Chapter 3
From Genome Sequences to Protein Structures: Comparative Modeling and Fold Assignment

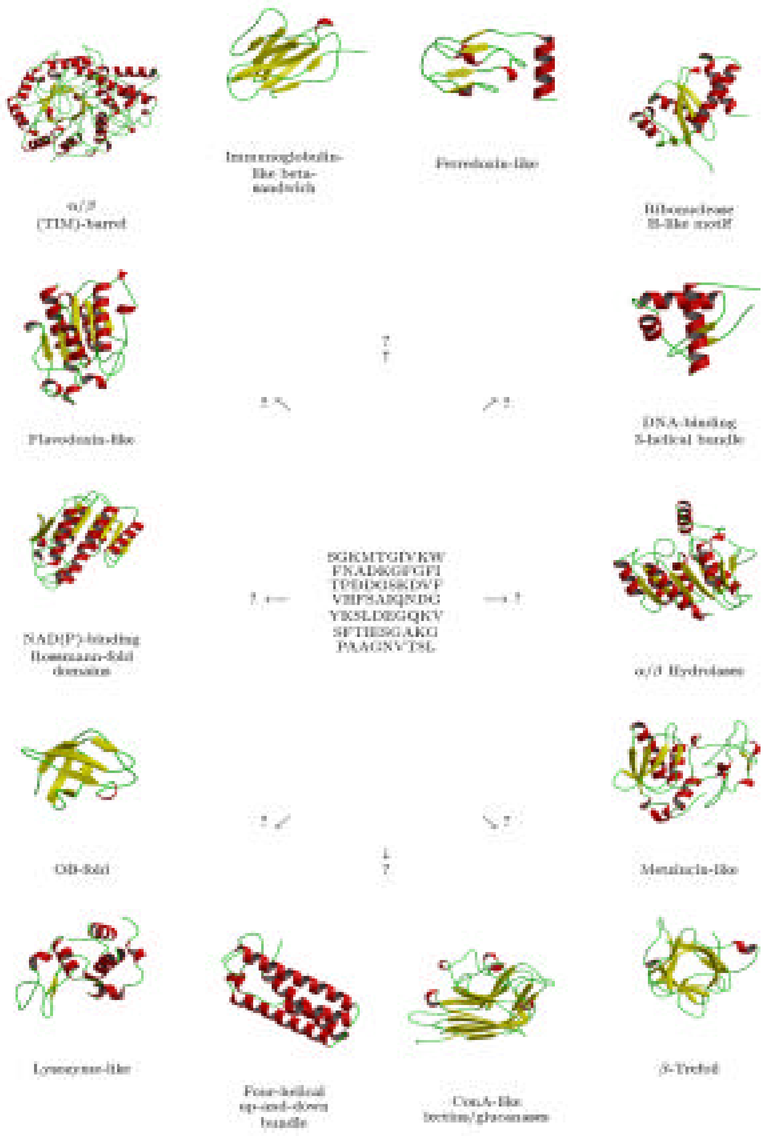
The key to understanding the inner workings of cells is to learn the three-dimensional atomic structures of the some 100,000 proteins that form their architecture and carry out their metabolism. These three-dimensional (3D) structures are encoded in the blueprint of the DNA genome. Within cells, the DNA blueprint is translated into protein structures through exquisitely complex machinery- itself composed of proteins. The experimental process of deciphering the atomic structures of the majority of cellular proteins is expected to take a century at the present rate of work. New developments in comparative modeling and fold recognition will short circuit this process, that is we can learn to translate the DNA message by computer.

The success of these methods rests on a fundamental experimental discovery of structural biology: the 3D structures of proteins have been better conserved during evolution than their genome sequences. When the similarity of a target sequence to another sequence with known structure is above a certain threshold, comparative modeling methods can often provide quantitatively accurate protein structure predict since a small change in the protein sequence usually results in a small change in its 3D structure**.** Even when the percentage identity of a target sequence falls below this level, then at least qualitative information about the overall fold topology can often be predicted.

In protein fold assignment, a genome sequence is computationally tested for compatibility with a library of known protein folds. Current estimates of the number of protein folds range between 800 and 15,000, but each estimate is more than a thousand times smaller than the number of proteins. The goal of fold assignment and comparative modeling is to assign each new genome sequence to the known protein fold or structure that it most closely resembles, using computational methods.

Fold assignment and comparative modeling techniques can then be helpful in proposing and testing hypotheses in molecular biology, such as in inferring biological function, predicting the location and properties of ligand binding sites, in designing drugs, testing remote protein-protein relationships. It can also provide starting models in X-ray crystallography and NMR spectroscopy

The experimental process of deciphering the atomic structures of the majority of cellular proteins is expected to take a century at the present rate of work. But there is now reason to think that new developments in computational protein "fold assignment" to genome sequences, coupled with comparative modeling, will short circuit this


α/β (TIM)-barrel
Immunoglobulin-like beta-sandwich
Ferredoxin-like
Ribonuclease H-like motif
Flavodoxin-like
DNA-binding 3-helical bundle
NAD(P)-binding Rossmann-fold domains
α/β Hydrolases
OB-fold
Metzincin-like
Lysozyme-like
Four-helical up-and-down bundle
ConA-like lectins/glucanases
β-Trefoil
SGKMTGIVKW
FNADKGFGFI
TPDDGSKDVF
VHFSAIQNDG
YKSLDEGQKV
SFTIESGAKG
PAAGNVTSL
?
?
?
?
?
?
?
?

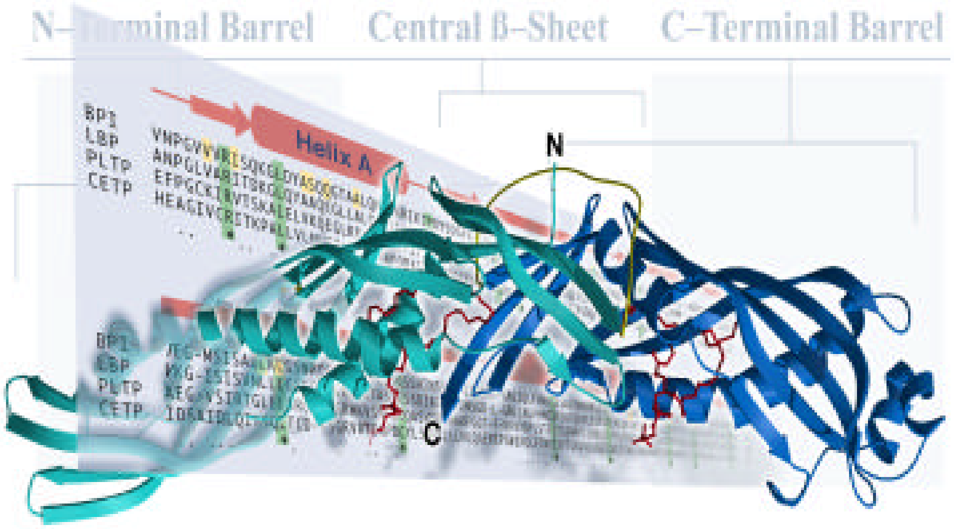
Protein fold assignment. A genome-encoded amino acid sequence (center) is tested for compatibility with a library of known 3D protein folds. An actual library would contain of the order of 1000 folds; the one shown here is a representation, illustrating the most common protein folding motifs. There are two possible outcomes of the compatibility test: that the sequence is most compatible with one of the known folds or that the sequence is not compatible with any known fold. The second outcome may mean either that the sequence belongs to one of the folds not yet discovered, or that the compatibility measures are not fully enough developed to detect the distant relationship of sequence to its structure.

BPI: A case study in assigning genome sequences to a known 3D protein structure.

Bactericidal permeability-increasing protein (BPI; Figure 1) from human white blood cells is a potent antimicrobial protein of 456 amino acid residues. Its structure, determined by X-ray crystallography, was found to be a new fold: an elongated, boomerang-shaped molecule, unlike any previously known structure. Prior to the publication of the 3D structure of BPI, its amino acid sequence was submitted to the 2nd meeting on Critical Assessment of Protein Structure Prediction methods (CASP2). Several methods of fold assignment correctly concluded that the BPI sequence was incompatible with any protein fold then in the database of known protein structures.

With the 3D structure of BPI available, it became possible to search databases of genome sequences to learn which other protein sequences are compatible with the BPI structure. In other words, it was then possible to assign other sequences to the BPI structure (Beamer, Fischer, & Eisenberg, 1998). This search uncovered 13 distant relatives of BPI in a diverse set of eurkaryotes, including rat, chicken, worm, and biomphalaria galbrata. The 13 new proteins share only 13-19 % sequence identity with BPI, below the "twilight zone" of marginal identification by sequence comparison methods.

The significance of this case study is that advanced computational methods can assign numerous genome sequences to the 3D structures by methods of fold assignment, short circuiting the laborious experimental determination of 3D structures. Chapter 3 discusses prospects for improving the sensitivity of fold assignment methods, so that even more distant sequence-structure relationships can be detected by computer.

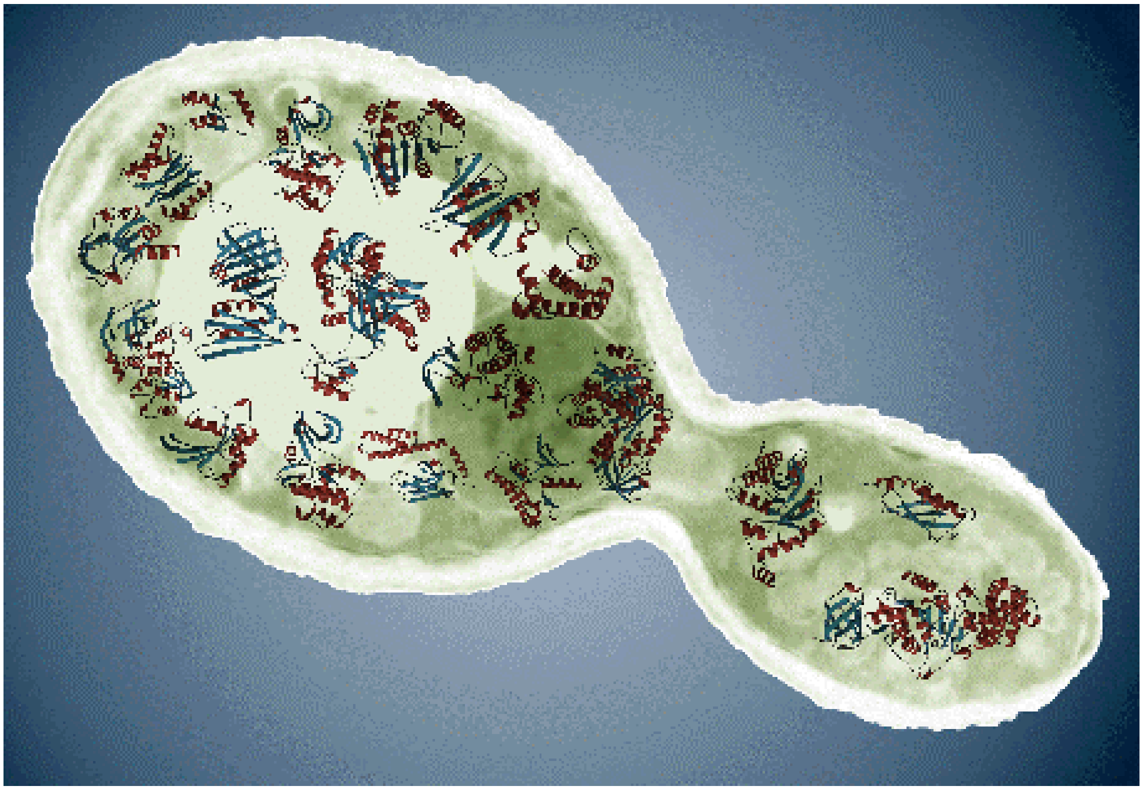


A ribbon diagram of human BPI. The N-terminal domain is aqua, and the C-terminal domain is blue. A proline-rich linker residues 230-250, which connects the two domains, is shown in yellow. The highly conserved disulfide bonds between Cys 135 and Cys 175 is shown as ball-and-stick atoms.

Large-scale comparative modeling of protein structures of the yeast genome

Recently, a large-scale comparative protein structure modeling of the yeast genome was performed (Sanchez and Sali, PNAS, 1998). Fold assignment, comparative protein structure modeling, and model evaluation were completely automated. As an illustration, the method was applied to the proteins in the Saccharomyces cerevisiae (baker's yeast) genome. It resulted in all-atom 3D models for substantial segments of 1071 (17%) of the yeast proteins, only 40 of which have had their 3D structure determined experimentally. Of the 1071 modeled yeast proteins, 236 were related clearly to a protein of known structure for the first time; 41 of these have not been previously characterized at all. Many of the models are sufficiently accurate to facilitate interpretation of the existing functional data as well as to aid in the construction of mutants and chimeric proteins for testing new functional hypotheses. This study shows that comparative modeling efficiently increases the value of sequence information from the genome projects, although it is not yet possible to model all proteins with useful accuracy.

The main bottlenecks are the absence of structurally defined members in many protein families and the difficulties in detection of weak similarities, both for fold recognition and sequence-structure alignment. However, while only 400 out of the total of a few thousand domain folds are known, the structure of most globular folds is likely to be determined in less than ten years. Thus, comparative modeling will conceivably be applicable to most of the globular protein domains close to the completion of the human genome project.



Large-scale protein structure modeling. A small sample of the 1,100 comparative models calculated for the proteins in the yeast genome is displayed over an image of a yeast cell.

Methods for Fold Assignment

Recent work on fold assignment (or "threading") involves two main approaches: developing potentials for fold assignment, and Hidden Markov Models (HHMs) or profile methods that descended from sequence alignment methods. The potentials can be contact potentials (potentials of mean force) or they can be more complex semi-empirical potentials, involving atomic areas and other properties. Fold assignment approaches further subdivide into two categories: (1) unipositional methods that consider probability distributions of amino acids at single sites and (2) those that consider distributions on pairs (or even triples) of amino acids within a contact distance in a given structure. Hidden Markov models to date have considered single site probability distributions. We discuss each approach in turn.

Contact potentials for threading. Unipositional fold assignment approaches score each residue position in a template structure using local 3D environmental information such as secondary structure propensity, degree of environmental polarity, and the fraction of the residue surface buried and inaccessible to solvent. The 3D environmental information for each residue then becomes a one-dimensional profile of the tertiary structure or fold, and the compatibility of the twenty common amino acids are evaluated for each position in the 1-D profile. Optimal 1-D alignments of a probe sequence to a given structure can be determined by dynamic programming, and the subsequent score of the aligned sequence against the template are determined by the global characteristics of the sequence-environment fit, thereby tolerating locally poor scores.

Unipositional methods demonstrated impressive ability for determining similar topological folds for proteins with less than 25% sequence identity in some cases. However, only 25% of genome sequences recognize their 3D protein fold with a sufficient threshold of confidence to be considered a successful fold assignment. One reason for this modest success rate of threading is that the repertoire of folds is thought to be incomplete. This repertoire (or library) is growing modestly through the current efforts of structural biologists, and a strong Structural Genomics Initiative will give a major boost to fold assignment in the future. It has been estimated that the success rate of fold assignment algorithms will increase to roughly 50% once these missing folds are identified structurally. For the remaining 50% of genome sequences to be assigned to folds, there must be advances in the directions discussed here.

The first advance is to move to multi-positional compatibility functions. Pairwise threading potentials typically consider the propensity of two amino acids to be within a specified distance using a score function compiled from a database of structures. Additional features can be used in addition to the identify of the amino acids, such as the secondary structure type, relative exposure to water, relative position, and local atomic density. The identification of the important features is emerging from interdisciplinary collaborations among protein scientists, physicists, and computer scientists, and often involves the use of computational benchmarks, as discussed below. Some attempts have been made to go beyond pairwise potentials, but determination of higher order probability distributions is limited by the data available from the present number of structures. New structures made available from the Structural Genomics Initiative would provide some assistance in this regard, however the number of new structures is unlikely to dramatically increase the order of probability distributions that can be reliably estimated. Therefore further improvements in pairwise and other potentials of mean force will rest on better identification of the relevant physical effects determining the relation of sequence to structure, and on improved algorithms to extract information about these effects from limited data.

Alignment of a genome sequence to 3D structure using pairwise potentials is more difficult than using unipositional potentials. Branch and bound algorithms have been shown to yield the optimal alignment when they converge, but since the general threading problem for multipositional potentials is NP complete, branch and bound algorithms will not converge in all cases. Nevertheless, they are often extremely useful. Approximations can also be employed, such as the frozen approximation, in which one assumes that the interaction of test sequence position j with the amino acid k' of the target structure would be similar to k in the sequence. Once the sequence is optimally aligned using the frozen approximation, the multipositional compatibility function is used to score the sequence-structure match. However, this current state-of-the-art alignment has proved insufficient in the blind prediction experiment CASP2, and improved methods are essential to reach accurate protein models.

Hidden Markov Models. HMM's consider single site probability distributions for amino acids, but have the added feature of a Markovian transition matrix between "hidden" states. The hidden states effectively perform a choice among a set of position dependent amino acid probability distributions. In contrast to threading methods, HMM's do not use an explicit scoring function to score the match of an amino acid with its environment, nor do they typically consider pairwise interactions. HMMs rely heavily on position specific scoring functions that, in combination with the hidden Markov states, match appropriate probability distributions to sequence positions. Prior knowledge about amino acid probability distributions can be incorporated in a Bayesian framework for HMM's using "Dirichlet prior" probability distributions.

HMM's can be used for fold identification by performing a standard sequence based homology search using the probe sequence to generate homologous sequences. These sequences can be used to construct an HMM based on the probe, and then the sequences from a library of folds can be matched against the HMM. Similarly, one can construct separate HMM models for each member of a library of folds, and then score the probe sequence against each model. Construction of HMM models is typically an iterative process involving successive periods of modeling building, searching with the given model, and model refinement. Alignment to a HMM can be performed in an efficient recursive manner, similar to dynamic programming.

A variety of methods has been applied to the problem of scoring the match of a sequence to a structure. These include both analytical methods such as Markov Random Fields, and neural networks, and highly empirical energy-like functions. These range from unipositional functions with three or greater environmental descriptors, to multipositional functions that consider the attributes of two or more amino acids at a time. Multipositional functions are potentially more sophisticated since a score is based on the compatibility of multiple amino acids in the test sequence with multiple positions in the target structure.

Even after perfect fold assignment can be achieved, there remains a computational bottleneck in providing a predicted 3D structure: proper alignment of the sequence to the structure. Alignment with unipositional compatibility functions, which add the independent contributions of a single position of a test sequence to a single position in the target fold, offers the advantage of using well-established dynamic-programming algorithms to find the optimal alignment, although poor gap and insertion penalty parameters can render this optimum somewhat arbitrary. HMM's offer effective position dependent insertion/deletion penal- ties as well as an efficient alignment procedure, but ignore more than single site probabilities, as do other unipositional compatibility functions. Pairwise threading potentials add an additional order of statistics (the pairwise probability distributions between amino acids), but with an increase in computational cost for the alignment step. Allowing only a limited number of gaps between secondary structure elements, and exhaustively enumerating all possible resulting threadings, has been implemented for multipositional compatibility functions and has been successful for a subset of interesting cases.

Methods for Comparative Modeling

Comparative modeling remains the only method at present that can provide models with an rms error lower than 2Å. All current comparative modeling methods consist of four sequential steps. The first step is to identify the proteins with known 3D structures that are related to the target sequence. The second step is to align them with the target sequence and to pick those known structures that will be used as templates. The third step is to build the model for the target sequence given its alignment with the template structures. In the fourth step, the model is evaluated using a variety of criteria. If necessary, the alignment and model building are repeated until a satisfactory model is obtained. The main difference between the different comparative modeling methods is in how the 3D model is calculated from a given alignment (step 3 above).

The original and still widely used method is modeling by rigid body assembly. The method constructs the model from a few core regions, and loops and side-chains, which are obtained from dissected related structures. This assembly involves fitting the rigid bodies on the framework, which is defined as the average of the C_α atoms in the conserved regions of the fold. Another family of methods, modeling by segment matching, relies on approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms. This is achieved by the use of a database of short segments of protein structure, energy or geometry rules, or some combination of these criteria. The third group of methods, modeling by satisfaction of spatial restraints, uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure. In addition to the methods for modeling the whole fold, numerous other techniques for predicting loops and side-chains on a given backbone have also been described. These methods can often be used in combination with each other and with comparative modeling techniques.

Perhaps the most promising comparative model building technique (step 3 above) is the comparative modeling by satisfaction of spatial restraints. The reason is that this approach is based only on optimization of an objective function, and it thus allows an efficient exploration of various representations of protein structure, methods of optimization, and objective function forms. The computational complexity of this approach is directly tied to methods such as global optimization described in the next chapter. This flexibility is essential for improving comparative protein modeling. It will also facilitate simultaneous use of different sources of information when calculating a model of a given protein. For example, a model may be constructed that is consistent with the template structures, potentials of mean force, NMR restraints, cross linking experiments, site-directed mutagenesis data, etc. Boundaries between comparative modeling, fold assignment, ab initio folding simulations, ligand docking, NMR and X-ray structure refinement will be blurred.

The best comparative techniques can generally produce models with good stereochemistry and overall structural accuracy that is slightly higher than the similarity between the template and the actual target structures, when the modeling alignment is correct. The errors in comparative models can be divided into five categories: (1) Side-chain packing errors. (2) Distortions and rigid body changes in regions that are aligned correctly (e.g., loops, helices). (3) Distortions and rigid body changes in insertions (e.g., loops). (4) Distortions in incorrectly aligned regions (loops and longer segments with low sequence identity to the templates). (5) Incorrect fold resulting from an incorrect choice of a template. The consequence of these errors is that the comparative method can result in models with a main-chain rms error as low as 1Å for 90% of the main-chain residues, if a sequence is at least 40% identical to one or more of the templates. In this range of sequence similarity, the alignment is mostly straightforward to construct, there are not many gaps, and structural differences between the proteins are usually limited to loops and side-chains. When sequence identity is between 30% and 40%, the structural differences become larger, and the gaps in the alignment are more frequent and longer. As a result, the main-chain r.m.s. error rises to ~ 1.5 Å for about 80% of residues. The rest of the residues are modeled with large errors because the methods generally cannot model structural distortions and rigid body shifts, and cannot recover from misalignments. Insertions longer than about 8 residues usually cannot be modeled accurately at this time, while shorter loops frequently can be modeled successfully. Model evaluation methods are frequently successful in identifying the inaccurately modeled regions of a protein. To put the errors into perspective, we list the differences among experimentally determined structures of the same protein: the 1.0Å accuracy of main-chain atom positions corresponds to X-ray structures defined at a low-resolution of about 2.5 Å and with an R-factor of about 25%, as well as to medium-resolution NMR structures determined from 10 inter-proton distance restraints per residue.

Future improvements of comparative modeling should aim to (1) model proteins with lower similarities to known structures (e.g. , less than 30% sequence identity), (2) to increase the accuracy of the models, and (3) to make modeling fully automated. The improvements are likely to include simultaneous optimization of side-chain and backbone conformations in side-chain modeling, simultaneous optimization of a loop and its environment in loop modeling, and simultaneous optimization of the alignment and the model. At the same time, better potential functions and possibly better optimizers are needed. The potential function should guide the model away from the templates in the direction towards the correct structure. An addition of atomic or residue based potentials of mean force to the homology-derived scoring could be one way of achieving this goal. This is a difficult problem, as illustrated by the fact that no present force field or potential of mean force can produce a model with a main-chain rmsd from the X-ray structure smaller than about 1Å, even when the starting conformation is the X-ray structure itself. For example, molecular dynamics simulations in solvent generally have a main-chain rmsd of more than 1Å, and the most detailed lattice folding simulations result in models with an rms error larger than 2Å. Since most of the main-chain atoms in two homologs with at least 40% sequence identity usually superpose with an r.m.s.d. of about 1Å, it is currently better to aim to reproduce the template structures as closely as possible rather than to venture away from the templates in the search for a better model.

The major factor that limits the use of comparative modeling in the cases of less than 30% sequence identity is the alignment problem, as discussed in the fold recognition problem (Figure 1A). In principle, the alignment can be derived by any of the sequence or sequence/ structure alignment methods, but in practice even careful manual editing frequently results in significant alignment errors. At 30% sequence identity, the fraction of incorrectly aligned residues is about 20% and this number rises sharply with further decrease in sequence similarity. This limits the usefulness of comparative modeling because no current modeling technique can recover from an incorrect input alignment. It would appear that fold recoginition methods are a natural solution to the alignment problem in comparative modeling. However, while these techniques are successful in identifying related folds, they appear to be somewhat less successful in generating correct alignments, although improvements in alignment for fold recognition is a goal of future work. To reduce the errors in the model stemming from the alignment errors, iterative changes in the alignment during the calculation of the model are needed. Provided the objective function is capable of distinguishing a good model from a bad one, the iterative realignment and re-selection of templates will minimize the effect of errors in the initial alignment and selection of templates. A case in point is provided by the generation of the RUVB model based on a remotely related E. coli δ structure.

The Need for Advanced Computing for Fold Assignment and Comparative Modeling

Several fundamental issues remain to amplify the effectiveness of fold assignment and comparative modeling. Both can be addressed with broader computational resources and better communication among protein scientists and computational scientists. A primary issue in fold assignment is the determination of better multipositional compatibility functions which will extend fold assignment into the "twilight zone" of sequence homology. In both fold assignment and comparative modeling, better alignment algorithms that deal with multipositional compatibility functions are needed. A move toward detailed empirical energy functions and increasingly sophisticated optimization approaches in comparative modeling will occur in the future. As these future directions develop, computational benchmarks will be important. These are sets of distantly related pairs of proteins, having similar folds, but very different amino acid sequences. New methods for fold recognition and comparative modeling are developed without the use of these pairs, and then tested on this set, with the goal of assigning each sequence to its proper fold, and further refining that fold for accurate structure. The value of computational benchmarks is that they permit unbiased development of new functions.

In fold assignment, an all to all comparison of sequence to structure using dynamic programming scales as L^2, and requires on order of 10^{13} FLOPs. The use of more sensitive multipositional functions that will scale as L^3 or L^4 will likely require 10^{15} to 10^{17} FLOPS. With increased complexity of modeling function and optimization approach, comparative modeling techniques used over the entire human genome will scale beyond 10^{17} Flops.

The availability of tera-scale computational power will further extend fold assignment in the following ways: the alignment by dynamic programming is performed in order L^2 time, where L is the length of the sequence/structure. For 100,000's of sequences and 10,000's of structures (each of order 10^2 amino acids long), an all to all comparison of sequence to structure using dynamic programming would require on order of 10^{13} operations. Genuine teraflop computing capability could make comparisons on this scale, and indeed one or two orders of magnitude bigger, routine today. The use of multipositional functions that will scale as L^3 or L^4 depending on the complexity of the compatibility function, and may require 10^{15} to 10^{17} FLOPS, and will likely need an effective search strategy in protein force fields used in ab initio global optimization prediction and protein folding (Chapter 3), and scales as M^2, where M is the number of atoms in the model. A typical calculation for a medium sized protein takes addition. The next generation of 100 teraflop computers addresses both fundamental issues in reliable fold assignment. First, an increase in complexity in alignment algorithms for multipositional functions that scale beyond L^2, and secondly the development of new multipositional compatibility functions whose parameters are derived by training on large databases with multiple iterations, resulting in an increase in sensitivity and specificity of these new models.

Availability of tera-flop computing will greatly benefit comparative protein structure modeling of both single proteins and the whole human genome. Once an alignment is determined, comparative modeling is formally a problem in molecular structure optimization. The objective function is similar in complexity to typical empirical in the order of 10^{12} Flops. More specifically, the increased computing power is likely to improve the accuracy of comparative models by aleviating the alignment problem and the loop modeling problem, at the very least.

It is probable that the impact of the alignment errors will be decreased by performing independent comparative modeling calculations for many different alignments. Perhaps as many as 1000 different alignments will be explored in this way. This would in essence correspond to a conformational search with soft constraints imposed by the alignment procedure. Such a procedure would increase the computational cost to 10^{15} Flops for a single protein, and to 10^{20} Flops for the human genome.

Specialized and time consuming loop modeling procedures can be used after the initial comparative models are obtained by standard techniques. Such specialized procedures typically need about 5000 energy function evaluations to obtain one loop conformation. It is standard to calculate an "ensemble" of 25-400 loop conformations for each loop sequence. Thus, the total number of function evaluations is on the order of 10^6 for prediction of one loop sequence. Applying these procedures to the whole human genome would take on the order of 10^{18} Flops.