Evolution and Physics in Comparative Protein Structure Modeling

ANDRAS FISER,† MICHAEL FEIG,‡
CHARLES L. BROOKS, III,‡ AND ANDREJ SALI*,†

Laboratory of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021, and The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Received September 17, 2001

ABSTRACT

From a physical perspective, the native structure of a protein is a consequence of physical forces acting on the protein and solvent atoms during the folding process. From a biological perspective, the native structure of proteins is a result of evolution over millions of years. Correspondingly, there are two types of protein structure prediction methods, de novo prediction and comparative modeling. We review comparative protein structure modeling and discuss the incorporation of physical considerations into the modeling process. A good starting point for achieving this aim is provided by comparative modeling by satisfaction of spatial restraints. Incorporation of physical considerations is illustrated by an inclusion of solvation effects into the modeling of loops.

Introduction

Three-dimensional (3D) structure of natural proteins is guided by two distinct sets of principles operating on vastly different time scales: the laws of physics and the theory of evolution (Figure 1). On one hand, according to the laws of physics, a protein molecule in solution is a system of atoms that interact through a variety of forces, such as chemical bonds, hydrogen bonds, Coulomb interactions, and Lennard-Jones forces. Under appropriate conditions, these forces fold almost any random starting conformation of a protein into a stable, well-defined 3D structure (i.e., the native state) in a matter of milliseconds or seconds. On the other hand, over millions of years, evolution resulted in families of proteins that share similar sequences, similar structures, and often related functions. Different proteins evolved through duplication, speciation, and horizontal transfer, followed by accumulation of mostly neutral mutations. In evolution, a protein changes gradually because it usually needs to retain its function, which requires the conservation of its structure and therefore also its sequence.[1]

Each of the two sets of principles that apply to the natural protein sequences gave rise to a class of protein structure prediction methods.[2] The first approach, de novo or ab initio methods, predicts the structure from sequence alone, without relying on similarity at the fold level between the modeled sequence and any of the known structures.[3] The de novo methods assume that the native structure corresponds to the global free energy minimum accessible during the lifespan of the protein and attempt to find this minimum by an exploration of many conceivable protein conformations. The two key components of the de novo methods are the procedure for efficiently carrying out the conformational search and the free energy function used for evaluating possible conformations. The second class of methods, including threading and comparative modeling, rely on detectable similarity spanning most of the modeled sequence and at least one known structure.[4,5] When the structure of one protein in the family has been determined by experiment, the other

András Fiser received his M.Sc. degree in chemistry from the Eotvos University of Budapest, Hungary, in 1991. During his subsequent studies at the Institute of Enzymology, he was supported by an award from the Scientific Qualificatory Committee of the Hungarian Academy of Sciences and by an award from the Foundation For Hungarian Science. He also studied at the Laboratories of Molecular Biophysics at the Oxford University, Oxford, England, from 1993 to 94, where he was supported by the Soros Foundation. He received his Ph.D. from the Hungarian Academy of Sciences in 1997, under the supervision of Prof. Istvan Simon. He focused on development of methods for prediction of spatial contacts between residues that are sequentially distant in the amino acid sequence. He then joined the group of Prof. Andrej Sali at The Rockefeller University, where he was a Burroughs Wellcome Fund Postdoctoral Fellow and is a Charles Revson Foundation Postdoctoral Fellow. His current interests include improvement and applications of comparative protein structure modeling.

Michael Feig received his Diplom degree in physics and computer science from the Technical University of Berlin, Berlin, Germany, in 1994. He went on to study nucleic acid structures and solvation with molecular dynamics simulations with Prof. B. Montgomery Pettitt at the University of Houston where he received his Ph.D. in chemistry in 1999. He subsequently joined the group of Prof. Charles Brooks III at The Scripps Research Institute as a postdoctoral research associate to work on multiscale modeling methods for protein structure prediction and refinement within the MMTSB (multiscale modeling tools in structural biology) NIH Research Resource.

Charles L. Brooks III received B.Sc. degrees in chemistry and physics from Alma College, Alma, MI, in 1978. He studied nonequilibrium statistical mechanics and received his Ph.D. in Physical Chemistry with Stephen A. Adelman in 1982. He was a postdoctoral associate at Harvard University from 1982 to 1985 working with Martin Karplus. Dr. Brooks was the recipient of an NIH postdoctoral fellowship at Harvard from 1983 to 1985. During his tenure at Harvard he developed the stochastic boundary simulation methods for biopolymer dynamics studies. He joined the faculty of the Department of Chemistry at Carnegie Mellon University in 1985, and he was promoted to the rank of Professor there in 1994. He joined the faculty of The Scripps Research Institute later that year. He currently directs research efforts in the areas of protein folding, multiscale modeling, drug design and discovery, and enzyme catalysis at TSRI. Professor Brooks was recognized as an Alfred P. Sloan Research Fellow and was recently elected a Fellow of the American Association for the Advancement of Sciences.

Andrej Sali received his B.Sc. degree in chemistry from the University of Ljubljana, Ljubljana, Slovenia, in 1987. He was awarded the Research Council of Slovenia Scholarship, the Overseas Research Students Award, and the Merck Sharpe and Dohm Academic Scholarship at Birkbeck College, University of London, where he received his Ph.D. in biophysics in 1991, under the supervision of Prof. Tom L. Blundell. He focused on development of methods for comparative modeling of protein three-dimensional structure and their implementation in the program MODELLER. He then went to the Department of Chemistry at Harvard University as a Jane Coffin Childs Memorial Fund Postdoctoral Fellow with Prof. Martin Karplus, where he continued to develop comparative modeling methods and also studied simple lattice Monte Carlo models of protein folding. Since 1995, Dr. Sali has been at The Rockefeller University, where he is now an Associate Professor. He was a Sinsheimer Scholar and an Alfred P. Sloan Research Fellow and is an Irma T. Hirschl Trust Career Scientist. He aims to improve and apply methods for (i) predicting the structures of proteins, (ii) determining the structures of macromolecular assemblies, and (iii) annotating the functions of proteins using their structures.

10.1021/ar010061h CCC: $22.00
Published on Web 04/10/2002

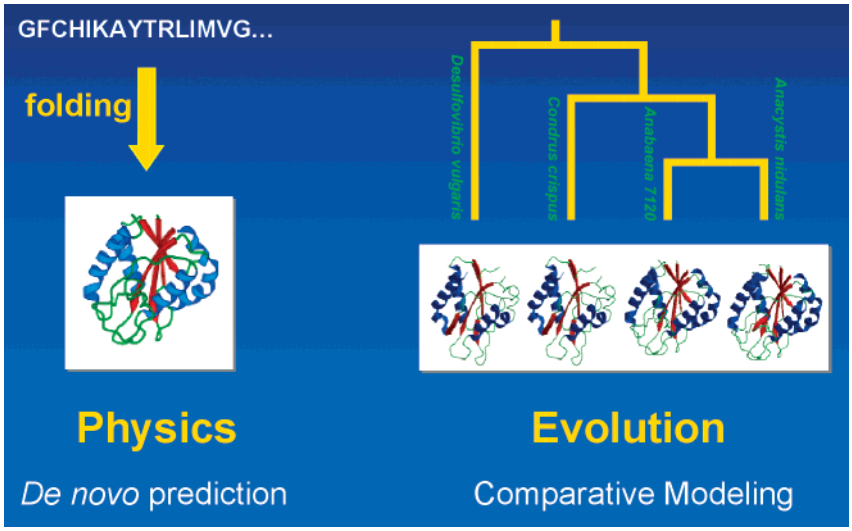

FIGURE 1. De novo structure prediction and comparative protein structure modeling. Proteins obey two distinct sets of principles, the laws of physics and the theory of evolution, each giving rise to the corresponding variety of protein structure prediction methods.

members of the family can be modeled on the basis of their alignment to the known structure.

It is useful to describe de novo prediction and comparative modeling within the same conceptual framework. Any protein structure prediction method can be seen as an optimization of a protein structure model with respect to a certain objective function.[6] The methods differ in the function optimized, in the model representation (including the degrees of freedom), and in the method of optimization (including the starting conformation). The individual terms of the objective function, such as Coulomb interactions in a molecular mechanics force field or side-chain dihedral angle restraints in comparative modeling, correspond to the individual spatial restraints. In this view, de novo methods attempt to find the most likely structure of a protein sequence given the forces between its atoms, while the comparative methods attempt to find the most likely structure of a protein sequence given primarily its relationship to the known similar structures. When the aim is to obtain the most accurate protein structure prediction, protein modeling by satisfaction of spatial restraints provides a framework for incorporating all spatial information about a given protein sequence, irrespective of its origin, be it physics-based, homology-based, or derived by experiment.

In this perspective, we begin by describing the essential features of comparative protein structure prediction. We introduce our approach to comparative modeling by satisfaction of spatial restraints derived mainly from the alignment of the modeled sequence with related protein structures. We then generalize this approach to spatial restraints from arbitrary sources and continue by exploring physical considerations in comparative modeling. Finally, we finish by describing our initial effort to include solvation effects into the modeling of loops.

* Corresponding author. E-mail: sali@rockefeller.edu. Tel: +1 (212) 327-7550. Fax: +1 (212) 327-7540. Web: http://guitar.rockefeller.edu.
† The Rockefeller University.
‡ The Scripps Research Institute.

Comparative Modeling and Threading

Modeling of a sequence on the basis of known structures generally consists of four steps: finding known structures related to the target sequence to be modeled (i.e., templates); aligning the sequence with the templates; building a model; assessing the model.[5] In this section, we list various comparative modeling approaches, errors in the resulting models, and their applications. We also point out the recent trends in large-scale comparative modeling of whole genomes and all known protein sequences.

The 3D structures of proteins from the same family are more conserved than their primary sequences.[1] Therefore, if similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed. Moreover, proteins that share low or even nondetectable sequence similarity often also have similar structures. The templates for modeling may be found by sequence comparison methods, such as PSI-BLAST,[7] or by sequence−structure threading methods[8] that can sometimes reveal more distant relationships than purely sequence-based methods. In the latter case, fold assignment and alignment are achieved by threading the sequence through each of the structures in a library of all known folds that are representative of the Protein Data Bank (PDB) of all available protein structures.[9] Each sequence−structure alignment is assessed by the energy of a corresponding coarse model and not by sequence similarity as in sequence comparison methods. While threading methods find known structures related to the input sequence and in the process also calculate an alignment between them, they do not result in an explicit atomic model of the sequence.

Comparative structure prediction produces an all-atom model of a sequence, based on its alignment to one or more related protein structures. Comparative model building includes either sequential or simultaneous modeling of the core of the protein, loops, and side chains. In the original comparative approach, a model is constructed

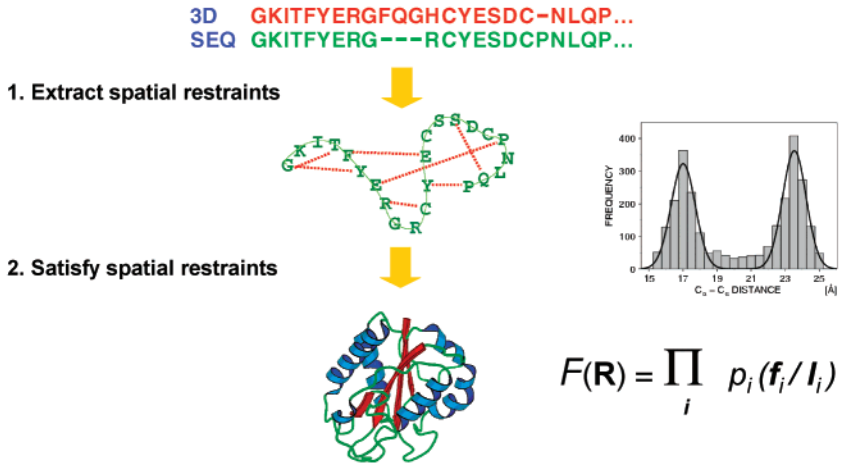

FIGURE 2. Comparative protein structure modeling by satisfaction of spatial restraints. First, spatial restraints are extracted from the input alignment, spatial preferences found in known protein structures, and a molecular mechanics force field. Second, all the restraints are combined into an objective function that is optimized to obtain the final model.

from a few template core regions and from loops and side chains obtained from either aligned or unrelated structures.[4,10,11] Another family of comparative methods relies on approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms.[12,13] A third group of methods uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the sequence−template alignment.[14−16] There are also many methods that specialize in the modeling of loops[17−19] and side chains[20,21] within the restrained environment provided by the rest of the structure.

The accuracy of comparative modeling is correlated with the percentage sequence identity on which the model is based, mimicking the correlation between the structural and sequence similarities of two proteins.[1,5,22] Errors include mistakes in side-chain packing, relatively small shifts and distortions in correctly aligned regions, errors in the unaligned regions (i.e., loops), alignment errors, and fold assignment mistakes. The alignment errors increase rapidly below 30% sequence identity and become the most significant origin of errors in comparative models. Errors in comparative modeling and threading are best quantified by continuous, automated, and large-scale assessment of automated prediction methods, such as that implemented by the LiveBench[23] and EVA web servers.[24]

Reasonable applications of any protein structure model depend on its accuracy, and even models with large errors can be helpful.[4,5] Comparative models have been used in studying catalytic mechanisms of enzymes, designing and improving ligands, docking of macromolecules, predicting interacting protein partners, virtual screening and docking of small ligands, defining antibody epitopes, molecular replacement in X-ray crystallography, designing chimeras, stable and crystallizable variants, supporting site-directed mutagenesis, refining NMR structures, fitting proteins into low-resolution electron density maps, finding functional sites by 3D motif searching, determining structure from sparse experimental restraints, annotating function from structural relationships, and finding patches of conserved surface residues.[5]

While the models can provide substantial insights, they can also be misleading. Thus, it is necessary to estimate the accuracy of a model before it is used. The accuracy of a comparative model can be estimated simply from sequence similarity to its template or more generally by a variety of model assessment methods.[25−27]

Domains in approximately half of all 600 000 known protein sequences were modeled with ModPipe,[22] relying on PSI-BLAST[7] and MODELLER,[15] and deposited into a comprehensive database of comparative models, ModBase (http://guitar.rockefeller.edu/modbase/).[28,29] While the current number of modeled proteins may look impressive, usually only one domain per protein is modeled (on the average, proteins have slightly more than two domains) and two-thirds of the models are based on less than 30% sequence identity to the closest template. The web interface to ModBase allows flexible querying for fold assignments, sequence−structure alignments, models, and model assessments of interest. An integrated sequence/structure viewer, ModView, allows inspection and analysis of the query results.[66] ModBase will be increasingly interlinked with other applications and databases such that structures and other types of information can be easily used for functional annotation.

The usefulness of comparative modeling is steadily increasing because the number of different structural folds that proteins adopt is limited and because the number of experimentally determined new structures is increasing rapidly.[30] This trend is accentuated by the recently initiated structural genomics project that aims to determine at least one structure for most protein families.[31] It is conceivable that structural genomics will achieve its aim in 5−10 years, making comparative modeling applicable to most protein sequences.

Comparative Modeling by Satisfaction of Spatial Restraints

We developed an automated approach to comparative protein structure modeling that is based on satisfaction

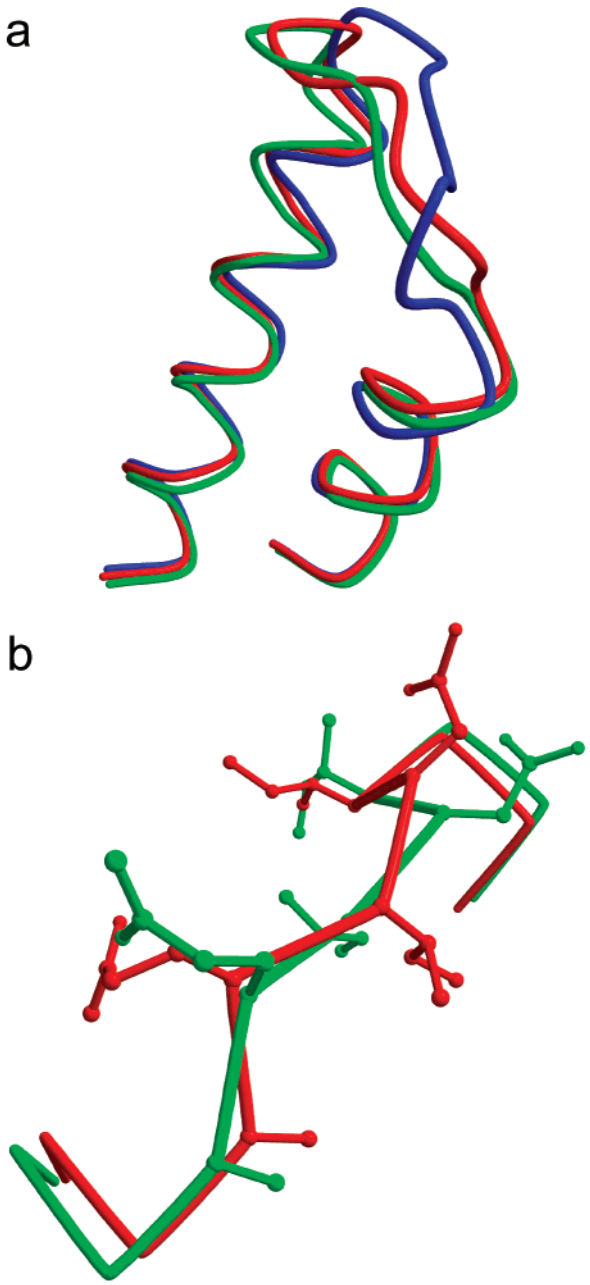

FIGURE 3. Loop modeling of the region 69−73 in *P. aerophilum* manganese superoxide dismutase. The target sequence is related strongly at 56% sequence identity to the template structure of *S. sulfolobus* iron superoxide dismutase, 1SSS (blue). The model (red) was constructed before the actual structure (target T0128 at CASP4) was defined by X-ray crystallography (green). The segment to be modeled de novo was identified from the divergence in sequence and structure among its three closest homologues of known structure. The loop model is significantly closer to the actual structure (the global main-chain rms error is 1.28 Å) than the best available template loop (2.69 Å). The loops are oriented by the superposition of the whole structures. Key: (a) C_α trace of the 5-residue loop spanned by the two stem regions, for the model, the actual structure, and the template; (b) all-atom representation of the loop residues in the model and the actual structure.

of spatial restraints[15,17] (Figure 3). It is implemented in the computer program MODELLER, which is freely available to academic researchers via the web at http://guitar.rockefeller.edu. In this section, we describe briefly the spatial restraints imposed on the target sequence and the optimization procedure that minimizes violations of the restraints to obtain a 3D model.

In the first step of model building, spatial restraints on the target sequence are calculated. In general, restraints are expressed as conditional probability density functions $P(\mathbf{f}/\mathbf{I})$ (pdf's) for the restrained spatial feature **f**, given several variables **I** that were found to be most predictive of the restrained feature. There are three types of a restraint, depending on the origin and nature of information **I**.

First, restraints are obtained for those target residues that are aligned with template residues. These homology-derived spatial restraints limit distances among the main-chain and side-chain atoms, as well as main-chain dihedral angles Φ, Ψ, and Ω and side-chain dihedral angles χ_i. The form of these restraints was obtained from a statistical analysis of 105 family alignments that included 416 structurally defined proteins.[32] For example, a restraint on a certain $C_\alpha-C_\alpha$ distance given equivalent distances in two related protein structures is described well by a weighted sum of two Gaussian functions corresponding to the two template distances, respectively (cf. the histogram in Figure 2).

The second class of restraints reflect statistical preferences extracted from known protein structures in general and are related to the statistical potentials of mean force.[33−36] The restraints depend only on the types of the restrained atoms or residues and not on the template structure. These restraints are applied to the main-chain and side-chain dihedral angles of target residues that are not aligned with template residues (i.e., in an inserted loop) and to distances between all nonbonded atom pairs. They are used because they were found to result in more accurate models than the corresponding terms from a molecular mechanics force field.[17]

The third type of a restraint is obtained from the molecular mechanics force field of CHARMM-22[37] and includes restraints on chemical bonds, angles, and improper dihedral angles. These molecular mechanics restraints enforce proper stereochemistry of the model.

After all pdf's are calculated, their logarithms are summed to obtain an objective function that depends on the model and gives its likelihood (see the equation below). Finally, the model containing all non-hydrogen atoms is calculated by optimizing the objective function in Cartesian space. The optimization is carried out by the variable target function method employing conjugate gradients and molecular dynamics with simulated annealing.

Protein Structure Modeling by Satisfaction of Spatial Restraints

We now describe the modeling approach implemented in MODELLER from a broader perspective; we outline how different kinds of spatial information can be used for protein structure prediction in general and not only for comparative modeling.

A 3D model of one or more molecules is obtained by minimizing the objective function F with respect to Cartesian coordinates of atoms $\mathbf{R}$:

$$F(\mathbf{R}) = -\ln \Pi P_i(\mathbf{f}_i/\mathbf{I}_i) = \Sigma E_i(\mathbf{f}_i, \mathbf{a}_i)$$

where P_i is a conditional probability density function for a geometric feature $\mathbf{f}_i$ that depends on information $\mathbf{I}_i$, E_i is the corresponding energy term, and $\mathbf{a}_i$ are parameters that generally vary from a term to a term and are related to $\mathbf{I}_i$. If the individual pdf's were statistically independent from each other, $F(\mathbf{R})$ would be the negative logarithm of the probability density of conformation $\mathbf{R}$. If the individual energy terms E_i were additive, $F(\mathbf{R})$ would be the energy of conformation $\mathbf{R}$. Clearly, the pdf and energy terms are related, $E_i = -\ln P_i$. The terms probability and energy are used loosely, without specifying the proportionality factors, the thermodynamic ensemble, or the type of energy.

Both the P_i and E_i terms can be seen as spatial restraints. For example, the "statistical" definition of restraints in terms of pdf's P_i is convenient when restraints are derived from a database of known related structures, and the "physical" definition of restraints in terms of the energy terms E_i is convenient when restraints are derived from the CHARMM force field. Although the statistical and physical definitions are equivalent, it may be easier to arrive at the correct restraint form using one or the other definition.[15]

In a typical comparative modeling calculation, there are on the order of 5000 atoms and 40 000 restraints. The form of E_i is simple; it can be a quadratic function, cosine, logarithm of the weighted sum of a few Gaussian functions, Coulomb law, Lennard-Jones potential, cubic spline function, and some other simple forms. The geometric features presently include a distance, an angle, a dihedral angle, and a pair of dihedral angles between two, three, four, and eight points, respectively. Points correspond either to real atoms or to pseudoatoms, such as a gravity center of several real atoms. A pair of dihedral angles can be used to restrain simultaneously such strongly correlated features as the main-chain dihedral angles Φ and Ψ of the same residue. Most terms in the CHARMM energy function are implemented in MODELLER. Molecular representations that correspond to any subset of an all-atom topology library of CHARMM (e.g., all-atom, non-hydrogen atoms, C_α-only) as well as a simplified side-chain model can be used.

Perhaps the main technical limitations presently are that the individual restrained features depend on a small number of atoms and that the first derivatives of the restraints with respect to Cartesian coordinates be calculated rapidly. These limitations make it difficult to optimize a model with respect to, for example, some representations of implicit solvation.

Combining Physics and Evolution To Improve Protein Structure Prediction

The similarity between the local fluctuations around the native state and the structural differences among the native structures of homologous proteins[38] suggests that accurate molecular dynamics simulations should improve the relatively inaccurate comparative models obtained with the existing methods. This hope is underpinned by the recent improvements in molecular mechanics force fields[37,39,40] and the ever increasing time scale of the molecular dynamics simulations.[41] Nevertheless, attempts to improve comparative models by molecular dynamics simulations have not yet been successful.[42]

A general approach to improving comparative modeling by physics-derived information is to add a physically correct model of interactions among atoms to the objective function that guides construction of a comparative model. The main role of the homology-derived terms in such an objective function would be to provide a suitable starting conformation and to restrain the conformational search to a manageable and relevant portion of the phase space, while the main role of the physical terms would be to allow the refinement of the model away from the template structure toward the actual structure of the target. Achieving a sufficient accuracy and a productive balance between these two types of a restraint is highly nontrivial, as evidenced by the difficulty of refining comparative protein structure models.[15,42] The emphasis on the objective function is justified because there is strong anecdotal evidence that the prediction accuracy is often limited by the accuracy of the energy function[15,17] and possibly by the accuracy of the corresponding protein and solvent representations. Thus, in the next three paragraphs, we comment on the current accuracy of the terms that describe local stereochemistry, nonbonded interactions between protein atoms, and nonbonded interactions between protein and solvent atoms (i.e., solvation effect).

Comparative methods use restraints on chemical bonds, bond angles, and dihedral angles that usually originate from a molecular mechanics force field or from general statistical preferences extracted from many known protein structures. It appears that these features of a protein are restrained with sufficient accuracy and do not require additional attention.

Comparative methods also generally apply nonbonded restraints on the protein atoms. These restraints are typically adopted from a molecular mechanics force field (e.g., Lennard-Jones interactions) and sometimes from statistical preferences obtained from known protein structures (e.g., atomic statistical potentials of mean force). For example, statistical potentials of mean force have been used successfully to assess comparative models,[26] to improve the accuracy of loop models,[17] and to improve the overall model accuracy.[16,27,43] In contrast to the sequentially local restraints, however, the nonbonded interactions between protein atoms are probably not modeled sufficiently accurately and their description needs to be improved.

And finally, comparative modeling lags significantly behind the state of the art in the description of solvation.[44−52] In fact, comparative methods generally do not even attempt to model solvation effects. There are only a

few studies that included solvation into assessment of models[53,54] and loop modeling.[55−58] Thus, it is most desirable to add to comparative modeling an accurate model of nonbonded interactions between the protein and solvent atoms.

Loop Modeling

In this section, we describe our preliminary results on using an implicit solvation model to improve the modeling of loops in protein structures, after we introduce the problem of loop modeling and our existing approach implemented in MODELLER.

Errors in loops are the dominant problem in comparative modeling above 35% sequence identity. In this range of overall similarity, loops among the homologues vary while the core regions are still relatively conserved and aligned accurately. There are two approaches to loop modeling.[17] First, the ab initio loop prediction is based on a conformational search or enumeration of conformations in a given environment, guided by a scoring or energy function. There are many such methods, exploiting different protein representations, energy function terms, and optimization or enumeration algorithms.[17,57,59] The second, database approach to loop prediction begins by finding segments of main chain that fit the two stem regions of a loop.[11,13,19,60] The search for such segments is performed through a database of many known protein structures and not only homologues of the modeled protein. The selected segments are then superposed and annealed on the stem regions and finally ranked according to some rule or a scoring function.

The loop modeling module in MODELLER implements the optimization-based approach.[17] The main reasons are the generality and conceptual simplicity of energy minimization, as well as the limitations imposed on the database approach by a relatively small number of known protein structures.[61] The method was tested on a large number of loops of known structure, both in the native and near-native environments. Loops of 8 residues predicted in the native environment have a 90% chance to be modeled with useful accuracy (i.e., rms error for superposition of the loop main-chain atoms is less than 2 Å). Even 12-residue loops are modeled with useful accuracy in 30% of the cases. It is possible to estimate whether a given loop prediction is correct, based on the structural variability of the independently derived lowest energy loop conformations. The method has been applied successfully in blind predictions of protein structure at meetings on critical assessment of protein structure prediction methods (CASP) 3 and 4 (Figures 3 and 4).

The current loop modeling method in MODELLER includes the solvation effect only indirectly, through the statistical potential of mean force used to restrain the nonbonded contacts between protein atoms. Since the environment of most loops is significantly solvent exposed, we aimed to improve loop modeling by including a more accurate description of the interactions between the protein and the solvent.

FIGURE 4. Loop modeling of the region 46−53 in *S. pombe* contractile ring protein Cdc4p. The target sequence is related remotely at 33% sequence identity to the template structure of *A. irradians* myosin, 1WDC (blue). The model (red) was constructed for CASP3 before the actual structure (1GGW) was defined by X-ray crystallography (green).[65] The 8-residue loop prediction has the global and local main-chain rms errors of 3.64 and 1.36 Å, respectively. The loops are in thick lines and are shown upon superposition of the whole structures.

The solvent could be taken into account explicitly by immersing the modeled loop into a bath of water molecules. However, this model representation would result in too costly an optimization, especially when many loop predictions need to be done, as in genome-scale comparative modeling. The large number of particles in the model would increase both the time required for a single function evaluation and the number of function evaluations needed for a given degree of optimization. This problem is significantly reduced when implicit solvent models based on approximate analytic solutions to continuum solvation theories are used.[45] As a result, we began exploring implicit solvation models instead of explicit solvent representations. In particular, we focused on the generalized Born (GB) approximation implemented in the CHARMM molecular mechanics and modeling package.[37,48,62] This approach has been demonstrated to assist in distinguishing native structures from misfolded decoys.[63]

The new protocol for modeling a given loop sequence was implemented using the multiscale modeling tools for structural biology (MMTSB, http://mmtsb.scripps.edu) and proceeds as follows.[64,65] First, 49 loop conformations were generated with MODELLER,[17] without any solvation terms. Next, the 49 conformations were minimized by

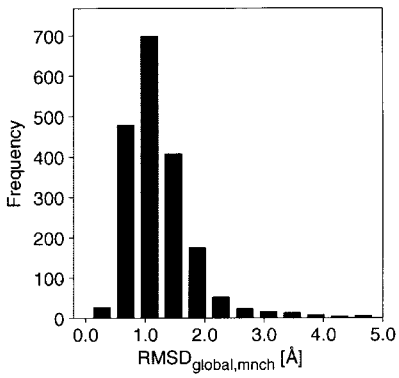

FIGURE 5. Magnitude of conformational change upon refinement of a MODELLER loop model with respect to the CHARMM/GB energy function. The distribution is shown for the 1910 refinements, corresponding to the 49 MODELLER runs for each of the 39 loop sequences.

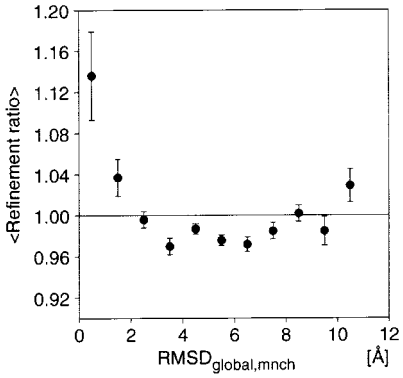

FIGURE 6. Success of refinement as a function of the initial model accuracy. The refinement ratio is defined as the ratio between the main-chain global rms errors of the refined and the initial loops. The distribution is shown for all of the 1910 refinements. The error bars indicate the standard error of the mean.

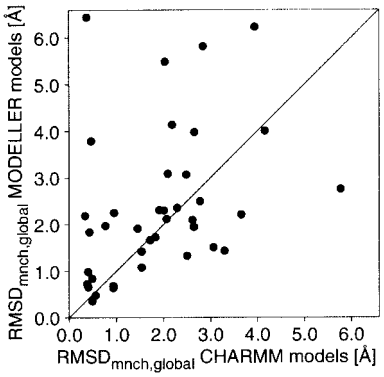

FIGURE 7. Correlation between the accuracy of the best scoring MODELLER and CHARMM/GB predictions. The scatter plot is shown for the 39 8-residue test loops.

CHARMM with respect to the standard PARAM19 force field and the generalized Born approximation to solvation, as well as harmonic restraints with force constants of 0.1 and 1 kcal/mol applied to the protein atoms within 9 Å of the loop atoms and to protein atoms 9−12 Å from the loop atoms, respectively (CHARMM/GB energy). The system was relaxed by 50 steps of the steepest descent minimization, followed by a more aggressive adopted basis Newton−Raphson minimization of up to 2000 steps or until the energy decrease between the steps became less than 10^{-4} kcal/mol. The 49 relaxed conformations were ranked by the CHARMM/GB energy.

At least for 8-residue loops, the new protocol is justified because MODELLER almost always produces at least a few accurate conformations in 49 independent runs. When MODELLER loop prediction fails, it fails because the MODELLER objective function is not able to identify the correct conformation and not because the optimizer is unable to sample it. We hoped that the CHARMM/GB energy function would improve identification of the best conformation among the 49 alternatives generated by MODELLER.

The accuracy of the new protocol was tested by modeling 39 of the 40 different 8-residue loops extracted from high-resolution protein structures.[17] Minimization of the CHARMM/GB energy of a MODELLER loop model usually does not move the loop by more than 2 Å (Figure 5). In 55% of the 1910 individual runs (39 loops times 49 runs/loop), minimization of the CHARMM/GB energy improved the accuracy of the initial MODELLER model. For the initial MODELLER loops that had the global main-chain rms error better than 5 Å and the local rms error better than 2 Å, the fractions of the MODELLER models improved by minimization of the CHARMM/GB energy were 57 and 63%, respectively. A global rms error is obtained by superposing the three stem residues on each side of a loop, whereas the local rms error is obtained by superposing the loops only. Refinement by minimization of the CHARMM/GB energy was particularly effective in improving relatively inaccurate initial models (global main-chain rms error > 2.0 Å) (Figure 6). In contrast, if MODELLER produced a relatively accurate model, minimization of the CHARMM/GB energy tends to slightly increase the error of the optimized model (Figure 6). A small but probably significant improvement is also observed in the ranking of loop conformations on the basis of the CHARMM/GB energy relative to that on the basis of the MODELLER energy. The average main-chain rms error of the best scoring models improved from 2.36 to 1.87 Å for global superposition and from 1.29 to 1.07 Å for local superposition (Figure 7).

Conclusions

The accuracy of comparative modeling is likely to be improved by an explicit consideration of the physical energy terms, especially of the solvation effect. We suggest to express both the comparative modeling rules and the energy terms as spatial restraints, combine them into a single objective function, and calculate a model by optimization of this function. The main role of the homology-derived terms in such an objective function would be to provide a suitable starting conformation and to restrain the conformational search to a manageable and relevant portion of the phase space, while the main role of the physical terms would be to allow the refinement of the model away from the template structure toward the actual structure of the target. These ideas are illustrated

by preliminary results from a loop modeling protocol that relies on the generalized Born approximation to solvation. Given a reasonable starting loop conformation and its environment produced by comparative modeling, a physical energy function with the generalized Born term forces the starting structure closer to the actual structure. In addition, the physical energy function is capable of ranking an ensemble of loop conformations more accurately than the comparative modeling objective function used to derive these conformations. These improvements encourage us to continue exploring different solvation models and optimization protocols for refining and ranking loop conformations as well as whole protein structures.

We are grateful to the members of our group for discussions about protein structure prediction. Research was supported by NIH Grant RR12255 (C.L.B.) and by NIH/GM Grant 54762, a Merck Genome Research Award, and a Mathers Fund Award (A.S.). A.F. was a Burroughs Wellcome Fund Postdoctoral Fellow and is a Charles Revson Foundation Postdoctoral Fellow. A.S. is an Irma T. Hirschl Trust Career Scientist. This perspective is based partly on previous papers.[2,5]

References

(1) Chothia, C.; Lesk, A. M. The Relation between the Divergence of Sequence and Structure in Proteins. *EMBO J.* **1986**, *5*, 823−826.
(2) Baker, D.; Sali, A. Protein Structure Prediction and Structural Genomics. *Science* **2001**, *294*, 93−96.
(3) Bonneau, R.; Baker, D. Ab Initio Protein Structure Prediction: Progress and Prospects. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 173−189.
(4) Blundell, T. L.; Sibanda, B. L.; Sternberg, M. J.; Thornton, J. M. Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules. *Nature* **1987**, *326*, 347−352.
(5) Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.
(6) Sanchez, R.; Sali, A. Comparative Protein Modeling As an Optimization Problem. *J. Mol. Struct. (THEOCHEM)* **1997**, *398*, 489−496.
(7) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI−BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(8) Torda, A. E. Perspectives in Protein-Fold Recognition. *Curr. Opin. Struct. Biol.* **1997**, *7*, 200−205.
(9) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(10) Browne, W. J.; North, A. C. T.; Phillips, D. C.; Brew, K.; Vanaman, T. C.; Hill, R. C. A Possible Three-Dimensional Structure of Bovine Lactalbumin Based Onthat of Hen's Egg-White Lysosyme. *J. Mol. Biol.* **1969**, *42*, 65−86.
(11) Greer, J. Comparative Model-Building of the Mammalian Serine Proteases. *J. Mol. Biol.* **1981**, *153*, 1027−1042.
(12) Levitt, M. Accurate Modeling of Protein Conformation by Automatic Segment Matching. *J. Mol. Biol.* **1992**, *226*, 507−533.
(13) Jones, T. A.; Thirup, S. Using Known Substructures in Protein Model Building and Crystallography. *EMBO J.* **1986**, *5*, 819−822.
(14) Havel, T. F.; Snow, M. E. A New Method for Building Protein Conformations From Sequence Alignments With Homologues of Known Structure. *J. Mol. Biol.* **1991**, *217*, 1−7.
(15) Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.
(16) Kolinski, A; Betancourt, MR.; Kihara, D; Rotkiewicz, P; Skolnick, J Generalized Comparative Modeling (GENECOMP): A Combination of Sequence Comparison, Threading, and Lattice Modeling for Protein Structure Prediction and Refinement. *Proteins* **2001**, *44*, 133−149.
(17) Fiser, A.; Do, R. K.; Sali, A. Modeling of Loops in Protein Structures. *Protein Sci.* **2000**, *9*, 1753−1773.
(18) Rufino, S. D.; Donate, L. E.; Canard, L. H.; Blundell, T. L. Predicting the Conformational Class of Short and Medium Size Loops Connecting Regular Secondary Structures: Application to Comparative Modelling. *J. Mol. Biol.* **1997**, *267*, 352−367.
(19) van Vlijmen, H. W.; Karplus, M. PDB−Based Protein Loop Prediction: Parameters for Selection and Methods for Optimization. *J. Mol. Biol.* **1997**, *267*, 975−1001.
(20) Dunbrack, R. L., Jr.; Karplus, M. Backbone-Dependent Rotamer Library for Proteins. Application to Side-Chain Prediction. *J. Mol. Biol.* **1993**, *230*, 543−574.
(21) Xiang, Z.; Honig, B. Extending the Accuracy Limits of Prediction for Side-Chain Conformations. *J. Mol. Biol.* **2001**, *311*, 421−430.
(22) Sanchez, R.; Sali, A. Large-Scale Protein Structure Modeling of the Saccharomyces Cerevisiae Genome. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 13597−13602.
(23) Bujnicki, J. M.; Elofsson, A.; Fischer, D.; Rychlewski, L. LiveBench-1: Continuous Benchmarking of Protein Structure Prediction Servers. *Protein Sci.* **2001**, *10*, 352−361.
(24) Eyrich, V.; Marti-Renom, M. A.; Przybylski, D.; Fiser, A.; Pazos, F.; Valencia, A.; Sali, A.; Rost, B. EVA: Continuous Automatic Evaluation of Protein Structure Prediction Servers. *Bioinformatics.* **2001**, in press.
(25) Luthy, R.; Bowie, J. U.; Eisenberg, D. Assessment of Protein Models With Three-Dimensional Profiles. *Nature* **1992**, *356*, 83−85.
(26) Sippl, M. J Recognition of Errors in Three-Dimensional Structures of Proteins. *Proteins* **1993**, *17*, 355−362.
(27) Melo, F.; Sanchez, R.; Sali, A. Statistical Potentials for Fold Assessment. *Protein Sci.* **2002**, *11*, 430−448.
(28) Pieper, U.; Eswar, N.; Ilyin, V. A.; Stuart, A.; Sali, A. ModBase, a Database of Annotated Comparative Protein Structure Models. *Nucleic Acids Res.* **2002**, *30*, 255−259.
(29) Sanchez, R.; Pieper, U.; Mirkovic, N.; de Bakker, P. I.; Wittenstein, E.; Sali, A. MODBASE, a Database of Annotated Comparative Protein Structure Models. *Nucleic Acids Res.* **2000**, *28*, 250−253.
(30) Holm, L.; Sander, C. Mapping the Protein Universe. *Science* **1996**, *273*, 595−603.
(31) Vitkup, D.; Melamud, E.; Moult, J.; Sander, C. Completeness in Structural Genomics. *Nat. Struct. Biol.* **2001**, *8*, 559−566.
(32) Sali, A.; Overington, J. P. Derivation of Rules for Comparative Protein Modeling From a Database of Protein Structure Alignments. *Protein Sci.* **1994**, *3*, 1582−1596.
(33) Tanaka, S.; Scheraga, H. A. Model of Protein Folding: Inclusion of Short-, Medium-, and Long-Range Interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 3802−3806.
(34) Finkelstein, A. V.; Badretdinov, A. Y.; Gutin, A. M. Why Do Protein Architectures Have Boltzmann-Like Statistics? *Proteins* **1995**, *23*, 142−150.
(35) Miyazawa, S.; Jernigan, R. L. Residue-Residue Potentials With a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.* **1996**, *256*, 623−644.
(36) Sippl, M. J. Calculation of Conformational Ensembles From Potentials of Mean Force. An Approach to the Knowledge-Based Prediction of Local Structures in Globular Proteins. *J. Mol. Biol.* **1990**, *213*, 859−883.
(37) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Muczera, K.; Lau, F. T. K.; Mattos, C.; Michnik, S.; Nguyen, D. T.; Ngo, T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modleing and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(38) Elber, R.; Karplus, M. Multiple Conformational States of Proteins: a Molecular Dynamics Analysis of Myoglobin. *Science* **1987**, *235*, 318−321.
(39) Nemethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. Energy Parameters in Peptides. Improved Geometrical Parameters and Non-Bonded Intercations for Use in the ECEPP/3 Algorithm, With Application to Proline-Containing Peptides. *J. Phys. Chem. B* **1992**, *96*, 6472−6484.
(40) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. J.; Fergusson, D. M.; Spellmeyer, D. C.; Fox, D. C.; Caldwelld, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins and Nucleic Acids. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(41) Duan, Y.; Kollman, P. A. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science* **1998**, *282*, 740−744.
(42) Koehl, P.; Levitt, M. A. Brighter Future for Protein Structure Prediction. *Nat. Struct. Biol.* **1999**, *6*, 108−111.

(43) Mohanty, D.; Dominy, B. N.; Kolinski, A.; Brooks, C. L., III; Skolnick, J. Correlation Between Knowledge-Based and Detailed Atomic Potentials: Application to the Unfolding of the GCN4 Leucine Zipper. *Proteins* **1999**, *35*, 447−452.
(44) Lazaridis, T.; Karplus, M. Effective Energy Function for Proteins in Solution. *Proteins* **1999**, *35*, 133−152.
(45) Bashford, D.; Case, D. A. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−152.
(46) Giesen, D. J.; Gu, M. Z.; Cramer, C. J.; Truhlar, D. G. A Universal Organic Solvation Model. *J. Org. Chem.* **2000**, *65*, 5886.
(47) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A. Biomolecular Simulations: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein−Ligand, Protein−Protein, and Protein-Nucleic Acid Noncovalent Interactions. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 211−243.
(48) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. a Fast Analytical Method for Th Ecalculation of Approximate Born Radii. *J. Phys. Chem.* **1997**, *101*, 3005−3014.
(49) Feig, M.; Rotkiewicz, P.; Kolinski, A.; Skolnick, J.; Brooks, C. L., III. Accurate Reconstruction of All-Atom Protein Representations From Side- Chain-Based Low-Resolution Models. *Proteins* **2000**, *41*, 86−97.
(50) Schaefer, M.; van Vlijmen, H. W.; Karplus, M. Electrostatic Contributions to Molecular Free Energies in Solution. *Adv. Protein Chem.* **1998**, *51*, 1−57.
(51) Gilson, M. K.; Honig, B. Calculation of the Total Electrostatic Energy of a Macromolecular System: Solvation Energies, Binding Energies, and Conformational Analysis. *Proteins* **1988**, *4*, 7−18.
(52) Friesner, R. A.; Dunietz, B. D. Large-Scale Ab Initio Quantum Chemical Calculations on Biological Systems. *Acc. Chem. Res.* **2001**, *34*, 351−358.
(53) Lazaridis, T.; Karplus, M. Discrimination of the Native From Misfolded Protein Models With an Energy Function Including Implicit Solvation. *J. Mol. Biol.* **1999**, *288*, 477−487.
(54) Gatchell, D. W.; Dennis, S.; Vajda, S. Discrimination of Near-Native Protein Structures From Misfolded Models by Empirical Free Energy Functions. *Proteins* **2000**, *41*, 518−534.
(55) Rapp, C. S.; Friesner, R. A. Prediction of Loop Geometries Using a Generalized Born Model of Solvation Effects. *Proteins* **1999**, *35*, 173−183.
(56) Tanner, J. J.; Nell, L. J.; McCammon, J. A. Anti-Insulin Antibody Structure and Conformation. II. Molecular Dynamics With Explicit Solvent. *Biopolymers* **1992**, *32*, 23−32.
(57) Moult, J.; James, M. N. An Algorithm for Determining the Conformation of Polypeptide Segments in Proteins by Systematic Search. *Proteins* **1986**, *1*, 146−163.
(58) Smith, K. C.; Honig, B. Evaluation of the Conformational Free Energies of Loops in Proteins. *Proteins* **1994**, *18*, 119−132.
(59) Bruccoleri, R. E.; Karplus, M. Conformational Sampling Using High-Temperature Molecular Dynamics. *Biopolymers* **1990**, *29*, 1847−1862.
(60) Martin, A. C.; Thornton, J. M. Structural Families in Loops of Homologous Proteins: Automatic Classification, Modelling and Application to Antibodies. *J. Mol. Biol.* **1996**, *263*, 800−815.
(61) Fidelis, K.; Stern, P. S.; Bacon, D.; Moult, J. Comparison of Systematic Search and Database Methods for Constructing Segments of Protein Structure. *Protein Eng.* **1994**, *7*, 953−960.
(62) Dominy, B. N.; Brooks, C. L., III. Development of a Generalized Born Model Parameterization for Proteins and Nucleic Acids. *J. Phys. Chem.* **1999**, *103*, 3765−3773.
(63) Dominy, B. N.; Brooks, C. L., III. Identifying Native-Like Protein Structures Using Physics-Based Potentials. *J. Comput. Chem.* **2001**, in press.
(64) Feig, M.; Fiser, A.; Sali, A.; Brooks, C. L., III. Implicit Solvation With Comparative Modeling Improves Loop Predictions Using MODELLER. Manuscript in preparation.
(65) Ursby, T.; Adinolfi, B. S.; Al Karadaghi, S.; De Vendittis, E.; Bocchini, V. Iron Superoxide Dismutase From the Archaeon Sulfolobus Solfataricus: Analysis of Structure and Thermostability. *J. Mol. Biol.* **1999**, *286*, 189−205.
(66) Ilyin, V. A.; Pieper, U.; Stuart, A. C.; Martí-Renom, M. A.; Sali, A. ModView, a web application for visualization of multiple protein sequences and structures. Submitted.

AR010061H