PROTEINS: Structure, Function, and Genetics, Suppl. 1:43–49 (1997)

Handling Context-Sensitivity in Protein Structures Using Graph Theory: Bona Fide Prediction

Ram Samudrala[1,2] **and John Moult**[1]*
[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*
[2]*Molecular and Cell Biology Program, University of Maryland at College Park, College Park, Maryland*

ABSTRACT **We constructed five comparative models in a blind manner for the second meeting on the Critical Assessment of protein Structure Prediction methods (CASP2). The method used is based on a novel graph-theoretic clique-finding approach, and attempts to address the problem of interconnected structural changes in the comparative modeling of protein structures. We discuss briefly how the method is used for protein structure prediction, and detail how it performs in the blind tests. We find that compared to CASP1, significant improvements in building insertions and deletions and sidechain conformations have been achieved. Proteins, Suppl. 1:43–49, 1997.**

Key words: graph theory; clique-finding; comparative modeling

INTRODUCTION

Comparative models of five structures, polyribonucleotide nucleotidyl s-transferase (pns1/-target 4; 76 residues[1]) from *E. coli*, neurocalcin delta (ncd/target 7; 193 residues) from *B. taurus*, cucumber stellacyanin (csc/target 9; 109 residues[2]) from *C. sativus*, ubiquitin conjugating enzyme (ubc9/target 24; 158 residues[3]) from *M. musculus*, and endoglucanase I (egi/target 28; 371 residues[4]) from *T. reesei*, were built. We used a graph-theoretic clique-finding (CF) method to build some sidechains and mainchain segments, including those that were thought to vary from the parent structure, after constructing an initial model by copying a subset of the atomic coordinates from the parent structure(s).[5]

METHODS

General Description of the Graph-Theoretic Clique-Finding Approach

Each possible conformation of a residue represents a node in the graph. Residues can have different mainchain and sidechain conformations. Edges are drawn between every pair of residue conformations if there are no clashes between atoms of the interacting residues and if the interaction between the two residues is covalently acceptable. A clash is said to occur if there are two nonhydrogen atoms, belonging to two different residues, with a contact of less than 2.0 Å. Contacts between pairs of atoms in the mainchain of neighboring residues are not evaluated for clashes. If the interaction weight of a sidechain with the local mainchain is extremely positive ($>$10.0), then an edge is not drawn between the nodes. If two residue positions are within one mainchain region being built, then both their conformations must be connected by a single covalently linked mainchain conformation before an edge can be drawn between them. Edges are also not drawn between different possible conformations of the same residue.

Nodes are weighted based on the strength of the interaction in pairs of atoms between the residue sidechain and the local mainchain (up to $\pm$four residues, total of nine). An edge between two nodes is weighted based on the strength of the interaction between pairs of atoms in the two residues. Nodes and edges are weighted using an all-atom distance-dependent conditional probability-based discriminatory function which provides a score related to the probability of observing a native conformation, given a set of distances between specific atom types.[6]

Once a graph representing the various possible sidechains and mainchains is constructed, we search for maximal completely connected subgraphs (cliques). Cliques the size of the target structure (which are the largest sized cliques that can be found) represent self-consistent arrangements of the individual amino acid conformations. The clique with the best weight is taken to represent the correct conformation. A full description of the method is given elsewhere.[5] In our application, clique-finding was accomplished using the Bron and Kerbosch algorithm.[7]

Search for Parent Sequences With Known Structure

Target protein sequences were obtained from the web page provided by the CASP2 organizers.[8] A basic

Contract grant sponsor: NIH; Contract grant number: GM41034.
*Correspondence to: Dr. John Moult, Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850.
E-mail: jmoult@carb.nist.gov
Received 29 May 1997; Accepted 25 August 1997

BLAST search,[9] using the program blastp and the default BLOSUM62 scoring matrix, was performed on the PDB[10] to identify parent sequences with known structures that are related to the target sequence.

In one case, pns1/t4, where no apparent homology could be detected by conventional sequence searches, distantly related sequences with known structure were found using the Hidden Markov Model (HMM) package HMMER.[11] The two highest scoring sequences returned by HMMER were considered to be distant homologs.

Sequence and Structure Alignment

Multiple sequence alignments were generated with the AMPS package.[12,13] The AMPS-derived alignments were was used to identify regions of sequence variability within the target sequence family. AMPS pairwise alignments were also used to determine the degree of sequence identity between the target sequences and the parent sequences with known structure. The default PAM250 mutation matrix and a length-independent gap penalty of 8.0 were used. In the case of target sequences with multiple parent structures, structural alignments between the parent structures were generated using the G program.[14] The structure and sequence conservation for each residue was examined to identify mainchain regions that might require rebuilding.

Visual inspection of the initial AMPS alignments revealed regions in two cases (pns1/t4 and egi/t28) where we thought the alignment was dubious. The alignment in these regions was adjusted manually.

In the case of pns1/t4, an insertion of two residues in pns1/t4 relative to 1csp in the sequence alignment was moved from residues 9–10 to residues 17–18, because the AMPS alignment placed the insertion in the middle of a beta strand. The single residue insertion at residue 21 was moved to residue 26 for the same reason (Fig. 1a).

For egi/t28, we noticed that aligning an identical stretch of four residues with sequence QNGV (residues 275–278 in egi/t28; see Figure 1e) between the target sequence and the parent sequence led to a higher degree of percentage sequence identity for entire alignment. We therefore made this correction by introducing an insertion and a deletion, as shown in Figure 1e.

Construction of an Initial Core Model

Following the sequence alignment, for each parent structure, an initial model was generated by copying atomic coordinates for the mainchain (excluding any insertions) and for the sidechains of residues that are identical in the target and parent structures. Residues that differ in type were constructed using a minimum perturbation (MP) technique implemented by the program MUTATE.[15] The MP method changes a given amino acid to the target amino acid preserving the values of equivalent χ angles between the two sidechains, where available. The other χ angles are constructed by MUTATE using an internally developed library based on residue type.[16]

Building Sidechain Conformations on the Core Model

Multiple sidechain conformations for a given residue position were generated by exploring all the possibilities in the rotamer set and selecting the most probable ones based on the interactions of a given conformation with the local mainchain. For each χ angle in a sidechain conformation, up to three rotamers were considered.[16] For each possible sidechain conformation, the interactions between the atoms of the sidechain and the local mainchain ($\pm$four residues, total of nine, where available) were evaluated using the conditional probability discriminatory function.[6] The sidechain conformations with the best score were taken to represent the most probable conformations. A detailed description of this sidechain sampling method is given elsewhere.[16]

Fifteen (in the case of csc/t9) to eighteen sidechains (in the case of ubc9/t24 and egi/t28) were identified by a preliminary environmental analysis of the initial model as positions for sampling. The environmental analysis was performed visually using interactive computer graphics, identifying sidechains with implausible packing, clashes, and unfavorable electrostatic interactions (hydrogen bonding, salt bridges) with other sidechains and/or mainchains. Between two to six different most probable sidechain conformations were considered for each such residue position. The optimal arrangement of the 15 to 18 sidechain conformations sampled was determined using the CF method in the context of the rest of the initial model.

Building the Remaining Mainchain and Sidechain Conformations

In two cases (csc/t9 and egi/t29), the initial model with the CF built sidechains was used as a template for building regions of insertions, deletions, and regions of suspected mainchain uncertainty. In one case (ubc9/t24), two initial models were created from the two different parent structures (PDB codes 1aak and 2uce). Mainchain regions selected for rebuilding were deleted from the initial models. A total of thirteen mainchain regions from the two models were mixed and matched using the CF method: A graph was constructed based on the possible mainchain and sidechain conformations and was searched for cliques representing plausible conformations of the model given the sidechain and mainchain choices per residue position. The conformation represented by the clique with the best score was used as a template for further building of mainchain regions.

For three regions (csc/t9 residues 1–2, 106–108; ubc9/t24 residues 164–166), mainchains were

(a) t4: pns1 vs. csc alignment differences (residues 1–76)

Correct: 19.6%

```
AEIEVGRVYTGKVTRIV-DFGAFVA-IG-GGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQGRIRL-SIKEA-
-------MLEGKVKWFNSEKGFGFIEVEGQDDVFV-HFSAIQGEGFK-T---LEEGQAVSFEI-VEGNRG-PQAANVT-KE
```

Final: 26.9%

```
AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQGRIRLSIKEA
--MLEGKVKWFNSEKG--FGFIEVE-GQDDVFVHFSAIQGEGFKT----LEEGQAVSFEIVEGNRGPQAANVTKEA
```

(b) t9: csc vs. 2cbp alignment differences (residues 60–83)

Correct: 32.6%

```
CNFVNSDNDVERTSPVIERLDELG
CNTPAGAKVY-TSGRDQIKL-PKG
```

Final: 33.6%

```
CNFVNSDNDVERTSPVIERLDELG
CNTPAGAKVYTSGRDQI-KLPK-G
```

(c) t24: ubc9 vs. 1aak alignment differences (residues 9–31)

Correct: 36.2%

```
-MSGIALSRLAQERKAWRKDHPFG
MSTPARKRLMRDFK-RLQQDPPAG
```

Final: 40.2%

```
MSGIALSRLAQERKAWRKDHPFG
MSTPARKRLMRDFKRLQQDPPAG
```

(d) t28: egi vs. 1cel alignment differences (residues 49–70)

Correct: 46.7%

```
CTVNGGV----NTTLCPDEATCGKNC
CYDGNTWSSTLCP---DNETCAK-NC
```

Final: 49.0%

```
CTVNGGVNTTLCPDEATCGKNC
CYDGNTWSSTLCPDNETCAKNC
```

(e) t28: egi vs. 1cel alignment correction (residues 259–302)

AMPS:

```
NGSPSGNLVSITRKYQQNGVDIPSAQPGGDTISSCPS------ASAY---GGL
SGAINRYYVQNGVTFQQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDKGGL
```

Correct:

```
NGSPSGNLVSITRKYQQNGVDIPS-AQ------PG-GDTISSCP----------SASAYGGL
-------G-AINRYYVQNGVTFQ-QPNAELGSYSGNELNDDYCTAEEAEFGGSSF-SDKGGL
```

Final:

```
NGSPSGNLVSITRKYQQNGVDIPSA-------QPGGDTISSCP---------SASAYGGL
-------SGAINRYYVQNGVTFQQPNAELGSYSGNELNDDYCTAEEAEFGGSSFSDKGGL
```

Fig. 1. Differences between the alignment used for the modeling exercise (labeled "Final") and the correct alignment based on a structural superposition (labeled "Correct") for various targets, and an example of an alignment correction. In (**a–c**), the final sequence-based alignment used to build the model is incorrect in comparison to the correct structure-based alignment. In (**d**), the mainchain region in egi/t28 residues 49–70 varies by more than 4.0 Å between the parent and the target structures, and a structural alignment in that region is not meaningful. In (**e**), an example of an hand-modified alignment that is correct is shown. The model constructed using the modified alignment (labeled "Final") is lower in C_α RMSD by more than 2.0 Å to the experimental structure compared to the model constructed using the AMPS-generated alignment, considering only mainchain regions that were copied from the parent. These regions are indicated by a thick black line for part of the correct and final alignments in (e).

TABLE I. Analysis of Sidechain Residues That Were Built Using the Clique-Finding (CF) Method*

Name of target	All MC Built SC	All MC Copied SC	Built MC Built SC	Built MC Copied SC	Copied MC Built SC	Copied MC Copied SC
egi/t28	46.5% (71)	52.3% (65)	49.0% (53)	53.2% (47)	38.9% (18)	50.0% (18)
ubc9/t24	45.2% (43)	46.0% (37)	56.0% (25)	63.2% (19)	33.3% (18)	33.3% (18)
csc/t9	47.4% (38)	40.0% (28)	69.6% (23)	46.2% (13)	13.3% (15)	33.3% (15)

*For each target (egi/t28, ubc9/t24, csc/t9), the percentage of χ_1 angles that deviate more than 30° for sidechains built using the CF method (labeled "Built SC") is shown. For comparison, the percentage error that would have resulted had those sidechains been built using the minimum perturbation (MP) method (labeled "Copied SC") is shown. The second and third columns (labeled "All MC") make this comparison for all sidechains built on any mainchain region, built or copied; the fourth and fifth columns make this comparison for sidechains that were built on mainchain regions not copied from a parent structure (labeled "Built MC"); and the last two columns make this comparison for sidechains that were built on mainchain regions that were copied from a parent structure (labeled "Copied MC"). Numbers in parentheses show the total number of χ_1 angles that were considered for the percentage error calculation.

sampled using a simple combinatorial mainchain grid search, with a 60° grid. All other mainchain regions were built by searching a database of mainchain regions by using distant constraints from the parent structure.[17] The matching main chain regions were positioned in the model structure using the method of Martin et al.[18]

Once the rebuilt mainchain regions were appropriately sampled, sidechain conformations within the mainchain and 2–10 sidechain conformations that were believed to be in contact with the segment being built were also sampled using the methods described in the previous section. In five cases, multiple regions of insertions and deletions were built simultaneously. The optimal arrangement of the possible sidechains and mainchains was determined using the CF method by selecting the conformation corresponding to the clique with the best score.

RESULTS

Sequence Alignment

The PDB codes of the parent structures and the percentage identity to the corresponding target sequences as determined by the alignment used for constructing the initial models are as follows: 1csp[19] and 1mjc[20] with percentage identities of 27.2% and 23.1% to pns1/t4; 2cbp[21] with a percentage identity of 33.6% to csc/t9; 1aak[22] and 2uce[23] with percentage identities of 40.4% and 37.8% to ubc9/t24; 1cel[24] with a percentage identity of 49.0% to egi/t28; and 1rec[25] with a percentage identity of 51.3% to ncd/t7. The experimental coordinates for ncd/t7 are not available to us at this time; the accuracy of model-building for that target will be evaluated at a later date.

To judge the accuracy of the alignments, we compare the alignment generated by a structural superposition of the parent structure and the target experimental structure to the sequence alignment used in the modeling exercise (Fig. 1).

For three of the proteins (pns1/t4, csc/t9, and ubc9/t24), neither the final alignments nor the initial AMPS alignments (which are identical in the case of csc/t9 and ubc9/t24) agree with those produced by structural superposition of the target experimental structures with the respective parent structures. A comparison of the alignment differences in nonloop regions identified by the comparative modeling evaluation program[26] is shown in Figure 1a–d. Figure 1e shows an example of a hand-corrected AMPS alignment that is correct.

In the case of pns1/t4 (Fig. 1a), the alignment used for model-building is incorrect for more than 50% of the residues, even though the proteins are related (the structural alignment between the parent and target structures results in a C_α RMSD of 2.52 Å for 64/67 residue positions that are aligned). Given such an alignment error, the rest of the model-building process is doomed to failure. The result of mainchain and sidechain building for pns1/t4 is thus not discussed in detail.

In two other cases (csc/t9 and ubc9/t24; Fig. 1b,c), the AMPS-generated alignments were incorrect for one region in each structure.

The "alignment difference" in egi/t28 (Fig. 1d), residues 49–70, illustrates that structure-based alignments are not necessarily meaningful. What is identified as an alignment error by the comparative modeling evaluation program is not really an error, but rather an example of a large mainchain shift (with a C_α RMSD of 4.85 Å for the 21 residues). The structural alignment between the parent and the target experimental structures is meaningless in this region.

The alignment correction in egi/t24 (Fig. 1e) underscores the importance of visual inspection. The C_α RMSD between the model constructed using the AMPS alignment and the target experimental structure is 4.24 Å for the 292 mainchain positions that were copied from the parent. The C_α RMSD between the model constructed using the hand-corrected alignment and the target experimental structure is 1.92 Å for the same number of residues. The hand-corrected alignment matches the structural one exactly for these residues.

Sidechain Building

Table I shows that in cases where the parent mainchain was copied, the percentage error in the χ_1

angles is significantly reduced in egi/t28 and csc/t9 by 11% and 20%, respectively, by building those sidechains with the CF method. In the case of ubc9/t24, the percentage error is similar regardless of the method used and the source of the mainchain. However, when we consider the columns labeled "All MC" in Table I, we see that the percentage error in the case of csc/t9 has risen (by 7%) by using the CF method. This presumably reflects the fact that the insertions in egi/t28 and ubc9/t24 were built relatively accurately, leading to better predictions with the sidechains, whereas the insertions in csc/t9 had large errors (C_α RMSDs greater than 3.0 Å) leading to inaccurate sidechain predictions. These observations are supported by the data under the columns labeled "Built MC" in Table I.

There were 15 sidechains built on mainchains copied from the parent experimental structure using the CF method that deviated by more than 30° in the χ_1 angles relative to the target experimental structure for the three proteins. Eight of the errors are associated with the presence of high (>30.0 Å^2) temperature factors in the sidechain atoms or a mainchain shift in the residue C_α (>1.0 Å) position in the model relative to the experimental structure. In six cases, the correct experimental conformation for those residues cannot be accommodated without clashes in the model structure because of mainchain and sidechain errors in the environment. In two cases, it appears as if the discriminatory function is unable to select the correct rotamer in the context of the model environment.[5]

Mainchain Building

Table II shows the details for the 22 mainchain regions that were selected using the CF method. There are five regions corresponding to insertions that represent accurate and *bona fide* blind predictions where simply copying the parent would not have sufficed (these rows are prefixed by an "*" in Table II). The sizes of these regions range from 4 to 10 residues (with sizes of the insertions ranging from one to five residues) with C_α RMSDs ranging from 0.77 Å (for a four residue region involving deletion) to 2.64 Å (for a 10 residue region involving a five residue insertion).

There are another five regions where copying the parent would have generally sufficed for building these regions (rows are prefixed by a "+" in Table II), but which were built using the CF method because we thought these regions would vary. However, these cases illustrate that the CF method works well and the C_α RMSDs range from 0.60 Å to 2.23 Å.

The last column in Table II makes a brief comment about the nature of problem for each mainchain that had a C_α RMSD greater than 3.0 Å between the model and the experimental structure. Out of the 12 regions that had large C_α RMSDs, nine of them were predicted incorrectly due to either lack of adequate sampling (no conformation with a C_α RMSD less than 3.0 Å), large C_α RMSDs for the two root residues (greater than 2.0 Å), or both. In two of the cases (csc/t9 residues 1–2 and residues 106–108), a technical error in which the mainchains returned by the grid search method were fitted incorrectly to the framework led to incorrect predictions. In one case (egi/t28 residues 155–161), we sample mainchains with C_α RMSDs between 1.29 Å and 5.55 Å, have a C_α RMSD of 0.95 Å in the root positions, but the predicted region has a C_α RMSD of 3.57 Å. This error is due to the fact that this region in egi/t28 interacts with residues 177–190, which could not have been predicted accurately due to inadequate sampling. These two regions are interconnected and cannot be built separately. If the mainchain in one region cannot be sampled adequately, then the other region is likely to be predicted incorrectly. This example illustrates the importance of handling context-sensitivity when building comparative models.

DISCUSSION

Alignment

At the first meeting on the Critical Assessment of protein Structure Prediction methods (CASP1), we learned that automated sequence alignment methods are inadequate and that a visual inspection is necessary to optimize the alignment. However, at CASP1, we were lucky that all our optimizations by hand based on sequence identity proved to be correct. Here, only one such optimization in egi/t28 produced the correct alignment (Fig. 1e). The other hand corrected alignment in pns1/t24 was wrong (Fig. 1a). This particular error could be attributed to the low level of global sequence identity in the target. However, in ubc9/t24 (Fig. 1c), the structural alignment differs significantly from the sequence-based one, and visual inspection of the alignment would have yielded no clues about the shift in the helix in that region. In fact, in all cases the correct structure-based alignments have a lower percentage sequence identity than the sequence alignments that were used (Fig. 1). This indicates that a sequence alignment that relies on percentage identity or homology alone cannot effectively produce the correct alignment, and that visual inspection and hand-optimization of alignments has its limits. As we suggested in Samudrala et al.,[27] better alignment methods that take structural information into account need to be developed.

Sidechains

Table I shows that for the CASP2 targets, there is a significant increase in the accuracy of sidechain construction using the CF method, compared with the MP method we relied on in CASP1, particularly for the portions of the mainchain copied from the

TABLE II. Analysis of the Predictions of 22 Mainchain Regions That Were Built 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
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
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
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 C_α RMSDs in Å and are based on a global superposition of the structures being compared. For each target (egi/t28, ubc9/t24, and csc/t9), the range of residues in the built region, the number of residues in the built region, the sequence of the region being built, the region type (a number greater than 0 indicates there was an insertion of that many residues, a "D" signifies a deletion, and a 0 signifies a region that is neither an insertion or a deletion but was built because we thought the mainchain conformation would differ from the parent), the C_α RMSD of the two root residues, the C_α RMSD for equivalent residues ("—" if there were no equivalent residues) 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 mainchain from the parent would have sufficed) are indicated by "+." The number of mainchain conformations sampled ranged from 78–1,013, with an average of ≃500. The number of sidechain conformations sampled *per mainchain conformation* ranged from 32 to 1,594,323, with an average of ≃370,000. The total number of conformations explored, considering both sidechain and mainchain conformations simultaneously, is generally in the order of 10^9–10^{10} conformations.

parents. Sidechain accuracy in all regions is still limited by the effect of errors in the mainchain.

Mainchains

Building mainchains in an interconnected manner (i.e., building multiple mainchains and sidechains in the environment simultaneously) has improved the predictability of insertions and deletions. At CASP1, none of the insertions and deletions were predicted accurately—in the case of models that we built, none of the insertions and residues flanking deletions had a C_α RMSD less than 3.0 Å.[27] At CASP2, five of the insertions and residues flanking deletions have C_α RMSDs less than 3.0 Å, and five regions that did not correspond to insertions or deletions were built by the CF method and predicted accurately with C_α RMSDs less than 2.0 Å to the experimental structure.

ACKNOWLEDGMENTS

Thanks to Jan Pedersen for help in using the AbM database method and for constructive advice, and Brett Milash, Michael Braxenthaler, and Rui Luo for valuable discussions. This work was supported in part by a Life Technologies Fellowship to Ram Samudrala and NIH grant GM41034 to John Moult. Some computations were performed using NIST computing resources.

REFERENCES

1. Bycroft, M., Hubbard, T., Proctor, M., Freund, S., Murzin, A.G. The solution structure of the S1 RNA binding domain: A member of an ancient nucleic acid-binding fold. Cell 88:235–242, 1997.
2. Hart, P., Nersissian, A., Herrmann, R., Nalbandyan, R., Valentine, J., Eisenberg, D. A missing link in cupredoxins: Crystal structure of cucumber stellacyanin at 1.6 Å resolution. Protein Sci. 5:2175–2183, 1996.

3. Hateboer, G., Perrakis, A., Bernards, R., Sixma, T. K. Crystal structure of murine/human Ubc9 provides insight into the variability of the ubiquitin-conjugation system. J. Biol. Chem. 272:21381–21387, 1997.
4. Kleywegt, G., Zou, J., Divne, C., Davies, G., Sinning, I., Stahlberg, J., Reinikainen, T., Srisodsuk, M., Teeri, T., Jones, T. The crystal structure of the catalytic core domain of endoglucanase I from *Trichoderma reesei* at 3.6 Å with related enzymes. J. Mol. Biol. 272:383–397, 1997.
5. Samudrala, R., Moult, J. A graph-theoretic clique finding approach to protein structure prediction. J. Mol. Biol., In press.
6. Samudrala, R., Moult, J. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol., In press.
7. Bron, C., Kerbosch, J. Algorithm 457: Finding all cliques of an undirected graph. Comm. ACM 16:575–577, 1973.
8. Bryant, S., Hubbard, T., Moult, J. Critical Assessment of protein Structure Prediction methods (2) web page. ⟨http://iris4.carb.nist.gov-casp2/⟩
9. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D. Basic local alignment search tool. J. Mol. Biol. 215:403–410, 1990.
10. Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T., Tsumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
11. Eddy, S., Mitchison, G., Durbin, R. Maximum discrimination hidden Markov models of sequence consensus. J. Comp. Biol. 2:9–23, 1995.
12. Barton, G.J. Protein multiple sequence alignment and flexible pattern matching. Methods Enzymol 183:403–428, 1990.
13. Barton, G.J., Sternberg, M.J.E. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. J. Mol. Biol. 198:327–337, 1987.
14. Pedersen, J.T.G. A molecular modelling program available upon request to the author, 1995.
15. Read, R. MUTATE, a program that implements the minimum perturbation method for building comparative models, 1984.
16. Samudrala, R., Moult, J. Determinants of side chain conformational preferences in protein structures, submitted.
17. Pedersen, J., Searle, S., Henry, A., Rees, A. Antibody modelling: Beyond homology. Immunomethods 1:126–136, 1992.
18. Martin, A., Cheetham, J., Rees, A. Modelling antibody hypervariable loops: A combined algorithm. Proc. Natl. Acad. Sci. USA 86:9268–9272, 1989.
19. Schindelin, H., M.A., M., Heinemann, U. Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-shock protein. Nature (London) 364:164–168, 1993.
20. Schindelin, H., Jiang, W., Inouye, M., Heinemann, U. Crystal structure of CspA, the major cold shock protein of Escherichia coli. Proc. Natl. Acad. Sci. USA 91:5119–5123, 1994.
21. Guss, J., Merritt, E., Phizackerley, R., Freeman, H.C. The structure of a phytocyanin, the basic blue protein from cucumber, refined at 1.8 A resolution. J. Mol. Biol. 262:686–705, 1996.
22. Cook, W., Jeffrey, L., Sullivan, M., Vierstra, R. Three-dimensional structure of a ubiquitin-conjugating enzyme (E2). J. Biol. Chem. 267:15116–15121, 1992.
23. Cook, W., Jeffrey, L., Xu, Y., Chau, V. Tertiary structures of class I ubiquitin-conjugating enzymes are highly conserved: Crystal structure of yeast Ubc4. Biochemistry 32:13809–13817, 1993.
24. Divne, C., Stahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J., Teeri, T., Jones, T. The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from Trichoderma reesei. Science 265:524–528, 1994.
25. Flaherty, K., Zozulya, S., Stryer, L., McKay, D. Three-dimensional structure of recoverin, a calcium sensor in vision. Cell 75:709–716, 1993.
26. Zemla, A., Venclovas, C., Fidelis, K., Moult, J. Ab initio protein structure prediction and comparative modeling evaluator. ⟨http://predictioncenter.llnl.gov/⟩
27. Samudrala, R., Pedersen, J., Zhou, H., Luo, R., Fidelis, K., Moult, J. Confronting the problem of interconnected structural changes in the comparative modeling of proteins. Proteins 23:327–336, 1995.