W50–W54 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh456

Consensus alignment server for reliable comparative modeling with distant templates

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

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

Received February 15, 2004; Revised March 19, 2004; Accepted April 26, 2004

ABSTRACT

Consensus is a server developed to produce high-quality alignments for comparative modeling, and to identify the alignment regions reliable for copying from a given template. This is accomplished even when target–template sequence identity is as low as 5%. Combining the output from five different alignment methods, the server produces a consensus alignment, with a reliability measure indicated for each position and a prediction of the regions suitable for modeling. Models built using the server predictions are typically within 3 Å rms deviations from the crystal structure. Users can upload a target protein sequence and specify a template (PDB code); if no template is given, the server will search for one. The method has been validated on a large set of homologous protein structure pairs. The Consensus server should prove useful for modelers for whom the structural reliability of the model is critical in their applications. It is currently available at http://structure.bu.edu/cgi-bin/consensus/consensus.cgi.

INTRODUCTION

Several sophisticated sequence alignment methods have been reported earlier, many of which yield very-high-quality alignments. However, in the context of comparative modeling (CM), even the best methods do not always result in highly accurate models. Structural divergence of some high sequence similarity regions in proteins, and alignment uncertainty in regions of low similarity are some of the reasons for this. We have developed an algorithm that consistently gives a high-quality target–template alignment, and predicts the regions that are structurally divergent between the target and template, regardless of the local sequence similarity. To identify these regions the algorithm does not require more than one template structure. The method follows from a benchmark analysis of the three-dimensional (3D) models generated by 10 alignment techniques (1–4) for a set of 79 homologous protein structure pairs. We selected the top five methods and developed a consensus algorithm to generate an improved alignment with a reliability measure for each alignment position. A set of criteria was implemented to identify zones of this consensus alignment that are suited for comparative modeling, i.e. regions that are structurally conserved between target and template. The algorithm was validated on an independent set of 48 homologous structure pairs. The average RMSD of the models obtained was 2.2 Å, while the average length of the alignments was ~75% of that found by standard structural superposition methods. The performance was consistent over a range of target–template sequence identity of 2–32%. (1). To our knowledge, Consensus is the only server that classifies alignment regions on the basis of alignment reliability as well as potential structural divergence between the target and the template. This should prove useful in the analysis of comparative models where the accuracy of the model has an impact on their application.

BRIEF DESCRIPTION OF THE ALGORITHM

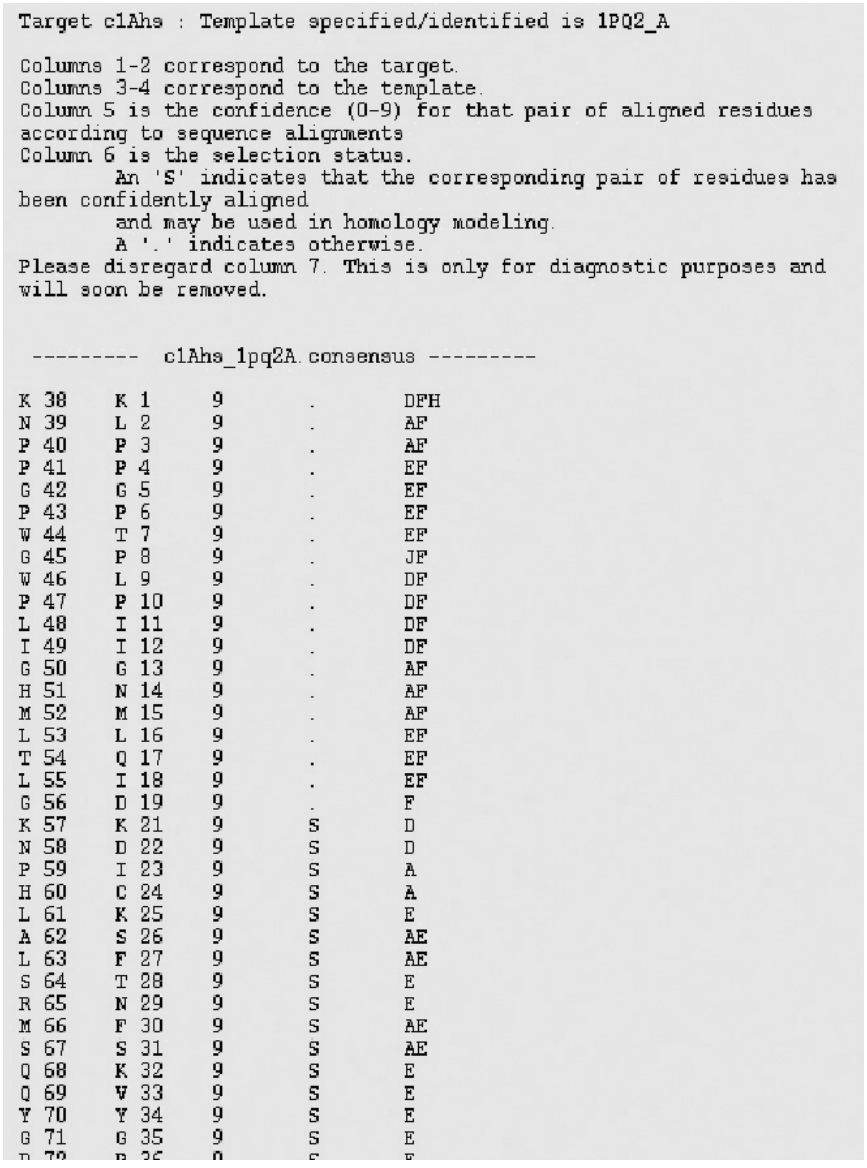
The following three types of information (not necessarily independent of each other) are required to generate high-quality alignments and to select the regions of modeling suitability:

(i) Evolutionary/family sequence information (profile/HMM)
(ii) Structural information (template and template family members)
(iii) Secondary structure prediction of the target

Alignment of target and template. Target and template sequences are aligned using the chosen five alignment methods—T99-BLAST, HMMER-BLAST, BLAST_PW, T99-HSSP, HMMER-HSSP based on (2–4) [as determined by a benchmark analysis in (1)].

Consensus of alignments. The five alignments generated are polled for each position to get a consensus alignment

*To whom correspondence should be addressed. Tel: +1 617 353 4842; Fax: +1 617 353 6766; Email: ccamacho@bu.edu



```
Target c1Ahs : Template specified/identified is 1PQ2_A

Columns 1-2 correspond to the target.
Columns 3-4 correspond to the template.
Column 5 is the confidence (0-9) for that pair of aligned residues
according to sequence alignments
Column 6 is the selection status.
       An 'S' indicates that the corresponding pair of residues has
been confidently aligned
       and may be used in homology modeling.
       A '.' indicates otherwise.
Please disregard column 7. This is only for diagnostic purposes and
will soon be removed.


 ---------  c1Ahs_1pq2A.consensus ---------

K 38    K 1     9       .       DFH
N 39    L 2     9       .       AF
P 40    P 3     9       .       AF
P 41    P 4     9       .       EF
G 42    G 5     9       .       EF
P 43    P 6     9       .       EF
W 44    T 7     9       .       EF
G 45    P 8     9       .       JF
W 46    L 9     9       .       DF
P 47    P 10    9       .       DF
L 48    I 11    9       .       DF
I 49    I 12    9       .       DF
G 50    G 13    9       .       AF
H 51    N 14    9       .       AF
M 52    M 15    9       .       AF
L 53    L 16    9       .       EF
T 54    Q 17    9       .       EF
L 55    I 18    9       .       EF
G 56    D 19    9       .       F
K 57    K 21    9       S       D
N 58    D 22    9       S       D
P 59    I 23    9       S       A
H 60    C 24    9       S       A
L 61    K 25    9       S       E
A 62    S 26    9       S       AE
L 63    F 27    9       S       AE
S 64    T 28    9       S       E
R 65    N 29    9       S       E
M 66    F 30    9       S       AE
S 67    S 31    9       S       AE
Q 68    K 32    9       S       E
Q 69    V 33    9       S       E
Y 70    Y 34    9       S       E
G 71    G 35    9       S       E
```

Figure 1. Screenshot of the alignment output from the Consensus server.

in a tabular format (Figure 1). Each residue–residue alignment position is assigned a consensus strength (CS) ranging from 0 to 9 based on its occurrence in the five initial alignments. In cases of conflicts between regions of alignments, the region with higher consensus strength receives priority. Ties between alignments are resolved according to the rankings obtained in the benchmark analysis earlier (1).

Identification of regions suitable for homology modeling. Consensus strength is a measure of alignment reliability. It does not indicate structural similarity. Identification of regions in the consensus alignment that are reliable for copying the backbone from the template is carried out by a selection procedure, wherein regions of highest CS that are buried in the template are selected first. This forms the core of the selection. It is then extended both sides subject to the CS, until a misaligned Glycine residue is encountered. Alignment regions corresponding to long template helices and sheets are selected subject to solvent exposure and percentage match of the predicted secondary structure of the target [using JNET (5)] with the 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 applied. Finally, 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 deselection steps. The final output of the server includes the selected consensus with reliable regions identified (Figure 1), and the corresponding models of the full consensus alignment and of the reliable regions predicted by the method (Figure 2).

Sequence pre-processing. To minimize potential misalignments arising from multi-domain sequences, we apply the

Figure 2. Superposition of our server's prediction of CAFASP target T0167. The actual predicted structure based on the selection is shown in blue. Segments of the full consensus alignment model that were left out by the selection procedure are shown in green, and the crystal structure is in red. The helix at the N-terminus was misaligned, and the server was correctly able to dismiss it from the high-accuracy model. One can also see that a loop (top-middle of the figure) that was left out by the server is actually broken even in the crystal structure.

following two-stage approach using a simplified preliminary alignment:

(i) If the two sequences differ significantly in length, and the smaller sequence completely aligns to the larger one, the larger one is cropped to the aligned region.
(ii) Otherwise, template domains are identified, and non-alignable ones are dropped.

If the template has multiple alignable domains, then each domain is considered separately, and all the selected domains merged together. We use a domain identification program that was developed in-house, DomainSplit, to identify loosely connected regions on the structure. It is also available as a server at http://structure.bu.edu/cgi-bin/domain/domainsplit.cgi.

The Consensus server is an independent server which does not rely on the output from any other servers [i.e. it is not a meta-server (6)]. It currently uses only one template. It can be run several times to analyze multiple templates. However, at this point, the integration of the different structures needs to be done manually.

PERFORMANCE

Validation

A set of 79 homologous protein pairs were used in the development of the algorithm (1). The server was then validated on an independent set of 48 homologous pairs (Figures 3 and 4). Both sets were chosen automatically, based on criteria explained in (1). We found that the server's predictions of the structurally reliable regions led to models that have C-alpha RMSDs on the order of 2.3 Å. However, the average alignment length was ~20% shorter than the optimal structural alignment as established by DALI (7).

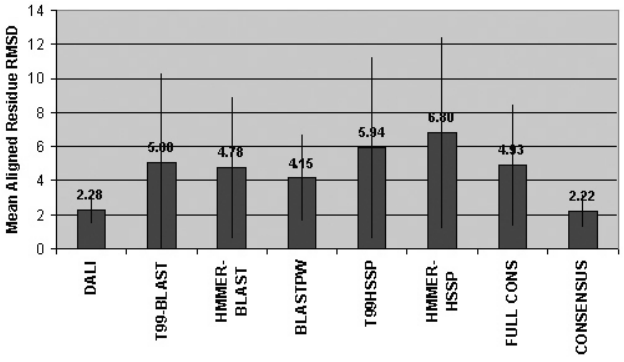


Figure 3. Aligned residues RMSD (C-alpha) for 48 target–template pairs in validation set. The average RMSD of the models generated by five alignment methods and both the full consensus alignment and the Consensus prediction of reliable regions with respect to DALI. Bars indicate one standard deviation from the average. This figure was first published in reference (1) and is reproduced here with permission.

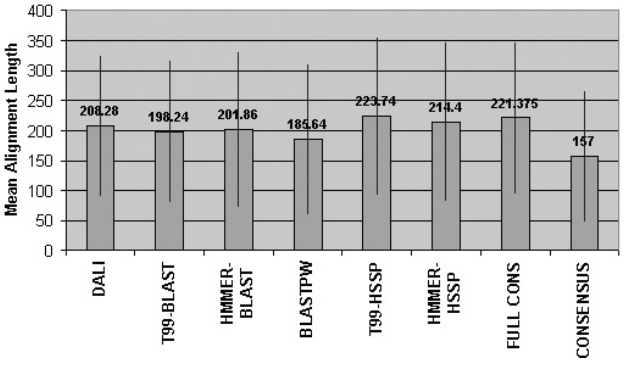


Figure 4. Alignment length for validation set. Average alignment length of five alignment methods and both the full consensus alignment and the Consensus prediction of reliable regions with respect to DALI. Bars indicate one standard deviation. This figure was first published in reference (1) and is reproduced here with permission.

Figures 3 and 4 show the means of RMSD and alignment length of the Consensus server, the structural superposition alignment (DALI database) and the 5 component alignment methods used; the average sequence identity of our validation set was 16.8%. 'Full consensus' stands for a completely selected consensus alignment, i.e. no regions have been removed by the selection criteria. These figures reflect the known fact that models generated based on most alignment methods tend to overextend to regions that are not suitable for structural modeling. At the same time, Figures 3 and 4 show that the consensus server consistently generated structurally reliable (low RMSD) models. The latter is the only goal of the server, i.e. to help users identify the regions of a target–template alignment that lead to low RMSD models, as well as identify the regions that are susceptible to lead to big differences with respect to the crystal structure.

In order to illustrate the performance of the server as a function of percentage identity, we computed a weighted RMSD parameter

$$\mathrm{WRMSD} = \mathrm{RMSD}/\sqrt{\text{Alignment length}}.$$

This parameter attempts to assess the structural quality of models with different alignment length. Figure 5 shows the

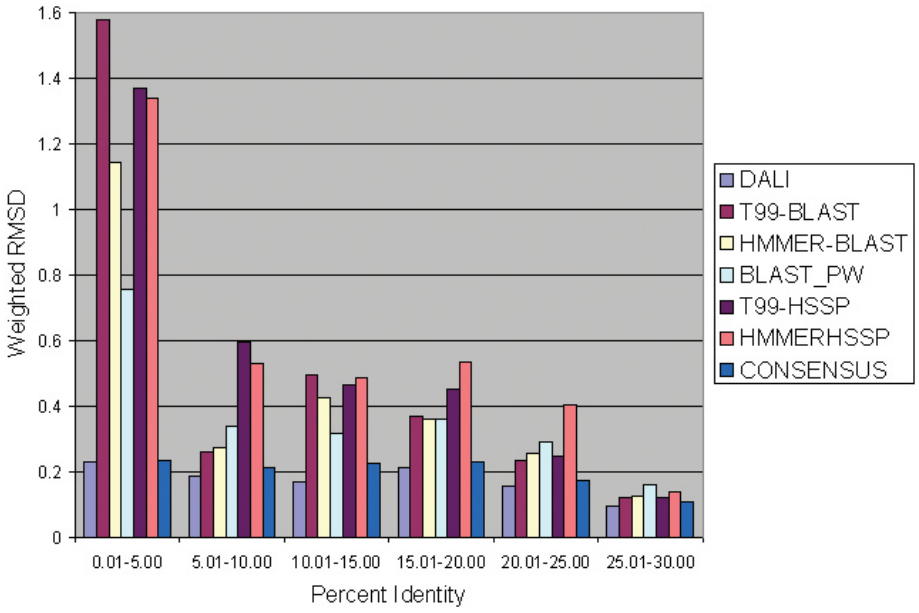


Figure 5. Validation set: weighted RMSD (C-alpha) versus percentage identity. Note that the framework models based on the Consensus alignment continue to be accurate below 10% sequence identity, as do the ones based on structural superposition (DALI).

performance of the different methods as a function of percentage identity. This is important since in the field of comparative modeling, it is often mentioned that an alignment between sequences with <5% identity is worthless, or nearly random. One can notice that the framework models constructed from the Consensus alignments continue to be accurate at low sequence identity (∼5%).

We would like to stress that structural assessment methodologies such as the ones used in CASP (8) are not appropriate to estimate structural accuracy because these methods do not penalize the inaccuracy of the structural predictions. Furthermore, although the weighted RMSD is not perfect, it is enough to convey the point that Consensus performs well across the whole range of sequence identity. Also, it is consistent with other such methods [see, e.g. (9)]; for further details see (1).

CAFASP3

To demonstrate an independent validation of the performance of Consensus in CM, Table 1 shows the predictions obtained by the server (ID number 98) at the CAFASP3 community-wide experiment. For the 40 CM targets that the Consensus server was able to automatically find a template in the PDB (10) (using a simple PDB-BLAST like template search method), we obtained a mean RMSD per residue of 2.36 Å for an average alignment length of 92 residues (the average target length was 160 residues; we did not compute the average length of the DALI alignment for this set). Figure 2 shows the submitted prediction for Target 167. This is one example of how the prediction of the server recognized regions of an alignment that do not superimpose well with the crystal structure. In the future, we expect to improve the performance of Consensus by incorporating more sophisticated searching techniques for templates and by implementing the ability to integrate multiple template information.

USING THE CONSENSUS SERVER

Input files required and format details

The server has a straightforward web interface. The target sequence is expected in pure text (with no header) or FASTA format. It may be entered in the text box provided or uploaded as a file. A five-lettered name (may contain numbers but no special characters) needs to be entered for the target. Template may be specified as a four-lettered pdb code followed by a chain identifier (a must if the template pdb has a chain identifier in it). If no template is specified, the method has been adapted to pick a template by using a method similar to PDB-BLAST (Godzik group, Burnham Institute). We would like to note, however, that template identification is not a strength of the server. Therefore, for distant homologues, it is best if the template is supplied. The results are sent by email.

The consensus polling and the selection algorithms that are the core of the server are very fast, taking no more than a few seconds to complete. However, the five component alignment methods that the server uses may take a up to a few hours to produce the results, depending on the size of the target and template proteins. The server has almost entirely been implemented in perl. It is currently running on a 1.7 GHz PC with 1 GB of RAM. Each request is sent to a queue and the results are sent back via email. Therefore typically, one can expect the results after a few hours, assuming that the queue is not very loaded.

Output format and explanation

The results returned are the alignment of target and template (if specified or identified) in Consensus format and FASTA format, and the framework model and a full model of the target. A typical Consensus alignment output is shown in Figure 1.

Table 1. CAFASP3 results of the Consensus server

CAFASP3 targets		Consensus server	
Target	NT[a]	NP[b]	CA[c]
133	293	148	3.26
137	133	105	1.09
140_1	87	24	2.23
141	187	55	2.96
142	280	212	3.33
143_1	121	99	3.68
143_2	95	86	3.18
149_1	201	49	11.11
150	96	85	2.14
151	106	74	2.07
152	198	68	2.47
153	134	94	1.19
154_2	103	84	1.78
155	117	103	0.79
160	125	84	1.45
165	318	123	1.94
167	180	131	1.47
169	156	96	2.78
172_1	192	34	4.2
176	100	68	4.51
177_1	57	55	1.4
177_2	88	86	1.57
177_3	75	68	1.79
178	219	170	1.48
179_1	56	50	0.8
179_2	218	211	3.06
182	249	229	1.01
183	247	185	1.31
184_2	72	46	5.24
185_1	101	47	2.22
185_2	197	134	2
185_3	130	23	1.97
186_1	77	35	0.77
186_2	250	36	4.69
188	107	68	1.62
189	319	70	3.03
190	111	97	2.09
191_2	143	95	3.88
192	170	37	1.33
195	290	130	2.8
Mean length	160	92.3	–
RMSD per al. res.	–	–	2.36

[a]NT—total number of residues.
[b]NP—number of aligned residues.
[c]CA—C-alpha RMSD from the X-ray structure of the target.

The consensus strength and selection

The reliability of each position in the alignment is assigned a score from 0 to 9. Higher score indicates a greater reliability of the alignment at that position. However, even a score of 9 does not necessarily mean that it is safe to model that region from the template. The score, along with other structural factors are used to select the regions reliable for modeling, indicated with an 'S' in the output. The term 'reliable' in this context implies that the model generated using this selection is expected to be within 3 Å RMSD from the native structure of the target.

3D Models

Two types of models of the target are generated (Figure 2). The full model is based on the entire alignment that contains the target residues which have a template residue aligned to them. The model is not necessarily full length, but it is the maximum that can be modeled from the template without resorting to any *ab initio* prediction. The framework model has only the regions that were selected by the server to be accurate for modeling. It does not contain any uncertain regions. Users may choose to model the uncertain regions on their own according to their requirements and resources. For example, the regions not present in the Consensus partial models can be visualy inspected and reassesed manually.

SUMMARY

The Consensus server consistently predicts high-quality target–template alignments even at low sequence identity. A measure of reliability is reported for each position in the sequence, which is useful both for accurate model building, and for assigning a measure of confidence to the different parts of the model. The server also predicts the regions of the alignment that are safe to model based on the template backbone. Consensus is 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

1. Prasad,J.C., Comeau,S.R., Vajda,S. and Camacho,C.J. (2003) Consensus alignment for reliable framework prediction in homology modeling. *Bioinformatics*, **19**, 1682–1691.
2. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
3. Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov Models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
4. Cuff,J.A. and Barton,G.J. (2000) Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. *Prot. Struct. Funct. Genet.*, **40**, 502–511.
5. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.
6. 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.
7. Fischer,D., Rychlewski,L., Dunbrack,R.L., Ortiz,A.R. and Elofsson,A. (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. *Prot. Struct. Funct. Genet.*, **53**, 503–516.
8. Zemla,A., Venclovas,C., Moult,J. and Fidelis,K. (2001) Processing and evaluation of predictions in CASP4. *Prot. Struct. Funct. Genet.*, **45**, (Suppl. 5), 13–21.
9. Carugo,O. and Pongor,S. (2001) A normalized root-mean-square distance for comparing protein three-dimensional structures. *Prot. Sci.*, **10**, 1470–1473.
10. 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.