Research Signpost, 37/661 (2) Fort P.O., Trivandrum - 695 023, Kerala, India.

Recent Research Developments in Protein Folding, Stability and Design. Editors: Dr. M. Michael Gromiha and Dr. S. Selvaraj., Research Signpost, Trivandrum, 2001

6 Lessons from blind protein structure prediction experiments

Ram Samudrala

Computational Genomics Group Department of Microbiology University of Washington School of Medicine Seattle, WA 98195

ABSTRACT:

The Critical Assessment of Protein Structure prediction methods (CASP) experiments have shown that structure prediction methods are slowly maturing and producing results that are useful in posing and answering biological questions about protein function. We have taken part in all four CASP meetings in both the comparative and ab initio modelling categories. In this paper, we describe the evolution of our methods from CASP1 to CASP4, present results from the most recent experiment, and explore future directions. We discuss the utility of the models produced by our methods in the context of ongoing structural and functional genomics efforts.

Keywords: protein structure prediction, CASP, comparative modelling, ab initio prediction, genomics

Correspondence/Reprint request: Ram Samudrala , P: (206) 732-6122 F: (206) 732-6055 E: ram@compbio.washington.edu
W: http://compbio.washington.edu

INTRODUCTION

The state of blind protein structure prediction

There are two primary categories of methods for predicting protein structure from sequence: comparative and ab initio modelling. In the comparative modelling category, the methodologies rely on the presence of one or more evolutionarily related template protein structures that are used to construct a model; the evolutionary relationship can be deduced from sequence similarity [1, 2, 3, 4] or by "threading" a sequence against a library of structures and selecting the best match [5, 6, 7]. In the ab initio category, there is no strong dependence on database information and prediction methods are based on general principles that govern protein structure and energetics [8, 9, 10, 11, 12]. The categories vary in difficulty, and consequently methods in each of these categories produce models with different levels of accuracy rel-ative to the experimental structure.

Protein structure prediction methods are rigourously evaluated by the Critical Assessment of Structure Prediction (CASP) experiments held every two years [13]. These experiments evaluate prediction techniques by asking modellers to construct models for a number of protein sequences before the experimental result is known, over a period of 3-4 months. We have taken part in all four CASP experiments, including the most recent one (CASP4) that finished in December 2000 [14]. The CASP4 results provide a benchmark as to what level of model accuracy we can currently expect from our approaches.

At CASP4, we made predictions for all of the 40 targets for which an experimental answer was determined [15]. The CASP4 results show that within each of the general structure prediction categories, some methods, including ours, are able to produce models with a fair amount of accuracy. Further improvements are necessary to overcome the limits of current approaches.

The CASP experiments also show that there is not one single algorithm that can "solve" the protein structure prediction problem. The most successful methods are those that combine and build upon the techniques developed by several researchers in the last thirty years (special issues of Proteins: Structure, Function, Genetics, 1995, 1997, 1999, and 2001). Generally the methods have incorporated different sampling techniques and a variety of scoring functions each of which aids prediction of structure only to a limited degree when used individually, but produce models useful for further biological study when combined together in a coherent manner.

Our own approaches combine monte carlo, simulated annealing, genetic algorithms, graph theory, and semiexhaustive searches with move sets consisting of fragments and discrete state models, and scoring functions consisting of all atombased pairwise functions, hydrophobicity indices, secondary structure preferences, and hydrogen bonding. Our goal is to continue to develop components that form a structure prediction engine, combining and innovating upon previously developed approaches by observing what methods work well at previous CASP experiments, and adding new components of our own.

METHODS OVERVIEW

The methods developed by the author for modelling proteins (during the course of the last nine years) is embodied in a suite of publicly available software programs [16] used by many other researchers around the world and by our group. In this work, we describe the methods that have worked well for us at CASP blind prediction experiments. The techniques used are divided based on the two major structure prediction categories, but methods developed for application in one category are useful in the other.

Comparative modelling and fold recognition

Alignment and template selection

A protein sequence that is evolutionarily related to sequences with known structure (determined by X-ray crystallography or NMR techniques) is modelled using comparative modelling techniques developed by us, which are among the most competitive at the CASP experiments. We use a combination of methodologies that are grouped together as shown in Figure 1. If the sequence relationship between the template structures and the target protein is unambiguous (usually when the sequence identity is > 40%), or if there is only one protein with known structure in the family, then the template structure serves as the sole initial model. If there are many possible template structures, models are constructed using all available templates.

Generation of multiple alignments

PSIBLAST alignments and other publicly available servers such as GenTHREADER [17] and SAM [18] are used to generate a variety of choices for alignments. These alternate alignments are also used to construct initial models. Thus, for a given protein in a family with at least one known representative structure, there could be many alignment choices for constructing the initial models.

Constructing initial models

Following the sequence alignment, for each template structure, an initial model is generated by copying atomic coordinates for the main chain (excluding any insertions/loops) and for the side chains of residues that are identical in the target and template proteins. Residues that differ in side chain type are constructed using a minimum perturbation (MP) technique [19]. The MP method changes a given amino acid to the target amino acid preserving the values of equivalent torsion angles between the two side chains, where available. The other angles are constructed using an internally developed library based on residue type [20]. In its current form, the MP method uses a simple library based on the most frequently observed values to determine torsion angles not present in the template structure.

Using initial models to refine alignments

An all-against-all structure comparison between all the initial models is used to produce a sequence alignment based on structural similarity for a given family. This alignment is used in

conjunction with sequence information to create a new multiple sequence alignment. which is compared to the initial set of alignments to check for consistency and make further refinements. This process is repeated until there is convergence.

Constructing variable side chains and main chains

Multiple side chain conformations for residue positions that differ in type between the template and target proteins are generated by exploring all the possibilities in a rotamer library [21]. The most probable conformations based on the interactions of a given conformation with the local main chain are selected using an all-atom distance dependent conditional probability discriminatory function [22].

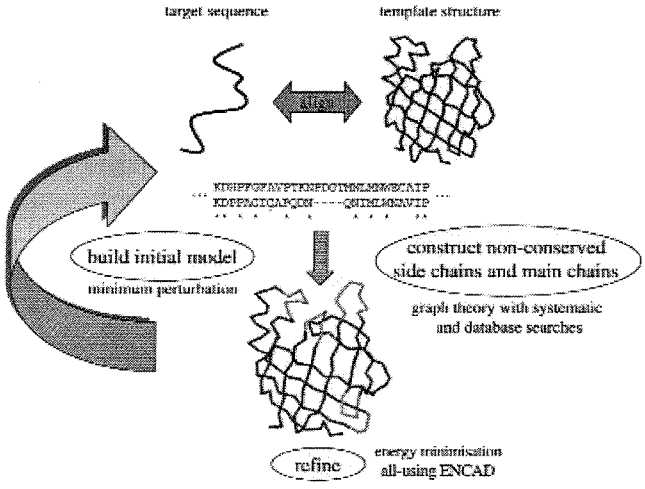

Figure 1: Methodology for comparative modelling. After clustering all sequences, families with members that have conformations determined by experiment are candidates for comparative modelling. Generally, alignments are constructed using one or more publicly available methods. Initial models are constructed and structure-based alignments are used in an iterative manner to refine alignments. Non-conserved side chains and main chains are built using a graph-theoretic approach with sampling provided by exhaustive and database searches. The final conformations are energy minimised to relieve bumps.

A set of possible conformations are generated for main chain regions (loops) considered to vary in the target with respect to the template structures, including insertions and deletions. Main chain sampling is performed using an exhaustive enumeration technique based on 14 discrete torsion angle states. For longer main chain regions, fragments from a database of protein structures are used to generate the torsion angle values. Developments in our ab initio sampling protocol are also incorporated into our loop sampling technique.

At CASP experiments, main chain regions and side chains selected for sampling were determined visually using interactive computer graphics. Some automation to this procedure was accomplished by developing programs to identify side chains with implausible packing, clashes, and unfavorable electrostatic interactions with other side chains and/or main chain.

All-atom conditional probability scoring function

The all-atom scoring function forms the core of any algorithm where identification of native-like conformations is required. The function calculates the probability of a conformation being native-like given a set of interatomic distances [22]. The conditional probabilities are compiled by counting frequencies of distances between pairs of atom types in a database of protein structures. All non-hydrogen atoms are considered, and a residue-specific description of the atoms is used, i.e., the C‰

of an alanine is different from the C‰ of a glycine. This results in a total of 167 atom types. The distances observed are divided into 1.0 ° A bins ranging from 3.0 ° A to 20.0 ° A. Contacts between atom types in the 0-3 ° A range are placed in a separate bin, resulting in a total of 18 distance bins. Distances within a single residue are not included in the counts.

We then compile tables of scores proportional to the negative log conditional probability that one is observing a native conformation given an interatomic distance for all possible pairs of the 167 atom types for the 18 distance ranges. Given a set of distances in a conformation, the probability that the conformation represents a "correct" fold is evaluated by summing the scores for all distances and the corresponding atom pairs.

Using graph theory to generate consistent conformations

We use a graphtheoretic approach to assemble the sampled side chain and main chain conformations together in a consistent and optimal manner: Each possible conformation of a residue is represented using the notion of a node in a graph. Each node is given a weight based on the degree of the interaction between its side chain atoms and the local main chain atoms. The weights are computed using the all-atom scoring function [22]. Edges are then drawn between pairs of residues/nodes that are consistent with each other (i.e., clashfree and satisfying geometrical constraints). The edges are also weighted according to the probability of the interaction between atoms in the two residues. Once the entire graph is constructed, all the maximal sets of completely connected nodes (cliques) are found using a clique-finding algorithm [23]. The cliques with the best total weights represent the optimal combinations of mixing and matching between the various possibilities, taking the respective environments into account [24].

Selecting the most nativelike conformations

All models produced are refined using the Energy Calculation and Dynamics (ENCAD) package [25]. For a given protein sequence, there could be more than one all-atom model produced. For such cases, all models are ranked using the all-atom pairwise scoring function [22] and the best scoring models are considered to be the most native-like ones. This approach is generally more effective than using sequence information alone.

Ab initio prediction

Sequence clusters without known homologues or analogues that are small in size and/or predicted to have largely helical content are modelled by our ab initio protocol. Such clusters may be subsequences of larger proteins, in which case they most likely represent domain boundaries [26]. Representative members of the clusters are targets for modelling, and consensus results are used to assign confidence levels to the models produced.

Our general paradigm for predicting structure involves sampling the conformational space (or generating "decoys") such that native-like conformations are observed, and then selecting them using a hierarchical filtering technique with many different scoring functions (Figure 2). This

approach is also among the most competitive at the CASP experiments. The two parts to our method are designed so they are completely automated and readily extendable to the genome-wide level. Generally, we explore combinations of different representations/move sets with two search methods for exploring protein conformational space, and combinations of a variety of scoring function "filters" to identify biologically relevant conformations.

Sampling protein conformational space

We initially start with an all-atom conformation where the torsion values for residues predicted to be in helix/sheet by secondary structure prediction [27] are set to idealised values. The remaining □/□ values are set in an extended conformation. Side chain conformations are predicted by simply using the most frequently observed rotamer in a database of protein structures [20]. New conformations are generated by perturbing the existing conformation at an arbitrary residue by one of three methods: (i) the torsion values for three residues with identical sequence from a known structure are used to modify the current conformation; (ii) one of possible 14 torsion (□/□) values derived based on the most frequently occurring torsion values for a given residue in a database of known structures; (iii) one of a possible 14 values based on a virtual coordinate frame where the C‰-C‰ positions of two neighbouring residues are represented by a single virtual bond [28].

The scoring function used for minimisation is primarily a combination of the all-atom function, a hydrophobic compactness function, and a bad contacts function [29]. The scores of the conformations are minimised by a combination of two approaches: a straight-forward monte carlo/simulated annealing approach similar in spirit to that of Baker and colleagues [30], and a genetic algorithm search where trajectories can communicate between each other and exchange substructures that have low energy [31, 32].

We use different move sets, minimisation techniques and parameters, and scoring functions to identify the combinations that sample protein conformational space efficiently and effectively. The methodology development in sampling conformational has impact in comparative modelling for building variable main chains (which represent small versions of the ab initio problem).

Selecting biologically relevant conformations

The conformations generated are minimised using ENCAD [25] and scored using a combination of scoring functions that hierarchically reduces the total number of conformations produced to one or a few final conformations. The scoring functions used for the final filtering include the all-atom function [22], hydrophobic compactness [29], a simple residue-residue contact function [33], a density-scoring function that is based on the distance of a conformation to all its relatives in the conformation pool [15], contact order [34], a secondary structure based scoring function that evaluates the match between the predicted structure and the secondary structure of a final energy-minimised conformation, and standard physics-based electrostatics and Van der Waals terms [35].

We are in the preliminary stages of investigating linear and nonlinear combinations of these

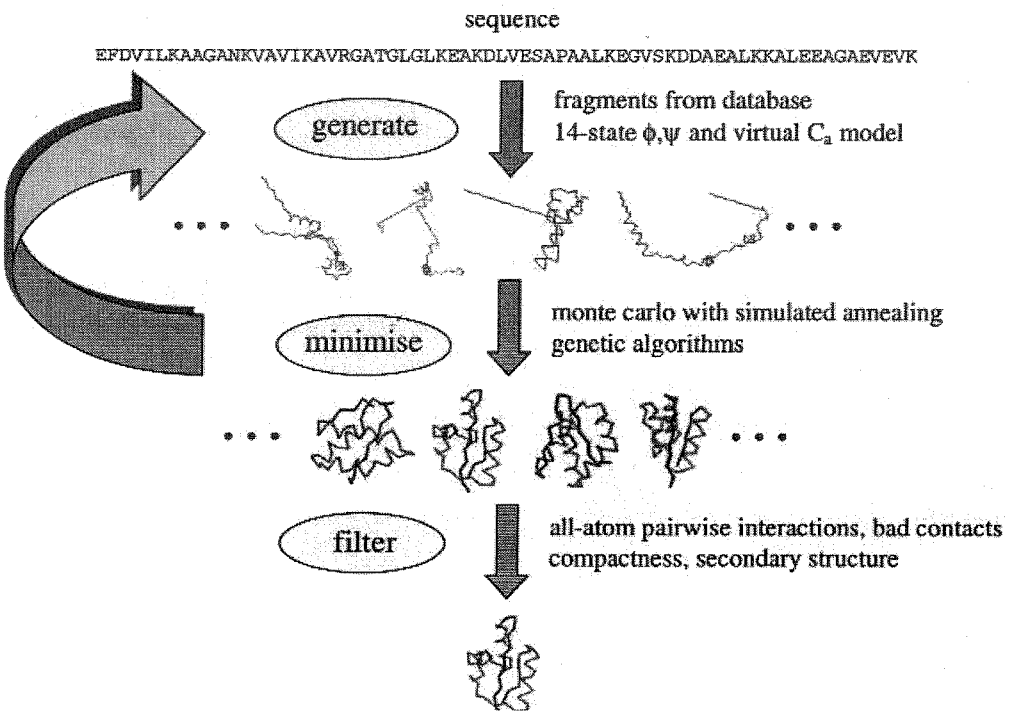

Figure 2: Methodology for ab initio prediction. We start with a sequence and generate conformations using three different move sets: fragments from a database with identical sequence, a 14-state □/€ model, and a 14-state model based on a virtual C‰ bond. Many trajectories are generated and minimised using two differ-ent protocols: monte carlo with simulated annealing and genetic algorithms. The minimisation function is primarily an all-atom conditional probability discriminatory function, an hydrophobic compactness function, and a bad contacts function. Once a set of conformations is generated, a hierarchical filtering technique is applied using many different filters/scoring functions to produce one or a few final conformations.

functions in discriminating native-like folds. Linear combinations of these functions are evaluated by optimising a least-squares function that finds the best of weights to achieve the best discrimination. In addition, we train neural networks on a test set of proteins using the different functions to determine the relative weights assigned to achieve maximum discrimination. We also use filters that include the use of experimental restraints such as cross-linking data and analyse preferences (in terms of volume, radius of gyration and range of scores for each of the scoring functions above) for certain classes and sizes of proteins to eliminate conformations not appropriate for that particular class.

Internal testing of our methods during the development phase

We initially run our algorithms on test sets consisting of 30-50 proteins. To minimise bias of a particular algorithm to a fixed test set, all or portions of the test sets are discarded and replaced with new ones every six months. For testing accuracy of alignments in comparative modelling, our goal are to match alignments produced by a structure comparison between the target and template proteins using only the template structure and the target sequence. In all other cases where a three-dimensional model must be compared to an experimental structure, we use the root mean square deviation (RMSD) between corresponding atoms of the prediction and the experimental answer (usually calculated using the C‰ atoms). In addition, we also use a Z-score that estimates the

number of standard deviations of the RMSD of a particular comparison, relative to an average comparison [36, 37]. The latter metric is more relevant when using models in conjunction with structure comparison to infer function [38].

SOME RESULTS

Comparative modelling

Table 1 shows a general estimate of how well our comparative modelling prediction methods have performed at different CASP experiments. Before the first CASP experiment, published results in the literature usually were obtained by applying structure prediction methods in the context of the exact experimental structure; for example, rebuilding side chains on the native main chain, or rebuilding regions of main chain keeping the rest of the experimental structure fixed. (This practice continues to this day.) CASP1 was an eyeopener in terms of understanding the difficulty of making accurate predictions on approximate templates [39].

Even though our methods produced mediocre results at CASP1, we realised that a major problem with accurate comparative modelling had to do with the interconnected nature of protein structures [40]: If a certain region of the protein varied with respect to the homologue, then it was likely that a structurally interacting region would also vary, even if that region was conserved in sequence. We therefore developed a graphtheory based approach to address this problem which demonstrated significant progress at CASP2 (Table 1) [19]. The CASP3 and CASP4 results do not represent significant improvements over the CASP2 results since the enhancements made to the graph theory method have been minimal.

Figure 3 shows some examples of the comparative modelling predictions with different difficulties made at CASP4. In the comparative modelling category, we made 29 predictions for targets that had sequence identities ranging from 50% to 10% to the nearest related protein with known structure. For 23 of these proteins, we produced models ranging from 1.0 to 6.0 ° A root mean square deviation (RMSD) for the C‰ atoms between the model and the corresponding experimental structure for all or large parts of the protein, with model accuracies scaling fairly linearly with respect to sequence identity (i.e., the higher the sequence identity, the better the prediction).

While the graphtheory methods have been fairly successful at handling the interconnectedness problem to build non-conserved side chains and main chains [19], other major problems preventing the construction of accurate comparative models have to do with inaccurate alignments and using the template structure as a static model upon which to build variable main chains. In the former case, if a region of the alignment is incorrect but is assumed to be correct, then no amount of further model building will fix this error. In the latter case, the loop and side chain construction methods, even if interconnectedness is taken into account, are limited by the approximate nature of the template framework.

Ab initio prediction

As seen in the comparative modelling category, the first CASP experiments did not live up to the results previously published in the ab initio field [41, 42]. It was not until CASP3 that the first consistent positive results were seen: several groups were able to predict the correct topologies for small proteins, or large fragments of a protein (Ÿ 60-80 residues to about 6.0 ° A RMSD relative to the experimental conformation) [43, 44]. CASP4 demonstrated further improvement [14].

Figure 4 illustrates some of our more successful predictions at CASP4 in the ab initio category. We made eleven predictions for targets that had no detectable sequence relationships. We produced nine models with accuracies ranging from 4.0 to 6.0 ° A C‰ RMSD for 60-100 residue proteins (or large fragments of a protein).

While these predictions are a significant improvement compared to the previous CASP results, we still have to make much progress before we can produce models rivalling that of experiment in accuracy. Given the range of RMSDs sampled for each of the proteins (average range for the 11

Category	CASP1	CASP2	CASP3	CASP4
Alignment quality	poor	fair	fair	fair
Side chains	50%	Ÿ 75%	Ÿ 75%	Ÿ 75%
Short loops (lo 6 aa)	Ÿ 3.0 ° A	Ÿ 1.0 ° A	Ÿ 1.0 ° A	Ÿ 1.0 ° A
Longer loops (> 6 aa)	> 5.0 ° A	Ÿ 3.0 ° A	Ÿ 2.5 ° A	Ÿ 1.0 ° A

Table 1: Qualitative assessment of our comparative modelling methods at CASP experiments. For evaluating side chain predictions, the percentage of .1 torsion angles predicted within 30" on average is given. For eval-uating variable main chain (loop) predictions, the average of the C‰ root mean square deviation (RMSDs) (calculated using a global superposition of the target and the model) is shown. The major improvement in our methods from CASP1 to CASP2 is from the use of manually-curated alignments and the development of a graph-theory approach to handle the interconnectedness problem in protein structures.

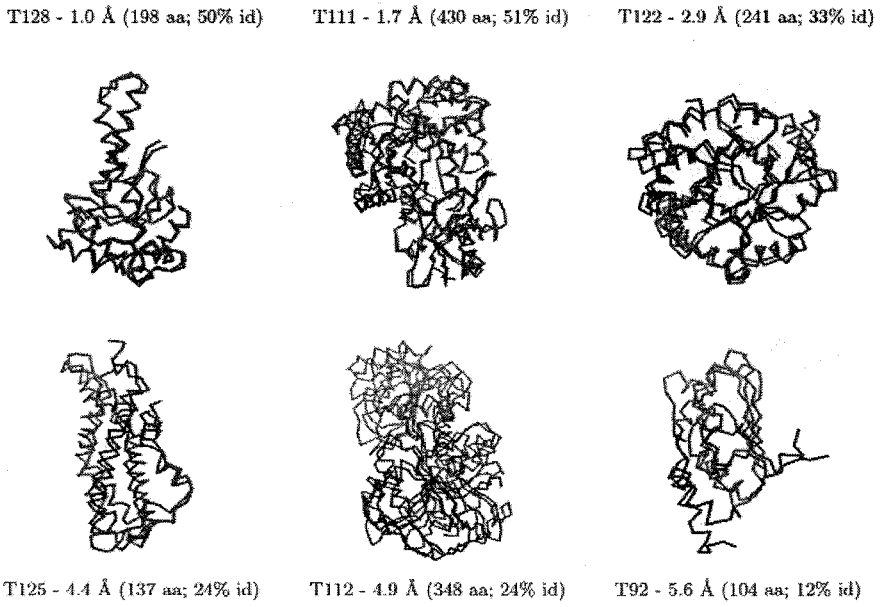

Figure 3: Six examples of our comparative modelling predictions at CASP4 for targets with different diffi-culties. The superposition of the model and the experimental structures is shown, along with the C‰ RMSD relative to the experimental structure and the percentage identity of the alignment between the target and template sequences. We made useful predictions for 23 out of 29 targets: sequences with high percentage identity to the template structures (L 50%) were modelled well (1-2 ° A RMSD) with model accuracy decreas-ing (4-6 ° A RMSD) fairly linearly as the relationship becomes more tenuous (10-25% sequence identity).

predictions was 9.3 - 17.6 ° A), it is clear that devising representations that will allow us to explore protein conformational space such that nearnative conformations are encountered is a major bottleneck. Our scoring functions generally do pick conformations from the lower end of the RMSD distribution (usually within the top 1%, and no worse than the 10%, of the conformations sampled), but further improvements can be made.

Computational times

Table 2 lists the times taken for the computational tasks outlined in this proposal. Times are given per 1000 MHz Pentium III processor and for a cluster of 64 such processors when the algorithm can run in parallel.

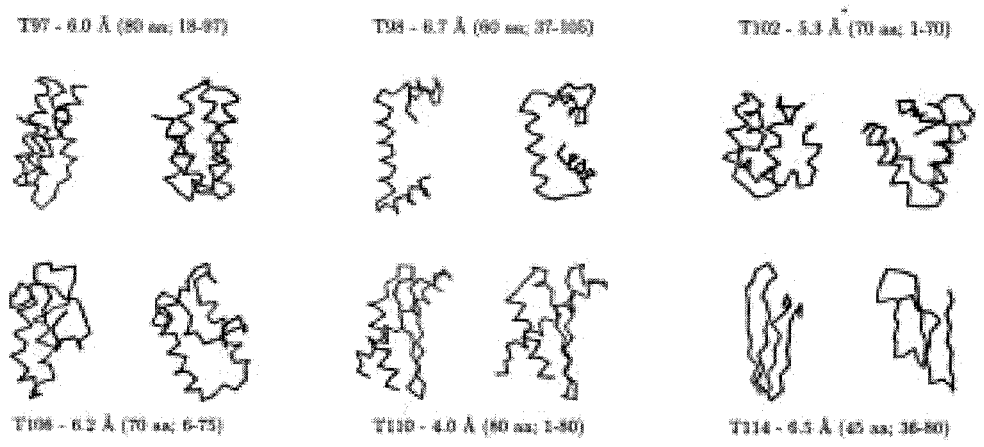

Figure 4: Examples of our ab initio predictions. Five of the examples were predictions submitted for CASP4; the sixth (T102/as48) is a "postdiction" using the actual secondary structure assignment which was available to all CASP predictors (our CASP4 submission for this target used predicted secondary structure which was only 60% accurate). The experimental structure is on the left and the model is on the right. We were able to make topologically accurate predictions for 9 out of 11 targets modelled for all or large parts of the protein. Targets with largely helical content are modelled well, with predictions as accurate as 4.0 ° A C‰ RMSD for 80 residues.

Task	Ÿ Time per CPU	Ÿ Time for cluster
Comparison of two protein sequences	< 1 sec	-
Clustering of sequence families for 3000 proteins	3 days...	1 day
Initial model building by minimum perturbation	< 1 sec	-
Graph-theory search with 30,000 nodes	24 hours	-
Refinement of single model using ENCAD for 200 steps	< 1 sec	-
Evaluationby all-atom function for one conformation	< 1 sec	-
Generating a three-dimensional conformation	< 1 sec	-
Trajectory of 10,000 steps to generate one decoy	1 minute	-
Generating 10,000 decoys	10000 minutes	3 hours

Table 2: Approximate times for certain calculations outlined in this proposal. Times are shown for a single 1000 MHz processor and for a cluster of 64 such processors if the algorithms used can run in parallel. ...indicates times can vary based on the quality of the results desired.

DISCUSSION

We propose to extend our methods to be more complete and rigourous, such that they can be applied to future blind prediction experiments, modelling proteins of particular interest to biology, and modelling whole genomes.

Estimating the reliability of the predictions

Protein structure prediction methods, while promising, are still in their infancy. Therefore, each prediction is internally annotated and assigned a confidence value. In the case of comparative modelling targets, the annotation includes the number of templates used, their scores (as ranked by the all-atom scoring function), and the match between the target and the template proteins (in the form of PSIBLAST evalues). For ab initio targets, the annotation includes a confidence valued based on the secondary structure class and the size of the protein, as well as the score of the final conformation relative to the average score observed for successful predictions with similar secondary structure class and size. In cases where homologous sequences from the same family are modelled, the degree of consensus between the different predictions are incorporated into the confidence assignment.

If given a protein sequence and asked to determine its structure, what can we expect? There will clearly be failures, even among the models assigned a high confidence value. Overall, based on our results from the CASP experiment, we can expect > 70 % of the models produced to be useful for the structure comparison analyses [45] and rational mutagenesis experiments [46] to ascertain function. A smaller fraction of these will yield higher quality models useful for microenvironment analyses to assign function [47]. In addition, the modelling techniques developed may yield novel structural and functional insights for proteins not readily amenable to experimental characterisation. Thus while not all models will be accurately predicted, useful structural and functional models using the methodologies described in this proposal can be produced for a large fraction of the organism's genome with a relatively low cost, compared to experimental approaches.

Application of structure prediction methods to whole genomes

Analyses of small genomes show that about 30-40% of the proteins within the genome can be modelled by comparative modelling methods [17, 45, 48, 49]. An additional 20-30% of the sequences are (or contain) small domains with simple secondary structures that are viable candidates for ab initio structure prediction [38]. The remaining proteins are usually not amenable to structure prediction and sometimes even structure determination (a significant fraction of the latter are membrane proteins).

It is thus possible to construct a "genome prediction engine" using the computational resources available where we can take the protein sequences encoded by an organism's genome and attempt to predict their structures, and use the modelled structures to predict functions. The goal of this endeavour is to improve existing methods and develop new ones to perform various facets of the genome/proteome modelling task.

Using predicted structures to predict function

The reason for obtaining structures for proteins encoded by a genome is so that they can be used to understand function and further our knowledge about the organism's biology. Even though structure prediction methods need further development, it is possible to produce models where functional hypotheses can be tested in a rational manner (for example, with mutagenesis experiments) through detailed analysis [46]. Additionally, structure comparisons can be used to

detect functional homology that cannot be detected by sequence information alone [45], and microenvironment analyses that parse models for particular three-dimensional motifs [47] can be used to discern molecular function. Both these structure-based approaches, used complementarily in conjunction with sequenceonly approaches like PROSITE [50] and experimental data, will enable to us better assign function to all or large parts of a proteome.

Dissemination of information to aid further biological study

The goal of our research is to provide a structural model through computational methods for a given protein with the intent of applying it to all the proteins encoded by an organism's genome. This information will be of greatest value when it is available publicly such that any researcher can use the annotations we make to guide their own experimental study.

We have created databases of the models generated and placed them on our webserver [51] for unrestricted download. The software used to create these models is also freely available so that researchers can make refined predictions of particular proteins of interest. Proteins not modelled by our software will generally be interesting targets for experimental structure determination, thus focusing the efforts of X-ray crystallographers and NMR spectroscopists in an optimal manner.

Future work

Our primary focus is on improving alignment and template selection techniques (for comparative modelling methods), and developing methods for moving an approximate conformation closer to the native structure (for comparative modelling and ab initio methods). Additionally, the lessons we learn from application of our ab initio methodologies will be incorporated to better construct non-conserved side chains and main chains.

Template selection and alignment

There are many methods that attempt to handle the alignment and template selection problems which are still unsolved as judged by the CASP experiments [14, 39, 52]. We propose to try one radically different approach (graphtheory) and one enhancement (PSTs) on an established approach (HMMs) with the aim of improving the current state-of-the-art. We will compare the use of HMMs, PSTs, and the graphtheory based approach for template selection and identification among themselves and to other approaches published in the literature, keeping in mind that at least one sequence-only method is necessary for fast application in a genome modelling scenario. The results produced by the initial application of the sequence-only methods will then be refined by alignment methods that directly incorporate structural information. In addition, we will combine the all-atom function with sequence-only metrics to determine if better discrimination can be achieved.

Refinement of nearnative conformations

Remarkably conspicuous at CASP experiments is the fact that there does not exist a method that can move an approximate conformation closer to the native conformation. The preliminary results indicate that a significant number of conformations must be sampled before generating one that is

closer to the native from the starting conformation. We will use two complementary approaches to address this problem.

In our prior work, we realised that discrimination by the all-atom function could be improved by performing a linear interpolation between the discrete probabilities [22]. We therefore will extend our all-atom function to have a continuous form, so that they can be used in an analytical manner for molecular dynamics and energy minimisation. We will compare the use of Fourier transforms, cublic splines, and polynomial interpolation to represent the discrete probabilities into continuous curves.

Complementarily, we will rigourously determine the degree of sampling required to generate conformations that are closer to the native one, ignoring the issue of selection. We will also use different representations (14-state models, fragments from a database, and a virtual C‰ representation [28]) to determine which ones are amenable to this type of neighbourhood sampling.

General relevance of predicting protein structure from sequence

The continually increasing amount of DNA and protein sequence data from genome projects makes it infeasible for NMR and x-ray crystallography techniques to rapidly provide information about the 3D structures of all the sequences determined [53]. Thus there is an urgent need for predicting structure from amino acid sequence.

There are several justifications for developing and improving protein structure prediction methods: The structure prediction problem is one of the most intellectually challenging problems in biology. Knowing the structure of a protein sequence enables us to probe the function of the protein [54, 55, 56, 57], understand substrate and ligand binding [58, 59, 60, 61], devise intelligent mutagenesis and biochemical protein engineering experiments that improve specificity and stability [62, 63, 64, 65], perform rational drug design [66, 67], and design novel proteins [68, 69, 70]. Understanding structure has potential applications in the various genome projects being undertaken, such as mapping the functions of proteins in metabolic pathways for whole genomes [71, 72] and deducing evolutionary relationships [73]. Understanding protein structure will allow us to design completely novel folds and functions with applications in other areas such as nanotechnology and biological computers.

Proteins in a cell do not work in isolation of one another. Thus to understand the function of multi-protein complexes, or whole proteomes, from a structural viewpoint, it is necessary to have a model for many proteins encoded by the genome of an organism. The CASP results indicate that structure prediction methods have matured to a point where they can be applied on a genomewide scale, and that these structures can be used with novel but straightforward approaches to understand molecular function [46, 47, 74]. The resulting models when combined with other genomic/proteomic data, including that from gene expression arrays [75], genomewide two-hybrid experiments [76], and other proteomics studies [77], will provide us with a dynamic picture of organismal structure, function, and evolution [78].

Availability of software and decoys

The ensembles of structures that were generated and much of the software used to generate them are available at <http://compbio.washington.edu>.

ACKNOWLEDGEMENTS

This work was supported in part by a Burroughs Wellcome Fund Fellowship from the Program in Mathematics and Molecular Biology to Ram Samudrala. The author wishes to thank John Moult and Michael Levitt and past and present members of the Moult and Levitt groups for their intellectual guidance and support that makes this review possible.

REFERENCES

[1] Doolittle R. Similar amino acid sequences: chance or common ancestry? Science. 214: 149–159 (1981).
[2] Greer J. Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins: Struct., Funct., Genet. 7: 317–334 (1990).
[3] Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: Struct., Funct., Genet. 9: 56–68 (1991).
[4] Murzin A, Bateman A. Distant homology recognition using structural classification of proteins. Proteins: Struct., Funct., Genet. 29S: 105–112 (1997).
[5] Bowie J, L" uthy R, Eisenberg D. Method to identify protein sequences that fold into a known three-dimensional structure. Science. 253: 164– 170 (1991).
[6] Jones D, Taylor W, Thornton J. A new approach to protein fold recognition. Nature. 258: 86–89 (1992).
[7] Fl" ockner H, Domingues F, Sippl M. Protein folds from pair interactions: a blind test in fold recognition. Proteins: Struct., Funct., Genet. S1: 129–133 (1997).
[8] Lee J, Liwo A, Ripoll D, Pillardy J, Scheraga J. Calculation of protein conformation by global optimization of a potential energy function. Proteins: Struct., Funct., Genet. S3: 204– 208 (1999).
[9] Ortiz A, Kolinkski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins: Struct., Funct., Genet. S3: 177–185 (1999).
[10] Osguthorpe D. Improved ab initio predictions with a simplified flexible geometry model. Proteins: Struct., Funct., Genet. S3: 186–193 (1999).
[11] Samudrala R, Xia Y, Huang E, Levitt M. Ab initio protein structure prediction using a combined hierarchical approach. Proteins: Struct., Funct., Genet. S3: 194–198 (1999).
[12] Simons K, Bonneau R, Ruczinski I, Baker D. Ab initio structure prediction of CASP3 targets using ROSETTA. Proteins: Struct., Funct., Genet. S3: 171–176 (1999).
[13] Moult J, Hubbard T, Fidelis K, Pedersen J. Critical Assessment of Methods of Protein Structure Prediction (CASP): Round III. Proteins: Struct., Funct., Genet. S3: 2–6 (1999).
[14] Critical Assessment of protein Structure Prediction methods <http://predictioncenter.- llnl.gov/>.
[15] Samudrala R, Levitt M. A comprehensive analysis of 40 blind protein structure predictions. In preparation (2001).
[16] RAMP: A suite of programs to aid in the modelling of protein structure and function <http:-//compbio.washington.edu/ramp/>.
[17] Jones D. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequence. J. Mol. Biol. 287: 797–815 (1999).
[18] Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R. Predicting protein structure using only sequence information. Proteins: Struct., Funct., Genet. S3: 121–125 (1999).
[19] Samudrala R, Moult J. Handling context-sensitivity in protein structures using graph theory: bona fide prediction. Proteins: Struct., Funct., Genet. 29S: 43–49 (1997).
[20] Samudrala R, Huang E, Koehl P, Levitt M. Side chain construction on nearnative main chains for ab initio protein structure prediction. Protein Eng. 7: 453–457 (2000).

[21] Samudrala R, Moult J. Determinants of side chain conformational preferences in protein structures. Protein Eng. 11: 991–997 (1998).
[22] Samudrala R, Moult J. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. 275: 895–916 (1998).
[23] Bron C, Kerbosch J. Algorithm 457: Finding all cliques of an undirected graph. Comm. ACM. 16: 575–577 (1973).
[24] Samudrala R, Moult J. A graphtheoretic algorithm for comparative modelling of protein structure. J. Mol. Biol. 279: 287–302 (1998).
[25] Levitt M, Hirshberg M, Sharon R, Daggett V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comp. Phys. Comm. 91: 215–231 (1995).
[26] Gouzy J, Corpet F, Kahn D. Whole genome protein domain analysis using a new method for domain clustering. Comp. and Chem. 23: 333–340 (1999).
[27] Jones D. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 195–202 (1999).
[28] Hunter C, Subramaniam S. A natural coordinate frame for protein backbone structures. Proteins: Struct., Funct., Genet. page submitted (2001).
[29] Samudrala R, Xia Y, Levitt M, Huang E. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. In Altman R, Dunker A, Hunter L, Klein T, Lauderdale K, editors, Proceedings of the Pacific Symposium on Biocomputing, pages 505–516, (1999).
[30] Simons K, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J. Mol. Biol. 268: 209–225 (1997).
[31] Pedersen J. T, Moult J. Folding simulation with genetic algorithms and a detailed molecular description. J. Mol. Biol. 269: 240–259 (1997).
[32] Dandekar T, Argos P. Applying experimental data to protein fold prediction with the genetic algorithm. Protein Eng. 10: 877–893 (1997).
[33] Huang E, Subbiah S, Levitt M. Recognising native folds by the arrangement of hydrophobic and polar residues. J. Mol. Biol. 252: 709–720 (1995).
[34] Plaxco K, Simons K, Baker D. Contact order, transition state placement, and the refolding rates of single domain proteins. J. Mol. Biol. 277: 985–994 (1998).
[35] Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4: 187–217 (1983).
[36] Holm L, Sander C. DALI/FSSP classification of three-dimensional protein folds. Nucleic Acids Res. 25: 231–234 (1996).
[37] Shindyalov I, Bourne P. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11: 739– 747 (1998).
[38] Bonneau R, Baker D. Ab initio protein structure prediction: Progress and prospects. Annu. Rev. Biophy. Biomol. Struct. 30: 173–89 (2001).
[39] Mosimann S, Meleshko R, James M. A critical assessment of comparative molecular modeling of tertiary structures in proteins. Proteins: Struct., Funct., Genet. 23: 301–317 (1995).
[40] Samudrala R, Pedersen J, Zhou H, Luo R, Fidelis K, Moult J. Confronting the problem of interconnected structural changes in the comparative modelling of proteins. Proteins: Struct., Funct., Genet. 23: 327–336 (1995).
[41] Defay T, Cohen F. Evaluation of current techniques for ab initio protein structure prediction. Proteins: Struct., Funct., Genet. 23: 431–445 (1995).
[42] Lesk A. CASP2: Report on ab initio predictions. Proteins: Struct., Funct., Genet. 29S: 151–166 (1997).
[43] Orengo C, Bray J, Hubbard T, LoConte L, Sillitoe J. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins: Struct., Funct., Genet. S3: 149–170 (1999).
[44] Koehl P, Levitt M. A brighter future for protein structure prediction. Nature Struct. Biol. 6: 108–111 (1999).
[45] Brenner S, Levitt M. Expectations from structural genomics. Protein Sci. 9: 197–200 (2000).
[46] Samudrala R, Xia Y, Levitt M, Cotton N, Huang E, Davis R. Probing structure-function relationships of the dna polymerase alpha-associated zinc-finger protein using computational approaches. In Altman R, Dunker A, Hunter L, Klein T, Lauderdale K, editors, Proceedings of the Pacific Symposium on Biocomputing, page (in press), (2000).
[47] Wei L, Huang E, Altman R. Are predicted structures good enough to preserve functional sites? Structure. 7: 643–650 (1999).

[48] Sanchez R, Sali A. Largescale protein structure modeling of the Saccharomyces cerevisiae genome. Proc. Natl. Acad. Sci. USA. 95: 13597- 13602 (1998).
[49] Martin-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophy. Biomol. Struct. 29: 291-325 (2000).
[50] Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. Nucleic Acids Res. 27: 215-219 (1999).
[51] Samudrala Computational Genomics Group <http://compbio.washington.edu>. [52] Jones T, Kleywegt G. CASP3 comparative modeling evaluation. Proteins: Struct., Funct., Genet. S3: 30-46 (1999).
[53] May A, Johnson M, Rufino S, Wako H, Zhu Z, Sowdhamini R, Srinivasan N, Rodionov M, Blun-dell T. The recognition of protein structure and function from sequence: adding value to genome data. Phil. Trans. Roy. Soc. Lond. 344: 373-381 (1994).
[54] Luecke H, Quiocho F. High specificity of a phosphate transport protein determined by hydrogen bonds. Nature (London). 347: 402-406 (1990).
[55] Hughes R, Hatfull G, Rice P, Steitz T, Grindley N. Cooperativity mutants of the gamma delta resolvase identify an essential interdimer interaction. Cell. 63: 1331-1338 (1990).
[56] Herzberg O. An atomic model for proteinprotein phosphoryl group transfer. J. Biol. Chem. 267: 24819-24823 (1992).
[57] Liu X, Zhu H, Huang B, Rogers J, Yu B, Kumar A, Jain M, Sundaralingam M, Tsai M. Phospholipase A2 engineering. Probing the structural and functional roles of N-terminal residues with site-directed mutagenesis, X-ray, and NMR. Bio-chemistry. 34: 7322-7334 (1995).
[58] Brick P, Bhat T, Blow D. Structure of tyrosyl-tRNA synthetase refined at 2.3 ° A resolution. Interaction of the enzyme with tyrosyl adenylate intermediate. J. Mol. Biol. 208: 83-98 (1988).
[59] Rould M, Perona J, Soll D, Steitz T. Structure of E. coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 A resolution. Science. 246: 1135-1142 (1989).
[60] Schulz G, Muller C, Diederichs K. Induced-fit movements in adenylate kinases. J. Mol. Biol. 213: 627-630 (1990).
[61] Cahoon E, Lindqvist Y, Schneider G, Shanklin J. Redesign of soluble fatty acid desaturases from plants for altered substrate specificity and double bond position. Proc. Natl. Acad. Sci. USA. 94: 4872-4877 (1997).
[62] Ulmner K. Protein engineering. Science. 219: 666-671 (1983).
[63] Pantoliano M, Whitlow M, Wood J, Dodd S, Hardman K, Rollence M, Bryan P. Large increases in general stability for subtilising BPN' through incremental changes in the free energy of unfolding. Biochemistry. 28: 7205-7213 (1988).
[64] Hua Q, Hu S, Frank B, Jia W, Chu Y, Wang S, Burke G, Katsoyannis P, Weiss M. Mapping the functional surface of insulin by design: structure and function of a novel A-chain analogue. J. Mol. Biol. 264: 390-403 (1996).
[65] Rezaie A, Olson S. Contribution of lysine 60f to S1' specificity of thrombin. Biochemistry. 36: 1026-1033 (1997).
[66] Hunter W, Bailey S, Habash J, Harrop S, Helli-well J, Aboagye-Kwarteng T, Smith K, Fairlamb. Active site of trypanothione reductase. A target for rational drug design. J. Mol. Biol. 227: 322- 333 (1992).
[67] Blundell T. Structure-based drug design. Nature (London). 384: 23-26 (1996).
[68] Bryson J, Betz S, Lu H, Suich D, Zhou H, O'Neil K, DeGrado W. Protein design: a hierarchic approach. Science. 270: 935-941 (1995).
[69] Smith C, Regan L. Guidelines for protein design: the energetics of beta sheet side chain interactions. Science. 270: 980-982 (1995).
[70] Cordes M, Davidson A, Sauer R. Sequence space, folding and protein design. Curr. Opin. Struct. Biol. 6: 3-10 (1996).
[71] Clegg M, Gaut B, Learn G.H. J, Morton B. Rates and patterns of chloroplast DNA evolution. Proc. Natl. Acad. Sci. USA. 91: 6795-6801 (1994).
[72] Holm L, Sander C. Mapping the protein universe. Science. 273: 595-603 (1996).
[73] Hubbard T, Murzin A, Brenner S, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res. 25: 236-239 (1997).
[74] Barnes A, Wynn C. Homology of lysozomol enzymes and related proteins: Prediction of posttranslational modification sites including phosphorylation of mannose and potential epitopic and substrate binding sites in the ‰ • and „ • submits of hexosaminidases, ‰ • glucosidase and rabbit and human isomaltase. Proteins: Struct., Funct., Genet. 4: 182-189 (1988).

[75] Lander E. Array of hope. Nature Genet. 21: 3 (1999).
[76] Schwikowski B, Uetz P, Fields S. A network of proteinprotein interactions in yeast. Nature Biotechnol. 18: 1242–1243 (2000).
[77] Gygi S, Rist B, Gerber S, Turecek F, Gelb M, Aebersold R. Quantitative analysis of complex protein mixtures using isotopecoded affinity tags. Nature Biotechnol. 17: 994–999 (1999).
[78] Ideker T, Thorsson V, Ranish J, Christmas R, Buhler J, Eng J, Bumgarner R, Goodlett D, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science. 292: 929–934 (2001).