ELSEVIER

Biophysical Chemistry 59 (1996) 1–32

Biophysical Chemistry

Review

New approaches in molecular structure prediction [1]

Gerald Böhm

Institut für Biotechnologie, Martin-Luther-Universität Halle-Wittenberg, Weinbergweg 16a, D-06120 Halle (Saale), Germany

Received 7 August 1995; revised 9 September 1995; accepted 11 September 1995

Abstract

In the past years, much effort has been put on the development of new methodologies and algorithms for the prediction of protein secondary and tertiary structures from (sequence) data; this is reviewed in detail. New approaches for these predictions such as neural network methods, genetic algorithms, machine learning, and graph theoretical methods are discussed. Secondary structure prediction algorithms were improved mostly by considering families of related proteins; however, for the reliable tertiary structure modeling of proteins, knowledge-based techniques are still preferred. Methods and examples with more or less successful results are described. Also, programs and parameterizations for energy minimisations, molecular dynamics, and electrostatic interactions have been improved, especially with respect to their former limits of applicability. Other topics discussed in this review include the use of traditional and on-line databases, the docking problem and surface properties of biomolecules, packing of protein cores, de novo design and protein engineering, prediction of membrane protein structures, the verification and reliability of model structures, and progress made with currently available software and computer hardware. In summary, the prediction of the structure, function, and other properties of a protein is still possible only within limits, but these limits continue to be moved.

Keywords: Computer simulations; Molecular modeling; Proteins; Molecular structure prediction; Knowledge-based methods; Review

1. Introduction

It is not yet proven that the protein folding problem, i.e. the definition of a unique (mathematical and physical) code that connects sequence-derived information with a three-dimensional structure, exists within our current mathematical and physical model of the macrocosm. Our low-resolution, anthropocentric description of nature and natural processes must not necessarily allow the denotation of a mathematical relationship between the sequence and the three-dimensional structure [2] of complex molecules; however, the application of more than three spatial dimensions for structural considerations of proteins is just emerging [1,2]. On the other hand, there are no serious doubts that in nature each protein sequence in its defined, natural environment (usually an aqueous solution) adopts a particular native structure

[1] This publication is dedicated to Professor Dr. Rainer Jaenicke, on the occasion of his 65th birthday.

[2] In the following, the term "structure" refers to the traditional three-dimensional representation of macromolecules. A description of objects in more spatial dimensions is usually not feasible with respect to the limited power of imagination in humans.

SSDI 0301-4622(95)00120-4

which we can measure in at least three dimensions. The depiction of an algorithm predicting these three-dimensional structures with a sufficient degree of reliability is a highly desirable objective in terms of rational design of proteins and drugs, for biotechnological as well as for medical purposes.

The current state of structure prediction has been summarized in a series of previous reviews [3–17]; for an introduction into the field of structure calculation methodology, see the excellent study of Scheraga [12] on peptide structure prediction and the comprehensive review by Eisenhaber et al. [16] on several aspects of protein tertiary structure prediction. Also, the fundamental obstacles in terms of computation have been addressed in a number of recent publications [13,18–21], and the mathematical basis for structure prediction methods with respect to their NP-completeness has been discussed [22,23].

During the past few years, special interest has been put on the modeling and accurate prediction of antibody variable regions [24–33]. Also, the prediction of segments of periodic secondary structure classes based on the respective sequence information solely, occasionally complemented by information derived from homologous proteins, is still an area of interest [4,6,34–41], although no significant improvement could be noted in this field. On the other hand, knowledge-based methods for the prediction of tertiary structures established from information of homologous proteins [42] has made considerable progress, and is now generally accepted as the most promising approach to date for protein structure prediction. Research regarding the design of new or improved drugs is another important field in molecular structure prediction and structural design which complements the efforts in modern molecular medicine and biotechnology [13,43–56]; however, this is beyond the scope of this review.

2. New algorithmic approaches

In the past, much work was carried out on the development and improvement of computer programs for secondary structure prediction, energy minimisation, and molecular dynamics; meanwhile, different ideas originally raised in related fields of modern science were adapted to the protein folding problem. Four techniques should be mentioned in this context: (i) machine learning; (ii) genetic algorithms; (iii) neural network methods; and (iv) graph theory.

In *machine learning*, rules are deduced from known relationships between information elements that are subsequently generalized. These rules are then the central component of the approach; unknown objects can serve as input to a program implementing these rules. Most work in this area has been performed in order to evaluate rules between the primary and secondary structure of proteins, i.e. in order to improve the quality of secondary structure prediction [57]. A program based on machine learning, PROMIS (PROtein Machine Induction System), has proven to be similarly useful compared with other secondary structure prediction approaches, with an average accuracy of 60% for the prediction of three states (helix, sheet, coil) for proteins of unknown domain type [58]. A similar method was developed for the deduction of significant properties of amino acids in predefined sequence patterns with known three-dimensional structures that serve as starting signals for α-helices of proteins, in order to improve secondary structure prediction by the precise assignment of bordering residues, at the same time avoiding redundant physical information [59].

A thorough study of the quality of machine learning compared to traditional methods and also to neural network algorithms has been presented by Sternberg et al. [60]. The resulting program GOLEM uses inductive logic programming, and has (as one of three tasks) shown success in predicting protein secondary structure elements from sequence data which is comparable to the more traditional methods, including neural networks. The authors conclude that rules deduced from machine learning algorithms, together with human intervention, are a powerful tool, especially for the investigation of stereochemical properties of biological macromolecules.

Apart from the structural predictions of native protein conformations, computer programs based on machine learning theory were also developed for predicting functional properties of enzymes [61] and for the structure–activity relationships of ligand binding to proteins [43], with the successful example of modeling the binding of trimethoprim analogs to the active site of dihydrofolate reductase.

Genetic algorithms are based on the observation that evolution usually proceeds gradually by the selective analysis of advantages and disadvantages of a mutation, i.e. a small change in the initial state of a system. The respective genetic algorithms iteratively find solutions to a problem by a permanent change in equation terms and parameters, and subsequent analysis of the results, using methods called "mutation", "recombination", and "selection"; optimisation is thus performed by stepwise improvements based on selection. Therefore, genetic algorithms represent efficient search methodologies for nearly optimal solutions of mathematical puzzles where a straightforward algorithmic code is not yet known.

Until now, only little work has been published in this area with respect to the protein folding problem [62–65]. Two-dimensional lattice simulations for conformational search performed with genetic algorithms have demonstrated to be significantly superior compared with traditional Monte Carlo methods [66]; populations of conformations of polypeptide chains are "mutated", here by conventional Monte Carlo search steps, and crossover takes place by exchanging parts of conformations between structural individuals. Genetic algorithms have also proven to possess high potential for the simulation of the evolution of motifs, i.e. the evolution of zinc finger structures starting from random structures, and the ab initio folding of a four β-strand protein maintaining a compact hydrophobic core of the protein [67]. The latter approach was subsequently used for the prediction of the main chain folding of small proteins, e.g. crambin, cytochrome, and hemerythrin [68].

An extraordinarily successful use of genetic algorithms has been reported for the docking in five distinct protein–ligand systems; here, the protein is considered to be semi-flexible, while the ligand is fully flexible [69]. Water molecules are included in the calculation in a way that the ligand must remove loosely bound water molecules from the docking site of the protein molecule in order to achieve binding. The results show excellent agreement with experimental data.

In contrast to genetic algorithms, numerous work has been published in the past few years on the usability of *neural network methods* in protein structure research. Computational neural networks mimic the animal neural system. In principle, simple neural networks consist of processing elements in several layers; the elements (called "neurons", in analogy to the respective biological units) are interconnected between the layers in a network-dependent fashion. Information and signals are transferred through these connections and processed by the neurons. The connections are numerically weighted; the weights are gradually changed and adapted in the "training phase" or "learning phase", until each pattern presented to the input layer of neurons is accurately projected onto the corresponding resulting pattern on the output layer. Predefined threshold values of the incoming signals have to accumulate in each respective processing element before an output signal is passed on to the connected neurons in the next layer. After an iterative optimisation regarding the optimal network topology, and the subsequent adaptation of the network parameters and weights to the problem under investigation, the network is ready to be used in the "recall phase", where patterns are presented to the input layer that were not part of the training phase. An important advantage of neural networks is their sensitivity to detect subtle patterns in the incoming data which may in some cases not be recognized by statistical or algorithmic methods. Also, neural networks may be applied to problems even without prior knowledge of an algorithmic correlation between the input and the output data, such as the relationship between the primary sequence and the native structure or the regular secondary structure elements of a protein.

The approach has gained much attention in several areas: (i) secondary structure prediction of proteins is the main application for neural network research at the moment, and will be discussed in detail later in this review [35,70–82]; (ii) predicting structural and functional features of proteins and nucleic acids [83–88] as well as sequence analysis, phylogeny, and rational design [89–91]; (iii) analysis of spectral properties of proteins, especially far-ultraviolet (far-UV) circular dichroism (CD) spectra [92–94] or nuclear magnetic resonance (NMR) spectra [95] (cf. Section 17); and (iv) tertiary structure prediction, often combined with energy minimisation [60,96,97]. On the other hand, a critical analysis of the performance of a neural network compared with sophisticated statistical methods [83] has demonstrated that the advantages of neural network ap-

proaches may eventually disappear when both types of analysis are carried out with greatest care. A summary on neural network methods in protein structure calculation, especially in the area of secondary structure prediction, can be found in [72,84].

Graph theoretical methods have been tested on predicting β-sheet topologies from protein sequences [98]; here, graph theory extends the topological description of proteins by the efficient incorporation of long-range interactions and connections. Four different notations characterizing the topology were derived by this approach; these may be used further to analyze and describe β-sheet topologies in proteins. In a different approach, graph theoretical aspects were used for the formulation of four rules describing commonly occurring kinetic processes in enzyme kinetics and in protein folding kinetics [99]. Apart from graph theory methods, rules of formal grammar were the basis for a linguistic approach to sequence analysis [100]. The high-order structure of biological sequences may be analyzed by general-purpose parsers for syntactic pattern recognition, i.e. in eukaryotic protein-encoding gene sequences; this helps in distinguishing species-specific signals from compositional or syntactic components in gene structure prediction.

3. Traditional databases

The most promising approach for a reliable and useful protein structure prediction from sequence data are knowledge-based approaches [101,102] (see Section 13). These algorithms rely on information derived from experimental data on the structures of biological macromolecules. The amount of data which is available today (and which is still increasing exponentially) has therefore led to increasing demands for an efficient storage and retrieval system for these data, i.e. commercially available relational database systems [103]. Apart from the traditional flat-file data storage system which is used in the common Brookhaven Protein Database (PDB) and in wide-spread sequence databases (EMBL, SwissProt, GenBank), there are special databases with a higher level of organisation either under development [104] or already available [105]. A searchable database for three-dimensional structures has been developed from the chemistry database of the National Cancer Institute (NCI) Drug Information System (DIS), containing approximately 450 000 compounds, which have been tested by the NCI for their anticancer activity [106]. Other useful databases were generated for structurally aligned folding families [107] and for protein structure–sequence alignments [103,108].

An interesting activity that is already commonly used is the connection of the major protein and nucleic acid sequence databases with relevant publications cited in MedLine, a database system named "ENTREZ" [104] which is maintained at the National Center for Biomedical Information (NCBI) and the National Library of Medicine (NLM). It may be obtained in a version on CD-ROMs (currently five) or accessed on-line via the Internet.

The definition of concepts for standardized molecular sequence data storage and access is still under development; a data description standard proposed at the NCBI (named ASN.1) may be expected to be the basis for sequence data formats for the future. Most of the databases in structural biochemistry and molecular biology follow different strategies and concepts, but as a common denominator many of them were intended as tools for knowledge-based approaches to protein structure prediction. A database on secondary structures (PSS: Protein Secondary Structures) that correlates sequence data derived from the PIR sequence database and (crystallographically derived) structures from the PDB has been developed [109]. A central feature of this database is that secondary structures, sites, regions, and domains of structural interest are displayed in conjunction with the respective sequence; this may be a helpful tool for the analysis of new sequences. The computer software also includes retrieval of corresponding peptide fragments, and hydrogen bonding patterns.

The database of known protein sequences is approximately two orders of magnitude larger than the database of known structures [108,110]. Therefore, it is useful to extend the experimental data in the structural database by homology-derived models calculated from sequence data: sequences collected by a special comparative algorithm are here held together with a homologous structure, and the derived secondary structure information of the resulting "secondary structure models" are compiled in a special

database (HSSP) containing the aligned sequences, secondary structure, sequence variability, and a sequence profile. Model structures of the aligned sequences may be implied, but are not explicitly stored. The database may be useful in determining the structural significance of matches in sequence database searches, and in deriving the structural role of conserved residues.

Another approach [111] dissects the database of known protein structures into (currently) 154 different families, with a sequence identity below 30%; related sequences (30–70% sequence identity) are clustered within the respective family, whereas highly related (> 70% sequence identity) protein chains are removed from the database due to the redundancy of the information. The database may be useful for understanding protein architecture as well as for the design of proteins, and is publicly available. A more recent approach used a different method for the generation of a non-redundant structure database [112]. First, sequence alignments for proteins from the structure database with more than 35% sequence identity were used to identify families of homologous proteins. From these families, one representative (highest resolution, best R-factor) was selected; structure comparisons between all members of this set revealed a dataset of homologous proteins. An extension of the method was applied to generate a dataset of analogous proteins, with related folds but more diverse structures. From 1410 protein chains used for this work, a set of 150 non-homologous families (and 112 non-analogous folds) were derived. This is in close agreement with the work described before. The "non-biased" redundancy in the protein structure database [113] is often an undesirable feature for statistical analysis; therefore, Hobohm and coworkers addressed this problem by two different algorithms, resulting in a dataset of 155 chains for the largest non-redundant set of non-homologous proteins (30% identical residues for alignment subsegments longer than 80 amino acids), again in close agreement with the work described above.

A database with information on hydrogen bonding in highly resolved protein structures has been published recently [114]. The authors examined 42 structures from the protein database in order to find rules and trends for hydrogen bonding and hydrogen bonding patterns: 68% of all hydrogen bonds are between backbone atoms; secondary structure elements have extended hydrogen bond networks, an average of 82%; almost all backbone hydrogen bonds comprise interactions within a single secondary structure element; sidechain to backbone hydrogen bonds are clustered at the termini of helices; helices often possess networks of hydrogen bonds, not only simple 1:1 relationships between donor and acceptor residues; the number of hydrogen bonds is roughly proportional to the content of periodical secondary structure in the protein, and linear proportional to the number of residues. These rules and trends could be valuable sources for further research, i.e. for machine learning approaches to structure prediction and verification.

A database with 316 different properties computed for 23 highly resolved protein crystal structures has been demonstrated to be useful in the mathematical definition of "frequent" features of proteins and serves as a valuable tool for protein engineering, structure verification for both crystal structures and homology-derived protein structures, and property definition, i.e. for dissecting extremophilic proteins from mesophilic ones [115]. A summary on methods for database analysis with respect to protein engineering has been published recently [116].

4. On-line information access via the Internet

The increasing amount of data in molecular biological and structural disciplines that are available today, together with the increased complexity of data verification, quality control, and hardware, software, and maintenance costs, make it highly desirable for the research community to have easy on-line access to centralized databases via public networks, such as the common Internet [117–122]; for an introduction into the history, structure, and services of the Internet, see [123–126]. The advantages of world-wide, distributed databases compared to (many) local database copies are obvious: (i) the data contents are always as actual as possible; (ii) storage and maintenance costs are minimal; (iii) optimal expertise on the databases can accumulate at the responsible site, thus generating "centers of excellence"; (iv) software and data contents or data format updates, and

Table 1
Some important mailservers for molecular biology and biomolecular sequence and structure research, taken from the list collected by Dr. Amos Bairoch

Service name	Respective organisation	Description	E-mail address
Biosafety	International Centre for Genetic Engineering and Biotechnology (ICGEB), Trieste (Italy)	Retrieval of documents that covers various safety aspects of biotechnology, especially: laboratory chemical and biosafety; use and release of genetically modified organisms; biodiversity issues pertinent to biotechnology	docserver@icgeb.trieste.it
BLAST E-mail server	National Center for Biotechnology Information, National Library of Medicine, NIH/Bethesda (USA)	The BLAST family of programs employs an heuristic algorithm to compare an amino acid query sequence against a protein sequence database or a nucleotide query sequence against a nucleotide sequence database, as well as other combinations of protein and nucleic acid searches	blast@ncbi.nlm.nih.gov
BLITZ electronic mail server (MPSRCH)	Edinburgh University Biocomputing Research Unit (UK), and European Bioinformatics Institute, Hinxton (UK)	MPsrch allows you to perform sensitive comparisons of your protein sequences against the SwissProt database using the Smith and Waterman best local similarity algorithm. Runs on the MasPar family of massively parallel machines; the fastest implementation of the algorithm available on any machine	blitz@ebi.ac.uk
GRAIL (Gene Recognition and Analysis Internet Link)	Oak Ridge National Laboratory (USA)	System for predicting protein coding regions in human DNA sequences using a neural network approach	grail@ornl.gov
HUGEMAP	CEPH-Genethon, Paris (France)	Allows to access Genethon's human physical map data on clones, YACs, or STSs	hugemap@genethon.fr
MOWSE	Imperial Cancer Research Fund, and SERC Daresbury Laboratory (UK)	Peptide mass fingerprint E-mail server service. MOWSE allows the identification of known proteins from a set of molecular weights (mass spec) determined after proteolytic digests	mowse@dl.ac.uk
NetGene mail server	Department of Physical Chemistry, Technical University of Denmark, Lyngby (Denmark)	Produce neural network predictions of splice sites in vertebrate genes	netgene@virus.fki.dth.dk
nnpredict	University of California, San Francisco (USA)	Analyze a protein sequence and send back a prediction of the secondary structure using a two-layer, feed-forward neural network method	nnpredict@celeste.ucsf.edu
PredictProtein	Protein Design Group, European Molecular Biology Laboratory, Heidelberg (Germany)	Analyze a protein sequence and send back a multiple sequence alignment performed by a weighted dynamic programming method (MaxHom) and a secondary structure prediction produced by a profile network method (PHD)	predictprotein@embl-heidelberg.de
Protein Sequence Analysis server (PSA)	Biomolecular Engineering Research Center, Boston University, Boston, MA (USA)	Analyze a protein sequence and determine which sequence–structure models are the most probable explanations of the input sequence. The analysis is particularly well suited for analyzing novel sequences that are unlike any others in the sequence databanks	psa-request@darwin.bu.edu
RETRIEVE dbSTS E-mail server	National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda (USA)	Allows the retrieval of Sequenced Tag Sites (STS) sequences from the dbSTS collection	retrieve@ncbi.nlm.nih.gov

The complete list may be obtained by using the standard Internet file transfer "ftp" to the computer "expasy.hcuge.ch"; in the directory "databases/info" the file "serv_ema.txt" always contains the latest catalogue of server. Examples shown are from the current version 3.11 of the database.

continuous support are easier to maintain. In the following, only two aspects of the broad spectrum of information access via public "data highways" are considered: simple access to mailserver by electronic mail (E-mail) [119,127], and comprehensive access via the world wide web system (WWW) [118,128].

Mailserver are special computer implementations that allow users to access data or services by simply sending an E-mail embedding special commands which the server is able to interpret [82,129]. A ubiquitous command for all mailserver is usually "HELP", which automatically causes the server to send a list of allowed commands to the requesting E-mail address. Mailservers are now usually substituted by the much more comfortable way of interactive information access via the WWW. In Table 1, some commonly used mailservers for molecular and structural biology are listed. The most comprehensive database of mailservers is maintained by Dr. Amos Bairoch from the University of Geneva; Table 1 shows some examples from this list.

The WWW consists of a dynamic network of computers that provide services to the community, i.e. database access. The hypertext network in WWW allows access to all services virtually via any entry point to the network; Table 2 lists some important servers for biomolecular structure research. There are currently more than 10000 WWW servers on the Internet, and the number is still growing. The WWW is therefore discussed to be one of the most important information tools for research in the future, with virtual unlimited access to any knowledge databases. Due to this increasing size of information sources, there is an increasing demand on "search robots". A most useful search entry point in this respect is the "WebCrawler" system (http://webcrawler.com).

5. Sequence-based calculations and the role of individual amino acids

The computational analysis of protein and nucleic acid sequences in order to elucidate structural or functional constraints, and for comparison of consecutive information contents in biological macromolecules, is still one of the fundamental tasks in molecular bioinformatics [18]. In studies regarding the sequence–structure relationship at the termini of parallel β-strands [130], there was significant conservation among the properties of the residues at

Table 2
A selection of useful sites at the World Wide Web that are relevant in the context of this review, and may be useful entry points into the web

Description	Hypertext link
Brookhaven Protein Database	http://www.pdb.bnl.gov/
Cambridge Crystallography Data Centre	http://csdvx2.ccdc.cam.ac.uk/
CERN Home Page	http://info.cern.ch/
EBI, the European Bioinformatics Institute	http://www.ebi.ac.uk/
EMBL Bioinformatics Resources	http://www.embl-heidelberg.de/
European WWW Server	http://www.tue.nl/maps.html
GenomeNet WWW server	http://www.genome.ad.jp/
Harvard University: Bio-services worldwide	http://golgi.harvard.edu/biopages/all.html
Harvard University: Biochemistry services	http://golgi.harvard.edu/biopages/biochem.html
Johns Hopkins BioInformatics Web Server	http://www.gdb.org/hopkins.html
Library of Congress WWW Home Page	http://lcweb.loc.gov/
NASA Jet Propulsion Laboratory	http://www.jpl.nasa.gov/
NCBI Blast Server	http://www.ncbi.nlm.nih.gov/Recipon/blast_search.html
NCBI Entrez (DNA/RNA, Protein, Medline)	http://www.ncbi.nlm.nih.gov/Search/Entrez/index.html
NCSA's Homepage	http://www.ncsa.uiuc.edu/General/NCSAHome.html
NIH/Natl. Center for Biomedical Information	http://www.ncbi.nlm.nih.gov/
NIH Molecular Modeling Home Page	http://www.nih.gov/molecular_modeling/mmhome.html
Recombinant DNA Technology Course	http://lenti.med.umn.edu/recombinant_dna/recombinant_flowchart.html
Restriction Enzyme Database	http://nearnet.gnn.com/wic/bio.12.html

For finding databases and information services on special topics, search robots should be used, e.g. the "WebCrawler" (http://webcrawler.com).

these sites (''β-breakers''). The authors demonstrate that these residues are conformationally homogeneous with respect to sidechain solvent accessibility, and backbone dihedral angle profile. Only a small subset of amino acids (with higher conservation than the rest of the residues under consideration) is observed at these sites, corresponding to the capping signals found for helices [131]. Here, the authors conclude that the helix signal (N-terminus) consists of a reciprocal backbone–sidechain hydrogen bonding ($N_{sidechain} \rightarrow N + 3_{backbone}$) interaction termed ''capping box''; the signal is found at the N-termini of helices in proteins (15 examples in a set of 161 helices from the protein database). These rules for capping of periodical secondary structure elements, together with studies of the natural amino acids on their helix-forming capacity in alanine-based peptides [132], may improve secondary structure prediction methods and machine learning approaches for protein architecture studies.

The special role of *cis*- and *trans*-proline and hydroxyproline, their structures in collagen, and the influence of neighboring chemical groups on the structure of proline has been examined by chemical synthesis and X-ray diffraction studies [133]. In order to determine the influence of proline residues on the conformation in their native environment in proteins, all proline residues in the Brookhaven protein database were examined with respect to local structures [134]. The outcome was – as expected – that proline has a highly unique role for the determination of local conformations. An extension of the approach with information theory applied to the probability for all natural amino acids to be in a distinct zone of the Ramachandran diagram demonstrated that the influence of local sequence information determines about 65% of the local conformation [135], with the rest (35%) influenced by long-range interactions. This result has been previously discussed to be the main limiting factor for the success rate of secondary structure prediction.

A major problem of structure prediction by homology (knowledge-based structure prediction) is still the alignment of two or more related sequences [108,110,136–138]. At low sequence homology, the best alignments found by scoring schemes may often not be the structurally correct ones; this has led to the development of approaches for the evaluation of suboptimal alignments [139]. Recently, filtering of suboptimal alignments based on core volume considerations and packing potential has been investigated [140]. The algorithm has proven to be superior to traditional optimizing alignment methods in cases of the difficult alignments of immunoglobins, distantly related globins, and plastocyanin with azurin. The resulting filtered alignments are demonstrated to correspond more closely to the structurally correct alignment than the alignments performed with traditional methods.

Ring et al. [141] have analyzed 432 loops (4–20 residues in length) from a total of 67 proteins, thereby generating a conformation-based classification scheme for loops in proteins. By using a special denotation derived from this structural analysis, they found that these loops have positional preferences for amino acid residues similar to those described for β-turns.

The correlation of secondary structure and sidechain torsion angles with the hydration of serine, threonine, and tyrosine [142], has revealed that both parameters have a significant influence on the pattern of solvation of serine and threonine, but not on tyrosine, since the hydroxyl group is too far away from the main chain to reflect secondary structure. Crystallographically resolved water molecules on the surface of proteins may bridge hydrogen bond donors and acceptors, thus adding a further problem to the prediction of protein structures from sequence. For a reliable prediction bound water molecules must be taken into account.

An important point mentioned in the past was the meaning of local similarity between unrelated proteins for structure prediction. The question is to what extent the local sequence influences its corresponding structure. Here, a careful approach revealed the important result that local sequence does *not* necessarily indicate a structural similarity except for evolutionary or functional reasons [136]. Similar, non-related sequence fragments from the protein database (often with 25% identical sequence, plus 25% ''conserved'' exchanges) do not exhibit any structural relationship. This means that prediction methods based on fragment built-up from functionally unrelated proteins are unlikely to be generally useful, despite the fact that these methods have been demonstrated to be successful in singular cases in the past.

The same holds for the use of patterns of physicochemical properties associated with a certain sequence for secondary structure prediction [143]; however, the current size of the protein database may pose severe limitations on the predictability of secondary structure elements (see above). The authors conclude that with a sufficiently large database the prediction of secondary structures may unravel segments which are important for early folding intermediates. In contrast, if homologous protein families and subfamilies are used as the primary database for sequence–structure correlations [144], there is convincing evidence that regions may be predicted to have a unique and unambiguous structure related to the local sequence. These common sites are thought to serve as folding nuclei for the assembly of surrounding elements since they are heavily involved in interactions with the rest of the protein. Experimental data available for some of the cases studied are in accordance with the hypothesis.

The prediction of surface regions from sequence has been improved markedly [145]. Tripeptides containing polar residues are used as markers for surface probability statistics; 83–86% of the observed surface regions were correctly predicted, with reduction of the number of wrong overpredictions compared to previous reports. In contrast to this, evidence is now accumulating that sequence statistics alone do not allow the prediction of e.g. the extremophilic properties of proteins from halophilic or hyperthermophilic sources [146]; the proposed "traffic rules of adaptation" to extreme environments do not hold when an unbiased database is used for the statistics, and when evolutionary diversity is mixed up with adaptation mechanisms.

6. Protein secondary structure prediction

Several reviews, methods, and publicly available computer programs have been published in the last few years on the prediction of protein secondary structures [3,4,6,37,39–41]. Secondary structure prediction methods may also be applied for the improvement of interpretation of low-resolution experimental results, i.e. for the prediction of the structure of the proteasome [147].

A promising approach in the area of prediction is the use of neural network methods [72,73,80] (see above). One of the first examples for this method [70] used 48 proteins in the learning dataset, in order to teach the relationship between primary sequence and secondary structure to the neural network. For the recall dataset of 14 proteins in a three-state study (helix, sheet, coil) the overall accuracy was 63%, with a peak accuracy of 79% for the strongest predictions (for 31% of the residues). This early approach therefore marks no major improvement compared with traditional methods of secondary structure prediction by statistical and knowledge-based methods. However, a complex, cascaded neural network for the prediction of secondary structure fractions of globular proteins [75] already showed improvements by the inclusion of information on the probability of residues to be buried in the protein core (or on its probability for the surface of the protein), and by inclusion of non-specific long-distance contact maps. The average success for this approach was 68.3% correct prediction for three different types of secondary structures.

Training of a two-layer, feed-forward neural network on 130 non-redundant proteins from the protein structure database [79] led to a final accuracy of prediction of 70.8% for globular, water-soluble proteins, again an improvement over the method described above. Additional improvement was gained by inclusion of related families of proteins (identified by multiple alignments) instead of single-member families. Extensive tests were performed on the value and accuracy of the method; of particular interest is the assignment of reliability indices for each position. Sites predicted with high reliability index had prediction accuracies of 82% and better.

In contrast, the comparison between prediction of all-helical proteins and prediction with the commonly used three-state model (helix, sheet, turn) [77] indicated that no significant advantage may be achieved by the reduction of predicted states; this holds for a neural network approach as well as for inductive logic programming. Results are also similar for different datasets, with 12 [57] or 130 [79] proteins, and different output layer characteristics. In a recent study, however, the introduction of constraints on structural classes of proteins led to significant improvements in prediction accuracy of a neural network, resulting in up to 79% correct prediction of all-helical proteins, despite the low number of

cases available for the study [71]; the limits on the all-helical prediction approach are discussed elsewhere [74]. In total, the neural network method for the prediction of secondary structure elements from sequence data has steadily been improving over the past five years, and is now a widespread method for prediction with average accuracies between 65 and 80% correctness.

Another technique uses predicted physical properties of proteins (instead of the text-like representation of sequences which is commonly used) as a marker for a certain sequence position and implements a digital encoding algorithm [148]; the correlation between sequence-derived profiles (but not the sequence itself) and the respective secondary structure can be predicted with an average accuracy of approximately 75% by this method.

A machine-learning approach (termed "GOLEM") using relational descriptions has been shown to be superior to traditional statistical prediction methods [57]. Again, physical and chemical properties are used as background knowledge by the program. Only α-helical proteins (12 proteins from the structure database) were used for the input data and learning set, and 4 non-homologous proteins were used as test cases. Prediction accuracy was 81% correctly predicted residues, which is significantly better than the standard Garnier–Osguthorpe–Robson (GOR) method (72%) and previously published, simple neural network approaches (76%). Slightly worse results were reported with a pure but sophisticated pattern-based approach [36] for α-helical proteins; a success rate of 71% for overall prediction and 78% for core helical features was observed. N-caps, helix core, and C-caps were investigated and predicted separately (and with markedly different success) in this model. A similar success was reported for the use of a more sophisticated nearest-neighbor algorithm [149]; success rates rank up to 71%.

In general, there are two major approaches for further improvement of secondary structure prediction independent of the algorithm and method used. (i) The usage of homologous sequence information for the prediction (reviewed in [34]); the success rate for the prediction of triosephosphate isomerase (TIM)-barrel like proteins has been reported up to 95% correct prediction in these cases [150]. (ii) Inclusion of knowledge from experimental data derived from far-UV CD (or related) spectroscopy [151,152]. Here, up to 5.3% improvement for a standard three-state prediction may be achieved compared to a solely sequence-based prediction. A more recent approach to predict protein secondary structure using CD spectra has been published [93]; here, two different types of artificial neural networks were tested: (i) a three-layer backpropagation network, and (ii) a hybrid self-organisation to backpropagation network. The training dataset consisted of the CD spectra of 22 proteins using the jack-knife technique for testing the prediction on each protein. The performance compares well with that obtained by other statistical and neural network methods, and may even increase with an expansion of the basis dataset.

A further procedure for the improvement of secondary structure prediction is the use of combination of several (mostly independent) algorithms and approaches, thereby hoping that disadvantages and weaknesses of the methods used cancel each other. A hybrid approach with three parallel expert systems has been described [76], using a neural network algorithm, a statistical method, and a memory-based reasoning module. First, the three modules learn predictions independent of each other (similar to the traditional methods); then, a "combiner" module makes final decisions based on the output of the three expert modules. Interestingly, for about 20% of the cases (107 proteins in the study), all of the three expert modules made identical but wrong predictions, thus suggesting that this barrier of prediction accuracy cannot be overcome by using local information only, and may indicate spots where non-local interactions play a significant role for the conformation. For 64% of the residues, two expert modules predicted the correct secondary structure, and for 77% of the residues at least one expert module made correct predictions, thus suggesting that there is still some space for improvement of the method.

Another combined approach used a two-step method [145,153]. The first step was the discovery of boundaries of periodical secondary structure elements by combining hydrophobicity, accessibility, and flexibility parameter-based predictions. Within these boundaries, five different statistical algorithms were used in the second step to predict the local class of secondary structure. After fine-tuning of the approach by elimination of less successful methods,

the method ended up with 80% correctly predicted regions in 30 proteins of known structure. Despite inherent problems of the method (discussed in [153]), the approach has been proven to be useful for further studies. The combination of neural network methods and evolutionary information [81], and the usage of multiple alignment of the sequences under consideration within their family of related sequences [38,154] can also improve otherwise weak secondary structure prediction attempts.

A joint prediction which combines the "best" aspects of six different methods, including statistical and neural network approaches, has been described in [35]. Several steps were taken in the adjustment of the method: (i) optimisation of each method individually; (ii) weighting of each method; (iii) combination of scores from the methods; (iv) discriminating three secondary structure type scores (helix, sheet, coil) at each residue position in order to indicate the secondary structure of highest score. Application of the program termed "Q7-JASEP" to 45 proteins from the database revealed excellent results and accuracy comparable to other combination approaches, and better performance than each of the six methods alone.

An optimisation of the widely used GOR method for the prediction of secondary structures with respect to β-structures has been published [155]. Internal and external β-sheets are discriminated based on their hydrogen bonding pattern; this is an extension of the method towards three-dimensional structure predictions, since additional information regarding the location of predicted sheets is presented. A critical analysis of several implementations of the GOR method is discussed in [41].

In conclusion, there is no doubt that secondary structure prediction methods tend to be refined and improved, with slowly increasing success rates over the last ten years. On the other hand, one should bear in mind that secondary structure predictions are only act 1 of the protein tertiary structure prediction drama, and are useful in this respect only when there is 100% accuracy and 100% reliability. Also, information on the secondary structural class of a certain residue only allows an approximation of its backbone conformational state, since there is still considerable conformational heterogeneity within the topological classes "helix" or "sheet".

7. Tertiary structure prediction

Some recent reviews and comments summarize the progression in the area of protein tertiary structure prediction from sequence data [16,156,157], especially with respect to improvements in the traditional approaches. Due to their small size, peptides are often used as primary models for the evaluation of the principal applicability of complex methods. In this respect, the conformation of the 29-residue rat galanin neuropeptide was studied using a Monte Carlo method combined with energy minimisation (MCM) and a further electrostatically driven Monte Carlo (EDMC) method [158]. The polypeptide chain is first treated in a "united-residue" approximation, in order to explore the virtually complete conformational space. Subsequently, the low-energy united-residue conformations are converted to an all-atom representation, and EDMC simulations are carried out for the all-atom polypeptide chains. The lowest-energy conformation had a non-helical N-terminal part packed against the non-polar face of a residual helix that extended from proline 13 toward the C-terminus. The final results are in qualitative agreement with the available NMR and CD data of galanin.

A novel approach for tertiary structure prediction of small proteins has been presented by Scheraga [159] and coworkers on the simulation of the folding pathway of bovine pancreatic trypsin inhibitor (BPTI). The protein is described as an ellipsoid characterized by three principal radii, and the pathway from the unfolded to folded state(s) are described by a dynamic equation. Constraining functions are introduced to bias the folding path to a compact structure. The result is a reproducible and unique pathway leading from an unfolded to a native-like structure [root mean square (r.m.s.) deviation between model and experimentally derived structure: 1.9–3.1 Å]; however, the final model lacks the important [30–51] disulfide bond. It is noteworthy that the compact, final structure is virtually independent of the starting conformation chosen for the unfolded state.

A sequence template method for the recognition and prediction of proteins with the TIM-barrel motif has shown only limited usefulness with respect to the accuracy and reliability of the prediction [160]. The $(\beta/\alpha)_8$-motif is investigated in terms of conserved

function of local residue positions, and global packing and volume. Motifs derived from the sequence of segments were found not to be reliable indicators of the fold when database searches and alignments were performed, at least for distantly related proteins. On the other hand, a profile method based on local environment information of residues may be more successful: it predicted three structurally yet unknown proteins which are important for the biosynthesis of aromatic amino acids as probable (β/α)-proteins [161].

A new algorithm describes the knowledge-based generation of protein backbone and sidechain coordinates from Cα-coordinates [162]. The algorithm is based on the initial use of suitable peptide fragments mapped onto the backbone trace; subsequently, an optimal path for the complete backbone chain is searched. The algorithm has been shown to generate correct atomic positions within 0.4–0.6 Å. Sidechains are added based on a rotamer library in conjunction with Monte Carlo dynamics and simulated annealing, with an average error of 1.6 Å (r.m.s. deviation) for residues in the protein core. Also, the "fragment built-up" procedure developed earlier by the group of Dr. Harold Scheraga has been modified and improved by the inclusion of statistical data of non-random structural pairs of residues [163]. The calculations resulted in coordinate sets with similar conformations of many residues to the native state.

On the other hand, the usability of peptide fragments for modeling has been questioned by Fidelis et al. [164]. A comparison between modeling based on a database of known structures and a systematically generated dataset that contains all possible conformations of a peptide revealed that the systematic search procedure (i) generates almost all structures of short segments found in proteins, and (ii) in contrast to the database method results in low r.m.s. error structures for a set of trial segments embedded in the rest of a protein structure. Therefore, the systematic search should be considered the method of choice in comparative modeling, since the current structure database is obviously too small to allow reliable extraction of structural fragments from it.

The deduction of a complete structure from a subset of the coordinates, i.e. Cα-atoms, is an area of great interest, and several methods have been described with respect to reconstruction of complete coordinate sets from Cα-coordinates [165–167]. Taking two sets of information, the amino acid sequence and a few atoms (i.e. Cα-atoms), Dr. Michael Levitt was able to accurately model the structure of eight test proteins from the database by segment matching [165]. The fast and fully automated method utilizes fragments from highly resolved protein structures which are fitted onto the frame of the target structure. Deviations of all atoms between models and experimentally derived structures were in the range of 0.9 to 1.7 Å. Even if only half of the Cα-coordinates are present, or Cα-atoms are shifted in the initial template (up to 1 Å), the models are still reliable. In this work, it has also been demonstrated that averaged coordinates from several independent models can significantly improve the quality of prediction. By the same author, a simple lattice model was developed that allows generation of all possible folds of a given chain for small protein molecules [168]. By using simple structural and energetical criteria, native-like folds are separated from the vast majority of non-native structures for five small, unrelated test cases. The method may be applied generally, without prior knowledge of structural details. The combination of the two approaches may yield a new methodology for ab initio tertiary structure prediction, with a lattice model in the first step and the structure refinement by segment matching in a second step.

Another method was developed to predict backbone tertiary structure folds from amino acid sequence alone [169]. The method uses a simplified representation of sidechain flexibility, and only seven backbone structural states are allowed in this model. In a further step, potentials of mean force are used to refine the model with respect to local environments of amino acids along the chain. The method allows fast computation of a number of low-energy structures, and results are satisfactory when compared with X-ray data, at least for short peptides where local interactions play a dominant role in structure determination.

Simple or complex energy minimisation methods alone, or free-energy calculation approaches, do not yet allow the deduction of a structure from a sequence; however, when a sufficient number of additional constraints exist, such as predefined atom–atom distances derived from experimental data

("constrained optimisation"), a unique structure may be predicted [8].

In contrast to the problem of folding globular proteins starting with a sequence and ending up in a (native-like) structure, the inverse protein folding approach addresses the question which sequences are compatible with a given (known) structure, i.e. which sequences can adopt a common fold, despite their respective sequence relationships. The method is an excellent supplement and verification for lattice models designed for the ab initio prediction of structures from sequences only. Reviews and key ideas on several approaches to the inverse protein folding problems are presented in [22,137,170–173]. Data discussed in [172] for artificial proteins are in good agreement with experimental data, and the structure of the four-helix dimer rop was predicted with 3–4 Å accuracy.

A mathematical formalism based on the theory of Markov random fields to the inverse protein folding problem is introduced in [174]. Here, a large set of (assumed) tertiary structures is used as starting point, and the structures most compatible with a new sequence are deduced. The algorithm allows explicit representations for the relevant amino acid position environments and the topologies of the structural contacts. The approach leads to a new, comprehensive scoring function for comparing different threadings of a sequence through the structure models. The scoring function is very important for the success of inverse protein folding approaches, since the comparison of alternative structure models with each other is usually the key step of the method.

The prediction of a structure from sequence by evaluation of the compatibility of a sequence with a hypothetical structure is presented in [175]. Here, native structures are described as intramolecular and solvent-directed contact interface vectors, and a database of native vector types is derived. Then, the sequence under consideration is fitted into a large number of possible tertiary folds, and the quality of each model is evaluated by the sequence preferences summarized over the folded chain. Finally, the most likely structure (most compatible fold) is selected. Applying several test cases and verification procedures, the correct fold in the correct alignment may well be identified amongst all alternatives, even at low sequence similarity. The method is clearly superior to traditional secondary structure prediction methods, and allows the structural prediction of unknown proteins with relatively high rate of reliability. However, the general applicability still has to be shown. This sequence–structure threading approach is now a method which has gained much interest [176]; for an overview regarding the success and pitfalls of the method, see [17].

Two different approaches for structure prediction are presented in [177]. The structure of Interleukin-4 was predicted using experimental data from mutagenesis and CD spectroscopy, together with heuristic prediction methods and constraints derived from disulfide bonding patterns. However, the predicted structure was similar (enantiomer) but not identical to the experimentally derived structure. On the other hand, proteinase structures important for medical therapy of infectious diseases were computed by comparative modeling, and used as templates for drug design. Experimental verification of the applicability of these non-peptidic drugs demonstrated their biological activity. This demonstrates an increasing trend in molecular bioinformatics: combining theoretical methods with experimental data and approaches on the protein under investigation allows significant progress in both areas of research (cf. Section 17).

Simple models of protein folds are often used in order to circumvent the problems of a precise algorithmic imitation of nature, and the current limit of computing time even with the most powerful machines; an excellent recent review describing the current state of lattice models has been published in [178]. The application of lattice models to structure prediction is demonstrated in [179]; the structure of simple, artificially designed proteins containing regular helices was predicted by fast lattice dynamics. By using sidechain rotamer library information and potentials of mean force for quantification of short-range and long-range interactions, four-helix bundles are predicted which are in agreement with experimental data. However, prediction of proteins containing β-sheet elements has failed so far with the method described above. A different work, using random walk on a two-dimensional square lattice [180] and using simple inter-residue contact potentials, allows the discrimination of native from non-native folds. Including structure fragments of crystal-

lographically resolved protein structures into the simulation allows the determination of 37 different proteins based on only 8 basis structures. In a recent review, lattice models for the simplified representation of protein backbone structures were discussed with respect to their quality, applicability, and reliability [173], and therefore lattice models are not discussed in detail in this review.

8. De novo design

The design of molecules ab initio, generally without knowledge of related structures, is still a complicated task which is currently feasible only for small or structurally simple molecules or motifs. Most of the work published in this area was therefore performed on small organic molecules [50,55,181,182], or on simple peptides and proteins of regular tertiary structure [63,91,172,183–188]. The design of helix bundles is a particularly interesting project which has been performed in several groups independently, either on three-helix bundles [189], on four-helix bundles [190–192], or on seven-helix bundles [193]. Some other simple structures consisting of few periodical secondary structures connected by loops were designed during a molecular modeling course at the EMBL, but were not yet subject to experimental verification of the respective structures [191].

The identification of a protease inhibitor based on the structure of the binding pocket of a homologous protein (papain) has been performed on the cathepsin L enzyme [194]. The structure of cathepsin L was modeled from the homologous protease papain by comparative modeling techniques (see later in this review). The closely related binding pockets were then used as frameworks for a search through a database of small organic molecules; the size of the pocket was taken as the primary key for the search. The molecule identified by this relatively simple method was shown to be a potent inhibitor of both the parent protein, papain, and the modeled enzyme, cathepsin L. In a different work, a computer program (DEZYMER) was implemented that introduces new ligand binding sites into known protein structures [195]. Here, sidechains of residues at the binding site were exchanged in order to remove steric hindrance and allow optimal packing and hydrogen bonding. The successful implantation of an artificial ion binding site for copper into *E. coli* thioredoxin indicates that the method may be of value for several other applications.

A novel, comprehensive approach to the creation of artificial and modified proteins has been elaborated [196]; this includes a sequence design based on protein secondary structure and folding patterns, gene expression in a cell-free system, and testing of structural properties of the synthesized polypeptide. A new synthetic protein called albebetin has been designed to form a fold which does not contradict any structural rule, but on the other hand has been never observed up to now in nature. It could be shown that the artificial protein is nearly as compact as natural proteins, unfolds cooperatively at high urea concentrations, and has some structural features of a defined structure consistent with the designed one; the experimental verification of the structure, however, has still to be done.

Another methodology for designing proteins de novo has been published [184] which automatically produces sequences that are compatible with a predefined tertiary structure, similar to the previously discussed inverse folding approach. The method incorporates (i) statistical information, (ii) a theoretical description of the protein structure, and (iii) structural motifs described in the literature. Two initial model systems have been tested for the phage 434 Cro, and fibronectin type III domain folds. The sequences selected by the algorithm are not homologous with known sequences from the respective families, and structural models based on the sequences appear reasonable.

9. Surfaces of proteins and docking

The accurate description of molecular surfaces is a most difficult task since these surfaces are highly complex, irregular, and dynamic in solution [197,198]; a description of protein surfaces in terms of fractal geometry instead of the traditional euclidic geometry may be useful [199]. On the other hand, these surfaces determine the chemical reactivity, i.e. the functionality, of the respective enzyme as well as molecular recognition [200] and regulation processes, and may even be used to discriminate native folds

from non-native ones via solvent interaction analysis [201,202]; an excellent review on this item may be found in [203].

The docking problem, i.e. the definition of complementary surfaces, binding orientations, and interaction energies of two or more distinct molecules, may be simplified by the description of the interaction of two surfaces with one fixed molecule and the other molecule rotated and translated (six degrees of freedom), in order to find optimal scores of interaction. Obviously, this simple rigid-body approach does not account for the dynamic properties of molecules and their surfaces, on one hand, and molecular effects like "induced fit" binding on the other. A counter-example for the applicability of rigid-body approaches for precise description of interactions is the binding of hirudin to thrombin, with major structural changes in hirudin observed upon association. The approximation of surfaces [197,198] may lead to simple insights into the properties of the molecule under investigation; however, in most cases a complex description is a more useful approach [115], although it is more time-consuming and computationally costly.

Noteworthy developments are the automation of protein/ligand docking based on surface complementarity [204–206] or with consideration of sidechain flexibility [207], the docking and recognition between protein and DNA [208,209], the modeling of interactions between lectin and oligosaccharides [210], the usage of conformational search methods together with precise algorithms for electrostatics [211], estimates of binding affinities [212], and the interaction analysis by molecular polarisation maps [213].

Surface analysis with respect to the interaction between protein and ligand may become a most important tool for the correct prediction of docking. In this respect, a recent work describing the interaction in HIV reverse transcriptase (HIV–RT) is a good example [214]. HIV–RT is a heterodimer with a complex tertiary domain structure; simplification of this domain structure by the alternative characterisation in terms of solvent-accessible surface areas allows the buried surface area of contact among the different subdomains and also the HIV–RT–DNA interactions to be described. Such analyses are important to extract features characteristic for specific interactions with the DNA sequence, and features that are not sequence-specific; here, non-specificity is correlated with an increased area of contact between DNA and protein.

Protein–protein interactions (i.e. interactions between an antibody and the respective peptide antigen) known from X-ray crystallographic analysis are often close-packed interfaces of relatively constant sizes (approximately on the order of 1.500 $Å^2$) and hydrogen bonding pattern; water is usually excluded from the interfaces [28,30–32,215,216]. A recently published approach for their prediction used a steric scoring scheme based upon a soft potential, and a simple electrostatic model for antigen–antibody docking [215]. Application to four model systems revealed a precision of 1.9–4.8 Å r.m.s. deviation for the prediction of the complexes. Computer simulations on similar interactions between molecules may be helpful for the identification of rules for the association of surfaces, especially for mutant proteins with altered association sites.

On the other hand, simulations performed with a grossly simplified molecular description (rigid-body models, crude energy functions) can – aside from correct predictions in some cases – lead to a misleading docking of molecules in other cases [25,216]. Especially the discrimination between the native or native-like complex (which is quite often among the most likely candidates of several docking models) and non-native docking orientations is a most formidable task. These non-native orientations often show similar contact areas, hydrogen bonds, and polar or electrostatic interactions similar or even numerically superior to the data for the native state. Thus, rational prediction of docking and association orientations between macromolecules continues to be an exciting area of research, with significant progress made in the past few years but not yet with a final, "universal" algorithm and code applicable to all kind of problems.

Also, the stability of a native, soluble protein depends severely on its interaction with solvent molecules, thus making the description of this term highly important for molecular modeling [217,218]. A simple model for the interaction between protein atoms and solvent atoms in the first hydration layer, the solvent contact model, is described in [219]. This model allows rapid estimates of physical properties

that depend on the number and type of protein–solvent nearest-neighbour contacts, allowing its use in the fast calculation of protein solvation energies, conformational energy calculations, and molecular dynamics simulations.

10. Packing in proteins

Early as well as recent work on the sidechain rotamer library approach for the prediction of residue conformation, and the important effect of local packing [6,220–225] has now been extended to hard-sphere models [226] of statistical and database approaches to structure prediction. Likewise, general tendencies of the fold of a protein are driven by the effort of the chain to assemble a densely packed core [223,227], a property that divides structured molecules like proteins from oil drops (both assemblies are thought to be driven by the hydrophobic effect).

The importance of tight packing on the correct folding of protein has also been demonstrated by recent, unsuccessful attempts in protein design [228]. The important features of closely packed cores for secondary structure formation has been confirmed by lattice and non-lattice methods [229]. It seems not always necessary to use full-atom model representations for the evaluation of the packing density; simple models of sidechains using one, two, or three spheres for the approximation of the respective volumes do allow a rapid estimate of densities [230]. With this method, residue contact pairs are also considered in order to indicate reliability and quality of a model structure.

The high impact of optimal packing in the core region on the structural folding process in proteins was also recognized in the past; a recent analysis suggests that this is achieved by clusters of hydrophobic amino acids, surprisingly without any statistically significant, preferred interaction between pairs of amino acid types [222]. On the other hand, a different work performed on identification and analysis of preferred interaction, packing, and orientation between pairs of residue types as well as atom types finds significant distributions [231] which now may be of value for the further development of modeling and verification techniques.

11. Energy minimisation and molecular dynamics

Energy minimisation and molecular dynamics (MD) calculations represent the traditional approaches for molecular structure calculations; however, there is still tremendous research in this area, with continuing progress in algorithms, methodologies, and experimental parameterisations, but also with respect to the now available enormous amount of adequate computer facilities and CPU time, i.e. on powerful workstations, parallel computers, and workstation networks [232] – at least in comparison to the "ancient" times of MD simulations. Recent reviews and related articles cover mainly the progress in force field theory and parameterisation [5,8,19,30,233–239]. Aside from the protein research area discussed below, other molecules of biological interest such as RNA [240] and DNA [241], lipid membranes [242], or ligands, e.g. cyclosporin A [243,244], were subject to investigations by MD.

Special emphasis has also been put on the simulation of unfolding kinetics and the properties of the denatured state of proteins by MD simulations [159,245–249]. Also, simulations of the folding pathways using a simple lattice model for globular proteins and Monte Carlo dynamics have been examined [250], with yet limited precision due to computational requirements and therefore also with only a poor value for the correlation with experimental data.

Energy embedding may be a suitable approach for the efficient determination of energetically low local minima in the potential energy hypersurface, eventually also the global potential energy minimum [2,251]. The common algorithm starts by locating a deep local energy minimum while the molecule is in a high-dimensional ($\gg 3$ dimensions) euclidean space, and then the molecule gradually smoothens down to the traditional three dimensions in a relaxed conformation. A variation of the method, called rotational energy embedding, has been presented where the descent into three dimensions is performed by internal rotations. The method was able to locate conformations close to the native state for avian pancreatic polypeptide and apamin, given only their amino acid sequences and a suitable potential function [2]. In addition, dimensional oscillation which is a computationally fast variant of energy embedding is introduced in [251]; it has also been found to

sample conformational space and local potential energy minima with high efficiency. However, there is no proof of the general and beneficial applicability of energy embedding methods yet.

A method for the construction of a potential function for approximate conformational calculations in singular cases of globular proteins has been developed [252]. The potential has been parameterised to reproduce the crystal structure of avian pancreatic polypeptide, and therefore is able to predict quite reasonably just this structure from sequence data (but, of course, no other protein or peptide structure). There is currently no evidence for a general applicability of this approach. Also, a very efficient Monte Carlo algorithm with additional usage of simulated annealing and simple potential energy functions was developed that generates reasonable models fast and reasonably accurate [253]. The algorithm could in general be used in automated, fast, and reproducible model building by homology. An empirical potential function for the recognition of protein folds has been developed in [254]. It is able – within its predefined limits – to select the native conformation for a sequence from a set of possible protein folds. None of these approaches could be extended in order to demonstrate its general usefulness yet.

A novel predictive method based on a digital encoding algorithm has been presented in [148]. Physical properties of an amino acid sequence instead of the sequence itself is used as input to the binary encoding algorithm. The property profiles are then used to predict secondary structures for proteins with an predictive accuracy over 75%; however, tertiary structures cannot be predicted yet, but the method may be extended with respect to this.

An alternative method to the knowledge-based approaches discussed later for the modeling of protein loops based on fast Monte Carlo calculations and simulated annealing was developed and subsequently tested on loops from immunoglobulin, BPTI, and trypsin [255]. In accordance with the immense reduction in the description of physical properties of the loop residues and the simple potentials that were used, the average accuracy of the modeling is only approximately 1 Å r.m.s. for the backbone and 2.3 Å r.m.s. for all heavy atoms. As a result, one can conclude that knowledge-based methods are still superior to this kind of MD simulation.

If MD simulations are carried out far away from any molecular equilibrium, even small differences in the starting conformations or parameterisation may add up during the simulation and may eventually result in significantly different results of otherwise identical simulations, a behaviour known as chaotic response [199]. Therefore, many publications in the past have set out to investigate protocols used in MD, parameterisations, solvent conditions, and more [97,256–265]. Of special interest is the calibration of dielectric properties of a protein by analysis of MD simulations [266]; however, depending on the exact model parameters and approximations, the estimates for the protein "dielectric constant" vary from 1.5 to 37. The results indicate that electrostatically precise models could treat the protein core as a low-dielectric medium, but the charged surface groups should – surprisingly – be considered as part of the solvent.

More selected examples on MD studies on distinct proteins are the human CD4 protein [267], bacteriorhodopsin [268], a MHC–peptide complex in order to predict potential T-cell epitopes [269], small helical peptides in solution or within biological membranes [270,271], subtilisin mutants [272], the antibody binding site [29], binding of inhibitors to dihydrofolate reductase [273], and anti-HIV drug design [274]; much work has also been performed on the dynamics of myoglobin [275–277]. Aside from these efforts on myoglobin to correlate MD methods with experimental values on the dynamics of a protein, the inclusion of thermal motions in crystallographically determined structures [278,279] and the hydration of cavities in proteins [280] is also under investigation.

12. Electrostatics and hydrophobicity

The theory of electrostatic interactions and the resulting hydrophobic effect in biological macromolecules was the subject of some recent reviews [223,281,282], and its relevance is still discussed controversially. In general, the electrostatic behaviour of macromolecules may be deduced from the analysis of properties of small molecules such as hydrocarbons [283,284]. Here, a model for the curvature dependence of the hydrophobic effect was de-

veloped, thereby showing that the macroscopic concept of interfacial free energy may also be applicable at the molecular level; this assumption is not trivial. Also, and again in the publications cited above, the hydrophobic effect is found to provide a major driving force for protein folding. The calculated strength of the hydrophobic effect (which is quantified approximately twice as high in this work as it had previously been described) is discussed also with respect to substrate binding and nucleic acid base stacking, and general interpretation of computer simulations. The results are in agreement with the outcome of site-directed mutagenesis experiments, and have led to a new, improved scale of hydrophobicity for the naturally occurring 20 amino acids.

The significance of the hydrophobic effect in the stabilisation of helical structures was investigated in [285]. Experiments on short peptides with modified, neutral lysine residues at various positions showed that methanol or low mole fraction mixtures of trifluoroethanol in water induce helical stability. This effect was, as expected, highest for a peptide with a pair of (modified) residues spaced by three other residues. Together with recent calculations on the electrostatic potentials in proteins and peptides in conjunction with conformational search strategies [262], and the analysis of the effect of point mutations of hydrophobic residues on the protein stability [286], these findings may be helpful for the design of energetically stable helices in proteins.

A recent, careful analysis of the electrostatics of a total of 141 protein structures from the protein database has led to a better understanding of proteins of different functional and folding type [287]. Here, the analysis of charge–charge interactions to the free energy of the native protein in comparison with randomly scattered charge distributions revealed that (i) charge–charge interactions are better optimised in proteins with enzymatic function than in proteins without enzymatic functions (i.e. structural proteins); (ii) β/α-proteins are electrostatically better optimised than pure α-helical or β-strand structures; (iii) proteins with disulfide bonds obviously show a lower degree of electrostatic optimisation; and (iv) the rejection of repulsive contacts may be a more important driving force for protein folding than previously recognised.

Also, there is growing evidence that a careful description of electrostatic effects at the interface between the macromolecule and the solvent is of highest interest for the analysis of protein stability and folding. In this area, several key publications have appeared in the last years, e.g. the comparison of water solvation of BPTI with theoretical models [288]. Neglect of electrostatic hydration energies can introduce significant errors in molecular mechanics calculations, and should therefore always be explicitly included in calculations, for example by a term based on the widely used finite difference Poisson–Boltzmann (FDPB) algorithm [289]. This term can be quite easily incorporated into existing force fields such as the popular CHARMm force field. A new, robust, and efficient approach for solving the full non-linear Poisson–Boltzmann equation has been described in [290]. This method allows quite precise calculation of the electrostatics of a protein within a sufficient short time for its integration into a molecular force field.

The protonation state of amino acids and of proteins as a whole are another area of interest for the application of electrostatic theory to macromolecules. Here, two important key publications have appeared recently: a theory on the calculation of pK_a in proteins [291] and on the pH dependence of the stability of proteins [292]. With respect to the pK_a calculation, the finite difference Poisson–Boltzmann method was extended to include the complete charge distribution of both the neutral and charged forms of each ionizable group in the protein. This is discussed to be mostly important for the correct treatment of salt bridges. Application of the method to BPTI and serine proteases demonstrates that the results are reliable within a certain limit; however, more experimental data seem to be necessary for the verification and refinement of the method. With respect to the pH dependence of protein stability [292], a new method has been developed for the prediction of the pH dependence of the denaturation free energy of a protein. Calculations include a statistical mechanical treatment similar to the determination of pK_a as discussed before. The overall shape of experimentally observed denaturation free energies as a function of pH are similar to the calculations made. One of the conclusions of this work is that due to desolvation effects, ionizable groups generally tend to a destabilisation of the native state, although this effect

is discussed to be highly dependent on the actual pH. On the other hand, there are strongly stabilising pairwise ionic interactions at the surface of the proteins. The free energies of stabilisation in proteins is thus – again – a complicated balance of stabilising and destabilising forces in the system of protein and solvent, a result which is not surprising.

Other applications derived from calculation of electrostatic interactions in proteins include the calculation of molecular polarisation maps as an indicator for chemical reactivity [213], the prediction of protein backbone coordinates with the help of peptide dipole alignments [293], and the deduction of the conformational free energies of loop segments in protein structures [294]. The electrostatics of the active site of proteins is of special interest if one wants to understand the principles governing functional properties of these enzymes. Two proteins should be mentioned here: the cysteine protease papain, which was examined mainly by site-directed mutagenesis [295] and the secretory protein phospholipase A2 where known X-ray structures were subject to electrostatic calculations that explain their electrostatic asymmetry [296].

13. Knowledge-based methods

Knowledge-based methods are the most reliable approach for the prediction of protein structures at the moment. This reflects the lack of a unique algorithm for the folding problem, on one hand, and the exorbitant increase in structural information by X-ray and NMR methods, on the other. Recent reviews in this area were published in [6,297,298]. Also, even the methods for the prediction of protein structures by knowledge-based methods were subject to a systematic, comparative analysis [299]. Here, a network is generated that represents a prototype of a knowledge-based system which in turn simulates the processes used in protein structure prediction.

Aside from efforts to predict simple protein conformations from backbone coordinates [166,300,301], the prediction of the side chain conformations by rotamer libraries has also been examined [167,302,303]. The definition and quantification of substitution tables are of utmost importance for the comparative modeling approach. These tables must obviously be different for lipid-facing residues in α-helical transmembrane domains [304] and for globular proteins [305].

Another knowledge-based approach for the generation of full datasets from limited coordinates is termed "segment match modeling" [165] (cf. Section 7), and is again based on a dataset of known structures, the amino acid sequence and at least a Cα atom backbone (real coordinates) of the unknown protein. Short segments of the sequence are fitted onto this target framework. The selection of segments from the fragment database follows three different criteria: amino acid similarity, conformational similarity, and Van der Waals interactions. The method is quite successful: for eight test proteins of various size, the r.m.s. deviation of the modeled structures is between 0.93 Å and 1.73 Å. The approach works fast, reliable, and completely automatic. This method may therefore be a useful extension to the modeling approaches discussed above that generate only a Cα trace of the protein under investigation.

Comparative analyses of three-dimensional structures of proteins provide useful rules for protein structure prediction [171]. However, no general rules concerning the quality and applicability of concepts and procedures used in homology modeling have been put forward yet. To achieve this, a large set of known structures from different conformational and functional classes, but various degrees of homology, was analyzed carefully [138]. Pairwise structure superpositions were calculated; it was shown that both the topological differences of the protein backbones and the relative positions of corresponding side chains diverge with decreasing sequence identity. Below 50% identity, the deviation in regions that are structurally not conserved continually increases, thus implying that with decreasing sequence identity modeling has to consider more and more structurally diverging loop regions that are more difficult to predict. A further study in this respect was performed on the level of singular residues, especially on structural constraints determining protein fold families and functional classes, by a comparative analysis of families of homologous globular proteins [306]. Here, the classification of the residues in the proteins showed that there are distinct patterns of substitution of key residues within classes of related enzymes,

mainly – and surprisingly – often in the case when residues are at the same time solvent inaccessible and hydrogen bonded.

A method to evaluate the structural similarity of proteins has been developed in [307]. Homologous protein fragments were extracted and a pairwise, stepwise optimal superposition was calculated. With this procedure, a systematic search for structural similarities of proteins was performed, resulting in the dissection of the complete protein database into 182 structural families. This conclusion is in qualitative agreement with other work. A different method was developed in order to compare protein structures and to combine them into a multiple structure consensus [308]. The algorithm is a fusion of the structural comparison program SSAP and the multiple sequence alignment program MULTAL. The progressive combination of coordinate sets implies a hierarchical "condensation" of the structures; the visual inspection of the coordinate superpositions demonstrated that the method was able to identify a reasonable core of amino acids for each structural family under consideration.

Fully automatic procedures for the generation of structures from sequences by knowledge-based methods are useful, especially for novices in the area of structure prediction. A new method uses rules that translate multiple sequence alignments into distance constraints for subsequent distance geometry calculations [309]. The success of the approach is demonstrated by correct prediction of structures of several trypsin inhibitors.

An interesting solution for matching globular protein sequences to the correct structural templates known from the protein database has been presented in [137]. In contrast to the traditional approach where the sequence homology between the protein under investigation and proteins from the Brookhaven database whose three-dimensional structure are known is calculated, a structural "fingerprint" library is used. This fingerprint is calculated for each fold based on the contact map of amino acids and the buried vs. solvent-exposed pattern of residues. The method is able to correctly identify the structural class for proteins having only weakly or unrelated sequences, as has been demonstrated by several examples. Another sequence-independent method uses a distance measure between the Cα-coordinates of structurally similar proteins; this allows the alternative (sequence-independent) deduction of phyletic classification between proteins [310].

A comparable approach to fold recognition where sequences are modeled onto the backbone coordinates of known protein structures has been published in [311]. The method includes automatic modeling of structures in a full three-dimensional space, based on a predefined sequence. The "correctness" of each model structure is evaluated by empirical potentials. This approach is also quite similar to the algorithm developed earlier by Dr. Manfred Sippl and his coworkers [298]. In order to improve the quality of these mean potential approaches, several types of statistical potentials (backbone dihedral angle, distance-dependent pairwise interactions, accessible surface area) were calculated from known protein structures, and the results were compared [312]. Here, residue interaction potentials from distances between sidechain centroids perform significantly better than those computed from inter-Cα or inter-Cβ distances, and also the combination of several potentials is advantageous. Interestingly, potentials derived from backbone dihedral angles recognise 68 protein structures from a total of 74. It was demonstrated that this highly encouraging result is not an artefact based on the usage of a limited dataset of coordinates.

Using structural information from five homologous members of the helical cytokine family, two alternative models of interleukin (IL)-13 are proposed in [313]. IL-13 has biological properties similar to those of IL-4 and, like the other interleukins, is a potentially important pharmaceutical target. Structures of the receptors for the four-helix bundle cytokines IL-2 and IL-4 receptors based on the structure of the complex of human growth hormone with its receptor were calculated [314]. The models offer structural explanations for the effects of mutation of the A- and D-helices of the cytokines. In addition, they may be of use in the identification of residues which interact in the ligand–receptor interfaces. Other examples for the structure prediction by comparative modeling are the *E. coli* tyrosine aminotransferase derived from the homologous aspartate aminotransferase [315], human glutathione S-transferase [316], and small α-helical proteins [65].

The hypothesis that five of the six hypervariable

regions (CDRs) in antibodies may adopt only a small subset of possible conformations ("canonical ensembles") has now been demonstrated by the successful prediction of structures of hypervariable regions in several antibodies [24]. Particular key residues in these regions are responsible for the conservation of the loop conformation, as has been shown by a comparative analysis of structures.

Homology modeling has also led to a new hypothesis on the origin of halophilic adaptation, i.e. the unusual denaturation of proteins from halophilic organisms at relatively low salt concentrations [317]. A model structure for dihydrofolate reductase (h-DHFR) from *Haloferax volcanii* derived from comparative modeling to the respective *E. coli* enzyme shows an unique asymmetrical charge distribution over the protein surface, with positively charged amino acids centered around the active site, and negative charges on the opposite side of the enzyme. This particular charge distribution seem to be functionally relevant. The negative charges on the surface form clusters which may be shielded at high salt concentrations; at low salt, however, they repulse each other, thus destabilising the protein. The results presented are in accordance with experimental data and, thus, provide an explanation for the exceptional stability properties of h-DHFR.

14. Verification and reliability of predictions

The verification of protein structural models is becoming more and more important; new methods that emerge from the scientific community should be checked with an independent test system for their general applicability and reliability. However, there is no standard protocol for this purpose at the moment [138]. A convincing method for structural verification that has widely been used during the last few years has been developed by Dr. Manfred Sippl and coworkers with their publicly available program PROSA II (see [298], and references cited therein).

One of the most successful methods for the discrimination between correctly and incorrectly folded protein model structures utilizes three-dimensional profiles indicating local environments (characterised by polarity and solvent accessibility) around residues [318–320], with tables indicating preferred physicochemical properties of these environments. It also allows the determination of sequences belonging to a structural class, despite their low sequence similarity to the members of this class, by profile similarity assessment; here, modified profile algorithms based on statistical neighbor preferences are used [161].

An alternative method for the discrimination of the native fold from thousands of alternative structures is based on a function describing interresidue contacts [321]. For 37 training proteins from the database, 10000 alternative structures were generated. In all cases of globular folds, the native structure was discriminated from all alternatives by significant margins; the method also allows discrimination of protein folds that were not used in the training set and thus demonstrates to be competent for generalisation. On the other hand, closely related structures, e.g. liganded crystal structures compared with non-liganded ones or proteins differing at one position by site-directed mutagenesis, have comparable function values but are still discriminated from non-native structures. The method convincingly allows the decision to be made whether a native-like fold has been generated during the course of a modeling process.

Another diagnostic tool for the assessment of the quality of modeled structures is based on atomic solvation preference analysis [202]. Clear discrimination between deliberately misfolded and native protein structures was observed. The results are, however, only feasible when complete structures are examined; the correct fold of parts of the molecule, substructures or domains, is more difficult to prove. Empirical solvation models were applied in the case of trypsin inhibitor structure; 39 alternative conformations that are more or less close to the native fold could be correlated to a scale derived mainly from NMR coupling constants, with a concordance of 0.939 [201]. In any case, there is a linear relationship between the calculated solvation free energy and the size of the protein [115,322] that allows the prediction and quality analysis of unknown proteins and model structures. Solvation parameters alone, however, should not be used as the only criterion for the reliability assessment of models, since incorrectly folded molecules may have similar values to native proteins in some cases.

The correlation between the resolution of X-ray

structures and clustering of backbone angles in the Ramachandran diagram is another observation that may aid in structure verification [323]; it suggests measures of stereochemical quality of protein structures. A contact quality index calculated from known highly resolved protein structures may be used as an alternative, independent measure of reliability in protein crystallography; results demonstrate that there is a correlation between the contact quality index (being a measure for correct packing) and the resolution or R-factor of the structure [324].

An elegant way for the examination of the correctness of a model structure by experiments has been proposed in [325]. Based on sequence data and localisation of possible disulfide bonds, a structure model for erythropoietin has been generated that consists mainly of a four-helix bundle with three interconnecting loops. This motif is a reasonable choice since it is common within the cytokine family. Site-directed mutants – mainly deletion mutants – were then generated and investigated in bioassays. Properties of all tested muteins were in accordance with the proposed model structure, thus the model may be termed reliable within the limits of the resolution of the method chosen.

Another approach for the verification of structural data uses a database of more than 300 statistical, geometrical, energetical, and surface-based properties computed for 23 proteins of known structure with high precision [115]. This database serves as a knowledge base of properties of real, highly resolved proteins structures and should allow the detection of anomalous properties in model structures and in non-trivial proteins like hyperthermophilic or halophilic proteins. The resulting tool will be complemented by a reliability index analysis based on the comparison with properties of (slightly) misfolded proteins.

15. Membrane proteins

The modeling of the structure of membrane proteins is a completely distinct task compared to the modeling of globular proteins. Therefore, strategies and concepts developed in one area are usually not transferable to the other. The main reason for the higher success of membrane protein structure prediction [7] is the constraint on possible topologies imposed by the lipid environment in biological membranes, and the targeting and insertion process of transmembrane segments with subsequent condensation into a compact folded state. The transmembrane segments are usually a stretch of highly hydrophobic residues flanked by charged or polar residues and are thus relatively simply to predict. Von Heijne [326] has successfully predicted the topology of proteins from bacterial inner membranes by a sophisticated hydrophobicity analysis with subsequent automatic generation of possible topologies. 23 of 24 inner membrane protein topologies, and 135 transmembrane segments (with only one wrong case) were predicted correctly by this method.

Even the de novo design of transmembrane α-helices [193,304] has been demonstrated to be successful. Most transmembrane seven-helix bundles such as opsins and G-protein coupled receptors are expected to be similar to the bacteriorhodopsin (BR) fold which is crystallographically resolved, and therefore used as a model for new algorithms [88,227,327]; the designed model for the β_2-adrenoreceptor has been shown to be physically plausible and in reasonable agreement with experimental data, thus providing a sound basis for additional experiments, which should be a primary goal of any modeling approach.

Protein sequence divergence was used to predict the structure of the transmembrane domain of seven-helix membrane proteins [227]. The key procedure is the calculation of a hydrophilic and lipophilic variability index for each amino acid in an alignment of a family of homologous proteins. This variability profile was applied to bacteriorhodopsin and the resulting model was compared with the known X-ray structure. Some features of the known BR structure were also found in the model structure, thus confirming the possible value of the method.

Artificial neural network models are also used for membrane protein structure prediction [88,327]. A neural net predicts putative transmembrane sequences and is usually also able to classify membrane/non-membrane transition regions in sequences of integral human membrane proteins with high accuracy [327].

The superfamily of bacterial porins [328,329] is of special interest since they are responsible for the

selective sieve properties of the outer membranes of gram-negative bacteria. Crystallographic studies have revealed a symmetrical trimeric association of molecules in the membrane with water-filled channels. A remarkable property of these proteins is their high content of β-strands; for the prediction of these β-sheet proteins, approaches different from the above for α-helical proteins were developed. Alignment studies in conjunction with hydrophobicity analysis suggest that porins usually possess 16 transmembrane strands, which is in agreement with three (non-related) crystal structures of porins. The protein maltoporin (LamB) has been investigated in terms of its general fold [330] and has been shown to be compatible with the usual 16-strand model of porins, despite its low sequence homology and significantly larger size.

A new method for the prediction of topology and secondary structure of membrane proteins based on the recognition of topological models has been described recently [331]. The algorithm utilizes statistical data derived from 83 well-known membrane protein structures and a dynamic programming method to describe membrane topology. The calculated tables indicate biases for certain amino acid types towards inside, middle, or outside of cellular membranes, and 64 out of 83 topologies were predicted successfully (77% correctness). An improvement of the method may be gained by multiple alignment methods. Additional enhancement for the prediction of membrane protein topologies results from the "positive inside" rule stating that positively charged amino acids occur more frequently at the cellular side of transmembrane proteins than on the extracytoplasmic side [332]. A combinatorial approach, connecting this information with other observations regarding the distribution of polar and charged residues across the membrane, has been demonstrated to be most valuable for the prediction of membrane proteins.

16. Protein engineering

The rational design of artificial proteins with purposely modified properties is a highly desirable goal in molecular bioinformatics and biotechnology; interesting aspects in this respect are (i) the improvement of the stability [333,334] or function [51,335–337] of native proteins, or the relationship between stability and function [338]; and (ii) the deliberate presentation of new properties, e.g. the implantation of an ion binding site in an existing protein [195,339]. With respect to the latter, thioredoxin from *E. coli* was taken as the target structure, and a copper-binding site which is crystallographically well known from copper-binding proteins was introduced into the buried hydrophobic core of thioredoxin (close to the surface) by four amino acid exchanges. However, in contrast to the design, copper did not bind in the predicted manner but used only two of the newly introduced residues of the coordination sphere, and two other residues that were not part of the design strategy. Mercury, however, seems to be coordinated in the way that was predicted.

Point mutations are generally used to investigate the structure and/or function of proteins. In a most elaborate work, a set of possible predictive rules for the effect of mutations on the structure, based on the comparison of crystal structures of point mutants and wild-types in a total of 83 cases, was derived [340]. Despite the simplicity of the rules, they describe well the conformational changes in 85% of all point mutant structures used in this work.

17. Correlation with experimental results

An important feature of most modeling approaches, aside from their reliability and quality, is the connection to (and verification by) experimental results [341]. There is not yet a standardized protocol for the "quality control" of molecular modeling methods; often, jack-knifing is used in knowledge-based approaches to verify the independence of the results on the basis data sets. Statistical methods for the prediction of secondary structural elements often quantify the success by analyzing correctly and wrongly predicted positions. A direct comparison of experimental results with predicted data is therefore a strong, if not the best, indicator for the quality of a method. This, however, requires that the protein is purified and available in high quantity; particular with respect to the immense efforts for the sequencing of complete genomes and expressed sequence tags (ESTs) performed by molecular biological techniques, this prerequisite is usually not fulfilled. On the other hand, combinatorial methods connecting

experimental methods with computer-aided predictions have gained particular attention. An excellent review on various aspects of the interpretation of biochemical data by computational techniques may be found in [19].

Circular dichroism spectra which are characteristic and unique for the secondary structure composition of proteins are often used for verification of structure prognoses; the deconvolution algorithms, however, are still under investigation [94,342]. A combination of statistical algorithms utilizing CD spectra for secondary structure prediction has been shown to improve the accuracy of prediction between 3.9 and 4.9% [151], on an average per-residue basis, for a three-state prediction. Ultraviolet circular dichroism (UVCD) and vibrational circular dichroism (VCD) spectra of 20 proteins were analyzed in terms of basis spectra, reflecting five independent secondary structure elements [152]. The advantages and disadvantages of different architectures of neural network approaches (a three-layer backpropagation network, and a hybrid self-organisation to backpropagation network) for the deconvolutions has been discussed [93].

A method for the deconvolution of the near-UV second-derivative spectra of proteins with respect to their aromatic amino acids, i.e. their tryptophan, tyrosine, and phenylalanine spectra, has been published [343]; here, the contributions of individual standard spectra of model solutions were fitted to the experimental protein spectrum. No correlation between spectral band positions and average solvent accessibility is observed for the three aromatic residues, thus suggesting a significant influence of other local effects on the spectra. This makes the near-UV spectroscopy a highly sensitive method for detection of subtle local changes on one hand, which, on the other hand, is still poorly understood in terms of mathematical tools for its interpretation.

Other methods utilize chemical shifts in ^{1}H NMR which are characteristic for secondary structure types and therefore allow a secondary structure assignment of the respective protein which is comparable to circular dichroism, Raman, and infrared spectroscopy [344], or molecular dynamics together with NMR data or crystallography [234,275,278]. A novel approach based on neural network methods for the depiction of secondary structure contents of proteins has been demonstrated to be highly accurate if the spectrum is extended to a wavelength of at least 180 nm [92], thus providing an additional tool for the verification of model structures of proteins. Precise methods which extend these efforts for the complementation of secondary structure and protein topology prediction by experimental data, however, are still to be developed.

The most exciting work at the interface between experimental and computational results was published recently by Lee and Levitt [345]. They were able to accurately predict the stability and activity behavior of 78 different triple-site sequence variants of the lambda repressor protein, in comparison with the wild-type protein. The calculated energies correlate well with the activities of the mutants determined experimentally, and even better with the thermal stability of these proteins. They could also (correctly) predict two mutants to be more stable than the wild-type. This work, if confirmed on other model systems such as T4 lysozyme which presents a rich source of data, could become a most useful tool for the design of proteins with altered stability and activity properties.

Interestingly, recent work presented by Van Gunsteren and coworkers [19,236] showed that free energy calculation methods which are state-of-the-art cannot accurately predict the stability differences for a point mutation in subtilisin (asparagine-218 to serine). They conclude that details of the computational procedure have a dominant influence on the results: These are mainly the complex description of the unfolded state of proteins, accurate long-range potentials, dielectric effects, and entropy considerations. This supports the view that the calculation of "small differences of large numbers", displayed as a complex interplay of compensating and enhancing factors, is a fundamental problem in the prediction of free energy values for proteins.

18. Hardware and software progress

This final section summarizes some of the approaches described above that have been published with respect to special hardware requirements or with new software developments. A list of computer programs related to molecular bioinformatics which may be of common interest for the scientific commu-

nity can be found in Table 3; the list is by no means complete. The programs listed there were published during the last five years, and are distributed by non-commercial groups; therefore they are mostly free of charge.

Although any brute force method for the calculation of protein structures ab initio was not even discussed in the past [21], recent work suggests that the computational complexity of the protein folding problem does not automatically rule out the existence of an efficient, effective, and precise algorithm as a "shortcut" solution; there may be methods to avoid exponential increase in computer time, taking into account specific properties of interactions in proteins [23]. Parallel to these considerations, the last decade has seen much progress in the assessment of computer hardware aspects and the development of special software for molecular modeling and prediction purposes. An overview of currently available computational sequence analysis databases and software tools may be found in [346]; aside from sophisticated programs, there are now a few new useful tools, i.e. for the display of molecular structures.

Parallel computing is expected to revolutionize molecular calculation techniques; most of the algorithms and methods used in structure prediction are dedicated for parallel computers. This reflects the highly parallel organisation and cooperativity of molecular events in nature; progress in this area has been reviewed recently [347]. Many parallel computing approaches have been tested in order to evaluate the long range interactions which is the most computationally expensive phase of molecular dynamics simulations. Aside from standard parallel processing facilities, a different approach uses networked workstations as a source of parallel processing, implemented by the machine-independent, parallel processing language LINDA, for the treatment of time-

Table 3
A list of some novel computer programs that are useful for visualization and/or biomolecular computing

Program name	Description	Reference
EYECHEM	Network-based, modular molecular visualisation toolkit within the Silicon Graphics-based "Iris Explorer" visualisation program	[349]
DRAWNA	Drawing of schematic views of nucleic acids; menu-driven	[350]
MOLSCRIPT	Drawing of schematic views of (mainly) proteins	[351]
n.n.	PC-based molecular visualisation by ray-tracing software	[352]
WINMGM	Molecular modeling program for Microsoft WINDOWS, with interactive molecule manipulation, energy computation, solvent-accessible surfaces, fast space-filling CPK images	[353]
n.n.	Computationally effective representation of surfaces by sparse critical points	[354]
n.n.	Molecular surface recognition by a computer vision-based technique	[355]
n.n.	Two-dimensional maps and the respective three-dimensional representation of a molecular surface	[356]
n.n.	Microsoft EXCEL worksheets for the prediction of structural characteristics in proteins; implements the Chou and Fasman algorithm, prediction of the protein structural class, sequential motifs, domain boundaries, loops, state of cysteines, hydropathy profiles, flexibility plot	[357]
SEQSEE	Program suite for molecular sequence analysis	[358]
PDBMOTIF	Identification of protein motifs; generates script for RASMOL; uses PROSITE pattern syntax	[359]
CAVEAT	De novo design of molecules	[49]
SPROUT	De novo design of molecules	[182]
GREEN	Interactive docking between ligand and a protein molecule, using the physical and chemical environment of the ligand-binding site; includes energy minimisation	[360]
GENSTAR	Automated de novo design of drugs	[181]
n.n.	Concomitant display of secondary structure predictions, multiple alignment, and motif and pattern searching in protein sequences	[361]
COMPOSER	Most commonly used tool for the modeling of proteins by homology	[362]
HERA	Program to draw schematic diagrams of protein structures, including hydrogen-bonding diagrams, helical wheels, and helical nets	[363]
SIRIUS	Several programs for the calculation and analysis of packing interactions (geometry) between segments in proteins	[231]
DALI	Superposition of protein structures in order to analyze their structural relationship	[364]

consuming molecular dynamics simulation [232,348]. The straightforward but effective algorithm was tested on both a shared memory parallel computer and on a network of high-performance Unix workstations. Especially with respect to the increasing number of networked, cheap workstations, this approach could help make computationally expensive, long lasting molecular dynamics simulations more feasible.

19. Conclusions

The past years have seen immense efforts in the structure prediction area, with a strong bias for the development of algorithms and computer programs that are used for practical applications (protein design, structure analysis). Knowledge based modeling techniques are still the best choice for structure prediction if reliable and reasonable model structures are needed within short time, and these methods continue to be investigated and refined. Also, there is a strong and increasing demand for world wide, high speed networking for information access, e.g. via the Internet. Publicly available computer programs, free of charge for the academic community, may speed up the spreading of new ideas and methodologies, and may also be useful for the assessment of the general applicability of these new approaches. However, there is still no unambiguous and faultless method for the prediction of the structure, function, and other properties of a protein or other biological macromolecule from its building blocks. The protein folding problem will thus continue to be an exciting, but often also discouraging, area of research.

Acknowledgements

The author is indebted to Prof. Dr. Rainer Jaenicke from the University in Regensburg for his continuing support. I also wish to thank the following colleagues for providing me with their most recent work and for sending manuscripts prior to publication: Profs. Fred E. Cohen, Gordon M. Crippen, William DeGrado, Russel F. Doolittle, David Eisenberg, J. Garnier, Barry Honig, Joel Janin, J.N. Jansonius, Martin Karplus, A. Kolinski, Michael Levitt, J.A. McCammon, Fred M. Richards, George D. Rose, Chris Sander, Harold A. Scheraga, Klaus Schulten, Jeffrey Skolnick, Michael J. Sternberg, Janet M. Thornton, Wilfried F. Van Gunsteren, Gunnar Von Heijne, Shoshana Wodak.

References

[1] R.C. van Schaik, H.J. Berendsen, A.E. Torda and W.F. van Gunsteren, J. Mol. Biol., 234 (1993) 751.
[2] G.M. Crippen and T.F. Havel, J. Chem. Inf. Comput. Sci., 30 (1990) 222.
[3] J. Garnier, Biochimie, 72 (1990) 513.
[4] J. Garnier, J.M. Levin, J.F. Gibrat and V. Biou, Biochem. Soc. Symp., 57 (1990) 11.
[5] G. Nemethy and H.A. Scheraga, FASEB J., 4 (1990) 3189.
[6] M.J. Sternberg and M.J. Zvelebil, Eur. J. Cancer, 26 (1990) 1163.
[7] G. von Heijne and C. Manoil, Protein Eng., 4 (1990) 109.
[8] J. Eisenfeld, S. Vajda, I. Sugar and C. DeLisi, Am. J. Physiol., 261 (1991) C376.
[9] F.M. Richards, Sci. Am., 264 (1991) 54.
[10] J. Garnier and J.M. Levin, Comput. Appl. Biosci., 7 (1991) 133.
[11] B. Rost, R. Schneider and C. Sander, Trends. Biochem. Sci., 18 (1993) 120.
[12] H.A. Scheraga, in K.B. Lipkowitz and D.B. Boyd (Eds.), Reviews in Computational Chemistry, Volume III, VCH Publishers, New York, 1992, p. 73.
[13] E. Sun and F.E. Cohen, Gene, 137 (1993) 127.
[14] S.A. Benner, T.F. Jenny, M.A. Cohen and G.H. Gonnet, Adv. Enzyme Regul., 34 (1994) 269.
[15] C.L. Verlinde, E.A. Merritt, F. Van den Akker, H. Kim, I. Feil, L.F. Delboni, S.C. Mande, S. Sarfaty, P.H. Petra and W.G. Hol, Protein Sci., 3 (1994) 1670.
[16] F. Eisenhaber, B. Persson and P. Argos, Crit. Rev. Biochem. Mol. Biol., 30 (1995) 1.
[17] S.H. Bryant and S.F. Altschul, Curr. Opin. Struct. Biol., 5 (1995) 236.
[18] G. von Heijne, Eur. J. Biochem., 199 (1991) 253.
[19] W.F. van Gunsteren and A.E. Mark, Eur. J. Biochem., 204 (1992) 947.
[20] T. Ito, T. Fukushige, J. Makino, T. Ebisuzaki, S.K. Okumura, D. Sugimoto, H. Miyagawa and K. Kitamura, Proteins, 20 (1994) 139.
[21] P. Argos and R. Abagyan, Comput. Chem., 18 (1994) 225.
[22] R.H. Lathrop, Protein Eng., 7 (1994) 1059.
[23] J.T. Ngo and J. Marks, Protein Eng., 5 (1992) 313.
[24] C. Chothia, A.M. Lesk, A. Tramontano, M. Levitt, S.J. Smith Gill, G. Air, S. Sheriff, E.A. Padlan, D. Davies, W.R. Tulip et al., Nature, 342 (1989) 877.
[25] J. Cherfils, S. Duquerroy and J. Janin, Proteins, 11 (1991) 271.
[26] J.M. Thornton, Ciba Found. Symp., 159 (1991) 55.
[27] J.F. Gibrat, J. Higo, V. Collura and J. Garnier, Immunomethods, 1 (1992) 107.
[28] T. Scherf, R. Hiller, F. Naider, M. Levitt and J. Anglister, Biochemistry, 31 (1992) 6884.

[29] X. de la Cruz, A.E. Mark, J. Tormo, I. Fita and W.F. van Gunsteren, J. Mol. Biol., 236 (1994) 1186.
[30] A. Di Nola, D. Roccatano and H.J. Berendsen, Proteins, 19 (1994) 174.
[31] B.C. Braden and R.J. Poljak, FASEB J., 9 (1995) 9.
[32] S. Lea and D. Stuart, FASEB J., 9 (1995) 87.
[33] V.A. Roberts and E.D. Getzoff, FASEB J., 9 (1995) 94.
[34] T. Niermann and K. Kirschner, Methods Enzymol., 202 (1991) 45.
[35] V.N. Viswanadhan, B. Denckla and J.N. Weinstein, Biochemistry, 30 (1991) 11164.
[36] S.R. Presnell, B.I. Cohen and F.E. Cohen, Biochemistry, 31 (1992) 983.
[37] M.J. Sternberg, Curr. Opin. Struct. Biol., 2 (1992) 237.
[38] J.M. Levin, S. Pascarella, P. Argos and J. Garnier, Protein Eng., 6 (1993) 849.
[39] C. Geourjon and G. Deleage, Protein Eng., 7 (1994) 157.
[40] B. Rost, C. Sander and R. Schneider, J. Mol. Biol., 235 (1994) 13.
[41] L.B. Ellis and R.P. Milius, Comput. Appl. Biosci., 10 (1994) 341.
[42] A. Sali, J.P. Overington, M.S. Johnson and T.L. Blundell, Trends' Biochem. Sci., 15 (1990) 235.
[43] R.D. King, S. Muggleton, R.A. Lewis and M.J. Sternberg, Proc. Natl. Acad. Sci. USA, 89 (1992) 11322.
[44] S.A. Gillmor and F.E. Cohen, Receptor, 3 (1993) 155.
[45] B. Lesyng and J.A. McCammon, Pharmacol. Ther., 60 (1993) 149.
[46] S. Coulton and I. Francois, Prog. Med. Chem., 31 (1994) 297.
[47] P.M. Dean, Bioessays, 16 (1994) 683.
[48] M. Gumbleton and W. Sneader, Clin. Pharmacokinet., 26 (1994) 161.
[49] G. Lauri and P.A. Bartlett, J. Comput. Aided Mol. Des., 8 (1994) 51.
[50] A.R. Leach and S.R. Kilvington, J. Comput. Aided Mol. Des., 8 (1994) 283.
[51] G.J. Moore, Trends Pharmacol. Sci., 15 (1994) 124.
[52] W.G. Richards, Q. J. Med., 87 (1994) 379.
[53] C.L. Verlinde and W.G. Hol, Structure, 2 (1994) 577.
[54] M. Perutz, Protein Sci., 3 (1994) 1629.
[55] B. Waszkowycz, D.E. Clark, D. Frenkel, J. Li, C.W. Murray, B. Robson and D.R. Westhead, J. Med. Chem., 37 (1994) 3994.
[56] P.J. Whittle and T.L. Blundell, Annu. Rev. Biophys. Biomol. Struct., 23 (1994) 349.
[57] S. Muggleton, R.D. King and M.J. Sternberg, Protein Eng., 5 (1992) 647.
[58] R.D. King and M.J. Sternberg, J. Mol. Biol., 216 (1990) 441.
[59] J. Selbig, F. Kaden and I. Koch, FEBS Lett., 297 (1992) 241.
[60] M.J. Sternberg, R.D. King, R.A. Lewis and S. Muggleton, Philos. Trans. R. Soc. London B, 344 (1994) 365.
[61] M.J. Sternberg, R.A. Lewis, R.D. King and S. Muggleton, Faraday Discuss., (1992) 269.
[62] S. Sun, Protein Sci., 2 (1993) 762.
[63] D.T. Jones, Protein Sci., 3 (1994) 567.
[64] A.C. May and M.S. Johnson, Protein Eng., 7 (1994) 475.
[65] J.U. Bowie and D. Eisenberg, Proc. Natl. Acad. Sci. USA, 91 (1994) 4436.
[66] R. Unger and J. Moult, J. Mol. Biol., 231 (1993) 75.
[67] T. Dandekar and P. Argos, Protein Eng., 5 (1992) 637.
[68] T. Dandekar and P. Argos, J. Mol. Biol., 236 (1994) 844.
[69] G. Jones, P. Willett and R.C. Glen, J. Mol. Biol., 245 (1995) 43.
[70] L.H. Holley and M. Karplus, Proc. Natl. Acad. Sci. USA, 86 (1989) 152.
[71] D.G. Kneller, F.E. Cohen and R. Langridge, J. Mol. Biol., 214 (1990) 171.
[72] L.H. Holley and M. Karplus, Methods Enzymol., 202 (1991) 204.
[73] M. Vieth and A. Kolinski, Acta Biochim. Pol., 38 (1991) 335.
[74] S. Hayward and J.F. Collins, Proteins, 14 (1992) 372.
[75] M. Vieth, A. Kolinski, J. Skolnick and A. Sikorski, Acta Biochim. Pol., 39 (1992) 369.
[76] X. Zhang, J.P. Mesirov and D.L. Waltz, J. Mol. Biol., 225 (1992) 1049.
[77] B. Rost and C. Sander, Protein Eng., 6 (1993) 831.
[78] B. Rost and C. Sander, Proc. Natl. Acad. Sci. USA, 90 (1993) 7558.
[79] B. Rost and C. Sander, J. Mol. Biol., 232 (1993) 584.
[80] F. Sasagawa and K. Tajima, Comput. Appl. Biosci., 9 (1993) 147.
[81] B. Rost and C. Sander, Proteins, 19 (1994) 55.
[82] B. Rost, C. Sander and R. Schneider, Comput. Appl. Biosci., 10 (1994) 53.
[83] J.D. Hirst and M.J. Sternberg, Protein Eng., 4 (1991) 615.
[84] J.D. Hirst and M.J. Sternberg, Biochemistry, 31 (1992) 7211.
[85] S.R. Presnell and F.E. Cohen, Annu. Rev. Biophys. Biomol. Struct., 22 (1993) 283.
[86] I. Mahadevan and I. Ghosh, Nucleic Acids Res., 22 (1994) 2158.
[87] T.M. Nair, S.S. Tambe and B.D. Kulkarni, FEBS Lett., 346 (1994) 273.
[88] G.W. Dombi and J. Lawrence, Protein Sci., 3 (1994) 557.
[89] N. Tolstrup, J. Toftgard, J. Engelbrecht and S. Brunak, J. Mol. Biol., 243 (1994) 816.
[90] C. Wu and S. Shivakumar, Nucleic Acids Res., 22 (1994) 4291.
[91] G. Schneider and P. Wrede, Biophys. J., 66 (1994) 335.
[92] G. Böhm, R. Muhr and R. Jaenicke, Protein Eng., 5 (1992) 191.
[93] B. Dalmas, G.J. Hunter and W.H. Bannister, Biochem. Mol. Biol. Int., 34 (1994) 17.
[94] N. Sreerama and R.W. Woody, J. Mol. Biol., 242 (1994) 497.
[95] B.J. Hare and J.H. Prestegard, J. Biomol. NMR, 4 (1994) 35.
[96] A.A. Rabow and H.A. Scheraga, J. Mol. Biol., 232 (1993) 1157.
[97] H. Kono and J. Doi, Proteins, 19 (1994) 244.

[98] I. Koch, F. Kaden and J. Selbig, Proteins, 12 (1992) 314.
[99] K.C. Chou, Biophys. Chem., 35 (1990) 1.
[100] S. Dong and D.B. Searls, Genomics, 23 (1994) 540.
[101] B.I. Cohen, S.R. Presnell and F.E. Cohen, Methods Enzymol., 202 (1991) 252.
[102] L. Holm and C. Sander, Proteins, 19 (1994) 165.
[103] M. Huysmans, J. Richelle and S.J. Wodak, Proteins, 11 (1991) 59.
[104] P. Aldhous, Science, 262 (1993) 502.
[105] J.R. Eccles and J.W. Saldanha, Comput. Methods Programs Biomed., 32 (1990) 115.
[106] G.W. Milne, M.C. Nicklaus, J.S. Driscoll, S. Wang and D. Zaharevitz, J. Chem. Inf. Comput. Sci., 34 (1994) 1219.
[107] L. Holm and C. Sander, Nucleic Acids Res., 22 (1994) 3600.
[108] C. Sander and R. Schneider, Nucleic Acids Res., 22 (1994) 3597.
[109] H. Suzuki, A.S. Kolaskar, S.L. Samuel, J. Otsuka and A. Tsugita, Protein Seq. Data. Anal., 4 (1991) 97.
[110] C. Sander and R. Schneider, Proteins, 9 (1991) 56.
[111] L. Holm, C. Ouzounis, C. Sander, G. Tuparev and G. Vriend, Protein Sci., 1 (1992) 1691.
[112] C.A. Orengo, T.P. Flores, W.R. Taylor and J.M. Thornton, Protein Eng., 6 (1993) 485.
[113] U. Hobohm, M. Scharf, R. Schneider and C. Sander, Protein Sci., 1 (1992) 409.
[114] D.F. Stickle, L.G. Presta, K.A. Dill and G.D. Rose, J. Mol. Biol., 226 (1992) 1143.
[115] G. Böhm and R. Jaenicke, Protein Sci., 1 (1992) 1269.
[116] J.M. Thornton and S.P. Gardner, Chem. Design. Automat. News, (1993) 18.
[117] R.M. Woodsmall and D.A. Benson, Bull. Med. Libr. Assoc., 81 (1993) 282.
[118] J.F. Aiton, Dis. Markers, 12 (1994) 3.
[119] C.J. DiGiorgio, C.A. Richert, E. Klatt and M.J. Becich, Semin. Diagn. Pathol., 11 (1994) 294.
[120] K. Heumann, D. George and H.W. Mewes, Comput. Appl. Biosci., 10 (1994) 519.
[121] J.A. Kruper, M.G. Lavenant, M.H. Maskay and T.M. Jones, Proc. Annu. Symp. Comput. Appl. Med. Care, (1994) 32.
[122] S.M. Powsner and N.K. Roderer, Bull. Med. Libr. Assoc., 82 (1994) 419.
[123] L.H. Nicoll, J. Nurs. Adm., 24 (1994) 9.
[124] L.H. Nicoll, J. Nurs. Adm., 24 (1994) 11.
[125] L.H. Nicoll, J. Nurs. Adm., 24 (1994) 15.
[126] F.J. Regennitter and J.E. Volz, Am. J. Orthod. Dentofacial. Orthop., 107 (1995) 339.
[127] R. Fuchs, Comput. Appl. Biosci., 10 (1994) 413.
[128] K.W. McEnery, Am. J. Roentgenol., 164 (1995) 469.
[129] S. Henikoff, Trends Biochem. Sci., 18 (1993) 267.
[130] N. Colloc'h and F.E. Cohen, J. Mol. Biol., 221 (1991) 603.
[131] E.T. Harper and G.D. Rose, Biochemistry, 32 (1993) 7605.
[132] A. Chakrabartty, T. Kortemme and R.L. Baldwin, Protein Sci., 3 (1994) 843.
[133] N. Panasik, Jr., E.S. Eberhardt, A.S. Edison, D.R. Powell and R.T. Raines, Int. J. Pept. Protein Res., 44 (1994) 262.
[134] M.W. MacArthur and J.M. Thornton, J. Mol. Biol., 218 (1991) 397.
[135] J.F. Gibrat, B. Robson and J. Garnier, Biochemistry, 30 (1991) 1578.
[136] M.J. Sternberg and S.A. Islam, Protein Eng., 4 (1990) 125.
[137] A. Godzik, A. Kolinski and J. Skolnick, J. Mol. Biol., 227 (1992) 227.
[138] M. Hilbert, G. Böhm and R. Jaenicke, Proteins, 17 (1993) 138.
[139] M.A. Saqi and M.J. Sternberg, J. Mol. Biol., 219 (1991) 727.
[140] M.A. Saqi, P.A. Bates and M.J. Sternberg, Protein Eng., 5 (1992) 305.
[141] C.S. Ring, D.G. Kneller, R. Langridge and F.E. Cohen, J. Mol. Biol., 224 (1992) 685.
[142] N. Thanki, J.M. Thornton and J.M. Goodfellow, Protein Eng., 3 (1990) 495.
[143] M.J. Rooman and S.J. Wodak, Proteins, 9 (1991) 69.
[144] M.J. Rooman and S.J. Wodak, Biochemistry, 31 (1992) 10239.
[145] J.M. Parker and R.S. Hodges, Pept. Res., 4 (1991) 347.
[146] G. Böhm and R. Jaenicke, Int. J. Pept. Protein Res., 43 (1994) 97.
[147] A. Lupas, A.J. Koster, J. Walz and W. Baumeister, FEBS Lett., 354 (1994) 45.
[148] R.J. Gilbert, J. Mol. Graph., 10 (1992) 112.
[149] S. Salzberg and S. Cost, J. Mol. Biol., 227 (1992) 371.
[150] T. Niermann and K. Kirschner, Protein Eng., 4 (1991) 359.
[151] E.A. Carrara, C. Gavotti, P. Catasti, F. Nozza, L.L. Berutti Bergotto and C.A. Nicolini, Arch. Biochem. Biophys., 294 (1992) 107.
[152] P. Pancoska and T.A. Keiderling, Biochemistry, 30 (1991) 6885.
[153] J.M. Parker and R.S. Hodges, Pept. Res., 4 (1991) 355.
[154] C.D. Livingstone and G.J. Barton, Int. J. Pept. Protein Res., 44 (1994) 239.
[155] R.C. Garratt, J.M. Thornton and W.R. Taylor, FEBS Lett., 280 (1991) 141.
[156] B. Robson and J. Garnier, Nature, 361 (1993) 506.
[157] J. Bajorath and A. Aruffo, Bioconjug. Chem., 5 (1994) 173.
[158] A. Liwo, S. Oldziej, J. Ciarkowski, G. Kupryszewski, M.R. Pincus, R.J. Wawak, S. Rackovsky and H.A. Scheraga, J. Protein Chem., 13 (1994) 375.
[159] M.H. Hao, M.R. Pincus, S. Rackovsky and H.A. Scheraga, Biochemistry, 32 (1993) 9614.
[160] S.D. Pickett, M.A. Saqi and M.J. Sternberg, J. Mol. Biol., 228 (1992) 170.
[161] M. Wilmanns and D. Eisenberg, Proc. Natl. Acad. Sci. USA, 90 (1993) 1379.
[162] L. Holm and C. Sander, J. Mol. Biol., 218 (1991) 183.
[163] I. Simon, L. Glasser and H.A. Scheraga, Proc. Natl. Acad. Sci. USA, 88 (1991) 3661.
[164] K. Fidelis, P.S. Stern, D. Bacon and J. Moult, Protein Eng., 7 (1994) 953.
[165] M. Levitt, J. Mol. Biol., 226 (1992) 507.
[166] P.W. Payne, Protein Sci., 2 (1993) 315.

[167] P. Koehl and M. Delarue, J. Mol. Biol., 239 (1994) 249.
[168] D.A. Hinds and M. Levitt, Proc. Natl. Acad. Sci. USA, 89 (1992) 2536.
[169] M.J. Rooman, J.P. Kocher and S.J. Wodak, J. Mol. Biol., 221 (1991) 961.
[170] C. Sander, Biochem. Soc. Symp., 57 (1990) 25.
[171] T.L. Blundell, Ciba Found. Symp., 161 (1991) 28.
[172] A. Godzik, A. Kolinski and J. Skolnick, J. Comput. Aided Mol. Des., 7 (1993) 397.
[173] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich and K.A. Dill, Proc. Natl. Acad. Sci. USA, 92 (1995) 325.
[174] J.V. White, I. Muchnik and T.F. Smith, Math. Biosci., 124 (1994) 149.
[175] C. Ouzounis, C. Sander, M. Scharf and R. Schneider, J. Mol. Biol., 232 (1993) 805.
[176] A. Godzik and J. Skolnick, Proc. Natl. Acad. Sci. USA, 89 (1992) 12098.
[177] C.S. Ring and F.E. Cohen, FASEB J., 7 (1993) 783.
[178] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas and H.S. Chan, Protein Sci., 4 (1995) 561.
[179] A. Kolinski, A. Godzik and J. Skolnick, J. Chem. Phys., 98 (1993) 7420.
[180] G.M. Crippen, Biochemistry, 30 (1991) 4232.
[181] S.H. Rotstein and M.A. Murcko, J. Comput. Aided Mol. Des., 7 (1993) 23.
[182] V.J. Gillet, W. Newell, P. Mata, G. Myatt, S. Sike, Z. Zsoldos and A.P. Johnson, J. Chem. Inf. Comput. Sci., 34 (1994) 207.
[183] S.J. Wodak, I. Lasters, F. Pio and M. Claessens, Biochem. Soc. Symp., 57 (1990) 99.
[184] S.E. Brenner and A. Berry, Protein Sci., 3 (1994) 1871.
[185] G. Schneider, T. Todt and P. Wrede, Comput. Appl. Biosci., 10 (1994) 75.
[186] T. Tanaka, H. Kimura, M. Hayashi, Y. Fujiyoshi, K. Fukuhara and H. Nakamura, Protein Sci., 3 (1994) 419.
[187] D.E. Robertson, R.S. Farid, C.C. Moser, J.L. Urbauer, S.E. Mulholland, R. Pidikiti, J.D. Lear, A.J. Wand, W.F. DeGrado and P.L. Dutton, Nature, 368 (1994) 425.
[188] H. Morii, S. Honda, S. Ohashi and H. Uedaira, Biopolymers, 34 (1994) 481.
[189] B. Lovejoy, S. Choe, D. Cascio, D.K. McRorie, W.F. DeGrado and D. Eisenberg, Science, 259 (1993) 1288.
[190] A. Rey and J. Skolnick, Proteins, 16 (1993) 8.
[191] C. Sander, G. Vriend, F. Bazan, A. Horovitz, H. Nakamura, L. Ribas, A.V. Finkelstein, A. Lockhart, R. Merkl, L.J. Perry et al., Proteins, 12 (1992) 105.
[192] T.M. Handel, S.A. Williams and W.F. DeGrado, Science, 261 (1993) 879.
[193] P. Cronet, C. Sander and G. Vriend, Protein Eng., 6 (1993) 59.
[194] S. Sudarsanam, G.D. Virca, C.J. March and S. Srinivasan, J. Comput. Aided Mol. Des., 6 (1992) 223.
[195] H.W. Hellinga and F.M. Richards, J. Mol. Biol., 222 (1991) 763.
[196] V.V. Chemeris, D.A. Dolgikh, A.N. Fedorov, A.V. Finkelstein, M.P. Kirpichnikov, V.N. Uversky and O.B. Ptitsyn, Protein Eng., 7 (1994) 1041.
[197] B.S. Duncan and A.J. Olson, Biopolymers, 33 (1993) 219.
[198] J. Gasteiger, X. Li and A. Uschold, J. Mol. Graph., 12 (1994) 90.
[199] G. Böhm, Chaos, Solitons, Fractals, 1 (1991) 375.
[200] J. Janin and C. Chothia, J. Biol. Chem., 265 (1990) 16027.
[201] J. Vila, R.L. Williams, M. Vasquez and H.A. Scheraga, Proteins, 10 (1991) 199.
[202] L. Holm and C. Sander, J. Mol. Biol., 225 (1992) 93.
[203] J. Cherfils and J. Janin, Curr. Opin. Struct. Biol., 3 (1993) 265.
[204] M. Helmer Citterich and A. Tramontano, J. Mol. Biol., 235 (1994) 1021.
[205] M.Y. Mizutani, N. Tomioka and A. Itai, J. Mol. Biol., 243 (1994) 310.
[206] R. Norel, S.L. Lin, H.J. Wolfson and R. Nussinov, Biopolymers, 34 (1994) 933.
[207] A.R. Leach, J. Mol. Biol., 235 (1994) 345.
[208] R.M. Knegtel, J. Antoon, C. Rullmann, R. Boelens and R. Kaptein, J. Mol. Biol., 235 (1994) 318.
[209] R.M. Knegtel, R. Boelens and R. Kaptein, Protein Eng., 7 (1994) 761.
[210] A. Imberty and S. Perez, Glycobiology, 4 (1994) 351.
[211] M. Zacharias, B.A. Luty, M.E. Davis and J.A. McCammon, J. Mol. Biol., 238 (1994) 455.
[212] J. Aqvist, C. Medina and J.E. Samuelsson, Protein Eng., 7 (1994) 385.
[213] I. Alkorta, J.J. Perez and H.O. Villar, J. Mol. Graph., 12 (1994) 3.
[214] J. Ding, A. Jacobo Molina, C. Tantillo, X. Lu, R.G. Nanni and E. Arnold, J. Mol. Recognit., 7 (1994) 157.
[215] P.H. Walls and M.J. Sternberg, J. Mol. Biol., 228 (1992) 277.
[216] S. Duquerroy, J. Cherfils and J. Janin, Ciba Found. Symp., 161 (1991) 237.
[217] V. Lounnas, B.M. Pettitt and G.N. Phillips, Jr. Biophys. J., 66 (1994) 601.
[218] Y. Kita, T. Arakawa, T.Y. Lin and S.N. Timasheff, Biochemistry, 33 (1994) 15178.
[219] F. Colonna Cesari and C. Sander, Biophys. J., 57 (1990) 1103.
[220] J.W. Ponder and F.M. Richards, J. Mol. Biol., 193 (1987) 775.
[221] J.W. Ponder and F.M. Richards, Cold Spring Harbor Symp. Quant. Biol., LII (1987) 421.
[222] M.J. Behe, E.E. Lattman and G.D. Rose, Proc. Natl. Acad. Sci. USA, 88 (1991) 4195.
[223] G.D. Rose and R. Wolfenden, Annu. Rev. Biophys. Biomol. Struct., 22 (1993) 381.
[224] A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky and H.A. Scheraga, Protein Sci., 2 (1993) 1715.
[225] N.G. Hunt, L.M. Gregoret and F.E. Cohen, J. Mol. Biol., 241 (1994) 214.
[226] R. Varadarajan, F.M. Richards and P.R. Connelly, Curr. Sci., 59 (1990) 819.

[227] P. Du and I. Alkorta, Protein Eng., 7 (1994) 1221.
[228] T. Tanaka, Y. Kuroda, H. Kimura, S. Kidokoro and H. Nakamura, Protein Eng., 7 (1994) 969.
[229] L.M. Gregoret and F.E. Cohen, J. Mol. Biol., 219 (1991) 109.
[230] L.M. Gregoret and F.E. Cohen, J. Mol. Biol., 211 (1990) 959.
[231] J. Singh and J.M. Thornton, J. Mol. Biol., 211 (1990) 595.
[232] M.A. Shifman, A. Windemuth, K. Schulten and P.L. Miller, Comput. Biomed. Res., 25 (1992) 168.
[233] R.E. Bruccoleri and M. Karplus, Biopolymers, 29 (1990) 1847.
[234] J. de Vlieg and W.F. van Gunsteren, Methods Enzymol., 202 (1991) 268.
[235] R.M. Brunne and W.F. van Gunsteren, FEBS Lett., 323 (1993) 215.
[236] S. Yun yu, A.E. Mark, W. Cun xin, H. Fuhua, H.J. Berendsen and W.F. van Gunsteren, Protein Eng., 6 (1993) 289.
[237] S.H. Bryant and C.E. Lawrence, Proteins, 16 (1993) 92.
[238] T.A. Halgren, Curr. Opin. Struct. Biol., 5 (1995) 205.
[239] C.L. Brooks, Curr. Opin. Struct. Biol., 5 (1995) 211.
[240] M. Zuker, Methods Mol. Biol., 25 (1994) 267.
[241] J. Miller, K. Miaskiewicz and R. Osman, Ann. N. Y. Acad. Sci., 726 (1994) 71.
[242] H. Haller, M. Schaefer and K. Schulten, J. Phys. Chem., 97 (1993) 8343.
[243] J. Lautz, H. Kessler, W.F. van Gunsteren, H.P. Weber and R.M. Wenger, Biopolymers, 29 (1990) 1669.
[244] R.C. van Schaik, W.F. van Gunsteren and H.J. Berendsen, J. Comput. Aided Mol. Des., 6 (1992) 97.
[245] V. Daggett and M. Levitt, Proc. Natl. Acad. Sci. USA, 89 (1992) 5142.
[246] V. Daggett and M. Levitt, J. Mol. Biol., 223 (1992) 1121.
[247] A.E. Mark and W.F. van Gunsteren, Biochemistry, 31 (1992) 7745.
[248] V. Daggett and M. Levitt, J. Mol. Biol., 232 (1993) 600.
[249] A. Caflisch and M. Karplus, Proc. Natl. Acad. Sci. USA, 91 (1994) 1746.
[250] A. Godzik, J. Skolnick and A. Kolinski, Proc. Natl. Acad. Sci. USA, 89 (1992) 2629.
[251] M.E. Snow and G.M. Crippen, Int. J. Pept. Protein Res., 38 (1991) 161.
[252] G.M. Crippen and M.E. Snow, Biopolymers, 29 (1990) 1479.
[253] L. Holm and C. Sander, Proteins, 14 (1992) 213.
[254] V.N. Maiorov and G.M. Crippen, Proteins, 20 (1994) 167.
[255] V. Collura, J. Higo and J. Garnier, Protein Sci., 2 (1993) 1502.
[256] T. Ichiye and M. Karplus, Proteins, 11 (1991) 205.
[257] C.A. Laughton, Protein Eng., 7 (1994) 235.
[258] J. Higo, V. Collura and J. Garnier, Biopolymers, 32 (1992) 33.
[259] J.J. Tanner, L.J. Nell and J.A. McCammon, Biopolymers, 32 (1992) 23.
[260] M. Levitt, J. Mol. Biol., 220 (1991) 1.
[261] K. Kuczera, J. Kuriyan and M. Karplus, J. Mol. Biol., 213 (1990) 351.
[262] R. Abagyan and M. Totrov, J. Mol. Biol., 235 (1994) 983.
[263] L. Wesson and D. Eisenberg, Protein Sci., 1 (1992) 227.
[264] M. Norin, F. Haeffner, K. Hult and O. Edholm, Biophys. J., 67 (1994) 548.
[265] J. Schlitter, M. Engels and P. Kruger, J. Mol. Graph., 12 (1994) 84.
[266] T. Simonson and D. Perahia, Proc. Natl. Acad. Sci. USA, 92 (1995) 1082.
[267] L.M. Ptaszek, S. Vijayakumar, G. Ravishanker and D.L. Beveridge, Biopolymers, 34 (1994) 1145.
[268] F. Zhou, A. Windemuth and K. Schulten, Biochemistry, 32 (1993) 2291.
[269] D. Rognan, L. Scapozza, G. Folkers and A. Daser, Biochemistry, 33 (1994) 11476.
[270] B. Roux and M. Karplus, Annu. Rev. Biophys. Biomol. Struct., 23 (1994) 731.
[271] Y. Okamoto, Biopolymers, 34 (1994) 529.
[272] A.P. Heiner, H.J. Berendsen and W.F. van Gunsteren, Protein Eng., 6 (1993) 397.
[273] P.R. Gerber, A.E. Mark and W.F. van Gunsteren, J. Comput. Aided Mol. Des., 7 (1993) 305.
[274] M.A. McCarrick and P. Kollman, Methods Enzymol., 241 (1994) 370.
[275] J. Smith, K. Kuczera and M. Karplus, Proc. Natl. Acad. Sci. USA, 87 (1990) 1601.
[276] K. Kuczera, J.C. Lambry, J.L. Martin and M. Karplus, Proc. Natl. Acad. Sci. USA, 90 (1993) 5805.
[277] G.R. Kneller and J.C. Smith, J. Mol. Biol., 242 (1994) 181.
[278] P. Gros, W.F. van Gunsteren and W.G. Hol, Science, 249 (1990) 1149.
[279] J. Kuriyan, K. Osapay, S.K. Burley, A.T. Brunger, W.A. Hendrickson and M. Karplus, Proteins, 10 (1991) 340.
[280] R.C. Wade, M.H. Mazor, J.A. McCammon and F.A. Quiocho, Biopolymers, 31 (1991) 919.
[281] K.A. Sharp and B. Honig, Annu. Rev. Biophys. Biophys. Chem., 19 (1990) 301.
[282] M.K. Gilson, Curr. Opin. Struct. Biol., 5 (1995) 216.
[283] A. Nicholls, K.A. Sharp and B. Honig, Proteins, 11 (1991) 281.
[284] K.A. Sharp, A. Nicholls, R. Friedman and B. Honig, Biochemistry, 30 (1991) 9686.
[285] J.S. Albert and A.D. Hamilton, Biochemistry, 34 (1995) 984.
[286] M. Prevost, S.J. Wodak, B. Tidor and M. Karplus, Proc. Natl. Acad. Sci. USA, 88 (1991) 10880.
[287] V.Z. Spassov, A.D. Karshikoff and R. Ladenstein, Protein Sci., 3 (1994) 1556.
[288] R.M. Brunne, E. Liepinsh, G. Otting, K. Wuthrich and W.F. van Gunsteren, J. Mol. Biol., 231 (1993) 1040.
[289] M.K. Gilson and B. Honig, J. Comput. Aided Mol. Des., 5 (1991) 5.
[290] M. Holst, R.E. Kozack, F. Saied and S. Subramaniam, J. Biomol. Struct. Dyn., 11 (1994) 1437.

[291] A.S. Yang, M.R. Gunner, R. Sampogna, K. Sharp and B. Honig, Proteins, 15 (1993) 252.
[292] A.S. Yang and B. Honig, J. Mol. Biol., 231 (1993) 459.
[293] A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky and H.A. Scheraga, Protein Sci., 2 (1993) 1697.
[294] K.C. Smith and B. Honig, Proteins, 18 (1994) 119.
[295] R. Menard, C. Plouffe, P. Laflamme, T. Vernet, D.C. Tessier, D.Y. Thomas and A.C. Storer, Biochemistry, 34 (1995) 464.
[296] D.L. Scott, A.M. Mandel, P.B. Sigler and B. Honig, Biophys. J., 67 (1994) 493.
[297] A.C. May, M.S. Johnson, S.D. Rufino, H. Wako, Z.Y. Zhu, R. Sowdhamini, N. Srinivasan, M.A. Rodionov and T.L. Blundell, Philos. Trans. R. Soc. London B, 344 (1994) 373.
[298] M. Sippl, Curr. Opin. Struct. Biol., 5 (1995) 229.
[299] D.A. Clark, G.J. Barton and C.J. Rawlings, J. Mol. Graph., 8 (1990) 94.
[300] H.S. Kang, N.A. Kurochkina and B. Lee, J. Mol. Biol., 229 (1993) 448.
[301] S. Srinivasan, C.J. March and S. Sudarsanam, Protein Sci., 2 (1993) 277.
[302] R.L. Dunbrack, Jr. and M. Karplus, J. Mol. Biol., 230 (1993) 543.
[303] C. Wilson, L.M. Gregoret and D.A. Agard, J. Mol. Biol., 229 (1993) 996.
[304] D. Donnelly, J.P. Overington, S.V. Ruffle, J.H. Nugent and T.L. Blundell, Protein Sci., 2 (1993) 55.
[305] C.M. Topham, A. McLeod, F. Eisenmenger, J.P. Overington, M.S. Johnson and T.L. Blundell, J. Mol. Biol., 229 (1993) 194.
[306] J. Overington, M.S. Johnson, A. Sali and T.L. Blundell, Proc. R. Soc. London B, 241 (1990) 132.
[307] U. Lessel and D. Schomburg, Protein Eng., 7 (1994) 1175.
[308] W.R. Taylor, T.P. Flores and C.A. Orengo, Protein Sci., 3 (1994) 1858.
[309] T.F. Havel and M.E. Snow, J. Mol. Biol., 217 (1991) 1.
[310] M.S. Johnson, M.J. Sutcliffe and T.L. Blundell, J. Mol. Evol., 30 (1990) 43.
[311] D.T. Jones, W.R. Taylor and J.M. Thornton, Nature, 358 (1992) 86.
[312] J.P. Kocher, M.J. Rooman and S.J. Wodak, J. Mol. Biol., 235 (1994) 1598.
[313] P. Bamborough, D. Duncan and W.G. Richards, Protein Eng., 7 (1994) 1077.
[314] P. Bamborough, C.J. Hedgecock and W.G. Richards, Structure, 2 (1994) 839.
[315] J. Jager, T. Solmajer and J.N. Jansonius, FEBS Lett., 306 (1992) 234.
[316] R.E. Cachau, J.W. Erickson and H.O. Villar, Protein Eng., 7 (1994) 831.
[317] G. Böhm and R. Jaenicke, Protein Eng., 7 (1994) 213.
[318] J.U. Bowie, R. Luthy and D. Eisenberg, Science, 253 (1991) 164.
[319] D. Eisenberg, J.U. Bowie, R. Luthy and S. Choe, Faraday Discuss., (1992) 25.
[320] R. Luthy, J.U. Bowie and D. Eisenberg, Nature, 356 (1992) 83.
[321] V.N. Maiorov and G.M. Crippen, J. Mol. Biol., 227 (1992) 876.
[322] L. Chiche, L.M. Gregoret, F.E. Cohen and P.A. Kollman, Proc. Natl. Acad. Sci. USA, 87 (1990) 3240.
[323] A.L. Morris, M.W. MacArthur, E.G. Hutchinson and J.M. Thornton, Proteins, 12 (1992) 345.
[324] G. Vriend and C. Sander, J. Appl. Crystallogr., 26 (1993) 47.
[325] J.P. Boissel, W.R. Lee, S.R. Presnell, F.E. Cohen and H.F. Bunn, J. Biol. Chem., 268 (1993) 15983.
[326] G. von Heijne, J. Mol. Biol., 225 (1992) 487.
[327] R. Lohmann, G. Schneider, D. Behrens and P. Wrede, Protein Sci., 3 (1994) 1597.
[328] D. Jeanteur, J.H. Lakey and F. Pattus, Mol. Microbiol., 5 (1991) 2153.
[329] T. Ferenci, Mol. Microbiol., 14 (1994) 188.
[330] T. Schirmer and S.W. Cowan, Protein Sci., 2 (1993) 1361.
[331] D.T. Jones, W.R. Taylor and J.M. Thornton, Biochemistry, 33 (1994) 3038.
[332] L. Sipos and G. von Heijne, Eur. J. Biochem., 213 (1993) 1333.
[333] G.E. Dale, C. Broger, H. Langen, A. D'Arcy and D. Stuber, Protein Eng., 7 (1994) 933.
[334] C. Lee, J. Mol. Biol., 236 (1994) 918.
[335] S. Zinn Justin, L. Pillet, F. Ducancel, A. Thomas, J.C. Smith, J.C. Boulain and A. Menez, Protein Eng., 7 (1994) 917.
[336] R.S. Farid, D.E. Robertson, C.C. Moser, D. Pilloud, W.F. DeGrado and P.L. Dutton, Biochem. Soc. Trans., 22 (1994) 689.
[337] Y. Reiter, U. Brinkmann, K.O. Webber, S.H. Jung, B. Lee and I. Pastan, Protein Eng., 7 (1994) 697.
[338] B.K. Shoichet, W.A. Baase, R. Kuroki and B.W. Matthews, Proc. Natl. Acad. Sci. USA, 92 (1995) 452.
[339] H.W. Hellinga, J.P. Caradonna and F.M. Richards, J. Mol. Biol., 222 (1991) 787.
[340] V. De Filippis, C. Sander and G. Vriend, Protein Eng., 7 (1994) 1203.
[341] L. Jin, F.E. Cohen and J.A. Wells, Proc. Natl. Acad. Sci. USA, 91 (1994) 113.
[342] J.D. Hirst and C.L. Brooks, J. Mol. Biol., 243 (1994) 173.
[343] H. Mach and C.R. Middaugh, Anal. Biochem., 222 (1994) 323.
[344] D.S. Wishart, B.D. Sykes and F.M. Richards, FEBS Lett., 293 (1991) 72.
[345] C. Lee and M. Levitt, Nature, 352 (1991) 448.
[346] M.S. Boguski, J. Lipid Res., 33 (1992) 957.
[347] R.C. Brower and C. DeLisi, Crit. Rev. Biomed. Eng., 20 (1992) 373.
[348] M.A. Shifman, A. Windemuth, K. Schulten and P.L. Miller, Proc. Annu. Symp. Comput. Appl. Med. Care, (1991) 414.
[349] O. Casher, S.M. Green and H.S. Rzepa, J. Mol. Graph., 12 (1994) 226.
[350] C. Massire, C. Gaspin and E. Westhof, J. Mol. Graph., 12 (1994) 201.
[351] P. Kraulis, J. Appl. Crystallogr., 24 (1991) 946.
[352] A.S. Dion, J. Mol. Graph., 12 (1994) 41.

[353] M. Rahman and R. Brasseur, J. Mol. Graph., 12 (1994) 212.
[354] S.L. Lin, R. Nussinov, D. Fischer and H.J. Wolfson, Proteins, 18 (1994) 94.
[355] R. Norel, D. Fischer, H.J. Wolfson and R. Nussinov, Protein Eng., 7 (1994) 39.
[356] A. Badel Chagnon, J. Nessi, L. Buffat and S. Hazout, J. Mol. Graph., 12 (1994) 162.
[357] J. Clotet, J. Cedano and E. Querol, Comput. Appl. Biosci., 10 (1994) 495.
[358] D.S. Wishart, R.F. Boyko, L. Willard, F.M. Richards and B.D. Sykes, Comput. Appl. Biosci., 10 (1994) 121.
[359] M.A. Saqi and R. Sayle, Comput. Appl. Biosci., 10 (1994) 545.
[360] N. Tomioka and A. Itai, J. Comput. Aided Mol. Des., 8 (1994) 347.
[361] C. Geourjon and G. Deleage, Comput. Appl. Biosci., 9 (1993) 87.
[362] C.M. Topham, P. Thomas, J.P. Overington, M.S. Johnson, F. Eisenmenger and T.L. Blundell, Biochem. Soc. Symp., 57 (1990) 1.
[363] E.G. Hutchinson and J.M. Thornton, Proteins, 8 (1990) 203.
[364] L. Holm and C. Sander, J. Mol. Biol., 233 (1993) 123.