# Macromolecular Modeling with Rosetta

## Rhiju Das[1] and David Baker[1,2]

[1]Department of Biochemistry, University of Washington, and [2]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195; email: rhiju@u.washington.edu, dabaker@u.washington.edu

## Key Words

protein design, phasing, proteins, RNA, protein structure prediction

## Abstract

Advances over the past few years have begun to enable prediction and design of macromolecular structures at near-atomic accuracy. Progress has stemmed from the development of reasonably accurate and efficiently computed all-atom potential functions as well as effective conformational sampling strategies appropriate for searching a highly rugged energy landscape, both driven by feedback from structure prediction and design tests. A unified energetic and kinematic framework in the Rosetta program allows a wide range of molecular modeling problems, from fibril structure prediction to RNA folding to the design of new protein interfaces, to be readily investigated and highlights areas for improvement. The methodology enables the creation of novel molecules with useful functions and holds promise for accelerating experimental structural inference. Emerging connections to crystallographic phasing, NMR modeling, and lower-resolution approaches are described and critically assessed.

## Contents

## INTRODUCTION

Biomolecules have evolved the fascinating property of folding into unique tertiary structures dictated by their chemical sequence. The prediction of these structures from sequence alone and the design of new functional molecules are classic problems in biophysics. Although general solutions to these formidable problems have not been achieved, recent years have seen much progress. In 2004, the principles and methods implemented in the Rosetta algorithm for de novo protein structure modeling were summarized (1), and since then, this method has led to a handful of blind predictions with back-

bone accuracies better than 2 Å (see, e.g., **Figure 1a,b**) (2–4). The first goal of the present review is to illustrate how, in these intervening four years, the same basic ingredients have proven successful in a wide range of biomolecular modeling problems beyond de novo protein structure prediction. The second goal is to outline how, perhaps in the next four years, these approaches may mature from largely academic pursuits into useful tools for characterizing and manipulating molecular systems. After briefly summarizing the main ingredients of the Rosetta methodology, we describe how these principles can be generally put into practice in applications ranging from loop modeling, to protein and ligand docking with backbone and side chain flexibility, to RNA folding. We end the review by highlighting new connections of these high-resolution molecular modeling efforts to experimental methods as well as current challenges in bringing these hybrid computational/experimental approaches into wide use.

## KEY INGREDIENTS OF MOLECULAR MODELING

Macromolecular structure prediction and design are based on the premise that the observed conformations of folded macromolecules are almost always the lowest free-energy states (see, however, Reference 5). Hence, structure prediction is generally the problem of finding the lowest-energy structure given the sequence of a biopolymer, and design is the problem of finding the lowest-energy sequence for the target structure. Critical to both molecular modeling problems are a reasonably accurate free-energy function and a sampling method capable of locating the minima of this function for the biomolecular system under study.

### Energy Function

The hallmark features of the folded structures of macromolecules are the burial of nonpolar
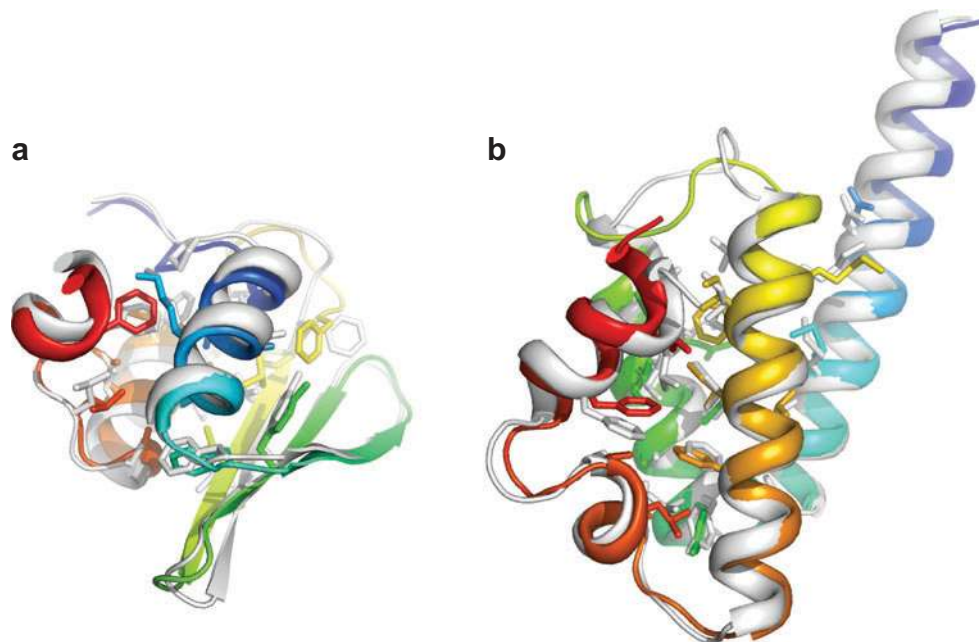
**Figure 1**

Blind de novo predictions of protein structure can achieve high resolution. (*a*) Rosetta prediction for the Critical Assessment of Techniques for Protein Structure Prediction (CASP)6 target, T0281, agrees with the subsequently released crystal structure shown with rainbow coloring from the N terminus (*blue*) to the C terminus (*red*) (*1whz*) with a backbone accuracy of 1.6 Å over 70 residues (3). (*b*) Rosetta prediction for the CASP7 target, T0283, (*white*) agrees with the subsequently released crystal structure also shown in rainbow coloring (*2hh6*) with a backbone accuracy of 1.4 Å over 90 residues (4). A nine-residue C-terminal helix that makes contacts with crystal neighbors is not shown. These panels and the following figures were prepared in PyMOL (Delano Scientific, Palo Alto, CA).

groups away from water; the close, nearly void-free packing of buried groups and atoms; and the formation of intramolecular hydrogen bonds by nearly all buried polar atoms (6, 7). The first feature is a direct consequence of the hydrophobic effect, recognized by Kauzmann (8) many years ago as the dominant driving force in protein folding. The second feature reflects van der Waals interactions between buried atoms and, perhaps more importantly, the strong size dependence of the free-energy cost of forming a cavity in solvent to accommodate the protein. The third feature follows from the significant free-energy cost of stripping water molecules from polar groups upon folding, which must be compensated by the formation of new hydrogen bonds within the protein or nucleic acid molecule. Recognition of these features, especially hydrogen bond-

ing, was fundamentally important for the earliest predictions of the basic secondary structure motifs in proteins by Pauling and colleagues (as reviewed in Reference 9) and, later, in nucleic acids by Watson & Crick (10).

A successful free-energy function must capture to some extent these dominant contributions to macromolecular stability. In Rosetta, atom-atom interactions are efficiently computed using a Lennard-Jones potential to describe packing, an implicit solvation model to describe the hydrophobic effect and the electrostatic desolvation cost associated with burial of polar atoms, and an explicit hydrogen-bonding potential to describe hydrogen bonding (1) (see **Figure 2c**). As we have discussed elsewhere (11–13), an explicit treatment of hydrogen bonding has advantages over the classical electrostatic models
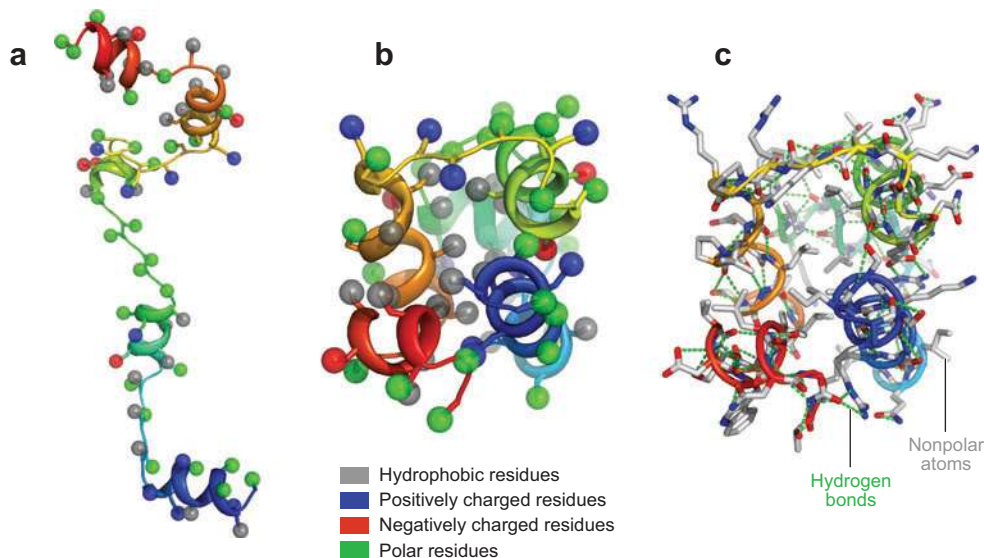
**Figure 2**

Schematic of Rosetta molecular modeling in the context of de novo structure prediction. (*a*) "Snapshot" of low-resolution fragment assembly: conformation after five nine-residue fragment insertions of the sequence of the phage 434 repressor protein. All backbone heavy atoms are simulated and illustrated here as a ribbon (*rainbow*). Side chains are represented in the simulation by interaction centers (*spheres*), with an energy function favoring burial of hydrophobic residues (*gray*) and the exposure of positively charged (*dark blue*), negatively charged (*red*), or other polar (*green*) residues (21). (*b*) Final low-energy conformation produced by fragment assembly. (*c*) All-atom model produced after high-resolution refinement. Pair-wise Lennard-Jones and solvation terms give attractive interactions between nonpolar atoms (*gray*); hydrogen bonds (*green dotted lines*) are also assigned attractive energies (1). For clarity, hydrogen atoms are not shown.

employed in most molecular mechanics potentials in that the orientation dependence is more correctly modeled. Furthermore, long-range electrostatic interactions, which are notoriously difficult to compute accurately owing to induced polarization effects, are strongly damped (12). An insightful comparison of the Rosetta and other energy functions used by different laboratories has been carried out recently (14).

Bonded interactions are treated in Rosetta for the most part with bond lengths and angles fixed at their ideal values. The remaining degrees of freedom are the bond torsion angles, and the associated torsional potentials are perhaps the most difficult aspect of any force field to represent accurately owing to the influence of inherently quantum mechanical effects, which cannot be rigor-

ously decomposed into independent classical contributions. These potentials are modeled empirically in Rosetta on the basis of torsion angle distributions observed in high-resolution crystal structures. This procedure is far from optimal because of the double counting of effects already captured in the nonbonded interaction terms and also on aesthetic grounds. Overall, the rigorous determination of bonded interactions continues to be a formidable challenge (see, e.g., the discussion in Reference 15).

The resulting energy function, encoding the basic physics of molecular interactions, is necessarily approximate. For example, the explicit structure of solvent, long-range electrostatics, and residual dynamics in the molecule have been ignored. Another striking omission is the massive entropy change of the molecule

upon attaining an ordered structure; we have assumed, to a first approximation, that the conformational entropies of different well-packed protein conformations are similar.

Nevertheless, it is important to recognize that success in prediction and design problems neither requires nor is an indicator of exceptionally high accuracy of the assumed energy function. Rather, the success of structure prediction partly stems from the very large energy gap that must exist between the experimentally observed conformation of a folded polymer and the vast majority of non-native conformations. In the context of protein structure prediction, a molecule being "folded" indicates that, at equilibrium, it has a very high probability of being in a single native state. If this probability exceeds 99.9%, the free-energy gap between the native state and the ensemble of nonnative states must be at least $\Delta G = k_\text{B}\text{T} \log (0.999/0.001) = $ 4 kcal/mol, by the Boltzmann relation. Indeed, this free-energy gap is typically measured at 3–10 kcal/mol (16). However, because of the huge decrease in entropy accompanying folding, the gap in energy (rather than free energy) must be much larger. Experimental and theoretical estimates of conformational entropy suggest that the energy of the native state must be on the order of 100 kcal/mol lower than that of any member of the ensemble of unfolded states (~1.4 kcal/mol per residue) (see, e.g., References 17 and 18). Assuming the native state can be located with reasonable confidence if the error in the energy function is on the order of 10% of the energy gap, this allowed error can thus be many kcal/mol. [This argument is oversimplified; the accuracy actually required for structure prediction is somewhat greater, given that there may be a small number of alternative conformations with energies within a few kcal/mol of the native structure (19).] Although errors in the range of several kcal/mol do not appear to compromise structure prediction, they do present problems for applications requiring the quantitative, high-accuracy estimation of free-energy

differences. Furthermore, the challenges associated with estimating conformational entropy changes make the accurate computation of absolute free energies of folding or binding exceptionally difficult.

Regardless of what approximations are made in the assumed all-atom energy function, a crucial aspect of discriminating native structures and viable designs is the maintenance of contacting atoms at their observed characteristic spacings (3, 20). Unfortunately, the short-range repulsion required to enforce these contact distances results in an exceptionally rugged energy landscape, with high barriers close to even the deepest minima. To make the sampling problem more tractable, it is useful to construct a smoother version of this all-atom potential whereby degrees of freedom associated with the rapid fluctuations are effectively "integrated out." For example, in Rosetta, the initial phase in many calculations is a search on a smoothed energy landscape where the side chain degrees of freedom are represented as soft interaction centers (21) (see also **Figure 2a,b**). For proteins, the dominant driving forces in this representation are the nonspecific burial of hydrophobic residues and the pairing of β-strands into sheets, with a smaller contribution from specific but averaged out interactions between side chain centers. For nucleic acids, the forces in this smoothed representation are coarse-grained potentials favoring base pairing and base stacking (see Reference 22).

## Conformational Sampling

The first stage in the search for the global minimum involves locating a large number of local minima in the coarse-grained lower-resolution potential (**Figure 2a,b**). The necessity of exploring as many local minima as possible is a consequence of coarse-grain smoothing, which necessarily introduces large errors owing to the missing critical contribution of interatomic packing to the true free energy. Indeed, given the approximate nature

of the low-resolution energy function, it is important wherever possible to bias the search with any additional information available. For example, the Rosetta low-resolution structure prediction for protein and RNA molecules is based on a picture of folding in which local chain segments sample from distributions of local conformations that are relatively low in energy given their sequences. Folding takes place when a combination of local conformations is sampled that makes possible low-energy tertiary interactions. This flickering between different local structures is modeled by assuming that the distribution of states sampled by a sequence segment in isolation is reasonably well approximated by the distribution of structures observed for the sequence in prior crystal structures.

The second stage in the search for the lowest free-energy minimum starts from each of the alternative minima identified in the initial low-resolution search and adds back the missing atomic detail (**Figure 2c**). In the protein folding case, this is carried out by first performing a simulated annealing search through combinations of discrete amino acid rotamers. In protein design calculations, the process is identical, except all rotamers of all amino acids are considered at each position rather than just the rotamers for a particular native sequence. Then, to further optimize the geometry, Rosetta employs a multistep Monte Carlo minimization procedure composed of an assortment of torsion angle perturbations, with each perturbation followed by efficient one-at-a-time rotamer optimization and then by continuous gradient-based minimization of side chain and backbone torsion angles (1). Beyond the side chain optimization protocol described above, the conformational perturbations include small changes to the backbone torsion angles (see, e.g., Reference 1) and shifts of the relative rigid-body orientations of multiple domains. The implementation of these moves for general modeling problems and the present capabilities and limitations of this challenging search procedure are discussed in the next section.

# A UNIFIED FRAMEWORK FOR BIOMOLECULAR MODELING AND DESIGN

The development over the past few years of an all-atom energy function and of an effective high-resolution search procedure has expanded molecular modeling work well beyond the de novo prediction of the structures of globular, soluble proteins. The same basic ingredients—and, indeed, the same core software routines—are now used to construct comparative models of large proteins, to predict protein-protein interfaces, to design new proteins, and even to explore polymers beyond proteins. In this section, we briefly describe how these seemingly diverse molecular modeling problems can be tackled within the same framework.

## Same Ingredients for Multiple Problems

Quite generally, any prediction or design challenge can be formulated as a global optimization problem with suitable degrees of freedom and constraints. The setup for all such problems is similar. First, the kinematic rules that the atoms follow upon changing any internal torsional or rigid-body degree of freedom are defined. Second, where relevant, alternative sets of discrete states for distinct subtrees are specified (different side chain rotamers at each sequence position, for prediction problems, or rotamers for all or a selected subset of amino acids, for design problems). Third, the values of the internal degrees of freedom are initialized, potentially with information from existing templates. Finally, a schedule of conformational moves is set into motion; this protocol and the associated energy function can be quite similar across many different problems.

The main difference between different classes of modeling problems is typically the first component, the definition of kinematics. In packages that simulate molecular dynamics, a file listing the simulated atoms and

their connections typically supplies this key information. Within Rosetta, these kinematic rules are encoded in an appropriate tree-like representation, as depicted in **Figure 3** [a protein-specific implementation has been described in References 23 and 24; tree-based representations are also used in several other programs (25–28)]. During each Rosetta Monte Carlo move, a subset of the internal degrees of freedom of the molecule, such as the backbone torsion angles of several randomly selected residues, is changed. Atoms that are marked as "ancestors" of the changed residues in the atom tree remain fixed, and the affected atoms and their "descendants" are translated and rotated to appropriately propagate the change. In addition to changes in torsion angles, the relative rigid-body orientations of two domains can be changed, and such moves are propagated across noncovalent connections encoded in the atom tree; see the thin, colored lines in **Figure 3**. Further, the current atom-tree framework allows bond lengths and bond angles to deviate from starting values, although this feature has not yet been widely explored.

As an illustration, the atom-tree diagrams for the low-resolution phase of de novo modeling and for rebuilding loops in comparative models are shown in **Figure 3a,b**. De novo prediction uses a simple atom tree, with each backbone atom connected to its neighbor; internal torsion angles are initialized to uniform values to be changed during the simulation. For loop modeling, backbone movements need to be carried out in the rebuilt segments (colored lines) without perturbing the rest of the structure, so temporary chain breaks are introduced into the atom tree [thin gray lines in **Figure 3b** (29–31, 42)]. After this setup, both de novo and loop rebuilding simulations proceed in a similar fashion, using essentially the same move sets except for additional loop closure steps in the latter case.

In the above framework, tackling a new problem is generally as simple as creating a new, appropriate atom tree. The power of this approach is further illustrated by

---------------------------------------------------------------------→
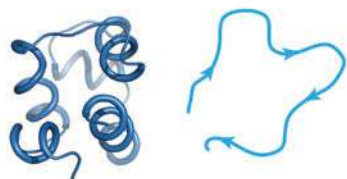
#### Figure 3

A unified framework for tackling multiple molecular modeling problems. Each panel depicts a problem in biomolecule structure prediction or design (*left*) and a diagram of the atom-tree representation used by Rosetta (*right*). (*a*) De novo structure prediction for the phage 434 repressor protein (PDB code: 1r69) (2, 82). (*b*) Loop modeling carried out during comparative modeling of the CASP7 target T0331, pyridoxamine 5′-phosphate oxidase-related protein (2hhz) (4). (*c*) Protein-protein docking for a host and viral major histocompatibility complex receptor,1p7q (83), with full flexibility of all degrees of freedom (backbone, side chain, rigid body). (*d*) Protein-protein docking with backbone flexibility limited to a hinge region (*red*) and to a loop (*blue*) (24). (*e*) Symmetric folding and docking to model the coiled-coil trimerization motif of coronin 1, 2akf (84; I. André, R. Das, D. Baker, unpublished results). (*f*) Symmetry-constrained modeling of the sequence NNQQNY from prion Sup35 (9, 35). (*g*) Small-molecule docking of a steroid with an antibody, 2dbl (36, 85). (*h*) RNA fold prediction for a pseudoknot with two known Watson-Crick pairings, 1l2x (22, 86). (*i*) Redesign of a protein-protein interface between colicin E7 DNase and the Im7 immunity protein, 1ujz (40). (*j*) Design of a novel retroaldolase enzyme into scaffold 1a53 (42, 87). (*k*) Redesign of an interface between a homing endonuclease and its DNA-binding site for altered cleavage specificity, 2fld (43). In all panels, colored segments represent degrees of freedom that are sampled; gray segments are held fixed during modeling. Thick lines represent torsional degrees of freedom in the polymer; arrows give the backbone direction (N terminus to C terminus for proteins; 5′ to 3′ for nucleic acids), and small white gaps represent temporary chain breaks required to maintain an acyclic tree representation. Each thin line represents six rigid-body degrees of freedom for the translation and rotation between the connected portions of the atom tree. Faded colors in (*e*) and (*f*) represent torsional and rigid-body degrees of freedom that are cloned from segments in dark colors. Side chain rotameric degrees of freedom in (*g*)-(*l*) are represented as solid circles; open circles represent the expansion of the sampled rotamer set to include all amino acid types.

flexible backbone protein-protein docking. Backbone conformational changes occur frequently upon protein binding; thus the fixed-backbone approximation used in most current docking algorithms typically precludes high-resolution prediction (32, 33). The most general representation of this problem allows all internal and rigid-body degrees of freedom to vary (**Figure 3c**), but this entails the search of a huge conformational space. Instead, with a combination of flexible and rigid atom-tree segments, it is straightforward to supplement

**a** Protein structure prediction

**b** Loop modeling
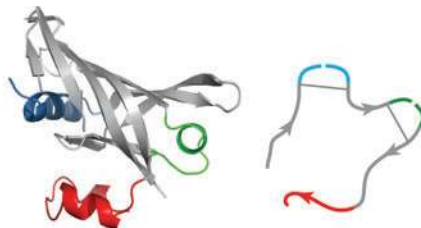
**c** Protein docking (fully flexible)
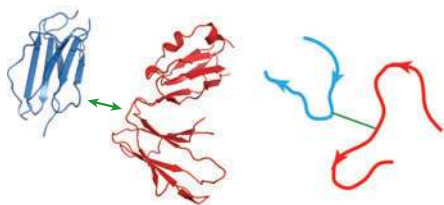
**d** Protein docking (partly flexible)
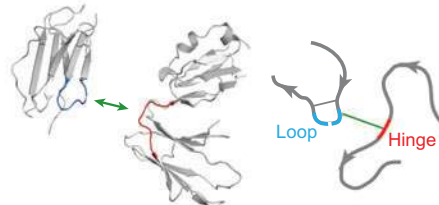
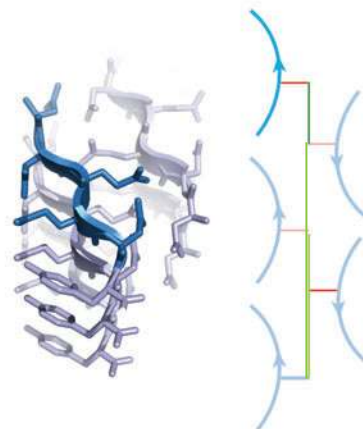**e** Symmetric complexes
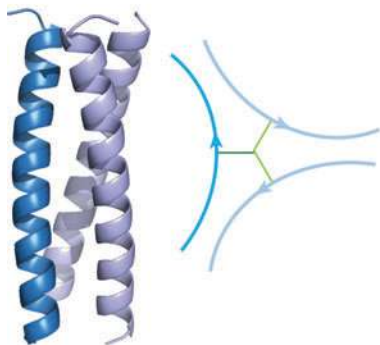
**f** Fibril modeling



**Figure 3**

(*Continued*)

the rigid-body and side chain degrees of freedom with movement in specific loops or hinge regions (**Figure 3d**) (24, 34). Alternatively, in special cases, a comprehensive search of all degrees of freedom can be carried out if symmetry constraints are available. **Figure 3e** shows the atom tree corresponding to the problem of modeling the structure of a three-helix coiled coil with cyclical symmetry (35; I. André, R. Das, D. Baker, unpublished results); a differ-

ent atom-tree topology (**Figure 3f**) describes the problem of modeling extended fibrils (35) and is used to model a variety of disease-associated amyloid structures. Moves carried out on the first protein chain are copied to the other chains, including torsion angle changes as well as shifts in the overall translation and rotation of the chain (thin colored connections in **Figure 3**). Without developing a large amount of additional code, the energy
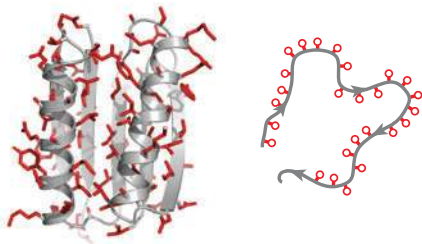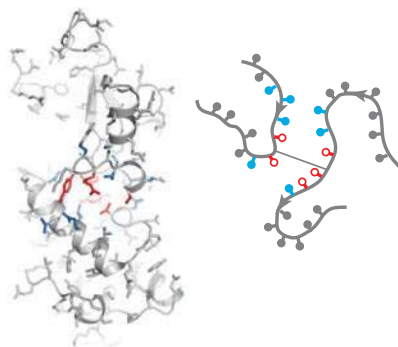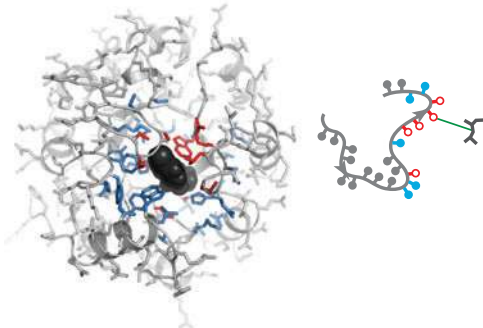
**g** Small-molecule docking
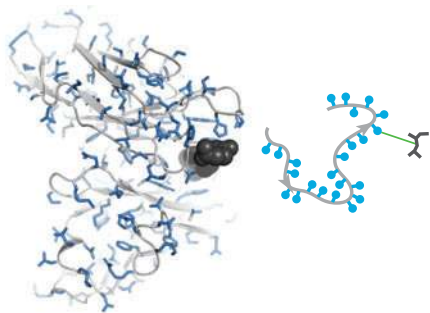
**h** RNA folding
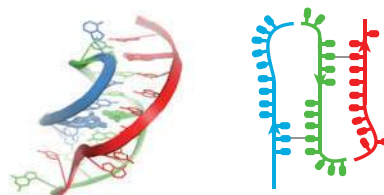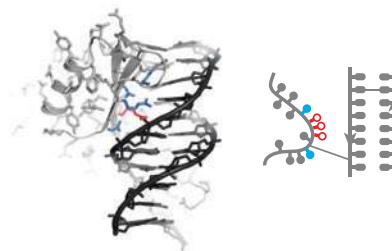
**i** Protein design

**j** Protein-protein interface design

**k** Enzyme design

**l** Protein-DNA interface design

function used in de novo folding and a hybrid of the folding and docking conformational search procedures can be transferred directly to this atom tree for both low-resolution and high-resolution phases.

Other molecular modeling problems are treated in Rosetta by following the same steps of defining an atom tree, initial conditions, accessible rotamers, and a schedule of Monte Carlo moves affecting the flexible degrees of freedom. Modeling the binding of small molecules to proteins is very similar to the problem of protein-protein docking (**Figure 3g**) (36). Modeling of membrane proteins follows a protocol similar to the modeling of soluble proteins, albeit with different low-resolution and high-resolution energy functions (37). Searches of nucleic acid conformations are also possible. Indeed, for RNA folds in which some Watson-Crick base pairings are known, this extra pairing information can be retained throughout the simulation with the appropriate atom tree and initial rigid-body transformations between paired residues (**Figure 3b**) (22, 23).

For a seemingly different molecular modeling application, the engineering of new proteins, the same framework is used. In design calculations, the simulated annealing move used to optimize side chain rotamers in structure prediction is simply expanded to include rotamers from all amino acid types (**Figure 3i**). If desired, such "fixed-backbone" design moves can be alternated with structure refinement—a strategy used to successfully design a protein with a novel α/β-fold to atomic resolution (20)—within a single protocol. Protein design methods have been used recently to design a sequence that can switch between two very different folded conformations (38), and further highlights and current challenges in protein design have been reviewed (39).

Beyond the design of globular proteins, the energy functions and search procedures developed for protein-protein docking, small-molecule docking, and protein/nucleic acid interactions can be converted into analo-

gous design applications. **Figure 3j** illustrates protein-protein interface design, which couples refinement of rigid-body orientation with optimization of the amino acids at the interface to create new pairs of interacting proteins (40). **Figure 3k** shows the atom tree for designing an enzyme by optimizing interactions with a model for the reaction transition state (41, 42); and **Figure 3l** illustrates protein-DNA interface design, recently used to reengineer endonuclease cleavage specificity for gene therapy and other applications (43).

## Similar Challenges Across Multiple Problems

An advantage of using the same overall framework to approach this wide range of problems is the tremendous amount of feedback that can be compiled to diagnose limitations and to improve the underlying free-energy function and sampling methodologies. In the Rosetta development community, there have been numerous occasions in which improvements in either the energy function or sampling methodology driven by findings in one area have led to payoffs in another area. For example, one crucial step that enabled both improved discrimination of native protein structures in prediction problems (44) and the creation of well-ordered protein folds in design problems (20, 37, 45) was to expand the atomic radii of the Rosetta energy function, decreasing the probability of artificially tight side chain packing. At the same time, having a unified framework clearly exposes the two formidable issues that still hamper the molecular prediction and design problems that can be currently modeled with Rosetta: the limits of conformational search and inaccuracies in the treatment of polar interactions in the energy function.

With current conformational sampling strategies and available computational power, every prediction problem discussed above becomes intractable at some point. The conformational space grows exponentially with the

number of degrees of freedom, roughly represented by the total length of colored segments in each atom-tree diagram in **Figure 3**. For example, the low-resolution phase of de novo structure prediction can only find folds within the 2-Å "radius of convergence" of the all-atom refinement procedure for protein lengths less than ~100 residues and, even then, only in favorable cases (2, 4). Similarly, the ability of high-resolution comparative modeling to improve over existing templates drops dramatically for protein domains with lengths greater than ~200 residues (4). Problems that additionally require modeling rigid-body degrees of freedom along with torsional degrees of freedom, e.g., protein-protein docking or protein/small-molecule docking, have only yielded high-resolution models in cases where the protein backbone remains close to a known template (46, 47) or the number of flexible residues (35; I. André, R. Das, D. Baker, unpublished data) is less than ~30. For larger problems, the conformational space that needs to be searched can be significantly contracted through the use of even limited experimental data, a frontier discussed in detail below.

In addition to making some de novo modeling problems essentially intractable, the current difficulty of conformational search greatly limits the throughput and availability of high-resolution modeling. In the recent Critical Assessment of Techniques for Protein Structure Prediction (CASP)7 trials, high-resolution de novo structure prediction simply could not be carried out by the automated Robetta server owing to the two-day limits on automated predictions. More practically, *in silico* screening of small-molecule inhibitors for proteins at high resolution requires a few hundred computer hours to be expended on each case. Given this daunting computational expense, large-scale high-resolution screening of thousands of small molecules with multiple protein targets allowing for flexibility of both the ligand and the protein—as would be desirable for most drug design applications— appears currently out of reach.

Even though conformational sampling limits most areas of high-resolution molecular modeling, there are known defects in the underlying energy function. As discussed above, quantitative estimates of free-energy changes that involve large changes in molecular order, e.g., for folding of whole protein domains or for the ordering of large loops, are currently intractable owing to the difficulty of estimating conformational entropy. Of further worry are intrinsic problems with treating polar interactions. For example, application of aggressive optimization methods to small proteins is beginning to reveal nonnative structures assigned lower all-atom energies than native structures for a subset of cases, including the chymotrypsin inhibitor and trp cage (R. Das & D. Baker, unpublished data). The native structures in these cases exhibit solvent-exposed hydrogen bonds, some involving charged residues; the energetics of these interactions may require taking into account explicit solvent, long-range electrostatic interactions and/or atomic polarizability. These types of polar interactions are more prevalent at active sites of proteins as hot spot interactions in protein-protein interfaces (48) and as mediators of protein/nucleic acid interactions (49). We have correspondingly found that prediction and design of polar interactions have been more challenging than problems for systems stabilized primarily by nonpolar interactions. Ongoing efforts to explicitly model discrete water molecules (50) and to develop polarizable force fields should contribute to improving prediction and design of challenging, highly polar systems (51).

## Toward RNA and Other Heteropolymers?

Although most of the work described above has focused on protein molecules, other polymers with intricate structures, notably RNA molecules, have important roles in modern-day viruses and cells and may have dominated biology in its primordial stages (52). More broadly, a vast range of nonbiological

heteropolymers with functional structures can now potentially be created and evolved, thanks to advances in combinatorial chemistry (as reviewed in Reference 53). Even though random sequences of these other polymers are not expected to have the energy gaps that aid current protein modeling, molecules under selection, either in vivo or in vitro, may evolve these properties in order to robustly carry out their selected function and thus be describable by the same basic ingredients developed for protein modeling in Rosetta.

Thus, naturally occurring RNA molecules that attain structure without protein cofactors, such as a recently discovered class of "riboswitches" (54), are potentially excellent targets. Initial studies of automated structure prediction with fragment assembly of RNA guided by a low-resolution energy function suggest that a fairly comprehensive conformational search can be carried out for sequences of less than 30 residues (22). For larger RNAs, the secondary structure (Watson-Crick base pairing pattern) of these molecules can typically be inferred from energy-based or phylogenetic analysis, dramatically decreasing the amount of conformational space that needs to be searched (55). By analogy with the history of protein structure prediction, an important next step is the development of a high-resolution energy function that provides a quantitative and reasonably accurate accounting of base stacking and hydrogen bonds. It will be interesting to see whether such energy terms will be accurate enough, and whether functional RNAs have evolved large enough energy gaps, to allow computational discrimination of functional structures from nonnative structures. The seeming prevalence of kinetic traps and alternative structures in the folding of large functional RNAs (56) suggests that the discrimination problem may not be as simple as for proteins.

Looking further in the future, nonnatural polymers such as nucleic acids with alternative bases (see, e.g., References 57 and 58), peptide nucleic acids (59), or peptoids [N-linked polyglycine (60)] may gain in importance. Long cooperatively folding sequences of these new heteropolymers may be evolvable in vitro to become enzymes, drugs, or nanotechnology scaffolds with useful properties orthogonal to existing biopolymers. The de novo prediction of these new molecules' secondary and tertiary structures—without prior extensive crystallographic information on sequences of the same polymer—will provide powerful tests of our understanding of molecular biophysics. The de novo prediction and discovery of the fundamental secondary and tertiary structure motifs for these new classes of heteropolymers—investigations that parallel the classic work on proteins by Pauling and colleagues or on nucleic acids by Watson and Crick—will be exciting studies. We expect that carrying out calculations on these new polymers will benefit greatly from the basic insights and powerful methods developed for protein structure prediction.

## CAN MOLECULAR MODELING BE A PRACTICAL TOOL?

The advent of a unified approach to multiple molecular modeling problems (**Figure 3**) suggests that computational approaches may soon have a broad impact on biology and medicine. Applications of computational design to molecular engineering are being explored widely and can be validated by testing whether the sequences carry out their desired function. However, for high-resolution structure prediction problems, high-resolution validation will typically not be available. How much can a practicing investigator trust any theoretical model? As described above, limitations of modeling produced by the approximate energy function and limited conformational search will likely bedevil structure predictions for many years to come. Nevertheless, molecular modeling can provide a reliable basis to guide and inform research if it can be coupled creatively and rigorously to experimental methods. Although these are early days, several initial investigations suggest

that hybrids between structure prediction and both high-resolution and low-resolution experimental approaches can accelerate the inference of trustworthy structural models.

## The Phase Problem in X-ray Crystallography

One particularly fruitful interface between high-resolution macromolecular modeling and high-resolution structure determination involves determining the phase estimates required for converting diffraction data into electron density maps. Although such phases are typically inferred from experiments on heavy-atom derivatives of proteins, more rapid solutions can be obtained by molecular replacement if a model with high structural similarity to the crystallized protein is available (61, 62). In the absence of prior sufficiently accurate structures, high-resolution comparative modeling is now able to bolster the success rate for molecular replacement in favorable cases in which the structures of even quite distant homologs are available (see, e.g., References 63–65). Recent work has also raised the exciting possibility of "de novo phasing" of diffraction data for proteins with de novo models (**Figure 4a**), although the successes so far have involved easy targets with significant noncrystallographic symmetry (66) or high solvent content crystals and few molecules in the asymmetric unit (65, 66). A largely unexplored frontier is the use of diffraction data earlier in the modeling; for example, likelihood scores for data without phases or with only weak phase information (67) may provide extra energy terms for refinement. Such enhancements will likely be important for making high-resolution computational modeling a practical addition to the arsenal available for crystallographic phasing.

## High-Resolution NMR Structures from Limited Data

Beyond applications in crystallographic phasing, molecular modeling may have even more

to contribute to high-resolution structural inference on the basis of NMR spectra. In addition to potentially increasing the rate of resonance assignment (68–70), energy-based refinement now appears capable of consistently improving the backbone accuracy and core side chain packing of medium-resolution NMR models (**Figure 4b**) (65). For small proteins, very high-throughput high-resolution structure determination using only chemical shift data during fragment assembly is another exciting prospect (71, 72). Establishing tests of refinement accuracy, e.g., using correlations of final structures to atom-atom contact data that are set aside during modeling or using packing and buried hydrogen bonding quality measures discussed below, is a critical challenge that needs to be addressed before these approaches to structural inference can come into wide use.

## High-Throughput Techniques

As is apparent from accurate reconstructions on the basis of limited NMR chemical shift measurements (71, 72), high-resolution molecular modeling can benefit from even small quantities of external data to constrain the conformational space that needs to be searched. It is thus exciting to note major progress in "lightweight," high-throughput experimental approaches that can potentially yield such constraints for every molecule or molecular complex that can be expressed and purified. Hydroxyl radical footprinting (73) and hydrogen-deuterium exchange (74) require small amounts of sample and sample preparation times. Via mass spectrometry readout, these methods give rapid information on burial of side chains and backbone hydrogen bonds, respectively (**Figure 4c**). Effective use of this burial data to guide modeling, for example, as a score in de novo fragment assembly, has not been demonstrated and remains an interesting challenge. Distance constraints from multiplexed cross-linking technologies (see, e.g., Reference 75) would be more easily incorporated into

structure predictions; the combined efforts of several groups will hopefully make high-throughput cross-linking a widely applicable method in the near future. If such ap-

proaches for residue-level structural information can be carried out rapidly and generally for protein domains and protein-protein interactions, these data will be of great use for



█ NMR model
█ Lowest energy structure
█ High-resolution crystal structure

**a**

**b**

**c**

**d**

█ Buried side chains
☐ Side chains exposed to solvent
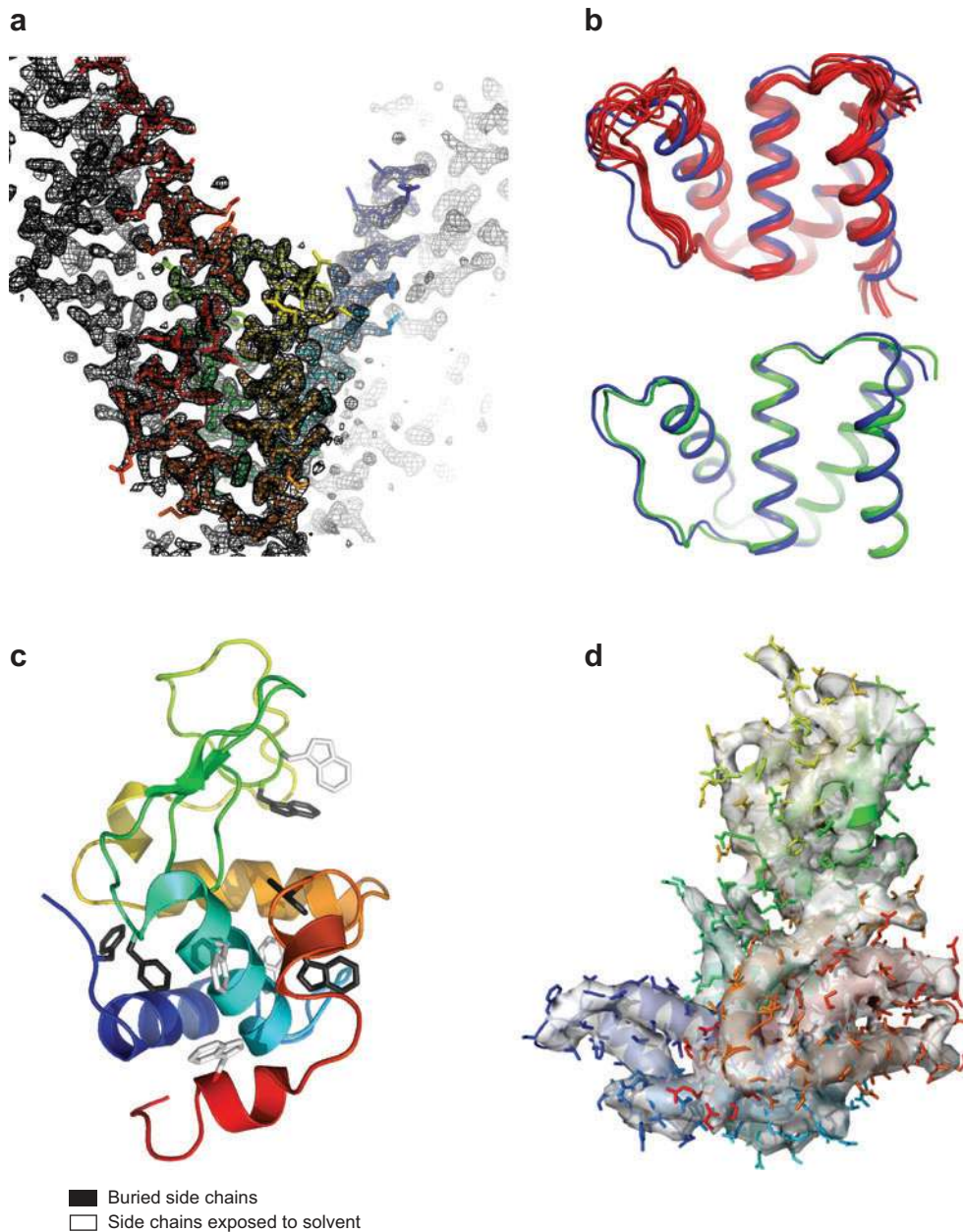
rapid validation of structure predictions; using these data to actually guide high-resolution structural inference is a nascent and exciting frontier.

## From Low-Resolution Maps to High-Resolution Models?

Perhaps the most rapidly growing new source of low-resolution information on biomolecular complexes is cryoelectron microscopy (76); this technique presents outstanding opportunities for all-atom molecular modeling methods. In the best of these maps, nucleic acid double helices and protein α-helices and β-sheets become clearly identifiable, and individual β-strands and connections between secondary structure elements can sometimes be inferred. Several groups are developing methods for automated or manual annotation of map features and connectivity (see, e.g., Reference 76). A natural next step is the high-resolution modeling of side chains and optimization of backbone coordinates, followed by energy-based refinement and selection (**Figure 4d**) (see Reference 77). Although maps typically involve large complexes, the conformational space that needs to be sampled can be greatly reduced if the secondary structure elements can be approximately located. For high-resolution refinement methods to be credible, strategies to avoid artifacts from map noise, from inaccuracies in the all-

atom energy function, and from incorrect or ambiguous secondary structure assignments need to be developed. If these challenges can be met, all-atom modeling may dramatically extend the resolution of cryoelectron microscopy from maps at subnanometer resolution (4–8 Å) to models of protein domains and protein-protein interactions approaching the high resolution offered by crystallography and NMR.

## Achieving Confident Models

We have outlined several exciting connections of experimental methods to molecular modeling. In each case, use of actual data bolsters the credibility of the final models. Still, how can one be sure that artifacts are not introduced by the incorrect energy terms or incomplete conformational searches in the modeling procedure? For example, molecular replacement phases for crystallographic data can introduce significant model bias, unless care is taken to cross-validate final coordinates with diffraction data not used during map refinement. Perhaps a more troubling scenario would be a high-resolution model derived from low-resolution cryoelectron map density because, generally, no independent data would be available to confirm model coordinates in atomic detail.

In the structure prediction field, assessing confidence in molecular models is not a precise science. In our own experience,

---

**Figure 4**

Connecting high-resolution molecular modeling to experimental structural inference. (*a*) Crystallographic phasing from a de novo model. Electron density map ($2m$F$_o$-DF$_c$; $2\sigma$ contour) for T0283 diffraction data (see **Figure 1b**), phased by molecular replacement with the Rosetta model and refined automatically, agrees with coordinates deposited in the Protein Data Bank using experimentally derived phases (*rainbow sticks*) (65). (*b*) Improving the accuracy of an NMR ensemble. Rebuilding and refinement of NMR models (*red*) (88) guided by the Rosetta energy function yields the lowest-energy structure (*green*) in better agreement with the high-resolution crystal structure of the same protein (*dark blue*) (65, 89). (*c*) Illustration of hydroxyl radical footprint information for hen egg lysozyme (1e8l) (90). Experimental data indicate which phenylalanine and tryptophan side chains are buried (*black*) or exposed to solvent (*white*) and may be useful in guiding structure prediction. (*d*) Medium-resolution cryoelectron microscopy maps may provide sufficient information for determining high-resolution models. This is a simulated map for a capsid protein from the rice dwarf virus (see Reference 91) overlaid on a high-resolution structure (92).

convergence of independent prediction runs to similar lowest-energy structures is generally a hallmark of accuracy (78), but there are exceptions, such as the highly polar chymotrypsin inhibitor case mentioned above. Interestingly, the tools commonly used to double-check the accuracy of NMR or crystal structures lose discrimination in assessing models produced by de novo molecular modeling. For example, completely inaccurate models from Rosetta all-atom structure predictions typically pass the tests for side chain clashes, side chain rotamers, and backbone configurations in the MolProbity package, essentially by construction (65; R. Das, D. Baker, S. Raman, unpublished results). One new possibility derives from the near universality of the fundamental features of folded macromolecular structures described above; we have found that statistical measures for the overall atomic packing and extent of hydrogen bonding of buried polar atoms are able to discriminate quite well between native structures and incorrect models (W. Sheffler, unpublished results). It is exceptionally difficult to produce an incorrect protein structure model with perfectly packed core side chains and no unsatisfied buried polar atoms. The development of robust metrics for evaluating model accuracy is an active area of current research (see, e.g., References 79 and 80). Nevertheless, we would expect no investigator to believe molecular models, even those estimated to have high accuracy, unless they are confronted with independent experimental tests. We therefore propose two steps that the molecular modeling field needs to take before de novo rebuilding and all-atom refinement tools are taken seriously and used widely.

First, investigators need to develop procedures that always set aside a fraction of the available data that will not be utilized during modeling and will thus give a reasonably independent measure of accuracy at the end of modeling. The obvious paradigm here is the computation, publication, and discussion of $R_{free}$ values (81) for diffraction data that is now carried out universally by the crystallography community. It is not difficult to imagine similarly rigorous protocols, dividing data into training and test sets, for de novo modeling methods guided by NMR chemical shift data or mass spectrometric data on residue burial. Envisioning the details of such a cross-validation approach for high-resolution modeling with cryoelectron microscopy maps is more difficult and remains an important challenge for the field.

Second, progress in experimentally coupled structural inference may be best catalyzed through blind trials, much as the CASP and the Critical Assessment of Prediction of Interactions (CAPRI) have promoted rapid progress in "experiment-blind" structure prediction. Perhaps the sequences of a subset of CASP targets with impending crystal structures could be given to predictors along with the available X-ray diffraction data but without experimental phases. For NMR cases, partial NMR chemical shift data—but not additional atom-atom contact data from the nuclear Overhauser effect or residual dipolar coupling data—might be made available. Independent assessors can then check whether the resulting structural models superimpose well on subsequently released structures solved with traditional structure determination methods that require more data, time, and experimental expense. Success in blind trials would lend much credibility to these new molecular modeling approaches that have the potential to increase the resolution and efficiency of structural inference.

## MACROMOLECULAR MODELING FOR THE COMMUNITY

The advances described in this review are the collective work of many scientists and research groups throughout the world who are collaborating on further developing and improving the approaches described in this review. All of the

resulting code is available freely to academic users (**http://www.rosettacommons.org/**). One of the current goals of this Rosetta development team is the creation of a streamlined version of the software that will allow the implementation of new protocols using new atom trees, via intuitive interaction with a graphical user interface as well as via simple scripts. We expect such a user-friendly toolkit for prediction and design to not only accelerate molecular modeling research for expert Rosetta users but also to help bring these tools into wider use throughout the biological community.

## SUMMARY POINTS

1. Many of the basic physical principles underlying biomolecular structure and interactions appear reasonably well understood. Success in molecular modeling requires fairly accurate energy functions which embody these principles as well as effective sampling methodologies for identifying very low-energy structures (in prediction problems) and very low-energy sequences (in design problems).

2. The energy function and sampling methodologies in the Rosetta program provide a unified framework for the prediction and design of macromolecular structures and interactions, with recent examples ranging from fibril structure prediction to RNA folding to the design of new enzyme catalysts.

3. In favorable cases, de novo structure prediction can approach atomic resolution, protein structure models can be refined to higher resolution, and novel proteins with new and useful functions can be designed.

4. Combination of high-resolution molecular modeling with experimental structural methods has the potential to improve crystallographic phasing, NMR structural inference, and lower-resolution approaches that make use of high-throughput mass spectrometric data and cryoelectron microscopy maps.

## FUTURE ISSUES

1. Improvements in conformational search strategies are required to go beyond the smallest biomolecules and the simplest modeling problems, along with making these approaches more widely available to investigators with access to modest computational power.

2. Major improvements in the treatment of polar interactions are critical for the accurate prediction and design of enzyme active sites, solvent-exposed loops, and protein/nucleic acid interfaces.

3. Feedback from applications of high-resolution molecular modeling to an ever widening array of prediction and design problems, including the modeling of long heteropolymers, such as RNA and polypeptoids, will continue to provide rigorous tests of the underlying physical principles and stimulate continued improvement of computational methods.

4. Precise statistical assessment measures and blind validation trials are important next steps for promoting experimentally coupled molecular modeling into a practical tool for accelerated structural inference.

## DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENT

## LITERATURE CITED

1. Rohl CA, Strauss CE, Misura KM, Baker D. 2004. *Methods Enzymol.* 383:66–93
2. Bradley P, Misura KM, Baker D. 2005. *Science* 309:1868–71
3. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, et al. 2005. *Proteins* 61(Suppl. 7):128–34
4. Das R, Qian B, Raman VS, Vernon R, Thompson J, et al. 2007. *Proteins.* 69:S118–28
5. Baker D, Sohl JL, Agard DA. 1992. *Nature* 356:263–65
6. Baldwin RL. 2007. *J. Mol. Biol.* 371:283–301
7. Dill KA. 1990. *Biochemistry* 29:7133–55
8. Kauzmann W. 1959. *Adv. Protein Chem.* 14:1–63
9. Nelson R, Sawaya MR, Balbirnie M, Madsen AØ, Riekel C, et al. 2005. *Nature* 435:773–78
10. Watson JD, Crick FH. 1953. *Nature* 171:737–38
11. Morozov AV, Kortemme T, Tsemekhman K, Baker D. 2004. *Proc. Natl. Acad. Sci. USA* 101:6946–51
12. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. 2005. *Science* 310:638–42
13. Kortemme T, Morozov AV, Baker D. 2003. *J. Mol. Biol.* 326:1239–59
14. Boas FE, Harbury PB. 2007. *Curr. Opin. Struct. Biol.* 17:199–204
15. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, et al. 1995. *J. Am. Chem. Soc.* 117:5179–97
16. Plaxco KW, Simons KT, Baker D. 1998. *J. Mol. Biol.* 277:985–94
17. Thompson JB, Hansma HG, Hansma PK, Plaxco KW. 2002. *J. Mol. Biol.* 322:645–52
18. Sullivan DC, Kuntz ID. 2004. *Biophys. J.* 87:113–20
19. Englander SW. 2000. *Annu. Rev. Biophys. Biomol. Struct.* 29:213–38
20. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. 2003. *Science* 302:1364–68
21. Simons KT, Kooperberg C, Huang E, Baker D. 1997. *J. Mol. Biol.* 268:209–25
22. Das R, Baker D. 2007. *Proc. Natl. Acad. Sci. USA* 104:14664–69
23. Bradley P, Baker D. 2006. *Proteins* 65:922–29
24. Wang C, Bradley P, Baker D. 2007. *J. Mol. Biol.* 373:503–19

25. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, et al. 2003. *Proteins* 53(Suppl. 6):491–96

26. Abagyan R, Totrov M, Kuznetsov D. 1994. *J. Comput. Chem.* 15:488–506

27. Rice LM, Brunger AT. 1994. *Proteins* 19:277–90

28. Schwieters CD, Clore GM. 2001. *J. Magn. Reson.* 152:288–302

29. Rohl CA, Strauss CE, Chivian D, Baker D. 2004. *Proteins* 55:656–77

30. Chivian D, Baker D. 2006. *Nucleic Acids Res.* 34:e112

31. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. 2006. *Proc. Natl. Acad. Sci. USA* 103:5361–66

32. Janin J. 2005. *Protein Sci.* 14:278–83

33. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, et al. 2003. *J. Mol. Biol.* 331:281–99

34. Wollacott AM, Zanghellini A, Murphy P, Baker D. 2007. *Protein Sci.* 16:165–75

35. André I, Bradley P, Wang C, Baker D. 2007. *Proc. Natl. Acad. Sci. USA* 104:17656–61

36. Meiler J, Baker D. 2006. *Proteins* 65:538–48

37. Barth P, Schonbrun J, Baker D. 2007. *Proc. Natl. Acad. Sci. USA* 104:15682–87

38. Ambroggio XI, Kuhlman B. 2006. *J. Am. Chem. Soc.* 128:1154–61

39. Butterfoss GL, Kuhlman B. 2006. *Annu. Rev. Biophys. Biomol. Struct.* 35:49–65

40. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D. 2004. *Nat. Struct. Mol. Biol.* 11:371–79

41. Zanghellini A, Jiang L, Wollacot A, Cheng G, Meiler J, et al. 2006. *Protein Sci.* 15:2785–94

42. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, et al. 2008. *Science* 319:1387–91

43. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ Jr, et al. 2006. *Nature* 441:656–59

44. Misura KM, Baker D. 2005. *Proteins* 59:15–29

45. Dantas G, Corrent C, Reichow SL, Havranek JJ, Eletr ZM, et al. 2007. *J. Mol. Biol.* 366:1209–21

46. Wang C, Schueler-Furman O, Andre I, London N, Fleishman SJ, et al. 2007. *Proteins* 69:758–63

47. Schueler-Furman O, Wang C, Baker D. 2005. *Proteins* 60:187–94

48. Kortemme T, Baker D. 2002. *Proc. Natl. Acad. Sci. USA* 99:14116–21

49. Havranek JJ, Duarte CM, Baker D. 2004. *J. Mol. Biol.* 344:59–70

50. Jiang L, Kuhlman B, Kortemme T, Baker D. 2005. *Proteins* 58:893–904

51. Ren P, Ponder JW. 2002. *J. Comput. Chem.* 23:1497–506

52. Gesteland RF, Cech TR, Atkins JF. 2006. *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*. Cold Spring Harbor, NY: Cold Spring Harb. Lab.

53. Wrenn SJ, Harbury PB. 2007. *Annu. Rev. Biochem.* 76:331–49

54. Winkler WC, Breaker RR. 2003. *ChemBioChem* 4:1024–32

55. Shapiro BA, Yingling YG, Kasprzak W, Bindewald E. 2007. *Curr. Opin. Struct. Biol.* 17:157–65

56. Treiber DK, Williamson JR. 1999. *Curr. Opin. Struct. Biol.* 9:339–45

57. Liu H, Gao J, Lynch SR, Saito YD, Maynard L, Kool ET. 2003. *Science* 302:868–71

58. Benner SA, Battersby TR, Eschgfaller B, Hutter D, Kodra JT, et al. 1998. *Pure Appl. Chem.* 70:263–66

59. Buchardt O, Egholm M, Berg RH, Nielsen PE. 1993. *Trends Biotechnol.* 11:384–86

60. Kirshenbaum K, Barron AE, Goldsmith RA, Armand P, Bradley EK, et al. 1998. *Proc. Natl. Acad. Sci. USA* 95:4303–8

61. Rossmann MG, Blow DM. 1962. *Acta Crystallogr.* 15:24–31
62. Read RJ. 2001. *Acta Crystallogr. D* 57:1373–82
63. Schwarzenbacher R, Godzik A, Grzechnik SK, Jaroszewski L. 2004. *Acta Crystallogr. D* 60:1229–36
64. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. 2005. *Proteins* 61(Suppl. 7):3–7
65. Qian B, Raman VS, Das R, Bradley P, Mccoy AJ, et al. 2007. *Nature* 450:259–64
66. Strop P, Brzustowicz MR, Brunger AT. 2007. *Acta Crystallogr. D* 63:188–96
67. Dauter Z. 2002. *Curr. Opin. Struct. Biol.* 12:674–78
68. Meiler J, Baker D. 2003. *Proc. Natl. Acad. Sci. USA* 100:15404–9
69. Bowers PM, Strauss CE, Baker D. 2000. *J. Biomol. NMR* 18:311–18
70. Rohl CA. 2005. *Methods Enzymol.* 394:244–60
71. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. 2007. *Proc. Natl. Acad. Sci. USA* 104:9615–20
72. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, et al. 2008. *Proc. Natl. Acad. Sci. USA* 105:4685–90
73. Takamoto K, Chance MR. 2006. *Annu. Rev. Biophys. Biomol. Struct.* 35:251–76
74. Woods VL Jr, Hamuro Y. 2001. *J. Cell Biochem. Suppl.* (Suppl. 37):89–98
75. Sinz A. 2006. *Mass Spectrom. Rev.* 25:663–82
76. Chiu W, Baker ML, Jiang W, Dougherty M, Schmid MF. 2005. *Structure* 13:363–72
77. Topf M, Baker ML, Marti-Renom MA, Chiu W, Sali A. 2006. *J. Mol. Biol.* 357:1655–68
78. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, et al. 2002. *J. Mol. Biol.* 322:65–78
79. Wallner B, Elofsson A. 2006. *Protein Sci.* 15:900–13
80. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. 2007. *Proteins* 69(Suppl. 8):175–83
81. Brunger AT. 1997. *Methods Enzymol.* 277:366–96
82. Mondragon A, Subbiah S, Almo SC, Drottar M, Harrison SC. 1989. *J. Mol. Biol.* 205:189–200
83. Willcox BE, Thomas LM, Bjorkman PJ. 2003. *Nat. Immunol.* 4:913–19
84. Kammerer RA, Kostrewa D, Progias P, Honnappa S, Avila D, et al. 2005. *Proc. Natl. Acad. Sci. USA* 102:13891–96
85. Arevalo JH, Taussig MJ, Wilson IA. 1993. *Nature* 365:859–63
86. Egli M, Minasov G, Su L, Rich A. 2002. *Proc. Natl. Acad. Sci. USA* 99:4302–7
87. Hennig M, Darimont BD, Jansonius JN, Kirschner K. 2002. *J. Mol. Biol.* 319:757–66
88. Andersen KV, Poulsen FM. 1993. *J. Biomol. NMR* 3:271–84
89. van Aalten DM, Milne KG, Zou JY, Kleywegt GJ, Bergfors T, et al. 2001. *J. Mol. Biol.* 309:181–92
90. Schwalbe H, Grimshaw SB, Spencer A, Buck M, Boyd J, et al. 2001. *Protein Sci.* 10:677–88
91. Zhou ZH, Baker ML, Jiang W, Dougherty M, Jakana J, et al. 2001. *Nat. Struct. Biol.* 8:868–73
92. Nakagawa A, Miyazaki N, Taka J, Naitow H, Ogawa A, et al. 2003. *Structure* 11:1227–38

# Contents

**Indexes**

**Errata**

An online log of corrections to *Annual Review of Biochemistry* articles may be found
at http://biochem.annualreviews.org/errata.shtml