

AB INITIO PROTEIN STRUCTURE PREDICTION: Progress and Prospects

Richard Bonneau and David Baker

*Department of Biochemistry, Box 357350, University of Washington, Seattle,
Washington, 98195; e-mail: dabaker@u.washington.edu, bonneau@u.washington.edu*

Key Words structural genomics, protein folding, scoring functions, fold prediction

■ **Abstract** Considerable recent progress has been made in the field of ab initio protein structure prediction, as witnessed by the third Critical Assessment of Structure Prediction (CASP3). In spite of this progress, much work remains, for the field has yet to produce consistently reliable ab initio structure prediction protocols. In this work, we review the features of current ab initio protocols in an attempt to highlight the foundations of recent progress in the field and suggest promising directions for future work.

CONTENTS

INTRODUCTION	174
REDUCED COMPLEXITY MODELS	175
Lattice Models	176
Discrete State Off-Lattice Models	176
Narrowing the Search with Local Structure Prediction	176
SCORING FUNCTIONS FOR REDUCED COMPLEXITY MODELS	177
Solvation-Based Scores	178
Pair Interactions	178
Sequence Independent Terms/Secondary Structure Arrangement	179
The Use of Multiple Sequence Alignments in Tertiary Prediction	180
High-Resolution Structure Prediction—Prospects for Refinement	181
THE ROSETTA METHOD	181
RELATIONSHIP TO PROTEIN FOLDING?	183
APPLICATIONS OF AB INITIO STRUCTURE PREDICTION	183
Genome Annotation	183
Homology Modeling	184
Structures from Limited Constraint Sets	184
CONCLUSION	185

INTRODUCTION

The goal of *ab initio* structure prediction is simple: Given a protein's amino acid sequence predict the structure of its native state. It is generally assumed that a protein sequence folds to a native conformation or ensemble of conformations that is at or near the global free-energy minimum. Thus, the problem of finding native-like conformations for a given sequence can be decomposed into two subproblems: (a) developing an accurate potential and (b) developing an efficient protocol for searching the resultant energy landscape.

To date, the most successful methods for structure prediction have been homology-based comparative modeling and fold recognition (58). When homologous or weakly homologous sequences of known structure are not available, the most successful structure prediction methods have been those that predict secondary structure and local structure motifs; these methods have been available for some time (12, 40, 43, 46, 74). This review focuses on current methods for predicting tertiary structure in the absence of homology to a known structure and discusses these local prediction methods only in the context of tertiary structure prediction.

Many of the methods today that predict protein structure use information from the protein data bank (PDB). This information can be found in the parameters of knowledge-based scoring functions, the training sets of machine learning approaches, and the coordinate libraries of methods that use fragments or templates from the PDB. In order to test the performance of any one of these approaches, one must carefully remove sequences that are homologous to proteins in the test set from all databases used by the method in question. Any errors or oversights made at this stage could lead to overestimates of success. For these reasons, the Critical Assessment of Structure Prediction (CASP), a biannual, community-wide blind test of prediction methods, was conceived and implemented (4, 6, 27). Three such tests have occurred and the fourth was undertaken this summer (i.e., 2000) (20, 39). Throughout this review, our assessment of the performance of various methods is influenced by the results of these community-wide tests, especially CASP3 (58). CASP1 and CASP2 showed that in spite of some success reported in the literature, little success was seen for proteins of the size and type being solved by structural biologists (50). The consensus after these first two experiments was that the *ab initio* structure prediction problem would most probably be solved too late to be applied to any real biological problems (20, 50). CASP3 showed some reversal of this consensus. Several methods made good predictions in the *ab initio* category, and some *ab initio* methods outperformed fold recognition methods for certain proteins in the fold recognition category (61, 63, 64).

In spite of recent progress, many issues must still be resolved if a consistently reliable *ab initio* prediction scheme is to be developed. No one method performs consistently across all classes of proteins (most methods perform worse on all-beta proteins), and all methods examined seem to fail totally on sequences longer than 150 residues in length (the longest contiguous blind predictions to date are ~100 residues in length). In addition, the successful prediction methods, as judged

by performance at CASP3 and by performance reported in the literature, show a large diversity in their formulation. Our goal here is to review the common features of these recent methods in order to highlight current challenges and future possibilities for the field of ab initio structure prediction.

The most natural starting point for simulating protein folding is standard molecular dynamics (MD) simulation (numerically integrating Newton's equations of motion for the polypeptide chain) using a physically reasonable potential function. This approach has a long history and is still popular, as illustrated perhaps most dramatically by IBM's Blue Gene project, which will apparently be devoted, at least in part, to such an endeavor. There are several rather obvious problems that have limited the success of such approaches thus far. First, MD is computationally expensive—with explicit representation of sufficient water molecules to minimally solvate the folding chain, a nanosecond MD simulation of a 100-residue protein takes ~400 hours on a current single processor. Advances in simulation strategy and increases in available computer power have considerably extended simulation times; for example, Kollman and coworkers carried out a microsecond simulation of a 36-residue peptide using a considerable amount of supercomputer time. However, simulating the folding of a 100-residue protein for the typical ~1 second required for a single folding transition requires more than six orders of magnitude more computing time. The second, and perhaps more serious, class of problems associated with MD are the inadequacies in current potential functions for macromolecules in water. Although important progress has been made, there is still a lack of consensus as to the best computationally tractable yet physically realistic model for water (a number of quite different models for water, both polarizable and nonpolarizable, have been used in current simulation methods), and some uncertainty exists as to the values of the parameters used in molecular mechanics potentials (partial atomic charges, Lennard-Jones well depths and radii). Accurate representation of electrostatics is also a considerable challenge given the high degree of polarizability of water, the large difference in the dielectric properties of the solvent and the protein, and the uncertainties in the magnitude and location of atomic partial charges. Because the free energy of a protein represents a delicate balance of large and opposing contributions, these problems significantly reduce the likelihood that the native state will be found at the global free-energy minimum using current potentials (2). The best current use of MD methods may be in refining and discriminating among models produced by lower-resolution methods (28a, 49).

REDUCED COMPLEXITY MODELS

To overcome the sampling problems mentioned above, most methods for fold prediction to date have involved some significant complexity reduction. Methods for reducing protein structure to discrete low-complexity models can be divided into two major classes: lattice and off-lattice models.

Lattice Models

Lattice models have a long history in the modeling of polymers due to their analytical and computational simplicity (22). The evaluation of energies on a lattice can be achieved quite efficiently (integer math can be made quite fast), and methods involving exhaustive searches of the available conformational space become feasible (30, 35, 85). However, lattice methods have a somewhat restricted ability to represent subtle geometric considerations (strand twist, secondary structure propensities, and packings) and can reproduce the backbone with accuracies no greater than approximately half the lattice spacing (73). The most common systematic error observed for a variety of lattice models is their inability to reproduce helices, and most lattice models exhibit various degrees of secondary structure bias (70). Given the importance of regular secondary structure in proteins, this is clearly a problem. Recent successful tests, however, have shown that the computational advantages of lattice models may outweigh the problems associated with their systematic biases (30, 47, 48, 73).

Discrete State Off-Lattice Models

Most off-lattice reduced complexity models fix all side chain degrees of freedom and all bond lengths. The most common practice is to limit the side chain to a single rotamer, or further to the C_{β} , or to one or more centroids plus backbone atoms (82, 87). Discrete state models of the protein backbone usually fix all side chain degrees of freedom and limit the backbone to specific Phi/Psi pairs: Models containing from 4 to 32 Phi/Psi states representing various strand, helix, and loop conformations have been described in the literature (70). Properly optimized six-state models (i.e., models that account for local features observed in proteins such as strands, helices, and canonical loops) can reproduce native contacts, preserve secondary structure, and fit the overall coordinates of the native state as well as 18-state lattice models that do not account for such protein-specific information (70).

Narrowing the Search with Local Structure Prediction

Local structures excised from proteins can fold independent of the full protein, demonstrating that strong local structural biases can exist for short sequence segments (8, 15, 54, 60). Several examples of excised fragments having little observable structure also exist, indicating that the strength and multiplicity of these local biases are highly sequence dependent. Some sequences are observed to fold to different conformations depending on their global sequence context, again demonstrating the possible multiplicity of local structure biases (17, 42). Bystroff et al developed a method that recognizes sequence motifs (ISITES) with strong tendencies to adopt a single local conformation that was used to make some good local structure predictions in CASP2 (12–14, 29). Despite the ambiguities in

local sequence-structure relationships, secondary structure prediction methods have been steadily improving (63, 64).

The above mentioned experiments and observations suggest that any method attempting to use local sequence-structure biases to guide complexity reductions will have to be adaptive to the strength and to the multiplicity of different sequence-structure patterns. The majority of methods proving successful at CASP3 used secondary structure predictions in one way or another. In one case, predicted secondary structure elements were fit to the results of initial lattice-based exhaustive enumeration, thus erasing any possible secondary structure bias in the initial lattice model prior to all-atom refinement (1). The Rosetta method used secondary structure to bias the selection of fragments of known structure from the PDB. Yet another paradigm is, given a secondary structure string, to reduce the problem of predicting the tertiary fold to the problem of how to assemble rigid secondary structure elements (24). Additional methods that determine local structure biases independent of secondary structure prediction algorithms (by calculating these biases during the folding simulation) have also been described (86).

There is likely to be an upper limit on the accuracy of secondary structure prediction methods owing to their failure to account for nonlocal interactions. The best secondary structure prediction algorithms have three-state accuracies of 76%–78%, and any *ab initio* method must account for this error rate to make consistently successful predictions (40, 74). A milestone for *ab initio* structure prediction, which takes such interactions into account, will be the production of models with secondary structure predicted more accurately than is possible with traditional secondary structure methods.

SCORING FUNCTIONS FOR REDUCED COMPLEXITY MODELS

Once a model for representing the protein is chosen that sufficiently reduces the complexity of the conformational search, a scoring or energy function that works in the chosen low-complexity space must be developed. The energy function must adequately represent the forces responsible for protein structure: solvation, strand hydrogen bonding, etc. Given that most low-complexity models do not explicitly represent all atoms and can reproduce even the native state backbone with only limited accuracy, any energy function designed to work in the low-complexity regime must represent these forces in a manner robust to such systematic error (the systematic limitations of the model). Last, these functions must be computationally efficient, for during the initial stages of any conformational search, huge numbers of energy evaluations are necessary. Because of the shortcomings of molecular mechanics-based potentials, and the considerations above, many methods developed in the last ten years utilize scoring functions derived from the protein database that in essence favor arrangements of residues frequently found in known protein structures and disfavor rarely seen arrangements.

Solvation-Based Scores

It has been long thought that the hydrophobic effect is the principal driving force behind protein folding (3). Many diverse methods for judging the fitness of conformations based on solvation or hydrophobic packing exist, and the debate over the proper functional form for representing solvation effects represents an open question of considerable importance and interest (69). A common approach is to classify sites in proteins according to their degree of solvent exposure (either through the surface area or the number of nearby residues) (11) and to determine the frequencies of occurrence of the amino acids in each type of site. The energy or score of an amino acid at a site is then taken to be the logarithm of the amino acid's frequency of occurrence at that type of site. This type of residue-environment term favors placement of hydrophobic amino acids at buried positions and of hydrophilic amino acids at exposed positions.

An additional commonly used class of functional forms consists of global measures of hydrophobic arrangement. One simple global quantity is a residue's distance from the entire conformation's center of mass, which can be used to calculate quantities analogous to the hydrophobic radius of gyration. Bowie & Eisenberg used this type of function, coined hydrophobic contrast, in combination with other terms, including a surface area-based term, to fold small alpha-helical proteins using an evolutionary algorithm (11). Huang et al used this type of function to recognize native structures (1, 34). One problem with the above global functions is that they assume that proteins are ideally spherical in shape when in actuality native proteins exhibit a much larger range of shapes. A more flexible approach uses an ellipsoidal approximation of the shape of the hydrophobic core that does not require a significant increase in computation and aids in the selection of near-native conformations from decoy sets containing a high number of protein-like yet incorrect compact conformations (9). The problem associated with these functions is that they will inevitably exclude a small percentage of protein structures that deviate from their assumptions concerning shape and thus fail when a protein is divided into small subdomains or contains large invaginations (1HQI is a toroid). In spite of this potential downfall, they have demonstrated their usefulness in several methods owing to their ability to recognize the majority of small hydrophobic cores, their simplicity, and the speed with which they can be computed.

Pair Interactions

Many low-resolution potential/scoring functions utilize an empirically derived pair potential in place of or in addition to the residue-environment term described above. The most common of these potentials are functions of the position of a single center per residue (C_α , C_β , or centroid/united atom center) and are thus quite computationally efficient; all-atom functions have also been used (77). Many variations of pair terms have been developed, with the two main branches of methods being

distance dependent and contact based (57, 84). Like the residue-environment term mentioned above, these scoring functions are sometimes justified by positing that the arrangements of residues in proteins follow a Boltzman distribution: $E(x) = kT \ln P(x)$, where x is a feature such as the occurrence of two residues separated by a distance less than r . Alternatively, these scoring functions may be seen purely as probability distribution functions (23, 83). In the former case, the optimization may be viewed as a search for the lowest energy configurations; in the latter, a search for the highest probability configurations. For most applications, there is little practical difference between the two viewpoints. The issue becomes more substantive, however, when such database-derived scoring functions are combined with physics-based potentials, as will likely become increasingly useful over the next several years.

Several problems are associated with statistically derived pair potentials. The assumption that free energies can be represented by summing over component interactions is not generally valid across all interactions present in proteins, and thus, the basic functional form may not be adequate to represent the free energy of a protein conformation (21, 53). The most significant problem with pair potentials is that they are dominated by hydrophobic/polar partitioning, which gives rise to anomalous effects such as a long-range repulsion between hydrophobic residues (89). This can be corrected by conditioning the pair distributions on the environments of the two residues, which largely eliminates these undesired effects (82). With the elimination of the otherwise overwhelming influence of hydrophobic partitioning, specific interactions such as electrostatic attraction between oppositely charged residues dominate the pair scoring/energy functions, and hydrophobic interactions make relatively modest contributions. The pair term, in this case, is perhaps best viewed as the second term in a series expansion for the residue-residue distributions in the database in which the residue-environment distributions are the first term.

Some of the earliest comprehensive tests of the discriminatory power of these pair potentials were done in the context of threading self-recognition (62). Later work demonstrated that the self-recognition problem was not a sufficiently challenging test of scoring functions and focused on the performance of multiple pairwise energy functions on larger, more diverse sets of conformations (24, 38, 41, 68, 69). The performance of the various energy functions at recognizing native-like structures in large ensembles of incorrect decoys is highly dependent on the methods used to create the decoy sets, highlighting the fact that an energy function that works well in the context of one method will not necessarily work well given a decoy set created using an orthogonal method.

Sequence Independent Terms/Secondary Structure Arrangement

Many features of proteins, such as the association of beta strands into sheets, can be described by sequence independent scoring functions. Several early approaches

to folding all-beta proteins were protocols dominated by initial low-resolution combinatorial searches of possible strand arrangements and were concerned only with the probability of different strand arrangements (16, 18, 19, 72). These early methods narrowed the conformational space by considering sequence specific effects only in the context of highly probable strand arrangements. Several of the relatively successful methods at CASP3 incorporated secondary structure packing terms (51, 81). Ortiz et al used an explicit hydrogen bond term in combination with a $\beta\alpha\beta$ and a $\beta\beta\alpha$ chirality term to ensure protein-like secondary structure formation. It cannot be expected that low-complexity lattice models produce the correct chirality or subtle higher-order effects such as strand twist, and these rules sensibly correct for these expected shortcomings (65). The Rosetta method used three terms that monitored strand-strand pairing, sheet formation, and helix-strand interactions to ensure protein-like secondary structure arrangements.

The Use of Multiple Sequence Alignments in Tertiary Prediction

The large number of homologous sequences often available for a protein family represents a potentially rich source of information useful to ab initio structure prediction methods. Correlated mutation analysis [contact prediction based on covariance patterns in multiple sequence alignments (MSAs)] was used as part of one relatively successful method at CASP3 (67). The accuracy of these covariance methods is not great (63, 64); various methods, including the fitting of secondary structure to initial constraints in order to obtain additional constraints, were used to increase the robustness of ab initio protocols based on such information to expected error levels (66). Previous methods have used MSAs to predict the burial of sequence positions in the query sequence based on hydrophobicity and patterns of variation (5, 7).

Another way to use MSAs is in the selection of low-resolution decoys by requiring that all decoys be consistent with all sequences in the family. In one procedure, the aligned sequences are mapped onto decoys one at a time and the energy of each decoy averaged over all aligned sequences is computed (2). In another approach, simulations of multiple aligned homologs are coupled and carried out simultaneously (45). In these two methods, the models with the lowest score are selected as the best models. In a recent method for applying MSA information to fold prediction, an independent simulation is carried out for each aligned sequence. Only after multiple simulations for each aligned sequence are completed is the multiple sequence alignment information utilized by simultaneously clustering all of the structures generated from the different aligned sequences. The largest and most diverse cluster often contains the best model (9). In all above described cases, significant improvements were seen in the ability of the prediction methods to select good models using highly simplified representations of proteins.

High-Resolution Structure Prediction—Prospects for Refinement

Reduced complexity approaches, as described above, cannot be expected to consistently generate predictions with resolutions of better than 3–7 Å. Low-resolution methods are perhaps best viewed as narrowing the possible conformations from an exponentially large number to a number small enough that more computationally expensive methods can be applied. High-resolution potentials must be improved in order for significant progress to be made in the field of ab initio protein fold prediction.

Some progress has been made in modeling two of the more difficult terms in the potential—solvation and electrostatics—in the context of full atom models. Because the transfer-free energies of small molecules from nonpolar solvents to water are correlated with solvent-accessible surface area, solvation energies are often described using surface area-based methods. Wesson & Eisenberg found that the addition of a solvent-accessible surface area-based term, the parameters for which were based on water-vapor partition experiments, stabilized molecular dynamics trajectories (90). Promising results were also obtained by combining a similar surface area-based empirical solvation term with molecular mechanics and entropic terms (36). One problem with surface area-based methods is that they do not resolve the considerable difference between the free energy of charges buried just below the surface and charges buried deep within the protein core. The generalized Born model remedies this problem by treating the nonpolar interactions with a surface area-based term while using a generalization of the Born equation to deal with the desolvation of charges and charge pairs in the protein interior (37, 52, 88). This method has proven to be useful for modeling loops as described above and is fast enough to be included in almost any discrimination or refinement procedure (71). An alternative implicit solvation model that is able to partially discriminate between native-like and incorrect decoys was developed by Lazaridis & Karplus (48a).

From the above discussion it is apparent that most current energy/scoring functions are based either on small-molecule parameterized functions with forms suggested by physical chemistry or on protein database statistics. We believe that progress will come from a combination of these two approaches.

THE ROSETTA METHOD

How can ab initio structure prediction methods produce reasonable models given the inadequacies of current potential functions? One possible solution is to limit the conformational space searched to that compatible with the local sequence-structure propensities of the protein sequence. Such a procedure is consistent with a view of the folding process in which local structure propensities influence the conformations sampled by short sequence segments, and folding involves a search

through these local conformations for a stable tertiary structure simultaneously consistent with these local biases and with nonlocal constraints (compactness, electrostatics, hydrophobic burial, etc.). Local biases are influenced by relatively subtle interactions (side-chain/main-chain hydrogen bonding, side chain configurational entropy losses, etc.) as witnessed by the debate over the origins of secondary structure propensities; current physical chemistry–based models do not capture all of the subtleties. To circumvent this problem, the Rosetta procedure makes the assumption that the distribution of configurations sampled by a peptide segment are reasonably well-approximated by the distribution of configurations observed in the protein database. Tertiary structures are generated using a Monte Carlo search of the possible combinations of likely local structures, minimizing a scoring function that accounts for nonlocal interactions such as compactness, hydrophobic burial, specific pair interactions (disulfides and electrostatics), and strand pairing (see Figure 1). With lists of 25 fragments per sequence position, many subtly different versions of essentially similar local structures may be represented; thus, the number of states per position is effectively much lower than 25. Optimization of nonlocal interactions within conformational space defined by the fragment sets produces structures with buried hydrophobic residues, paired beta strands, and other protein-like features that are, by construction, consistent with

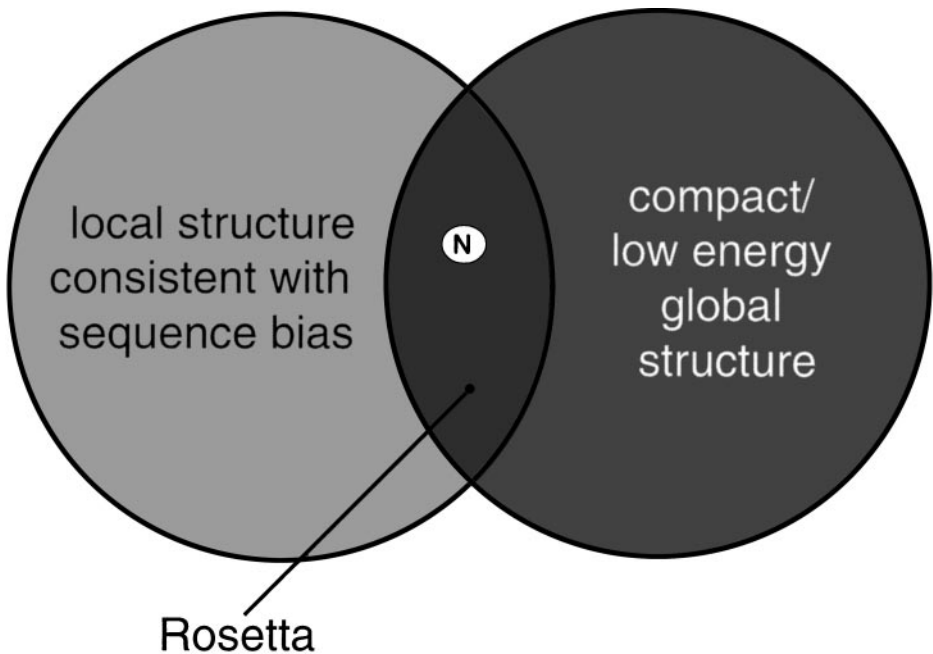


Figure 1 Venn diagram illustrating the conceptual basis for Rosetta. Near-native structures are labeled N.

local sequence-structure biases. This method produced relatively good predictions during CASP3 (81).

RELATIONSHIP TO PROTEIN FOLDING?

Do ab initio folding methods have any resemblance to how proteins actually fold?

It is interesting to note that low-resolution models have been surprisingly successful in accounting for both protein folding rates and the distribution of structure in the folding transition state from knowledge of the final folded state (1, 28, 59). Because of the considerable conformational averaging that takes place before the native state is reached (either when a protein folds in solution or during a folding simulation), the level of complexity with which the interactions responsible for folding must be represented may be relatively low until after the folding transition state. Thus, low-resolution models can provide reasonable results for the broad outlines of the folding process (the rate, mechanism, and low-resolution structure), whereas higher-resolution models are likely to be required for properties depending on the detailed structure of the native state such as stability and high-resolution structure prediction.

APPLICATIONS OF AB INITIO STRUCTURE PREDICTION

As structures are rapidly determined and novel folds become increasingly rare, what are the likely applications of ab initio structure prediction methods? Three promising areas are the annotation of open reading frames (ORFs) and genes with unknown functions and structures, the elaboration and extension of homology-based models, and the rapid generation of structural models from low-resolution structural information.

Genome Annotation

Many factors reduce the ability of sequence homology searches to identify distant homologs (75). Domain insertions and extensions, circular permutations, and the exchange of secondary structure elements have all been observed in cases where structural and functional relationships were not clear based on sequence homology. In order to reliably interpret the flood of sequence currently entering databases, we must have at our disposal methods that can deal with these difficult cases as well as the clearer evolutionary relationships detectable at the sequence level. One recently solved genome (*Mycoplasma genitalium*) showed sequence homology to proteins of known function and/or structure for 38% of its proteins (76). For *Saccharomyces cerevisiae*, ~1/3 of the ORFs in the genome show homology to proteins of known structure (80). Annotation of ORFs

lacking sequence homology to proteins of known function represents one of the most promising potential uses for ab initio prediction. Current methods can make reasonable predictions for small alpha and alpha-beta proteins; the Rosetta method in particular has been successful in blind tests and extensive in house tests on this class of proteins. Of the ~6000 ORFs in baker's yeast, ~300 have at least 15% of their residues predicted to be helical with a total length less than 110 residues and no link to proteins of known structure (220 of these 300 also lack functional annotation) (56). Models can also be produced for modular domains of up to ~150 residues that occur in sufficiently diverse sequence contexts for their boundaries to be readily evident from multiple sequence alignments.

These low-resolution models can be analyzed using several different methods. First, the structures may be compared against PDB using a structure-structure comparison method such as Dali (31–33). Promising preliminary results have been obtained with such an approach using models produced by Rosetta. Dali frequently matches Rosetta models to protein structures related to the native structure for the sequence. Second, the structures may be probed for the presence of sets of residues in specified geometrical arrangements that are indicative of specific protein functions (25, 26). Third, the structures may be used to increase the reliability of matches to sequence motif libraries such as PROSITE—Taylor's and Thornton's groups have shown that structural consistency can be used quite effectively to filter through weak sequence matches to PROSITE patterns (44, 55).

Homology Modeling

Clearly important applications exist for homology modeling in whole genome structure prediction. For example, Sanchez & Sali have constructed models automatically for all ORFs in *S. cerevisiae*, showing significant homology to proteins of known structure (80). Frequently, however, only a portion of the sequence being modeled is represented in the template structure, and thus, homology models often leave a significant fraction of the sequence not modeled (frequently large insertions and N- and C-terminal flanking sequences). Ab initio methods are well-suited for adding these missing regions to homology models, thereby producing much more complete (to the extent the elaborations are accurate!) sets of models.

Structures from Limited Constraint Sets

The obvious drawback of current ab initio structure prediction methods is their relatively low accuracy and reliability. Even limited amounts of experimental data on the structure of a protein can remedy this considerably. For example, quite accurate structures were produced by Rosetta in conjunction with NMR chemical shift data (to enhance fragment selection) and sparse NMR constraints (10). Distance constraints from cross-linking, followed by mass spectrometry, could also be readily incorporated into such an approach and could be obtained on a high-throughput scale.

CONCLUSION

Ab initio protein folding has traditionally been an area of purely academic interest characterized by relatively slow progress. There is hope that the next years will see considerable improvements in ab initio structure prediction methods, and that this will provide both considerable basic insight into folding and a valuable resource for interpreting genome sequence information.

NOTE ADDED IN PROOF

The recent CASP4 structure prediction experiment showed a dramatic improvement in ab initio structure prediction methods; in particular, the Rosetta method produced reasonably correct large (~100 residues) fragments for 16 of 22 domains under 300 residues for which ab initio predictions were made.

Visit the Annual Reviews home page at www.AnnualReviews.org

LITERATURE CITED

1. Alm E, Baker D. 1999. Matching theory and experiment in protein folding. *Curr. Opin. Struct. Biol.* 9:189–96
2. Badretdinov A, Finkelstein AV. 1998. How homologs can help to predict protein folds even though they cannot be predicted for individual sequences. *J. Comp. Biol.* 5:369–76
3. Baldwin RL. 1999. Protein folding from 1961 to 1982. *Nat. Struct. Biol.* 6:814–17
4. Barton GJ, Russell RB. 1993. Protein structure prediction. *Nature* 361:505–6
5. Benner SA, Badcoe I, Cohen MA, Gerloff DL. 1994. Bona fide prediction of aspects of protein conformation. Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. *J. Mol. Biol.* 235:926–58
6. Benner SA, Cohen MA, Gerloff D. 1992. Correct structure prediction? *Nature* 359: 781
7. Benner SA, Gerloff D. 1991. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul.* 31:121–81
8. Blanco FJ, Rivas G, Serrano L. 1994. A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nat. Struct. Biol.* 1:584–90
9. Bonneau R, Strauss CEM. 2000. Improving the performance of ROSETTA using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins: Struct. Funct. Genet.* In press
10. Bowers P, Strauss CEM, Baker D. 2000. De novo protein structure determination using sparse NMR data. *J. Biomol. NMR.* In press
11. Bowie JU, 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: 4436–40
12. Bystroff C, Baker D. 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281:565–77
13. Bystroff C, Simons KT, Han KF, Baker D. 1996. Local sequence-structure correlations in proteins. *Curr. Opin. Biotechnol.* 7:417–21
14. Bystroff C, Thorsson V, Baker D. 2000.

- HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* 301:173–90
15. Callihan DE, Logan TM. 1999. Conformations of peptide fragments from the FK506 binding protein: comparison with the native and urea-unfolded states. *J. Mol. Biol.* 285:2161–75
 16. Chothia C. 1984. Principles that determine the structure of proteins. *Annu. Rev. Biochem.* 53:537–72
 17. Cohen BI, Presnell SR, Cohen FE. 1993. Origins of structural diversity within sequentially identical hexapeptides. *Protein Sci.* 2:2134–45
 18. Cohen FE, Sternberg MJ, Taylor WR. 1980. Analysis and prediction of protein beta-sheet structures by a combinatorial approach. *Nature* 285:378–82
 19. Cohen FE, Sternberg MJ, Taylor WR. 1982. Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* 156:821–62
 20. Defay T, Cohen FE. 1995. Evaluation of current techniques for ab initio protein structure prediction. *Proteins: Struct. Funct. Genet.* 23:431–45
 21. Dill KA. 1997. Additivity principles in biochemistry. *J. Biol. Chem.* 272:701–4
 22. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, et al. 1995. Principles of protein folding—a perspective from simple exact models. *Protein Sci.* 4:561–602
 23. Domingues FS, Koppensteiner WA, Jaritz M, Prlic A, Weichenberger C, et al. 1999. Sustained performance of knowledge-based potentials in fold recognition. *Proteins: Struct. Funct. Genet.* 37:112–20
 24. Eyrich VA, Standley DM, Friesner RA. 1999. Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J. Mol. Biol.* 288:725–42
 25. Fetrow JS, Godzik A, Skolnick J. 1998. Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* 282:703–11
 26. Fetrow JS, Skolnick J. 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* 281:949–68
 27. Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, et al. 1999. CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins: Struct. Funct. Genet. Suppl.* 3, pp. 209–17
 28. Galzitskaya OV, Finkelstein AV. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA* 96:11299–304
 - 28a. Gatchell DW, Dennis S, Vejda S. 2000. Discrimination of near-native proteins structures from misfolded models by empirical free energy functions. *Proteins: Struct. Funct. Genet.* 41:518–34
 29. Han KF, Bystroff C, Baker D. 1997. Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci.* 6:1587–90
 30. Hinds DA, Levitt M. 1994. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* 243:668–82
 31. Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233:123–38
 32. Holm L, Sander C. 1995. Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* 20:478–80
 33. Holm L, Sander C. 1997. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* 25:231–34
 34. Huang ES, Subbiah S, Levitt M. 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* 252:709–20
 35. Ishikawa K, Yue K, Dill KA. 1999.

- Predicting the structures of 18 peptides using Geocore. *Protein Sci.* 8:716–21
36. Janardhan A, Vajda S. 1998. Selecting near-native conformations in homology modeling: the role of molecular mechanics and solvation terms. *Protein Sci.* 7:1772–80
 37. Jayaram B, Fine R, Sharp K, Honig B. 1989. Free energy calculations of ion hydration: an analysis of the born model in terms of microscopic simulations. *J. Phys. Chem.* 93:4320–27
 38. Jernigan RL, Bahar I. 1996. Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* 6:195–209
 39. Jones DT. 1997. Progress in protein structure prediction. *Curr. Opin. Struct. Biol.* 7:377–87
 40. Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292:195–202
 41. Jones DT, Thornton JM. 1996. Potential energy functions for threading. *Curr. Opin. Struct. Biol.* 6:210–16
 42. Kabsch W, Sander C. 1984. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA* 81:1075–78
 43. Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R. 1999. Predicting protein structure using only sequence information. *Proteins: Struct. Funct. Genet.* 37:121–25
 44. Kasuya A, Thornton JM. 1999. Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.* 286:1673–91
 45. Keasar C, Tobi D, Elber R, Skolnick J. 1998. Coupling the folding of homologous proteins. *Proc. Natl. Acad. Sci. USA* 95:5880–83
 46. King RD, Saqi M, Sayle R, Sternberg MJ. 1997. DSC: public domain protein secondary structure prediction. *Comput. Appl. Biosci.* 13:473–74
 47. Kolinski A, Skolnick J. 1994. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins: Struct. Funct. Genet.* 18:338–52
 48. Kolinski A, Skolnick J. 1994. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins: Struct. Funct. Genet.* 18:353–66
 - 48a. Lazaridis T, Karplus M. 1999. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* 288:477–87
 49. Lee MR, Duan Y, Kollman PA. 2000. Use of MM-PB/SA in estimating the free energies of proteins: application to native, intermediates, and unfolded villin headpiece. *Proteins: Struct. Funct. Genet.* 39:309–16
 50. Lesk AM. 1997. CASP2: report on ab initio predictions. *Proteins: Struct. Funct. Genet. Suppl.* 1, pp. 151–66
 51. Lomize AL, Pogozheva ID, Mosberg HI. 1999. Prediction of protein structure: the problem of fold multiplicity. *Proteins: Struct. Funct. Genet.* 37:199–203
 52. Luo R, David L, Hung H, Devaney J, Gilson MK. 1999. Strength of solvent-exposed salt-bridges. *J. Phys. Chem. B* 103:727–36
 53. Mark AE, van Gunsteren WF. 1994. Decomposition of the free energy of a system in terms of specific interactions. *J. Mol. Biol.* 240:167–76
 54. Marqusee S, Robbins VH, Baldwin RL. 1989. Unusually stable helix formation in short alanine-based peptides. *Proc. Natl. Acad. Sci. USA* 86:5286–90
 55. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, et al. 1998. Protein folds and functions. *Structure* 6:875–84
 56. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, et al. 2000. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 28:37–40
 57. Miyazawa S, Jernigan RL. 1999. An

- empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins: Struct. Funct. Genet.* 36: 357–69
58. Moulton J, Hubbard T, Fidelis K, Pedersen JT. 1999. Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins: Struct. Funct. Genet. Suppl.* 3, pp. 2–6
 59. Munoz V, Eaton WA. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA* 96:11311–16
 60. Munoz V, Serrano L. 1996. Local versus nonlocal interactions in protein folding and stability—an experimentalist's point of view. *Fold. Des.* 1:R71–77
 61. Murzin AG. 1999. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins: Struct. Funct. Genet.* 37:88–103
 62. Novotny J, Brucoleri R, Karplus M. 1984. An analysis of incorrectly folded protein models. Implications for structure predictions. *J. Mol. Biol.* 177:787–818
 63. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. 1999. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins: Struct. Funct. Genet.* 37:149–70
 64. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. 1999. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins: Struct. Funct. Genet. Suppl.*(3):149–70
 65. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. 1999. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins: Struct. Funct. Genet. Suppl.* 3, pp. 177–85
 66. Ortiz AR, Kolinski A, Skolnick J. 1998. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.* 277:419–48
 67. Ortiz AR, Kolinski A, Skolnick J. 1998. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc. Natl. Acad. Sci. USA* 95:1020–25
 68. Park B, Levitt M. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* 258:367–92
 69. Park BH, Huang ES, Levitt M. 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* 266:831–46
 70. Park BH, Levitt M. 1995. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* 249:493–507
 71. Rapp CS, Friesner RA. 1999. Prediction of loop geometries using a generalized Born model of solvation effects. *Proteins: Struct. Funct. Genet.* 35:173–83
 72. Reva BA, Finkelstein AV. 1996. Search for the most stable folds of protein chains: II. Computation of stable architectures of beta-proteins using a self-consistent molecular field theory. *Protein Eng.* 9:399–411
 73. Reva BA, Finkelstein AV, Sanner MF, Olson AJ. 1996. Adjusting potential energy functions for lattice models of chain molecules. *Proteins: Struct. Funct. Genet.* 25:379–88
 74. Rost B, Sander C. 1993 Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232:584–99
 75. Russell RB, Ponting CP. 1998. Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* 8:364–71
 76. Rychlewski L, Zhang B, Godzik A. 1998. Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold. Des.* 3:229–38
 77. Samudrala R, Xia Y, Huang E, Levitt M. 1999. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins: Struct. Funct. Genet. Suppl.* 3, pp. 194–98

78. Deleted in proof
79. Samudrala R, Xia Y, Levitt M, Huang ES. 1999. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. *Pac. Symp. Biocomput.* 505–16
80. Sanchez R, Sali A. 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* 95:13597–602
81. Simons KT, Bonneau R, Ruczinski I, Baker D. 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct. Funct. Genet.* 37:171–76
82. Simons KT, Kooperberg C, Huang E, Baker D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268:209–25
83. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Struct. Funct. Genet.* 34:82–95
84. Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5:229–35
85. Skolnick J, Kolinski A. 1991. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J. Mol. Biol.* 221:499–531. Erratum. 1992. *J. Mol. Biol.* 223:583
86. Srinivasan R, Rose GD. 1995. LINUS: a hierarchic procedure to predict the fold of a protein. *Proteins: Struct. Funct. Genet.* 22:81–99
87. Sternberg MJ, Cohen FE, Taylor WR. 1982. A combinatorial approach to the prediction of the tertiary fold of globular proteins. *Biochem. Soc. Trans.* 10:299–301
88. Still WC, Tempczyk A, Hawley RC, Hendrickson T. 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* 112:6127–29
89. Thomas PD, Dill KA. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* 257:457–69
90. Wesson L, Eisenberg D. 1992. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.* 1:227–35