BIOINFORMATICS

Vol. 19 no. 13 2003, pages 1682–1691
DOI: 10.1093/bioinformatics/btg211

Consensus alignment for reliable framework prediction in homology modeling

J.C. Prasad[1], S.R. Comeau[1], S. Vajda[2] and C.J. Camacho[1,2,*]

[1] Bioinformatics Graduate Program and [2] Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

Received on December 16, 2002; revised on December 27, 2002; accepted on March 7, 2003

ABSTRACT

Motivation: Even the best sequence alignment methods frequently fail to correctly identify the framework regions for which backbones can be copied from the template into the target structure. Since the underprediction and, more significantly, the overprediction of these regions reduces the quality of the final model, it is of prime importance to attain as much as possible of the true structural alignment between target and template.

Results: We have developed an algorithm called *Consensus* that consistently provides a high quality alignment for comparative modeling. The method follows from a benchmark analysis of the 3D models generated by ten alignment techniques for a set of 79 homologous protein structure pairs. For 20-to-40% of the targets, these methods yield models with at least 6 Å root mean square deviation (RMSD) from the native structure. We have selected the top five performing methods, and developed a consensus algorithm to generate an improved alignment. By building on the individual strength of each method, a set of criteria was implemented to remove the alignment segments that are likely to correspond to structurally dissimilar regions. The automated algorithm was validated on a different set of 48 protein pairs, resulting in 2.2 Å average RMSD for the predicted models, and only four cases in which the RMSD exceeded 3 Å. The average length of the alignments was about 75% of that found by standard structural superposition methods. The performance of *Consensus* was consistent from 2 to 32% target–template sequence identity, and hence it can be used for accurate prediction of framework regions in homology modeling.

Availability: The algorithm is available as a server at http://structure.bu.edu/cgi-bin/consensus/consensus.cgi

Contact: ccamacho@bu.edu

INTRODUCTION

Over the last decade the exponential growth of sequenced genes has prompted the development of several methods for large-scale prediction of protein structures. Homology modeling is based on the structural conservations of the framework regions between the members of a protein family. The 3D structures are more conserved in evolution than sequence, and hence even the best sequence alignment methods frequently fail to correctly identify the framework regions that possess the desired level of structural similarity. Since the quality of the alignment is the single most important factor determining the accuracy of the 3D model (Fiser *et al*., 2001), it is of substantial interest to develop methods that can both provide highly accurate sequence alignment, and identify and remove segments leading to structural dissimilarity (Cline *et al*., 2002). Indeed, even isolated structural errors seriously limit the value of models in biological applications where, for example, one or two misplaced residues can substantially reduce a protein's affinity for its substrate (Kimura *et al*., 2001).

At 40% sequence identity, alignments by pair-wise methods are only 80% correct on average (Marti-Renom *et al*., 2000), and this number drops sharply at lower similarity ranges (Rost, 1999) as gaps and localized regions of total dissimilarity increase in size and number. Two potential approaches to improving the quality of the alignment in homology modeling are utilizing the evolutionary relationships between all available homologs, and accounting for the known structure of the template. For example, the profile–profile matching algorithm by Jaroszewski *et al*. (2000) involves a weighting scheme that takes into account the topology of the evolutionary tree of all the proteins in the homologous family. Sanchez and Sali (1997a,b) employ position dependent gap penalties based on structural information of the template for generating alignments in their program MODELLER. In an extensive study on aligning protein sequences, Elofsson (2002) compared several sequence alignment methods, including those based on Hidden Markov Models (HMMs), and fold recognition methods. He observed that despite varying results with different protein pairs, methods based on sequence profiles and utilizing predicted secondary structure performed best overall. Fiser *et al*. (2001) also developed approaches that involve extensive usage of both sequence profiles and secondary structure information.

[*] To whom correspondence should be addressed.

In this paper, we develop and test a consensus alignment algorithm for the prediction of the framework regions that are structurally conserved between two proteins. The best available algorithms are used to align target and template sequences. Then, we use the consensus of these alignments, as well as additional structural constraints, to identify the regions on which the alignment is likely to result in substantial structural similarity. The methodology was developed based on a diverse set of 79 pairs of homologs with an average sequence identity of 18.5%, and was validated using a different set of 48 target–template pairs. We use the RMSD between the target and its model as our primary measure of alignment accuracy. The RMSD is a well-established quantity to determine whether the model will be accurate enough to be used in applications such as protein–protein docking (Camacho and Vajda, 2002). We find that the consensus alignment is generally more accurate than the individual methods. On an average, our method predicts models that deviate from the native structures by about 2.5 Å, and extend to almost 80% of the regions that are structurally aligned in the FSSP database (Holm and Sander, 1996). The latter, referred to as the DALI alignment, will be used as the golden standard in this work.

Several recent papers address the problem of selecting fragments of the sequence on which the alignment is likely to yield models that remain structurally similar when the two structures are superimposed. Venclovas (2001) confronted this problem by manually assessing the convergence of target–template alignments extracted from different PSI-BLAST (Altschul *et al.*, 1997) profiles to remove unreliable alignment regions. SCORE (Deane *et al.*, 2001) is another approach that attempts to predict the structurally similar regions given an accurate alignment. The method does not address the problem of alignment itself. In contrast, the algorithm presented here attempts to both get as accurate an alignment as possible, and to identify the reliable regions. Cline *et al.* (2002) have proposed the removal of unreliable regions in the alignment using a neural network based approach. Their method was able to retain 86% of the accurately aligned regions, while still 30% of substantially misaligned positions were retained. The main differences with our study compared to this one are that our goal is to remove *all* misaligned regions, and that we did not restrict our analysis to structures with high structural similarity. Indeed, for many of our target/template pairs, only limited regions superimpose in the structural alignment of the two proteins.

METHODOLOGY

Benchmark and validation sets

A total of 127 target–template pairs were selected from the FSSP database, and divided into a training set of 79 pairs and a validation set of 48 pairs. The selection was governed by the 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 the shorter sequence, must be less than 35%. The percent identity for the training set is between 5 and 32%, averaging 18.5%, whereas for the validation set is between 2 and 29%, averaging 16.8%. The full list of targets and templates is available at http://structure.bu.edu/consdoc.html

Alignment methods

We have tested ten widely used alignment methods in the context of comparative modeling. Seven of the ten methods are HMM based approaches, as implemented in SAM-T99 (Karplus *et al.*, 1998) and HMMER (Eddy, 1998) packages. The other three were based on BLAST and CLUSTALW. The following is a brief description of the application of each method. The name of a method is only indicative of the main program(s) used, and not of the original method/package the programs were part of.

The first step involves compiling a non-redundant set of PSI-BLAST hits for target and template. This set is then supplied as the homologs for the following hidden Markov model based alignment methods:

(1) T99: The target99 script takes these homologs, generates an alignment, and then iteratively improves the multiple sequence alignment by successive HMM generation and alignment of sequences to the model (referred to as ‘tuning up’). A model is constructed, and target and template sequences are aligned to it to get a pair-wise alignment.

(2) SAM: A model is built using the combined hits as family members. Target and template sequences are then aligned to this model.

(3) T99-BLAST: A PSI-BLAST 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.

(4) 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.

As already mentioned, in addition to evolutionary information, alignment quality can be improved by accounting for the secondary and tertiary structure of the template. Thus, we incorporate family alignments of the template structure, available from the FSSP database (Holm and Sander, 1996). Methods using this information are referred below as HSSP-based (Homology derived Secondary Structure Prediction) methods.

(5) T99-HSSP: Family alignments around the template (downloaded from the FSSP database) are used to build the initial profile. 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.

(6) 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 pair-wise alignment.

(7) SAM-HSSP: Similar to T99-HSSP except that the alignment of combined hits is not tuned up. This method was included simply for its speed advantage over T99-HSSP.

The above combinations of T99 and HMMER with BLAST and HSSP alignments were also tested as two-step alignment methods. In Step 1, all hits are aligned to the first combined hits model (and tuned up as well, in case of T99), and a model based on this new alignment, is constructed. In Step 2, the target and template are aligned to this model to get the final pair-wise alignment. However, preliminary benchmarking results showed very marginal improvement, if at all, while taking substantially more processing time. Therefore some of the two stage procedures were not included in the final round of benchmarking. T99 alone as a method was also not included in the final round for the same reason.

(8) BLAST-PW: Target sequence is BLASTed (pair-wise) against the template sequence to get an alignment.

During algorithm development we also considered the following methods.

(9) CLUSTALW-pairwise (Thompson *et al.*, 1994) to simply align the two sequences, i.e. no evolutionary information was used.

(10) CLUSTALW-MSA, where a multiple sequence alignment (MSA) of target, template and their combined hits is constructed. CLUSTALW-MSA was excluded from the final round due to poor performance during preliminary benchmarking.

Benchmarking and measure of accuracy

The benchmarking of the methods was done on the 79 homologous protein pairs of the training set. Following the 3D structure of the template, we build the corresponding model for the aligned regions of the target. The model is then fitted to the same regions in the target PDB (Berman *et al.*, 2000) structure, and the RMSD is calculated and compared to the structural alignment provided by the DALI server.

For benchmarking, we require a measure of accuracy that (a) reflects the structural superposition of the homology model with the crystal structure, (b) that is alignment-dependent such that to preserve the biological relevance of the model, and (c) that incorporates alignment length into the assessment. Because of the two first constraints, we moved away from using known measures like shift-score (Cline *et al.*, 2002) or GDT_TS (Zemla *et al.*, 2001). On the other hand, since RMSD is a well-established quantity to determine model accuracy, we settle on a simple weighted RMSD (WRMSD) parameter,

Table 1. Hierarchical ranking of alignment methods

Alignment method	Number of times it has the lowest WRMSD	Number of times within 0.02 from lowest WRMSD	Hierarchical clustering (no overlaps)
T99-BLAST	21	37	37
HMMER-BLAST	12	29	17
BLAST_PW	17	27	13
T99-HSSP	8	22	5
HMMER-HSSP	11	17	4
SAM	4	12	1
SAM-HSSP	4	12	1
HMMER-BLAST 2 iter	2	12	1
CLUSTALW-PW	2	6	0

$$\mathrm{WRMSD} = \mathrm{RMSD}/\sqrt{Alignment\_length}$$

The WRMSD seems to provide a fair comparison of the different methods, perhaps with a small bias towards prioritizing the RMSD. For example, a 2.5 Å RMSD alignment of 50 residues would be equivalent to a 3.5 Å RMSD alignment over 100 residues. Note that the straightforward ratio of RMSD and alignment length would prioritize alignment length, undermining our goal to find core regions with very low alignment-dependent backbone RMSD. Our working upper bound for a model was 6 Å RMSD. The above notwithstanding, other accuracy measures (Sander *et al.*, 2000) would likely result in similar outcomes. The main benefit of the WRMSD is certainly its simplicity, and hence usefulness in the process of algorithm development. As expected, the structural superposition based alignment from DALI, the gold standard we use in this work, always has the lowest WRMSD compared to all other alignment methods.

The benchmarking of the final nine methods arranged in a hierarchical manner, so that we cover a broad range of alignments, is summarized in Table 1. For 37 out of 79 cases, T99-BLAST produced the largest number of low WRMSD (within 0.02 of the lowest). For 17 of the remaining 42, HMMER-BLAST gets the largest number of low WRMSD, and so on. Surprisingly, leaving aside the cases for which BLAST-PW method failed to produce an alignment, this fast and simple method had the best overall average WRMSD. T99-HSSP did not perform better overall than HMMER-HSSP. However, as a function of the hierarchically ranked number of targets for which each method obtained the best WRMSD, the best performers for the training set were T99-BLAST, HMMER-BLAST, BLAST-PW, T99-HSSP and HMMER-HSSP. Together, they produced the best alignment(s) for 75 out of 79 cases. These five methods were then selected to be used in the consensus methodology.

Cropping and splitting protein structures

It is well known that the quality of the alignment is seriously undermined when a template (target) have multiple domains and one of the corresponding target (template) domains is either not homologous, absent, or connected by a flexible hinge. To minimize potential misalignments arising from multi-domain sequences, we apply the following two-stage approach: first, the target sequence is aligned to template sequence using BLAST. If the highest-scoring segment pair spans the entire target (template) sequence, and the unaligned fraction of template (target) is more than 30% larger than the target (template), then the unaligned region of the template (target) is removed. Second, we use our domain splitting program, *DomainSplit*, to identify loosely connected regions on the structure. Each template domain is BLASTed against the target sequence (or its cropped version, depending on stage one). If there are easily identifiable domains that align to the target with a reasonable E-value, then these domains are designated as 'alignable' domains, and the alignment is performed between these regions only.

Several approaches for identification of protein structural domains have been documented in literature (Holm and Sander, 1994; Swindells, 1995; Siddiqui and Barton, 1995; Sowdhamini and Blundell, 1995; Taylor, 1999). There also exist well-known public databases online such as VAST (Madej *et al*., 1995; Gibrat *et al*., 1996) and 3Dee (Siddiqui *et al*., 2001) that provide domain information for PDB structures. However, here we require that *DomainSplit* further split loosely connected domains, and not necessarily those that share a large interface. The basic algorithm involves clustering the tertiary structure of the molecule such that regions that do not have a significant contact area are split.

Consensus and selection

For all our training set targets, at least one of the aforementioned top five performing methods was able to produce a low RMSD model. However, as expected, consensus alone did not eliminate all the structurally dissimilar regions, nor did it produce alignment lengths comparable to the DALI alignment. Hence, we developed a selection procedure for determining whether or not two aligned residues can be included for accurate homology modeling. This process invokes further structural considerations, e.g. secondary structure and solvent exposure information and to a lesser degree secondary structure prediction of the target, as well as the consensus strength (explained below). A flow chart of the overall algorithm is depicted in Figure 1.

Consensus. As sketched in Figure 2A, 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 $X_{\mathrm{tar}}X_{\mathrm{tem}}$ 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.

Selection. Since consensus strength does not eliminate all the regions of potential structural dissimilarity, the following selection method (see Fig. 2B) is applied. If CS is 9 and template residue X_{tem} is buried, $X_{\mathrm{tar}}X_{\mathrm{tem}}$ 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 (Cuff and Barton, 2000)] 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 uncertain regions with high number of gaps are deselected. Consensus strength is always considered in all selection and de-selection steps.

The output of the algorithm consists of the full selected consensus. If the template had multiple alignable domains, then the output would include the selected consensus for each domain separately, and all the selected domains merged together.

AVAILABILITY

A beta version of the *Consensus* program is available at http://structure.bu.edu/cgi-bin/consensus/consensus.cgi If no template is given, the method has been adapted to pick a template by using a method similar to PSI-BLAST. The training set and the validation set of target and templates used in this study are also linked to this webpage. The automated splitting domain program *DomainSplit* is available

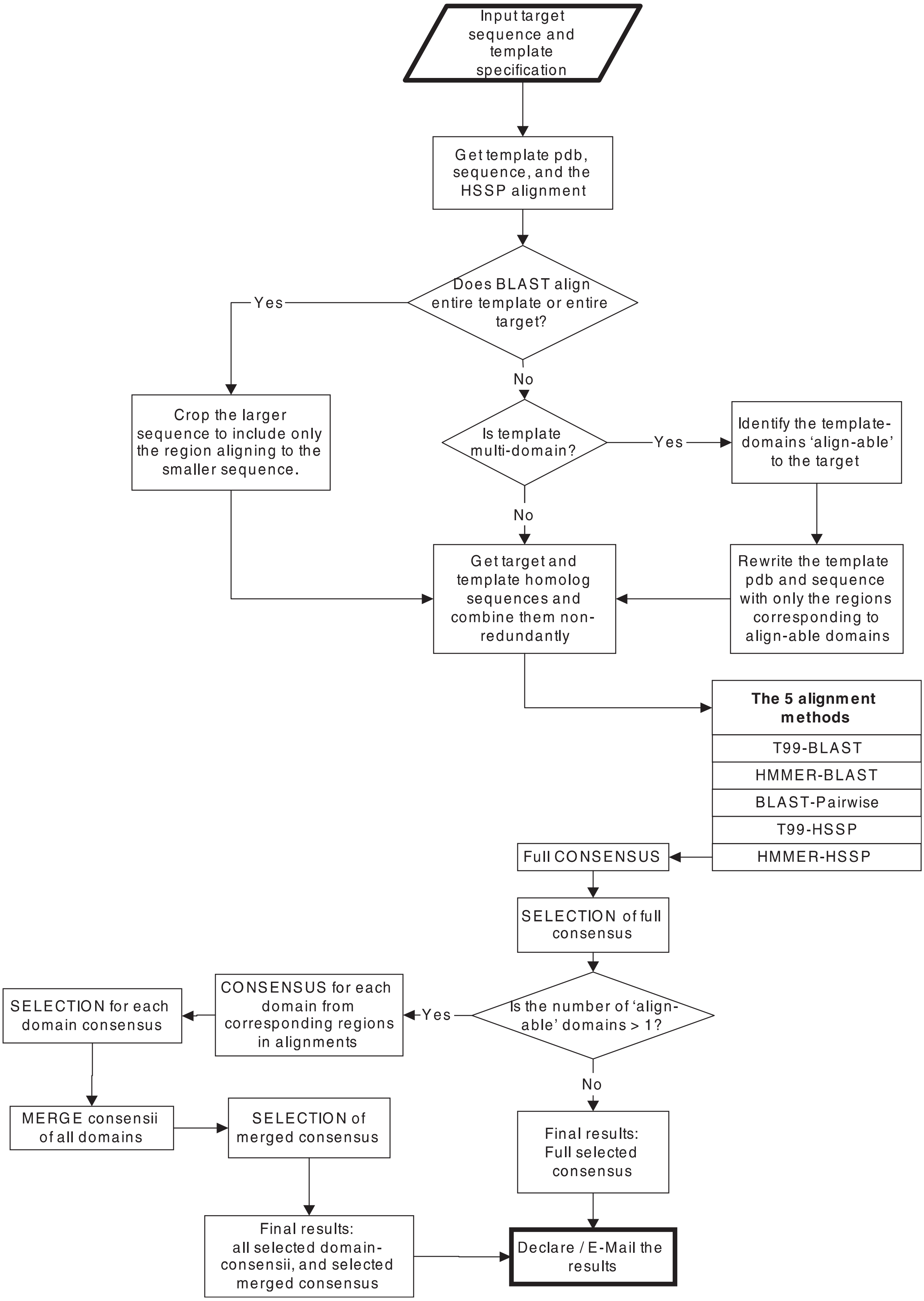
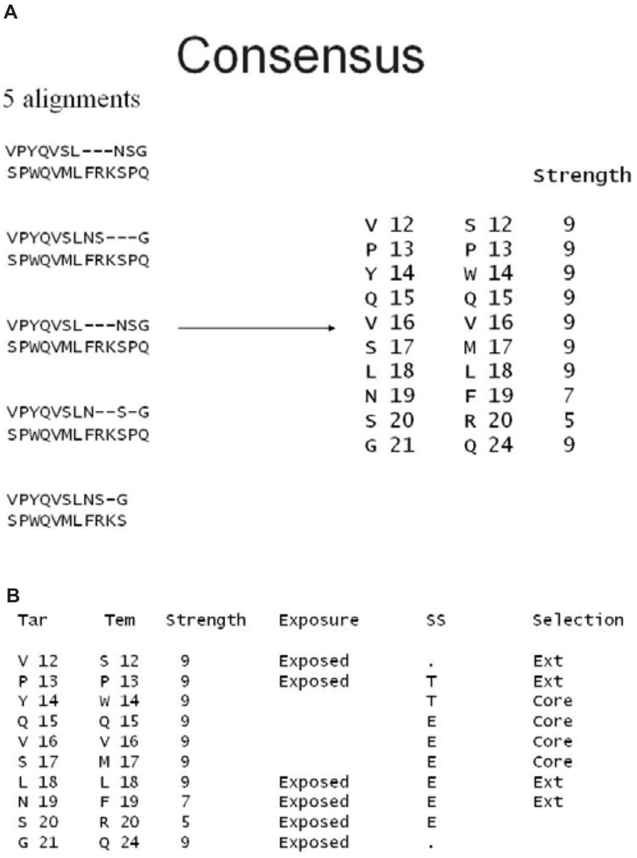

Input target sequence and template specification
Get template pdb, sequence, and the HSSP alignment
Does BLAST align entire template or entire target?
Yes
No
Crop the larger sequence to include only the region aligning to the smaller sequence.
Is template multi-domain?
Yes
No
Identify the template-domains 'align-able' to the target
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 'align-able' domains > 1?
Yes
No
CONSENSUS for each domain from corresponding regions in alignments
SELECTION for each domain consensus
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.



```
Consensus

5 alignments

VPYQVSL---NSG
SPWQVMLFRKSPQ                                        Strength

VPYQVSLNS---G               V 12    S 12     9
SPWQVMLFRKSPQ               P 13    P 13     9
                            Y 14    W 14     9
                            Q 15    Q 15     9
VPYQVSL---NSG  ---------->  V 16    V 16     9
SPWQVMLFRKSPQ               S 17    M 17     9
                            L 18    L 18     9
                            N 19    F 19     7
VPYQVSLN--S-G               S 20    R 20     5
SPWQVMLFRKSPQ               G 21    Q 24     9

VPYQVSLNS-G
SPWQVMLFRKS
```

Tar	Tem	Strength	Exposure	SS	Selection
V 12	S 12	9	Exposed	.	Ext
P 13	P 13	9	Exposed	T	Ext
Y 14	W 14	9		T	Core
Q 15	Q 15	9		E	Core
V 16	V 16	9		E	Core
S 17	M 17	9		E	Core
L 18	L 18	9	Exposed	E	Ext
N 19	F 19	7	Exposed	E	Ext
S 20	R 20	5	Exposed	E	
G 21	Q 24	9	Exposed	.	

Fig. 2. Consensus and selection of optimal alignment. (**A**) Assigning a consensus value to each residue of the target sequence by pooling the output from the five selected alignment methods. (**B**) Selection of the core residues on the basis of the consensus core, solvent exposure, and secondary structure. Unexposed residues with high consensus score are first selected, and this core is extended on both sides subject to exposure and secondary structure conditions.

upon request and as a server at http://structure.bu.edu/cgi-bin/domain/domainsplit.cgi The consensus and selection programs are available for academic use upon request.

RESULTS

Training set

We apply our selection rules to the consensus alignment between target and template after domain splitting. Figure 3 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 comparisons the individual alignment methods have benefited from the automatic splitting and cropping of domains which do not align with the target sequence, otherwise the average RMSD for, say, T99-BLAST would be higher than 6 Å. On the other hand, as shown in Figure 4, the length of the alignments produced by the consensus method is obviously smaller than that of its compound methods, corresponding to around 79% of the DALI alignment. It should be noted that some methods yielded alignment lengths longer than DALI—a clear sign of the overreaching properties of the alignments.

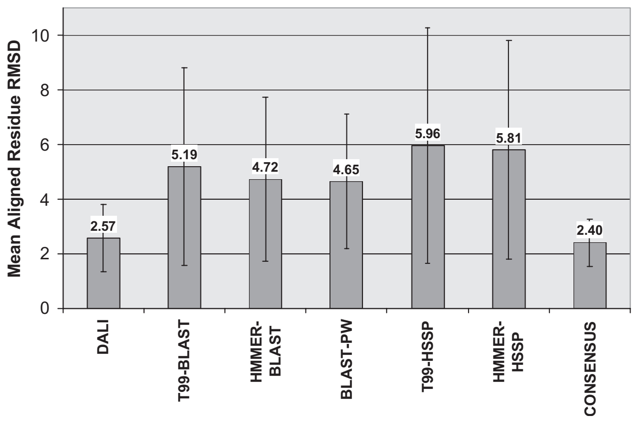


Fig. 3. Aligned residues RMSD for 79 target–template pairs in the training set. Comparison of five alignment methods and the *Consensus* prediction of reliable regions, compared to DALI structural alignment. Bars indicate one standard deviation.

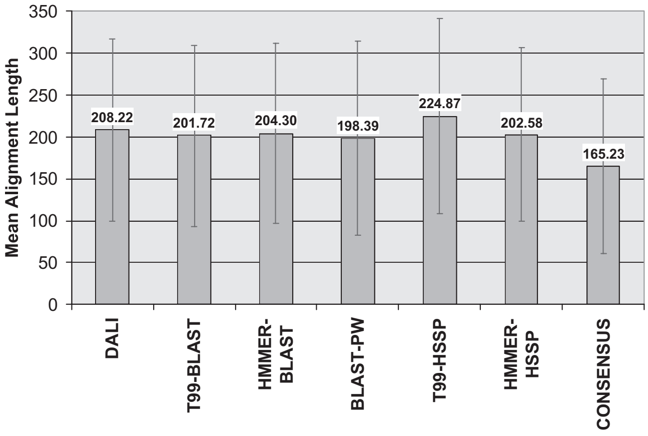


Fig. 4. Average alignment length for training set. Performance of five methods and the *Consensus* prediction of reliable regions, compared to DALI. Bars indicate one standard deviation.

We have also confirmed that *Consensus* generates reliable models over the whole range of percent identity. Figure 5 shows the WRMSD as a function of percent identity. We recall that a WRMSD value of 0.2 is equivalent to a 2 Å RMSD

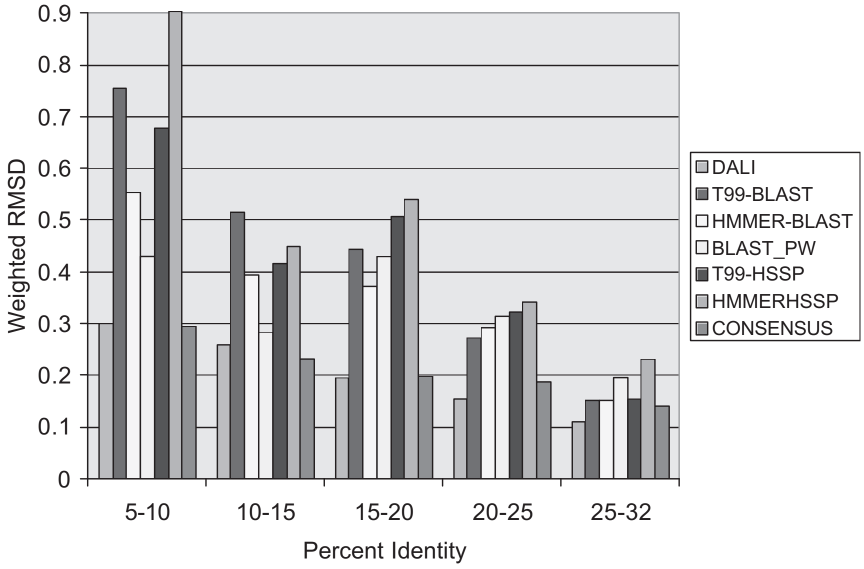


Fig. 5. Weighted RMSD (WRMSD) values as function of percent identity, obtained by five alignment methods, *Consensus*, and DALI for the 79 pairs in the training set. DALI and *Consensus* perform the best at all levels of percent identity. We note that WRMSD = 0.2 is equivalent to 2 Å RMSD for a segment 100 aligned residues.

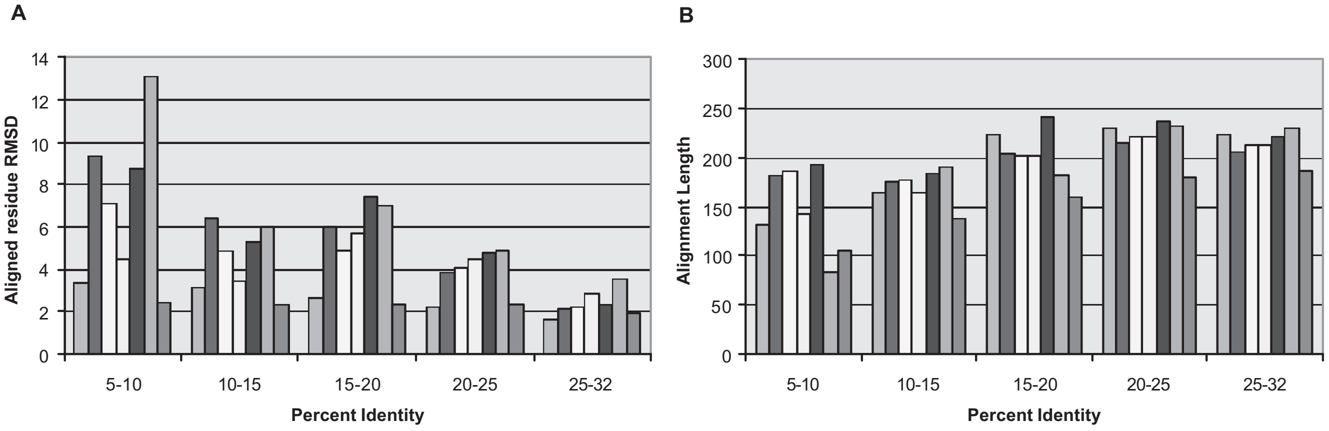


Fig. 6. Homology modeling properties as a function of percent identity for training set. Average (A) aligned residue RMSD, and (B) alignment length for five alignment methods, *Consensus*, and DALI. The legend is the same as in Figure 5.

model for a 100 residue long alignment. To further confirm the consistency of the WRMSD measure, Figure 6 shows an expanded analysis of the performance of the methods versus % identity. As expected, the DALI alignment performs better than any of the five component methods for all 127 homolog pairs studied here (79 training and 48 validation pairs).

Validation set

Despite the fact that the validation set had a somewhat lower sequence identity than the training set, the average RMSD obtained by *Consensus* was 2.22 Å, 3% lower than that obtained by DALI structural alignment (Fig. 7). The average alignment length is 75% of the optimal alignment as given by DALI (see Fig. 8), i.e. somewhat smaller than the 79% found for the training set. As for the training set, the *Consensus* algorithm again generates reliable models over the whole range of percent identity (data not shown). As shown in Figures 7 and 8, the low RMSD models also translate into improved alignments. Indeed, the average RMSD and alignment length for the full consensus alignment (without removing unreliable regions) is somewhat better than any of the individual methods. Perhaps, the only exception is BLAST_PW that has shorter alignments (84% of the consensus length) and a better average RMSD.

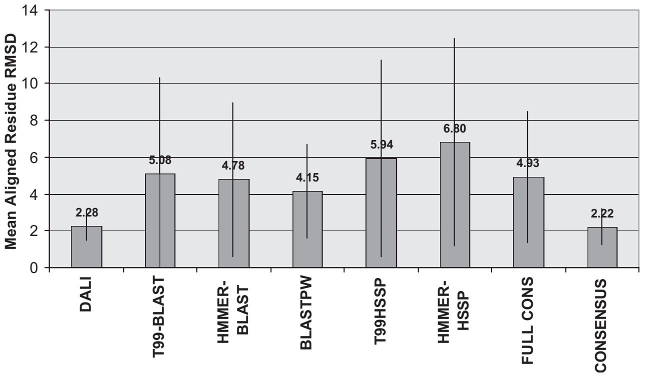


Fig. 7. Aligned residues RMSD for 50 target–template pairs in validation set. Comparison of the models generated by five alignment methods and both the full consensus alignment (without removing unreliable regions) and the *Consensus* prediction of reliable regions with respect to DALI. Bars indicate one standard deviation from the average.

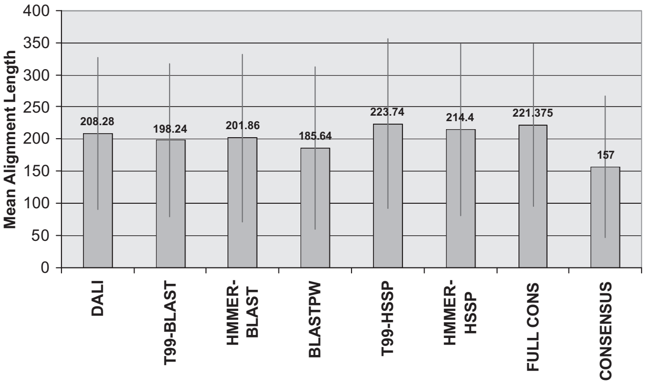


Fig. 8. Alignment length for validation set. Performance of five alignment methods and both the full consensus alignment and the *Consensus* prediction of reliable regions with respect to DALI. Bars indicate one standard deviation.

Table 2. Validation set

Target	Template	Explanation
(cases where the consensus method performed well.)		
1e6wA	1cvl	The consensus method identified the reliable regions of the alignment. Selection reduced the RMSD by avoiding regions with large gaps.
1fchA	1hxiA	Selection procedure correctly identifies the long exposed C-terminal helix as unreliable for modeling.
1cjcA	1qjdA	The consensus strength was generally low, so most unreliable regions were removed from the alignment. However, the final alignment length was far lower than that of DALI structural alignment.
1eggA	1ixxB	Despite the consensus alignment, a large loop was identified correctly as a region of potential structural dissimilarity.
(cases where the consensus method performed poorly)		
1uok	1bvzA	Two of the component methods outperformed the consensus method in terms of alignment length, whereas T99-BLAST got the wrong alignment distorting the consensus. The *DomainSplit* program did not properly identify the domains of this protein.
2tnfA	1aly	T99-BLAST significantly outperforms our method. Two other methods got the wrong alignment, distorting the overall consensus alignment.
1iakB	1hdmB	The consensus method performed slightly poorer than 4 out of the 5 component methods due to poor agreement between methods in a long helix region.
1tf4A	1nbcA	RMSD is greater than 5 Å. The consensus alignment is better than any of the individual methods, but worse than DALI.
1a0p	1aihA	RMSD is greater than 5 Å. Consensus does relatively better than all the five component methods.

We found that only four cases had RMSDs greater than 3 Å, and none above 6 Å. Only two cases, PDB codes 1tf4A-1nbcA and 1a0p-1aihA, have RMSDs above 5 Å. These are worst than the DALI alignment, but still significantly better than the component methods. Finally, Table 2 lists some of the best and worst performing cases.

DISCUSSION

An important step in homology modeling is to identify the framework regions that are structurally conserved between a template and the target, and to align these regions with the highest achievable accuracy. Since no current homology modeling method can recover from an incorrect alignment, it is crucial to correctly identify the regions that translate into good structural alignment. In our method we seek the consensus of five well-tested alignment algorithm. This approach has two advantages. First, averaging the results from several methods seems to provide the best results in homology modeling, as the success of meta-servers demonstrated at the CASP5 meeting (December 1–5, 2002, Pacific Grove, CA). Notice, however, that we use the consensus idea to generate the best alignment rather than creating a consensus 3D structure from a number of models. Second, employing five alignment algorithms provides an excellent tool to assign a measure of reliability to each position in the sequence. This information is very important both for accurate model building, and in biological applications in which it is frequently critical to assign a measure of confidence in the different parts of the model. No current metaserver for homology modeling provides such measure of reliability.

The *Consensus* algorithm is very fast, taking no more than a few seconds in selecting the reliable regions of the alignment. We find that the method produces alignments that lead to models which have RMSDs on the order of 2.3 Å, much lower than standard methods available in the literature. Indeed, the obtained RMSDs are even lower than those obtained by structural superposition, though the alignment lengths are around 20% shorter than the optimal structural alignment as established by DALI.

For the most part, the structurally similar regions not selected by *Consensus* entailed either terminii or, less often, alignment shifts within secondary structures. Terminal regions are usually more flexible than the rest of the protein structure, thus we imposed more stringent constraints than in the rest of the sequence. The end result is that these motifs are sometimes unnecessarily excluded from the alignment. In such cases, visual inspection can readily re-instate the alignment. However, failing to remove these regions when they are structurally misaligned inmediately brings the RMSD to double digits. A similar situation occurs with gap-broken secondary structures, these gaps can usually be filled from either direction yielding a 50–50 chance of getting the right alignment. *Consensus* does not fill these gaps but, in principle, they could be filled manually after obtaining the *Consensus* alignment. Further improvements of the method involves adding knowledge based information that would allow us to improve our selection criteria in these regions.

If no template is given, the server provides the option to search for a template using a method like PDB-BLAST (Burnham Institute). Using this setting, *Consensus* participated in the most recent Critical Assessment of Fully Automated Structure Prediction (CAFASP 2002). The results obtained in this community-wide blind experiment were consistent with the validation results presented here. Namely, for the 44 comparative model targets that *Consensus* was able to automatically find a template in the PDB, we obtained a mean RMSD (per residue) of 2.4 Å for an average alignment length of 89 residues (the average target length was 158 residues). *Consensus* also obtained the best contact predictions among all 52 servers participating in CAFASP (see http://www.pdg.cnb.uam.es/eva/cafasp3/).

Our results compare well with those available in the literature, with RMSDs comparable to those obtained by structural alignment methods. The method provides an improved alignment and a measure of sequence alignment reliability, or consensus strength, for each region of the sequence. The low RMSD models predicted by the method can then be used for either biological applications or as starting points for further extension and refinement. With regards to template selection, our preliminary evidence suggests that the template that yields the longest selected alignment is always the best for homology modeling. Furthermore, we are also compiling evidence to show how multiple templates that produce alignments for different regions can be combined to build a full protein model. In summary, *Consensus* provides a reliable tool for researchers to obtain highly confident alignments for homology modeling.

ACKNOWLEDGEMENTS

This research has been supported by grants GM61867 from the National Institute of Health, and P42 ES07381 from the National Institute of Environmental Health.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Camacho,C.J. and Vajda,S. (2002) Protein–protein association kinetics and protein docking. *Curr. Opin. Struct. Biol.*, **12**, 36–40.

Cline,M., Hughey,R. and Karplus,K. (2002) Predicting reliable regions in protein sequence alignments. *Bioinformatics*, **18**, 306–314.

Cuff,J.A. and Barton,G.J. (2000) Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. *PROTEINS: Structure, Function and Genetics*, **40**, 502–511.

Deane,C.M., Kaas,Q. and Blundell,T.L. (2001) SCORE: predicting the core of protein models. *Bioinformatics*, **17**, 541–550.

Eddy,S.R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Elofsson,A. (2002) A study on how to best align protein sequences. *PROTEINS: Structure, Function and Genetics*, **46**, 330–339.

Fiser,A., Sanchez,R., Melo,F. and Sali,A. (2001) Comparative protein structure modeling. *Computational Biochemistry and Biophysics*. Marcel Dekker, New York, pp. 275–312.

Gibrat,J.-F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Holm,L. and Sander,C. (1994) Parser for protein folding units. *PROTEINS: Structure, Function and Genetics*, **19**, 256–268.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.

Jaroszewski,L., Rychlewski,L. and Godzik,A. (2000) Improving the quality of twilight-zone alignments. *Protein Sci.*, **9**, 1487–1496.

Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Kimura,R., Brower,R., Vajda,S. and Camacho,C.J. (2001) Dynamical view of the positions of key side chains in protein–protein recognition. *Biophys. J.*, **80**, 635–642.

Madej,T., Gibrat,J.-F. and Bryant,S.H. (1995) Threading a database of protein cores. *PROTEINS: Structure, Function and Genetics*, **23**, 356–369.

Marti-Renom,M.A., Stuart,A.C., Fiser,A., Sanchez,R., Melo,F. and Sali,A. (2000) Comparative protein structure modeling of genes and genomes. *Ann. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Sander,J.M., Arthur,J.W. and Dunbrack,R.L. (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *PROTEINS: Structure, Function and Genetics*, **40**, 6–22.

Sanchez,R. and Sali,A. (1997a) Advances in protein-structure comparative modeling. *Curr. Opin. Struct. Biol.*, **7**, 206–214.

Sanchez,R. and Sali,A. (1997b) Evaluation of comparative protein structure modeling by MODELLER-3. *PROTEINS: Structure, Function and Genetics*, Suppl. 1, 50–58.

Siddiqui,A.S. and Barton,G.J. (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.*, **4**, 872–884.

Siddiqui,A.S., Dengler,U. and Barton,G.J. (2001) 3Dee: a database of protein structural domains. *Bioinformatics*, **17**, 200–201.

Sowdhamini,R. and Blundell,T.L. (1995) An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci.*, **4**, 506–520.

Taylor,W.R. (1999) Protein structural domain identification. *Protein Eng.*, **12**, 203–216.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Venclovas, C. (2001) Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. *PROTEINS: Structure, Function and Genetics*, Suppl. 5, 47–54.

Zemla,A., Venclovas,C., Moult,J. and Fidelis,K. (2001) Processing and evaluation of predictions in CASP4. *PROTEINS: Structure, Function and Genetics*, Suppl. 5, 13–21.