The Protein Folding Problem: A Biophysical Enigma

J. S. Fetrow[a,*], A. Giammona[a], A. Kolinski[b,c] and J. Skolnick[c]

[a]*GeneFormatics, Incorporated, 5830 Oberlin Drive, Suite 200, San Diego, CA 92121, USA;* [b]*Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland and* [c]*Laboratory of Computational Genomics, Donald Danforth Plant Science Center, 975 North Warson Road, St. Louis, MO 63132, USA*

Abstract: Protein folding, the problem of how an amino acid sequence folds into a unique three-dimensional shape, has been a long-standing problem in biology. The success of genome-wide sequencing efforts has increased the interest in understanding the protein folding enigma, because realizing the value of the genomic sequences rests on the accuracy with which the encoded gene products are understood. Although a complete understanding of the kinetics and thermodynamics of protein folding has remained elusive, there has been considerable progress in techniques to predict protein structure from amino acid sequences. The prediction techniques fall into three general classes: comparative modeling, threading and *ab initio* folding. The current state of research in each of these three areas is reviewed here in detail. Efforts to apply each method to proteome-wide analysis are reviewed, and some of the key technical hurdles that remain are presented. Protein folding technologies, while not yet providing a full understanding of the protein folding process, have clearly progressed to the point of being useful in enabling structure-based annotation of genomic sequences.

I. OVERVIEW OF THE PROTEIN FOLDING PROBLEM

The protein folding problem was conceptualized with the publication of the Anfinsen results in 1961 [1]. This set of experiments showed that all the information necessary for a protein to fold into its functional, three-dimensional shape is contained in its primary amino acid sequence. These results set the stage for years of research devoted to understanding how the amino acid sequence, comprising a small set of 20 different amino acids, could encode the complex, three-dimensional structure of a protein.

In 1968 Levinthal described what has come to be called the "Levinthal paradox" [2]: the observation that it would be impossible for a protein to fold at observed rates by randomly searching all possible conformations of the polypeptide chain. Even for a small protein of 100 amino acids the number of conformations to be searched in folding the backbone would be well in excess of 9^{100} and this estimate does not take side chain conformations into account. Zwanzig resolved the paradox in 1992 by showing that even a slight bias in the folding potential would allow a stable conformation to be found quickly [3], nevertheless the kinetics and thermodynamics of folding remain an enigma.

Solving the protein folding problem has often been compared in difficulty to "cracking the second genetic code". An understanding of this problem and a computational solution to it would greatly enhance the ability to utilize the enormous amount of data being generated by genome sequencing projects. Researchers would no longer need to rely on resource-intensive experimental methods for determining protein structures, but could determine them computationally. Important processes, such as drug discovery, could be accelerated and greatly enhanced, saving significant resources. However, the time scale of the process complicates using computation to solve the enigma of protein folding. Current methods of simulation by *ab initio* protein folding are feasible and robust over a range of 10 to 100 ns on a Cray computer, but protein folding occurs on the millisecond to minutes timescale [4]. This has led to the development of partial solutions that seek to predict the structure of proteins, rather than delineate the folding pathway. Even this more narrowly defined problem presents difficult challenges: What are good models for the potential energy surface? How can native conformations be found and recognized?

Computational solutions to the prediction of protein folds are often divided into three main categories: comparative modeling algorithms, threading algorithms and *ab initio* folding algorithms. Comparative modeling tools are used to build a model based on a previously determined structure of a related sequence. To apply comparative modeling tools successfully, the two proteins must be minimally related by structural similarity. Usually, there is significant sequence similarity between the two proteins and a sequence alignment is used as a starting point. The second method, threading, attempts to identify proteins that are structurally similar to one another, even if sequence similarity is negligible. Threading alignments may be used as starting points for comparative modeling algorithms. Finally, *ab initio* folding algorithms are used to fold the proteins according to basic physico-chemical principles, without the use of a structural template. Each of these methods will be reviewed in this paper in detail.

II. OVERVIEW OF COMPARATIVE MODELING METHODOLOGIES

Comparative modeling procedures generate three-dimensional models for amino acid sequences of unknown

*Address correspondence to this author at GeneFormatics, Incorporated, 5830 Oberlin Drive, Suite 200, San Diego, CA 92121, USA; Tel: (858) 882-5903; Fax: (858) 450-1138; E-mail: jacquefetrow@geneformatics.com

1389-2010/02 $35.00+.00 © 2002 Bentham Science Publishers Ltd.

structure (probes) by comparison with similar sequences of known structure (templates). This approach is possible because small changes in protein sequence generally produce only small changes in protein structure [5]. Conversely, since protein structure is more highly conserved than protein sequence [6], a discernible sequence similarity is a likely indicator of structural similarity.

Comparative modeling procedures comprise four steps: a) fold assignment by selection of structural templates, b) alignment of the probe to the templates, c) construction of the three-dimensional model and d) evaluation of the model. To make comparative modeling feasible on a genome-wide scale, these steps must be automated and integrated. This section will review the techniques used in each of these steps and discuss the application of comparative modeling on a genome-wide scale.

a. Fold Assignment

Alignment methods are often used to select structural templates. These methods fit into three categories: pairwise sequence comparison methods that produce a separate alignment for the probe with each template sequence (e.g., BLAST [7]), multiple sequence alignment methods that produce a single alignment for the probe against a set of template sequences (e.g., PSI-BLAST [8]) and threading methods that use a structure-dependent score to optimize the sequence-structure alignment of the probe with a library of fold templates (e.g., PROSPECTOR [9]). The pairwise sequence alignment methods are most successful when sequence identity is greater than 30% although recent efforts to derive improved amino acid interchange matrices from structural superposition data show some promise in improving the ability of pairwise methods to detect remote homologues [10]. The multiple sequence alignment methods have been shown to identify twice to thrice as many homologues as the pairwise methods [11]. By relying on a combination of sequence and structure comparison, the threading methods are the most sensitive in detecting distant homologues, and their success appears to be more directly related to the level of structural similarity between template and probe [12]. The disadvantage of the more sensitive methods, particularly as more distant homologues are detected, is the rapid increase in the false positive rate.

b. Alignment Optimization

The quality of a model generated by comparative modeling depends significantly on the quality of the alignment of the probe and its template. Sequence-only alignments for the selection of templates are tuned to recognize distant sequence similarity and may not be optimal for model building. Before model building, it is usually necessary to apply a method that optimizes the alignments (e.g., CLUSTAL [13]). In cases for which sequence identity between the probe and template sequences is above 30%, alignment procedures are robust and comparative modeling can produce models with accuracy approaching that of structures determined at low resolution by crystallography or NMR spectroscopy [14]. The resulting models can be of the quality required for use in docking and ligand-design exercises [15, 16].

Below 30% sequence identity, the "twilight zone" of sequence similarity, traditional sequence-only alignment techniques may produce significant misalignment between the probe and template sequences [14, 17]. Models built from such alignments may have significant errors in backbone structure. The application of threading algorithms for sequence alignment has extended the reach of comparative modeling into the twilight zone by using structural data to guide the alignment [9, 12, 18-20]. For example, optimizing the burial status of residues or evaluating residue pair interactions may improve alignments. Successful alignments in the twilight zone produced by threading algorithms identify the overall fold but may still have substantial uncertainty in loop conformations and in the size and packing of secondary structure elements. Structural similarity may be limited to only part of the structure while the remainder of the structure is different from that of the template [19]. A recently developed method termed "generalized comparative modeling" [21] aims to refine modeled structures derived from such alignments by performing *ab initio* folding in the vicinity of the template structure, in effect sampling a larger piece of conformational space near the template structure and moving the modeled structure closer to the native structure of the probe. In some cases, these lower-quality models have even been used for docking of small molecules [22].

c. Model Building

Once an optimized probe-template alignment has been generated by sequence alignment or by a threading program, the next step in comparative modeling is to generate atomic level structural details for the probe sequence. Methods for model construction have recently been reviewed [23]. This step comprises three areas: a) generation of coordinates for the protein backbone from the aligned portions of the templates, b) generation of coordinates for the unaligned portions of the backbone and c) generation of coordinates for the side chain atoms. One approach for modeling the aligned segments of the probe sequence generates a framework for the probe by averaging the Cα positions from the structurally conserved regions of the aligned templates and adding the remaining main chain atom positions for the conserved regions by comparison with the template with the highest sequence similarity to the probe sequence [24]. A second approach identifies and assembles short, all-atom segments from the probe sequence that "match" the atomic positions in the template structures [25-28]. The segments are generated by scanning known structures or by conformational search. A third approach derives spatial constraints from the template structure that are then used to generate a structure for the probe sequence by distance geometry [29]. For example, intramolecular distances between residues in the aligned template structure can be used as restraints for modeling the probe structure. The advantage of this approach is that restraints can be added from sources other than the original template alignment, such as NMR data.

Approaches to the generation of loop conformations for the non-aligned portions of the probe fall into two classes: *ab initio* folding methods [21, 30-32] and database scanning methods [5, 33-36]. The *ab initio* folding methods employ conformational searches restrained by the structurally conserved regions modeled in the first step. The database scanning methods search protein structure databases for segments that "fit" the structures already generated for the aligned portions of the probe. The database scanning methods are limited by the incomplete representation of loop conformations in known protein structures.

Finally, after structurally conserved and structurally unconserved regions have been modeled, side chains must be added. Side chain conformations are modeled by comparison to similar structures or derived from rotamer library conformations, but the method used must take into account the coupling between main chain and side chain conformations [37-41]. Side chain conformations can deviate significantly from rotamer conformations under the influence of a rigid backbone. Likewise, side chains can cause significant changes, even in regions of the backbone that would otherwise be structurally conserved. At this time, it is difficult to determine *a priori* if a side chain will move to accommodate a rigid backbone, or if a backbone will move to accommodate a side chain. These effects have been recently reviewed [23].

d. Model Evaluation

Once models are created, they must be evaluated to be sure one has produced a representative model. For alignments above the twilight zone of sequence identity, confidence in the correctness of the modeled fold is derived from the high level of sequence similarity between probe and template. For alignments within the twilight zone (generally between 25-35% sequence identity), a combination of the calculation of an energy-based Z score and the ability of functional descriptors to identify the active site of the probe can be used to assess confidence in the model [9, 21, 42-47]. A method such as PROCHECK can be used to evaluate the basic stereochemistry of the model [48]. Models can also be evaluated against statistical analyses of structural databases for characteristics such as intramolecular packing, formation of the hydrophobic core, solvent accessibility and distribution of charged groups [23, 49-52]. For these models produced from alignments in the twilight zone of sequence identity, it can be difficult to determine absolutely whether one has a representative model. Use of these knowledge-based approaches can, however, improve the confidence in the final model produced.

e. Genome-wide Comparative Modeling

Models from two methods for genomic-scale comparative modeling, Modeller [53-55] and GEM (published as GENECOMP [21]) were recently compared. Modeller generates an objective function for a given probe that comprises restraints derived from distances and dihedral angles in the aligned templates and restraints that enforce good stereochemistry. The function is then optimized in Cartesian space and a model generated. GEM uses threading to select structural templates, then generates a structure for the probe sequence by performing conformational searching in the vicinity of the aligned templates for the aligned segments and applying an *ab initio* folding method for the unaligned segments (Fig. **1**).

Modeller is less computationally intensive than GEM, however GEM was found to produce, on average, qualitatively better molecular models. For the comparison, each method was used to generate models for 68 proteins from the Fischer database, a standard benchmark. This data set contains a variety of protein pairs of similar structure, but low sequence identity. The starting point for both methods was the same set of templates and alignments generated by the threading program, PROSPECTOR. A comparison between the model and the known structure for both methods is shown in Fig. (**2**). When templates were very good (allowing at least one method to build a model with a Cα coordinate-root-mean-square-deviation (RMSD) to the known structure of less than 3.5 Å), the methods performed comparably with one notable exception. For probe 1onc, the GEM model had an RMSD of 3.5 Å while the Modeller model had an RMSD of 5.14 Å. As the sequence similarity between the probe and template sequences decreases, the GEM models are more similar to the experimentally determined structures than the Modeller models. There were 14 probes for which at least one method produced a model with a Cα coordinate RMSD to the known structure below 5 Å. In seven of these cases, GEM produced the better model, in two cases Modeller produced the better model and in five cases the two methods performed equally well (Fig. **2**). Both cases in which Modeller performed better were for probe-template pairs whose level of sequence identity is in the upper portion of the twilight zone at approximately 21% (1hip, 20.9%, and the A chain of 1chr, 21.1%). In both cases, the Modeller models were only slightly better than the GEM models (Cα RMSD to the known structure of 4.3 Å for GEM vs. 4.1 Å for Modeller for 1hip, and 4.9 Å for GEM vs. 4.6 Å for Modeller for the A chain of 1chr). Of the seven models on which GEM performed better, three were created on templates with a level of sequence identity to the probe deep in the twilight zone, 1ten (1.5% sequence identity), 3chy (4.6%) and 1mup (18.2%). GEM produced significantly better models for these probes (Cα RMSDs to known structures of 3.6, 4.4, and 4.4 Å respectively) than did Modeller (Cα RMSDs to known structures of 5.2, 6.2 and 4.9 Å, respectively). Fig. (**3a**) is an example of a GEM model for the probe, 1aba, superimposed on the known structure. The probe-template pair has 24.5% sequence identity and the overall Cα RMSD between the probe model and known structure is 3.85 Å. All secondary structure elements (three helices and a pair of layered β sheets) have been placed correctly, although the boundaries of the secondary structural elements and the loops connecting them show greater variation from the known structure.

f. Outlook for the Future

As the number of protein folds is finite (estimated at 2,000-4,000 folds [56]) and knowledge of fold space is increasing rapidly, the success rate of comparative modeling

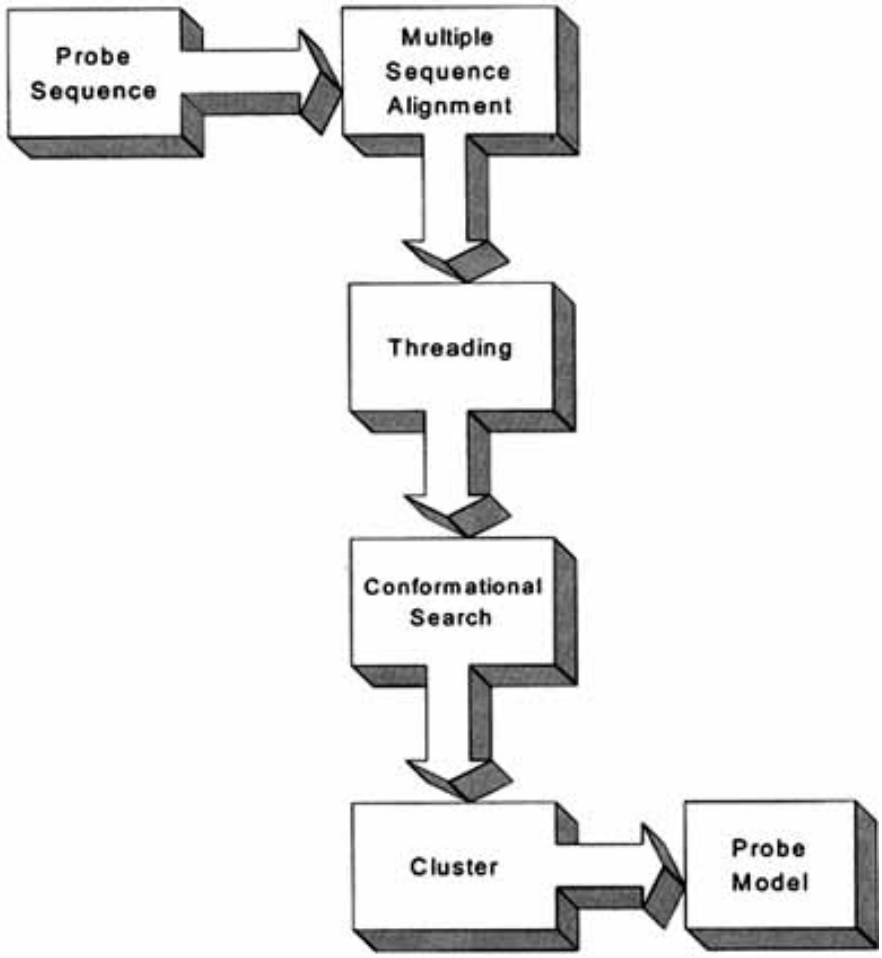

Fig. (1). Modules of GEM Comparative Modeling Process: Near and distant sequence profiles guide the initial probe-sequence to template-structure alignment from which a starting set of predicted pair interactions is calculated. Consensus interactions are pooled and converted into a potential of mean force used in subsequent rounds of threading. For the aligned segments, REMC is used to explore conformational space for a lattice representation of the probe in the vicinity of the template backbone. For the unaligned segments *ab initio* folding generates backbone structures. The ensemble of results from the conformational search is clustered and an all-atom model of the probe is generated for the average structure from each cluster.

techniques is expected to grow. A recent effort to describe a strategy for optimizing information return from structural genomics efforts estimates there are about 8,000 protein domain families and a minimum of 16,000 structure determinations will be necessary to permit non-twilight zone comparative modeling of 90% of protein structural space [57]. Such a goal can be achieved in as little as a decade but could also take three times as long if proteins for structure determinations are chosen randomly [57].

Improvements in modeling methods to recognize homology in the twilight zone will also expand the applicability of comparative modeling. A hard lower limit for comparative modeling of 10% sequence identity arises from the observation that, at that level of similarity, accuracy of pairwise alignments is indistinguishable from chance [10]. Nevertheless, lowering the threshold of the twilight zone from 30% to 20% sequence identity would cut in half the minimum number of structure determinations necessary to allow comparative modeling of 90% of protein structural space [57]. The use of threading to generate alignments for comparative modeling (as in GEM) is already moving in this direction. Comparative modeling methods will become more successful as alignment techniques, side chain and loop model building and model evaluation are improved.

III. OVERVIEW OF THREADING TECHNIQUES

The goal of threading is to find the closest matching structure in a library of known folds for a given probe sequence [58, 59]. In principle, if not in practice, accurate construction of such alignments should extend comparative modeling techniques. For example, threading should be able

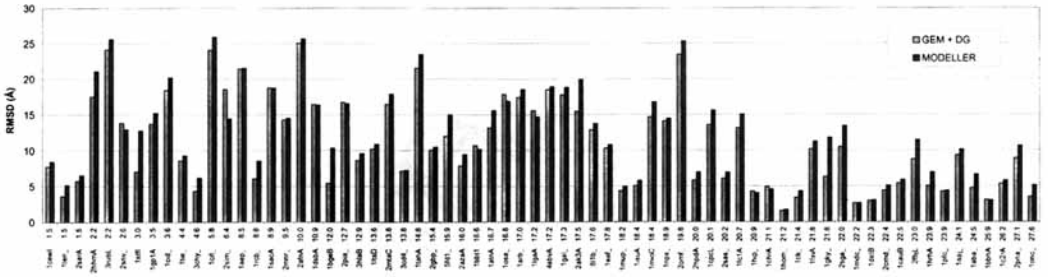

Fig. (2). Histogram of Cα RMSD between model and known structure for GEM and Modeller models. For 68 probes in the Fischer database decoy set, the Cα RMSD in the aligned region between the model and known structure is shown for the models built with GEM (light) and Modeller (dark). Each pair of bars is labeled with the PDB name of the probe and the percent sequence identity between the probe and the template. Each pair of models was built from the same template alignment generated by PROSPECTOR. For 14 probes for which at least one model RMSD was < 5 Å, in seven cases GEM produced the better model, in two cases Modeller produced the better model and in five cases the difference in performance of the two methods was insignificant.

to recognize not only distantly related (homologous or evolutionarily related) proteins, but also analogous folds where the proteins are evolutionarily unrelated, but have converged to the same fold. This is complicated by the fact that analogous proteins may share a common structural core over a fraction of their sequences with the remainder adopting a significantly different fold.

Key characteristics of threading algorithms include: a) the kind of scoring function or "energy" used to assess the probe sequence-template structure fitness, b) the level of detail describing the protein (side chains, backbone, Cαs, etc.) and c) if multibody interactions are included, the kind of optimization algorithm used.

a. Scoring Functions

Among the terms that have been used in scoring functions for describing the compatibility of the probe sequence to the specified template are the burial status of residues, secondary structure propensities and/or predicted secondary structure, additional penalty terms [60, 61] (e.g., a penalty that depends on the difference in the length of the probe and template sequences), and pair or higher order interactions between side chains. When the scoring function contains more than one term, their relative weights must be established. The best scoring functions also include an evolutionary component related to the sequence similarity between the template and the probe [19]. This term significantly improves both the fold recognition and alignment ability of a threading algorithm [61-65]. In a structure-based approach, such terms should not be required as chemistry and structural information should be enough to define the fold; unfortunately, in practice they are quite important.

b. Representation of the Protein

When pair interactions are considered, the type of interaction center must be selected. Among the standard choices are the backbone atoms [66, 67], the alpha carbons [68, 69], side chain centers of mass, other interaction centers [58, 70], or any side chain heavy atom [71]. Given a set of interaction centers, the functional form of their interactions must be specified. These include contact potentials [71, 72], distance-dependent potentials [68, 73], and interaction environments [20].

c. Optimization Methods

Having specified the scoring function and protein representation, the optimal alignment between the probe sequence and each structural template must be generated. Dynamic programming [74] is the best approach when the scoring function is purely local. When a non-local scoring function is used (e.g., pair interactions), the situation is more complicated. As part of structure optimization, the interactions in the template structure should utilize the actual partners present in the probe sequence, but in dynamic programming, these specific partners are unknown. To retain speed (essential if entire genomes are to be scanned), among the approximations made is the use of dynamic programming with the "frozen" approximation where the interaction partners or local environmental preferences are taken from the template protein [71, 75]. Then, iterative updating is done [20, 71, 76]. Others use double dynamic programming, where a subset of interactions recognized as being the most important in the first pass of the dynamic programming algorithm [68] are updated. There are also more computationally intensive, but exact approaches which fully evaluate the non-local scoring function and search for the optimal probe-template alignment by Monte Carlo [70] or branch-and-bound approaches [58].

d. Challenges and Improvements in Threading Algorithms

Threading follows a similar paradigm to comparative modeling: a) choosing the structural template, b) generating an alignment and c) constructing a model. As such, its

Fig. (3a).

Fig. (3b).

Figs. (3a) and **(3b).** GEM model for the probe, 1aba, a glutaredoxin mutant. Figure 3a shows a ribbon representation of the GEM model (dark) superimposed on the known structure (light). The Cα RMSD between the model and known structure is 3.85 Å. The template for this model, 1ego, has 24.5% sequence identity to the probe. The threading Z score for alignment of the template to the probe is 7.4. Figure 3b shows the key residues of the functional site identified by the FFF in space-filling models (model in dark gray, known structure in light gray). The FFF identifies the functional site as a disulfide oxidoreductase catalytic site.

limitations are similar to those of comparative modeling. First, an example of the probe's structure must have been solved already or the method will fail. Second, the quality of the model depends on the extent of structural similarity between the probe and template structure. Third, while alignment quality improved from CASP1 (Critical

Assessment of Protein Structure) to CASP3 [77] and now CASP4, it nevertheless remains problematic, and until recently, the alignment could not be adjusted to fix errors [78].

Almost all threading approaches freeze the template structure (i.e., the frozen approximation) and do not allow it to adjust to the probe sequence. For close homology modeling, this is a good approximation, but as sequence identity between the probe and template sequences moves into the twilight zone of sequence identity, or if the two proteins are analogous rather than homologous, substantial backbone rearrangements may be ignored. The ability to recognize analogous structures is precisely the realm where threading should be the most valuable as compared to pure sequence-based methods, thus it is clear that the frozen approximation introduces a severe limitation.

Because threading uses structure, it should be superior to one-dimensional sequence-based approaches that assess the evolutionary relationship between sequences by inferring a structural relationship, such as PSI-BLAST, which has been used for genome-wide modeling exercises. In practice, however, many of the successful fold-recognition approaches in CASP3 and CASP4 were pseudo one-dimensional and used evolutionary information that contributed a significant fraction of the selectivity [79]. The Jones [80] and Koretke groups [65], among others, employed this type of approach. The Nishikawa group [81] also employed a hierarchy of local scoring functions to describe hydration, secondary structure, hydrogen bonding and side chain packing.

Successful approaches in CASP3 where structure played a more prominent role included that of the Sippl group [82]. They employed a burial energy and the frozen approximation to evaluate pair interactions, but used a single sequence rather than a sequence profile; thus all interactions are pseudo one-dimensional.

The Bryant group [83] explicitly treated pair interactions within a structural core identified using the conservation of structure across each protein family. This approach has three limitations: a) the core identification required by this method limits applicability to protein families where a number of structures in the family have been solved, b) a BLAST sequence-profile component is used, which weights sequence similarity and evolutionary relationships heavily and c) since the method uses a non-local scoring function and a Monte Carlo search procedure to find the best probe-template score, these calculations are very CPU intensive, thereby precluding its application on a genomic scale.

The general consensus is that CASP3 saw some progress in threading, with alignment quality improving from CASP2 [77, 79, 84]. Nevertheless, threading does better on distant homology sequence pairs than on analogous pairs. A similar conclusion was reached for CASP4. Thus, threading techniques to address structurally similar, but non-homologous, pairs of proteins are still required.

These observations motivated the development of a new threading algorithm called PROSPECTOR (PROtein Structure Predictor Employing Combined Threading to Optimize Results). PROSPECTOR is both hierarchical and iterative. In the first stage, a sequence profile [85-87] is used to generate an initial probe sequence to template structure alignment. Then, this alignment is used to calculate the partners for the evaluation of the pair interactions known as partly defrosted approximation. Both near and distant sequence profiles are used for a total of four scoring functions. The consensus contacts found in at least weakly scoring structures are pooled, converted into a potential of mean force and used in a subsequent round of threading. The process is repeated and a third round of threading is done.

The method was tested on a standard benchmark, the Fischer database [61] comprised of 68 probe sequences and 301 template structures. In 59 cases, the top scoring match from PROSPECTOR aligned the probe protein to the correct template structure [9]. It is superior to earlier efforts including the alternative hybrid method [88], BLAST [7] and PSI-BLAST [8, 87]. It might be argued that since PROSPECTOR uses four scoring functions, and the hybrid method only uses three, this is not a strictly fair comparison. If the results obtained from one scoring method, such as the "distant" sequence-profiles, are eliminated then 58 correct matches in the top scoring position are identified, as compared to 52 alignments identified by Gonnet with a pairwise sequence alignment method that includes predicted secondary structure [9]. Then, in a second pass of PROSECTOR that uses predicted contacts from the first pass, 61 proteins are identified in the top scoring position[1].

e. Genome-wide Threading

Following benchmarking on the Fischer database, PROSPECTOR was applied on a genome-wide scale. The first genome considered was *Mycoplasma genitalium*, consisting of 480 open reading frames (ORFs) [89]. The first pass of PROSPECTOR assigns 230 proteins to a structure in the Protein Databank when a minimum Z score of 7 is used as a cutoff. The second pass assigns 260, and the third pass assigns 300, again with a Z score of 7. All assignments are made using an automated protocol based on the score significance. In contrast, Fischer and Eisenberg [90] assigned the folds of 103 of 468 proteins using their threading algorithm. Gerstein reported identification of 211 proteins using BLAST [91, 92] while Genethreader assigns 200 proteins, but 15 appear to be incorrect [18] as assessed by a consensus of Gerstein's results (http://bioinfo.mbb.yale.edu/genome/MG/) and three-pass PROSPECTOR results.

PROSPECTOR was also applied to additional genomes. The *Escherichia coli* genome contains 4,289 ORFs [93], for which three-pass PROSPECTOR assigns 2,611 ORFs to structures in the Protein Data Bank. Similarly, three-pass PROSPECTOR assigns 2071 of 4101 ORFS in the *Bacillus subtilis* genome and 972 ORFS of 1,530 ORFs in the *Aquifex aeolicus* genome to known protein structures. Thus, for a typical small genome, over 50% of the structures can be assigned to folds with some degree of confidence. All

[1] J. Skolnick and coworkers, unpublished results.

assignments had threading Z scores greater than 7. Note that fold and function assignment are not the same; in only about 50% of the cases are fold and function degenerate, presenting significant challenges for threading algorithms [94].

IV. OVERVIEW OF *AB INITIO* FOLDING METHODS

Proteins fold in milliseconds to minutes [4]. Classical molecular mechanics simulations (with all atomic details treated in an explicit way) of a protein [95, 96] submerged in an appropriate number of water molecules over such a long time are currently impractical [97, 98]. Consequently, for the purpose of *ab initio* protein folding, the problem needs to be simplified. Simplification can be achieved by reducing the number of explicitly treated degrees of freedom [99] of the polypeptide chain and by simplifying the model force field [100]. Simplification can also include treating solvent in an implicit fashion. In simplified protein models, groups of atoms are replaced by single "united atoms" [101, 102]. Internal degrees of freedom for these united atoms are ignored or treated in a pre-averaged fashion. A simplified representation of the polypeptide chain conformation leads to a simplified interaction scheme [103]. The complexity of the resulting energy landscape of the model is significantly reduced in comparison to the detailed atomic models. Using some or all of these simplifications, the search for the lowest energy conformation (the native protein fold according to Anfinsen's postulate [104]) becomes more feasible. The following sections review these various methods of simplifying the *ab initio* folding problem.

a. Continuous Reduced Models

The first non-trivial reduced models were proposed about 25 years ago. For example, the model studied by Levitt and Warshel [101, 102] assumed two centers of interaction per residue: one representing the main chain segment centered on the alpha carbon, and one representing the side chain. A single degree of freedom per amino acid was assumed, rotation around the Cα-Cα virtual bond, while the planar angles for the reduced backbone were assumed to be constant. A knowledge-based statistical potential controlled the short-range conformational propensities, while the side chains interacted via a Lennard-Jones potential. With this model, molecular dynamics simulations of the small protein, bovine pancreatic trypsin inhibitor, sometimes produced a native-like fold of low resolution. The best structures had a RMSD from native in the range of 6.5 Å.

The model proposed by Levitt and Warshel [101] inspired other analogous simplifications of protein representation. Kuntz *et al.* [105, 106], Hagler and Honig [107], and Wilson and Doniach [108] studied somewhat similar continuous models, with results of a comparable quality.

Reduced continuous-space models with more structural details have also been proposed. Sun designed a model with an all-atom representation of the main chain and a single united atom representation of the side groups [109]. Knowledge-based statistical potentials controlled the interactions between the side groups, and a genetic algorithm was employed as a sampling tool. For small peptides, quite accurate structures were predicted whose RMSD from the native structure ranged from 1.66 Å to 4.5 Å, depending on peptide size. A similar model was studied by Wallqvist and Ullner [103]. More accurate representation of the side chains (two united atoms per side chain for larger amino acids) resulted in slightly more accurate structural predictions.

A very different approach to protein structure prediction was proposed by Pedersen and Moult [110]. They assumed an all-heavy atom representation of the protein with knowledge-based potentials describing interactions between these atoms. A combination of Monte Carlo and genetic algorithms was employed as a sampling tool; and a set of trial structures generated by means of the Monte Carlo method provided the starting population for the genetic algorithm. Approximate structures have been successfully predicted by this method for a number of small proteins.

Very recently, Osguthorpe applied a continuous model and molecular dynamics simulated annealing to structure prediction [111]. The very flexible chain geometry of the model enabled efficient sampling in spite of the detailed representation of the proteins. A force field for the model was derived from statistics of known protein structures. This method was tested during the CASP3 experiment [77], and large fragments of the attempted probe were correctly predicted. For one of the difficult probes, this method produced the most accurate prediction.

A method developed by Scheraga and coworkers enabled exceptionally good predictions for a fraction of CASP3 probes [112]. This off-lattice protein model has a united atom representation of the alpha carbons, side chains and peptide bond, with fixed bond lengths and variable bond angles. The interaction potentials between united atoms describe the mean free energy of interactions and account in an implicit way for the average solvent effect and cooperativity of the hydrogen bonds [113]. The lowest energy models (obtained via the Conformational Space Annealing technique [112, 114]) can be converted into all-atom models and optimized by electrostatically driven Monte Carlo simulations [115].

b. Lattice Models

To further increase sampling efficiency and to allow treatment of larger proteins, discrete or lattice models were proposed and explored. Early studies of the lattice-based protein models, pioneered by Go, *et al.* [116-118] and then followed by Krigbaum and Lin, [119, 120] Skolnick and Kolinski [121-127], Sikorski and Skolnick [128-131], Chan and Dill [132-134], Dill *et al.* [135, 136], Sali *et al.* [137, 138], Shakhnovich *et al.* [139-145], and others [146-148] focused not on structure prediction but rather on basic aspects of protein folding thermodynamics and mechanism. This work has contributed significantly to a general understanding of the forces controlling protein folding, the reasons and requirements for the uniqueness of the native state [149], an explanation of the mechanism of the folding

process and the nature of its cooperativity [150]. In addition, the early studies of very simplified models of proteins provided the knowledge necessary for the subsequent application of lattice models to protein structure (and even function) prediction.

Probably the first attempt at *ab initio* prediction of protein fold via lattice modeling was done by Dashevskii [151]. In his work, the polypeptide conformations were restricted to a diamond lattice. Compact structures resembling native folds of small polypeptides were generated by means of a chain-growth algorithm controlled by a simple force field.

Covell investigated a simple cubic lattice model of proteins [152]. The interaction scheme consisted entirely of long-range interactions that included a pairwise, knowledge-based potential, a surface term and a potential that corrects the local packing of the model chain. Interestingly, the quality of crude folds generated by this method was comparable to the quality of folds obtained using early continuous models.

Covell and Jernigan [153] enumerated all possible compact conformations of a body-centered cubic lattice chain representing small globular proteins, and a simple knowledge-based interaction scheme was used to rank-order these compact conformations [152, 153]. The closest to native conformation could always be found within the top 2% of the lowest energy structures.

An interesting lattice representation of proteins was proposed by Hinds and Levitt [154, 155], where a single lattice vertex of the diamond lattice corresponded to several residues of a real protein. A complex statistical potential was employed to mimic the mean interactions between such fragments of the model protein chain. Frequently, qualitatively correct folds of low resolution were generated by this model.

Kolinski and Skolnick [127, 156-174] developed a series of increasing resolution lattice models of globular proteins. High coordination number lattices were employed to mimic the conformation of the Cα-trace of proteins. In the later models a single sphere, multiple rotamer representation of the side chains was assumed. The force field consisted of several terms mimicking the short-range interactions; explicitly cooperative hydrogen bonds; and one body, pairwise and multibody long-range interactions with an implicit averaged effect of the solvent water. For several small globular proteins and simple multimeric molecular assemblies such models generated folds of low- to moderate-accuracy (high-accuracy in the case of leucine zippers [175, 176]) [162-164, 173, 174]. Monte Carlo simulated annealing was used as a sampling method.

c. Hierarchical Approaches to Structure Prediction

More recently, a number of new approaches to *ab initio* protein structure prediction were developed that combine various modeling and fold selection procedures. Very innovative is the ROSETTA method proposed by Baker and coworkers [177]. The fold prediction procedure consists of several steps. First, a prediction of secondary structure is made using the PHD server based on Rost and Sander's method [178-180]. Subsequently, the predicted secondary structure and multiple sequence alignments guide the selection (from the PDB) of the most plausible 3- to 9-residue structural fragments for the protein of interest. Next, a Monte Carlo algorithm builds a large number of structures from the set of predefined fragments, which are screened with a scoring function containing a hydrophobic burial term, elements of electrostatics, a disulfide bond bias and a sequence-independent term that evaluates the packing of secondary structure elements. The top scoring structures frequently contain the proper fold. The best structures are chosen in a somewhat arbitrary fashion using compactness of the hydrophobic core as a selection criterion. Nevertheless, for eighteen probes of the CASP3 experiment, four predictions were globally correct (with an RMSD range of 4-6 Å from the native structure [181]), and the majority of the predictions contained significant fragments of structure that were qualitatively correct. In a later version of the method, fold selection is done by clustering a very large number of trial structures. A somewhat similar method based on predefined fragments and the Monte Carlo method was earlier investigated by Jones [80].

Ortiz *et al.* [182-185] used a combination of sequence analysis and lattice Monte Carlo simulations for *ab initio* prediction of native-like protein structures. A multiple sequence alignment was used for secondary structure prediction and for prediction of a fraction of long-range side chain contacts via the correlated mutations analysis [186, 187]. These predictions guided the fold assembly. A high coordination lattice model developed by Kolinski and Skolnick [166-171] was used for the protein representation. Monte Carlo simulations started from random expanded conformations and the lowest energy conformations were selected from the large number of final structures from the simulated annealing experiments. This method performed well in test predictions and was capable of assembling low-resolution novel folds during the CASP3 experiment [185].

An interesting hierarchical procedure has been developed by Samudrala, *et al.* [188]. First, the very simple lattice model of Hinds and Levitt [154] (described above) was used to enumerate all compact conformations of the protein of interest. The lowest energy structures were then selected for further consideration. For these, the all-atom structures were reconstructed by fitting the predicted secondary structure fragments to the lattice models. Next, these distorted all-atom structures were subjected to energy minimization using an all-atom force field with spatial restraints taken from the original lattice models, and the optimized structures were rank-ordered according to a combination of all-atom and residue-based knowledge-based potentials. Subsequently, the consensus model obtained via distance geometry was regularized and minimized. Final predictions were made according to conformational energy ranking within a set of structures obtained in this procedure. A number of qualitatively correct protein fragments of significant size were correctly predicted by this method during the CASP3 exercise. This combination of sequence methods, lattice simulations, distance geometry and off-lattice refinement

appears to be very robust. The very crude and low-accuracy model employed for the generation of the initial compact conformations is probably a weak point of this approach.

Recently, Kolinski, Skolnick and coworkers developed a unified approach to structure prediction (Fig. **4**) [189]. This method is termed "unified" for two reasons. First, the method constitutes a well defined hierarchy of sequence analysis techniques; threading; lattice Monte Carlo modeling; fold selection via a combination of distance geometry clustering; off-lattice refinement and minimization and all-atom reconstruction of the final molecular model (or models). Second, the methodology is essentially the same regardless of the level of homology that might be detected between the probe and the protein sequences from the structural database. In other words, the same (automated) methodology is applied to cases that could be classified as suitable for distant homology modeling, threading and *ab initio* (or novel fold) structure prediction. A multiple sequence alignment is used in the derivation of protein-dependent statistical short-range potentials, orientation-dependent pairwise side chain potentials [190], and secondary structure predictions. Predicted secondary structure provides a weak bias for short-range

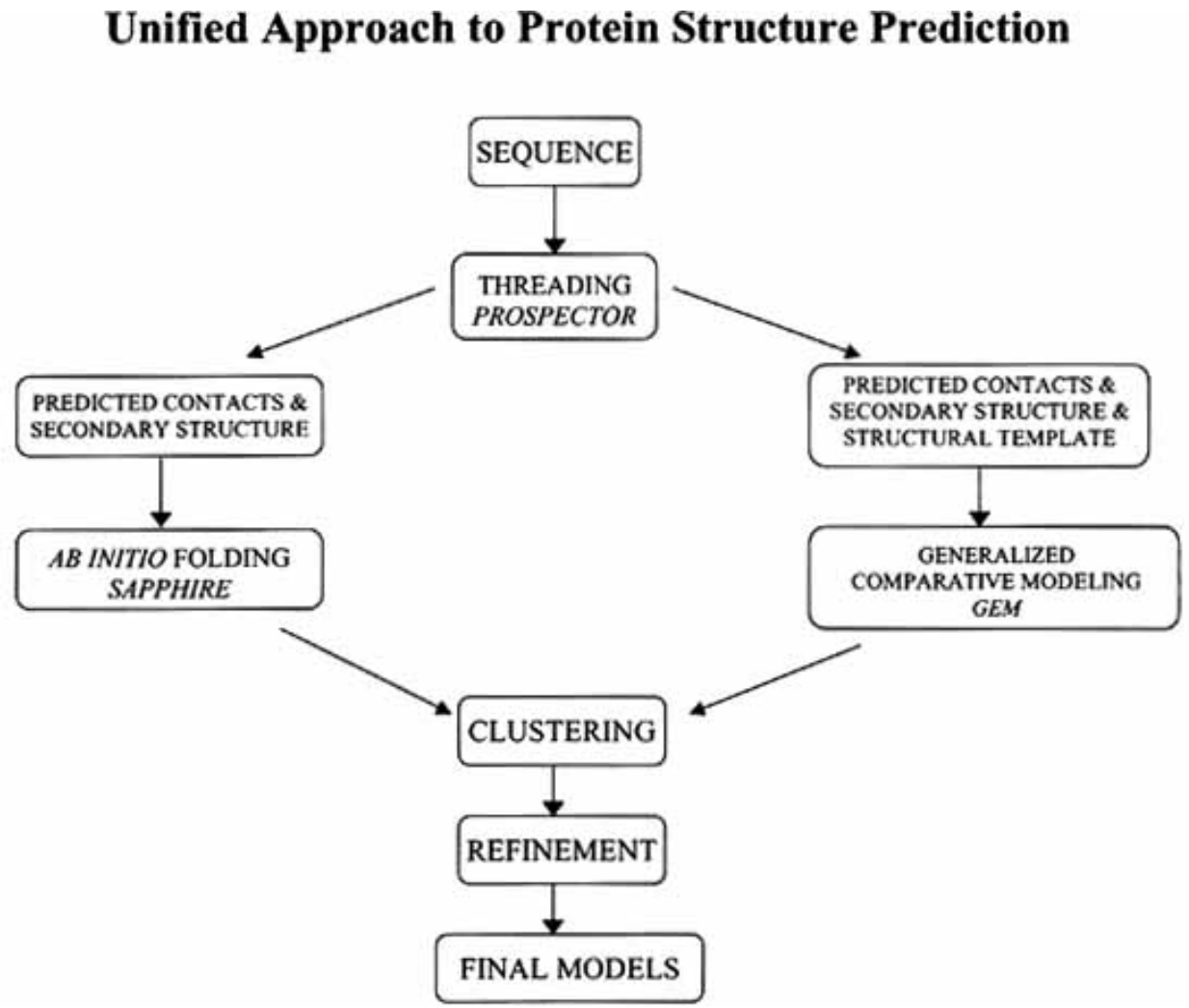

Fig. (4). Unified approach to protein structure prediction. The unified method generates protein models by applying a well-defined hierarchy of sequence analysis tools: threading, lattice Monte Carlo conformational searching, clustering and refinement in a well-defined, automatable process. The methodology is essentially the same regardless of the level of homology that might be detected between the probe sequence and the sequences in the template library. A multiple sequence alignment is used in the derivation of protein-dependent statistical short-range potentials, orientation dependent pairwise side group potentials, and secondary structure predictions. Predicted secondary structure provides a weak bias for short-range conformational propensities and some weak restrictions on the hydrogen bond network for the lattice simulations. Threading provides a prediction of the short-range distances, and prediction of the long-range side chain contacts. When PROSPECTOR finds a homologous protein with a significant similarity score, the resulting structural template is used as an additional source of spatial restraints for the lattice folding simulations.

conformational propensities and some weak restrictions on the hydrogen bond network for the lattice simulations. These parameters are first used in the recently developed threading algorithm (PROSPECTOR [9]). Threading provides a prediction of the short-range distances (an additional source of restraints for the folding stage), and prediction of the long-range side chain contacts (that are on average 70% correct within a limit of ±1 or ±2 residues). When PROSPECTOR finds a homologous protein with a significant similarity score, the resulting structural template is used as an additional source of spatial restraints for the lattice folding simulations. Folding is carried out by means of Monte Carlo simulations using the SICHO (SIde CHain Only) lattice representation of polypeptide chains [78, 191, 192] and REMC (Replica Exchange Monte Carlo) sampling technique [193]. SICHO is a lattice model of 1.45 Å resolution that emphasizes side chain packing instead of the commonly employed Cα reduced representation. In SICHO side chains are modeled as clusters of lattice knots on the underlying simple cubic lattice, and the polypeptide model is a chain of virtual bonds connecting centers of mass of the side chains in their actual rotational isomeric states. Thus, several degrees of freedom of a single residue in a polypeptide are reduced to a single degree of freedom in the model, leading to a significant increase in sampling efficiency and allowing simulation of much longer chains. Folding simulations start from 20-50 random lattice chains (replicas) and a large number of simulations is performed. In the cases when a structural template is provided by PROSPECTOR, the starting chains are built in the spatial proximity of the template chain (for example within GEM, [21] Generalized Comparative Modeling, described above). Otherwise, the entire procedure does not depend on the level of similarity of the probe sequence to the sequences of the known structures. Subsequently, a large number of structures obtained from lattice REMC simulations are subjected to a clustering algorithm [194] that contains elements of distance geometry [195]. The best (in respect to conformational energy) cluster centroids are subject to off lattice refinement and all-atom reconstruction [196]. In test folding experiments for a set of small proteins this procedure produces correct folds for about 80% of the cases. A significant fraction (ca. 25%) of predictions lead to structures whose accuracy is close to experimental quality [197]. During the CASP4 experiment, a preliminary and incomplete version of this methodology was used. The method performed well: in two cases the best models in the competition were obtained using this method [198].

d. Methods for Sampling Conformational Space

The enormous conformational space [2, 3] of proteins defines a very complex energy landscape, and the problem of finding the global energy minimum in this landscape is very difficult [199-201], even for reduced, but not trivial, models. Thus the choice of algorithm that searches conformational space is (almost) as important as the model design and the quality of the force field used. For continuous models, variants of Molecular Dynamics could be used; however this is rarely the optimal choice. Other sampling schemes, including a variety of Monte Carlo methods [193, 202-209], genetic algorithms [210-213], and combinations of these methods could be applied to continuous as well as to the discrete (or lattice) models. We focus here on the Monte Carlo schemes, since they form the majority of the present approaches to *ab initio* structure prediction.

Recently, significant progress in Monte Carlo techniques has been achieved. There are two key characteristics defining the Monte Carlo schemes: first, the method of conformational updating, and second, the choice of acceptance criteria to some extent resulting from the assumed statistical ensemble.

Conformational updates [214] can be global or local. Global updates are employed in the chain growth algorithms, where the sample consists of uncorrelated chains built from scratch. Other algorithms employ pivot moves of a large part of the model chain. Usually, the trial modifications are local, involving only a small portion of the chain, or a small distance displacement of a larger part of the chain. Finally, the local and global modifications can be combined in the same algorithm. The choice of updating strategy in Monte Carlo algorithms (also in genetic algorithms and in hybrid minimization/optimization algorithms) depends on the aim of the studies. Different strategies are needed for the study of protein folding dynamics and thermodynamics than are suitable for those procedures that aim to find the lowest energy conformation. With a proper selection of local updates, an isothermal Monte Carlo simulation with a simple Metropolis [215] acceptance scheme could be considered a numerical solution of a stochastic equation of motion. For sufficiently long time intervals, trajectories from such Monte Carlo simulations mimic trajectories from Molecular Dynamics or Brownian Dynamics. Consequently, properly designed Monte Carlo algorithms can be used in studies of protein long-time dynamics and folding pathways [216-218].

For finding the lowest energy state, the traditional simulated annealing Metropolis scheme might not be the best choice. There are several Monte Carlo schemes that could be much more efficient. Monte Carlo with local minimization could be more efficient for a higher resolution models than a simple Metropolis scheme. For example, multicanonical ensemble methods [219, 220] (one of the variants is known as Entropy Sampling Monte Carlo, ESMC [221-223]) more easily surmount local energy barriers and are quasi-deterministic in the sense that a subsequent simulation uses information from the preceding simulations and (on average) improves the previous results. Moreover, when converged such simulations provide not only a good guess for the lowest energy conformation but also a full thermodynamic description of the model system in a wide range of temperatures. A disadvantage of these methods (sometimes very serious) is their high computational cost.

The REMC technique [193, 224] (or its variants [213]) may be a method of choice for an efficient search for the global minimum of the conformational energy. In the REMC method a number of copies of the model system, placed at various temperatures, are simulated by means of a standard Metropolis scheme. The range of temperatures should cover the region of denatured and folded states of the model protein. At high temperatures, the energy barriers can be surmounted easily. Occasionally, the replicas are exchanged

between various temperatures according to a criterion that depends on temperature difference and energy difference. Consequently, low-energy conformations at a higher temperature have a chance to be moved to a lower temperature. At low temperatures they search a narrower window of the conformational space "looking" for deep minima. Thus, the copies of the system sample not only the conformational space but also move between various temperatures. In spite of the necessity to maintain several copies (replicas) of the model system, the REMC method seems to be qualitatively more efficient (and faster) in finding the lowest energy state than the previously mentioned Monte Carlo schemes [225]. Moreover, a reasonably accurate full (conformational energy and entropy from the same simulation) thermodynamic description can be extracted from relatively inexpensive simulations [226]. Recently, the REMC method was successfully employed in several studies of proteins and protein-like systems [204-206, 225, 226] and also used in protein structure prediction procedures [21, 197, 198].

V. ROLE OF STRUCTURE PREDICTION IN THE GENOMIC ERA

a. Genome-wide Protein Folding Studies of All Sequences from a Genome

To date, several groups have attempted computational protein folding on a genome-wide scale. These efforts include modeling of the yeast genome [227], analysis of folds in the worm genome [228], and modeling of a number of bacterial genomes [229, 230]. Yokoyama and colleagues have initiated a search for all "natively-folded" proteins on a large scale [231], and Baker and colleagues have accomplished a proof-of-concept for *ab initio* folding [232], but neither of these has yet been applied to a complete genome. The Sali, Gerstein and Godzik efforts have utilized some kind of alignment algorithm, either PSI-BLAST or threading alignment methods, to obtain alignments between the genome sequences and known structures. Comparative modeling programs have then been used to build three-dimensional models from the alignments. Using this approach, Sali was able to build models for 1,071 yeast proteins, 17% of the proteome [227]. Gerstein was able to match 250 known folds to 8,000 domains in 4,500 ORFs in the worm genome [228]. In *H. pylori*, Godzik was able to recognize over 40% of the proteins encoded by that genome [229].

As all of these methods start with alignments to known structures, the results are necessarily limited to the structures that are found in the current structural database. Baker's *ab initio* proof-of-concept project [232] is interesting, in that it is a first attempt to propose an *ab initio* approach to large scale protein folding. Such methods have the advantage of not relying on known protein folds, but have the disadvantage of being limited to smaller proteins. An analysis of several genomes that have been sequenced indicates that this method could be applied to 15-25% of proteins in a number of genomes (Table **1**). These methods could be extended to more ORFs in any given genome by applying them selectively to independently folded domains within larger proteins. However, such approaches are computationally intensive and results on a genome-wide scale have not yet been published.

Table 1. Percentage of Protein ORFs Containing Less Than 150 Amino Acids

Genome	Percentage
Homo sapiens (Build 22)	16%
Drosophila melanogaster	15%
Escherichia coli	21%
Saccharomyces cerevisiae	17%
Mycobacterium tuberculosis (h37rv)	19%
Bacillus subtilis	25%
Ureaplasma urealyticum	21%
Mycoplasma genitalium	17%
Mycoplasma pneumoniaem (129)	20%
Average	19%

Using Approximate Models in Functional Site Identification and Small Molecule Screening

Are the models that come from today's state-of-the-art folding algorithms sufficient to identify functional sites, including enzyme active sites, co-factor and small molecule binding sites and site of protein interactions? Sali suggested that the functional sites in his homology models were better predicted than the remainder of the protein [227]. Fetrow and coworkers made a similar observation for a set of proteins folded by *ab initio* methods [42]. This suggests the appropriate structural descriptors might indeed be useful for function identification. Clearly, such descriptors cannot rely on the atomic detail that would be found in a typical structure determined by x-ray or NMR techniques. Structural motifs called Fuzzy Functional Forms™ (FFFs) were designed to meet this need [45]. FFFs have been shown to be very successful in identifying functional sites in genomes (Fig. **3b**) folded by today's threading and comparative modeling algorithms [43] suggesting that, indeed, these approximate models are useful, at least for biochemical function determination.

Are the approximate models good enough to be used for ligand or lead identification? As with functional site identification, today's standard methods, e.g., DOCK [22], rely on the structural quality found in experimentally determined structures. Clearly a different approach is required for identifying ligands that bind to approximate models produced by protein folding algorithms, given the lack of atomic detail. Recently, two groups have developed different types of approaches to this problem. Hoffman and coworkers have developed a method that uses the FFF fingerprint of functional sites to screen large libraries of

small molecules[2]. This is a profiles approach that does not utilize docking methodologies, so it is significantly faster computationally than standard docking methods. As such, it would be appropriate for initial very high throughput screening of large, small-molecule libraries. Skolnick and coworkers have also created a "fuzzy docking" approach, which is applicable to computational models, for the prediction of the conformation of receptor-small ligand complexes [22]. This approach uses only approximate, discretized models of both protein and small molecule, and identifies steric and quasi-chemical complementarity between a ligand and the receptor. In the application of this method to test cases from the Protein Data Bank, not only is the localization of the binding site on the receptor surface correctly identified, but also the proper orientation of the bound ligand is reasonably well reproduced within the level of accuracy of the modeled receptor itself[3]. To maximize use of the computational models that come from the large scale structural proteomics projects, these types of approaches to small molecule screening and docking must be further developed.

VI. OUTLOOK FOR FUTURE PROGRESS

These methodologies for protein structure prediction, while partially successful, need further improvement. An outstanding problem is the need to develop potentials where the native state is the lowest in energy. At present, the native state is often one of the low energy answers (in itself major progress), but in general, it is not the best-scoring solution. Similarly, global conformational search schemes need improvement so larger proteins can be treated and the problem of quaternary structure prediction addressed. There have been few studies that predict quaternary structure from sequence alone; this problem too must be addressed.

The future will likely see a confluence of homology modeling, threading and *ab initio* approaches. Already, the methods of Baker [177] and Skolnick [197, 198] are proceeding along this direction. Such methods must be applied and tested on a genomic scale to demonstrate their applicability. Homology modeling [53] and threading methods [18] have been applied to screen entire genomes, and *ab initio* folding approaches are not far behind.

Turning to threading, better integration of sequence- and structure-based approaches needs to be developed. The problem of alignment quality, even when the topology is correctly identified, is not yet solved. For tractability, perhaps an intermediate step where explicit interactions are considered but the backbone is frozen might be undertaken once a significant scoring template is identified. Since the resulting starting structure will be closer to the native state, this might increase the convergence of generalized comparative modeling methods [21]. Efforts to accomplish this intermediate refinement step are underway.

For *ab initio* folding, a better means of fold selection is needed. As mentioned above, quite often the folding simulations produce a fraction of very good, low-to-moderate resolution structures. Unfortunately, the force field does not always recognize these as being lowest in energy. Perhaps fold generation and selection should be done separately. One promising possibility is to generate the structures using a reduced model for computational tractability and then to select the resulting structures using a detailed atomic model [197].

Does more compute power solve the problem? Researchers at IBM believe that significantly more computing power could help solve the protein folding enigma and have begun working on Blue Gene, a massively parallel computing machine that could be applied to the study of the protein folding problem [233]. This ambitious project requires the development of machine architecture, programming models, and algorithmic techniques customized to protein folding by researchers at IBM. Only after this development can the machine's compute cycles be applied to studying the protein folding algorithm. Blue Gene will allow simulations that are orders of magnitude larger than can be accomplished with current technology. Such an advance would allow researchers to explore protein folding over significant time intervals, calculating large numbers of trajectories, with different atomic potentials. The increased computational power will allow researchers to collect meaningful statistics, a goal that is limited with today's available compute power. Even if Blue Gene does not "solve" the protein folding problem, data obtained from the effort will be quite useful to further understanding of the problems and issues surrounding this problem.

Sparse experimental data could be used to extend the range of applicability of threading and *ab initio* folding. Some examples of NMR-based experimental data include secondary structure, sparse tertiary restraints, and residual dipolar coupling information. Fluorescence data, double mutagenesis data or crosslinking experiments could also provide some information about side chain contacts. Mutation experiments or NMR methods can help identify residues that are involved with ligand binding. Information about the spatial arrangement of these residues could be easily incorporated into the folding algorithm. Alternatively, since *ab initio* folding often provides a few folds, experiments could be designed to select among these few possible structures. Such methods are currently being implemented in our laboratories.

The large scale prediction of the biochemical function of a protein using a structure-based approach requires an extensive active site library. Once available, the assignment of biochemical function can be done with a far smaller false positive rate than alternative sequence-based approaches [42, 234]. While three-dimensional active site descriptors can be built by hand, this is very time consuming, and automated approaches are needed. Among these is the use of PDB descriptors to assign active site residues [235]. Alternatively, BLOCKS [236] or Pfam could be used to identify conserved positions and then build a three-dimensional descriptor [237]. Other functional sites, such as ligand co-factor binding sites, might be built in the same way.

[2]B. Hoffman, unpublished results.

[3]J. Skolnick and coworkers, unpublished results.

To date, no large-scale refinement of the threading-generated structures has been done. If the alignment is in error with active site residues incorrectly aligned, then a false negative will result. Thus, GEM [21] is being applied to demonstrate the stability of correct alignments (i.e. true positives do not become false negatives) as well as to convert false negatives into true positives. The method is being tested on the weakly significant alignments (Z score >1) in both *M. genitalium* [89] and *E. coli* [93]. If the Fischer data set is a guide, this will provide a set of better models for a significant fraction of both genomes and allow the biochemical function of additional proteins to be assigned.

In conclusion, while techniques for the prediction of low-resolution structures have improved, structure prediction is not yet routine. Nevertheless, the low-resolution structures produced by contemporary algorithms are of considerable utility both in the identification of biochemical function and in ligand docking. Such efforts will have to be applied on a genomic scale if structure-based approaches to function prediction are to play a significant role in the post genomic era. A number of such efforts are underway and undoubtedly there will be more in the future. Thus, while the protein folding problem is not yet solved, it is becoming less of an enigma, and more of a practical approach to genomic scale, structure-based function annotation.

ACKNOWLEDGEMENTS

We wish to thank Deepak Singh for generating GEM models and for helpful discussions on comparative modeling, and Ruth Feldblum for designing figures and significant contributions in the preparation of the manuscript. This research was supported in part (JS, AK) by grants GM 37408 and GM 48835 of the Division of General Medical Sciences of the National Institutes of Health.

REFERENCES

[1] Anfinsen, C.B., Haber, E., Sela, M. and White, F.H. (1961) *Proc. Natl. Acad. Sci. USA*, **47**, 1309-1314.

[2] Levinthal, C. (1968) *Chem.Phys.*, **65**, 44-45.

[3] Zwanzig, R., Szabo, A. and Bagchi, B. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 20-22.

[4] Creighton, T.E. (1990) *Biochem. J.*, **270**, 1-16.

[5] Chothia, C. and Lesk, A.M. (1987) *J. Mol. Biol.*, **196**(4), 901-17.

[6] Lesk, A.M. and Chothia, C. (1980) *J. Mol. Biol.*, **136**(3), 225-70.

[7] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**(3), 403-10.

[8] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucleic Acids Res.*, **25**(17), 3389-402.

[9] Skolnick, J. and Kihara, D. (2001) *Proteins*, **42**(3), 319-31.

[10] Blake, J.D. and Cohen, F.E. (2001) *J. Mol. Biol.*, **307**(2), 721-35.

[11] Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. (1998) *J. Mol. Biol.*, **284**(4), 1201-10.

[12] Bryant, S.H. (1996) *Proteins*, **26**(2), 172-85.

[13] Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G. and Gibson, T.J. (1998) *Trends Biochem. Sci.*, **23**(10), 403-5.

[14] Sanchez, R. and Sali, A. (1997) *Curr. Opin. Struct. Biol.*, **7**(2), 206-14.

[15] Makino, S., Ewing, T.J. and Kuntz, I.D. (1999) *J. Comput. Aided Mol. Des.*, **13**(5), 513-32.

[16] Ewing, T.J., Makino, S., Skillman, A.G. and Kuntz, I.D. (2001) *J. Comput. Aided Mol. Des.*, **15**(5), 411-28.

[17] Rost, B. (1999) *Protein Eng.*, **12**(2), 85-94.

[18] Jones, D.T. (1999) *J. Mol. Biol.*, **287**(4), 797-815.

[19] Panchenko, A.R., Marchler-Bauer, A. and Bryant, S.H. (2000) *J. Mol. Biol.*, **296**(5), 1319-31.

[20] Wilmanns, M. and Eisenberg, D. (1995) *Protein Eng.*, **8**(7), 627-39.

[21] Kolinski, A., Betancourt, M., Kihara, D., Rotkiewicz, P. and Skolnick, J. (2001) *Proteins*, **44**(2), 133-149.

[22] Wojciechowski, M. and Skolnick, J. (2002) *J. Comput. Chem.*, **23**(1), 189-197.

[23] Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F. and Sali, A. (2000) *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291-325.

[24] Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L. (1987) *Protein Eng.*, **1**(5), 377-84.

[25] Bassolino-Klimas, D. and Bruccoleri, R.E. (1992) *Proteins*, **14**(4), 465-74.

[26] Claessens, M., Van Cutsem, E., Lasters, I. and Wodak, S. (1989) *Protein Eng.*, **2**(5), 335-45.

[27] Holm, L. and Sander, C. (1991) *J. Mol. Biol.*, **218**(1), 183-94.

[28] van Gelder, C.W., Leusen, F.J., Leunissen, J.A. and Noordik, J.H. (1994) *Proteins*, **18**(2), 174-85.

[29] Havel, T.F. and Snow, M.E. (1991) *J. Mol. Biol.*, **217**(1), 1-7.

[30] Bruccoleri, R.E. and Karplus, M. (1987) *Biopolymers*, **26**(1), 137-68.

[31] Moult , J. and James, M.N. (1986) *Proteins*, **1**(2), 146-63.

[32] Galaktionov, S., Nikiforovich, G.V. and Marshall, G.R. (2001) *Biopolymers*, **60**(2), 153-68.

[33] Jones, T.A. and Thirup, S. (1986) *EMBO. J.*, **5**(4), 819-22.

[34] Deane, C.M. and Blundell, T.L. (2001) *Protein Sci.*, **10**(3), 599-612.

[35] D'Alfonso, G., Tramontano, A. and Lahm, A. (2001) *J. Struct. Biol.*, **134**(2-3), 246-56.

[36] Al-Lazikani, B., Jung, J., Xiang, Z. and Honig, B. (2001) *Curr. Opin. Chem. Biol.*, **5**(1), 51-6.

[37] Summers, N.L. and Karplus, M. (1989) *J. Mol. Biol.*, **210**(4), 785-811.

[38] Dunbrack, R.L. Jr. and Karplus, M. (1993) *J. Mol. Biol.*, **230**(2), 543-74.

[39] Laughton, C.A. (1994) *J. Mol. Biol.*, **235**(3), 1088-97.

[40] Reid, L.S. and Thornton, J.M. (1989) *Proteins*, **5**(2), 170-82.

[41] Vasquez, M. (1996) *Curr. Opin. Struct. Biol.*, **6**(2), 217-21.

[42] Skolnick, J. and Fetrow, J. (2000) *TIBTECH*, **18**, 34-39.

[43] Fetrow, J.S., Godzik, A. and Skolnick, J. (1998) *J. Mol. Biol.*, **282**(4), 703-11.

[44] Fetrow, J.S., Siew, N. and Skolnick, J. (1999) *FASEB J.*, **13**(13), 1866-74.

[45] Fetrow, J.S. and Skolnick, J. (1998) *J. Mol. Biol.*, **281**(5), 949-68.

[46] Zhang, L. and Skolnick, J. (1998) *Protein Sci.*, **7**(5), 1201-7.

[47] Zhang, L., Godzik, A., Skolnick, J. and Fetrow, J.S. (1998) *Fold. Des.*, **3**(6), 535-48.

[48] Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Crystallogr.*, **26**, 283-91.

[49] Gregoret, L.M. and Cohen, F.E. (1991) *J. Mol. Biol.*, **219**(1), 109-22.

[50] Bryant, S.H. and Amzel, L.M. (1987) *Int. J. Pept. Protein Res.*, **29**(1), 46-52.

[51] Koehl, P. and Delarue, M. (1994) *Proteins*, **20**(3), 264-78.

[52] Bryant, S.H. and Lawrence, C.E. (1991) *Proteins*, **9**(2), 108-19.

[53] Sanchez, R., Pieper, U., Mirkovic, N., de Bakker, P.I., Wittenstein, E. and Sali, A. (2000) *Nucleic Acids Res.*, **28**(1), 250-3.

[54] Sali, A. and Blundell, T.L. (1993) *J. Mol. Biol.*, **234**(3), 779-815.

[55] Sali, A., Potterton, L., Yuan, F., van Vlijmen, H. and Karplus, M. (1995) *Proteins*, **23**(3), 318-26.

[56] Govindarajan, S., Recabarren, R. and Goldstein, R.A. (1999) *Proteins*, **35**(4), 408-14.

[57] Vitkup, D., Melamud, E., Moult, J. and Sander, C. (2001) *Nat. Struct. Biol.*, **8**(6), 559-66.

[58] Lathrop, R. and Smith, T.F. (1996) *J. Mol. Biol.*, **255**, 641-665.

[59] Miller, R.T., Jones, D.T. and Thornton, J.M. (1996) *FASEB J.*, **10**, 171-178.

[60] Wilmanns, M. and Eisenberg, D. (1993) *Proc. Natl. Acad. Sci. USA*, **90**(4), 1379-83.

[61] Fischer, D., Elofsson, A., Rice, D. and Eisenberg, D. (1996) *Pac. Symp. Biocomput.*, 300-18.

[62] Matsuo, Y. and Nishikawa, K. (1994) *FEBS. Lett.*, **345**(1), 23-6.

[63] Yi, T.-M. and Lander, E.S. (1994) *Protein Sci.*, **3**, 1315-1328.

[64] Jones, D.T. (1999) *J. Mol. Biol.*, **292**(2), 195-202.

[65] Koretke, K.K., Russell, R.B., Copley, R.R. and Lupas, A.N. (1999) *Proteins*, Suppl. (3), 141-8.

[66] Maiorov, V.N. and Crippen, G.M. (1992) *J. Mol. Biol.*, **227**(3), 876-88.

[67] Tropsha, A., Singh, R.K., Vaisman, I.I. and Zheng, W. (1996) Statistical geometry analysis of proteins: Implications for inverted structure prediction, in Pacific Symposium on Biocomputing '96, Hunter, L. and Klein, T.E. Editors. World Scientific: Singapore. p. 614-623.

[68] Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) *Nature*, **358**, 86-89.

[69] Koretke, K.K., Luthey-Schulten, Z. and Wolynes, P.G. (1996) *Protein Sci.*, **5**, 1043-1059.

[70] Bryant, S.H. and Lawrence, C.E. (1993) *Proteins*, **16**, 92-112.

[71] Godzik, A., Skolnick, J. and Kolinski, A. (1992) *J. Mol. Biol.*, **227**, 227-238.

[72] Selbig, J. (1995) *Protein Eng.*, **8**, 339-351.

[73] Sippl, M.J. and Weitckus, S. (1992) *Proteins*, **13**, 258-271.

[74] Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.*, **48**(3), 443-53.

[75] Bowie, J.U., Luthy, R. and Eisenberg, D. (1991) *Science*, **253**(5016), 164-70.

[76] Thiele, R., Zimmer, R. and Lengauer, T. (1995) *ISMB*, **3**, 384-392.

[77] Moult, J., Hubbard, T., Fidelis, K. and Pedersen, J.T. (1999) *Proteins*, Suppl. (3), 2-6.

[78] Kolinski, A., Rotkiewicz, P., Ilkowski, B. and Skolnick, J. (1999) *Proteins*, **37**, 592-610.

[79] Murzin, A.G. (1999) *Proteins*, **37**(S3), 88-103.

[80] Jones, D.T., Tress, M., Bryson, K. and Hadley, C. (1999) *Proteins*, Suppl. (3), 104-111.

[81] Ota, M., Kawabata, T., Kinjo, A.R. and Nishikawa, K. (1999) *Proteins*, Suppl. (3), 126-32.

[82] Domingues, F.S., Koppensteiner, W.A., Jaritz, M., Prlic, A., Weichenberger, C., Wiederstein, M., Floeckner, H., Lackner, P. and Sippl, M.J. (1999) *Proteins*, Suppl. (3), 112-20.

[83] Panchenko, A., Marchler-Bauer, A. and Bryant, S.H. (1999) *Proteins*, Suppl. (3), 133-40.

[84] Zemla, A., Venclovas, C., Moult, J. and Fidelis, K. (1999) *Proteins*, Suppl. (3), 22-9.

[85] Ogiwara, A., Uchiyama, I., Takagi, T. and Kanehisa, M. (1996) *Protein Sci.*, **5**, 1991-1999.

[86] Ouzounis, C., Sander, C., Scharf, M. and Schneider, R. (1993) *J. Mol. Biol.*, **232**, 805-825.

[87] Altschul, S.F. and Koonin, E.V. (1998) *Trends Biochem. Sci.*, **23**(11), 444-7.

[88] Jaroszewski, L., Rychlewski, L., Zhang, B. and Godzik, A. (1998) *Protein Sci.*, **7**, 1431-1440.

[89] Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J., Dougherty, B.A., Bott, K.F., Hu, P. and Lucier, C.T. (1995) *Science*, **270**(5235), 397-403.

[90] Fischer, D. and Eisenberg, D. (1997) *Proc. Natl. Acad. Sci. USA*, **94**(22), 11929-34.

[91] Teichmann, S.A., Chothia, C. and Gerstein, M. (1999) *Curr. Opin. Struct. Biol.*, **9**(3), 390-9.

[92] Gerstein, M. (1998) *Proteins*, **33**(4), 518-34.

[93] Blattner, F.R., Plunkett, G. 3rd., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) *Science*, **277**(5331), 1453-74.

[94] Koppensteiner, W.A., Lackner, P., Wiederstein, M. and Sippl, M.J. (2000) *J. Mol. Biol.*, **296**(4), 1139-52.

[95] Karplus, M. and Petsko, G.A. (1990) *Nature*, **347**, 631-639.

[96] McCammon, J.A. (1984) *Rep. Prog. Phys.*, **47**, 1-46.

[97] Brooks, C.L.I., Karplus, M. and Pettitt, B.M. (1988) *Adv. Chem. Phys.*, **71**, 1-259.

[98] Brooks, C.L.I. (1993) *Curr. Opinion Struct. Biol.*, **3**, 92-98.

[99] DeWitte, R.S. and Shakhnovich, E.I. (1994) *Protein Sci.*, **3**, 1570-1581.

[100] Monge, A., Lathrop, E.J.P., Gunn, J.R., Shenkin, P.S. and Friesner, R.A. (1995) *J. Mol. Biol.*, **247**, 995-1012.

[101] Levitt, M. and Warshel, A. (1975) *Nature*, **253**(27), 694-698.

[102] Levitt, M. (1976) *J. Mol. Biol.*, **104**, 59-107.

[103] Wallqvist, A. and Ullner, M. (1994) *Proteins*, **18**, 267-289.

[104] Anfinsen, C.B. (1973) *Science*, **181**, 223-230.

[105] Kuntz, I.D. (1975) *J. Am. Chem. Soc.*, **97**, 4362-4366.

[106] Kuntz, I.D., Crippen, G.M., Kollman, P.A. and Kimelman, D. (1976) *J. Mol. Biol.*, **106**, 983-994.

[107] Hagler, A.T. and Honig, B. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 554-558.

[108] Wilson, C. and Doniach, S. (1989) *Proteins*, **6**, 193-209.

[109] Sun, S. (1993) *Protein Sci.*, **2**, 762-785.

[110] Pedersen, J.T. and Moult, J. (1997) *Proteins*, Suppl. (1), 179-184.

[111] Osguthorpe, D.J. (1999)*Proteins*, Suppl. (3), 186-193.

[112] Lee, J., Liwo, A., Ripoll, D.R., Pilardy, J. and Scheraga, H.A. (1999) *Proteins*, Suppl. (3), 204-208.

[113] Liwo, A., Kazimierkiewicz, R., Czaplewski, C., Groth, M., Oldziej, S., Wawak, R.J., Rackovsky, S., Pinkus, M.R. and Scheraga, H.A. (1988) *J. Comput. Chem.*, **19**, 259-276.

[114] Lee, J., Liwo, A. and Scheraga, H.A. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 2025-2030.

[115] Ripoll, D.R. and Scheraga, H.A. (1990) *Biopolymers*, **30**, 165-176.

[116] Go, N., Abe, H., Mizuno, H. and Taketomi, H. (1980) *Protein folding*. ed. Jaenicke, N. Elsevier/North Holland: Amsterdam. 167-181.

[117] Go, N. and Taketomi, H. (1978)*Proc. Natl. Acad. Sci. USA*, **75**, 559-563.

[118] Ueda, Y., Taketomi, H. and Go, N. (1978) *Biopolymers*, **17**, 1531-1548.

[119] Krigbaum, W.R. and Komoriya, A. (1979) *Biochim. Biophys. Acta*, **576**, 204-246.

[120] Krigbaum, W.R. and Lin, S.F. (1982) *Macromolecules*, **15**, 1135-1145.

[121] Skolnick, J., Kolinski, A. and Yaris, R. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 5057.

[122] Skolnick, J. and Kolinski, A. (1989) *Annu. Rev. Phys. Chem.*, **40**, 207-235.

[123] Skolnick, J., Kolinski, A. and Yaris, R. (1989) *Biopolymers*, **28**, 1059-1095.

[124] Skolnick, J., Kolinski, A. and Yaris, R. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 1229-1233.

[125] Skolnick, J. and Kolinski, A. (1990) *J. Mol. Biol.*, **212**, 787-817.

[126] Skolnick, J. and Kolinski, A. (1990) *Science*, **250**, 1121-1125.

[127] Skolnick, J. and Kolinski, A. (1991) *J. Mol. Biol.*, **221**, 499-531.

[128] Sikorski, A. and Skolnick, J. (1989) *Biopolymers*, **28**, 1097-1113.

[129] Sikorski, A. and Skolnick, J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 2668-2672.

[130] Sikorski, A. and Skolnick, J. (1990) *J. Mol. Biol.*, **215**, 183-198.

[131] Sikorski, A. and Skolnick, J. (1990) *J. Mol.Biol.*, **212**, 819-836.

[132] Chan, H.S. and Dill, K.A. (1989) *Macromolecules*, **22**, 4559-4573.

[133] Chan, H.S. and Dill, K.A. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 6388-6392.

[134] Chan, H.S. and Dill, K.A. (1991) *Annu. Rev. Biophys. Biophys. Chem.*, **20**, 447-490.

[135] Dill, K.A. (1993) *Current Biol.*, **3**, 99-103.

[136] Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. and Chan, H.S. (1995) *Protein Sci.*, **4**, 561-602.

[137] Sali, A., Shakhnovich, E. and Karplus, M. (1994) *J. Mol. Biol.*, **235**, 1614-1636.

[138] Sali, A., Shakhnovich, E. and Karplus, M. (1994) *Nature*, **369**, 248-251.

[139] Shakhnovich, E.I. and Finkelstein, A.V. (1989) *Biopolymers*, **28**, 1667-1680.

[140] Shakhnovich, E.I. and Gutin, A.M. (1989) *Biophys. Chem.*, **34**, 187-199.

[141] Shakhnovich, E., Farztdinov, G. and Gutin, A.M. (1991) *Phys. Rev. Lett.*, **67**, 1665-1668.

[142] Shakhnovich, E.I. and Gutin, A.M. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 7195-7199.

[143] Shakhnovich, E.I. and Gutin, A.M. (1993) *Protein Engng.*, **6**, 793-800.

[144] Shakhnovich, E.I. (1994) *Phys. Rev. Lett.*, **72**, 3907-3910.

[145] Shakhnovich, E.I. (1996) *Fold. Des.*, **1**, R50-R54.

[146] Dinner, A.R., Sali, A. and Karplus, M. (1996) *Proc. Natl. Acad. Sci. USA*, **93**(16), 8356-61.

[147] Brower, R.C., Vasmatiz, G., Silverman, M. and Delisi, C. (1993) *Biopolymers*, **33**, 329-334.

[148] Sun, Z., Xia, X., Guo, Q. and Xu, D. (1999) *J. Protein Chem.*, **18**, 39-46.

[149] Finkelstein, A.V. and Reva, B.A. (1991) *Nature*, **351**, 497-499.

[150] Levitt, M. (1991) *Current Opinion Struct. Biol.*, **1**, 224-229.

[151] Dashevskii, V.G. (1980) *Molekulyarnaya Biologiya* (Translation from), **14**(1), 105-117.

[152] Covell, D.G. (1992) *Proteins*, **14**, 409-420.

[153] Covell, D.G. and Jernigan, R.L. (1990) *Biochemistry*, **29**, 3287-3294.

[154] Hinds, D.A. and Levitt, M. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 2536-2540.

[155] Hinds, D. and Levitt, M. (1994) *J. Mol. Biol.*, **243**, 668-682.

[156] Kolinski, A. and Skolnick, J. (1986) *Proc. Natl. Acad. Sci., USA*, **83**, 7267-7271.

[157] Kolinski, A., Skolnick, J. and Yaris, R. (1986) *J. Chem. Phys.*, **85**, 3585.

[158] Kolinski, A., Skolnick, J. and Yaris, R. (1987) *Biopolymers*, **26**, 937-962.

[159] Kolinski, A., Milik, M. and Skolnick, J. (1991) *J. Chem. Phys.*, **94**, 3978-3985.

[160] Kolinski, A. and Skolnick, J. (1992) *J. Phys. Chem.*, **97**, 9412-9426.

[161] Kolinski, A. and Skolnick, J. (1998) *Proteins*, **32**(4) 475-94.

[162] Kolinski, A., Godzik, A. and Skolnick, J., A (1993) *J. Chem. Phys.*, **98**, 7420-7433.

[163] Kolinski, A. and Skolnick, J. (1994) *Proteins*, **18**, 353-366.

[164] Kolinski, A. and Skolnick, J. (1994) *Proteins*, **18**, 338-352.

[165] Kolinski, A., Milik, M., Rycombel, J. and Skolnick, J. (1995) *J. Chem. Phys.*, **103**, 4312-4323.

[166] Kolinski, A. and Skolnick, J. (1996) *Lattice models of protein folding, dynamics and thermodynamics*. Austin, TX.: R. G. Landes. 200.

[167] Kolinski, A., Galazka, W. and Skolnick, J. (1996) *Proteins*, **26**, 271-287.

[168] Kolinski, A. and Madziar, P. (1997) *Biopolymers*, **42**, 537-548.

[169] Kolinski, A. and Skolnick, J. (1998) *Acta Biochimica Polonica*, **44**, 389-422.

[170] Kolinski, A., Rotkiewicz, P. and Skolnick, J. (1998) Application of high coordination lattice model in protein structure prediction, in Monte Carlo approaches to biopolymers and protein folding, Grassberger, P., Barkema, G.T. and Nadler, W. Editors. World Scientific: Singapore. p. 110-130.

[171] Kolinski, A. and Skolnick, J. (1998) *Proteins*, **32**, 475-494.

[172] Kolinski, A., Galazka, W. and Skolnick, J. (1998) *J. Chem. Phys.*, **108**, 2608-2617.

[173] Skolnick, J., Kolinski, A., Brooks III, C.L., Godzik, A. and Rey, A. (1993) *Curr. Biol.*, **3**, 414-423.

[174] Skolnick, J. and Kolinski, A. (1996) Monte Carlo lattice dynamics and the prediction of protein folds, in computer simulations of biomolecular systems. Theoretical and experimental studies., van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. Editors. ESCOM Science Publ.

[175] Vieth, M., Kolinski, A., Brooks III, C.L. and Skolnick, J. (1994) *J. Mol. Biol.*, **237**, 361-367.

[176] Vieth, M., Kolinski, A., Brooks III, C. L. and Skolnick, J. (1995) *J. Mol. Biol.*, **251**, 448-467.

[177] Simons, K.T., Bonneau, R., Ruczinski, I. and Baker, D. (1999) *Proteins*, Suppl. (3), 171-176.

[178] Rost, B. and Sander, C. (1993) *J. Mol. Biol.*, **232**, 584-599.

[179] Rost, B. and Sander, C. (1994) *Proteins*, **19**, 55-72.

[180] Rost, B. and Sander, C. (1996) *Proteins*, **23**, 295-300.

[181] Baker, D. (2000) *Nature*, **405**, 39-42.

[182] Ortiz, A.R., Kolinski, A. and Skolnick, J. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 1020-1025.

[183] Ortiz, A.R., Kolinski, A. and Skolnick, J. (1998) *J. Mol. Biol.*, **277**, 419-448.

[184] Ortiz, A.R., Kolinski, A. and Skolnick, J. (1998) *Proteins*, **30**, 287-294.

[185] Ortiz, A.R., Kolinski, A., Rotkiewicz, P., Ilkowski, B. and Skolnick, J. (1999) *Proteins*, Suppl. (3), 177-185.

[186] Gobel, U., Sander, C., Schneider, R. and Valencia, A. (1994) *Proteins*, **18**, 309-314.

[187] Thomas, D.J., Casari, G. and Sander, C. (1996) *Protein Engng.*, **9**, 941-948.

[188] Samudrala, R., Xia, H., Huang, E. and Levitt, M. (1999) *Proteins*, Suppl. (3), 194-198.

[189] Skolnick, J. and Kolinski, A. (2002) *Adv. Chem. Phys.*, **120**, 131-192

[190] Skolnick, J., Kolinski, A. and Ortiz, A.R. (2000) *Proteins*, **38**, 3-16.

[191] Kolinski, A., Jaroszewski, L., Rotkiewicz, P. and Skolnick, J. (1998) *J. Phys. Chem.*, **102**, 4628-4637.

[192] Kolinski, A., Rotkiewicz, P., Ilkowski, B. and Skolnick, J. (2000) *Progress of Theoretical Physics* (Kyoto), Suppl. (138), 292-300.

[193] Swendsen, R.H. and Wang, J.S. (1986) *Phys. Rev. Lett.*, **57**, 2607-2609.

[194] Betancourt, M. and Skolnick, J. (2001) *Biopolymers*, **59**(5), 305-309.

[195] Havel, T.F. and Wuthrich, K. (1985) *J. Mol. Biol.*, **182**, 281-294.

[196] Feig, M., Rotkiewicz, P., Kolinski, A., Skolnick, J. and Brooks, C.L.I. (2000) *Proteins*, **41**, 86-97.

[197] Kihara, D., Lu, H., Kolinski, A. and Skolnick, J. (2001) *Proc. Natl. Acad. Sci. USA*, **98**, 10125-10130.

[198] Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P. and Boniecki, M. (2001) *Proteins*, Suppl. (5), 149-156.

[199] Piela, L., Kostrowicki, J. and Scheraga, H.A. (1989) *J. Phys. Chem.*, **93**, 3339-3346.

[200] Ripoll, D.R., Piela, L., Velasquez, M. and Scheraga, H.A. (1991) *Proteins*, **10**, 188-198.

[201] Scheraga, H.A. (1996) *Biophys. Chem.*, **59**, 329-339.

[202] Binder, K. (1995) Monte Carlo and molecular dynamics simulations in polymer science, New York: Oxford.

[203] Hansmann, U.H.E. and Okamoto, Y. (1993) *J. Comput. Chem.*, **14**, 1333-1338.

[204] Hansmann, U.H.E. (1997) *Chem. Phys. Lett.*, **281**, 140-150.

[205] Hansmann, U.H.E. and Okamoto, Y. (1997) *J. Comput. Chem.*, **18**, 920-933.

[206] Hansmann, U.H.E. and Okamoto, Y. (1999) *Current Opin. Struct. Biol.*, **9**, 177-181.

[207] Ferrenberg, A.M. and Swendsen, R.H. (1988*) Phys. Rev. Lett.*, **61**, 2635-2637.

[208] Ferrenberg, A.M. and Swendsen, R.H. (1989) *Phys. Rev. Lett.*, **63**, 1195-1198.

[209] Scheraga, H.A. and Hao, M.H. (1999) *Adv. Chem. Phys.*, **105**, 243-272.

[210] Dandekar, T. and Argos, P. (1992) *Protein Eng.*, **5**, 637-645.

[211] Dandekar, T. and Argos, P. (1994) *J. Mol. Biol.*, **236**, 844-861.

[212] Unger, R. and Moult, J. (1993) *J. Mol. Biol.*, **231**, 75-81.

[213] Sugita, Y. and Okamoto, Y. (1999) *Chem. Phys. Lett.*, **314**, 141-151.

[214] Baumgaertner, A. (1995) Simulation of macromolecules, in the Monte Carlo method in condensed matter physics, K. Binder, Editor. Springer: Heidelberg.

[215] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) *J. Chem. Phys*., **51**, 1087-1092.

[216] Rey, A. and Skolnick, J. (1991) *Chemical Physics*, **158**, 199-219.

[217] Ilkowski, B., Skolnick, J. and Kolinski, A. (2000) *Theory and Simulations*, **9**, 523-533.

[218] Kolinski, A., Ilkowski, B. and Skolnick, J. (1999) *Biophys. J.*, **77**, 2942-2952.

[219] Berg, B.A. and Neuhaus, T. (1991) *Phys. Rev. Lett.*, **68**, 9-12.

[220] Lee, J. (1993) *Phys. Rev. Lett.*, **71**, 211-214.

[221] Hao, M.H. and Scheraga, H.A. (1994) *J. Phys. Chem.*, **98**, 4940-4948.

[222] Hao, M.H. and Scheraga, H.A. (1994) *J. Phys. Chem.*, **98**, 9882-9893.

[223] Hao, M.H. and Scheraga, H.A. (1995) *J.Chem. Phys.*, **102**, 1334-1348.

[224] Hukushima, K. and Nemoto, K. (1996) *J. Phys. Soc. (Jap.)*, **65**, 1604-1608.

[225] Gront, D., Kolinski, A. and Skolnick, J. (2000*) J. Chem. Phys.*, **113**, 5065-5071.

[226] Gront, D., Kolinski, A. and Skolnick, J. (2001) *J. Chem. Phys*., **115**, 1569-1574.

[227] Sanchez, R. and Sali, A. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 13597-13602.

[228] Gerstein, M., Lin, J. and Hegyi, H. (2000) *Pac. Symp. Biocomput.*, 30-41.

[229] Pawlowski, K., Zhang, B., Rychlewski, L. and Godzik, A. (1999) *Proteins*, **36**(1), 20-30.

[230] Rychlewski, L., Zhang, B. and Godzik, A. (1999) *Protein Sci.*, **8**(3), 614-24.

[231] Kuroda, Y., Tani, K., Matsuo, Y. and Yokoyama, S. (2000) *Protein Sci*., **9**(12), 2313-21.

[232] Simons, K.T., Strauss, C. and Baker, D. (2001) *J. Mol. Biol.*, **306**(5), 1191-9.

[233] Allen, F., Almasi, G., Andreoni, W., Beece, D., Berne, B.J., Bright, A., Brunheroto, J., Cascaval, C., Castanos, J.,

Coteus, P., Chrumley, P., Curioni, A., Denneau, M., Donath, W., Eleftheriou, M., Fitch, B., Fleischer, B., Georgiou, C.J., Germain, R., Giampapa, M., Gresh, D., Gupta, M., Haring, R., Ho, H., Hochschild, P., Hummel, S., Jonas, T., Lieber, D., Martyna, G., Maturu, K., Moreira, J., Newns, D., Newton, M., Philhower, R., Picunto, T., Pitera, J., Pitman, M., Rand, R., Royyuru, A., Salapura, V., Sanomiya, A., Shah, R., Sham, Y., Singh, S., Snir, M., Suits, F., Sweltz, R., Swope, W.C., Vishnumurthy, N., Ward, T.J.C., Warren, H. and Zhou, R. (2001) IBM Systems Journal, in Blue Gene: A vision for protein science using a petaflop supercomputer. International Business Machines. p. 310-326.

[234] Skolnick, J., Fetrow, J.S. and Kolinski, A. (2000) *Nat. Biotechnol.*, **18**(3), 283-7.

[235] Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J.S., Skolnick, J. and Godzik, A. (1999) *Protein Sci*., **8**(5), 1104-15.

[236] Henikoff, S., Henikoff, J.G. and Pietrokovski, S. (1999) *Bioinformatics*, **15**(6), 471-9.

[237] Henikoff, J.G., Pietrokovski, S., McCallum, C.M. and Henikoff, S. (2000) *Electrophoresis*, **21**(9), 1700-6.

Copyright © 2003 EBSCO Publishing