13 *De Novo* Protein Structure Prediction

Ling-Hong Hung, Shing-Chung Ngan, and Ram Samudrala

13.1 Introduction

An unparalleled amount of sequence data is being made available from large-scale genome sequencing efforts. The data provide a shortcut to the determination of the function of a gene of interest, as long as there is an existing sequenced gene with similar sequence and of known function. This has spurred structural genomic initiatives with the goal of determining as many protein folds as possible (Brenner and Levitt, 2000; Burley, 2000; Brenner, 2001; Heinemann et al., 2001). The purpose of this is twofold: First, the structure of a gene product can often lead to direct inference of its function. Second, since the function of a protein is dependent on its structure, direct comparison of the structures of gene products can be more sensitive than the comparison of sequences of genes for detecting homology. Presently, structural determination by crystallography and NMR techniques is still slow and expensive in terms of manpower and resources, despite attempts to automate the processes. Computer structure prediction algorithms, while not providing the accuracy of the traditional techniques, are extremely quick and inexpensive and can provide useful low-resolution data for structure comparisons (Bonneau and Baker, 2001). Given the immense number of structures which the structural genomic projects are attempting to solve, there would be a considerable gain even if the computer structure prediction approach were applicable to a subset of proteins.

There are two approaches to predicting protein structure. Template-based methods identify one or more homologues on which the structure is based. *Ab initio* or *de novo* methods obtain a structure more directly from sequence, without the need for a template. *De novo* techniques are much more computationally intensive than template methods and are limited to smaller proteins ($<$100–150 residues). Template-based methods can be applied to larger proteins and are generally more accurate than *de novo* methods. However, this is only true when a template exists [$<$2000 known folds in SCOP 1.61 (Murzin et al., 1995; Andreeva et al., 2004) out of an estimated 10,000 that are possible (Koonin et al., 2002)] and can be found and properly aligned. *De novo* methods are necessary when no template exists and competitive in accuracy when templates cannot be identified or aligned with confidence. Even when a good template and alignment are found, *de novo* methods are still necessary to build the nonhomologous "loop" regions.

Perhaps more so than for other methodologies, the development of *de novo* methods has been greatly aided by the blind tests provided biannually by the Critical

Assessment of Methods for Protein Structure Prediction (CASP) (Moult et al., 1995, 1997, 2001, 2003, 2005; Moult, 1999). The diversity of proteins is extremely large and it is easy to overoptimize and obtain methods that perform well on small test sets but fail when given a new unknown target. CASP1 revealed the depth of this problem and quickly dispelled any illusions about the protein structure prediction problem being solved (Moult et al., 1995). By CASP3, however, predictors had adapted their methodologies and both lattice and fragment assembly methods began to make predictions with the correct fold for small proteins (Moult, 1999; Moult et al., 1999; Venclovas et al., 1999). The continued steady improvement in the performance of the methods can be seen in the results of the 6th iteration of CASP (*http://predictioncenter.org/casp6/*).

There are a multitude of *de novo* protein methodologies and algorithms but all of them can be viewed as search algorithms attempting to find the conformation with the global minimum folding energy. It is the size of the search space and complexity of the energy function that make the problem so very difficult. We will review the approaches to calculating folding energies, the different methods used to represent and generate conformations to efficiently search for the lowest energies, and finally the methods used to select the best minimized conformers. The object is not to enumerate the different simulation methods that exist but to review the principles that all prediction protocols share. PROTINFO (Hung and Samudrala, 2003; Hung et al., 2005), the *de novo* program developed by us, is used throughout for illustration purposes.

13.2 Methods and Algorithms

13.2.1 Energy Functions

Very accurate energies can be calculated *ab initio* for small organic and inorganic molecules using quantum-mechanical (QM) methods (Hartree, 1957; Hohenberg and Kohn, 1964). Unfortunately, proteins are much more difficult systems, not only because of the size and flexibility of the protein molecule but also because of the presence of solvent molecules. The nonuniform polar aqueous environment complicates the calculation of electrostatic energies. In addition, the largest driving force for protein folding is the hydrophobic effect (Kauzmann, 1959; Dill, 1990) which is dependent not only on solvent–protein interactions, but also on higher-order solvent–solvent interactions (Frank and Evans, 1945).

Although a complete QM treatment for a complete protein is not feasible, approximations and simplifications can be made to derive empirical physics-based energies. For example, QM calculations of simple systems give hydrogen bond geometries that are applicable to those found in proteins (Morozov et al., 2004). Electrostatic calculations can be approximated using classical point charges and modifying the dielectric constant to approximate polarizability of the protein and solvent. Lennard-Jones potentials can be used to approximate van der Waals

interactions. The first use of these functions was in molecular dynamics simulations where fast, easily calculated, and differentiable energies were required to determine the force fields. Some prototypes for these types of energies are AMBER (Weiner and Kollman, 1981), CHARMM (Brooks et al., 1983), and ENCAD (Levitt et al., 1995). Parameters for these energies have been obtained by fitting to experimental data. For perturbations around a known native conformation (Levitt, 1983a; Daggett et al., 1995) these energies perform adequately, since the electrostatic and solvent-dependent information is implicit in the initial conformation itself. In combination with experimental constraints from NMR (Levitt, 1983b; Brunger et al., 1986), these force fields give rise to accurate structures, as long as the constraints are sufficient to define the fold. However, in isolation, the weaknesses of the solvent and electrostatic modeling become important and simulations attempting to fold proteins *de novo* from physics-based energies alone perform poorly.

13.2.2 Knowledge-Based Energies

Physics-based functions, while empirical, still derive their basic formulation from an underlying physical model. In contrast, knowledge-based functions are derived from properties observed in known folded proteins which are not observed in unfolded or misfolded peptides (Moult, 1997). The bases of the knowledge-based propensities are of course physical. However, the black-box approach to weighting of physical effects has proven to be more effective than explicitly specifying the form and calculating the constants in traditional physics-based energies. Most of the current successful *de novo* techniques have some knowledge-based component.

An example of a simple heuristic energy is the hydrophobic moment (Eisenberg et al., 1982) which is analogous to the physical moment of inertia except that mass term is replaced by a measure of the hydrophobicity of the residue. Minimization of this function results in compact structures with hydrophobic residues in the center. In principle, any property that is differentially observed in folded proteins and unfolded proteins can be converted to an energy using a log-odds Bayesian formulation. HMMs, neural nets, SVMs, and trial and error have been used to find such properties (for a review see Melo et al., 2002). A particularly useful class of knowledge-based functions has been pairwise distance preferences (Jones et al., 1992; Sippl and Weitckus, 1992; Samudrala and Moult, 1998) which reflect proper packing. Examples are found in many of the top performing *de novo* methods including ROSETTA (Simons et al., 1999), FRAGFOLD (Jones, 2001), TASSER (Zhang and Skolnick, 2004c), CABS (Boniecki et al., 2003), and PROTINFO (Hung and Samudrala, 2003). Some of the recent advances in understanding and improving knowledge-based functions based on pairwise distances are described in the work of Zhou and Zhou (2002) and Zhang et al. (2004a,b). Combinations of different knowledge-based energies are used to take into account different properties of folded protein. Our PROTINFO protocol uses a combination of hydrophobic moment, an all-atom pairwise distance function, RAPDF (Samudrala and Moult, 1998) and a bad contacts function. ROSETTA's complex energy function is an amalgam of many

different physics- and knowledge-based terms (Bonneau et al., 2002; Bradley et al., 2003). FRAGFOLD uses a pairwise short-range distance potential, solvation, steric, and hydrogen bonding terms. CABS and TASSER use potentials for secondary structure propensity, hydrogen bonding, short-range pairwise correlations, and homology information in the form of distance constraints and preferred side-chain contacts. Thus, while all of the above prediction protocols employ some sort of energy functions based on pairwise distance preferences, they differ in the degree to which homology information is used. Also, the number of physics-based and knowledge-based terms employed varies considerably from one protocol to another.

In-depth discussions of the mathematical formulations, limitations, and possibilities of the physics-based and knowledge-based energy functions can be found in Chapters 2 and 3 of this text, respectively.

13.2.3 Simplified Representations

A simple Cartesian representation of a protein conformation gives rise to $3N$ degrees of freedom where N is the number of atoms. The fact that bond lengths are nearly constant allows representation of a protein by its torsional angles, reducing the dimensionality of the space threefold. Even in torsional coordinates, the size of conformational space is enormous (Levinthal, 1968), too large for even nature to search exhaustively. To find the conformation with lowest energy, algorithms used by both nature and humans need to sample the canonical conformational space selectively, quickly, and efficiently.

Reductions in the size of the conformational space can be achieved using simplified representations. Ignoring the side chains is a common simplification for both Cartesian and torsional representations. In torsional space, a further simplification is made by assuming that the peptide bond is planar, eliminating the ω angle as a parameter and leaving only the φ and ψ angles to represent the main chain. Alternatively, there can be a limited representation of side chains. This can be in the form of pseudoatoms which are the weighted average in position and/or size of several real atoms in Cartesian space. Increasing computational power has made all-(heavy)-atom representations more common. For all-atom representations in torsional space, usually the most common of the observed side-chain angle combinations (rotamers) are used to reduce computational costs. It is also not unusual to use multiple representations: a coarse representation at the beginning and a finer representation at the end after most of the computationally expensive search has been accomplished (Samudrala et al., 1999a; Kolinski et al., 2001).

13.2.4 Lattice Methods

Even with simplified structural representations, the conformational space is still very large. A straightforward method to reduce search space is to digitize the possible conformations onto a regular grid or lattice. This is typically done in Cartesian space although some attempts have been made using lattice-type methods in torsional space

(Park and Levitt, 1995). Grid geometries vary, and include tetrahedral (Hinds and Levitt, 1992) and face center cubic (Kolinski et al., 2003) geometries. For maximum speed, lattice models tend to use very simple representations containing one (e.g., Kolinski et al., 2001) or two (e.g., Levitt and Warshel, 1975) atoms, one for the main chain and one for the side chain. Because the side chain sizes and distances relative to the main chain vary with residue type, the side-chain positions are off lattice, i.e., not restricted to the grid. Lattice searches are extremely fast, allowing for the exhaustive enumeration of all possibilities for small proteins in coarser representations. This is an important advantage given the difficulty of finding the global minimum by heuristic searches due to the chaotic nature of protein folding energy landscape. Unfortunately, most proteins are too large for exhaustive enumeration of all states and heuristic searches are required. Specialized grid-based move sets are used in conjunction with standard global optimization techniques such as Monte Carlo simulated annealing (Kolinski et al., 2001) and genetic algorithms (Rabow and Scheraga, 1996) to search the lattice space for the energy minimum.

13.2.5 Fragment Assembly

Template methods use the structure of one or more fragments of closely related proteins to build the model of the target protein. Fragment assembly (Bowie and Eisenberg, 1994) is an extension of this idea except that the fragments can be smaller, come from multiple sources, and need not transfer the parent fold to the target. There are two major benefits of using fragments. The first is the transfer of homology-based information. For larger fragments, this occurs at the level of supersecondary structures (Jones and McGuffin, 2003; Zhang and Skolnick, 2004a). For example, TASSER (Zhang and Skolnick, 2004a) uses large fragments exclusively as minitemplates and uses lattice sampling to build the regions between them. For smaller fragments, secondary structure information may be conveyed whereas for the smallest fragments (typically tripeptides), the sequence match is not significant and little or no information based on evolutionary conservation is transferred. However, the second advantage of using fragments is that they are derived from real folded proteins and thus implicitly contain useful knowledge-based information, i.e., no clashes, good geometry, and good local packing. Move sets based on substitution of random fragments automatically bias the search space to conformations with good local structure.

Fragment assembly can be divided into two stages: choosing the library of fragments to use and the minimization of the conformations formed by substitutions of fragments from the library into the target conformer (see Fig. 13.1). Fold recognition techniques are used to select larger homologous fragments when available. Although they provide homology-based structural information, large fragments with a good sequence match to the target are too few in number to provide a sufficiently large library of fragments required to generate the diverse structures to efficiently sample conformational space. Smaller fragments are used as they can be more easily matched to the target sequence being replaced during move generation.

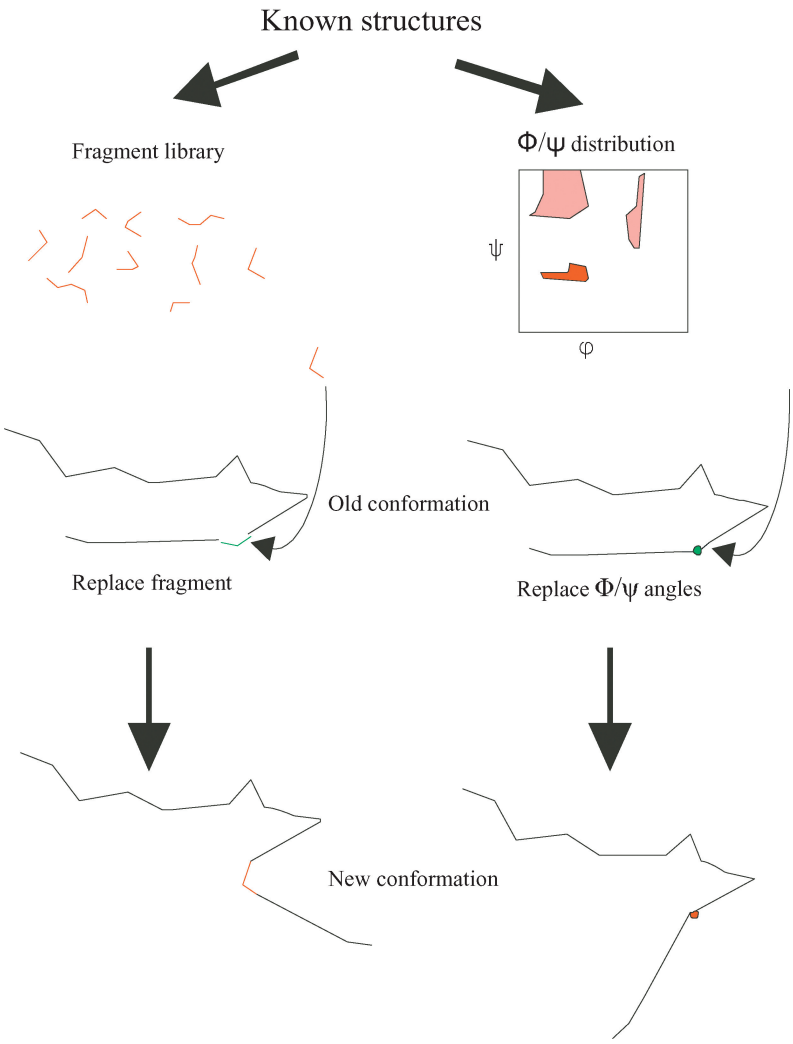

Fig. 13.1 Fragment assembly and continuous torsional methods for generating new structures to explore conformational space. Fragment assembly takes pieces from known protein structures to generate a library of fragments. By replacing part of the old conformer with a fragment from the library, a new conformation is generated. The continuous torsional moves are similar except that instead of using actual fragment coordinates or angles, a database of known structures is used to generate smoothed distributions of φ/ψ angles. These distributions in turn provide torsional angles which are substituted into a conformer to generate a new conformation. The major advantage of this method over fragment assembly is that the angles are not limited to the values in a library while still reflecting the distribution of angles observed in known proteins

For tripeptides, sufficient numbers exist in the PDB so that exact sequence matches are possible even for rare triplet sequences. Fragment libraries used for move sets are usually filtered to match the secondary structure of the target, particularly in regions where the secondary structure is known with confidence.

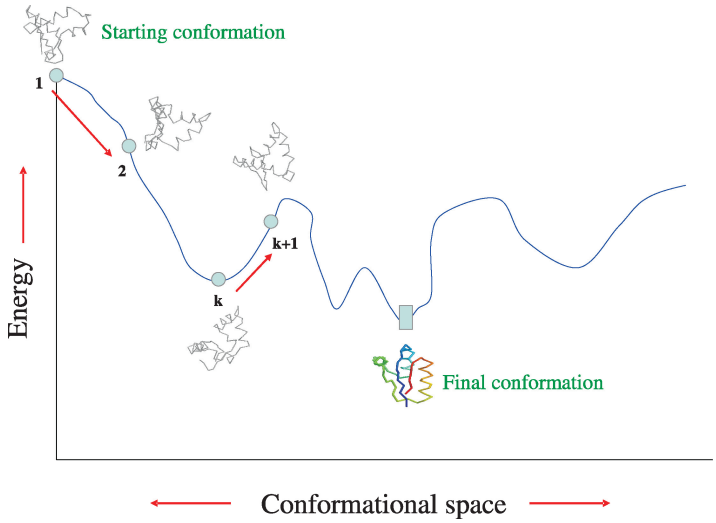

Fig. 13.2 Conformational searching using Monte Carlo simulated annealing simulation. Conformational changes or "moves" are generated as shown in Fig. 13.1 using fragment assembly or continuous torsional distributions. The energy is evaluated after each move and is compared with the previous energy, obtaining the energy difference ΔE. The move and the associated conformational change is either accepted or rejected based on Boltzmann probability $P \propto \exp(-\Delta E/kT)$, where k is Boltzmann's constant and T is the temperature. In the early stage of the simulated annealing simulation, the temperature is set high so that many energetically unfavorable (uphill) moves (for example, from step k to step k+1) can be accepted to allow the search to climb out of shallow minima. As the simulation progresses, the temperature is gradually decreased and the chance of accepting an uphill move becomes progressively lower, allowing the search to find the lowest point of the final deep minimum

Unlike lattice-based methods, complete enumeration of all states is not possible and heuristic techniques are used to search for energy minima. Because moves are based on the substitution of a finite set of fragments, there is a danger of fragment assembly searches becoming trapped in local minima. Monte Carlo simulated annealing (MCSA) with its ability to move "uphill" during the search has been used to reduce the likelihood of this happening. Figure 13.2 illustrates the essential ideas behind MCSA. Fragment assembly methods have been among the most successful of sampling techniques for *de novo* simulations with many of the best performers at CASP6, ROSETTA (Simons et al., 1999) and ROBETTA (Chivian et al., 2003), FRAGFOLD (Jones and McGuffin, 2003), Undertaker (Karplus et al., 2003), and TASSER (Zhang and Skolnick, 2004a), using some form of fragment assembly.

13.2.6 Continuous Torsional Distributions

Protein energy landscapes are very complex with multiple local minima. Exhaustive searches, heuristic search techniques such as simulated annealing/genetic algorithms, small local moves, and multiple simulations have been used to avoid being trapped in local minima. However, a problem inherent in the discrete representations

used in lattice models and fragment assembly is that the resolution of the search is limited by the resolution of the representation and move set. Searches become trapped when the discrete moves are too coarse to further explore the space around the minimum. Continuous representations avoid these problems. We have implemented a continuous version of fragment assembly. Rather than using the torsional angles of real fragments, move sets are based on continuous distributions of angles derived from known structures. Unlike earlier attempts (Lee et al., 1996), we have developed a system to mimic three-residue fragment replacement taking into account secondary structure. For each possible three-residue sequence with each possible secondary structure, a continuous basis φ/ψ angle distribution for the central residue is determined based on the observed angles in known structures. For a given target sequence, the secondary structure propensities are estimated using PsiPred (Jones, 1999). For each residue, these propensities are used to calculate a linear combination of the secondary structure-dependent basis distributions. This combined distribution is then used to bias the choice of a new set of φ/ψ angles. Rather than replacing part of a fragment, angles from the existing conformation are replaced by these new angles to make a move. The protocol is as effective as fragment assembly in cases where the structure can only be resolved coarsely. However, the advantages of a continuous representation become readily apparent when the energies are able to define the structure more precisely as shown in Fig. 13.3. The results in Fig. 13.3A are generated using the default PROTINFO energy function of RAPDF, hydrophobic moment, and bad contacts. In Fig. 13.3B, ambiguous NMR constraints have been included in the energy function.

13.2.7 Selection of the Best Conformers

Because of the problem of multiple minima, most structure prediction protocols consist of two stages: generation of a set of multiple conformers from the minimization searches and selection of the best conformer(s) from this group. Energy functions can be used to choose the best conformer. Computationally more expensive functions are available to accomplish this because the number of minimized conformations is much smaller than the number of conformations evaluated during the minimization process. However, any function related to the minimized energy will also be minimized and less informative. Orthogonal, unrelated functions are difficult to find, especially for knowledge-based functions which are the sum of many substituent effects. Nevertheless, the cumulative effect of multiple small enrichments can be effective for selection as shown in Fig. 13.4.

The energy functions used for the MCSA simulations as well as those subsequently used for conformer selection are not precise. Thus, incorrect folds will sometimes have good scores according to these approximate energy functions. However, conformers with correct folds will be similar to other conformers with correct folds. Therefore, unless there is a systematic error (e.g., an incorrect starting secondary structure), it is unlikely that multiple conformers will make the same mistake and conformers with the incorrect fold will be in general dissimilar both to those with

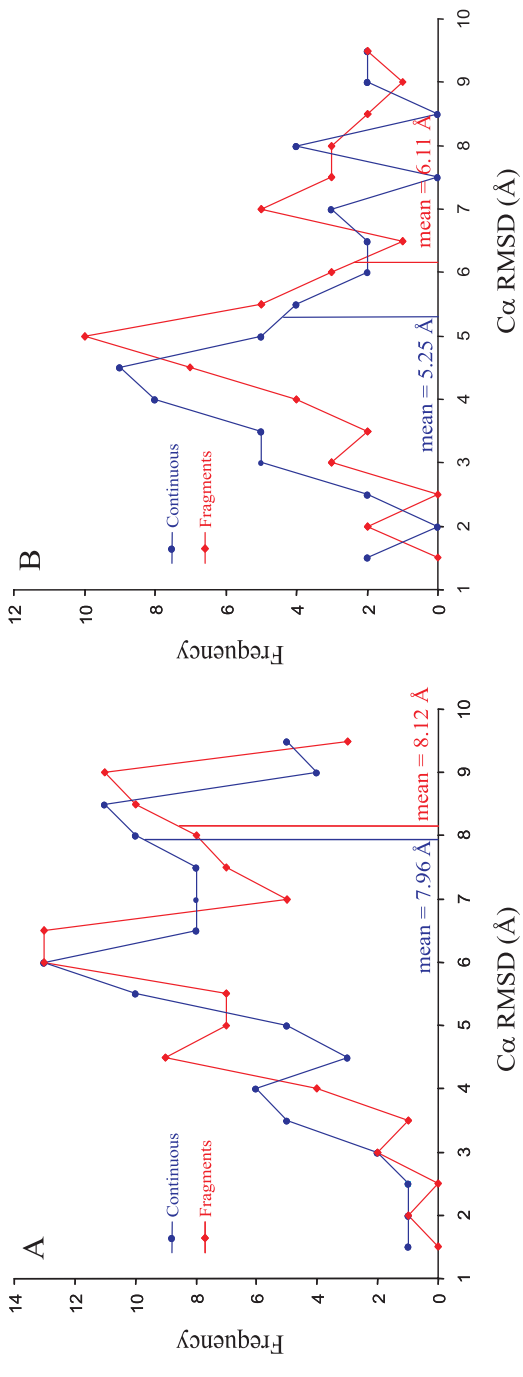

Fig. 13.3 Histograms show the Cα RMSD distributions of the best generated structures using continuous angle distributions and fragment assembly.
(A) The histogram shows the Cα RMSD distribution for a set of 129 proteins using the usual PROTINFO energy function of RAPDF, hydrophobic moment, and bad contacts. Although the overall average difference in Cα RMSD is small, the shift in the histogram is marked for cases where the energy function is able to drive the search close to the native fold.
(B) To demonstrate more clearly the superiority of the continuous method for good energy functions, the histogram shows the distribution of the Cα RMSD best generated conformers for 55 proteins where ambiguous NMR constraints have been included in the energy function. The improvement in average Cα RMSD is much larger and the difference between the Cα RMSD distributions is more distinct for the continuous method.

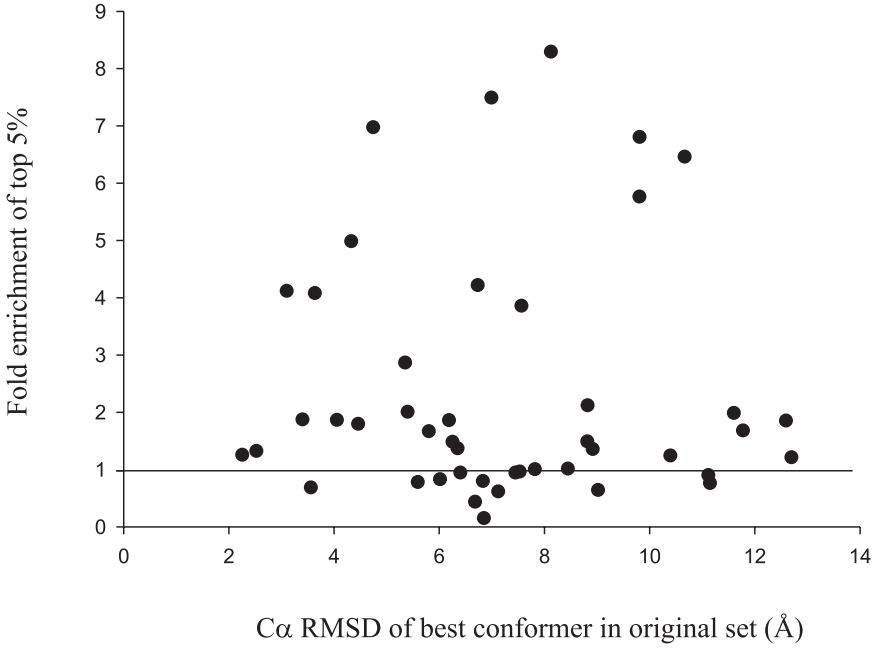

Fig. 13.4 The enrichment of conformations after selection using a set of energy functions for CASP6 targets. Fourteen different energy functions are used to iteratively rank the minimized conformers. The number of conformers retained and the order in which the energies are applied were previously optimized for a large set of different target sequences. The enrichment index is computed as follows: Let *a* be the number of the selected conformations which are in the top 10% in terms of their Cα RMSD relative to the native structures. Let *b* be the expected number in a random distribution. The enrichment index is the ratio *a*/*b*. Although the energy functions are related to the target function minimized during generation, the use of multiple functions produces significant enrichment of the top conformers in most cases

the correct fold and to each other. Thus, the conformers that are the most similar to the others, i.e., near the center of the conformational distribution, will tend to be the correct ones. The conformational center is found by using clustering methods. Metrics used include pairwise root-mean-square deviation (RMSD), pairwise RMSD with cutoffs, and number of neighbors (Simons et al., 1999; Wang et al., 2004).

When there are no systematic errors, treating the entire ensemble as a single cluster gives the best signal. Even in a relatively uniform distribution, outliers can still bias and skew the determination of the conformational center. Our iterative density protocol addresses the problem by removing outliers, recalculating the conformational center of the remaining set, and repeating until the set is well-conditioned and no outliers remain. Alternatively, patchiness due to systematic errors can be addressed by dividing the conformers into different groups using K-means or hierarchical clustering. Iterative sampling schemes can also be used to allow very large sets to be clustered quickly (Zhang and Skolnick, 2004b). When multiple clusters

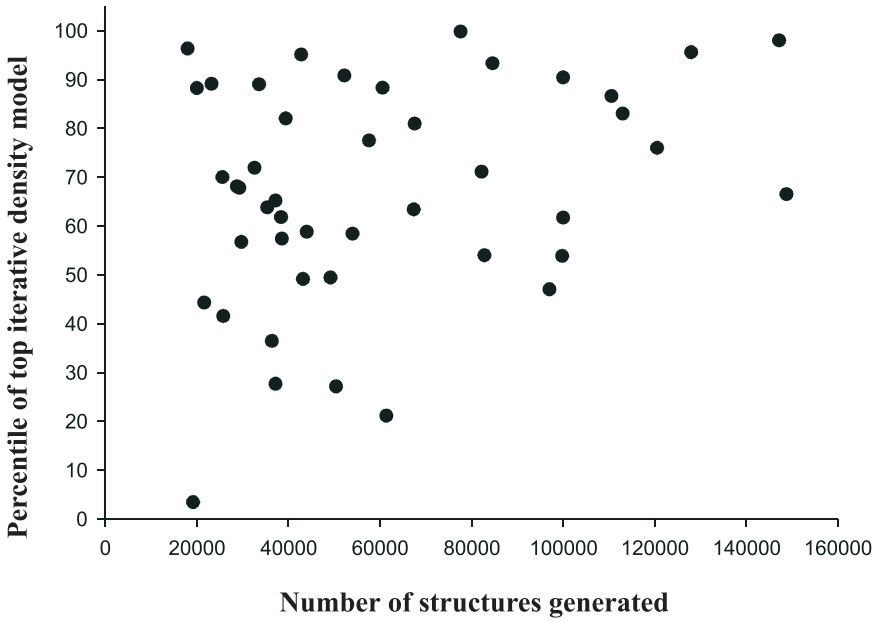

Fig. 13.5 Ranking of filtered conformations by iterative density. The percentile ranking (relative to this filtered set of conformers) for the single best conformer chosen by iterative density is shown. Even though all the conformers have been minimized for the original target function and ranked by 14 selection functions at this stage, iterative density is sufficiently orthogonal to these functions to be able to choose a top conformer. The accuracy and consistency of the selection also improves with the number of conformers generated

are used, the conformers at the center of the largest cluster or the cluster with the lowest energies are then chosen. Clustering can be used in conjunction with an initial energy-based screening of conformers. This combination is particularly effective because clustering is relatively orthogonal to energy functions. In addition, the noise reduction achieved by clustering increases with the number of conformations generated and thus automatically takes advantage of increased computational resources. This observation is illustrated in Fig. 13.5, which shows a positive correlation between the percentile ranking of conformations selected by iterative density and the number of candidate conformations generated.

13.2.8 PROTINFO, an Example *de Novo* Prediction Protocol

When given a target sequence, PROTINFO first uses PsiPred to predict the secondary structure. The secondary structure propensities are used to derive a set of φ/ψ distributions used to construct the conformations. A separate set of move likelihoods is generated which makes angle substitutions more likely in coil regions and least likely in helical regions. Move likelihoods are also biased by the standard deviation of the φ/ψ distribution. MCSA simulations are run for 10,000 iterations

using the move sets based on the angle distributions and move likelihoods. A small percentage of moves (local moves) are restricted not only by the distribution but also by the change in global RMSD. When the search nears an energy minimum, the likelihood of local moves is increased to allow the search to find the lowest point in the minimum. Brent minimization is also used, in conjunction with local moves, to pinpoint the position of the nadir. The target energy function used is a combination of hydrophobic moment to drive compactness (Eisenberg et al., 1982; Samudrala et al., 1999b), a bad-contacts function to prevent overlapping van der Waals radii, and a fast version of RAPDF (Samudrala and Moult, 1998) to ensure good packing. The fast RAPDF function has a 0.99 correlation with RAPDF but is 10–15 times faster. After 10,000 steps, the energy function is changed to use the full RAPDF for highest resolution and a further 1000 steps of the simulation is run. A flowchart illustrating the protocol is shown in Fig. 13.6.

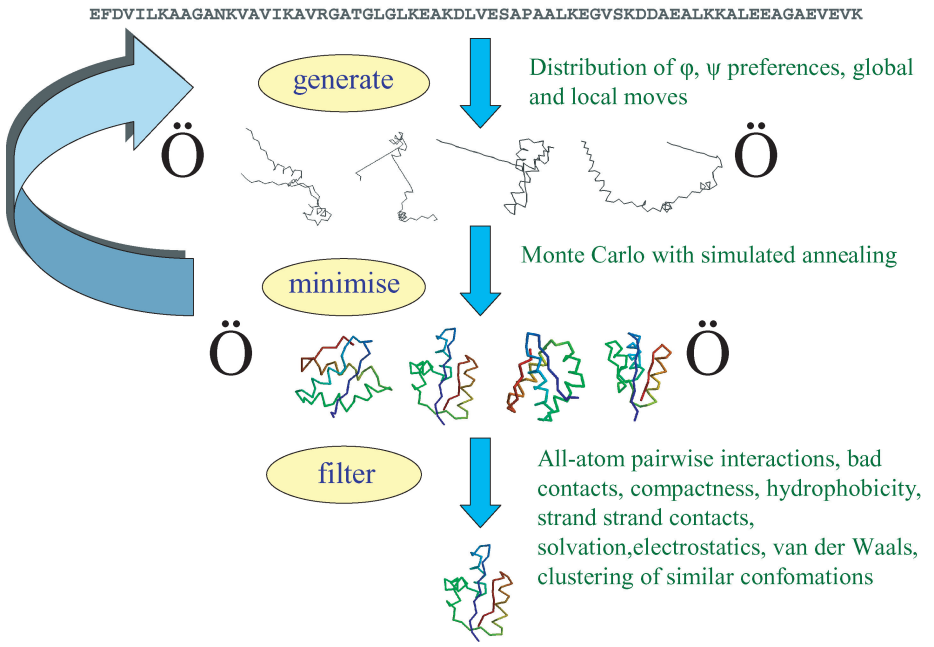

Fig. 13.6 Flowchart illustrating the PROTINFO *de novo* prediction protocol. We start with a sequence and generate main chain conformations derived from distribution of φ,ψ preferences for residues in a protein. The move sets include a combination of moves where the magnitude of change in the fold is not restricted and local moves where only small conformational changes (as measured by RMSD) are allowed. Many trajectories are generated and minimized using Monte Carlo simulated annealing. The minimization is primarily a fast version of RAPDF, a hydrophobic compactness function, and a bad contacts function. Once a set of conformations is generated, filtering is applied using many different filters and scoring functions to produce one or a few final conformations

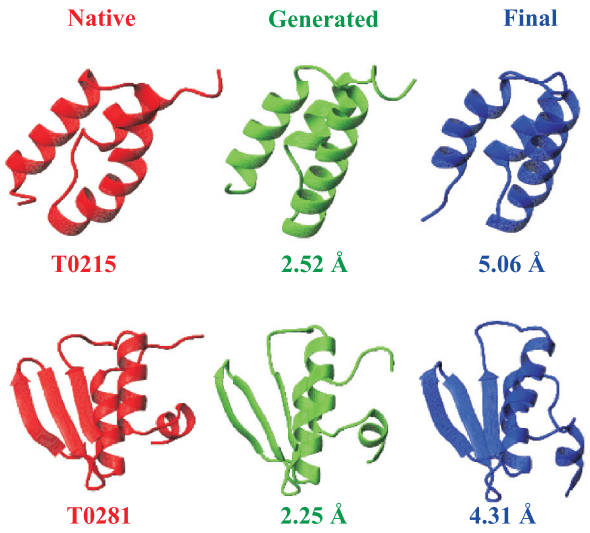

Fig. 13.7 Performance of PROTINFO on two CASP6 targets, T0215 and T0281. Cα RMSDs for the entire protein are shown for the best model generated during the simulations and the final model picked. In both cases the major structural elements are accurately placed and the major source of error is in the terminal floppy coil regions.

Selection occurs using a combination of 14 different physics-based and knowledge-based energy functions. The functions are used as filters where the best scoring conformers are retained. The cutoffs and order of application for each of the functions were determined by optimization on a large set of simulated conformers of different targets. The enrichment of this protocol for CASP6 targets is shown in Fig. 13.4. After filtering, iterative density is used to rank the remaining conformers and to choose the top three models. Two more conformers are obtained by K-means clustering and choosing the center of the two largest clusters. Figure 13.7 shows the quality of some models obtained using the procedure at CASP6.

13.2.9 Other *de Novo* Structure Prediction Protocols

Several other *de novo* structure prediction protocols were judged to perform well in the most recent meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP-6) (Moult et al. 2005). These include the Rosetta method of the Baker Group (Bradley et al., 2005a), FRAGFOLD3 of the Jones-UCL Group (Jones et al., 2005), the SAM-T04 method of Karplus and colleagues (Karplus et al., 2005), the "FRankenstein's Monster" method of the GeneSilico Group (Kosinski et al., 2005), and the CABS method of Kolinski and Bujnicki (Kolinski and Bujnicki, 2005). As mentioned earlier, a commonality among this category of methods is the emphasis on using fold-recognition techniques to generate libraries of fragments, from small fragments of fewer than five consecutive residues to large fragments up

to the level of supersecondary structures, for a given protein sequence. A variety of MCSA techniques are then used to select and assemble these fragments in a jigsawlike fashion to generate a complete conformation. In addition, in the Rosetta method, a technique based on identifying and constraining residue-pair orientation enables advances in sampling nonlocal beta-sheet structures, a difficult subproblem in *de novo* prediction. Their high-resolution refinement protocol further allows refinement of small, nativelike low-resolution structures to near-native resolution. As a result of these additional ingredients, Rosetta can produce high-resolution prediction of less than 1.5 Å for small protein domains of less than 85 residues in some cases (Bradley et al., 2005b), and is currently judged the most successful among this category of *de novo* prediction methods.

Many difficulties in consistently producing reliable *de novo* structure prediction remain. For example, prediction of large proteins and proteins of complex topologies with many nonlocal residue–residue contacts still presents major challenges. Dependency on reasonably accurate domain parsing and secondary structure prediction is the Achilles' heel of most prediction protocols. Target T0281 of CASP6 is a good example of this. This was the best modeled target of the ones that did not have obvious templates and ROSETTA was able to achieve a 1.6-Å model. The secondary structure predictions for this target were ambiguous and initially almost all groups, including the ROBETTA server, submitted some models using a two-strand topology rather than the correct three-strand topology, which resulted in models >6 Å from the experimental structure.

Finally, examples of *de novo* structure prediction methods that are more theoretical and physics-based can be found in the works of Wolynes and colleagues (Onuchic and Wolynes, 2004; Wolynes, 2005), Daggett and colleagues (Daggett and Fersht, 2003; Beck and Daggett, 2004), and Head-Gordon and colleagues (Crivelli et al., 2002; Head-Gordon and Brown, 2003). Compared with the approaches described in the previous paragraph, this category of methods encounters substantially greater challenges because they have much less reliance on fold recognition and other bioinformatic information.

13.3 Discussion

13.3.1 Faster Computers and Larger Databases

An often-stated tenet of modern computing is Moore's law which postulates that computational power increases exponentially with time (Moore, 1965). This has been paralleled by the exponential growth of the PDB (Berman et al., 2004). Both of these developments have been instrumental in improving the effectiveness of protein structure prediction techniques. Increased computational power permits all-atom representations and more numerous, longer, and finer searches. Increases in the size of the PDB increase the number of fragments available for fragment assembly libraries. Knowledge-based energies, such as RAPDF, also benefit from a larger PDB

as do secondary structure predictions. A good example of these effects is the evolution of FRAGFOLD which started as a simple fragment assembly method using a reduced representation (Jones, 2001; Jones and McGuffin, 2003). Its present incarnation uses secondary structure selected fragments with an all-atom representation and optimizes side-chain rotamers in parallel with the fragment assembly.

13.3.2 Future Directions

Of the two major components of protein structure prediction, conformational search methodology has been helped more by the computational and database technology improvements, compared to energy function methodology. The major limitation of *de novo* structure prediction techniques remains the quality of the energy functions. One area where this is much less of a problem is the marriage of *de novo* protein structure prediction techniques and limited experimental data to obtain structures. For example, in the case of NMR, with good unambiguous data, structures have long been obtained even with physics-based energies and molecular dynamics searches (Brunger et al., 1986). Hybrid methods (Rohl and Baker, 2002; Li et al., 2003) using knowledge-based energies and more efficient global search techniques promise to lower threshold of quality and quantity of NMR data required to obtain a structure. This idea of bypassing the energy function bottleneck by using limited experimental data can be applied in principle to any methodology and is becoming an important application of theoretical *de novo* techniques to practical problems.

Due to the increased size of the PDB and the ongoing structure genomics initiatives, the chance of finding a template to a given target is steadily increasing and will continue to do so. The refinement of template-derived structures is thus becoming an important problem. However, until recently (Qian et al., 2004), refinement of template-based structures using *de novo* techniques has been largely limited to building loops between homologous elements. Refining the homologous elements themselves has proven to be difficult to do consistently. The problem is a bit different from *de novo* prediction since the starting point is a relatively good conformation, making it more of a local rather than global optimization problem. Some of the global energies and search strategies used for *de novo* prediction remain effective for local optimization (e.g., RAPDF and continuous move sets). Other energies and search strategies that are not effective or feasible for the global case (e.g., physics-based energies and molecular dynamics) may work better in the local case.

Finally, the energy functions themselves are slowly improving. The increase in the size of the PDB combined with increased computational power means that more complex knowledge-based energies are now feasible. For example, rather than a simple pairwise distance function, a secondary structure-dependent distance function or a ternary distance function is now possible. Improved physics-based energies are also being developed with more sophisticated electrostatic and solvent models which, while still too computationally expensive for global structure prediction, may be useful in refinement. The protein structure prediction problem has not yet been

solved, but after 40 years in the wilderness, the promised land may finally be within sight.

Acknowledgments

This work was supported in part by a Searle Scholar Award, NSF CAREER Grant IIS-0448502, NIH Grant GM068152, and the University of Washington's Advanced Technology Initiative in Infectious Diseases.

References

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. 2004. SCOP database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acids Res*. 32:D226–D229.

Beck, D.A.C., and Daggett, V. 2004. Methods for molecular dynamics simulations of protein folding/unfolding in solution. *Methods* 34:112–120.

Berman, H.M., Bourne, P.E., and Westbrook, J. 2004. The Protein Data Bank: A case study in management of community data. *Curr*. *Proteomics* 1:49–57.

Boniecki, M., Rotkiewicz, P., Skolnick, J., and Kolinski, A. 2003. Protein fragment reconstruction using various modeling techniques. *J. Comput. Aided Mol. Des*. 17:725–738.

Bonneau, R., and Baker, D. 2001. Ab initio protein structure prediction: Progress and prospects. *Annu. Rev. Biophys. Biomol. Struct*. 30:173–189.

Bonneau, R., Strauss, C.E., Rohl, C.A., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., and Baker, D. 2002. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol*. 322:65–78.

Bowie, J.U., and Eisenberg, D. 1994. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. USA* 91:4436–4440.

Bradley, P., Chivian, D., Meiler, J., Misura, K.M., Rohl, C.A., Schief, W.R., Wedemeyer, W.J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C.E., and Baker, D. 2003. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. *Proteins* 53 (Suppl. 6):457–468.

Bradley, P., Malmstrom, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D.E., Meiler, J., Misura, K.M.S., and Baker, D. 2005a. Free modeling with Rosetta in CASP6. *Proteins* 61 (Suppl. 7):128–134.

Bradley, P., Misura, K.M.S., and Baker, D. 2005b. Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871.

Brenner, S.E. 2001. A tour of structural genomics. *Nat. Genet*. 210:801–809.

Brenner, S., and Levitt, M. 2000. Expectations from structural genomics. *Protein Sci*. 9:197–200.

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

Brunger, A.T., Clore, G.M., Gronenborn, A.M., and Karplus, M. 1986. Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: Application to crambin. *Proc. Natl. Acad. Sci. USA* 83:3801–3805.

Burley, S.K. 2000. An overview of structural genomics. *Nat. Struct. Biol.* 7 (Suppl.):932–934.

Chivian, D., Kim, D.E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C.E., Bonneau, R., Rohl, C.A., and Baker, D. 2003. Automated prediction of CASP-5 structures using the Robetta server. *Proteins* 53 (Suppl. 6):524–533.

Crivelli, S., Eskow, E., Bader, B., Lamberti, V., Byrd, R., Schnabel, R., and Head-Gordon, T. 2002. A physical approach to protein structure prediction. *Biophys. J.* 82:36–49.

Daggett, V., and Fersht, A.R. 2003. The present view of the mechanism of protein folding. *Nat. Rev. Mol. Cell Biol.* 4:497–502.

Daggett, L.P., Sacaan, A.I., Akong. M., Rao, S.P., Hess, S.D., Liaw, C., Urrutia, A., Jachec, C., Ellis, S.B., Dreessen, J., et al. 1995. Molecular and functional characterization of recombinant human metabotropic glutamate receptor subtype 5. *Neuropharmacology* 34:871–886.

Dill, K.A. 1990. Dominant forces in protein folding. *Biochemistry* 29:7133–7155.

Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. 1982. The helical hydrophobic moment: A measure of the amphiphilicity of a helix. *Nature* 299:371–374.

Frank, H.S., and Evans, M.W. 1945. Free volume and entropy in condensed systems. III. Entropy in binary liquid mixtures; partial molal entropy in dilute solutions; structure and thermodynamics in aqueous electrolytes. *J. Chem. Phys*. 13:507–532.

Hartree, D.R. 1957. *The Calculation of Atomic Structure*. New York, John Wiley & Sons.

Head-Gordon, T., and Brown, S. 2003. Minimalist models for protein folding and design. *Curr. Opin. Struct. Biol*. 13:160–167.

Heinemann, U., Illing, G., and Oschkinat, H. 2001. High-throughput three-dimensional protein structure determination. *Curr. Opin. Biotechnol*. 12:348–354.

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

Hohenberg, P., and Kohn, W. 1964. Inhomogeneous electron gas. *Phys. Rev*. 136:864.

Hung, L.-H., Ngan, S.-C., Liu, T., and Samudrala, R. 2005. PROTINFO: New algorithms for enhanced protein structure predictions. *Nucleic Acids Res*. 33: (in press).

Hung, L.-H., and Samudrala, R. 2003. PROTINFO: Secondary and tertiary protein structure prediction. *Nucleic Acids Res*. 31:3296–3299.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*. 292:195–202.

Jones, D.T. 2001. Predicting novel protein folds by using FRAGFOLD. *Proteins* Suppl. 5:127–132.

Jones, D.T., Bryson, K., Coleman, A., McGuffin, L.J., Sadowski, M.I., Sodhi, J.S., and Ward, J.J. 2005. Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins* 61 (Suppl. 7):143–151.

Jones, D.T., and McGuffin, L.J. 2003. Assembling novel protein folds from super-secondary structural fragments. *Proteins* 53 (Suppl 6):480–485.

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

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., and Hughey, R. 2003. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* 53 (Suppl 6):491–496.

Karplus, K., Katzman, S., Shackleford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M., and Hughey, R. 2005. SAM-T04: what is new in protein-structure prediction for CASP6. *Proteins* 61 (Suppl. 7):135–142.

Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem*. 14:1–64.

Kolinski, A., Betancourt, M.R., Kihara, D., Rotkiewicz, P., and Skolnick, J. 2001. Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* 44:133–149.

Kolinski, A., and Bujnicki, J.M. 2005. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins* 61 (Suppl. 7):84–90.

Kolinski, A., Gront, D., Pokarowski, P., and Skolnick, J. 2003. A simple lattice model that exhibits a protein-like cooperative all-or-none folding transition. *Biopolymers* 69:399–405.

Koonin, E.V., Wolf, Y.I., and Karev, G.P. 2002. The structure of the protein universe and genome evolution. *Nature* 420:218–223.

Kosinski, J., Gajda, M.J., Cymerman, I.A., Kurowski, M.A., Pawlowski, M., Boniecki, M., Obarska, A., Papaj, G., Sroczynska-Obuchowicz, P., Tkaczuk, K.L., Sniezynska, P., Sasin, J.M., Augustyn, A., Bujnicki, J.M., and Feder, M. 2005. FRankenstein becomes a cyborg: The automatic recombination and realignment of fold-recognition models in CASP6. *Proteins* 61 (Suppl. 7):106–113.

Lee, B., Kurochkina, N., and Kang, H.S. 1996. Protein folding by a biased Monte Carlo procedure in the dihedral angle space. *FASEB J*. 10:119–125.

Levinthal, C. 1968. Are there pathways for protein folding? *J. Chim. Phys*. 65: 44.

Levitt, M. 1983a. Molecular dynamics of native protein. I. Computer simulation of trajectories. *J. Mol. Biol*. 168:595–617.

Levitt, M. 1983b. Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol*. 170:723–764.

Levitt, M., Hirshberg, M., Sharon, R., and Daggett, V. 1995. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comput. Phys. Commun*. 91:215–231.

Levitt, M., and Warshel, A. 1975. Computer simulation of protein folding. *Nature* 253:694–698.

Li, W., Zhang, Y., Kihara, D., Huang, Y.J., Zheng, D., Montelion, G.T., Kolinski, A., and Skolnick, J. 2003. TOUCHSTONEX: Protein structure prediction with sparse NMR data. *Proteins* 53:290–306.

Melo, F., Sanchez, R., and Sali, A. 2002. Statistical potentials for fold assessment. *Protein Sci*. 11:430–448.

Moore, G.E. 1965. Cramming more components onto integrated circuits. *Electronics* 38:114–117.

Morozov, A.V., Kortemme, T., Tsemekhman, K., and Baker, D. 2004. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. USA* 101:6946–6951.

Moult, J. 1997. Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol*. 7:194–199.

Moult, J. 1999. Predicting protein three-dimensional structure. *Curr. Opin. Biotechnol.* 10:583–588.

Moult, J., Fidelis, K., Tramontano, A., Rost, B., and Hubbard, T. 2005. Critical assessment of methods of protein structure prediction (CASP)—round VI. *Proteins* (accepted preprint). **Au: update ref**

Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. 2001. Critical assessment of methods of protein structure prediction (CASP): Round IV. *Proteins* Suppl. 5:2–7.

Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. 2003. Critical assessment of methods of protein structure prediction (CASP)—round V. *Proteins* 53 (Suppl. 6):334–339.

Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K., and Pedersen, J.T. 1997. Critical assessment of methods of protein structure prediction (CASP): Round II. *Proteins* Suppl. 1:2–6.

Moult, J., Hubbard, T., Fidelis, K., and Pedersen, J.T. 1999. Critical assessment of methods of protein structure prediction (CASP): Round III. *Proteins* Suppl. 3:2–6.

Moult, J., Pedersen, J.T., Judson, R., and Fidelis, K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins* 23: ii–v.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. 247:536–540.

Onuchic, J.N., and Wolynes, P. G. 2004. Theory of protein folding. *Curr. Opin. Struct. Biol*. 14:70–75.

Park, B.H., and Levitt, M. 1995. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol*. 249:493–507.

Qian, B., Ortiz, A.R., and Baker, D. 2004. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proc. Natl. Acad. Sci. USA* 101:15346–15351.

Rabow, A.A., and Scheraga, H.A. 1996. Improved genetic algorithm for the protein folding problem by use of a Cartesian combination operator. *Protein Sci*. 5:1800–1815.

Rohl, C.A., and Baker, D. 2002. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.* 124:2723–2729.

Samudrala, R., and Moult, J. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol*. 275:895–916.

Samudrala, R., Xia, Y., Huang, E., and Levitt, M. 1999a. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins* Suppl. 3:194–198.

Samudrala, R., Xia, Y., Levitt, M., and Huang, E.S. 1999b. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pac. Symp. Biocomput*. pp. 505–516.

Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* Suppl. 3:171–176.

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

Venclovas, C., Zemla, A., Fidelis, K., and Moult, J. 1999. Some measures of comparative performance in the three CASPs. *Proteins* Suppl. 3:231–237.

Wang, K., Fain, B., Levitt, M., and Samudrala, R. 2004. Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct. Biol*. 4:8.

Weiner, P.K., and Kollman, P.A. 1981. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J. Comput. Chem*. 2:287–303.

Wolynes, P. G. 2005. Energy landscapes and solved protein folding problems. *Philos. Trans. R. Soc. London Ser. A* 363:453–464.

Zhang, C., Liu, S., Zhou, H., and Zhou, Y. 2004a. The dependence of all-atom statistical potentials on training structural database. *Biophys. J*. 86:3349–3358.

Zhang, C., Liu, S., Zhou, H., and Zhou, Y. 2004b. An accurate residue-level pair potential of mean force for folding and binding based on the distance-scaled ideal-gas reference state. *Protein Sci*. 13:400–411.

Zhang, Y., and Skolnick, J. 2004a. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA* 101:7594–7599.

Zhang, Y., and Skolnick, J. 2004b. SPICKER: A clustering approach to identify near-native protein folds. *J. Comput. Chem.* 25:865–871.

Zhang, Y., and Skolnick, J. 2004c. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophys. J*. 87:2647–2655.

Zhou, H., and Zhou, Y. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*. 11:2714–2726.