M., and Onuchic, J. N., The protein folding funnels: a kinetic approach to the sequence-structure relationships. *Proc. Natl. Acad. Sci. U.S.A.*, **89,** 8721–8725 (1992).

Lesk, A. M. and Chothia, C., How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.*, **136,** 225–270 (1980).

Lessel, U. and Schomburg, D., Similarities between protein 3-D structures. *Prot. Eng.*, **7,** 1175–1187 (1994).

Levin, J. M., Pascarella, S., Argos, P., and Garnier, J., Quantification of secondary structure prediction improvement using multiple alignments. *Prot. Eng.*, **6,** 849–854 (1993).

Levin, J. M. and Garnier, J., 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,** 283–295 (1988).

Levinthal, C., Are there pathways for protein folding? *J. Chem. Phys.*, **65,** 44–45 (1968).

Levitt, M., A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, **104,** 59–107 (1976).

Levitt, M., Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, **226,** 507–533 (1992).

Levitt, M. and Chothia, C., Structural patterns in globular proteins. *Nature*, **261,** 552–558 (1976).

Levitt, M. and Sharon, R., Accurate simulation of protein dynamics in solution. *Proc. Natl. Acad. Sci. U.S.A.*, **85,** 7557–7561 (1988).

Li, Z. and Scheraga, H. A., Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, **84,** 6611–6615 (1987).

Lim, V. I., Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.*, **88,** 873–894 (1974).

Lis, H. and Sharon, N., Protein glycosylation. Structural and functional aspects. *Eur. J. Biochem.*, **218,** 1–27 (1993).

Liwo, A., Pincus, M. R., Wawak, R. J., Rackovsky, S., and Scheraga, H. A., Calculation of protein backbone geometry from α-carbon coordinates based on peptide-group dipole alignment. *Prot. Sci.*, **2,** 1697–1714 (1993).

Luo, Y., Jiang, X., Lai, L., Qu, C., Xu, X., and Tang, Y., Building protein backbones from C_α coordinates. *Prot. Eng.*, **5,** 147–150 (1992).

Luo, Y., Lai, L., Xu, X., and Tang, Y., Defining topological equivalences in protein structures by means of a dynamic programming algorithm. *Prot. Eng.*, **6,** 373–376 (1993).

Luthardt, G. and Frömmel, C., Local polarity analysis: a sensitive method that discriminates between native proteins and incorrectly folded models. *Prot. Eng.*, **7,** 627–631 (1994).

Lüthy, R., McLachlan, A. D., and Eisenberg, D., Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins,* **10,** 229–239 (1991).

Lüthy, R., Bowie, J. U., and Eisenberg, D., Assessment of protein models with three-dimentional profiles. *Nature,* **356,** 83–85 (1992).

Lüthy, R., Xenarios, I., and Bucher, P., Improving the sensitivity of the sequence profile method. *Prot. Sci.*, **3,** 139–146 (1994).

Madej, T. and Mossing, M. C., Hamiltonians for protein tertiary structure prediction based on three-dimensional environment principle. *J. Mol. Biol.*, **233,** 480–487 (1993).

Maiorov, V. N. and Crippen, G. M., Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.*, **227,** 876–888 (1992).

Mandal, C. and Linthicum, D. S., PROGEN: an automated modeling algorithm for the generation of complete protein structures from the α-carbon atomic coordinates. *J. Comp.-Aid. Mol. Des.*, **7,** 199–224 (1993).

Mao, B., Chou, K.-C., and Zhang, C.-T. Protein folding classes: a geometric interpretation of the amino acid composition of globular proteins. *Prot. Eng.*, **7,** 319–330 (1994).

Mao, B. and Friedman, A. R., Molecular dynamics simulation by atomic mass weighting. *Biophys. J.* **58,** 803–805 (1990).

Maple, J. R., Hwang, M.-H., Stockfish, T. P., Dinur, U., Waldmann, M., Ewig, C. S., and Hagler, A. T., Derivation of class II force fields. I. Methodology and quantum force field for the alkyl functional group and the alkane molecules. *J. Comp. Chem.*, **15,** 162–180 (1994).

Mark, A. E., van Gunsteren, W. F., and Berendsen, H. J. C., Calculation of relative free energy via indirect pathways. *J. Chem. Phys.*, **94,** 3808–3816 (1991).

Matsuo, Y. and Kanehisa, M., An approach to systematic detection of protein structural motifs. *Comput. Appl. Biosci.*, **9,** 153–159 (1993).

Matsuo, Y. and Nishikawa, K., Protein database search and structure prediction by 3D-1D compatibility method. *Prot. Eng.*, **7,** 1163 (1994).

Matthew, J. B., Electrostatic effects in proteins. *Ann. Rev. Biophys. Biophys. Chem.*, **14,** 387–417 (1985).

May, A. C. W. and Johnson, M. S., Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. *Prot. Eng.*, **7,** 475–485 (1994).

Mazur, A. K., Dorofeev, V. E., and Abagyan, R. A., Derivation and testing of explicit equations of motion for polymers described in internal coordinates. *J. Comp. Phys.*, **92,** 261–272 (1991).

Mazur, A. K. and Abagyan, R. A., New methodology for computer-aided modeling of biomolecular structure and dynamics. I. Non-cyclic structures. *J. Biomol. Struct. Dyn.*, **6**(4), 815–832 (1989).

McGarrah, D. B. and Judson, R. S., Analysis of the genetic algorithm method for molecular conformation determination. *J. Comp. Chem.*, **14**, 1385–1395 (1993).

McGregor, M. J., Flores, T. P., and Sternberg, M. J. E., Prediction of beta-turns in proteins using neural networks. *Prot. Eng.*, **2(7)**, 521–526 (1989).

McLachlan, A. D., A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Cryst.*, **A28**, 656–657 (1972).

McLachlan, A. D., Rapid comparison of protein structures. *Acta Cryst.*, **A38**, 871–873 (1982).

Mehler, E. L. and Solmajer, T., Electrostatic effects in proteins: comparison of dielectric and charge models. *Prot. Eng.*, **4**, 903–910 (1991).

Metfessel, B. A., Saurugger, P. N., Connelly, D. P., and Rich, S. S., Cross-validation of protein structural class prediction using statistical clustering and neural networks. *Prot. Sci.*, **2**, 1171–1182 (1993).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H., Equation of state calculations for fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092 (1953).

Mezei, M., Calculation of solvation free-energy differences for large solute change from computer simulations with quadrature-based nearly linear thermodynamic integration. *Mol. Simul.*, **10**, 225–239 (1993).

Mezei, M., A heuristic procedure for the detection of locally similar substructures of two equivalent structures. *Prot. Eng.*, **7**, 331–333 (1994).

Mitchell, E. M., Artymiuk, P. J., Rice, D. W., and Willett, P., Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.*, **212**, 151–166 (1989).

Miyamoto, S. and Kollman, P. A., SETTLE: an analytical version of SHAKE and RATTLE algorithm for rigid water models. *J. Comp. Chem.*, **13**, 952–962 (1992).

Miyazawa, S. and Jernigan, R. L., Estimation of interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).

Monge, A., Friesner, R. A., and Honig, B., An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 5027–5029 (1994).

Moore, C. and Fasman, G. D., The random coil conformation of proteins. *Chemtracts _ Biochem. Mol. Biol.*, **4**, 67–74 (1993).

Mottonen, J., Strand, A., Symerski, J., Sweet, R. M., Danley, R. E., Geoghegen, K. F., Gerard, R. D., and Goldsmith, E. J., Structural basis of latency in plasminogen activator inhibitor-1. *Nature*, **355**, 270–273 (1992).

Moult, J. and James, M. N. G., An algorithm for determining the conformation of polypeptide segemnts in proteins by systematic search. *Proteins*, **1**, 146–163 (1986).

Muggleton, S., King, R. D., and Sternberg, M. J. E., Protein secondary structure prediction using logic-based machine learning. *Prot. Eng.*, **5**, 647–657 (1992).

Muggleton, S., King, R. D., and Sternberg, M. J. E., Corrigenda: protein secondary structure prediction using logic-based machine learning. *Prot. Eng.*, 6, 549 (1993).

Murthy, M. R. N., A fast method of comparing protein structures. *FEBS Lett.*, **168**, 97–102 (1984).

Musacchio, A., Gibson, T. J., Rice, P., Thompson, J. D., and Saraste, M., The PH-domain: a common piece in the structural patchwork of signalling proteins. *Trends Biochem. Sci.*, **18**, 343–348 (1993).

Muskal, S. M. and Kim, S.-H., Predicting protein secondary structure content: a tandem neural network approach. *J. Mol. Biol.*, **225**, 713–727 (1992).

Nagy, P. I., Bitar, J. E., and Smith, D. A., Comparison of the molecular mechanics + generalized Born/surface area and the *ab initio* + Monte Carlo simulation methods in estimating conformational equilibria in aqueous solution. *J. Comp. Chem.*, **15,** 1228–1240 (1994).

Nakashima, H., Nishikawa, K., and Ooi, T., The folding type of a protein is relevant to the amino acid composition. *J. Biochem.*, **99,** 153–162 (1986).

Nayeem, A., Vila, J. and Scheraga, H. A., A comparative study of the simulated-annealing and Monte-Carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [Met]-enkephalin. *J. Comp. Chem.*, **12,** 594–605 (1991).

Nemethy, G., Pottle, M. S., and Scheraga, H. A., Energy parameters in polypeptides. IX. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occuring amino acids. *J. Phys. Chem.*, **87,** 1883–1887 (1983).

Nemethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S., and Scheraga, H. A., Energy parameters in polypeptides. X. Improved geometrical parameters and non-bonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.*, **96,** 6472–6484 (1992).

Ngo, T. J. and Marks, J., Computational complexity of a problem in molecular structure prediction. *Prot. Eng.*, **5,** 313–321 (1992).

Nishikawa, K., Kubota, Y., and Ooi, T., Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. *J. Biochem.*, **94,** 997–1007 (1983a).

Nishikawa, K., Kubota, Y., and Ooi, T., Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. *J. Biochem.*, **94,** 981–995 (1983b).

Nishikawa, K. and Matsuo, Y., Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Prot. Eng.*, **6,** 811–820 (1993).

Nishikawa, K. and Ooi, T., Correlation of the amino acid composition of a protein to its structural and biological characters. *J. Biochem.*, **91,** 1821–1824 (1982).

Noguti, T. and Go, N., Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins. *Biopolymers*, **24,** 527–546 (1985).

Norman, G. E., Podlipchuk, V. Y., and Valuev, A. A., Equations of motion and energy conservation in molecular dynamics. *Mol. Simul.*, **9,** 417–424 (1994).

Novotny, J., Bruccoleri, R. E., and Karplus, M., An analysis of incorrectly folded protein models. Implications for structure prediction. *J. Mol. Biol.*, **177,** 787–818 (1984).

Novotny, J., Rashin, A. A., and Bruccoleri, R. E., Criteria that discriminate between native proteins and incorrectly folded models. *Proteins*, 4, 13–30 (1988).

Oberoi, H. and Allewell, N. M., Multigrid solution of the nonlinear Poisson-Boltzmann equation and the calculation of titration curves. *Biophys. J.* **65,** 48–55 (1993).

Okamoto, Y., Helix-forming tendencies of non-polar amino acids predicted by Monte Carlo simulated annealing. *Proteins*, **19,** 14–23 (1994).

Okunbor, D. I. and Skeel, R. D., Canonical numerical methods for molecular dynamics simulations. *J. Comp. Chem.*, 15, 72–79 (1994).

Olszewsky, K. A., Piela, L., and Scheraga, H. A., Mean field theory as a tool for inter-molecular conformational optimization. 1. Tests on terminally blocked alanine and Met-enkephalin. *J. Phys. Chem.*, **96,** 4672–4676 (1992).

Ooi, T., Oobatake, M., Nemethy, G., and Scheraga, H. A., Accessible surface areas as measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. U.S.A.*, 84, 3086–3090 (1987).

Orengo, C. A., Brown, N. P., and Taylor, W. R., Fast structure alignment for protein databank searching. *Proteins*, **14**, 139–167 (1992).

Orengo, C. A., Flores, T. P., Taylor, W. R., and Thornton, J. M., Identification and classification of protein fold families. *Prot. Eng.*, **6**, 485–500 (1993).

Orengo, C. A. and Taylor, W. R., A rapid method of protein structure alignment. *J. Theor. Biol.*, **147**, 517–551 (1990).

Oroczo, M. and Luque, F. J., *Ab initio* study of bond stretching: implications in force-field parametrization for molecular mechanics and dynamics. *J. Comp. Chem.*, **14**, 881–894 (1993).

Ouzounis, C., Sander, C., Scharf, M., and Schneider, R., Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.*, **232**, 805–825 (1993).

Overington, J. P., Donelly, D., Johnson, M. S., Sali, A., and Blundell, T. L., Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Prot. Sci.*, **1**, 216–226 (1992).

Pabo, C. O., Designing proteins and peptides. *Nature*, 301, 200 (1983).

Paine, G. H. and Scheraga, H. A., Prediction of the native conformation of a polypeptide by a statistical-mechanical procedure. III. Probable and average conformations of enkephalin. *Biopolymers*, **26**, 1125–1162 (1987).

Palmer, K. A. and Scheraga, H. A., Standard geometry chains fitted to X-ray-derived structures: validation of the rigid geometry approximation. I. Chain closure through a l;imited search of "loop" conformations. *J. Comp. Chem.*, **12**, 505–526 (1991).

Palmer, K. A. and Scheraga, H. A., Standard geometry chains fitted to X-ray-derived structures: validation of the rigid-geometry approximation. II. Systematic searches for short loops in proteins: Application to bovine pancreatic ribonuclease A and human lysozyme. *J. Comp. Chem.*, **13**, 329–350 (1992).

Pancoska, P., Blazek, M., and Keiderling, T. A., Relationships between secondary structure fractions for globular proteins. Neural network analysis of crystallographic data sets. *Biochemistry*, **31**, 10250–10257 (1992).

Parker, J. M. R., Guo, D., and Hodges, R. S., New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry*, **25**, 5425–5432 (1986).

Pascarella, S. and Argos, P., Analysis of insertions/deletions in proteins. *J. Mol. Biol.*, **224**, 461–471 (1992a).

Pascarella, S. and Argos, P., A data bank merging related protein structures and sequences. *Prot. Eng.*, **5**, 121–137 (1992b).

Pastor, R. W., Analyses of statistical errors in dynamics simulations. Renugopalakrishnan, V., Carey, P. R., Smith, I. C. P., Huang, S. G., and Storer, A. C., Eds. *Proteins Structure, Dynamics and Design*, pp. 229–233, ESCOM, Leiden (1991).

Pastore, A., Saudek, V., Ramponi, G., and Williams, R. J. P., Three-dimensional Structure of acylphosphatase. *J. Mol. Biol.*, **224**, 427–440 (1992).

Patthy, L., Exons — original building blocks of proteins? *Bioessays*, **13**, 187–192 (1991).

Payne, P. W., Reconstruction of protein conformations from estimated positions of the C_α coordinates. *Prot. Sci.*, **2**, 315–324 (1993).

Pearlman, D. A., A comparison of alternative approaches to free energy calculations. *J. Phys. Chem.*, **98**, 1487–1493 (1994).

Perczel, A., Hollósi, M., Tusnády, G., and Fasman, G. D.,Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Prot. Eng.*, **4**, 669–679 (1991) .

Persson, B., Flinta, C., von Heijne, G., and Jörnwall, H., Structures of N-terminally acetylated proteins. *Eur. J. Biochem.*, **152,** 523–527 (1985).

Persson, B. and Argos, P., Prediction of transmembrane regions in proteins utilising multiple sequence alignments. *J. Mol. Biol.*, **237,** 182–192 (1994).

Pickett, S. D., Saqi, M. A. S., and Sternberg, M. J. E., Evaluation of the sequence template method for protein structure prediction. Discrimination of the (beta/alpha) 8-barrel fold. *J. Mol. Biol.*, **228,** 170–187 (1992).

Pickett, S. D. and Sternberg, M. J. E., Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.*, **231,** 825–839 (1993).

Picot, D., Loll, P. J., and Garavito, M., The X-ray crystal structure of the membrane protein prostaglandin H_2 synthase-1. *Nature,* **367,** 243–249 (1994).

Piela, L., Kostrowicki, J., and Scheraga, H. A., The multiple-minima problem in the conformational analysis of molecules. Determination of the potential energy surface by the diffusion equation. *J. Phys. Chem.*, 93, 3339–3346 (1989).

Pletnev, V. Z., Popov, E. M., and Kadymova, F. A., Approximated potential functions of non-bonded interactions of methyl groups. *Theoret. Chim. Acta,* **35,** 93–96 (1974).

Ponder, J. W. and Richards, F. M., Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193, 775–791 (1987).

Preissner, R. and Bork, P., On α-helices terminated by glycine. 1. Identification of common structural features. *Biophys. Biochem. Res. Comm.*, **180,** 660–665 (1991).

Presnell, S. R., Cohen, B. I., and Cohen, F. E., A segment-based approach to protein secondary structure prediction. *Biochemistry,* **31,** 983–993 (1992).

Purisima, E. O. and Scheraga, H. A., Conversion from a virtual-bond chain to a complete polypeptide backbone. *Biopolymers,* **23,** 1207–1224 (1984).

Purisima, E. O. and Scheraga, H. A., An approach to the multiple-minima problem in protein folding by relaxing dimensionality. Tests on enkephalin. *J. Mol. Biol.*, **196,** 697–709 (1987).

Qian, N. and Sejnowski, T. J., Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202,** 865–884 (1988).

Rao, J. K. and Argos, P., A conformational preference parameter to predict helices in integral membrane proteins. *Biochim. Biophys. Acta.*, **869,** 197–214 (1986).

Reczko, M., Bohr, H., Subramaniam, S., Pamigighantam, S., and Hatzigeorgiou, A., Fold-class prediction by neural networks. Bohr, H. and Brunak, S., Eds. *Protein structure by distance analysis*, pp. 277–286. IOS Press, Amsterdam, Tokyo (1994).

Reczko, M. and Bohr, H., The DEF data base of sequence based protein fold class predictions. *Nucl. Acid. Res.*, **22,** 3616–3619 (1994).

Remington, S. J. and Matthews, B. W., A systematic approach to the comparison of protein structures. *J. Mol. Biol.*, **140,** 77–99 (1980).

Resh, M. D., Myristylation and palmitylation of Src family members: the fats of the matter. *Cell,* **76,** 411–413 (1994).

Rey, A. and Skolnick, J., Efficient algorithm for reconstruction of a protein backbone from the α-carbon coordinates. *J. Comp. Chem.*, **13,** 443–456 (1992).

Reynolds, J. A., Gilbert, D. B., and Tanford, C., Emprical correlation between hydrophobic free energy and aqueous cavity surface area. *Proc. Natl. Acad. Sci. U.S.A.*, **71,** 2925–2927 (1974).

Richards, F. M. and Kundrot, C. E., Identification of structural motifs from protein coordinate data: secondary and first-level supersecondary structure. *Proteins*, **3**, 71–84 (1988).

Richardson, J. S., The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.*, **34**, 168–339 (1981).

Richardson, J. S., Richardson, D. C., Tweedy, N. B., Gernet, K. M., Quinn, T. P., Hecht, M. H., Ericson, B. W., Yan, Y., McClain, R. D., Donlan, M. E., and Surles, M. C., Looking at proteins: representations, folding, packing, and design. *Biophys. J.*, **63**, 1186–1209 (1992).

Richmond, T. J., Solvent accessible surface area and excluded volume in proteins. *J. Mol. Biol.*, 178, 63–89 (1984).

Ring, C. S., Kneller, D. G., Langridge, R., and Cohen, F. E., Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.*, **224**, 685–699 (1992).

Ripoll, D. R., Piela, L., Vasquez, M., and Scheraga, H. A., On the multiple minima problem in the conformational analysis of polypeptides. V. Application of the self-consistent electrostatic field and the electrostatically driven Monte Carlo methods to bovine pancreatic trypsin inhibitor. *Proteins*, **10**, 188–198 (1991).

Robson, B. and Platt, E., Refined models for computer calculations in protein engineering. Calibration and testing of atomic potential functions with ore efficient calculations. *J. Mol. Biol.*, **188**, 259–281 (1986).

Rohde, K. and Bork, P., A fast, sensitive pattern-matching approach for protein sequences. *Comput. Appl. Biosci.*, **9**, 183–189 (1993).

Roitberg, A. and Elber, R., Modeling side chains in peptides and proteins: application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.*, **95**, 9277–9287 (1991).

Rooman, M. J., Kocher, J.-P. A., and Wodak, S. J., Prediction of protein backbone conformations based on seven structure assignments. *J. Mol. Biol.*, **221**, 961–979 (1991).

Rooman, M. J., Kocher, J.-P. A., and Wodak, S. J., Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformations in the absence of tertiary interactions. *Biochemistry*, **31**, 10226–10238 (1992).

Rooman, M. J. and Wodak, S. J., Identification of predictive sequence motifs limited by protein structure database size. *Nature*, **335**, 45–49 (1988).

Rose, J. and Eisenmenger, F., A fast unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm. *J. Mol. Evol.*, **32**, 340–354 (1994).

Rossmann, M. G. and Argos, P., Exploring structural homology of proteins. *J. Mol. Biol.*, **105**, 75–95 (1976).

Rossmann, M. G. and Argos, P., The taxonomy of protein structure. *J. Mol. Biol.*, **109**, 99–129 (1977).

Rost, B., Sander, C., and Schneider, R. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, 235, 13–26 (1994).

Rost, B. and Sander, C., Secondary structure prediction of all-helical proteins in two states. *Prot. Eng.*, **6**, 831–836 (1993).

Rost, B. and Sander, C., Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72 (1994a).

Rost, B. and Sander, C., Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226 (1994b).

Rost, B. and Sander, C., 1D secondary structure prediction through evolutionary profiles. Bohr, H. and Brunak, S., Eds. ***Protein structure by distance analysis***, pp. 257–276, IOS Press, Amsterdam, Oxford, Tokyo (1994c).

Roterman, I. K., Gibson, K. D., and Scheraga, H. A., A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. I. Conformational predictions for the tandemly repeated peptide (Asn-Ala-Asn-Pro)$_9$. *J. Biomol. Struct. Dyn.*, **7**, 391–419 (1989a).

Roterman, I. K., Lambert, M. H., Gibson, K. D., and Scheraga, H. A., A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. phi-psi maps for N-acetyl alanine N′-methyl amide: comparisons, contrasts and simple experimental tests. *J. Biomol. Struct. Dyn.*, **7**, 421–453 (1989b).

Rudnicki, W. R., Lesyng, B., and Harvey, S. C., Lagrangian molecular dynamics using selected conformational degrees of freedom, with application to pseudorotation dynamics of furanose rings. *Biopolymers*, **34**, 383–392 (1994).

Rufino, S. D. and Blundell, T. L., Structure-based identification and clustering of protein families and superfamilies. *J. Comp.-Aid. Mol. Des.*, **8**, 5–27 (1994).

Russel, R. B. and Barton, G. J., Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts, secondary structure and accessibility. *J. Mol. Biol.*, **244**, 332–350 (1994).

Russel, R. B. and Barton, G. J., Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323 (1992).

Russel, R. B. and Barton, G. J., The limits of protein secondary structure prediction accuracy from multiple sequence alignments. *J. Mol. Biol.*, **234**, 951–957 (1993).

Saitoh, S., Nakai, T., and Nishikawa, K., A geometrical constraint approach for reproducing the native backbone conformation of a protein. *Proteins*, **15**, 191–204 (1993).

Sali, A., Shakhnovich, E., and Karplus, M. Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.*, **235**, 1614–1636 (1994).

Sali, A. and Blundell, T. L., Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.*, **212**, 403–428 (1990).

Sali, A. and Blundell, T. L., Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815 (1993).

Salzberg, S. and Cost, S., Predicting protein secondary structure with a nearest-neighbor algorithm. *J. Mol. Biol.*, **227**, 371–374 (1992).

Samorjai, R. L., Novel approach for computing the global minimum of proteins. I. General concepts, methods and approximations. *J. Phys. Chem.*, **95**, 4141–4146 (1991).

Sander, C. and Schneider, R., Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68 (1991).

Schaefer, M. and Frömmel, C., A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solutions. *J. Mol. Biol.*, **216**, 1045–1066 (1990).

Schiffer, C. A., Caldwell, J. W., Kollman, P. A., and Stroud, R. M., Prediction of homologous protein structures based on conformational searches and energetics. *Proteins*, **8**, 30–43 (1990).

Schiffer, C. A., Caldwell, J. W., Stroud, R. M., and Kollman, P. A., Inclusion of solvation free energy with molecular mechanics energy: alanyl dipeptide as test case. *Prot. Sci.*, **1**, 396–400 (1992).

Schiffer, C. A., Caldwell, J. W., Kollman, P. A., and Stroud, R. M., Protein structure prediction with a combined solvation free energy-molecular mechanics force field. *Mol. Simul.*, **10**, 121–149 (1993).

Schmitz, U., Ulyanov, N. B., Kumar, A., and James, T. L., Molecular dynamics with weighted time-averaged restraints for a DNA octamer. *J. Mol. Biol.*, **234**, 373–389 (1993).

Schrauber, H., Eisenhaber, F., and Argos, P., Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol.*, **230**, 592–612 (1993).

Schreiber, H. and Steinhauser, O., Taming cut-off induced artefacts in molecular dynamics studies of solvated polypeptides. *J. Mol. Biol.*, **228,** 909–923 (1992).

Schroll, W. K., Van Zandt, L. L., and Saxena, V. K., Low-frequency parametrization of hydrogen bonding. *J. Biomol. Struct. Dyn.*, **8,** 1057–1067 (1991).

Scully, J. L. and Hermans, J., Multiple time steps: limits on the speedup of molecular dynamics simulations of aqueous systems. *Mol. Simul.*, **11,** 67–77 (1993).

Shalloway, D., Application of the renormalization group to deterministic global minimization of molecular conformation energy functions. *J. Glob. Optim.*, **2,** 281–311 (1992).

Sheridan, R. P., Dixon, J. S., Venkataraghavan, R., Kuntz, I. D., and Scott, K. P., Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. *Biopolymers*, **24,** 1995–2023 (1985).

Shestopalov, B. V., Prediction of protein secondary stucture by a doublet code. *Mol. Biol.*, **24,** 1117–1125 (1990).

Shin, J. K. and Jhon, M. S., High directional Monte Carlo procedure coupled with the temperature heating and annealing as a method to obtain the global energy minimum structure of polypeptides and proteins. *Biopolymers*, **31,** 177–185 (1991).

Shrake, A. and Rupley, J. A., Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, **79,** 351–371 (1973).

Sibanda, B. L. and Thornton, J. M., Conformation of beta-hairpins in protein structures: classification and diversity in homologous structures. *Meth. Enzymol.*, **202,** 59–82 (1991).

Sibbald, P. and Argos, P., Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.*, **216,** 813–818 (1990).

Simon, I., Glasser, L., and Scheraga, H. A., Calculation of protein conformation as an assembly of stable overlapping segments: application to bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. U.S.A.*, **88,** 3661–3665 (1991).

Sipos, L. and von Heijne, G., Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.*, **213,** 1333–1340 (1993).

Sippl, M. J., Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213,** 859–883 (1990).

Sippl, M. J., Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comp.-Aid. Mol. Des.*, **7,** 473–501 (1993a).

Sippl, M. J., Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17,** 355–362 (1993b).

Sippl, M. J. and Weitckus, S., Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformation. *Proteins*, **13,** 258–271 (1992).

Sklenar, H., Etchebest, C., and Lavery, R., Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins*, **6,** 46–60 (1989).

Sklenar, H., Eisenhaber, F., Poncin, M., and Lavery, R., Including solvent and counterion effects in the force field of macromolecular mechanics: the field integrated electrostatic approach (FIESTA). Beveridge, D. L. and Lavery, R., Eds. *Theoretical Biochemistry & Molecular Biophysics*, pp. 317–335. Adenine Press, New York (1990).

Smith-Brown, M. J., Kominos, D., and Levy, R. M., Global folding of proteins from a limited number of distance constraints. *Prot. Eng.*, **6,** 605–614 (1993).

Snow, M., A novel parametrization scheme for energy equations and its use to calculate the structure of protein molecule. *Proteins,* **15,** 183–190 (1993).

Solovyev, V. V. and Makarova, K. S., A novel method of protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization. *Comput. Appl. Biosci.,* **9,** 17–24 (1993).

Sonnhammer, E. L. L. and Durbin, R., A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.,* **10,** 301–307 (1994).

Sonnhammer, E. L. L. and Kahn, D., The modular arrangement of proteins as inferred from analysis of homology. *Prot. Sci.,* **3,** 482–492 (1994).

Sreerama, N. and Woody, R. W., Protein secondary structure from circular dichroism spectroscopy. *J. Mol. Biol.,* **242,** 497–507 (1994).

Starzyk, R., Webster, T., and Schimmel, P., Evidence for disposable sequences inserted into a nucleotide fold. *Science,* **237,** 1614–1618 (1987).

Steinbach, P. J. and Brooks, B. R., New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comp. Chem.,* **15,** 667–683 (1994).

Sternberg, M. J. E. and Chickos, J. S., Protein side-chain conformational entropy derived from fusion data- comparison with other empirical scales. *Prot. Eng.,* **7,** 149–155 (1994).

Stevens, R. C., Gouaux, J. E., and Lipscomb, W. N., Structural consequences of effector binding to the T state of aspartate carbamoyltransferase: crystal structure of the unligated and ATP- and CTP-complexed enzymes at 2.6-Å resolution. *Biochemistry,* **29,** 7691–7701 (1990).

Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T., (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem., Soc.,* **112,** 6127–6129.

Subbarao, N. and Haneef, I., (1991) Defining topological equivalence in macromolecules. *Prot. Eng.,* **4,** 877–884.

Summers, N. L. and Karplus, M., Construction of side-chains in homology modeling. Application to the C-terminal lobe of rhizopuspepsin. *J. Mol. Biol.,* **210,** 785–811 (1989).

Sun, S., Reduced representation model pf protein structure prediction: statistical potential and genetic algorithm. *Prot. Sci.,* **2,** 762–785 (1993).

Sutcliffe, M. J., Hayes, F. R. F., and Blundell, T. L., Knowledge-based modeling of homologous proteins. II: Rules for conformations of substituted side chains. *Prot. Eng.,* **1,** 385–392 (1987).

Swindells, M. B. and Thornton, J. M., A study of structural determinants in the interleukin-1 fold. *Prot. Eng.,* **6,** 711–715 (1993).

Tanaka, S. and Scheraga, H. A., Medium and long-range interaction parameters between amino acids for predicting three-dimensional structure of proteins. *Macromolecules,* **9,** 945–950 (1976).

Tanford, C. and Kirkwood, J. G., Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Am. Chem. Soc.,* **79,** 5333–5339 (1957).

Taylor, W. R., Protein fold refinement: building models from idealized folds using motif constraints and multiple sequence data. *Prot. Eng.,* **6,** 593–604 (1993).

Taylor, W. R. and Orengo, C., A holistic approach to protein structure alignment. *Prot. Eng.,* **2,** 505–519 (1989).

Taylor, W. R. and Orengo, C. A., Protein structure alignment. *J. Mol. Biol.,* **208,** 1–22 (1989).

Thompson, J. D., Higgins, D. G., and Gibson, T. J., Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.,* **10,** 19–29 (1994).

Toma, K., Protein three-dimensional structure generation with an empirical hydrophobic penalty function. *J. Mol. Graph.*, **11,** 222–232 (1994).

Topham, C. M., McLeod, A., Eisenmenger, F., Overington, J. P., Johnson, M. S., and Blundell, T. L., Fragment ranking in modeling of protein structure. Conformationally constrained environmental aminoacid substitution tables. *J. Mol. Biol.*, **229,** 194–220 (1993).

Tramontano, A., Chothia, C., and Lesk, A. M., Structural determinants of the conformations of medium-sized loops in proteins. *Proteins,* **6,** 382–394 (1989).

Tramontano, A., Chothia, C., and Lesk, A. M., Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the V_H domains of the immunoglobulins. *J. Mol. Biol.*, **215**, 175–182 (1990).

Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R., A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.*, **8,** 1267–1289 (1991).

Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R., A critical comparison of search algorithms applied to the optimization of protein side-chain conformations. *J. Comp. Chem.*, **14,** 790–798 (1993).

Tunon, I., Silla, E., and Pascual-Ahuir, J. L., Molecular surface area and hydrophobic effect. *Prot. Eng.*, **5,** 715–716 (1992).

Unger, R. and Moult, J., Genetic algorithms for protein folding simulations. *J. Mol. Biol.*, **231,** 75–81 (1993).

Unger, R. and Moult, J., Finding the lowest free energy conformation is an NP-hard problem: proof and implications. *Bull. Math. Biol.*, **55,** 1183–1198 (1994).

Vajda, S. and DeLisi, C., Determining minimum energy conformations of polypeptides by dynamic programming. *Biopolymers,* **29,** 1755–1772 (1990).

van Gelder, C. W. G., Leusen, F. J. J., Leunissen, J. A. M., and Noordik, J. H., A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. *Proteins,* **18,** 174–185 (1994).

van Gunsteren, W. F. and Berendsen, H. J. C., Algorithms for macromolecular dynamics and constraint dynamics. *Mol. Phys.*, **34,** 1311–1327 (1977).

van Gunsteren, W. F. and Berendsen, H. J. C., Computer simulation of molecular dynamics: methodology, application and perspectives in chemistry. *Angew. Chem. Int. Ed. Engl.*, **29,** 992–1023 (1990).

van Heel, M., A new family of powerful multivariate statistical sequence analysis (MSSA) techniques. *J. Mol. Biol.*, **216,** 877–887 (1992).

van Schaik, R. C., Berendsen, H. J. C., Torda, A. E., and van Gunsteren, W. F., A structure refinement method based on molecular dynamics in four spatial dimensions. *J. Mol. Biol.*, **234,** 751–762 (1993a).

van Schaik, R. C., van Gunsteren, W. F., and Berendsen, H. J. C., Conformational search by potential energy annealing: algorithm and application to cyclosporin A. *J. Comp.-Aid. Mol. Des.*, **6,** 97–112 (1993b).

Vanderbilt, D. and Louie, S. G., A Monte Carlo simulated annealing approach to optimization over continuous variables. *J. Comp. Phys.*, **56,** 259–271 (1984).

Vasmatzis, G., Brower, R., and DeLisi, C., Predicting immunoglobulin-like hypervariable loops. *Biopolymers,* **34,** 1669–1680 (1994).

Vasquez, M. and Scheraga, H. A., Use of buildup and energy minimization procedures to compute low-energy structures of the backbone of enkephalin. *Biopolymers,* **24,** 1437–1447 (1985).

Veenstra, D. L., Ferguson, D. M., and Kollman, P. A., How transferable are hydrogen parameters in molecular mechanics calculations? *J. Comp. Chem.*, **13,** 971–978 (1992).

Venneri, G. D. and Hoover, W. G., Simple exact test for well-known molecular dynamics algorithms. *J. Comp. Phys.*, 73, 468–475 (1987).

Vieth, M., Kolinski, A., Brooks, C. L., III and Skolnick, J., Prediction of the folding pathways and the structure of the GNC4 leucine zipper. *J. Mol. Biol.*, **237**, 361–367 (1994).

Vila, J., Williams, R. L., Vasquez, M., and Scheraga, H. A., Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. *Proteins*, **10**, 199–218 (1991).

von Freyberg, B., Richmond, T. J. and Braun, W., Surface area included in energy refinement of proteins. *J. Mol. Biol.*, **233**, 275–292 (1993).

von Freyberg, B. and Braun, W., Minimization of empirical energy functions including hydrophobic surface area effects. *J. Comp. Chem.*, **14**, 510–521 (1993).

von Heijne, G., The distribution of positively charged residues in bacterial inner membrane proteins correlates with trans-membrane topology. *EMBO J.*, **5**, 3021–3027 (1986).

von Heijne, G., Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.*, **225**, 487–494 (1992).

Vorobyev, Y. N., Grant, J. A., and Scheraga, H. A., A combined iterative and boundary element approach for solution of the nonlinear Poisson-Boltzmann equation. *J. Am. Chem. Soc.*, **114**, 3189–3196 (1992).

Vriend, G. and Eijsink, V., Prediction and analysis of structure, stability and unfolding of thermolysin-like proteases. *J. Comp.-Aid. Mol. Des.*, **7**, 367–396 (1993).

Vriend, G. and Sander, C., Detection of common three-dimensional substructures in proteins. *Proteins*, **11**, 52–58 (1991).

Vriend, G. and Sander, C., Quality control of protein models: directional atomic contact analysis. *J. Appl. Cryst.*, **26**, 47–60 (1993).

Wako, H. and Blundell, T. L., Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologues proteins. I. Solvent accessibility classes. *J. Mol. Biol.*, **238**, 682–692 (1994a).

Wako, H. and Blundell, T. L., Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *J. Mol. Biol.*, **238**, 693–708 (1994b).

Wallqvist, A., On the implementation of Friedman boundary conditions in liquid water simulations. *Mol. Simul.*, **10**, 13–17 (1993).

Wallqvist, A. and Ullner, M., A simplified amino acid potential for use in structure prediction of proteins. *Proteins*, **18**, 267–280 (1994).

Walther, D., (1994) personal communication.

Warshel, A. and Åqvist, J., Electrostatic energy and macromolecular function. *Ann. Rev. Biophys. Chem.*, **20**, 267–298 (1991).

Weiner, S. J., Kollman, P. A., Nguyen, D. T., and Case, D. A., An all atom force field for simulations of proteons and nucleic acids. *J. Comp. Chem.*, **7**, 230 (1986).

Weiss, M. S. and Schulz, G. E., Structure of porin refined at 1.8 Å resolution. *J. Mol. Biol.*, **227**, 493–509 (1992).

Wendoloski, J. J. and Salemme, F. R., PROBIT: a statistical approach to modeling proteins from partial coordinate data using substructure libraries. *J. Mol. Graph.*, **10**, 124–126 (1992).

Wesson, L. and Eisenberg, D., Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Prot. Sci.*, **1**, 227–235 (1992).

Wilmanns, M. and Eisenberg, D., Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alfa-barrel fold. *Proc. Natl. Acad. Sci. U.S.A.*, **90,** 1379–1383 (1993).

Wilson, C., Gregoret, L. M., and Agard, D. A., (1993) Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.*, **229,** 996–1006.

Wilson, C. and Cui, W., Application of simulated annealing to peptides. *Biopolymers,* **29,** 149–157 (1990).

Wilson, C. and Doniach, S., A computer model to dynamically simulate protein folding: studies with crambin. *Proteins,* **6,** 193–209 (1989).

Wodak, S. J. and Rooman, M. J., Generating and testing protein folds. *Curr. Opinion. Struct. Biol.*, **3,** 247–259 (1993).

Wootton, J. C., Sequences with 'unusual' amino acid compositions. *Curr. Opin. Struct. Biol.*, **4,** 413–421 (1994).

Ycas, M., Computing tertiary structures of proteins. *J. Prot. Chem.*, **9,** 177–200 (1990).

Yee, D. P. and Dill, K. A., Families and the structural relatedness among globular proteins. *Prot. Sci.*, **2,** 884–899 (1993).

Yi, T.-M. and Lander, E. S., Protein structure prediction using nearest-neighbor methods. *J. Mol. Biol.*, **232,** 1117–1129 (1993).

Yip, V. and Elber, R., Calculations of a list of neighbors in molecular dynamics simulations. *J. Comp. Chem.*, **10,** 921–927 (1989).

Zhang, C.-T. and Chou, K.-C., Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. *Biophys. J.*, **63,** 1523–1529 (1992a).

Zhang, C.-T. and Chou, K.-C., An optimization approach to predicting protein structural class from amino acid composition. *Prot. Sci.*, **1,** 401–408 (1992b).

Zhang, G. and Schlick, T., LIN: a new algorithm to simulate the dynamics of biomolecules by combining implicit-integration and normal mode techniques. *J. Comp. Chem.*, **14,** 1212–1233 (1993).

Zhang, K. Y. J. and Eisenberg, D., The three-dimensional profile method using residue preference as a continuous function of residue environment. *Prot. Sci.*, **3,** 687–695 (1994).

Zhang, T., Bertelsen, E., and Alber, T., Entropic effects on disulfide bonds on protein stability. *Nature Struct. Biol.*, **1,** 434–438 (1994).

Zhang, X., Mesirov, J. P., and Waltz, D. L., Hybrid system for protein secondary structure prediction. *J. Mol. Biol.*, **225,** 1049–1063 (1992).

Zheng, Q., Rosenfeld, R., Vajda, S., and DeLisi, C., Loop closure via bond scaling and relaxation. *J. Comp. Chem.*, **14,** 556–565 (1993a).

Zheng, Q., Rosenfeld, R., Vajda, S., and DeLisi, C., Determining protein loop conformation using scaling-relaxation techniques. *Prot. Sci.*, **2,** 1242–1248 (1993b).

Zheng, Q. and Kyle, D. J., Multiple copy sampling: rigid versus flexible protein. *Proteins,* **19,** 324–329 (1994).

Zhong, L. and Johnson, W. C., Jr., Environment affects amino acid preference for secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, **89,** 4462–4465 (1992).

Zimmermann, K., When awaiting 'Bio' Champollion: dynamic programming regularization of the protein secondary structure predictions. *Prot. Eng.*, **7,** 1197–1202 (1989).

Zuker, M. and Somorjai, R. L., The alignment of protein structures in three dimensions. *Bull. Math. Biol.*, **51,** 55–78 (1989).