Journal of Computational Physics **151,** 388–401 (1999)
Article ID jcph.1999.6200, available online at http://www.idealibrary.com on IDEAL®

Comparative Protein Structure Modeling in Genomics

Roberto Sánchez and Andrej Šali

Laboratories of Molecular Biophysics, The Rockefeller University, 1230 York Avenue, New York, New York 10021
E-mail: sali@rockefeller.edu

Received September 9, 1998; revised December 24, 1998

The function of a protein is generally determined by its three-dimensional (3D) structure. Thus, it would be useful to know the 3D structures of the thousands of protein sequences that are emerging from the many genome projects. This is the aim of *structural genomics*. The aim will be achieved by a focused, large-scale determination of protein structures by X-ray crystallography and nuclear magnetic resonance spectroscopy, combined efficiently with accurate protein structure modeling techniques. In particular, comparative or homology-based protein structure modeling is expected to play a major role in this effort. Comparative modeling calculates a 3D model of a given protein sequence from the previously determined structures of related proteins. It involves fold assignment, sequence–structure alignment, model building, and model evaluation. To enable large-scale modeling, these steps are being assembled into a completely automated pipeline. The methods involved in the pipeline and their performance are reviewed. The errors in the resulting models are described and their uses in biology are discussed.

1. INTRODUCTION

In a few years, the genome projects will have provided us with the amino acid sequences of more than a million proteins—the catalysts, inhibitors, messengers, receptors, transporters, and building blocks of the living organisms. The full potential of the genome projects will only be realized once we assign and understand the function of these new proteins. The biochemical function of a protein is defined by its interactions with other molecules and the biological function is a consequence of these interactions. While protein function is best determined experimentally [1], it can sometimes be predicted by matching the sequence of a protein with proteins of known function [1–3]. One way to improve such sequence-based predictions of function is to rely on the known native three-dimensional (3D)[1] structure of proteins [4]. The 3D structure of a protein generally provides more

[1] Abbreviations used: 3D, three-dimensional; NMR, nuclear magnetic resonance; PDB, Protein DataBank.

0021-9991/99 $30.00

information about its function than its sequence because interactions of a protein with other molecules are determined by amino acid residues that are close in space but are frequently distant in sequence.

To determine or predict 3D structure of all the proteins encoded in the genomes is the aim of *structural genomics* [5]. Unfortunately, experimental methods for protein structure determination are time consuming and not successful for all proteins; consequently, 3D structures have been determined for only a fraction of proteins for which the amino acid sequence is known; while there are approximately 356,000 protein sequences in GENPEPT (December 14, 1998; URL ftp://ncbi.nlm.nih.gov/genbank/genpept.fsa), there are only 8876 known protein structures in the Brookhaven Protein Databank (PDB) (March 17, 1999; URL http://www.pdb.bnl.gov/statistics.html) [6]. However, a useful 3D model can frequently be obtained by comparative or homology protein structure modeling, which can construct all-atom 3D models for those proteins that are related to at least one known protein structure. Even though comparative modeling is applicable only to the members of structurally characterized protein families, it is the most appropriate modeling method for structural genomics. The reason is that it results in the most accurate, detailed, and explicit models of protein structure. This maximizes the usefulness of the models in biological applications such as interpretation of the existing functional data, design of ligands, and construction of mutants and chimeric proteins for testing new functional hypotheses [7]. Studies on model genomes indicate that currently up to 40% of the known protein sequences have at least one segment related to one or more known structures [8, 9, 73, 74]. Thus, the number of sequences that can be modeled relatively accurately by comparative modeling is already an order of magnitude larger than the number of experimentally determined protein structures. This ratio is likely to increase in the future and underscores the need for an efficient combination of experimental and theoretical efforts in structural genomics. An efficient structural genomics project will put every protein sequence within a "modeling distance" of at least one known protein structure while minimizing the total cost of the project. This can be achieved by focusing X-ray crystallography and magnetic resonance spectroscopy on proteins with new folds and on carefully selected representative structures in more divergent or important protein families.

In this review, we emphasize our own work and experience, although we have profited greatly from the contributions of many others, cited in the list of references. We introduce the technique of comparative protein structure modeling (Section 2), discuss it in the context of large-scale modeling of thousands of proteins (Section 3), describe some applications of the many resulting models in biology (Section 4), and conclude with future trends (Section 5).

2. COMPARATIVE PROTEIN STRUCTURE MODELING

Comparative or homology protein modeling uses experimentally determined protein structures (templates) to predict conformation of another protein with a similar amino acid sequence (target). The necessary conditions for calculating a useful model are (i) that the similarity between the target sequence and the template structures be detected and (ii) that the correct alignment between them be constructed. For reviews of comparative modeling see Refs. [7, 10–13]. This approach to protein structure modeling is possible because a small change in the protein sequence usually results in a small change in its three-dimensional structure [14, 15]. Comparative modeling remains the only modeling method that can provide models with an rms error lower than 2 Å.

A traditional classification of protein structure prediction methods includes two other major classes in addition to comparative modeling [16, 17], *ab initio* protein structure prediction and fold assignment. Each one of these classes includes a large variety of different methods. The defining feature of the *ab initio* methods is that they attempt to predict the native structure only from the sequence of the target protein, using an objective function which may depend on the interaction energies and sometimes on the knowledge of other related sequences. Unfortunately, such methods have so far produced models with the correct fold and an rms error of approximately 4 Å for only a few simple and small protein structures [18]. The defining feature of the fold assignment methods is that they assign a fold to the target sequence by aligning the target sequence with the most compatible known protein structure in the set of alternatives [19]. As such, the fold assignment methods are best seen as the first, and in many cases the most important, step in comparative protein structure modeling (see below).

All current comparative modeling methods consist of four sequential steps [11]. 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 select 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.

A major 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 the most widely used method is modeling by rigid body assembly [20–22]. 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 [23–26]. 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 [27, 28] or optimization techniques [29] to satisfy spatial restraints obtained from the alignment of the target sequence with homologous templates of known structure. As this restraint-based modeling can use many different types of information about the target sequence, it is perhaps the most promising of all comparative modeling techniques. In addition to the methods for modeling the whole fold, numerous other techniques for predicting loops [30, 31] and side chains [32, 33] on a given backbone have also been described. These methods can often be used in combination with each other and with comparative modeling techniques.

3. COMPARATIVE PROTEIN STRUCTURE MODELING ON A LARGE SCALE

Large-scale comparative modeling is an automated application of comparative modeling to thousands of protein sequences, not only a few. Since many computer programs for performing each of the operations in comparative modeling already exist, it may seem trivial to construct a pipeline that completely automates the whole process. In fact, it is not easy to do so in a robust manner. For a good reason, most of the tasks in modeling of individual proteins, including template selection, alignment, and model evaluation, are

typically performed with significant human intervention. This allows the use of the best tool for a particular problem at hand and consideration of many different sources of information that are difficult to take into account entirely automatically. Because large-scale modeling can only be performed in a completely automated manner, the main challenge is to build an automated and robust pipeline that approaches the performance of a human expert as much as possible.

Recently, two applications of comparative modeling to complete genomes have been described. For the sequences encoded in the *E. coli* genome, models were built for 10–15% of the proteins using the SWISS-MODEL web server [34, 35]. Another such study was our own modeling of five procaryotic and eucaryotic genomes [36]. The flowchart for the modeling and some technical details are given in Fig. 1. Our calculation resulted in the models for substantial segments of 17.2%, 18.1%, 19.2%, 20.4%, and 15.7% of all proteins in the genomes of *Saccharomyces cerevisiae* (6218 proteins in the genome; Fig. 2), *Escherichia coli* (4290 proteins), *Mycoplasma genitalium* (468 proteins), *Caenorhabditis elegans* (7299 proteins, incomplete), and *Methanococcus jannaschii* (1735 proteins), respectively. An important feature of this study was an evaluation of all the models by a statistical potential function [37] (Fig. 1). This allowed identification of those models that were likely to be based on correct templates and at least approximately correct alignments. As a result, 236 yeast proteins without any prior structural information were assigned to a particular fold family; 40 of these proteins did not have any prior functional annotation. All the alignments and models of the five genomes are available on the Internet at URL http://guitar.rockefeller.edu, as is our program MODELLER used for sequence–structure alignment, model building, and model evaluation. The models are also accessible through the Saccharomyces Genome Database (SGD) (URL http://genome-www.stanford.edu/Saccharomyces/).

We now discuss each of the steps in the pipeline individually, as applied so far by others and us, either in large-scale fold assignment or comparative modeling.

3.1. *Template Search and Selection*

Traditionally, the selection of template structures is done by programs that detect sequence similarity only, including FASTA [38], BLAST [39], and programs based on dynamic programming methods [40, 41]. These methods are generally rapid, automated, and useful for detection of relatively close relationships between proteins. However, in order to maximize the usefulness of the database of known protein structures, it is also necessary to detect remotely related sequence–structure pairs. This is usually done with more sophisticated methods that rely on structural information or multiple sequences from the family of interest. These methods include threading and 3D profile matching [42–44], Hidden Markov Models [45–47], and iterative sequence similarity searches such as PSI-BLAST [48]. Detection of remote relationships can sometimes also be achieved by relaxing the similarity cutoffs in the simple sequence comparison schemes, albeit at the cost of a higher number of false positives; these may then be eliminated by 3D model building and model evaluation [36].

Both simple sequence similarity searches and more sophisticated methods have been used for fold assignment of protein sequences in whole genomes. Sequence similarity searches have been used to assign templates for 10–15% of the proteins in the *E. coli* genome [34]. We have used the program ALIGN [49] for pairwise sequence–sequence alignment with local dynamic programming to find suitable templates for up to 20% of the sequences in the genomes of *S. cerevisiae*, *E. coli*, *C. elegans*, *M. jannaschii*, and

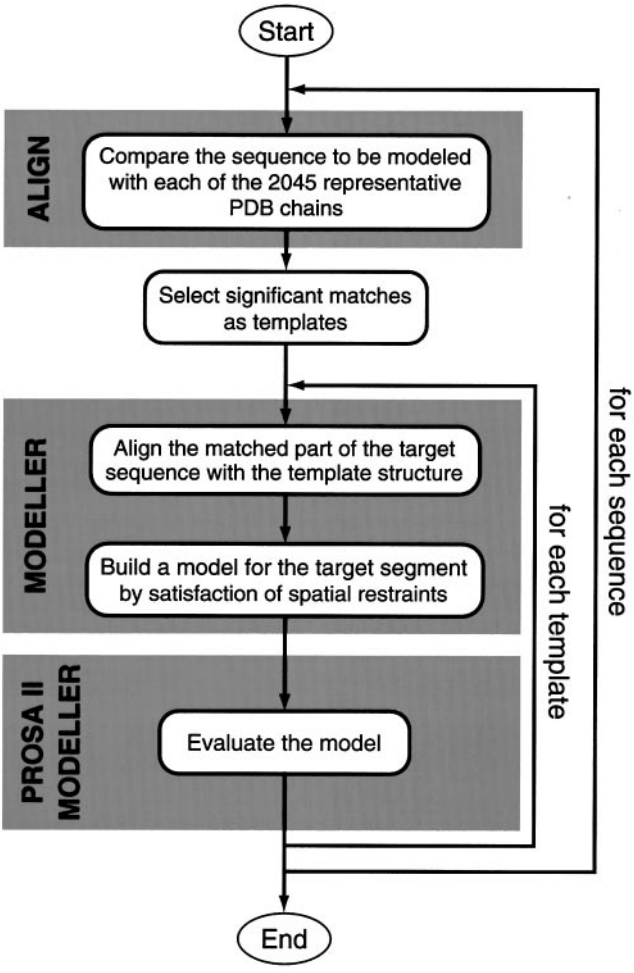

FIG. 1. Flowchart for comparative protein structure modeling on the genome scale [36]. To find template structures for modeling of the protein sequences, each of the sequences is compared with each of the 2045 potential templates corresponding to the protein chains representative of the Brookhaven Protein Databank (PDB) of known protein structures [6]. The representative PDB proteins have at most 95% sequence identity to each other, or have length difference of at least 30 residues or 30%; they are also the highest quality structures within each group. The matching is done by the program ALIGN, which implements the local dynamic programming method with a new gap penalty function and has a search sensitivity higher than that of BLAST [71]. Each sequence–structure matching is run with the default gap penalty parameters first. A match is considered significant or insignificant if the alignment score is more than 22 or less than 19 nats, respectively, where the nat is a unit for measuring significance of a match [49]. All the pairs with intermediate matches with scores between 19 and 22 nats are realigned using 600 combinations of the gap penalty parameters. The match is finally considered significant if the best of the 600 alignments has a score of at least 22 nats. The PDB chain from a significant match is used as the template structure for the corresponding region of the sequence. To obtain the target–template alignment for comparative modeling, the matching parts of the template structure and the protein sequence are realigned by the use of the ALIGN2D command (R.S and A.Š., in preparation) of the MODELLER program [16, 29, 54]. This command implements a global dynamic programming method for comparison of two sequences, but also relies on the observation that evolution tends to place residue insertions and deletions in the regions that are solvent exposed, curved, outside secondary structure segments, and between two C_α positions close in space. Gaps in these structurally reasonable positions are favored by a variable gap penalty function that is calculated from the template structure alone. As a result, the alignment errors are reduced by approximately one third relative to the

(*Continued*)

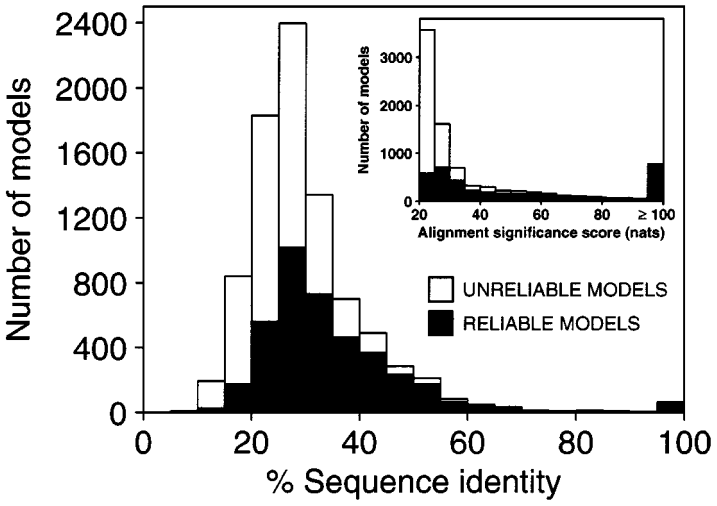

FIG. 2. Distribution of the sequence identity between yeast protein models and corresponding templates [36]. The 3992 reliable models for substantial segments of 1071 different proteins that are predicted to be based on a correct template and approximately correct alignment are represented by the filled bars, and the 4588 unreliable models that are predicted to be based on a mostly incorrect alignment or an incorrect template are represented by the empty bars. The inset shows the corresponding distribution of the alignment significance score calculated by the program ALIGN [71]. The unreliable models with sequence identity to the templates higher than 35% correspond mostly to sequences shorter than 50 residues.

M. genitalium [36] (see above; Fig. 1). Fold recognition by 3D profile matching assigned folds to 22% of the proteins encoded by the *M. genitalium genome* [50]. A new profile–profile sequence alignment method was able to find homologues of known structure for 38% of the *M. genitalium* proteins [8]. Similar results [9] were obtained with PSI-BLAST [48], which also relies on multiple sequence information in finding related proteins. Even though the latter two studies were performed a year after the first three reports, they clearly demonstrate

---

(*Continued*)—standard sequence alignment techniques. The refined sequence–structure alignment is used by MODELLER to construct a 3D model of the matched protein sequence region, containing all main chain and side chain non-hydrogen atoms. Model building begins by extracting distance and dihedral angle restraints on the target sequence from its alignment with the template structure. These template-derived restraints are combined with most of the CHARMM energy terms [72] to obtain a full objective function. Finally, this function is optimized to construct a model that satisfies all the spatial restraints as well as possible. The overall accuracy of the resulting model is predicted by a procedure that relies on a *Z*-score from the program PROSAII [37]. The PROSAII *Z*-score approximates the difference in free energy of an evaluated model and the mean free energy of the same sequence threaded through unrelated folds, expressed in units of standard deviation. The free energies are calculated with statistical potentials of mean force for single residues and pairs of residues [37]. By use of many models of proteins with known structure, the distributions of the PROSAII *Z*-score were obtained for good models, which have more than 30% of their C_α atoms within 3.5 Å of their actual positions, and for bad models. These distributions are used with the Bayesian theorem to calculate the probability that a given model with a certain *Z*-score is either good or bad. Once a model is predicted to be good, its overall accuracy is evaluated more precisely based on an empirical relationship between the fraction of the correctly modeled C_α atoms and the percentage sequence identity to the template [36]. The modeling flowchart in this figure can result in duplicate and overlapping models of some sequence regions. The flowchart has been implemented in a UNIX PERL script that calls the appropriate programs for the individual tasks. Program CLUSTOR is used to distribute efficiently smaller jobs on many workstations, without having to adapt the individual programs for parallel execution (URL http://www.activetools.com).

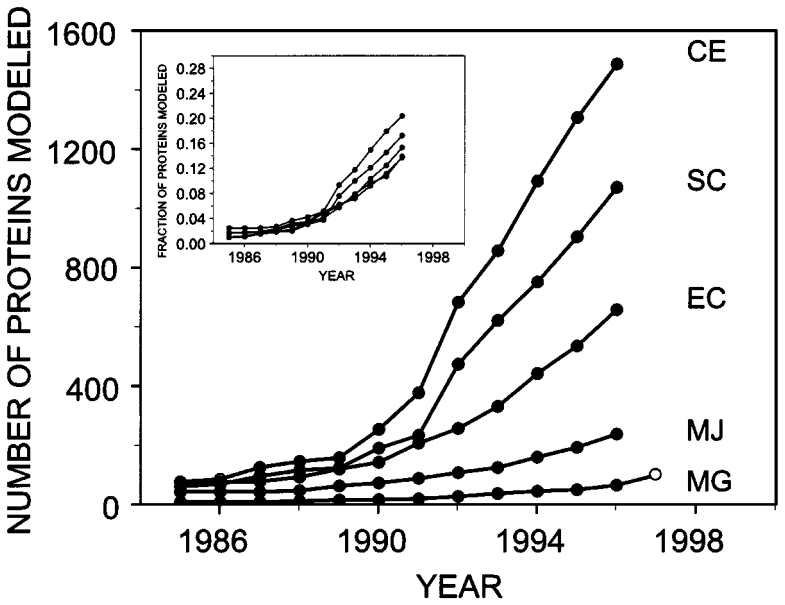

FIG. 3. Simulated effect of the growth of PDB on the number of modeled sequences in various genomes. The data for the plot were obtained from the large-scale modeling study performed with the March 1997 version of PDB [36]. Each point represents the number of protein sequences that would have at least one reliable model using only structures deposited in PDB by the end of the corresponding year. CE, *C. elegans*; SC, *S. cerevisiae*; EC, *E. coli*; MJ, *M. jannaschii*; MG, *M. genitalium*. The empty circle indicates the actual number of the MG proteins modeled at the end of 1997. The inset shows the growth of the fraction of the proteins in each genome that could be modeled as a function of time.

the increased sensitivity of matching a sequence against multiple sequences, compared to the matching against a single sequence or even threading against a single structure; at most a few percentage points of the difference are due to the increase in the number of known structures (Fig. 3).

Although fold recognition is useful for template selection and frequently for functional annotation, it is not the ultimate goal of structural genomics. If a full understanding of the function of a protein is to be achieved, a detailed, full-atom 3D model needs to be obtained and the whole comparative modeling procedure has to be applied. A model based on a remotely related template structure is more likely to be grossly inaccurate because of the errors in the alignment and the structural differences between the template structure and the actual structure of the target sequence [51]. For this reason, the number of proteins with a reliable comparative model will always be smaller than the number of correct fold assignments.

3.2. *Target–Template Alignment*

Since no human intervention is possible in a large-scale effort and since no existing model-building method can recover from an incorrect input alignment, it is particularly important that the automated alignment method be as accurate and robust as possible. Simple pairwise sequence–sequence alignment methods, such as dynamic programming approaches [40, 41], can be used. However, at least two sources of additional information can be incorporated to improve the alignments, similarly to the methods used for template identification. First, when several homologous sequences are known, they can be used to construct a family sequence

profile [8]. Second, structural information from the template structure(s) can also be used to guide the alignment [36]. For example, the gap penalty function in the standard sequence alignment programs can be modified to favor gaps in structurally reasonable contexts (A.Š., R.S., in preparation; 52, 53) (Fig. 1). Three-dimensional profile and threading methods [51] can also be used, although it is not clear whether or not their alignments are more accurate than multiple sequence alignments [8] or alignments from Hidden Markov Models [46].

3.3. *Model Building*

The method of choice for calculating atomic coordinates from a target–template alignment in large-scale modeling must be automated for building the core regions, loops, and side chains in the target sequence. It should permit the use of several templates at the same time, since this significantly increases the accuracy of the final models [54]. One such method is modeling by satisfaction of spatial restraints as implemented in program MODELLER [29, 54]. It was used for large-scale modeling by Adam Godzik (personal communication) and us [36]. Comparative modeling is not CPU time intensive; it typically takes only a few minutes per model. However, application of specialized methods for loop and side chain modeling can be so time consuming that it is not yet possible to apply them on the genome scale. Another important methodological improvement, which will require increased computer power and/or better algorithms, involves automating the cycle of alignment, modeling, and model evaluation for a single protein sequence [54, 55]. This approach can decrease the effect of errors in the input alignment on the final model, but is computationally intensive, requiring from several hours to several days of CPU time for a single target sequence.

3.4. *Model Evaluation*

Model evaluation should serve two roles to facilitate the use of the models in biology: First, it needs to distinguish the models that have at least approximately correct fold (reliable models) from those that do not (unreliable models). Second, it needs to indicate which smaller regions of a reliable model are potentially in error. Unreliable models are obtained when incorrect templates are used; in addition, they also result from mostly incorrect alignments, even when the fold assignment is correct. Incorrect templates occur more frequently when a low similarity cutoff is used in the template selection, which is needed to detect the remote relationships and to minimize the number of missed templates. Comparative models obtained from large-scale modeling have been assigned into the reliable or unreliable class by a procedure [36] (Fig. 1) that relies on the statistical potential function from PROSAII [37]. They have also been evaluated more precisely using a calibrated relationship between the model accuracy and the percentage sequence identity on which the model is based [36].

4. USING COMPARATIVE PROTEIN STRUCTURE MODELS

In general, mistakes in comparative modeling include side chain packing errors, small distortions and rigid body shifts in correctly aligned regions, errors in inserted regions (loops), incorrect alignments, and incorrect templates [54]. The magnitude and prediction of errors in comparative models have been discussed [36, 54]. Fortunately, a 3D model does not have to be absolutely perfect to be helpful in biology, as illustrated by a large number of successful studies that relied on comparative modeling [7]. The type of question that can be addressed with a particular model clearly depends on its accuracy. A convenient and simple predictor of model accuracy is the percentage sequence identity to the template on which

the model was based. Although sequence identity is a useful predictor in many cases, the accuracy of the models based on the same degree of similarity to the templates can vary significantly (Fig. 1 in [36]). One reason is that a large change in structure can be caused by a small change in sequence, binding of a ligand (i.e., induced fit), quaternary interactions, and changes in the environment (e.g., crystal packing, solvent) [56, 57]. This highlights the importance of using the templates whose structures were determined in the environment and with the ligands that pertain to the target model. For example, the calcium binding proteins of the calmodulin type consist of two globular domains, with a pair of EF-hand calcium binding motifs each. The two domains are connected by a flexible helix. The binding of the calcium ions can induce shifts of secondary structure segments within the domains as well as large rigid body movement of the two domains relative to each other [57]. In such cases, even comparative models based on very similar sequences of known structure will have large errors. Fortunately, such cases are relatively rare. Comparative modeling of all the known structures in the Brookhaven Protein Databank indicated that less than 5% of the models based on more than 80% sequence identity have main chain rms errors larger than 2 Å (see the error bars at high sequence identity in Fig. 1B in [36]).

At the low end of the accuracy spectrum, there are models that are based on less than 25% sequence identity and sometimes have less than 50% of the C_α atoms within 3.5 Å of their correct positions. However, such models still have the correct fold and even knowing only the fold of a protein is frequently sufficient to predict its approximate biochemical function. More specifically, only 9 of 80 fold families known in 1994 contained proteins (domains) that were not in the same functional class, although 32% of all protein structures belonged to one of the 9 superfolds [58]. Explicit 3D modeling and model evaluation provide the best way of either confirming or rejecting a remote match [36, 54]. This is important because most of the related protein pairs share less than 30% sequence identity (Fig. 2).

In the middle of the accuracy spectrum are the models based on approximately 35% sequence identity, corresponding to 85% of the C_α atoms modeled within 3.5 Å of their correct positions. Almost half of the 1071 reliably modeled proteins in the yeast genome share more than approximately 35% sequence identity with their templates (Fig. 2). In such cases, it is frequently possible to predict correctly important features of the target protein that do not occur in the template structure. For example, the location of a binding site can be predicted from clusters of charged residues [59], and the size of a ligand can be predicted from the volume of the binding site cleft [60].

Another use of 3D models is that some binding and active sites, which cannot possibly be found by searching for local sequence patterns [61, 62], frequently should be detectable by searching for small 3D motifs that are known to bind or act on specific ligands [63, 64]. This is a consequence of the facts (i) that structure is more conserved than sequence [65]; (ii) that 3D motifs tend to consist of residues distant in sequence; and (iii) that there are some 3D motifs whose residues do not follow the same order in sequence, even though they have the same arrangement in space. An example of this is the serine catalytic triad that almost certainly arose by convergent evolution in serine proteases of the trypsin and subtilisin type, and also in some lipases [63].

In general, medium resolution models frequently allow a refinement of the functional prediction based on sequence alone because ligand binding is most directly determined by the structure of the binding site rather than its sequence. Even when the conserved binding sites are present in the templates, comparative models can still add value to the sequence-based analysis. For example, they can be used to construct site-directed mutants

with altered or destroyed binding capacity, which in turn could test hypotheses about the sequence–structure–function relationships. Other problems that can be addressed with medium resolution comparative models include designing proteins that have compact structures without long tails, loops, and exposed hydrophobic residues for better crystallization; or designing proteins with added disulfide bonds for extra stability.

The high end of the accuracy spectrum corresponds to models based on 50% sequence identity or more. The average accuracy of these models approaches that of low resolution X-ray structures (3 Å resolution) or medium resolution nuclear magnetic resonance (NMR) structures (10 distance restraints per residue) [54]. The alignments on which these models are based generally contain almost no errors. In addition to the already listed applications, high quality models can be used for docking of small ligands into a protein [66] or for docking of a protein to a protein [67, 68].

Large-scale comparative modeling opens new opportunities for tackling existing problems by virtue of providing many protein models from many genomes. One example is the selection of a target protein for which a drug needs to be developed. A good choice is a protein that is likely to have high ligand specificity; specificity is important because specific drugs are less likely to be toxic. Large-scale modeling facilitates imposing the specificity filter in target selection by enabling a structural comparison of the ligand binding sites of many proteins, either human or from other organisms. Such comparisons may make it possible to select rationally a target whose binding site is structurally most different from the binding sites of all the other proteins that may potentially react with the same drug. For example, when a human pathogenic organism needs to be inhibited, it may be possible to select as the target that pathogen's protein that is structurally most different from all the human homologues. Alternatively, when a human metabolic pathway needs to be regulated, the target identification could focus on that particular protein in the pathway that has the binding site most dissimilar from its human homologues.

5. FUTURE DIRECTIONS

It seems likely that in the immediate future the largest improvements in the accuracy and number of comparative models will come from more sensitive template identification, more accurate alignments, more accurate loop modeling, and the growth of the structure and sequence databases. For example, large-scale comparative modeling based on multiple sequence information in template identification, alignment, and model building has not been implemented yet, although it is clear that this will increase both the number and accuracy of the resulting models. A case in point is that the use of multiple sequences increases the rate of fold assignment for almost a factor of two to approximately 38% [8, 9].

Another important factor that determines the degree of structural coverage of a genome is the size of the database of known protein structures. We simulated the impact of the database growth on the number of reliable models (Fig. 3). The fraction of a genome for which relatively accurate models can be calculated with the current modeling procedure (Fig. 1) has grown approximately 3% yearly over the last 2 years; this corresponds to a yearly increase in the number of modeled proteins by approximately 20%. The database of known protein structures grows increasingly faster. At the moment, the doubling rate is approximately 18 months. This progressive growth will undoubtedly continue because of the improvements in the techniques for protein cloning, expression, purification, crystallization, and structure determination by X-ray crystallography and NMR spectroscopy.

Modeling of some proteins is an alternative to direct experimental determination by X-ray crystallography or NMR spectroscopy, even though the models are less accurate than experimentally determined structures. The factors favoring modeling are that it is applicable to all proteins in a family containing at least one known structure, that it is relatively fast (hours instead of months), and that it is inexpensive. Given current modeling techniques, it seems reasonable to require models based on at least 30% sequence identity, corresponding to one experimentally determined structure per *sequence family* rather than fold family. Since there are between 1000 and 5000 fold families and perhaps about five times as many sequence families [69], the experimental effort in structural genomics has to deliver on the order of 10,000 protein domain structures. As an alternative, it has also been suggested that 100,000 protein structures need to be determined by experiment [5]; this would allow calculation of models with higher accuracy than is possible with 10,000 known structures. These are large numbers, but they could be reduced significantly by a relatively small improvement in the comparative modeling techniques. The reasons are (i) that the errors in models increase rapidly as the target–template sequence identity drops below 30% and (ii) that most related protein pairs share less than 30–35% sequence identity (Fig. 2). For example, if the current average model accuracy corresponding to 30% sequence identity is accepted as sufficient, a new comparative modeling method that is capable of delivering equally accurate models based on only 25% sequence identity would decrease the number of needed experimental structures by about 25%. On the scale of the "minimalist" structural genomics project, this corresponds to approximately 2500 structures and justifies a significant investment in the development of new comparative modeling methods and in multi-processor computers for using these methods.

6. CONCLUSIONS

The fraction of protein sequences that can be modeled with useful accuracy by comparative modeling is increasing rapidly. The main reasons for this improvement are the increases in the numbers of known folds and the structures per fold family [69] as well as the improvement in the fold recognition and comparative modeling techniques [16]. It has been estimated that globular protein domains cluster in only a few thousand fold families, approximately 800 of which have already been structurally defined [75]. Assuming the current growth rate in the number of known protein structures, the structure of at least one member of most globular folds will be determined in less than 10 years [69]. According to this argument, comparative modeling would be applicable to most of the globular protein domains before the expected completion of the human genome project. However, there are some classes of proteins, including membrane proteins, that will not be amenable to modeling without improvements in structure determination and modeling techniques. For example, it has been predicted that 839 (13.9%) of the yeast proteins have at least two transmembrane helices [70]. To maximize the number of proteins that can be modeled reliably, a concerted effort toward structural determination of the new folds by X-ray crystallography and NMR spectroscopy is in order (http://genome5.bio.bnl.gov/Proteome/) [5]. A combination of a more complete database of known protein structures with accurate modeling techniques will efficiently increase the value of sequence information from the genome projects.

ACKNOWLEDGMENTS

We are grateful to Drs. Azat Badretdinov, Stephen K. Burley, David Cowburn, John Kuriyan, Richard L. Stevens, Otto Ritter, William Studier, and Joel Sussman for discussions about this project, to Dr. Stephen F. Altschul for

the ALIGN program, to Dr. Manfred Sippl for the PROSAII program, and to Dr. Rok Sosič of ActiveTools for the CLUSTOR program. R.S. is a Howard Hughes Medical Institute predoctoral fellow. A.Š. is a Sinsheimer Scholar and an Alfred P. Sloan Research Fellow. The investigation has also been aided by grants from NIH (GM 54762) and NSF (BIR-9601845). The computations were done on Silicon Graphics, SUN, DEC, and Linux PC computers. This review is based on [13, 36, 54].

REFERENCES

1. S. G. Oliver, From DNA sequence to biological function, *Nature* **379**, 597 (1996).
2. E. V. Koonin and A. R. Mushegian, Complete genome sequences of cellular life forms: Glimpses of theoretical evolutionary genomics, *Curr. Opin. Gen. Dev.* **6**, 757 (1996).
3. B. Dujon, The yeast genome project: What did we learn? *Trends Genet.* **12**, 263 (1996).
4. P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan, Predicting function: From genes to genomes and back, *J. Mol. Biol.* **283**, 707 (1998).
5. A. Šali, 100,000 protein structures for the biologist, *Nature Struct. Biol.* **5**, 1029 (1998).
6. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F Koetzle, and J. Weng, Protein data bank, in *Crystallographic Databases—Information, Content, Software Systems, Scientific Applications*, edited by F. H. Allen, G. Bergerhoff, and R. Sievers (Data Commission of the International Union of Crystallography, Bonn/ Cambridge/Chester, 1987), pp. 107–132.
7. M. S. Johnson, N. Srinivasan, R. Sowdhamini, and T. L. Blundell, Knowledge-based protein modelling, *CRC Crit. Rev. Biochem. Mol. Biol.* **29**, 1 (1994).
8. L. Rychlewski, B. Zhang, and A. Godzik, Fold and function predictions for mycoplasma genitalium proteins, *Fold. Des.* **3**, 229 (1998).
9. M. Huynen, T. Doerks, F. Eisenhaber, C. Orengo, S. Sunyaev, Y. Yuan, and P. Bork, Homology-based fold predictions for mycoplasma genitalium proteins, *J. Mol. Biol.* **280**, 323 (1998).
10. J. Bajorath, R. Stenkamp, and A. Aruffo, Knowledge-based model building of proteins: Concepts and examples, *Protein Sci.* **2**, 1798 (1994).
11. A. Šali, Modelling mutations and homologous proteins, *Curr. Opin. Biotech.* **6**, 437 (1995).
12. B. Rost and C. Sander, Bridging the protein sequence–structure gap by structure predictions, *Annu. Rev. Biophys. Biomol. Struct.* **25**, 113 (1996).
13. R. Sánchez and A. Šali, Advances in comparative protein-structure modeling, *Curr. Opin. Struct. Biol.* **7**, 206 (1997).
14. A. M. Lesk and C. H. Chothia, The response of protein structures to amino-acid sequence changes, *Philos. Trans. R. Soc. London Ser. B* **317**, 345 (1986).
15. T. J. P. Hubbard and T. L. Blundell, Comparison of solvent inaccessible cores of homologous proteins: Definitions useful for protein modelling, *Protein Eng.* **1**, 159 (1987).
16. R. L. Dunbrack Jr., D. L. Gerloff, M. Bower, X. Chen, O. Lichtarge, and F. E. Cohen, Meeting review: The second meeting on the critical assessment of techniques for protein structure prediction (CASP2), Asilomar California, December 13–16, 1996, *Folding & Design* **2**, R27 (1997).
17. J. Moult, T. Hubbard, S. H. Bryant, K. Fidelis, and J. T. Pedersen, Critical assessment of methods of protein structure prediction (CASP): Round II, *Proteins* (Suppl.) **1**, 2 (1997).
18. A. Lesk, CASP2: Report on ab initio predictions, *Proteins* (Suppl.) **1**, 151 (1997).
19. D. T. Jones, Progress in protein structure prediction, *Curr. Opin. Struct. Biol.* **7**, 377 (1997).
20. W. J. Browne, A. C. T. North, D. C. Phillips, K. Brew, T. C. Vanaman, and R. C. Hill, A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme, *J. Mol. Biol.* **42**, 65 (1969).
21. J. Greer, Comparative model-building of the mammalian serine proteases, *J. Mol. Biol.* **153**, 1027 (1981).
22. T. L. Blundell, B. L. Sibanda, M. J. E. Sternberg, and J. M. Thornton, Knowledge-based prediction of protein structures and the design of novel molecules, *Nature* **326**, 347 (1987).
23. T. H. Jones and S. Thirup, Using known substructures in protein model building and crystallography, *EMBO J.* **5**, 819 (1986).

24. R. Unger, D. Harel, S. Wherland, and J. L. Sussman, A 3-D building blocks approach to analyzing and predicting structure of proteins, *Proteins* **5**, 355 (1989).
25. M. Claessens, E. V. Cutsem, I. Lasters, and S. Wodak, Modelling the polypeptide backbone with "spare parts" from known protein structures, *Protein Eng.* **4**, 335 (1989).
26. M. Levitt, Accurate modeling of protein conformation by automatic segment matching, *J. Mol. Biol.* **226**, 507 (1992).
27. T. F. Havel and M. E. Snow, A new method for building protein conformations from sequence alignments with homologues of known structure, *J. Mol. Biol.* **217**, 1 (1991).
28. S. Srinivasan, C. J. March, and S. Sudarsanam, An automated method for modeling proteins on known templates using distance geometry, *Protein Sci.* **2**, 227 (1993).
29. A. Šali and T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.* **234**, 779 (1993).
30. T. V. Borchert, R. A. Abagyan, K. V. R. Kishan, J. P. Zeelen, and R. K. Wierenga, The crystal structure of an engineered monomeric triosephosphate isomerase, monotim: The correct modelling of an eight residue loop, *Structure* **1**, 205 (1993).
31. H. W. T. van Vlijmen and M. Karplus, PDB-based protein loop prediction: Parameters for selection and methods for optimization, *J. Mol. Biol.* **267**, 975 (1997).
32. F. Eisenmenger, P. Argos, and R. Abagyan, A method to configure protein side-chains from the main-chain trace in homology modelling, *J. Mol. Biol.* **231**, 849 (1993).
33. M. Vásquez, Modeling side-chain conformation, *Curr. Opin. Str. Biol.* **6**, 217 (1996).
34. M. C. Peitsch, M. R. Wilkins, L. Tonella, J. C. Sánchez, R. D. Appel, and D. F. Hochstrasser, Large-scale protein modelling and integration with the SWISS-PROT and SWISS-2DPAGE databases: The example of *Escherichia coli*, *Electrophoresis* **18**, 498 (1997).
35. M. C. Peitsch, Promod and swiss-model—Internet-based tools for automated comparative protein modeling, *Biochem. Soc. Trans.* **24**, 274 (1996).
36. R. Sánchez and A. Šali, Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 13597 (1998).
37. M. J. Sippl, Recognition of errors in three-dimensional structures of proteins, *Proteins* **17**, 355 (1993).
38. W. R. Pearson, Rapid and sensitive comparison with FASTA and FASTP, *Methods Enzymol.* **183**, 63 (1990).
39. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* **215**, 403 (1990).
40. S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.* **48**, 443 (1970).
41. T. F. Smith and M. S. Waterman, Identification of common molecular subsequences, *J. Mol. Biol.* **147**, 195 (1981).
42. J. U. Bowie, R. Lüthy, and D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, *Science*, **253**, 164 (1991).
43. D. T. Jones, W. R. Taylor, and J. M. Thornton, A new approach to protein fold recognition, *Nature* **358**, 86 (1992).
44. A. Godzik, A. Kolinski, and J. Skolnick, Topology fingerprint approach to the inverse protein folding problem, *J. Mol. Biol.* **227**, 227 (1992).
45. A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, Hidden Markov models in computational biology: Applications to protein modeling, *J. Mol. Biol.* **235**, 1501 (1994).
46. E. L. L. Sonnhammer, S. R. Eddy, and R. Durbin, Pfam, A comprehensive database of protein domain families based on seed alignments, *Proteins* **28**, 405 (1997).
47. S. R. Eddy, Hidden Markov models, *Curr. Opin. Struct. Biol.* **6**, 361 (1996).
48. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang Zhang, W. Miller, and D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucl. Acids Res.* **25**, 3389 (1997).
49. S. F. Altschul and W. Gish, Local alignment statistics, *Methods Enzymol.* **266**, 460 (1996).
50. D. Fischer and D. Eisenberg, Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 11929 (1997).

51. M. Levitt, Competitive assessment of protein fold recognition and alignment accuracy, *Proteins* (Suppl.) **1**, 92 (1997).
52. K. K. Koretke, Z. Luthey-Schulten, and P. G. Wolynes, Self-consistently optimized statistical mechanical energy functions for sequence structure alignment, *Protein Sci.* **5**, 1043 (1996).
53. A. M. Lesk, M. Levitt, and C. Chothia, Alignment of the amino acid sequences of distantly related proteins using variable gap penalties, *Protein Eng.* **1**, 77 (1986).
54. R. Sánchez and A. Šali, Evaluation of comparative protein structure modeling by MODELLER-3, *Proteins* (Suppl.) **1**, 50 (1997).
55. B. Guenther, R. Onrust, A. Šali, M. O'Donnell, and J. Kuriyan, Crystal structure of the δ' subunit of the clamp-loader complex of *E. coli* DNA polymerase, III, *Cell* **91**, 335 (1997).
56. H. R. Faber and B. W. Matthews, A mutant T4 lysozyme displays five different crystal conformations, *Nature* **348**, 263 (1990).
57. K. Pawlowski, A. Bierzyński, and A. Godzik, Structural diversity in a family of homologous proteins, *J. Mol. Biol.* **258**, 349 (1996).
58. C. A. Orengo, D. T. Jones, and J. M. Thornton, Protein superfamilies and domain superfolds, *Nature* **372**, 631 (1994).
59. R. Matsumoto, A. Šali, N. Ghildyal, M. Karplus, and R. L. Stevens, Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells. A cluster of histidines in mouse mast cell protease-7 regulates its binding to heparin serglycin proteoglycan, *J. Biol. Chem.* **270**, 19524 (1995).
60. L. Z. Xu, R. Sánchez, A. Šali, and N. Heintz, Ligand specificity of brain lipid binding protein, *J. Biol. Chem.* **271**, 24711 (1996).
61. A. Bairoch, PROSITE: A dictionary of sites and patterns in proteins, *Nucl. Acids Res.* **20**, 2013 (1992).
62. T. Pawson, Protein modules and signalling networks, *Nature* **373**, 573 (1995).
63. A. C. Wallace, N. Borkakoti, and J. M. Thornton, TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites, *Protein Sci.* **6**, 2308 (1997).
64. J. S. Fetrow and J. Skolnick, Method for prediction of protein function from sequence using the sequence-to-structure-to function paradigm with application to glutaredoxins/thioredoxins and T_1 ribonucleases, *J. Mol. Biol.* **281**, 949 (1998).
65. C. Chothia and A. M. Lesk, The relation between the divergence of sequence and structure in proteins, *EMBO J.* **5**, 823 (1986).
66. C. S. Ring, E. Sun, J. H. McKerrow, G. K. Lee, P. J. Rosenthal, I. D. Kuntz, and F. E. Cohen, Structure-based inhibitor design by using protein models for the development of antiparasitic agents, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 3583 (1993).
67. M. Totrov and R. Abagyan, Detailed *ab initio* prediction of lysozyme–antibody complex with 1.6 Å accuracy, *Nature Struct. Biol.* **1**, 259 (1994).
68. I. A. Vakser, Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin–antibody complex, *Proteins* (Suppl.) **1**, 226 (1997).
69. L. Holm and C. Sander, Mapping the protein universe, *Science* **273**, 595 (1996).
70. H. W. Mewes, K. Albermann, M. Bähr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S. G. Oliver, F. Pfeiffer, and A. Zollner, Overview of the yeast genome, *Nature* **387** (6632 Suppl), 7–65 (1997).
71. S. F. Altschul, Generalized affine gap costs for protein sequence alignment, *Proteins* **32**, 88 (1998).
72. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, CHARMM: A program for macromolecular energy minimization and dynamics calculations, *J. Comp. Chem.* **4**, 187 (1983).
73. S. A. Teichmann, J. Park, and C. Chothia, Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements, *Proc. Natl. Acad. Sci. U.S.A.* **22**, 14658 (1998).
74. R. Grandori, Systematic fold recognition analysis of the sequences encoded by the genome of *Mycoplasma pneumoniae*, *Protein Eng.* **11**, 1129 (1998).
75. T. J. P. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, and C. Chothia, SCOP: A structural classification of proteins database, *Nuc. Acids Res.* **27**, 254 (1999).