ELSEVIER

Biochimica et Biophysica Acta 1343 (1997) 1–15

BIOCHIMICA ET BIOPHYSICA ACTA
BBA

Review

Computational methods for the prediction of protein folds

Thomas Dandekar *, Rainer König

EMBL, Postfach 102209, D-69012 Heidelberg, Germany

Received 28 July 1997; accepted 7 August 1997

Contents

1. Introduction

Structure and function in proteins are closely related. Despite rapid growth of known protein se-

* Corresponding author. Fax: 49 6221 387 518; E-mail: dandekar@embl-heidelberg.de

PII S0167-4838(97)00132-5

quences, direct experimental determination of their structure by NMR or X-ray crystallography is still quite time consuming and often limited by the protein size (NMR) or the availability of crystals [1]. However, for most applications, such as protein design and pharmacology, and for a thorough understanding of the function, knowledge not just of the protein sequence but of the three dimensional protein structure is highly desirable. The much faster computer-based methods of predicting protein structure from sequence are thus a center of current research [2].

Ideally, such methods start from the sequence only (ab initio) and calculate from it the final fold. However, the conformational space accessible for even a medium sized peptide is enormous. This is often called the protein folding problem. Levinthal as early as 1968 contended that in real protein folding only a minimal subset of this space is searched [3]. A number of computer studies on protein folding have tried to delineate further important factors. Sali et al. [4] suggested that protein folds are selected by nature to have a good energetic separation between energy minimum and alternative accessible conformations, as otherwise such a protein structure is not sufficiently stable to be selected during evolution. Funnels or pathways may direct the folding and show an efficient way to prune out solutions from the otherwise fathomless conformational space. There is also evidence that natural protein sequences are again selected by nature to achieve sure and direct pathways and eliminate metastable states or many undefined conformations, which are found often in simulations with random sequences [5].

Computational methods for fold prediction have to find their own approach to reduce and efficiently search conformational space. A second complication is that the exact physicochemical potentials governing fold formation and separating native fold from wrongly folded are not yet fully understood [6]. Despite sound progress in the field [7], ab initio prediction is still a challenging task [8]. However, profiting from the growing number of known three dimensional structures, comparative modeling [9] and fold recognition [10] have prospered and are the methods of choice when a structure homologue can be successfully identified.

Our synopsis will first introduce different aspects of the protein folding problem, than review different strategies and finally discuss success, obstacles and new theoretical concepts.

2. The protein folding problem

2.1. The minimum of free energy

The native protein structure might be closely related to the minimum of free energy. This basic assumption is, for instance, supported by hydrogen exchange experiments, which show that the low energy native state is indeed found most of the time. However, such methods detect also partially unfolded states for shorter time periods [11]. The classical in vitro renaturation experiments [12,13] imply that all the information required to determine protein folding is contained in the amino acid sequence. Nevertheless, biological factors can be important to control the kinetics of protein folding and subunit assembly such as the peptidyl-prolyl-isomerase, disulfide isomerase and molecular chaperones [14,15].

Protein folding invokes the ‘multiple minima’ problem of finding the global energy minima among a vast multitude of alternative energy minima [7]. Similarly, this is also the challenge for computational approaches, particularly those more ambitious that start from sequence only (ab initio). Further features to consider are breathing of the protein [11] and chaperone action on unfolded proteins [15]. However, for most applications these complications can be ignored and the protein fold prediction problem can be reduced to find the lowest energy conformation. This ‘simplified’ prediction problem is nevertheless quite challenging, in computational mathematics it is classified as NP complete ([16], Box 1, Appendix A). Astonishingly, an alternative strategy to identify the correct fold called ‘threading’ is also NP complete. In threading the sequence with unknown structure is forced to adopt the conformation of each of the known structures in a protein structure database. The structure with the lowest strain after the sequence has been forced to adopt this fold or ‘threaded’ onto this structure, is taken as the predicted structure. For optimal results with this strategy (i) variable-length gaps have to be admitted in the alignment and (ii)

interactions between amino acids from the sequence have to be considered in the scoring function. This yields a similarly difficult computational problem [17].

2.2. The components of the energy function

To find the global energy minimum, a detailed energy calculation has to take the following considerations into account:

- Intramolecular energy (in vacuo) includes covalent terms (bond energy, strain, stretching, angle bending, planar and dihedral torsions) and non covalent terms such as van der Waals interactions, modeled for instance with the quantum mechanical Buckingham potential $E(r) = Ae^{-Br} - C/r^6 - D/r^8$ (r distance, A, B, C, D are fitted parameters) or the simplified (but artificial stiff) Lenard-Jones potential (substituting the exponential by A'/r^{12} and omitting the final term describing dipole–quadrupole interactions). Furthermore, hydrogen bond energy and the non-electrostatic component (Lippincott-Schroeder potential) are non-covalent terms. The conformational entropy (side and mainchain) is difficult to calculate. Exact coulomb electrostatics of atomic charges and dipoles are computationally intensive for the protein surface [18–20] and are usually simplified. The in vacuo energy components have been modeled in various atomic force fields such as ECEPP [21], GROMOS [22], CHARMM [23], AMBER [24], DISCOVER [25] and the rigid geometry model of Robson and Platt [26].
- Solvation energy with short- (cavity formation, solute–solvent dispersion, surface) and long-range components, volume related terms and dielectric media polarization [27,28].

Apart from the computational expense of an atomic level energy function, a serious problem is the accurate modeling of the potentials and solvent. The exact physico-chemical interactions are currently incompletely understood [6] and thus the accuracy of molecular mechanical potentials and solvent models is a controversial subject, including the electrostatic fields and the simulation of many atoms in the solvent in a suitable, often simplified way. Atomic level energy functions are a frontier of research where realistic results to present are only obtainable in carefully selected settings and protein examples.

2.3. Helpful simplifications

The rather complex energy function, and also the computational expensive exact geometric representation of the molecule and the search space itself have been simplified in a number of ways and approaches to increase computational speed.

Pairwise interactions are often reduced by cutoff distances [29,30]. Long-range interactions between distant atoms may be approximated by mean field approximations [31] and can partly compensate for the errors caused by cutoff [32]. Hydrogens can be treated as united atoms with their corresponding carbon [23], though one has to remember that the actual protein surface is almost exclusively formed by hydrogens [33]. Fixed covalent bonds, angles and the use of rotamers are also simplifications often used to gain computational speed [34]. Experimental constraints such as distance constraints may be used to bias the energy function and to restrict conformational search space, for instance [35]. Possibilities for reduced representation of the molecule are numerous, for instance reducing the peptide group [36], having a 2–3 point representation of the side chain [37] or having no side chain at all [38].

The energy function can be simplified to empirical pseudopotentials in many different ways, such as knowledge based potentials of mean force [39,40]. An application of simplified models is their use in generating low-resolution models from experimental information [35,41]. Such models are not only directly usable and testable by experimentalists, but at the same time, the limits (the coarseness in its resolution) and accuracy (number of satisfied distance constraints etc.) are apparent.

Lattice models reduce the search space drastically. A drawback is positional inaccuracy, e.g. in the coordination of helices with the rest of the protein. Godzik et al. suggest diamond, hybrid and ultra lattices as a good compromise between precise representation (1–2 Å) and computational costs [42]. However, these claims have not been verified by direct CPU (and RMSD) comparisons between off lattice models and the many neighbours to consider, such as in an ultra lattice model. On the other hand, even

very small ($3 \times 3 \times 3$) cubic lattices can offer important general insights into protein folding [4].

3. Overview on different strategies

3.1. Homology modelling on a known fold

Since the time of Brown et al. (1969) homology modeling has been the favored method of identifying the three-dimensional structure if the three-dimensional fold of a related protein is known [43,44]. Typical steps are:

1. To define structurally conserved regions (SCR), on the basis of 3D structural comparisons and/or multiple sequence alignments (see below) within the protein family, and correctly align the sequence to be modeled. This can be done by 3D superposition if both 3D structures are known to reveal homologies with low residue identities [45]. However, in real modeling a new sequence of unknown structure has to be aligned. Multiple alignments are a helpful tool for this, as they collect sequences from proteins with closely related function and align identical residues, while revealing sequence variation despite similar function in other positions. Correct multiple alignment is non-trivial. There are many alternatives, in particular some identities in one position can only be achieved by sacrificing identities in other places. Elaborate schemes and comparison matrices are required to evaluate different degrees of similarity in amino acids aligned at one position in alternative multiple alignments. Gaps have to be taken into account by dynamic programming [46] as do additional criteria such as solvent exposure and secondary structural assignments [47]. There are self-consistency tests for suboptimal alignments [48], and genetic algorithms can be used for stochastic optimization of the alignment without presupposing an initial alignment [49]. In more complicated cases, threading techniques may also be used.
2. Having the alignment, construct the SCR backbone of the homologous protein to be modeled [50].
3. Construct the loop regions and other structurally different regions (e.g. by best fitting loop regions from known 3D structures [51] or by conformational search [52] and subsequent energy refinement in most cases. Unfortunately, this problem is also computational demanding and NP complete [16]. Loop construction is difficult. Current loop construction methods do not work reliably. They often fail if the structures are not closely related [2].
4. Construct and place side-chains using either known structures (e.g. with similar local residue environment [53]), a rotamer library [54] or molecular mechanical placement [50].
5. Verify the structure, check the plausibility of the model.
6. Refine the structure by energy minimization or related techniques (molecular mechanics).

Helpful software packages for homology modeling tasks include Whatif [55], Insight (Biosym Technologies, San Diego) and Modeller [56] but require some experience for proper use and reliable results.

A well known application is the prediction of a structure after mutagenesis [57]. However, we have to alert the reader that mutation of critical points in the structure may sometimes cooperatively change the domain arrangement of the whole structure (for instance [58]). These limitations have to be taken into account when using homology modeling. If sequence identity between predicted structure and a known 3D structure is sufficiently high (above 30%) and the protein to be modeled has similar function as the template and occurs like this unmutated in nature, homology modeling is a reliable approach to create a first view of the arrangement of the secondary structural elements in the unknown structure and to compare different homologous proteins or to design experiments. Databases of sequences with homologous structures such as HSSP are available [59] as are rotamer libraries (e.g. [54]). Side chain placement can be further optimized using a genetic algorithm and exploiting the dead-end elimination theorem [60,61]. Recently molecular modeling of glycoproteins by homology with non-glycosylated protein domains and computer simulated glycosylation has also been tried [62].

3.2. Recognition methods for a tertiary fold

A number of methods try to recognize a tertiary fold by analyzing the sequence. Critical protein fea-

tures correlate sufficiently with amino acid or amino acid profile type at conserved positions in the protein to allow by this a recognition of the general fold topology, its protein family membership or its basic function. Though this approach is the opposite strategy to the detailed, full energy treatment outlined above, it has become surprisingly powerful. It has many applications including rapid characterization of proteins in large scale sequencing projects.

3.2.1. Pattern searches and neural networks

- Simple recognition is already possible using global and local [63] *sequence alignment*. Sequence identity above 25–30% implies structural similarity with a high probability if the alignment exceeds a threshold of about 70 residues [64,59].
- Alignment with *amino acid property profiles* is even more sensitive (e.g. [65]) in detecting homologies such as the pleckstrin homology domain [66]. Compilations of specific amino acid patterns or signatures point to protein families (PROSITE, [67]), and sets of motifs can be combined (PRINTS; [68]).
- *Neural networks* mimic computational pattern recognition by neurons in living organisms. Connections between the modeled neurons are not preprogrammed but instead selected for optimal performance in the recognition of a training set where the correct answers are all known. Subsequently the optimal trained network is used to detect as yet unrevealed patterns. Trained neural networks may even do better than profile search, for instance in revealing integrases, DNA-polymerases and immunoglobulins [69]. They allow good secondary structure and structural class prediction [70], belong to the most accurate secondary structure predictors [71], and are fully available and easy to use. An average of more than 70% accuracy (72%, [72]) may be approached exploiting multiple alignments (see above) and neural networks. However, depending on the protein example, even such secondary structure predictors are still surprisingly poor (sometimes many elements are mispredicted), and claims of 90% accuracy are only tenable if prediction concentrates on known regions within a selected set of proteins.

Further applications of neural networks include a novel neural network prediction technique based on consideration of both local and long-range interactions between amino acid residues, and has been used to detect a new member of the thioredoxin fold-containing family (DSBC protein) [73]. Hypervariable antibody loop structures were also recently predicted by using such networks [74].

- A particularly successful *pattern search* technique derives patterns of conserved physical properties of amino acids from multiple sequence alignments comprising a structural family or structural feature (MAKPAT/PROPAT; [75]). These and related techniques for identification of functional patterns or domains are powerful, however, they rely further on experienced and intelligent use of typical motif search techniques such as Blast and reverse Blast screens, detailed in [76]. Recent examples have revealed internal repeats in a breast cancer gene [77] and relations between sperm–egg binding protein and proto-oncogenes [78].
- *Amino acid composition* discriminates strongly between protein structural classes. This can also be used to train neural networks to identify protein structural families with 80% accuracy [79]. Dipeptide frequencies may work even better [80]. A further improvement is the global description of the amino acid sequences [81].
- *Functional residues* may also be predicted by a vector representation of the entire protein and sequence residues in a generalised sequence space [82].
- In a similar vein, intra/extracellular ratios calculated for different amino acid types allow topology prediction of the *membrane-spanning segments* in membrane proteins (TMAP, useful and available on the net; [83]).

3.2.2. Threading

Sequence threading to distinguish whether or not a certain sequence matches a known fold seems to be currently the most sophisticated of the fold recognition approaches. Nevertheless, the distinction between the fold recognition methods is somehow arbitrary. For instance, a neural network may do a threading task, the folding class, secondary structure or motif recognition performed by the same program may exploit amino acid profiles too.

At first glance, threading should be ideal for fold

recognition as there exists only a limited number of folds. Alexandrov et al. [84] estimated 6700 different folds taking into account uneven representation in PDB. Taking a less precise measurement for similarity, ρ(based on optimal rigid body superposition) Crippen and Maiorov [85] calculated that there are about 128 folding motifs for single chain proteins (that have no more than about 170 residues and are not beta-barrels) of which about 100 had their structure determined experimentally. Hence, the conformational space seems to be reduced to a couple of hundred possibilities. This is a deception: threading is NP complete (see box 1, Appendix A) as gaps of different lengths must be allowed and long range interactions have to be taken into account [17]. Most algorithms thus use different simplifications:

(a) Only the local environment is taken into account [86].

(b) Non-local contacts are assumed with generic bulk peptide [87].

(c) first all suboptimal alignments are compiled by looking only locally, then potential favored alignments are pairwise assessed [88].

(d) Interaction preferences are always evaluated with the original structure [89].

(e) No gaps are allowed in the alignment [90].

After different procedures to obtain the (hopefully) correct alignment of test sequence and known 3D structure, scoring functions must judge the quality of the match [91]. These include three dimensional profiles [92,93], empirical functions estimating solvent exposed area [94] or the frequency of contacts between hydrophobic groups [95] and knowledge based potentials [39,40], sometimes combined with a reduced residue representation. As in ab initio approaches, complex minimization techniques such as multidimensional minimization [96] can be applied too. Jones [97] used pairwise potentials and efficiently optimized alignment and detection using a genetic algorithm. Calculating residue–residue specific interactions is non-trivial [98,99] and, quite importantly, threading has to take into account distant interactions correctly.

Already Novotny et al. [100] noted that incorrectly folded models may arrive at potential energy values comparable to correct structures. The absence of non-bonded contacts is a first requirement for correctly assigned structures. Larger solvent-accessible surface and a greater fraction of non-polar side-chains exposed are some indications for a mispredicted fold. However, a thorough quantitative evaluation of the results is always required, rigorous blind testing is a good validation. Thus 7 out of 11 chains could be correctly blind predicted by [101] while having no predictions for 7 other folds and concentrating on different alpha–beta barrels. Proteins with low residue identity but identical folds show an inherent risk of failure for threading methods. In such examples there is low conservation of the residue–residue interactions, residue accessibility and secondary structure [102].

Algorithms that discriminate compact non-native structures from known native structures of globular proteins mostly do not allow specific identification of a structure and only indicate the plausibility of the predicted structure. Nevertheless, they have good applications, for instance in experimental structure determination [103]. Another interesting and promising variation of the threading theme is the algorithm [104], which screens different structural models according to functional data, as in this case antibody epitope data, to decide on the correct structure.

3.3. Searching for the correct fold in the conformational space of protein folding

Ab initio prediction is more challenging than homology modeling or threading, and is the only way to derive a prediction when no similar test fold is known. Furthermore, such simulations allow insights into the forces governing protein folding. Finally progress here can be directly exploited for the problem areas in homology modeling (modeling the loop regions and less homologues structures) and threading (nonphysical interaction potentials; detection of similar folds in unrelated sequences).

Due to the complexity of the prediction task, several simplifications have been studied.

3.3.1. Simplified models

Simplified models or selected parts of the protein allow a complete enumeration of all conformations. This has been largely done in simplified lattice models, e.g. [105] or in many studies on a minimal 2D square lattice using only two types of amino acids (‘hp-model’; e.g. [106]). In side-chain placement

complete enumeration of all rotamers is more easily possible, e.g. [107] (CONGEN package).

3.3.2. Build-up techniques

Build-up techniques from building blocks are attractive to restrain the exponential growth of possibilities [7]. However, tertiary interactions, as required for instance to delineate the complete protein topology, are inherently more difficult to take into account. This is strikingly illustrated by sequence-identical penta- or hexapeptides found in different conformations in the complete protein, e.g. [108].

3.3.3. Deterministic methods

Deterministic methods of global optimization achieve a flattening and smoothing of the potential energy hypersurface by averaging the potential over a suitably wide range. Protocols based on Schroedinger's equation (e.g. [109]) and the diffusion equations [110,111] have been developed. In the first case, the correct basis set has to be chosen, but it may be softened by a mean field approximation. In the second approach, the relationship between the global minimum of the smoothed function and of the real energy function has to be resolved. This has been successful e.g. for terminally blocked alanine and Met-enkephalin [110] and requires about 100 times less computation time than Monte Carlo (MC) [7]. Regulation of the degree of hyperspace smoothing by a neural network has been applied successfully to the folding of melittin [112]. However, despite being known for some time now and in contrast to the peptide simulations just mentioned, the diffusion equation approach has not yet been successfully applied to proteins.

Protein topology prediction through parallel constraint logic programming is a further interesting approach in the deterministic direction [113] as well as different approaches predicting protein topology by optimizing many distance constraints. This is a classical NMR approach [114], but new theoretical insights are also possible [35,115].

3.3.4. Stochastic searches: Monte Carlo

MC methods are able to localize the local minimum, moreover a number of additional strategies have been applied to improve their searching ability. This includes the combination with energy minimization MCM, for instance, to search reliably in peptides for the most stable alternative structures in solution (e.g. for met-enkephalin). Uneven sampling balanced by this is especially an efficient method if analytical gradients of the energy are available to find the local minimum. Other enhancements include build-up procedures from subsets of the structure [7]. These are quite successful in cases such as poly(Gly-Pro-Pro) or other fibrous peptides related to collagen. Self-consistent electrostatic field (SCEF) methods first neglect all components of the energy except the electrostatic field and then successively align the dipole moment of each residue optimally in the field, then minimize the energy of the whole molecule and repeat the whole process. MCM and SCEF may be combined (electrostatically driven MC, EDMC).

Biased probabilities for sterically more permissible regions improve performance [29]. Fixed secondary structure is helpful to reduce the huge conformational space, particularly important if proteins are to be tackled. Thus myohemerythrin and cytochrome b562 (two four-helix bundles) could be predicted with RMSDs of 4.1 Å to the crystal structure, specifying the known secondary structure and keeping it fixed ([116]; compare this to results below using only secondary structure prediction and the genetic algorithm [34]). Subsequently, applying known secondary structure, myoglobin (eight helices, and with 146 residues already of considerable size) was delineated with 6.2 Å RMSD to observed, as well as a β-strand containing domain (CTF, 66 residues) with 5.0 Å RMSD. On R69 (60 residues, five helices) the potential used was not as successful [117]. A useful conformational and energetical analysis of these results and on different reduced protein models is given in [117,118]. Results utilizing known and fixed secondary structure with evolutionary prediction algorithms are described (and were published) later [36,119]. MC searches in higher dimensional spaces can be gradually projected to three dimensions [7] or analysed with heating and cooling or a progressive decrease in MC step size [120]. The MC search space can again be reduced using grids. A recent example is prediction of the quaternary structure of coiled coils in mutants of the GCN4 leucine zipper [121].

3.3.5. Stochastic searches: Evolutionary strategies

Genetic algorithms (GAs) use the optimization procedures of natural evolution: mutation, crossover

and replication operating on strings encoding solution trials [122]. The optimum protocol is quite problem dependent [123], furthermore, the optimal choice of recombination and mutation is important [124]. For many optimization problems, including specific protein folding applications, GAs are superior search methods to MC approaches [106] as they search the space in parallel by using a whole population of solution trials, select the fittest by suitable fitness criteria and exchange information on the whole search space by cross-over. A drawback in terms of computing time is that a whole population has to be calculated. Hence, GAs are often used in reduced representation.

Genetic algorithms have great application potential for protein engineering, evolution simulation and three dimensional structure prediction [125–127]. Sequence and simple protein folding principles as selection criteria and standard secondary structure predictions have been sufficient to obtain the C_α-trace fold of three different four-helix-bundle proteins ([34], RMSD to observed around 6 Å). With known secondary structure and specific criteria for strand selection 19 different protein mainchain topologies (helical, β-strand, mixed) have been successfully delineated ([119]; RMSD 1.8–6.9 Å).

Sun [36] applied a genetic algorithm to obtain grid free two smaller protein folds assembling a peptide fragment library also containing the two structures and with the radius of gyration as an additional constraint. However, a subsequent paper [128] leaves this approach and achieves, by the use of a simplified energy function (compare with other similar examples in this review), known as secondary structure and S–S bonds (as a further guide) low root mean square deviations for seven smaller helical protein folds (RMSD 2.2–4.3 Å); only four larger or strand rich folds still have some problems (RMSD 6.5–10.7 Å). Initial conformations came from a library of observed mono- and dipeptide conformations for different residue types.

LeGrand and Merz [129] used a full atom representation, a rotamer library and the AMBER potential for fitness evaluation to fold several small peptides and crambin. They used point charge electrostatics and accessible surface area for fitness evaluation, with no mutation and extensive annealing of sidechain conformations after cross-over. Bowie and Eisenberg [130] selected nine-residue segments from a library of fragment conformations on the basis of their environment according to amino acid profiles. These fragments were optimally assembled using a genetic algorithm and an empirical fitness function evaluating profile fit, hydrophobicity, accessible surface area, atomic overlap and sphericalness of the structure. The weighing was biased by the experimental structure but under these conditions several native-like structures could efficiently be generated, along with several competing low-energy incorrect structures.

Pedersen and Moult achieved a native-like structure for a 22-residue fragment (a folding initiation helix in clotting factor VIII) in three *blind predictions* (their two less successful trials on independent folding initiation sites in a neurotoxin show that true blind testing is still a challenge) [123]. Other genetic-algorithm-guided topology predictions include amoebapore and NK-lysin [131] and predicting flp-protein topology [132]. The latter example has been tested and complemented by different structural probing methods. Examining the forthcoming results from the recently concluded CASP2 contest will offer many more applied and blind test predictions by different groups and methods of comparison. Rabov et al. [133] introduced Cartesian recombination operators to genetic algorithm simulations on protein folding as an interesting extension. Another genetic algorithm application is sidechain placement (which requires properly positioned backbone conformations; [134]) including efficient pruning methods [61]. A genetic algorithm has also been used to construct initial conformations for loop regions in proteins, using four possible conformations for each residue [135].

3.3.6. Molecular dynamics

These simulations are the most ambitious in computation of protein folding because the detailed evolution of the protein structure in time is investigated by integration of Newton's equations of motion in more or less sophisticated force fields. The upper time limit of the time step size is set by $1/15$ of the highest frequency vibration modes of atomic movement (typically a few femto seconds [136]) defined by the stiffness of the energy function [137] to achieve a reasonable integration accuracy. High-temperature

molecular dynamics may aid in conformational searches [138].

The main problem area is the accuracy of the force fields used. Smith and van Gunsteren [139] explain methods for molecular dynamics calculation for translational and rotational diffusion in proteins, while pointing to limitations in the calculation of protein–water interactions in current force fields. Further, sampling errors have to be considered as only comparatively short time scales and can be computationally treated. For instance low-frequency atomic displacements of the order of 1 ns are probably currently undersampled [140]. Novel methods in the development of molecular dynamics during the last year [141] include faster algorithms to pursue molecular dynamic simulations, advances in the design of new optimization algorithms guided by molecular dynamics protocols and techniques to calculate quantum spectra of protein vibrations and simulated annealing by restrained molecular dynamics and flexible restraint potentials [142].

One application of molecular dynamics is to study different conformational motions within a molecule. Problems and tools in extracting useful information about such conformational motions from a molecular dynamics trajectory are reviewed in [143]. Further, useful suggestions for protein (e.g. enzyme) movements can be obtained, but they need agreement and validation with experimental data. Molecular dynamics simulations are also used as a tool for theoretical studies of protein folding and unfolding. Studies investigating initial stages in protein unfolding are summarized in [144].

4. Prediction success

Moult [2] concluded from the CASP I prediction competition that in comparative modeling many of the numerical methods did not perform as well as expected, but the resulting structures are still of great practical use. The new methods for fold identification (‘threading’) have been partially successful, and show considerable promise for the future. Except for secondary structure data, results from traditional ab initio methods have been poor. However, any of the current problems in *homology* or comparative *modeling* are minor versions of the global protein folding problem. It has already been shown that if the sequence alignment is in error, then the comparative model is guaranteed to be wrong. In addition, loops, insertions and deletions are exceedingly difficult to model. Where sequence identity between target and template structure is high (> 70%) comparative modeling is highly successful, but even more advanced schemes do fail when the sequence identity is low (around 30%) [9].

Though *threading methods* are capable of identifying the correct fold in many cases among the top ten suggestions (from, e.g., a total of 203 folds stored in the data base), they are not yet reliable enough. Common folds such as TIM barrels are more easily recognized. However, the quality of the sequence-structure alignments is generally very poor. Lemer et al. suggest that current threading procedures achieve fold recognition just because they maximize hydrophobic interactions in the protein core [10]. The potential energy functions for threading seem to be indeed the critical point [91]. Despite promising initial results, the methods are clearly not yet fully mature in this respect.

Defay and Cohen are still quite pessimistic on ab initio fold *prediction* of novel folds, however, protein fold and motif prediction are possible when the motif is recognizably similar to another known structure [8]. Scheraga is more optimistic, as the multiple-minima problem has been surmounted for small open-chain and cyclic peptides, and for regular-repeating fibrous proteins, and as progress is being made in resolving this problem for globular proteins [7]. However, for practical purposes pure ab inito predictions of proteins have still to be developed further to yield reliable results. Only if cross-verified by experimental data and applied to carefully selected protein examples [7] will current algorithms yield useful topological suggestions on specific protein structures.

Molecular dynamics simulations are inherently very complex. Karplus and Sali claim slightly ironically that protein unfolding studies are currently rather tractable and this is what most investigators in this area analyze in the moment [133]. Nevertheless, especially simplified protein models have generated results of considerable interest, particularly in relation to the properties of molten globules, the nature of transition states in folding and the significance of the

energetics of the native fold relative to those of alternative folds.

This is only a very brief summary of expert opinions on different prediction methods, but evidently the consensus is that the computational protein folding problem is far from being solved. A large blind test competition has been conducted in 1994 (CASP) by Moult and co-workers [2] and a second, similar trial has again been hosted by them (CASP2, finished in December 1996). Homology modeling is constantly improving; however, a high degree of sequence identity is still important for success and even in 1996 the loop regions remained problem areas under such conditions. Threading clearly has evolved, several participating groups correctly assigning five of the seven test structures (though these structures might not yet be completely representative). Proper alignments were achieved much earlier than two years ago. Correct structure identification could be achieved, but it did not necessarily coincide with this. The challenge for ab initio prediction in the blind tests conducted in CASP2 has remained tough beyond peptides (see also [7]). The reader is alerted to the detailed, qualified reports on the CASP2 meeting.

5. New theoretical concepts

This is a very active area of research and only a few interesting examples are illustrated. The first two are directly applicable for the design of experiments on homologous structures, the others are more theoretical and stress also on forefront problems in different research areas.

Wilmanns et al. [145] improved the correctness of alignments of structurally compatible sequences in *homology modeling*, applying inverse protein folding and combining environmental profiles and pair preference profiles, and used their novel approach to predict the structure of HisA.

Functional interfaces can be identified by extracting functionally important residues from sequence conservation patterns in homologous proteins and mapping these onto the protein surface. Lichtarge et al. provide a simple and versatile approach for this to direct mutagenesis studies in related protein structures and to avoid the need to calculate each of them [146].

Crippen is optimistic about deriving correct *folding potentials* [147]. Using a 2D lattice and two-residue-type simplified protein chains he has demonstrated that the correct potential can be derived from knowing the native and the non native structures for a given sequence. However, for more realistic applications additional complications arise. Thus the accuracy of statistical potentials, used for threading, folding, docking and protein fold recognition, as extracted from protein structures, was examined recently in a very nice study [98]. Using exact lattice models and an exactly known energy function they showed that although statistical potentials extracted from their test structures often correctly rank-ordered the strength of inter-residue interactions, they did not reflect the true underlying energies because of systematic errors arising from the neglect of excluded volume in proteins. Complex residue–residue distance dependencies observed in statistical potentials, even those among charged groups, can be largely explained as an indirect consequence of the burial of non-polar groups.

Local moves in dihedral space for a localized segment in the protein, while the rest remains unchanged, seem to be a new enhancement both for MC and genetic algorithm simulations of ab initio protein folding [148].

For more detailed models and *molecular dynamics*, better empirical solvation models are highly needed. However, to yield any improvement in comparison to elaborated and well validated models such as [27], new approaches such as the recent continuum solvation model by Fraternali and van-Gunsteren [149] need further evaluation and research.

Several computational studies aim at insights into *the folding process* itself. Dinner et al. [150] concluded from their $5 \times 5 \times 5$ MC heteropolymer model on a cubic lattice that folding begins with a rapid collapse followed by a slow search through the semi-compact globule for a sequence-dependent stable core (30 of the 176 of native contacts). An efficient search for the core is dependent on structural features of the native state: large amounts of stable, cooperative structure that is accessible through short-range initiation sites. Sali et al. speculated that a pronounced energy minimum is a necessary and sufficient condition to ensure folding of a sequence to its lowest energy conformation [4]. Unger and Moult

show that this result strongly depends on the particular temperature scheme chosen [151]. Instead the strength of possible interactions between residues that are close together in the sequence seem to be a dominant factor determining stable folding. Sequences with many possible strong local interactions (either favorable or a mixture of strong favorable and unfavorable ones) are easier to fold in their models. Though these conclusions may partly be model dependent, their findings strengthen the notion that initial formation of local substructures is important for the foldability of real proteins.

6. Obstacles and perspectives

The interconnectedness of protein structure is a major problem area even in *comparative modeling*, most strikingly illustrated by the cooperative effects of single mutations on the whole protein (e.g. [58]). Samudrala et al. [152] expect improvement in the near future through the development of structure-based sequence alignment tools, side chain interconnectedness, rotamer choice algorithms and a better understanding of the context sensitivity of conformational features. In all these areas there is in fact constant progress, however this only slowly pushes down the twilight zone (below 30% sequence identity) where the risk of failure for homology modeling starts and rapidly increases.

Better integration of experimental data and structural prediction is a safe route to more accurate models. Thus for the exact calculation of side chain placement, the latest version of the CONGEN package includes a new functional representation of NMR-derived distance constraints [153].

Improved search strategies come up, e.g. enhanced dead-end elimination [61]. Yue and Dill fold proteins ab initio and obtain a first speed up by using very few well selected energy parameters [154]. A common trend for simple but well chosen parameters seems to be emerging here (compare with other studies [34,39,116,128,130]). Moreover, they have developed a new constraint-based and *exhaustive searching* algorithm, which appears to be quite fast. They successively add one residue, trying out all possible rotamer conformations. A major drawback is their use of predefined bounded regions around the native structure to prune out wrong directions. The method gives low-energy conformations, among which (so actually not yet clearly separated from the other conformations) there is also the native state for crambin, pancreatic polypeptide, melittin and apamine. The *diffusion equation method* [7] is another promising new algorithm to find the energy minimum in peptides, though it has not yet been extended to proteins (a potential exciting area for new research). A rapid breakthrough in computational protein folding by novel search strategies may be possible but is rather unlikely. Nevertheless, the combination of steadily increasing computational power and continuously refined and more intelligent search algorithms led to the decisive improvement in computational protein prediction evident during the last decade.

However, errors in the *energy parameters* used for protein structure prediction remain a major obstacle. Improving the modeling of the electrostatic field exemplifies one of these challenges [155]. Finkelstein et al. have suggested a way to reduce energy errors by relying on statistical mechanics. Carrying out calculations at a temperature T^*, somewhat below that of protein melting, should give the best results [156]. Computational protein prediction has many more facets, the following three areas give a flavour on the variety future developments may bring.

In *pattern matching* approaches, the screening of many sequences to predict critical features of the structure and function on a genomic scale becomes feasible, for instance in screening yeast chromosome VIII [157]. This is a hot area of research. Further analyses on completely sequenced genomes are rapidly following and yield directly usable results on involved functions of new protein sequences.

Foldons, kinetically competent, quasi-independent folding units of a protein is another new research area. Panchenko et al. [158] predicted kinetic foldons and compared these with exons and structural modules. For 16 of 30 proteins studied, they correlated with geometrically defined structural modules, but only weakly with exons. At least for gamma II crystalline, myoglobin, barnase, a-lactalbumin and cytochrome c, the foldons and some noncontiguous clusters of foldons compared well with intermediates actually observed in experiments.

A further frontier of research are *protein interac-*

tions. Thus Vieth et al. [121] made testable predictions for the state of association for the GCN4 leucine zipper and its fragments, fos and jun using their discretized protein models. It will be interesting to see the results of experimental tests. Though even more demanding in its full complexity, calculation and prediction of the interaction between two proteins can yield good progress: Molecular *docking* programs successfully predict the binding of a β-lactamase inhibitory protein to the enzyme [159]. Though docking in this example of course starts from two known three dimensional structures and this is only one specific example, the result that six independent groups were each successful in this task clearly shows that this is an area where computational prediction of protein structure is solidly advancing.

7. Conclusion

Computational protein folding has to cope with a vast conformational space. This is apparent in ab initio calculations, but homology modeling and threading are also NP hard problems; further unavoidable limitations such as imperfections in model representation and energy function are inherent to any simulation.

Useful and available predictions for the experimentalists involve: homology models if sequence identity is sufficiently high (loops are not yet accurate), prediction of secondary structure (about 70% accuracy, much worse on some proteins) and functional sequence analysis including important functional hints on amino acids motifs. Threading and even more so ab initio folding and molecular dynamics are clearly not yet sufficiently reliable for an independent identification of the three dimensional fold but showed constant improvement during the last years. These are central areas of research in computational protein prediction. They provide important theoretical insights into protein folding but may also yield directly useful structural information in concrete protein examples, if well selected and professionally used by an expert and critically compared with and evaluated by experimental data.

Progress in the field has been achieved partly due to increased computational power and better understanding and more data available on protein structures — but very much due to the many creative and new ideas put into this exciting field too. This review could necessarily present only an appetizing fraction of them.

Acknowledgements

We thank Dr. Frank Eisenhaber and Dr. Burkhard Rost for stimulating discussions, Dr. Thomas Creighton and Dr. Toby Gibson for critical reading of the manuscript, and Dr. Michélle Stewart as well as Dr. Mark Nichols for stylistic corrections.

Appendix A. Box 1: NP-complete

An algorithm solves a problem correctly if it terminates correctly on every instance of the problem. It is customary to identify the computational complexity of a problem with that of the most efficient algorithm that solves it. The class of problems that can be solved in polynomial time is named 'P'. Problems which belong to this class are formally tractable. The class of problems, whose solution, once found or *guessed* may be *verified* in polynomial time is named 'NP' [17,160]. Problems in NP are solvable in non-deterministic polynomial time, meaning that if one could check and guess all possible solutions simultaneously, the solution would be obtained in polynomial time. The practical importance of this theoretical condition is that the search for a solution, not the verification step, determine whether a polynomial time solution is possible or not. Clearly, P is a subset of NP; but it is unknown whether P = NP.

References

[1] M. MacArthur, P.C. Discroll, J.M. Thornton, Trends Biotechnol. 12 (1994) 149–153.
[2] J. Moult, Curr. Op. Biotech. 7 (1996) 422–427.
[3] C. Levinthal, J. Chem. Phys. 65 (1968) 44–45.
[4] A. Sali, E. Schakhnovich, M. Karplus, Nature 369 (1994) 248–251.
[5] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, Proteins: Structure, Function and Genetics 21 (1995) 167–195.

[6] S. Yun-Yu, A.E. Mark, W. Cun-xin, H. Fuhuae, H.J.C. Berendsen, W.F. van Gunsteren, Protein Eng. 6 (1993) 289–295.
[7] H.A. Scheraga, Biophys. Chem. 59 (1996) 329–339.
[8] T. Defay, F.E. Cohen, Proteins 23 (1995) 431–445.
[9] S. Mosimann, R. Meleshko, M.N.G. James, Proteins 23 (1995) 301–317.
[10] C.M.R. Lemer, M.J. Rooman, S.J. Wodak, Proteins 23 (1995) 337–355.
[11] Y. Bai, S.W. Englander, Proteins: Structure, Function and Genetics 24 (1996) 145–151.
[12] C.B. Anfinsen, Science 181 (1973) 223–230.
[13] T. Creighton, Protein Folding, Freeman, New York, 1992.
[14] M.J. Gething, J. Sambrook, Nature 355 (1992) 33–45.
[15] F.U. Hartl, Nature 371 (1994) 557–559.
[16] T.J. Ngo, J. Marks, Protein Eng. 5 (1992) 313–321.
[17] R.H. Lathrop, Protein Eng. 7 (1994) 1059–1068.
[18] M.E. Davis, J.A. McCammon, Chem. Rev. 90 (1990) 509–521.
[19] A.H. Juffer, E.F.F. Botta, B.A.M. van Keulen, A. van der Ploeeg, H.J.C. Berendsen, J. Comp. Phys. 97 (1991) 144–171.
[20] M. Holst, R.E. Kozack, F. Saied, S. Subramaniam, J. Biomol. Struct. Dyn. 11 (1994) 1437–1445.
[21] G. Nemethy, K.D. Gibson, K.A. Palmer, C.N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, H.A. Scheraga, J. Phys. Chem. 96 (1992) 6472–6484.
[22] W.F. van Gunsteren, H.J.C. Berendsen, Angew. Chem., Int. Ed. Engl. 29 (1990) 992–1023.
[23] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, J. Comp. Chem. 4 (1983) 187–217.
[24] S.J. Weiner, P.A. Kollman, D.T. Nguyen, D.A. Case, J. Comp. Chem. 7 (1986) 230.
[25] P. Dauber-Osguthorpe, V.A. Roberts, D.J. Osguthorpe, J. Wolff, M. Genest, A.T. Hagler, Proteins 4 (1988) 31–47.
[26] B. Robson, E. Platt, J. Mol. Biol. 188 (1986) 259–281.
[27] W.C. Still, A. Tempzczyk, R.C. Hawley, T. Hendrickson, J. Am. Chem. Soc. 112 (1990) 6127–6129.
[28] C.J. Cramer, D.G. Truhlar, Science 256 (1992) 213–217.
[29] R. Abagyan, M. Totrov, J. Mol. Biol. 235 (1994) 983–1002.
[30] P.J. Steinbach, B.R. Brooks, J. Comp. Chem. 15 (1994) 667–683.
[31] H.Q. Ding, N. Karasawa, W.A. Goddard III, J. Chem. Phys. 97 (1992) 4309–4315.
[32] H. Schreiber, O. Steinhäuser, J. Mol. Biol. 228 (1992) 909–923.
[33] J.S. Richardson, D.C. Richardson, N.B. Tweedy, K.M. Gernet, T.P. Quinn, M.H. Hecht, B.W. Ericson, Y. Yan, R.D. McClain, M.E. Donlan, M.C. Surles, Biophys. J. 63 (1992) 1186–1209.
[34] T. Dandekar, P. Argos, J. Mol. Biol. 236 (1994) 844–861.
[35] A. Aszodi, W.R. Taylor, Folding and Design 1 (1996) 325–334.
[36] S. Sun, Protein Sci. 2 (1993) 762–785.
[37] A. Wallqvist, M. Ullner, Proteins 18 (1994) 267–280.
[38] A. Aszodi, W.R. Taylor, Biopolymers 34 (1994) 489–505.
[39] M.J. Sippl, Curr. Op. Struct. Biol. 5 (1995) 229–235.
[40] M.J. Sippl, S. Weitckus, Proteins 13 (1992) 258–271.
[41] O. Lund, J. Harsen, S. Brunak, J. Bohr, Protein Sci. 5 (1996) 2217–2225.
[42] A. Godzik, A. Kolinski, J. Skolnick, J. Comp. Chem. 14 (1993) 1194–1202.
[43] W.J. Brown, A.C.T. North, D.C. Phillips, K. Brew, T.C. Vanaman, R.A. Hill, J. Mol. Biol. 42 (1969) 65–86.
[44] M.S. Johnson, N. Srinivasan, R. Sowdhamini, T.L. Blundell, Crit. Rev. Biochem. Mol. Biol. 29 (1994) 1–68.
[45] L. Holm, C. Sander, Nucl. Acid Res. 22 (1994) 3600–3609.
[46] W.R. Taylor, C.A. Orengo, J. Mol. Biol. 208 (1989) 1–22.
[47] T.P. Flores, C.A. Orengo, D.S. Moss, J.M. Thornton, Protein Sci. 2 (1993) 1811–1826.
[48] Y. Luo, L. Lai, X. Xu, Y. Tang, Protein Eng. 6 (1993) 373–376.
[49] A.C.W. May, M.S. Johnson, Protein Eng. 7 (1994) 475–485.
[50] C.W.G. van Gelder, F.J.J. Leusen, J.A.M. Leunissen, J.H. Noordik, Proteins 18 (1994) 174–185.
[51] C.M. Topham, A. McLeod, F. Eisenmenger, J.P. Overington, M.S. Johnson, T.L. Blundell, J. Mol. Biol. 229 (1993) 194–220.
[52] S.J. Hubbard, F. Eisenmenger, J.M. Thornton, Protein Sci. 3 (1994) 757–768.
[53] C.A. Laughton, J. Mol. Biol. 235 (1994) 1088–1097.
[54] R.L. Dunbrack, M. Karplus, Nature Struct. Biol. 1 (1994) 334–340.
[55] G. Vriend, J. Mol. Graph. 8 (1990) 52–56.
[56] A. Sali, T.L. Blundell, J. Mol. Biol. 234 (1993) 779–815.
[57] V. De Fillipis, C. Sander, G. Vriend, Protein Eng. 7 (1994) 1203–1208.
[58] A. Nordhoff, U.S. Bücheler, D. Werner, R.H. Schirmer, Biochemistry 32 (1993) 4060–4066.
[59] R. Schneider, C. Sander, Nucleic Acids Res. 24 (1996) 201–205.
[60] J. Desmet, M. DeMayer, B. Hazes, I. Lasters, Nature 356 (1992) 539–542.
[61] I. Lasters, M. DeMaeyer, J. Desmet, Protein Eng. 8 (1995) 815–822.
[62] D.V. Renouf, E.F. Hounsell, Adv. Exp. Med. Biol. 376 (1995) 37–45.
[63] E.L.L. Sonnhammer, R. Durbin, Comput. Appl. Biosci. 10 (1994) 301–307.
[64] C. Chothia, A.M. Lesk, EMBO J. 5 (1986) 823–826.
[65] S. Henikoff, J.G. Henikoff, J. Mol. Biol. 243 (1994) 574–578.
[66] T.J. Gibson, M. Hyvönen, A. Musacchio, M. Saraste, E. Birney, Trends Biochem. Sci. 19 (1994) 349–353.
[67] A. Bairoch, Nucl. Acids. Res. 21 (1993) 3097–3103.
[68] T.K. Attwood, M.E. Beck, Protein Eng. 7 (1994) 841–848.
[69] D.I. Frishman, P. Argos, J. Mol. Biol. 228 (1992) 951–962.
[70] J.M. Chandonia, M. Karplus, Protein Sci. 4 (1995) 275–285.
[71] B. Rost, Methods Enzymol. 266 (1996) 525–539.

[72] B. Rost, C. Sander, Proteins 23 (1995) 295–300.
[73] D. Frishman, Biochem. Biophys. Res. Commun. 219 (1996) 686–689.
[74] M. Reczko, A.C. Martin, H. Bohr, S. Suhai, Protein Eng. 8 (1995) 389–395.
[75] K. Rohde, P. Bork, Comput. Appl. Biosci. 9 (1993) 183–189.
[76] P. Bork, T. Gibson, Methods Enzymol. 266 (1996) 162–184.
[77] P. Bork, N. Blomberg, M. Nilges, Nat. Genet. 13 (1996) 22–23.
[78] P. Bork, Science 271 (1996) 1431–1432.
[79] I. Dubchak, S.R. Holbrook, H.S. Kim, Proteins 16 (1993) 79–91.
[80] M. Reczko, H. Bohr, Nucl. Acids Res. 22 (1994) 3616–3619.
[81] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, PNAS 92 (1995) 8700–8704.
[82] G. Casari, G. Sander, A. Valencia, Nat. Struct. Biol. 2 (1995) 171–178.
[83] B. Persson, P. Argos, Protein Sci. 5 (1996) 363–371.
[84] N.N. Alexandrov, N. Go, Protein Sci. 3 (1994) 866–875.
[85] G.M. Crippen, V.N. Maiorov, J. Mol. Biol. 252 (1995) 144–151.
[86] K.Y.J. Zhang, D. Eisenberg, Protein Sci. 3 (1994) 687–695.
[87] C. Ouzounis, C. Sander, M. Scharf, R. Schneider, J. Mol. Biol. 232 (1993) 805–825.
[88] J. Gracy, L. Chiche, J. Sallantin, Protein Eng. 6 (1993) 821–829.
[89] R. Abagyan, D.I. Frishman, P. Argos, Proteins 19 (1994) 132–140.
[90] S.H. Bryant, C.E. Lawrence, Proteins 16 (1993) 92–112.
[91] D.T. Johnson, J.M. Thornton, Curr. Op. Struct. Biol. 6 (1996) 210–216.
[92] J.U. Bowie, K. Zhang, M. Wilmanns, D. Eisenberg, Methods Enzymol 266 (1996) 598–616.
[93] J.U. Bowie, R. Luthy, D. Eisenberg, Science 253 (1991) 164–170.
[94] P. Koehl, M. Delarue, Proteins 20 (1994) 264–278.
[95] C. Colovos, T.O. Yeates, Protein Sci. 2 (1993) 1511–1519.
[96] V.N. Maiorov, G.M. Crippen, J. Mol. Biol. 227 (1992) 876–888.
[97] D.T. Jones, Protein Sci. 3 (1994) 567–574.
[98] P.D. Thomas, K.A. Dill, J. Mol. Biol. 257 (1996) 457–469.
[99] D.T. Jones, J.M. Thornton, Curr. Op. Struct. Biol. 6 (1996) 210–216.
[100] J. Novotny, R. Bruccoleri, J. Newell, D. Murphy, E. Haber, M. Karplus, J. Biol. Chem. 258 (1983) 14433–14437.
[101] D.T. Jones, R.T. Miller, J.M. Thornton, Proteins 23 (1995) 387–397.
[102] R.B. Russel, G.J. Barton, J. Mol. Biol. 244 (1994) 332–350.
[103] Y. Wang, H. Zhang, W. Li, RA Scott, PNAS 92 (1995) 709–713.
[104] L. Jin, F.E. Cohen, J.A. Wells, Proc. Natl. Acad. Sci. 91 (1994) 113–117.
[105] A.V. Finkelstein, B. Reva, Nature 351 (1991) 497–499.
[106] R. Unger, J. Moult, J. Mol. Biol. 231 (1993) 75–81.
[107] R.E. Bruccoleri, Mol. Simul. 10 (1993) 151–174.
[108] B.I. Cohen, S.R. Presnell, F.E. Cohen, Protein Sci. 2 (1993) 2134–2145.
[109] K.A. Olszewsky, L. Piela, H.A. Scheraga, J. Phys. Chem. 96 (1992) 4672–4676.
[110] J. Kostrowicki, H.A. Scheraga, J. Phys. Chem. 96 (1992) 7442–7449.
[111] D. Shalloway, J. Glob. Optim. 2 (1992) 281–311.
[112] T. Head-Gordon, F.H. Stillinger, Biopolymers 33 (1993) 292–303.
[113] D.A. Clark, C.J. Rawlings, J. Shirazi, A. Veron, M. Reeve, Ismb. 1 (1993) 83–91.
[114] M. Nilges, J. Mol. Biol. 245 (1995) 645–660.
[115] P. Sibbald, J. Theo. Biol. 173 (1995) 361–375.
[116] A. Monge, R.A. Friesner, B. Honig, Proc. Natl. Acad. Sci. USA 91 (1994) 5027–5029.
[117] M. Monge, E. Lathrop, J.R. Gunn, P.S. Shenkin, R.D. Friesner, J. Mol. Biol. 247 (1995) 995–1012.
[118] R.A. Friesner, J.R. Gunn, Annu. Rev. Biophys. Biomol. Struct. 25 (1996) 315–342.
[119] T. Dandekar, P. Argos, J. Mol. Biol. 256 (1996) 645–660.
[120] Y. Okamoto, Proteins 19 (1994) 14–23.
[121] M. Vieth, A. Kolinski, J. Skolnick, Biochemistry 35 (1996) 955–967.
[122] D.B. Fogel, Evolutionary Computation: Towards a New Philosophy of Machine Intelligence, IEEE Press, New York, 1995
[123] J.T. Pedersen, J. Moult, Proteins 23 (1995) 454–460.
[124] D.B. McGarrah, R.S. Judson, J. Comp. Chem. 14 (1993) 1385–1395.
[125] T. Dandekar, P. Argos, Protein Eng. 5 (1992) 637–645.
[126] J.T. Pedersen, J. Moult, Curr. Op. Struct. Biol. 6 (1996) 227–231.
[127] D.E. Clark, D.R. Westhead, J. Comp.-Aided Mol. Design 10 (1996) 337–358.
[128] S. Sun, P.D. Thomas, K.A. Dill, Protein Eng. 8 (1995) 769–778.
[129] S.M. LeGrand, K.M. Merz Jr., Mol. Simulat. 13 (1994) 299–320.
[130] J.U. Bowie, D. Eisenberg, Proc. Natl. Acad. Sci. USA 91 (1994) 4436–4440.
[131] T. Dandekar, M. Leippe, Folding and Design 2 (1997) 47–52.
[132] P. Saxena, I. Whang, Y. Voziyanov, C. Harkey, P. Argos, M. Jayaram, T. Dandekar, Biochim. Biophys. Acta 1340 (1997) 187–204.
[133] A.A. Rabov, H.A. Scheraga, Protein Sci. 5 (1996) 1800–1815.
[134] P. Tuffery, C. Etchebest, S. Hazout, R. Lavery, J. Biomol. Struct. Dyn 8 (1991) 1267–1289.
[135] C.S. Ring, F.E. Cohen, Isr. J. Chem. 34 (1994) 245–252.
[136] D. Bouzida, S. Kumar, R.H. Swendson, Phys. Rev. A 45 (1992) 8894–8901.
[137] R.W. Harrison, J. Comp. Chem. 14 (1993) 1112–1122.

[138] R.E. Bruccoleri, M. Karplus, Biopolymers 29 (1990) 1847–1862.
[139] P.E. Smith, W.F. van-Gunsteren, J. Mol. Biol. 235 (1994) 629–636.
[140] J.B. Clarage, T. Romo, B.K. Andrews, B.M. Pettitt, G.N. Phillips Jr., Proc. Natl. Acad. Sci. USA 92 (1995) 3288–3292.
[141] R. Elber, Curr. Op. Struct. Biol. 6 232–235
[142] D. Bassolino-Klimas, R. Tejero, S.R. Krystek, W.J. Metzler, G.T. Montelione, R.E. Bruccoleri, Protein Sci. 5 (1996) 593–603.
[143] P. Dauber-Osguthorpe, C.M. Maunder, D.J. Osguthorpe, J. Comput. Aided Mol. Des. 10 (1996) 177–185.
[144] M. Karplus, A. Sali, Curr. Op. Struct. Biol. 5 (1995) 58–73.
[145] M. Wilmanns, D. Eisenberg, Protein Eng. 8 (1995) 627–639.
[146] O. Lichtarge, H.R. Bourne, F.E. Cohen, J. Mol. Biol. 257 (1996) 342–358.
[147] G.M. Crippen, Easily searched protein folding potentials, J. Mol. Biol. 260 (1996) 467–475.
[148] A. Elofsson, S.M. Le-Grand, D. Eisenberg, Proteins 23 (1995) 73–82.
[149] F. Fraternali, W.F. van-Gunsteren, J. Mol. Biol. 256 (1996) 939–948.
[150] A.R. Dinner, A. Sali, M. Karplus, Proc. Natl. Acad. Sci. USA 93 (1996) 8356–8361.
[151] R. Unger, J. Moult, J. Mol. Biol. 259 (1996) 988–994.
[152] R. Samudrala, J.T. Pedersen, H.B. Zhou, R. Luo, K. Fidelis, J. Moult, Proteins 23 (1995) 327–336.
[153] D. Bassolino-Klimas, R. Tejero, S.R. Krystek, W.J. Metzler, G.T. Montelione, R.E. Bruccoleri, Protein Sci. 5 (1996) 593–603.
[154] K. Yue, K.A. Dill, Protein Sci. 5 (1996) 254–261.
[155] J. Warwicker, J.Mol. Biol. 236 (1994) 887–903.
[156] A.V. Finkelstein, A.M. Gutin, A.Y. Badretdinov, Proteins 23 (1995) 151–162.
[157] C. Ouzounis, P. Bork, G. Casari, C. Sander, Protein Sci. 4 (1995) 2424–2428.
[158] A.R. Panchenko, Z. Luthey-Schulten, P.G. Wolynes, Proc. Natl. Acad. Sci. USA 93 (1996) 2008–2013.
[159] N.C. Strynadka, M. Eisenstein, E. Katchalski-Katzir, B.K. Shoichet, I.D. Kuntz, R. Abagyan, M. Totrov, J. Janin, J. Cherfils, F. Zimmerman, A. Olson, B. Duncan, M. Rao, R. Jackson, M. Sternberg, M.N. James, Nature Structural Biology 3 (1996) 209–210.
[160] R. Unger, J. Moult, Bull. Math. Biol. 55 (1994) 1183–1198.