JMB

A Graph-theoretic Algorithm for Comparative Modeling of Protein Structure

Ram Samudrala[1,2] and John Moult[1]*

[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600, Gudelsky Drive, Rockville, MD 20850 USA*

[2]*Molecular and Cell Biology Program, University of Maryland at College Park College Park, MD 20742, USA*

*Corresponding author

The interconnected nature of interactions in protein structures appears to be the major hurdle in preventing the construction of accurate comparative models. We present an algorithm that uses graph theory to handle this problem. Each possible conformation of a residue in an amino acid sequence 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. Edges are then drawn between pairs of residue conformations/nodes that are consistent with each other (i.e. clash-free and satisfying geometrical constraints). The edges are weighted based on the interactions between the atoms of the two nodes. Once the entire graph is constructed, all the maximal sets of completely connected nodes (cliques) are found using a clique-finding algorithm. The cliques with the best weights represent the optimal combinations of the various main-chain and side-chain possibilities, taking the respective environments into account. The algorithm is used in a comparative modeling scenario to build side-chains, regions of main chain, and mix and match between different homologs in a context-sensitive manner. The predictive power of this method is assessed by applying it to cases where the experimental structure is not known in advance.

© 1998 Academic Press Limited

Keywords: graph theory; clique finding; comparative modeling; context-sensitivity; inter-connectedness

Introduction

The rapidly increasing number of known protein structures has resulted in a situation where approximate structures corresponding to new sequences are often available from one of two sources. First, when the sequence of interest is clearly related to those of one or more known structures, then the overall folds are the same (Chothia & Lesk, 1986). This is the case now for about 30% of the general sequences entering the databases (Schneider & Sander, 1996), and about 10% of genome sequences (Scharf *et al.*, 1994). Second, even when there is no detectable sequence

relationship, the techniques of threading make it possible to recognize the fold in many cases (Levitt, 1997). Extrapolating this trend, it appears that the routine generation of approximate models of protein structure from sequence may soon become a reality.

To be of much practical use, these approximate structures need to be refined into detailed accurate models. The technique for doing this is usually termed comparative or homology modeling. In contrast to progress in generating approximate structures, this process has turned out to be more difficult. Objective testing of the methods (Mosimann *et al.*, 1995; Martin *et al.*, 1997) shows that the current numerical techniques offer little advantage over simply copying large parts of the related structures. There are two fundamental difficulties to be overcome. First, methods must deal with the compact states of the polypeptide chain, where steric exclusion effects makes the energy surface extremely discontinuous, so that search methods that make semi-random moves such as Monte Carlo (Chen, 1989; Skolnick & Kolinski, 1990; Wilson & Cui, 1990; Okamoto *et al.*, 1991;

Abagyan & Totrov, 1994; Avbelj & Moult, 1995) and genetic algorithms (Sun, 1993; Unger & Moult, 1993; Pedersen & Moult, 1997) have difficulty finding acceptable conformations, while continuous search methods such as molecular dynamics appear to get stuck in local parts of the space (Venclovas *et al.*, 1997). Second, in many instances the details of the conformation are highly context-dependent: for example, the conformation adopted by a particular stretch of polypeptide chain can be different, depending on the environment it is in (Samudrala *et al.*, 1995). Thus, to be effective, an algorithm must be able to cope with the discontinuous nature of the search space, and to take into account an extensive web of interactions.

In computing science, the notion of a "graph" has been used to describe many systems that are made up of such interconnected networks (Harel, 1992). These include laying out the shortest combination of railroad segments between a network of cities (finding minimal spanning trees), finding the shortest paths between any two cities in a network of cities, and finding the shortest path in a city network, passing through all the cities exactly once (the famous Travelling Salesman problem). In computational chemistry and biology, graph-theoretic approaches have been used to enumerate chemical isomers (National Research Council, 1995) and for protein structure comparison (Grindley *et al.*, 1993; Artymiuk *et al.*, 1995).

Our goal is to find the best set of interactions in a protein structure, given a variety of side-chain and main-chain conformational choices for each position in the structure. We present an algorithm based on graph theory that will find the optimal arrangement of all these choices, as measured by some discriminatory function, while adequately considering the context-sensitivity seen in protein structures.

Specifically, we present conformations of parts of protein molecules, usually single amino acid residues, as nodes in a graph. Edges are drawn between self-consistent sub-conformations and nodes and edges are weighted with some fitness function. The maximal completely connected graphs (cliques) then represent possible conformations of the molecule, and those with the best weight are assumed to be the most native-like. In principle, the method can be applied to any structure prediction problem. In practice, computational limitations on the number of combinations of conformations that can be considered make it most suitable for comparative modeling applications.

There are three principle advantages to the graph-theoretic representation: (1) it provides a simple framework in which to consider the combinations of possible sub-conformations systematically, avoiding the need for following a trajectory through the rough energy landscape. (2) It provides control over which sub-conformations to include, allowing resources to be focused on the more uncertain aspects of the structure. (3) Pre-calculation of the fitness of each node, and of the

interaction between pairs of nodes, greatly reduces the computational cost of evaluating a conformation, allowing many more combinations to be considered compared to a conventional fitness calculation.

Comparative modeling can be regarded as a series of steps. Generally, an alignment between the sequence to be modeled (the target) and a related sequence with known structure (the parent of the template) is first constructed (Greer, 1990; Mosimann *et al.*, 1995). An initial partial model is then built by copying the main-chain coordinates from the parent structure(s) for equivalent residues in the alignment. Some side-chain conformations may be inferred from the parent structures. Remaining parts of the structure (insertions in the target sequence relative to the parent, rejoining of the chain around deletions, regions of chain with low levels of sequence homology between the target and parent, regions where an alternative parent structure may result in a more accurate model, and other side-chain conformations) must then be built.

We describe how these building steps can be accomplished with the graph-theoretic clique-finding method. We summarize the results from the second experiment on the Critical Assessment of protein Structure Prediction methods (CASP2), where the method was applied to build comparative models of sequences for which the experimental structures were not then known (Samudrala & Moult, 1997).
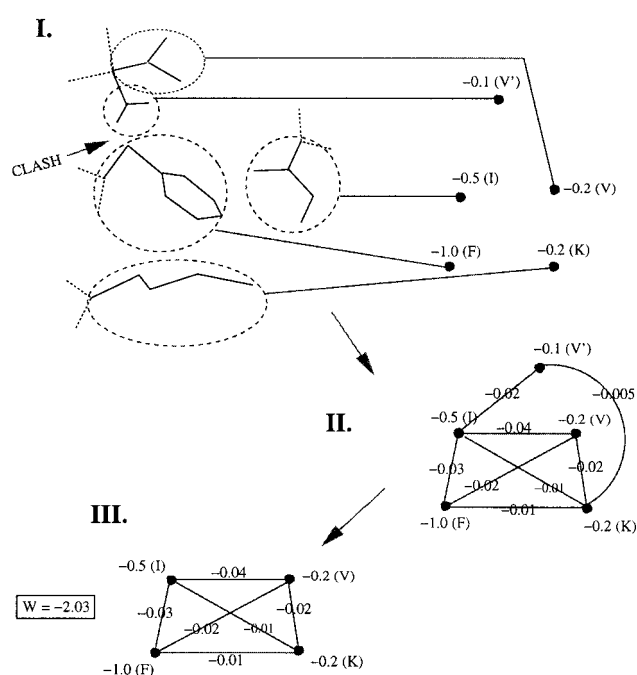
Methods

General description

Each possible conformation of a residue in an amino acid sequence is represented as a node in a graph. Edges are then drawn between pairs of residues/nodes that are consistent with each other. Edges and nodes are weighted according to some fixed criteria. Once the entire graph is constructed, all the maximal sets of completely connected nodes (cliques) are found using a clique-finding (CF) algorithm. The cliques with the best weight are considered to be similar to the native structure. Figure 1 illustrates how the CF method is used to model structures.

Description of nodes

Each possible conformation of a residue (side-chain and main-chain) represents a node in the graph. Nodes have weights based on the strength of the interaction between the side-chain atoms and the local main-chain atoms. The main-chain atoms up to four residues on either side of the residue position representing the node, and the main-chain atoms within the residue, are considered for calculating the weights.

Figure 1. Illustration of the graph-theoretic, clique-finding (CF) method for protein structure prediction. In the first step, possible side-chain and main-chain conformations of individual residues are represented as nodes in a graph, and each node is weighted according to the interactions between the side-chain and the local main chain. In this idealized example, three residue positions (isoleucine (I), lysine (K), phenylalanine (F)) with a single possible conformation and one residue (valine) with two possible conformations (V and V′) are shown, resulting in five nodes with different weights. In step II, edges are drawn between consistent nodes. In the example, the inconsistent pairs of nodes are those representing the two different valine conformations V and V′ (an amino acid cannot have two conformations simultaneously) and a clash that occurs between V′ and F. Edges are not drawn between these pairs of nodes. Edges are drawn between all other pairs of nodes and each edge is assigned a weight based on the interaction between the pair of residue conformations (nodes). In the third step, all maximal completely connected subgraphs, or cliques, the size of the region considered, where every node is connected to every other node, are found and the total weights of the cliques are calculated by summing the weight of the nodes and the edges. Each clique represents a plausible conformation of the entire region of protein considered and the clique with the best weight is assumed to represent the correct structure. In this example, there is only one clique with nodes {I,V,K,F}. A potential clique I,V′,K,F is incomplete because of the clash between V′ and F.

Description of edges

Edges are drawn between pairs of nodes. Edges are weighted based on the strength of interaction between the atoms of the pair of residues representing the nodes. Edges are drawn in a consistent manner. (1) Packing consistency is maintained by

not drawing edges between nodes whose atoms clash with each other. (2) Main-chain consistency is maintained by partitioning the complete protein main-chain conformation into segments. Each segment may have one or more conformations. If two nodes represent residue conformations within the same main-chain segment, then both conformations must be part of the same segment conformation for an edge to be drawn between them (Figure 2). (3) Edges not drawn between different possible side-chain conformations of the same residue.

Description of the clique-finding method

The clique-finding (CF) algorithm that we use was developed by Bron & Kerbosch (1973). This algorithm combines a recursive backtracking procedure with a branch and bound technique to eliminate searches that cannot lead to a clique. The recursive procedure is self-referential: finding a clique of length n is accomplished by finding a clique of length $n-1$ and finding another node that is connected to all the nodes in that clique. The branch and bound technique makes use of rules that allow us to determine in advance certain cases for which possible combinations of nodes and edges will never lead to a clique.

There are three sets that are essential for this algorithm: (1) `potential-clique`; this is a set of nodes where every node is connected to every other node. Each recursive call will either extend this set by one node or reduce it by one node. (2) `candidates`; this is the set of nodes that are eligible for addition to the `potential-clique` set. (3) `Already-found`; this is the set of nodes that have already served as an extension to the present configuration of `potential-clique` and are now explicitly excluded. That is, all possible extensions of `potential-clique` containing any point in this set have already been generated.

The algorithm operates recursively on each of the sets by generating all extensions of a given configuration of `potential-clique` that it can make with the given set of `candidates` and that do not contain any of the nodes in `already-found`, as described in the simplified pseudocode representation (Scheme 1):

Initially, the set `candidates` contains all the nodes in the graph and the sets of `potential-clique` and `already-found` are empty. Bron & Kerbosch (1973) select their nodes in a clever manner by choosing nodes with the largest number of edges to reach the branch and bound condition (see pseudocode implementation above) as soon as possible. This leads to the larger cliques being found first and sequentially generates cliques having a large common intersection. More details of this algorithm, including a more detailed pseudocode implementation, are given by Bron & Kerbosch (1973).

```
begin procedure find-cliques(potential-clique, candidates, already-found)

   if a node in already-found is connected to all nodes in candidates then
     no clique can ever be found (branch and bound step)
   else
     foreach candidate-node in candidates do
       move candidate-node to potential-clique
       create new-candidates by removing nodes in candidates not connected to candidate-node
       create new-already-found by removing nodes in already-found not connected to candidate-node
       if new-candidates and new-already-found are empty then
           potential-clique is a maximal-clique
       else
         find-cliques(potential-clique, new-candidates, new-already-found)
       endif
       move candidate-node from potential-clique to already-found
     endfor
   endif

end procedure find-cliques
```

Scheme 1

All-atom, distance-dependent conditional probability discriminatory function

We use an all-atom, distance-dependent conditional probability-based discriminatory function to calculate the probability of a native structure, given a set of distances between pairs of atoms. A full description can be found in Samudrala & Moult (1998). Briefly, the required 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 the description of the atoms is residue-specific, i.e. the C^α of the alanine residue is different from the C^α of a glycine residue. This results in a total of 167 atom types. We divide the distances observed into 1.0 Å bins ranging from 3.0 Å to 20.0 Å. Contacts between atom types in the 0.0 to 3.0 Å range are placed in a separate bin, resulting in total of 18 distance bins. For observations of distances between pairs of atoms between the atoms of a side-chain and the main-chain atoms of that residue, a separate table of frequencies is compiled using 18 1.0 Å bins ranging from 0.0 Å to 18.0 Å.

We compile tables of scores s proportional to the negative log conditional probability that we are observing a native conformation given an interatomic distance d for all possible pairs of the 167 atom types, a and b, for the 18 distance ranges, $P(C \mid d_{ab})$:

$$s(d_{ab}) = -\ln \frac{P(d_{ab}|C)}{P(d_{ab})} \propto -\ln P(C|d_{ab}) \qquad (1)$$

where $P(d_{ab} \mid C)$ is the probability of observing a distance d between atom types a and b in a correct structure, and $P(d_{ab})$ is the probability of observing such a distance in any structure, correct or incorrect. The required ratios $P(d_{ab} \mid C)/P(d_{ab})$ are obtained as follows:

$$\frac{P(d_{ab}|C)}{P(d_{ab})} = \frac{N(d_{ab})/\Sigma_d N(d_{ab})}{\Sigma_{ab}N(d_{ab})/\Sigma_d \Sigma_{ab}N(d_{ab})} \qquad (2)$$

where $N(d_{ab})$ is the number of observations of atom types a and b in a particular distance bin d, $\Sigma_d N(d_{ab})$ is the number of a-b contacts observed for all distance bins, $\Sigma_{ab}N(d_{ab})$ is the total number of contacts between all pairs of atom types a and b in a particular distance bin d, and $\Sigma_d \Sigma_{ab}N(d_{ab})$ is the total number of contacts between all pairs of atom types a and b summed over all the distance bins d.
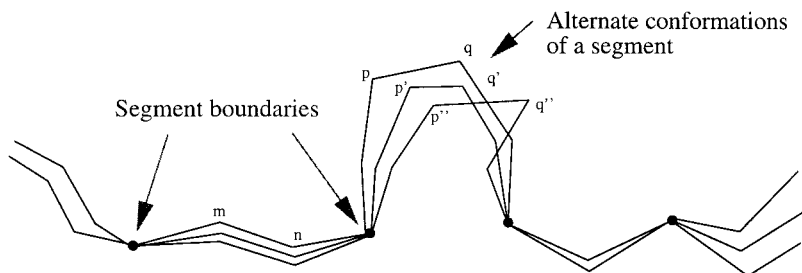
Figure 2. Alternative main-chain conformations used in building the graph. The main-chain is divided into segments, and each segment can have one or more conformations. Different conformations of a segment may represent alternative parent main-chains, or alternative loop conformations. Junctions between segments are positions where only one conformation is allowed. Edges may be drawn between nodes representing different side-chain conformations within the same main-chain segment (for example, between *m* and *n*). Edges may also be drawn between nodes representing different conformations in different segments (for example, between *n* and any *p*). Edges cannot be drawn between nodes on different conformations of a segment (for example, between *q* and *q'*).

Intra-residue distances are not included in the summation.

The tables of scores are compiled from a set of 265 non-homologous (less than 30% sequence identity between any proteins in the set) high-resolution (less than 3.0 Å) X-ray structures (Orengo *et al.*, 1993 available at http://www.biochem.ucl.ac.uk/bsm/cath/).

Evaluation of the probability that a clique represents a native conformation

Given a clique of n nodes and m edges, the total score representing the probability that the corresponding conformation is correct is expressed as a sum over the nodes and the edges:

$$S(\text{clique}) = \sum_n S(\text{node}) + \sum_m S(\text{edge}) \qquad (3)$$

where $S(\text{node})$ is the sum of the scores for the distances between all atoms p of the side-chain and all atoms q of the total main-chain (±four residues, total of nine where available):

$$S(\text{node}) = \sum_{pq} s(d_{ab}^{pq}) \qquad (4)$$

and $S(\text{edge})$ is the sum of the scores for the distances between atoms r of one residue and atoms s of the other:

$$S(\text{edge}) = \sum_{rs} s(d_{ab}^{rs}) \qquad (5)$$

If r and s are within four residues, then only the side-chain atoms are considered for calculating the score. All node and edge scores, $S(\text{node})$ and $S(\text{edge})$ are computed only once, greatly reducing the cost of calculating the total score for any conformation represented by a clique.

Application to a comparative modeling scenario

We focus on applications to comparative modeling, where large sections of main-chain conformation are taken from one or more related parent structures. Only those main-chain and side-chain conformations that are thought to vary significantly (>2.0 Å RMSD) from the parent structures are sampled using the CF method. Side-chain conformations thought to be conserved between parent and target are built using the minimum perturbation (MP) method implemented by the program MUTATE (written by R. Read). The MP method changes a given amino acid to the target amino acid preserving the equivalent χ angles, as determined by an equivalence table between the two side-chains. The χ angles not present in the model are constructed using a library based on the residue type (Samudrala *et al.*, 1995 and unpublished).

Multiple main-chain conformations are sampled in regions where alternative parents might provide a more complete model, and for insertions and regions surrounding deletions. Possible conformations for these latter regions are generated using a database search. For each side-chain for which the conformation is uncertain, possible conformations are selected using the rotamer library. Each main-chain/side-chain conformation combination is a node in the graph.

Main-chain sampling

The first step in constructing the graph is the selection of the main-chain conformations of each residue. The main chain is divided into segments. Each segment can consist of one or more possible conformations. The junctions between segments are positions on the main chain where only one conformation is allowed (see Figure 2 for an illustration). The possible conformations of main chains representing a single segment are obtained from alternative parents or by database search (Pedersen *et al.*, 1992).

Description of the database method to build loops

One of the most difficult problems in comparative modeling is the construction of regions of chain where no approximate conformation is available from the parent structure(s). Determination of the conformation of these regions (often referred to as loops) is aided by an approximate knowledge of the adjacent main chain (referred to as the roots of a loop) and the more general environment. The conformation is often determined by other features of the environment including adjacent loops.

Main-chain conformations for loop segments are generated using a database of fragment conformations in protein structures. C^α distance constraints from the roots of a loop are used to find a set of compatible main-chain conformations in a database of 520 protein structures (Pedersen *et al.* (1992). Three constraints are used, and their specification is the same as that given by Pedersen *et al.* (1992): if the main-chain region being built is n residues spanning residue positions p to q, then the constraints used are $d(p, q)$, the C^α distance between residues p and q, $d(p, q - 1)$, and $d(p + 1, q)$.

A database conformation is considered to fit the distance constraints if $d(p, q)$ differs by less than ±1.0 Å from the constraint value, and $d(p, q - 1)$ and $d(p + 1, q)$ differ by less than ±2.0 Å. The root residues, i.e. residues flanking the region being built, are defined to be residues $p - 1$ and $q + 1$.

Selected conformations are clustered by ϕ/ψ angles, and sets where all values agree with some cutoff (generally 30°) are replaced by the most representative member of the cluster. The main-chain conformations found are then positioned in the initial model using the methods described by Martin *et al.* (1989) and Pedersen *et al.* (1992). At this point, the conformations of the guiding residues in the initial model used to generate the

distance constraints and for fitting purposes (residues $p, p + 1, q, q - 1$) are removed and the database conformations are used for these regions (residues p through q). A preliminary screening is done to exclude any main-chain conformation that clashes (any interatomic contact less than 2.0 Å) with the main chain of the rest of the initial model.

Side-chain sampling

Given a local main-chain conformation, a set of the most probable side-chain conformations is generated by exploring all conformations allowed by a rotamer library, and calculating the value of the discriminatory function including all atom-atom interactions between the side-chain and the local main chain. Interactions with the main chain up to four residues on either side are included, if available. The rotamer library, by Samudrala & Moult (unpublished), contains up to three values for each χ angle.

Up to six best-scoring conformations per residue are selected to form nodes. We have shown that using an experimental main chain, the correct side-chain rotamer is present in one of the top five conformations more than 80% of the time (unpublished results).

Relationship between clique scores and conformation score

The discriminatory function used to score cliques is similar to but not identical with the one we have extensively tested against a wide range of decoys (Samudrala & Moult, 1997). There are two differences from the carefully tested function: we include intra-residue contributions between side-chains and main chain, and we exclude contributions between main-chain atoms less than four residues apart in the sequence.

To see how the rankings of the conformations are affected by these differences, we compare the scores of conformations obtained by summing weight of the nodes and edges of its clique with the scores of the same conformations obtained by calculating the conditional probabilities of the interatomic contracts as in (Samudrala & Moult, 1997). Figure 3 shows such a comparison for 100,000 conformations of residues 21 to 32 in the α-lactalbumin structure (PDB code 1alc). The fragment in α-lactalbumin is proposed to be an independent folding unit as determined by local hydrophobic burial and experimental evidence (Unger & Moult, 1991; Avbelj & Moult, 1995; Pedersen & Moult, 1997). The conformations represent 100,000 cliques with the best weight obtained after exploring up to six conformations per residue position with a fixed main chain. That is, each of the 100,000 conformations represents a different side-chain arrangement for the 12 residues in the independent folding unit.

Figure 3 shows that even though the correspondence between the two types of score calculation is
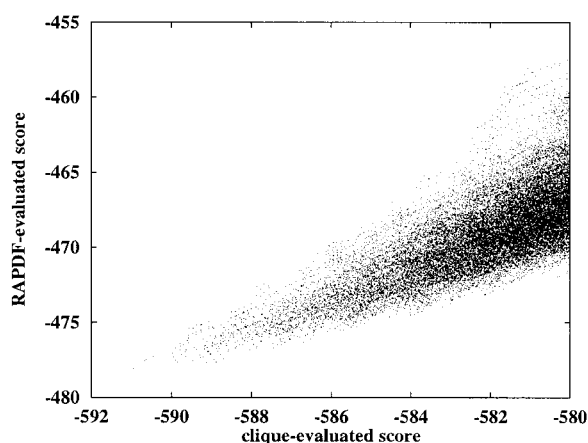
Figure 3. Comparison of the total scores obtained by summing the weights of nodes and edges in a clique (horizontal axis) to those obtained by summing up the probabilities of inter-residue atomic contacts in the 3D conformation represented by the clique. The scores are for 100,000 side-chain conformations/cliques of an independent folding unit, α-lactalbumin (residues 21 to 32; Unger & Moult, 1991; Avbelj & Moult, 1995; Pedersen & Moult, 1997). The two methods of evaluation produce similar but not identical ranking of conformations for the best-scoring structures. In practice, the best scoring 100 cliques are re-evaluated using the corresponding conformations and the residue-specific all-atom conditional probability discriminatory function (Samudrala & Moult, 1997). The best-scoring conformation is considered the best structure.

not perfect, the conformation with the best score is in the set of the ten best clique scores. In our implementation, we retain the 100 cliques with the best scores, and then re-evaluate them by calculating the score for the corresponding conformations. The best-scoring conformation is selected as representing the correct structure.

Implementation issues

The graphs are stored as a set of nodes and an edge matrix of size $n \times n$, where n is the number of nodes. The size of a single element in the matrix, representing an edge, is one byte.

Sometimes, cliques the size of the protein cannot be found because all possible conformations of some residues are inconsistent with the rest of the nodes. There are two options in these cases: either additional possible conformations for those residues are generated by further sampling until the full cliques are found. Or, smaller cliques are used to produce partial structures that are completed by other means, generally by manual intervention.

The algorithm includes zero-weight edges between pairs of residues separated by a large distance in the protein. For proteins that are larger than ≃200 residues, this results in a large number of edges per node and increases the running time of the program. In such cases, we consider only a subset of the protein and omit all residues beyond

a certain cutoff (typically 20.0 Å) from the region of interest.

Results

Building side-chains in a comparative modeling scenario

We illustrate how the clique finding method performs for building side-chains using a comparative target and corresponding model from the first experiment on the Critical Assessment of protein Structure Prediction methods (CASP1). The target is the histidine-containing phosphocarrier protein (hpr) from *Mycoplasma capricolum*, an 89 residue protein (Pieper *et al.*, 1995). In the model of this structure we built for CASP1, 27 of 67 χ_1 angles deviated more than $30°$ from the experimental structure. We rebuild the 27 side-chains using the discriminatory function, side-chain sampling, and clique-finding methods described above. We compare the accuracy of building the 27 side-chains on the experimentally determined main chain and on the approximate model main chain, which is copied over from the parent (PDB code 2hpr; Liao & Herzberg, 1994). When building side-chains on the experimental structure main chain, the experimental side-chain conformations are used for the residues not built by the CF method. When building side-chains on the CASP1 model, side-chain conformations built by us at CASP1 are used for residues not built by the CF method.

Since this is a test of the search method, we ensure that a conformation close to the correct one is included in the sample space. For each of the 27 residues, we generate all possible side-chain conformations and select different numbers of conformations per residue based on their score in such a way that, as far as possible, at least one conformation is within $30°$ of the experimental χ angle. For two residues, the rotamer library does not provide adequate sampling. Thus the maximum accuracy that we can achieve in terms of the fraction of χ_1 angles correctly built is 25/27. To achieve this level of accuracy, 19 of the 27 residue positions must be sampled with the two top-scoring side-chain conformations, seven positions are sampled with three side-chain conformations, and one position is sampled with four conformations. This is the equivalent of systematically exploring $4^1 \times 3^7 \times 2^{19} \simeq 5 \times 10^9$ possibilities.

In another experiment, we include the exact experimental structure rotamers in the sample space for all 27 residues, replacing the closest rotamer library value to the experimental conformation, so that 100% accuracy is possible. Table 1 summarises the results of the side-chain construction for the 27 residues that were built using the CF method in the histidine-containing phosphocarrier protein (hpr).

Considering that in the original model all of these side-chains were incorrect, there is substantial improvement through using the CF method.

Table 1. Results of the simultaneous construction of 27 "difficult" residues using the clique finding (CF) method for the histidine-containing phosphocarrier protein (hpr).

	$\chi_1 < 30°$ (%)	all $\chi < 30°$ (%)
CASP1 original model	0.0	0.0
Correct main chain with only library rotamers	70.4	57.9
Correct main chain including correct rotamers	74.1	68.4.
Model main chain with only library rotamers	59.3	50.9
Model main chain including correct rotamers	59.3	57.9

All 27 side-chains had χ_1 conformations that deviated by more than $30°$ from the experimental structure in the original model built by us for CASP1. Side-chains were built using a rotamer library (Samudrala & Moult, 1997 unpublished) and with that library supplemented by the correct angles. When building on the experimental main chain, the environment was completed with the remaining experimental side-chains. When building on the model main chain, CASP1 model side-chains were used to complete the environment. The probable causes of the remaining incorrect conformations are described in Table 2.

Still, even when the exact experimental structure rotamer is included in the sample space, we are unable to build the conformations of 7/27 χ_1 angles and 18/57 all χ angles. In Table 2, we analyze the possible causes of these errors.

Thirteen of the χ rotamers incorrectly built have at least one atom with a temperature factor of more than 30.0 Å^2, raising the possibility of experimental errors. In 12 cases, the side-chains are involved in intermolecular crystallographic contacts of less than 4.0 Å. In nine cases, atoms involved in the rotamers are close to water molecules or the sulfate ion in the experimental structure (which are not taken into account by our discriminatory function in a direct manner). All the rotamers built inaccurately may be affected by one or more of these factors. This is not to suggest that the experimental structure is incorrect or that the discriminatory function is not failing, but that it is difficult to assess what the cause of failure is. Most incorrectly built side-chains are very exposed to solvent.

Mixing and matching between different parent homolog structures

We found after CASP1, for one of the targets, cellular retinoic acid binding protein I (crabpi), that certain regions in the closest homolog (muscle fatty acid binding protein; PDB code 2hmb) did not match the experimental structure as well the next-to-closest homolog (cellular retinol binding protein II; PDB code 1opa-A). The C^{α} RMSD between 2hmb, the closest homolog, and the experimental structure is 2.03 Å for the 130 residues that are superimposable. The C^{α} RMSD between 1opa-A, the next-to-closest homolog, and the experimental structure is 1.87 Å for 130 residues. The C^{α} RMSD between the final model generated by us at CASP1 (which involved subjectively mixing and matching

Table 2. Analysis of χ angles that were incorrectly built for the histidine-containing phosphocarrier protein (hpr) using the clique-finding (CF) algorithm

χ angle	Residue	Largest B ($\mathring{A}^2$)	Number of xtal contacts	Observation
χ_2	I7	24.3	1	Residue on surface of protein
χ_1	L14	20.5	3	Deviation of 33°
χ_2	L14	22.7	3	Intermolecular contacts
χ_1	S30	32.6	4	High B; intermolecular contacts; 2 H_2O molecules within 3.6 $\mathring{A}$
χ_2	I36	40.0	0	High B; residue on surface of protein
χ_2	N38	42.2	6	Intermolecular contacts
χ_2	E39	35.0	6	Intermolecular contacts
χ_3	E39	52.2	6	High B
χ_1	I47	40.5	3	High B; SO_4 ion within 3.8 $\mathring{A}$
χ_2	I47	40.5	3	High B; SO_4 ion within 4.0 $\mathring{A}$
χ_3	M48	69.1	5	High B; 6.2 $\mathring{A}$ to SO_4 ion
χ_1	D66	41.0	0	High B
χ_2	D66	45.6	0	High B; 2 H_2O molecules within 3.8 $\mathring{A}$
χ_1	N68	46.3	0	High B; H_2O molecule within 4.0 $\mathring{A}$
χ_3	Q72	32.9	8	Intermolecular contacts
χ_1	I87	22.8	0	4 H_2O molecules within 3.0-5.0 $\mathring{A}$
χ_2	I87	22.8	0	4 H_2O molecules within 3.0-5.0 $\mathring{A}$

Largest B is the largest temperature factor of any atom in the χ angle. Crystal contacts include atom-atom distances less than 4.0 $\mathring{A}$ to another molecule. Observation identifies factors in the environment not taken into account but that may affect side-chain conformation.

between 2hmb and 1opa-A) is 1.81 $\mathring{A}$ for the same 130 residues (we exclude regions that represent insertions in the calculation of this RMSD).

We ask the following question: given the two parent homolog structures to the crabpi sequence, how effective is the graph-theoretic, clique-finding method at mixing and matching between the structures to obtain the best possible model?

To answer this question, we first define crossover points where mixing between different parent structures can occur. We do this by performing a structural superposition between the 2hmb and 1opa-A structures. Ranges of main chain where the C^α atoms are less than 1.0 $\mathring{A}$ from each other define the crossover points. Exceptions to the 1.0 $\mathring{A}$ limit are handled in a subjective manner by visual inspection of the 2hmb and 1opa-A structures. We define seven crossover points, leading to eight mix and match regions: 1 to 20, 21 to 41, 42 to 52, 53 to 73, 74 to 98, 99 to 107, 108 to 122 and 123 to 140.

We built two initial models by copying the main chain for 130 residues from the two parent structures, 2hmb and 1opa-A. Side-chain conformations representing identities were copied from the parent structures, and all other side-chains were initially constructed on both models using the minimum perturbation (MP) method. Some of the side-chains were found to clash in each of the models. For these 14 residues, we determined the three best-scoring conformations per side-chain in the two separate initial models and all combinations of the main-chain and the side-chain possibilities were then explored using the CF method. This equates to exploring $2 \times 3^{14} \times 2^8 \simeq 2 \times 10^9$ conformations systematically.

The theoretical limit for the C^α RMSD of mixing and matching between the main chains given the designated crossover points is 1.54 $\mathring{A}$. This is determined by considering the C^α RMSDs of the confor-

mations created by mixing and matching between the two crabpi homologs for all the $2^8 = 256$ possibilities. Our CASP1 model of the cellular retinoic acid binding protein I (crabpi) has a C^α RMSD of 1.81 $\mathring{A}$ relative to the experimental structure for the 130 non-insertion positions. The corresponding C^α RMSD of the conformation built by mixing and matching between the two templates using the CF method is 1.66 $\mathring{A}$, with six out of eight segments correctly selected. The two correct segments not selected are affected by side-chain clashes.

Building loops in an interconnected manner

We apply the CF method to a classic problem in building main-chain regions, that of determining the conformation of antibody complementarity determining regions (CDRs). In one experiment, we built four CDRs on the Fv fragment of the D1.3 antibody (PDB code 1vfa; Bhat *et al*., 1994) simultaneously, sampling only the single best-scoring side-chain conformation per residue position (see Table 3 for details about the CDRs built). In another experiment, we built two of the CDRs, H3 and L3, simultaneously, sampling the two best-scoring side-chains per residue position for all residues except the proline residue in L3 (a total of 15). In the former case, the number of possible choices available is the product of the number of main chains generated for all four CDRs using the database method. In the latter case, the number of possible choices available is the product of the number of main chains generated using the database method for the H3 and L3 CDRs times 2^{15}. In both these cases, the environment of the experimental structure was used to build the CDRs. The database search found 168, 216, 176 and 166 main-chain conformations for the H2, H3, L2 and L3 CDRs (Table 4). This is the equivalent of systemati-

Table 3. Details of the four complementarity determining regions (CDRs) built in the D1.3 antibody (PDB code 1vfa) using the clique-finding (CF) method

CDR	Residue range	Number of residues	Sequence
H2	158–166	9	MIWGDGNTD
H3	205–212	8	RERDYRLD
L2	47–55	9	LVYYTTTLA
L3	90–97	8	HFWSTPRT

cally exploring $168 \times 216 \times 176 \times 166 \simeq 10^9$ conformations when building the four CDRs simultaneously with only one side-chain per residue and $216 \times 2^8 \times 166 \times 2^7 \simeq 10^9$ conformations when building the H3 and L3 CDRs with two side-chains per residue.

Table 4 shows the global RMSDs for the four CDRs built simultaneously with only a single side-chain conformation sampled per residue, and the RMSD values for the H3 and L3 CDRs built simultaneously sampling two side-chain conformations per residue. The main-chain conformations of the H3 and L3 loops selected in the two different experiments are identical. The largest C^α RMSD of any CDR is 1.33 Å and the largest main-chain RMSD is 1.42 Å.

All-atom RMSD values range from 1.94 Å to 2.70 Å for any single CDR when the four are built simultaneously and are slightly higher when building only the H3 and L3 CDRs sampling two side-chain conformations per residue. The C^α and all-atom RMSD values for the set of four CDRs (34 residues) are 1.10 Å and 2.46 Å (see Figure 4). The best all-atom RMSD that could be obtained, given our rotamer library approximation, and using the database selected main-chain conformations, for all the four CDRs, is 1.71 Å.

Although a few of the side-chains do not have correct rotamer choices, we consider this a moderately satisfactory result: good main-chain conformations are selected in all cases, in a highly combinational manner. Another experiment, using the experimental main chain, selected the correct
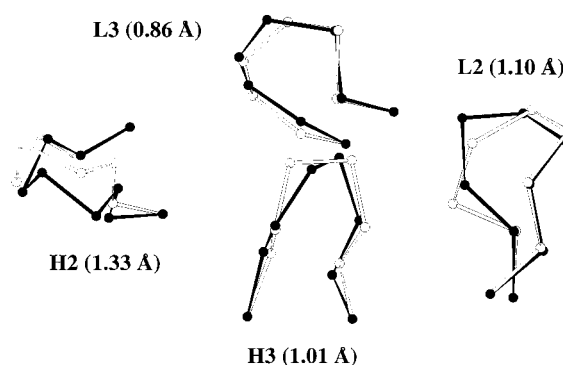
Figure 4. Comparison of conformations built simultaneously using the clique-finding (CF) method (white) to the experimental structure (black) for four complementarity determining regions (CDRs) in the D1.3 antibody. Shown are C^α traces of the four CDRs, H2 (residues 158 to 166), H3 (205 to 212), L2 (47 to 55) and L3 (90 to 97), together with the individual C^α RMSDs. The overall C^α RMSD is 1.10 Å for all the 34 residues, relative to the experimental structure. The C^α RMSDs do not include the root residues and are based on a global superposition.

side-chain rotamers for all but one very flexible arginine residue. Thus the higher all-atom RMSD values are a result of the approximate main chain.

The data base used to search for main-chain conformations contained several antibody conformations but did not include D1.3 antibody structures. The sources of the conformations selected by the CF method for each of the four loops in the two different situations are given in Table 5. A loop from another antibody is selected for only the L3 CDR.

Bona fide **prediction of comparative modeling targets**

The predictive ability of the graph-theoretic clique finding method was rigorously tested at the second meeting on the Critical Assessment of pro-

Table 4. Results of building multiple complementarity determining regions (CDRs) simultaneously in the D1.3 antibody structure using the clique-finding (CF) method

CDR	Number of main-chain conformations	C^α RMSD range (Å)	C^α RMSD (Å)	Main-chain RMSD (Å)	All-atom RMSD (Å)
A *One side-chain per residue, four CDRs*					
H2	168	0.40–6.23	1.33	1.42	1.94
H3	216	0.59–5.43	1.01	1.20	2.43
L2	176	0.66–5.28	1.10	1.54	2.67
L3	166	0.70–5.24	0.86	1.13	2.70
B. *Two side-chains per residue, two CDRs*					
H3	216	0.59–5.43	1.01	1.20	2.65
L3	166	0.70–5.24	0.86	1.13	2.78

The number of main-chain conformations sampled, the C^α RMSD range of these sampled conformations, and the C^α, main-chain (N, C^α, C, O), and all-atom RMSDs of the conformation selected are shown. The results are given for two experiments, where four CDRs (H2, H3, L2 and L3) were built simultaneously with one side-chain conformation per residue and where two CDRs (H3 and L3) were built simultaneously with two side-chain conformations per residue (see Figure 4). All RMSDs are based on a global superposition of the complete proteins.

Table 5. Sources for the main chain conformations selected by the CF method for the four complementarity determining regions (CDRs). A loop from another immunoglobulin is selected for only one CDR

CDR	Loop source (PDB code and name of protein)	Residue range in source	Sequence of source	Sequence of CDR
A. One side-chain per residue				
H2	1tgs Trypsinogen	143–151	NTKSSGTSY	MIWGDGNTD
H3	1npx NADH peroxidase	329–336	LAVFDYKF	RERDYRLD
L2	1aaz Glutaredoxin	36–44	IMPEKGVFD	LVYYTTTLA
L3	1rei Immunoglobulin	90–97	QYQSLPYT	HFWSTPRT
B. Two side-chains per residue				
H3	1npx NADH peroxidase	329–336	LAVFDYKF	RERDYRLD
L3	1rei Immunoglobulin	90–97	QYQSLPYT	HFWSTPRT

tein Structure Prediction methods (CASP2), where we made blind predictions for three targets for which experimental structures are now available (Samudrala & Moult, 1998). The three targets are cucumber stellacyanin (csc/target 9; 109 residues; Hart *et al.*, 1996) from the *Cucumis sativus*, ubiquitin conjugating enzyme (ubc9/target 24; 158 residues; Tong & Sixma, unpublished) from *Mus musculus*, and endoglucanase I (egi/target 28; 371 residues; Kleywegt *et al.*, unpublished) from *Trichoderma reesei*. For each target, an initial model was first constructed from the parent homolog structure(s) as described in Methods. Further details regarding the alignment and construction of the initial models are given by Samudrala & Moult (1998).

Bona fide prediction: side-chains

For each protein, between 15 and 18 side-chains were built using the CF method on portions of the main chain that were copied over from the parent structure. Up to six side-chain conformations were explored per residue. The remaining side-chains were built using the MP method. Table 6 summarises the results of side-chain construction. The percentage of correct χ_1 angles is given between the model and experimental structure for the relevant side-chains. For comparison purposes, the percentage of correct χ_1 angles had they been constructed using the minimum perturbation (MP) method is also given. For two targets, egi/t28 and csc/t9, the percentage of χ_1 angles built accurately increases

significantly when the CF method is used. In one case, ubc/t24, the percentage remains the same.

Table 7 shows an analysis of side-chains that had an error in the χ_1 angles of more than 30°. Out of the 15 such side-chains, nine of the errors are associated with the presence of high (>30.0 Å^2) temperature factors in the side-chain atoms, a large number of intermolecular crystallographic contacts, or a main-chain shift in the residue C$^\alpha$ (>1.0 Å) position in the model relative to the experimental structure. In four other cases, the longer-range approximate environment of the model makes the experimental conformation unlikely to be selected. In the two remaining cases, the problem appears to be due to the failure of the discriminatory function.

Bona fide prediction: main chains

A total of 22 main-chain regions, with lengths ranging from two to 14 residues, were built using the CF methods in the three CASP2 targets. Eighteen corresponded to insertions or deletions. Some of these main-chain regions were built in an interconnected manner (i.e. two or three main-chain regions were built simultaneously), and in all cases, some of the side-chains in the environment of the main-chain region being built were also varied. Main-chain conformations for 19 main-chain regions were sampled using the database search method described above. Conformations for three of the main-chain regions were sampled using the simple combinatorial main-

Table 6. Analysis of side-chain residues that were built using the clique-finding (CF) method for CASP2

Name of target	Number of side-chains	Number of conformations	Built CF % $\chi_1 < 30°$	Built MP % $\chi_1 < 30°$
egi/t28	18	$6^3 \times 4^3 \times 3^7 \times 2^5 \simeq 10^9$	61.1	50.0
ubc/t24	18	$6^2 \times 5^2 \times 4^2 \times 3^7 \times 2^5 \simeq 10^9$	66.7	66.7
csc/t9	15	$6^4 \times 5^2 \times 3^9 \simeq 6 \times 10^8$	86.7	66.7

For each target (egi/t28, ubc/t24, csc/t9), the number of side-chains, the number of conformations explored, and the percentage of χ_1 angles that deviate less than 30° is shown. For comparison, the percentage correct had those side-chains been built using the minimum perturbation (MP) method (Built MP) is also shown. All side-chains were built on main chain that was copied from the parent structure. The number of side-chain arrangements considered is the product of the number of side-chain conformations explored per residue (specified by the mantissas in column 3) of all residues whose side-chains were built using the CF method (specified by the sum of the exponents in column 3). For example, in the case of egi/t28, three residues with six conformations each, three residues with four conformations, seven residues with three conformations and two residues with five conformations each were used to construct the graph that was handed over to the CF method. For two of the models, accuracy is improved substantially by using the CF method.

Table 7. Analysis of side-chains with an error of more than 30° in the χ_1 angle built using the clique-finding method for CASP2 targets

Residue	C^α-C^α distance (Å)	Largest B (Å^2)	Number of xtal contacts	Observations
A. *egi/t28*				
W36	0.87	20.7	>10	Experimental conformation clashes with model I72 built using the minimum perturbation method; many intermolecular xtal contacts;
E73	0.90	23.6	>10	Experimental conformation O$^{\epsilon 1}$ and model G4 carbonyl at 2.7 Å (unfavorable electrostatics); G4 main-chain B is 55.8 Å^2; many intermolecular xtal contacts
Y94	0.30	16.0	0	Experimental conformation clashes with model L349 built using the minimum perturbation method; L349 has a main-chain shift of 2.62 Å in the target relative to the parent structure
V119	0.46	49.0	0	High B
Q149	0.31	47.0	0	High B
E342	2.40	73.6	0	Main-chain shift; high B
T355	0.55	16.3	0	Discriminatory functions fails
B. *ubc/t24*				
R21	4.36	19.3	>10	Region of alignment error in model; many intermolecular xtal contacts;
R25	1.60	31.0	0	Main-chain shift in target relative to parent
C51	0.22	17.5	1	Discriminatory function fails
L89	0.80	23.0	1	Shift in surrounding main chain in target relative to the parent structure
Y142	1.33	21.3	>10	Main-chain shift in target relative to parent; many intermolecular xtal contacts;
C. *csc/t9*				
T11	0.93	20.1	0	Interacts with region 14-24 (C^α RMSD 5.23)
D66	4.72	95.9	0	Main-chain shift; high B

For each residue with an error in the χ_1 angle, the distance between the C^α atoms of the corresponding residues in the experimental structure and the model, the largest temperature factor (B) of any of the atoms determining the χ_1 rotamer, intermolecular crystal contacts of less than 4.0 Å to another molecule, and a brief comment about the nature of the error is shown. In all but two cases, the side-chains predicted incorrectly have high (>30 Å^2) temperature factors in the side-chain atoms, a large number of intermolecular crystallographic contacts, or a main-chain shift in the residue C^α (>1.0 Å) position in the model relative to the experimental structure, or the correct side-chain conformation is excluded because of the longer-range approximate environment of the model.

chain grid search, with a 60° grid. Further details on the main-chain sampling are given by Samudrala & Moult (1998).

Table 8 gives the details of the main-chain region building process, including the interconnected manner in which they were built (i.e. combining main-chain and side-chain possibilities simultaneously). Table 9 shows the accuracies of the regions built to the experimental structure, along with a comment about the nature of the problem in cases where an unsuccessful prediction was made.

Out of 22 main-chain regions, ten are considered to be built successfully (C^α RMSD less than 3.0 Å, ranging from 0.60 Å for a region of main-chain variation to 2.64 Å for a ten-residue region containing a five-residue insertion). One of the more dramatic predictions is the construction of three regions in ubc/t24 (residues 37 to 46, 73 to 79 and 106 to 111) simultaneously with an overall C^α RMSD of 2.22 Å for the 23 residues (Figure 5).

Nine of the regions were built incorrectly due to inadequate main-chain sampling or because of large C^α deviations at the root residue positions. In two cases (csc/t9 residues 1 and 2, and 106 to 109), a technical error in the fitting of the main-chain conformations obtained from the grid search led to incorrect predictions. In one case (egi/t28 residues 155 to 161) the error could be due to sampling and

root positioning problems in predicting the interconnecting region (egi/t24 residues 177 to 190). Both these regions were built simultaneously but, since a low RMSD conformation was never sampled for residues 177 to 190, it is unlikely that an accurate conformation could have been predicted for the other interacting region. In no case is the discriminatory function an apparent cause for error.

Computation times

The size of problems that can be handled by the CF method (Tables 6, 8 and 10) generally depends on the number of residues being built, and the number of main-chain and side-chain possibilities considered. The finding of cliques by the Bron & Kerbosch (1973) algorithm is much faster than evaluating the weight of a clique, so that the time taken is proportional to the number of cliques. The times spent in the execution, determined using the Unix "time", for some of the experiments are given in Table 10. All times are for a multi-user Silicon Graphics (SGI) Challenge workstation with a R1000 processor. In general, finding the structural arrangement with the best score, sampling 10^9 to 10^{10} possible conformations, can be accomplished in a 24 to 48 hour period. If we assume that sampling of a conformation using conventional

Table 8. Computational details of 22 main chains that were built using the clique-finding (CF) method for CASP2

Regions built	Number of main-chain conformations	Number of side-chain conformations per main chain
A. *egi/t28*		
42–48	364	$6^2 \times 5^2 \times 3^6 \times 1^1 \simeq 6 \times 10^6$
78–81,98–103	$1013 \times 586 \simeq 6 \times 10^6$	$3^1 \times 2^8 \times 1^6 = 768$
155–161,177–190	$591 \times 96 \simeq 5 \times 10^5$	$2^{13} \times 1^8 = 8192$
214–219	468	$4^3 \times 3^6 \times 2^4 \times 2^1 \simeq 10^5$
240–244	497	$6^4 \times 3^4 \times 1^1 \simeq 10^5$
256–268	294	$3^{12} \times 1^3 \simeq 5 \times 10^5$
282–287	973	$6^5 \times 3^1 \times 1^4 \simeq 2 \times 10^4$
293–301	991	$6^2 \times 3^8 \times 5^1 \simeq 10^6$
B. *ubc/t24*		
37–46,73–79,106–111	$517 \times 451 \times 461 \simeq 10^8$	$2^5 \times 1^{23} = 32$
56–62	461	$6^6 \times 3^2 \times 1^2 \simeq 4 \times 10^5$
164–166	78	$6^4 \times 3^1 \times 1^2 = 3888$
C. *csc/t9*		
42–54,90–93	$456 \times 311 \simeq 14 \times 10^4$	$3^6 \times 1^4 = 726$
77–83,106–108	$205 \times 102 \simeq 2 \times 10^4$	$3^{10} \times 1^3 \simeq 6 \times 10^4$
1–2	1256	$6^6 \times 3^1 \simeq 14 \times 10^4$
14–24	595	$3^{13} \times 1^2 \simeq 10^8$
51–57	133	$6^2 \times 3^{10} \simeq 2 \times 10^6$

Simultaneous building of main-chain regions is shown in the first column. The number of main-chain and side-chain arrangements explored is the number of side-chain conformations times the number of side-chain conformations per main chain (see Table 6). The total number of conformations explored, considering both side-chain and main-chain conformations simultaneously, is generally of the order of 10^9 to 10^{10} conformations.

Table 9. Analysis of the predictions of 22 main-chain loop regions that were built for CASP2 using the clique-finding (CF) method

	Region built	No.	Sequence	Region type	Root RMSD (Å)	Parent RMSD (Å)	Sample range (Å)	Region RMSD (Å)	Problem
A. *egi/t28*									
	42–48	7	HDANYNS	D	2.53	2.14	1.76-7.43	3.12	Roots
*	78–81	4	AASG	D	0.60	1.15	0.68-2.49	0.77	
	96–103	8	PSSSGGYS	2	2.86	6.60	6.20-8.26	7.43	Sampling/roots
	155–161	7	GANQYNT	D	0.95	2.16	1.29–5.55	3.57	Context
	177–190	14	VQTWRNGTLNTSHQ	D	2.31	5.63	10.36-16.30	11.39	Sampling/roots
*	214–219	6	CTATAC	D	1.02	2.76	1.02-3.54	1.14	
+	240–244	5	GDTVD	D	0.77	1.15	1.78–3.70	2.23	
	256–268	13	NTDNGSPSGNLVS	7	0.46	1.85	4.07–13.58	5.36	Sampling
	282–287	6	SAQPGG	D	5.64	6.23	3.28-7.29	5.02	Sampling/roots
	293–301	9	CPSASAYGG	D	2.82	2.31	3.66-10.50	8.70	Sampling/roots
B. *ubc9/t24*									
*	37–46	10	TKNPDGTMNL	5	0.85	2.32	1.72–9.20	2.64	
+	56–62	7	KKGTPWE	0	0.57	0.53	0.60–5.45	0.60	
+	73–79	7	KDDYPSS	0	0.83	1.20	1.13–4.78	1.18	
*	106–111	6	EEDKDW	2	0.66	1.44	1.32–4.77	2.38	
	164–166	3	APS	1	4.19	6.05	4.57–6.47	6.29	Sampling/roots
C. *csc/t9*									
	1–2	2	GS	2	0.68	–	1.46–5.20	4.53	Fitting error
	14–24	11	SVPSSPNFYSQ	4	1.15	2.45	4.07–9.40	5.23	Sampling
+	42–45	4	PANA	0	0.45	1.92	1.33–2.64	1.90	
*	51–57	7	METKQSF	1	0.50	1.55	1.07–5.18	1.57	
	77–83	7	ERLDELG	1	2.71	1.45	2.62–3.82	3.56	Roots/alignment
+	90–93	4	TVGT	0	0.43	0.82	0.66–2.41	0.83	
	106–108	3	VAA	2	0.67	0.46	3.07–6.90	5.49	Fitting error

All RMSDs shown are on C^α atoms and are based on a global superposition of the structures compared. The range of residues, the number of residues, the sequence, the region type in parentheses (a number greater than 0 indicates there was an insertion of that many residues, a D signifies a deletion, and a zero signifies a region that is neither an insertion nor a deletion but was built because we thought the main-chain conformation would differ from the parent), the C^α RMSD of the two root residues, the C^α RMSD for equivalent residues ('-' if there was no equivalent residue) between the parent structure and the target experimental structure, the range of C^α RMSDs that were sampled, the C^α RMSD of the built region (not including the roots) between the model and the target experimental structure, and a brief comment about the nature of the problem in building the region accurately (if there was one). *Bona fide* successful predictions where copying the parent would not have sufficed are indicated by * and cases where the CF method works well (even though copying the main chain from the parent would have sufficed) are indicated by +.

106-111 (2.64 Å)   **37-46 (2.38 Å)**

E106

T37

K73

S79

W111   L46

73-79 (0.60 Å)
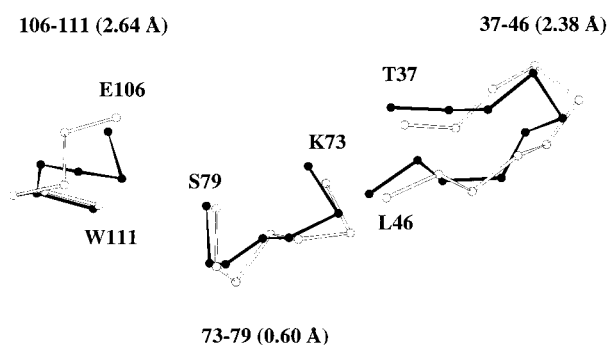
Figure 5. Comparison of conformation predicted (white) using the clique-finding (CF) method to the experimental structure (black) for three loop regions that were built simultaneously in the CASP2 target ubiquitin enzyme (ubc9). Individual loop C^α RMSDs, relative to the experimental structure, are shown. The overall C^α RMSD is 2.22 Å for all the 23 residues. The RMSDs do not include the root residues and are based on a global superposition.

methods takes just as long as finding a clique, then the number of conformations that can be sampled in the same amount of time, using the same scoring function, decreases by a factor of at least 10.

Discussion

General effectiveness of the algorithm

The graph-theoretic clique-finding method is a member of the class of algorithms for protein structure prediction that are based on a partial enumeration and evaluation of the possible structures. There are three interlocked components to such a procedure: adequate sampling of the conformations of substructures, filtering out combinations of substructures that are clearly non-viable, and scoring complete structures for fitness. Performance is affected by all these factors: a poor quality scoring function will make it difficult to select native-like conformations, and poor sampling of component conformations will exclude the native structure from the solution sets. In looking at the results, we try to assess which factors limit performance. Rather surprisingly, the discriminatory function we have used (Samudrala &

Moult, 1998), although far from perfect, is rarely the obvious cause of failures. The most severe limitation encountered in the comparative modeling experiments is sampling. Limits on sampling are dictated by the maximum size of graph that can be handled by the clique finding algorithm, and by the time taken to score a clique. In general, we find it practical to evaluate up to about 10^{10} conformations. Even though still limiting, evaluating such a large number of conformations is a major strength of the algorithm, and is possible because the scores of the nodes and edges of a graph are pre-compiled. The ability to select which substructures will be allowed to vary in setting up the graph provides a powerful means of reducing the sampling necessary: any regions of main chain or side-chains for which the conformation is reasonably certain can be entered into the graph with only one node per residue, allowing sampling to be focused on those regions of uncertain conformation. Filtering out edges between incompatible residue conformations substantially reduces the number of cliques resulting from a given level of sampling. In spite of these advantages, in most of the tests we have restricted sampling further than we should have liked to keep the graph manageable. As discussed below, limitations on performance in each of the experiments can be traced to these restrictions.

Building side-chain conformations

The ability of the CF method to build side-chains has been assessed in three ways. First, building side-chains in the experimental structures of the histidine-containing phosphocarrier protein (hpr). Only the 27 side-chains that were incorrectly built by us in CASP1 are considered, so that we have selected a "difficult" set of residues. Using the standard rotamer library on the model side-chain, 19 of the 27 χ_1 angles, and 33/57 of all χ angles, are correct (Table 1, row 3). Analysis of the probable cause of errors (Table 2) shows that most are associated with crystal contacts or uncertain experimental conformations. Although there certainly may also be some failures of the discriminatory function, these results are probably near the upper limit of possible accuracy.

Table 10. Computational times of the clique-finding (CF) method for the three comparative modeling experiments described

Experiment	Number of nodes	Edges per node	Number of conformations	Time (hh:mm:ss)
Side-chain building on hpr	125	59	$\simeq 5 \times 10^9$	31:02:14
Mixing and matching crabpi templates	316	128	$\simeq 2 \times 10^9$	18:01:42
Building all four CDRs simultaneously with one conformation per residue	6265	2170	$\simeq 10^9$	33:12:37
Building two CDRs with two conformations per residue	6063	1565	$\simeq 10^9$	37:29:34

The time spent in execution of the command is from the "time" command on a Silicon Graphics (SGI) Challenge R10000 workstation.

Although a useful test, modeling side-chains on the experimental main chain does not represent a realistic scenario. We know from previous studies that side-chain building accuracy in a comparative modeling situation decreases rapidly as the main chain varies (Chung & Subbiah, 1995). The same set of hpr side-chains was built in the CASP1 hpr model. As expected, accuracy falls under these conditions, because of the effect of the approximate model environment. Nevertheless, only three additional χ_1s and four additional χ angles, among all χ values are incorrect. Considering that all 27 side-chains were incorrect in the original model, this is a significant improvement. Greater accuracy would require much more extensive sampling of the environment. In the hpr experiments, the set of side-chain conformations was chosen to ensure the probability of a correct solution while keeping the computation tractable.

Side-chain construction was also tested by use of the CF method in CASP2. Overall, accuracy is improved for two out of the three proteins evaluated, and is unaltered for the remaining protein (Table 6). Again, although a few errors may be due to the discriminatory function, significantly improved accuracy would require very much more extensive sampling of the environment.

Mixing and matching

It is well established that the use of multiple parent main-chain conformation is often useful in comparative modeling (Greer, 1990). However, it is not easy to choose which parent to use for which region. We tested the usefulness of the CF method for this purpose on one of the CASP1 targets, the cellular retinoic acid binding protein I (crabpi), where two different proteins, the muscle fatty acid binding protein (PDB code 2hmb) and the cellular retinol binding protein II (PDB code 1opa-A), need to be mixed and matched to produce the optimum main chain. A close to optimum result is obtained, with limitations of side-chain sampling preventing the very best solution. Greater accuracy would require much more extensive sampling of the side-chain conformations together with the mix and match of the main-chain segments. The CF method was applied without any manual intervention, and so this test suggests it should be effective in a comparative modeling scenario where multiple template/parent structures are available for modeling.

Building main-chain loop regions

An advantage of the CF method is that it allows multiple loops to be built simultaneously, together with some side-chains in the environment, thus allowing for some of the context sensitivity that often determines the conformation. We tested this in two ways: rebuilding antibody CDRs and building loops in *bona fide* predictions for CASP2.

The results shown in Tables 4 and 5 compare favorably with the results for building the antibody complementarity determining regions (CDRs) on a D1.3 antibody structure using the most homologous canonical loops in other antibody structures (Pedersen *et al.*, 1992). The only case where the CF method selects a main-chain conformation from another antibody (PDB code 1rei) is with L3, where it finds a match of the same CDR with a similar sequence (Table 5). All selections were made on the basis of the best-scoring cliques and no homology information was included.

The antibody loop building test is demanding, but does make use of knowledge of the correct environment for the surrounding structure. For the *bona fide* prediction of the CASP2 target loops, about half of the 22 regions built have reasonable conformations. Limitations in accuracy are partly imposed by the loop main-chain sampling of the data base method (Fidelis *et al.*, 1994). Increased systematic exploration of main-chain conformations may therefore result in an improvement. More serious are the errors resulting from incorrect root positioning and errors in the general environment, requiring much more extensive sampling to correct. Nevertheless, these results do represent a substantial improvement over those of CASP1.

Tractability and complexity of clique finding

Clique finding in a graph in an NP-hard problem with a worst-case estimate of O ($3^{n/3}$), where n is the number of nodes in the graph (Moon & Moser, 1965; Bron & Kerbosch, 1973; Tarjan & Trojanowski, 1977). The big-O estimate indicates that even the best algorithm for finding all the cliques in a graph will take at least $k \times 2^{n/3}$ time, where k is some constant, in the worst case. For building the main chains and side-chains at CASP2, a typical graph had around 5000 nodes and we were able to search graphs with up to 30,000 nodes using an SGI Challenge R10000 workstation within a 24 hour period. None of the graphs we have encountered represents a worst case scenario, i.e. they do not take time of the order of 3^n, where n is the number of nodes. This is presumably due to the nature of the representation and its relation to protein structure, i.e. the number of cliques per node is not of the order of 3^n, and illustrates that big-O and NP-hard estimates, which apply in the worst case scenarios, are not necessarily relevant to particular problems.

Choice of the Bron & Kerbosch algorithm for clique-finding

There is no rigorous proof of the time for the Bron & Kerbosch algorithm in the average case scenario, but plots given by Bron & Kerbosch (1973) show that it works well in practice. In one test case, the authors generate a number of random graphs and the computing time per clique remains linear in the size of the graph. In a second test case where special graphs of size $3n$, which contain the largest number of cliques per node, are used, the

computing time is proportional to 3.14^n ms, where 3^n is the theoretical limit for these graphs (Moon & Moser, 1965; Bron & Kerbosch, 1973; Tarjan & Trojanowski, 1977).

In the case of a practical application involving graph-theoretical techniques to compare protein structures, this algorithm is reported to produce the best performance among several different clique-finding algorithms (Artymiuk *et al.*, 1995). Also, as demonstrated at CASP2, this algorithm performs well in the case of realistic homology modeling problems (Samudrala & Moult, 1998).

Advantages of this method compared to conventional search methods

There are three primary advantages of this method compared to traditional methods that search conformational space in proteins. First, the calculation of the fitness of a conformation is extremely fast. Weights of nodes and edges are effectively pre-compilations of the scores for all interactions in small substructures. The total score of a conformation is calculated by summing the scores for all nodes and edges in a clique and is very fast compared to calculating a score or energy based on the full set of interatomic distances. Second, a large number of unacceptable conformations are never evaluated for their weights and are rejected in advance; i.e. they are never found as cliques in the graphs. The extent to which this is true depends on the effective application of filters to eliminate edges before clique finding occurs. Third, the conformations represented by the cliques are found independently and do not depend on a continuous search through the conformational space. This allows the method to ''jump through'' the space without regard to energy barriers or local minima.

Limitations of the method

The foremost limitation of this method is the cost of enumerating all the cliques. Even though the worst-case big-*O* estimate does not apply in the cases we encounter, the size of problems that can be solved with current computing abilities is limited to the equivalent of exploring 10^{10} conformations of a medium-sized protein. Secondly, the discriminatory function used must be able to represent weights of nodes and edges independently of other nodes and edges (Samudrala & Moult, 1997). This prevents the use of many established functions, such as those that include the accessible surface area of atoms (Avbelj & Moult, 1995; Bowie *et al.*, 1991; Holm & Sander, 1992).

Conclusion

We consider that overall the results obtained are encouraging, and do demonstrate that the method already provides a powerful basis for tackling comparative modeling problems. In particular, we obtain improved accuracy for building side-chains,

are able to make better choices as to which parent structures are most appropriate for different regions of main chain, and the ability to handle multiple loops and parts of the environment simultaneously leads to usefully accurate loop conformations in some cases in blind tests.

The chief issue for the future is whether the limitations on sampling imposed by the present algorithm can be overcome. Improved side-chain and main-chain sampling methods that further narrow the choices for a residue conformation will help in reducing the number of nodes. Improvement in clique-finding algorithms by using approximation algorithms, and filtering based on weights of nodes and edges, will enable us to build a greater numbers of side-chains and larger main-chain regions simultaneously, compared to the size of problems we handle in this work.

References

Abagyan, R. & Totrov, M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002.

Artymiuk, P., Poirrette, A., Rice, D. & Willett, P. (1995). Comparison of protein folds and sidechain clusters using algorithms from graph theory. In *Proceedings of the CCP4 Study Weekend, SERC, Daresbury Laboratory, Daresbury*.

Avbelj, F. & Moult, J. (1995). Determination of the conformation of folding initiation sites in proteins by computer simulation. *Proteins: Struct. Funct. Genet.* **23**, 129–141.

Bhat, T., Bentley, G., Boulot, G., Green, M., Tello, D., Dallacqua, W., Souchon, H., Schwarz, F., Mariuzza, R. & Poljak, R. (1994). Bound water molecules and conformational stabilisation help mediate an antigen-antibody association. *Proc. Natl Acad. Sci. USA,* **91**, 1089–1093.

Bowie, J., Lüthy, R. & Eisenberg, D. (1991). Method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164–170.

Bron, C. & Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM,* **16**, 575–577.

Chen, R. (1989). Monte Carlo simulations for the study of hemoglobin-fragment conformations. *J. Comput. Chem.* **10**, 488–494.

Chothia, C. & Lesk, A. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Chung, S. & Subbiah, S. (1995). The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins. *Protein Sci.* **4**, 2300–2309.

Fidelis, K., Stern, P., Bacon, D. & Moult, J. (1994). Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* **7**, 953–960.

Greer, J. (1990). Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins: Struct. Funct. Genet.* **7**, 317–334.

Grindley, H., Artymiuk, P., Rice, D. & White, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.* **229**, 707–721.

Harel, D. (1992). *Algorithmics,* Garland Publishing Inc., New York.

Hart, P., Nersissian, A., Herrmann, R., Nalbandyan, R., Valentine, J. & Eisenberg, D. (1996). A missing link in cupredoxins: crystal structure of cucumber stellacyanin at 1.6 Å resolution. *Protein Sci.* **5**, 2175–2183.

Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 93–105.

Levitt, M. (1997). Competitive assessment of protein fold recognition and threading accuracy. *Proteins: Struct. Funct. Genet.* **Sup.1**, 92–104.

Liao, D. & Herzberg, O. (1994). Refined structures of the active ser83-cys and impaired ser46-asp histidine-containing phosphocarrier proteins. *Structure,* **2**, 1203–1216.

Martin, A., Cheetham, J. & Rees, A. (1989). Modeling antibody hypervariable loops: a combined algorithm. *Proc. Natl Acad. Sci. USA,* **86**, 9268–9272.

Martin, A., MacArthur, M. & Thornton, J. (1997). Assessment of comparative modeling in CASP2. *Proteins: Struct. Funct. Genet.* **Sup.1**, 14–28.

Moon, J. & Moser, L. (1965). On cliques in graphs. *Israel J. Math.* **3**, 23–28.

Mosimann, S., Meleshko, R. & James, M. (1995). A critical assessment of comparative molecular modeling tertiary structures in proteins. *Proteins: Struct. Funct. Genet.* **23**, 301–317.

National Research Council (1995). *Mathematical Challenges from Theoretical/Computational Chemistry,* National Academy Press, Washington, DC.

Okamoto, Y., Fukugita, M., Nakazawa, T. & Kawai, H. (1991). α-Helix folding by Monte Carlo simulated annealing in isolated C-peptide of ribonuclease A. *Protein Eng.* **4**, 639–647.

Pedersen, J. T. & Moult, J. (1997). Folding simulation with genetic algorithms and a detailed molecular description. *J. Mol. Biol.* **269**, 240–259.

Pedersen, J., Searle, S., Henry, A. & Rees, A. (1992). Antibody modeling: beyond homology. *Immunomethods,* **1**, 126–136.

Pieper, U., Kapadia, G., Zhu, P., Peterkofsky, A. & Herzberg, O. (1995). Structural evidence for the evolutionary divergence of mycoplasma from Gram-positive bacteria: the histidine-containing phosphocarrier protein. *Structure,* **3**, 781–790.

Samudrala, R. & Moult, J. (1998). An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275**, 893–914.

Samudrala, R. & Moult, J. (1998). Handling context-sensitivity in protein structures using graph theory: bona fide prediction. *Proteins: Struct. Funct. Genet.* **Sup.1**, 43–49.

Samudrala, R., Pedersen, J., Zhou, H., Luo, R., Fidelis, K. & Moult, J. (1995). Confronting the problem of interconnected structural changes in the comparative modeling of proteins. *Proteins: Struct. Funct. Genet.* **23**, 327–336.

Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C. & Sander, C. (1994). GeneQuiz: a workbench for sequence analysis. In *Proceedings of the Second International Conference on Intelligent Systems in Molecular Biology* (Altmann, R., Brutlag, D., Karp, P., Lathrop, R. & Searls, D., eds), pp. 348–353, AAAI Press, Menlo Park, CA.

Schneider, R. & Sander, C. (1996). The hssp database of protein structure-sequence alignments. *Nucl. Acids Res.* **24**, 201–205.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science,* **250**, 1121–1125.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.

Tarjan, R. & Trojanowski, A. (1977). Finding a maximum independent set. *SIAM J. Comput.* **6**, 537–546.

Unger, R. & Moult, J. (1991). An analysis of protein folding pathways. *Biochemistry,* **30**, 3816–3823.

Unger, R. & Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol.* **231**, 75–81.

Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. (1997). Numerical criteria for evaluating protein structures derived from comparative modeling. *Proteins: Struct. Funct. Genet.* **Sup.1**, 7–13.

Wilson, S. & Cui, W. (1990). Applications of simulated annealing to peptides. *Biopolymers,* **29**, 225–235.

Edited by F. Cohen