Homology Modeling of Proteins Using Multiple Models and Consensus Sequence Alignment

Jahnavi C. Prasad[1], Michael Silberstein[1], Carlos J. Camacho[2], and Sandor Vajda[2]

[1] Program in Bioinformatics, Boston University, Boston MA 02215
[2] Department of Biomedical Engineering, Boston University, Boston MA 02215
vajda@bu.edu

Abstract. Homology modeling predicts the three-dimensional structure of a protein (target), given its sequence, on the basis of sequence similarity to a protein of known structure (template). The main factor determining the accuracy of the model is the alignment of template and target sequences. Two methods are described to improve the reliability of this step. First, multiple alignment are produced, converted into models, and then the structure with the lowest free energy is chosen. The method performs remarkably well for targets for which a good template is available. In the second approach, the alignment is based on the consensus of five popular methods. It provides reliable prediction of the structurally conserved framework region, but the alignment length is reduced. A homology modeling tool combining the two methods is in preparation.

1 Introduction

Over the last decade the exponential growth of sequenced genes has prompted the development of several methods for the prediction of protein structures. The most successful prediction method to date is homology modeling (also known as comparative modeling), which predicts the three-dimensional structure of a protein (target), given its sequence, on the basis of sequence similarity to proteins of known structure (templates) [1,2]. The approach is based on the structural conservations of the framework regions between the members of a protein family. Since the 3D structures are more conserved in evolution than sequence, even the best sequence alignment methods frequently fail to correctly identify the regions that possess the desired level of structural similarity, and the quality of alignment remains the single most important factor determining the accuracy of the 3D model [3]. Therefore, it is of substantial interest to develop methods that can provide highly accurate sequence alignment, and possibly identify regions were the similarity is too low for building a meaningful model on the basis of the template structure [4].

In this paper we describe two approaches to reduce the uncertainty of the alignment. The first approach to dealing with this uncertainty is based on the use of multiple models [5]. A number of pairwise alignments (using simple dynamic programming with variation of parameters), is generated. This is followed

G. Benson and R. Page (Eds.): WABI 2003, LNBI 2812, pp. 389–401, 2003.

by energy based discrimination of the generated models [6,7]. The approach was tested at the CASP4 (Comparative Assessment of Structure Prediction) competition in 2000 (see http://predictioncenter.llnl.gov/casp4/). As we will show, in view of its relative simplicity the approach provides surprisingly good result for the easy targets, i.e., for targets with a good template available, but the dynamic programming is too rudimentary to obtain any good alignment for difficult targets. Thus, the free energy ranking algorithm had too choose one among such inferior models, and the method was unable to compete with approaches based on more sophisticated sequence alignment algorithms which employed evolutionary relationships between all homologues, and accounted for the known structure of the template.

The second approach involves a consensus alignment algorithm for the prediction of the framework regions that are structurally conserved between two proteins [8]. The target and template sequences are aligned by the five best algorithms currently available, and each position is assigned a confidence level (consensus strength) based on the consensus of the five methods. The regions reliable for homology modeling are predicted by applying criteria involving secondary structure and solvent exposure profile of the template, predicted secondary structure of the target, consensus confidence level, template domain boundaries and structural continuity of the predicted region with other predicted regions. The methodology was developed based on a diverse set of 79 pairs of homologues with an average sequence identity of 18.5%, and was validated using a different set of 48 target-template pairs. On the average, our method predicts structures that deviate from the native structures by about 2.5 Å, and the predictions extend to almost 80% of the regions that are structurally aligned in the FSSP database [9]. The approach was tested at the as an automatic server, participating in the CAFASP3 competition of such servers, described on the webpage http://www.cs.bgu.ac.il/∼dfischer/CAFASP3/.

2 Methods

2.1 Multiple Model Approach to Homology Modeling

The basic idea of the method is to generate a large number of alignments, construct a homology model for each, and rank the models according to their free energies. The current implementation of the procedure starts with traditional template selection using BLAST and PSI-BLAST [10]. The Domain Profile Analysis developed in Temple Smith's lab (http://bmerc-www.bu.edu/bioinformatics/profile_request.html) has also been consulted. One or (infrequently) several proteins have been selected as templates for the comparative modeling. In the second step of the algorithm, we generate multiple alignments between target and template sequences by varying the alignment parameters (gap-opening, gap-extension, and scoring matrix) for producing semiglobal alignments by standard dynamic programming. The blosum62 and gonnet matrices were used with gap opening penalty values 5, 6, 7, 8, 9, 10, 12, 14, 17, 20, 25, and gap extension penalty values 0.1, 0.2, 0.3, 0.5, 0.75, 1.0, 1.25, 1.6, 2.0, 2.5, 3, 4, 5, 7, 10. We produced only one alignment for each set of parameters using a single trace-back path in the dynamic programming matrix, thus resulting in 330 alignments for each template-target pair. Any alignment was deleted if it was a duplicate, or less than 75% of the target residues were aligned to the template, generally resulting in 80 to 150 retained alignments.

In the third step, all alignments are used for model construction via the MODELER program developed by Sali and co-workers [2,11]. The resulting models were minimized for 200 steps using the Charmm potential [12], and ranked by using an empirical free energy function [6,7]. The function combines molecular mechanics with empirical solvation/entropic terms to approximate the free energy G of the system consisting of the protein and the solvent, the latter averaged over its own degrees of freedom. The free energy is given by $\mathrm{G} = \mathrm{E}_{conf} + \mathrm{G}_{solv}$. The conformational energy E_{conf} is calculated by Version 19 of the Charmm potential, $\mathrm{E}_{conf} = \mathrm{E}_{elec} + \mathrm{E}_{int}$, where the internal (bonded) energy, E_{int}, is the sum of bond stretching, angle bending, torsional, and improper terms, $\mathrm{E}_{int} = \mathrm{E}_{bond} + \mathrm{E}_{angle} + \mathrm{E}_{dihedral} + \mathrm{E}_{improper}$[12]. The electrostatic energy, E_{elec}, is calculated using neutral side chains and the distance-dependent dielectric εg= 4r. G_{solv} is the solvation free energy, obtained by the atomic solvation parameter model of Eisenberg and McLachlan [13].

Notice that the function does not include the van der Waals energy term [6,7]. This approximation is based on the concept of van der Waals cancellation which assumes that the solute-solute and solute-solvent interfaces are equally well packed, and hence the van der Waals contacts lost between solvent and solute are balanced by new solute-solute contacts formed upon protein folding. This cancellation is promoted by a procedure called van der Waals normalization, prior to the free energy calculations. Van der Waals normalization impliesthat all conformations are minimized for a moderate number of steps, the structure with the lowest van der Waals energy is selected, and all other structures are further minimized to attain the same van der Waals energy value. The van der Waals cancellation implies that we can remove both the solute-solvent and the solute-solute van der Waals terms from the free energy function.

2.2 Consensus Alignment

In a benchmarking analysis [8], we have tested ten widely used methods and selected five of them in a hierarchical manner so that we cover a broad range of alignments. The five methods are as follows:

(1) BLAST-Pairwise: Target sequence is blasted against the template sequence to get an alignment.

(2) T99-BLAST: A PSI-BLAST [10] alignment of the target and template hits is supplied as the ‘seed alignment’ to Target99 script. It is tuned up, an HMM is built using this alignment, and target and template are aligned to it [14].

(3) HMMER-BLAST: The PSI-BLAST generated alignment of target and template hits is used to build a model. Target and template sequences are then aligned to it [15].

(4) T99-HSSP: Family alignment around the template (downloaded from the FSSP database) is used to build the initial model. The combined hits of target and template sequences are then aligned to this model and "tuned up" using the target99 script. A model is then constructed from this alignment and target and template are aligned to it.

(5) HMMER-HSSP: A model is built using the template family alignment. The combined hits of target and template are aligned to it to get a multiple sequence alignment. Another model is then constructed from this alignment and used to get a target-template pairwise alignment.

For our training set targets [8], at least one of the above methods was able to produce an alignment resulting in a low RMSD model, but it was not possible to predict which of the methods would perform well for a particular problem. Therefore we developed a selection procedure for determining whether or not two aligned residues can be included for accurate homology modeling. This process also invokes structural considerations, e.g., secondary structure and solvent exposure information, on the template, and to a lesser degree secondary structure prediction of the target. A flow chart of the overall algorithm is depicted in Fig. 1.

As shown in Fig. 2, the consensus strength (CS) is a measure of the agreement between the five alignment methods calculated for all target-template residue pairs. If all three methods T99-BLAST, HMMER-BLAST and BLAST-PW align target residue X_{tar} to template residue X_{tem} then CS = 9. If only two of the above three methods concur in aligning X_{tar} to X_{tem}, then CS = 6 for the $X_{tar}X_{tem}$ pair. If any three out of the five methods concur then CS = 7, and concurrence of only any 2 out of 5 means CS = 5. Consensus strengths between 4 and 0 are assigned to the residue pairs aligned by only one method, the methods respectively being T99-BLAST, HMMER-BLAST, BLAST-PW, T99-HSSP and HMMER-HSSP. Obviously, the methods will differ in certain regions. Consensus among certain alignment methods for a certain region may be incompatible with consensus among other methods for a different region. In such a case, the region with higher consensus strength receives priority.

Since consensus strength does not eliminate all the regions of potential structural dissimilarity, the following selection method is applied. If CS is 9 and template residue X_{tem} is buried, $X_{tar}X_{tem}$ pair is selected. If there are no pairs with a CS of 9, then pairs with a CS of 7 that are buried are selected. This forms the core of the selection. Subsequently the selected regions are extended towards the N and C termini as long as neighboring residues have CS of 7, or until a misaligned GLY residue occurs. Moreover, alignment regions where the template has long helices and sheets are selected subject to their CS, solvent exposure and percentage match of the predicted secondary structure of the target (using JNET of Cuff and Barton [16]) with the actual secondary structure of the corresponding template region. Other structural criteria such as single beta-sheet pairing, taut regions in template with limited potential for conformational variation are also used for selection. Regions corresponding to potentially loose termini, and

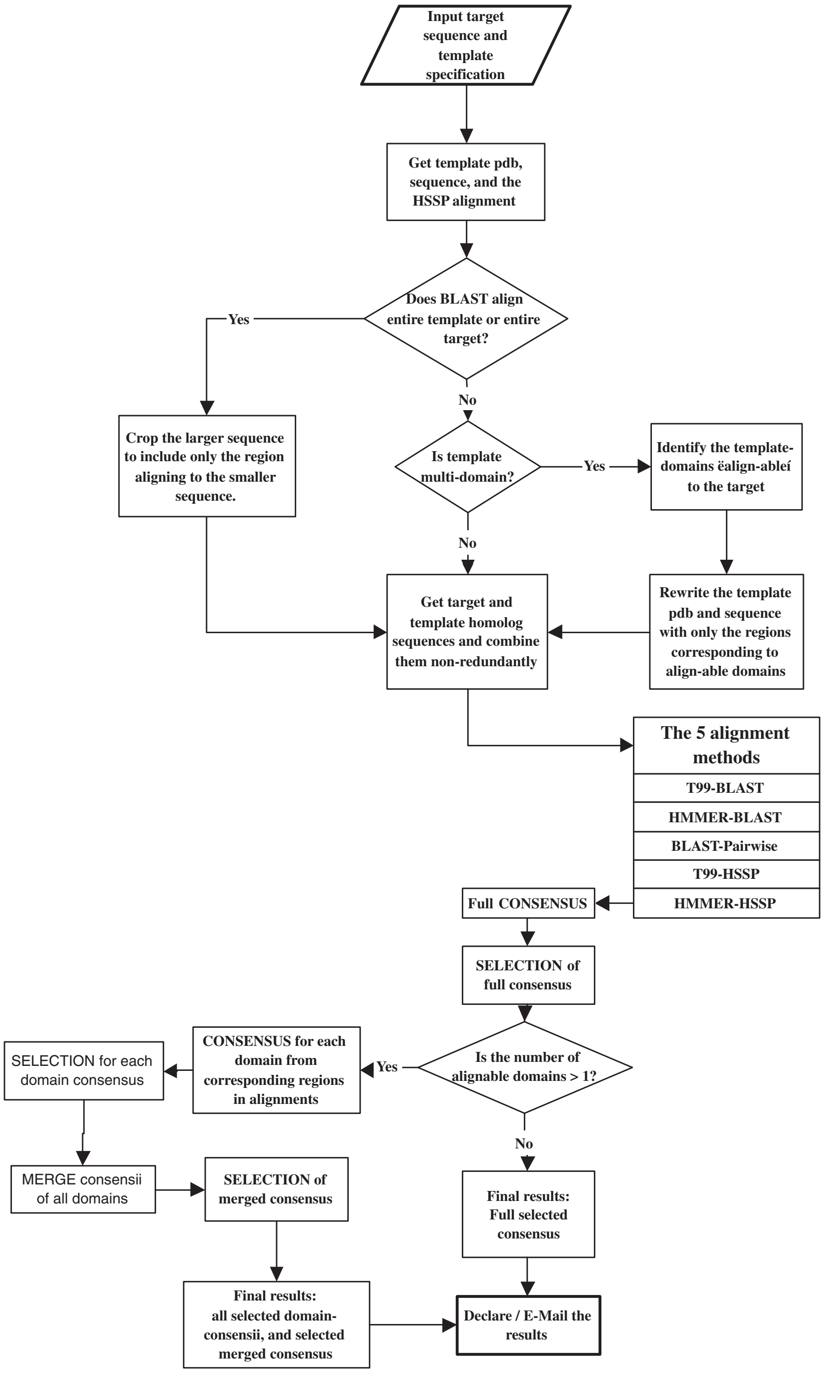
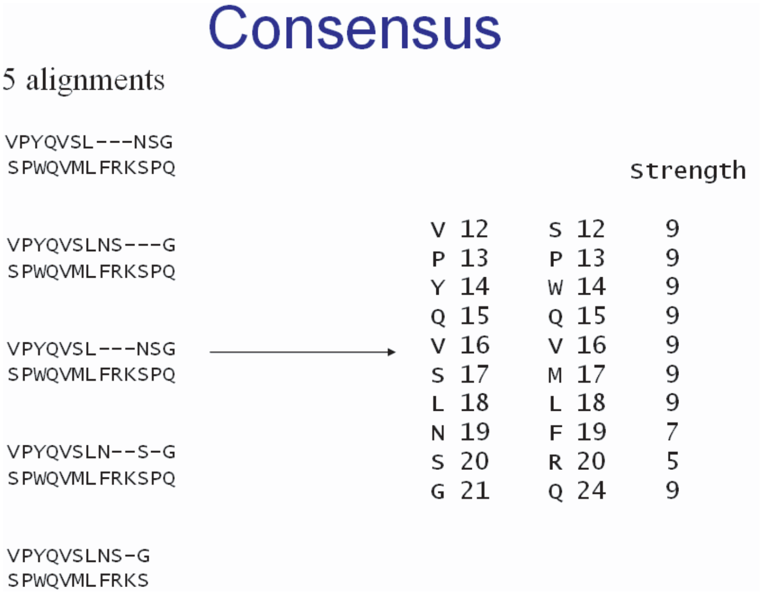

Input target sequence and template specification
Get template pdb, sequence, and the HSSP alignment
Does BLAST align entire template or entire target?
Yes
No
Crop the larger sequence to include only the region aligning to the smaller sequence.
Is template multi-domain?
Yes
Identify the template-domains ëalign-ableí to the target
No
Get target and template homolog sequences and combine them non-redundantly
Rewrite the template pdb and sequence with only the regions corresponding to align-able domains
The 5 alignment methods
T99-BLAST
HMMER-BLAST
BLAST-Pairwise
T99-HSSP
HMMER-HSSP
Full CONSENSUS
SELECTION of full consensus
Is the number of alignable domains > 1?
Yes
CONSENSUS for each domain from corresponding regions in alignments
SELECTION for each domain consensus
No
MERGE consensii of all domains
SELECTION of merged consensus
Final results: Full selected consensus
Final results: all selected domain-consensii, and selected merged consensus
Declare / E-Mail the results

Fig. 1. A summary flow chart of the method, as implemented in the Consensus server



Consensus

5 alignments

```
VPYQVSL---NSG
SPWQVMLFRKSPQ

VPYQVSLNS---G
SPWQVMLFRKSPQ

VPYQVSL---NSG
SPWQVMLFRKSPQ

VPYQVSLN--S-G
SPWQVMLFRKSPQ

VPYQVSLNS-G
SPWQVMLFRKS
```

		Strength
V 12	S 12	9
P 13	P 13	9
Y 14	W 14	9
Q 15	Q 15	9
V 16	V 16	9
S 17	M 17	9
L 18	L 18	9
N 19	F 19	7
S 20	R 20	5
G 21	Q 24	9

Fig. 2. Consensus of the output from five alignment methods

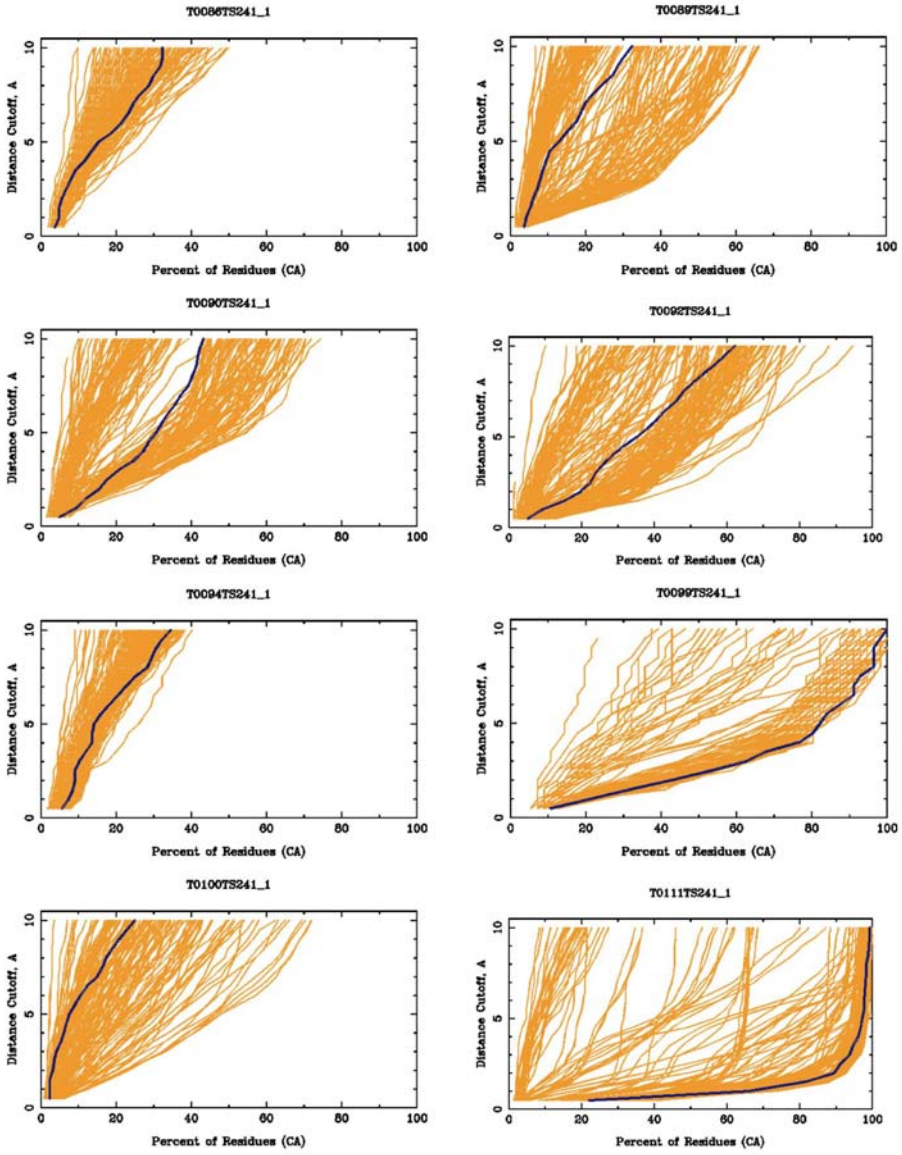
uncertain regions with high number of gaps are deselected. Consensus strength is always considered in all selection and deselection steps.

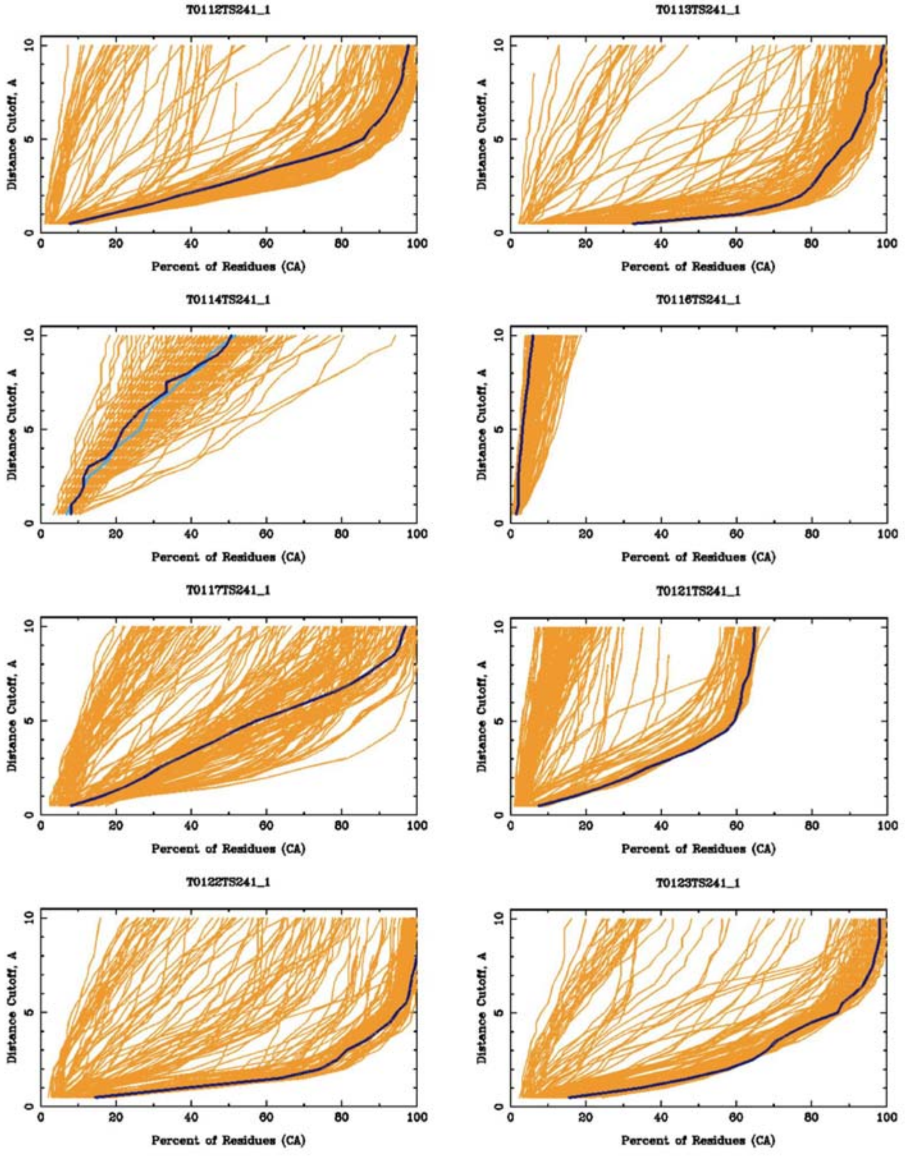
The output of the algorithm consists of the full selected consensus. The selected regions of target are then predicted by simply following the backbone of corresponding regions of the template. If the template is multi-domain, target regions corresponding to each domain are predicted separately. In such cases, full predictions were submitted as first models and the domains as subsequent models. The entire method is automatic and available as a server at the webpage http://structure.bu.edu/.

3 Results and Discussion

3.1 Multiple Model Approach

Out of 43 targets in the fourth Critical Assessment of Techniques for Protein Structure Prediction (CASP4), we submitted models for 20 targets, eight of them not in the comparative modeling category. Figure 3 shows the prediction results in terms of Global Distance Test (GDT). The GDT algorithm identifies in the prediction the sets of residues deviating from the target by not more than specified C_a distance cutoff using many different superpositions. Each residue in a prediction is assigned to the largest set of the residues (not necessary continuous), deviating from the target by no more than a specified distance cutoff on the x axis of the plot. The figures show submissions by all groups, participating


T0086TS241_1
T0089TS241_1
T0090TS241_1
T0092TS241_1
T0094TS241_1
T0099TS241_1
T0100TS241_1
T0111TS241_1
Distance Cutoff, A
Percent of Residues (CA)
0
5
10
0
20
40
60
80
100


T0112TS241_1
T0113TS241_1
T0114TS241_1
T0116TS241_1
T0117TS241_1
T0121TS241_1
T0122TS241_1
T0123TS241_1
Distance Cutoff, A
0
5
10
Percent of Residues (CA)
0
20
40
60
80
100

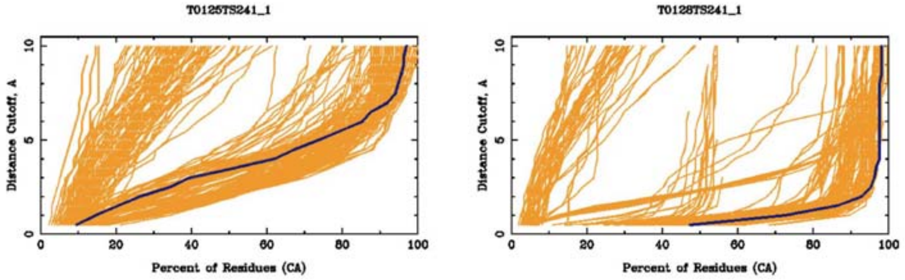


Fig. 3. Results of homology modelling at CASP4 in terms of the Global Distance Test (GDT). The GDT algorithm identifies in the prediction the set of residues deviating from the target by not more than a specified distance cutoff. The figures show the percent of such residues vs. the distance cutoff. All predictions for each target are shown, our prediction is indicated by the dark blue curve. For Target 114 we also submitted a second model, shown in light blue.

in CASP4, our submission being shown in dark blue. For target 114 there were two submissions from us, shown in dark and light blue. Results for targets 88, 93 and 119 have not been published, and hence are not shown here.

According to the above Figure 3, the multiple model approach provides good result for the relatively easy targets, i.e., for the targets were at least one groups predicted over 80 % of the residues below 5 Å distance cutoff (targets 99, 111, 112, 113, 117, 122, 123, 125, and 128). Indeed, with the exception of targets 117 and 125, for these easy cases our predictions are among the bests, but they are above average even for targets 117 and 125. Our prediction is also very good for target 121, a more difficult case. In addition, we had one of the largest prediction lengths for these targets. However, for most of the really difficult targets for which no group was able to predict at least 80% of residues below 5 Å distance cutoff (targets 86, 89, 90, 92, 94, 100, 114, 116, and 117) our method yields average or below average results. The main reason is that for these targets any available template had low sequence similarity, and the simple alignment by dynamic programming produced poor results for any of the parameters. Thus, the free energy ranking algorithm had too choose one among inferior models, and the method was unable to compete with approaches based on more sophisticated sequence alignment algorithms, utilizing the evolutionary relationships between all available homologues, and accounting for the known structure of the template.

3.2 Homology Modeling Using Consensus Alignment

For the validation of the method, 48 target-template pairs were selected from the FSSP database. The selection was governed by following criteria: (a) each target belongs to a different family in FSSP; (b) the length of the structural alignment must be greater than 100 residues; (c) the percent identity, defined as the number of identical residues divided by the length of shorter sequence, must be less than 35%. The percent identity is between 2 and 29%, averaging 16.8%. The full list of targets and templates is available at http://structure.bu.edu/consdoc.html.

Figure 4 compares, in terms of the average RMSD of the aligned residues, the structural superposition alignment from the DALI database, the homology models obtained from the five methods used in our analysis, and the consensus based method. Also shown are the standard deviations for all the models. We find that, for the training set, the Consensus algorithm not only provides the lowest RMSD but also has the smallest standard deviation. It should be emphasized that in these comparison the individual alignment methods have benefited from an automatic splitting and cropping of domains which do not align with the target sequence [8], otherwise the average RMSD for, say, T99-BLAST would be higher than 6 Å. The main result to consider here is that the RMSD has been brought down to about 2.2 Å (lower than the average DALI RMSD) while keeping the alignment length to about 75% of the DALI alignment.

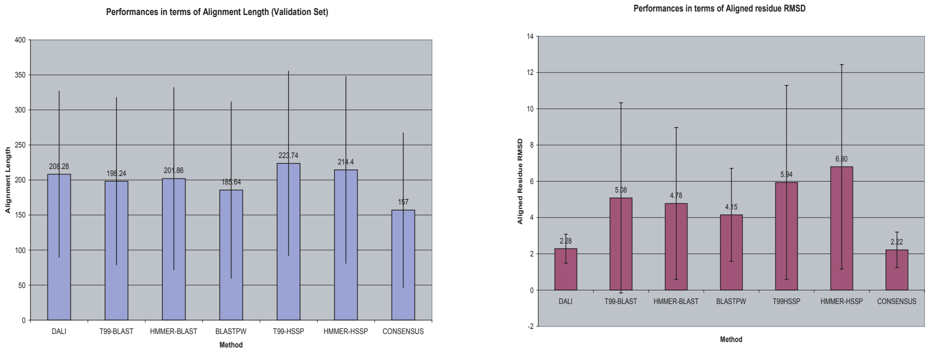
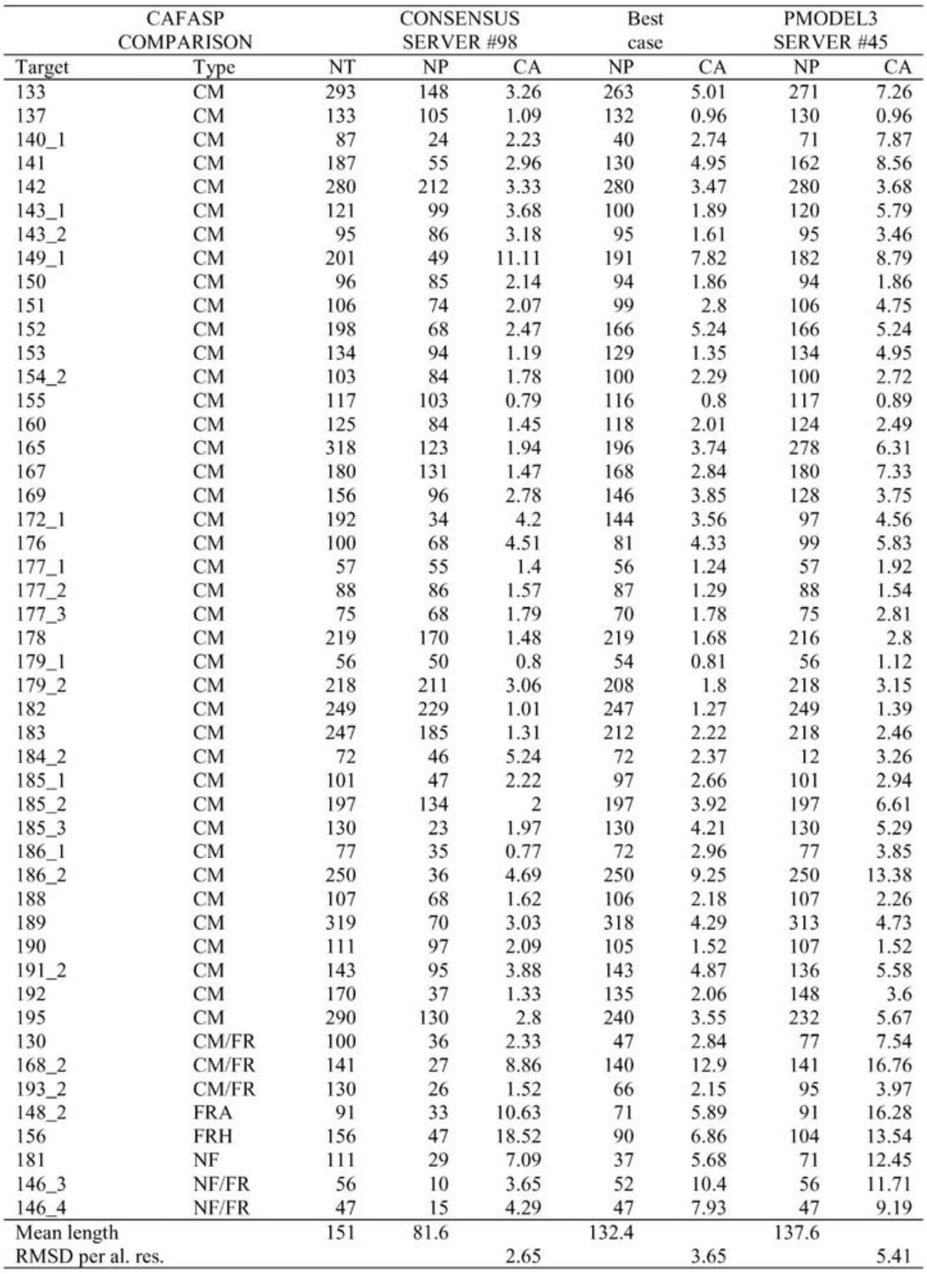


Fig. 4. Comparison of the models generated from five alignment methods and their selected consensus with respect to DAL1 in terms of RMSD and the number of aligned residues. Bars indicate one standard deviation from the average.

Figure 5 compares the CASP5 predictions we have obtained using the automatic consensus server (server #98) to the best result for each target, as well as to the output of server #45, that was deemed to produce the best overall results in the homology modeling competition. Since we restrict consideration to the highly reliable regions of the proteins, the direct comparison is of somewhat limited value, because the reliable regions, selected by the consensus method, constitute about 60 % of the total length, thus substantially smaller than for the competing methods. However, the average RMSD on the reliable regions is 2.65 Å, much lower than for the other methods. Thus, homology modeling based on consensus alignment is a reasonable first step, provided the alignment can be extended to the less reliable regions of the target (see below).



CAFASP COMPARISON			CONSENSUS SERVER #98		Best case		PMODEL3 SERVER #45	
Target	Type	NT	NP	CA	NP	CA	NP	CA
133	CM	293	148	3.26	263	5.01	271	7.26
137	CM	133	105	1.09	132	0.96	130	0.96
140_1	CM	87	24	2.23	40	2.74	71	7.87
141	CM	187	55	2.96	130	4.95	162	8.56
142	CM	280	212	3.33	280	3.47	280	3.68
143_1	CM	121	99	3.68	100	1.89	120	5.79
143_2	CM	95	86	3.18	95	1.61	95	3.46
149_1	CM	201	49	11.11	191	7.82	182	8.79
150	CM	96	85	2.14	94	1.86	94	1.86
151	CM	106	74	2.07	99	2.8	106	4.75
152	CM	198	68	2.47	166	5.24	166	5.24
153	CM	134	94	1.19	129	1.35	134	4.95
154_2	CM	103	84	1.78	100	2.29	100	2.72
155	CM	117	103	0.79	116	0.8	117	0.89
160	CM	125	84	1.45	118	2.01	124	2.49
165	CM	318	123	1.94	196	3.74	278	6.31
167	CM	180	131	1.47	168	2.84	180	7.33
169	CM	156	96	2.78	146	3.85	128	3.75
172_1	CM	192	34	4.2	144	3.56	97	4.56
176	CM	100	68	4.51	81	4.33	99	5.83
177_1	CM	57	55	1.4	56	1.24	57	1.92
177_2	CM	88	86	1.57	87	1.29	88	1.54
177_3	CM	75	68	1.79	70	1.78	75	2.81
178	CM	219	170	1.48	219	1.68	216	2.8
179_1	CM	56	50	0.8	54	0.81	56	1.12
179_2	CM	218	211	3.06	208	1.8	218	3.15
182	CM	249	229	1.01	247	1.27	249	1.39
183	CM	247	185	1.31	212	2.22	218	2.46
184_2	CM	72	46	5.24	72	2.37	12	3.26
185_1	CM	101	47	2.22	97	2.66	101	2.94
185_2	CM	197	134	2	197	3.92	197	6.61
185_3	CM	130	23	1.97	130	4.21	130	5.29
186_1	CM	77	35	0.77	72	2.96	77	3.85
186_2	CM	250	36	4.69	250	9.25	250	13.38
188	CM	107	68	1.62	106	2.18	107	2.26
189	CM	319	70	3.03	318	4.29	313	4.73
190	CM	111	97	2.09	105	1.52	107	1.52
191_2	CM	143	95	3.88	143	4.87	136	5.58
192	CM	170	37	1.33	135	2.06	148	3.6
195	CM	290	130	2.8	240	3.55	232	5.67
130	CM/FR	100	36	2.33	47	2.84	77	7.54
168_2	CM/FR	141	27	8.86	140	12.9	141	16.76
193_2	CM/FR	130	26	1.52	66	2.15	95	3.97
148_2	FRA	91	33	10.63	71	5.89	91	16.28
156	FRH	156	47	18.52	90	6.86	104	13.54
181	NF	111	29	7.09	37	5.68	71	12.45
146_3	NF/FR	56	10	3.65	52	10.4	56	11.71
146_4	NF/FR	47	15	4.29	47	7.93	47	9.19
Mean length		151	81.6		132.4		137.6	
RMSD per al. res.				2.65		3.65		5.41

Fig. 5. CAFASP3 results of the consensus server, http://structure.bu.edu. NT – total number of nucleotides, NP – number of aligned residues, CA – C_αRMSD from the x-ray structure of the target.

4 Conclusions

The multiple model approach, using a simple dynamic programming alignment with variation of the parameters, yields very good models for the relatively easy homology modeling targets, but its performance deteriorates if good template is not available. The consensus method keeps the RMSD values low (2.65 Å), but models are constructed only for 60 % of all residues. We are in the process of developing a homology modeling procedure that will integrate the two methods, and will perform the following steps:

1. The consensus alignment method is used to identify and align regions on which the alignment is highly reliable.

2. Multiple alignments are generated with the reliable regions constrained. This would result in far fewer alignments than without constraints, and will also reduce the false positive problem.

3. The models are ranked using a simple free energy evaluation expression, and the model with the lowest free energy is used as the prediction. Alternatively, the generated models are clustered on the basis of the pairwise RMSD, and the largest clusters are retained as predictions.

Acknowledgement This research has been supported by grants DBI-0213832 from the National Science Foundation, and P42 ES07381 from the National Institute of Environmental Health.

References

1. Marti-Renom,M.A., Stuart,A.C., Fiser,A., Sanchez,R., Melo,F., Sali,A.: Comparative protein structure modeling of genes and genomes. Ann. Rev. Biophys. Biomol. Struct. **29** (2000) 291-325
2. Sanchez, R., Sali, A.: Advances in protein-structure comparative modeling. Curr. Opinion in Struct. Biol. **7** (1997) 206-214
3. Fiser, A., Sanchez, R., Melo, F., Sali, A. Comparative protein structure modeling. In: M. Watanabe, M., Roux, B., MacKerell, A., Becker, O (eds.): Computational Biochemistry and Biophys. Marcel Dekker . (2001) 275-312
4. Cline,M., Hughey,R. and Karplus,K.: Predicting reliable regions in protein sequence alignments. Bioinformatics **18** (2000) 306-314.
5. Jaroszewski, L., Rychlewski, L., Godzik, A.: Improving the quality of twilight-zone alignments. Protein Science **9** (2000) 1487-1496
6. Janardhan, A., Vajda, S.: Selecting near-native conformations in homologymodeling: The role of molecular mechanics and solvation terms. Protein Science **7** (1997) 1772-1780, 1997.
7. Gatchell, D., Dennis, S., Vajda, S.: Discrimination of near-native protein structures from misfolded models by empirical free energy functions. Proteins **41** (2000) 518-534
8. Prasad, J.C., Comeau, S.R., Vajda, S. and Camacho, C.J.: Consensus alignment for reliable framework prediction in homology modeling. Bioinformatics, in press.
9. Holm,L., Sander,C.: Mapping the protein universe. Science **273** (1996) 595-602
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., Lipman,D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res. **25** (1997) 3389-3402

11. Sanchez,R., Sali,A.: Evaluation of comparative protein structure modeling by MODELLER-3. PROTEINS: Structure, Function and Genetics, **Suppl. 1** (1997) 50-58
12. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D. J., Swaminathan, S., Karplus. M.: CHARMM: A Program for Macromolecular Energy, Minimization, andDynamics Calculations, J. Comp. Chem. **4** (1983) 187-217
13. Eisenberg, D., McLachlan, A.D.: Solvation energy in protein folding and binding. Nature **319** (1986) 199-203
14. Karplus, K., Barrett, C., Hughey, R.: Hidden Markov models for detecting remote protein homologies. Bioinformatics **14** (1998) 846-856
15. Eddy S. R. : Profile hidden Markov models. Bioinformatics **14** (2001) 755-763
16. Cuff, J.A., Barton, G.J.: Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40 (2000) 502-511