BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 7 2012, pages 1038–1039
doi:10.1093/bioinformatics/bts070

Structural bioinformatics

Advance Access publication February 4, 2012

Improving your target-template alignment with MODalign

Alessandro Barbato[1], Pascal Benkert[2,3], Torsten Schwede[2,3], Anna Tramontano[1,4] and Jan Kosinski[1,*]

[1]Department of Physics, Sapienza University P.le A. Moro, 5, 00185 Rome, Italy, [2]Biozentrum, University of Basel, [3]SIB Swiss Institute of Bioinformatics, Basel, Switzerland and [4]Center for Life Nano Science @Sapienza, Istituto Italiano di Tecnologia, Sapienza University, P.le A. Moro 5, 00185 Rome, Italy

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: MODalign is an interactive web-based tool aimed at helping protein structure modelers to inspect and manually modify the alignment between the sequences of a target protein and of its template(s). It interactively computes, displays and, upon modification of the target-template alignment, updates the multiple sequence alignments of the two protein families, their conservation score, secondary structure and solvent accessibility values, and local quality scores of the implied three-dimensional model(s). Although it has been designed to simplify the target-template alignment step in modeling, it is suitable for all cases where a sequence alignment needs to be inspected in the context of other biological information.

Availability and implementation: Freely available on the web at http://modorama.biocomputing.it/modalign. Website implemented in HTML and JavaScript with all major browsers supported.

Contact: jan.kosinski@uniroma1.it

Received on November 2, 2011; revised on January 26, 2012; accepted on February 1, 2012

1 INTRODUCTION

Protein sequence alignment is a key step in most, if not all, applications of protein bioinformatics. Evolutionary analysis, functional assignment and comparative modeling projects all heavily rely on an accurate sequence alignment. The alignment plays a particularly pivotal role in comparative modeling: if the target-template alignment contains errors, it is extremely difficult to correct them in either the model building or the refinement stage.

In the case of difficult comparative modeling tasks, there can be several regions that have structurally diverged between the target and the template(s). To optimize the alignment in such regions, expert modelers take into account as much information as possible. This includes sequence conservation patterns in multiple sequence alignments of the target and template evolutionary families, secondary structure states, and often their expert biological knowledge. Moreover, to find potential errors in the alignment, several models based on different alignment versions are usually built and analyzed in terms of parameters such as unfavorable burial of charged residues. Combining all this information during alignment optimization is often difficult and time-consuming and without thorough inspection of the sequence and structural parameters can lead to unspotted errors and inaccurate models.

*To whom correspondence should be addressed.

Here we describe an interactive web-based alignment editor that makes alignment optimization simpler. It works as a dashboard for computing, displaying, inspecting and updating alignment related information in real time. The tool also automatically builds models according to the alignment that is being inspected (without modeling insertions) and allows the quantitative assessment of their global and local quality.

Finally, the edited alignment can be directly used for building a full length three-dimensional model using Modeller (Sali and Blundell, 1993).

2 TOOL DESCRIPTION

Given a target and template alignment (multiple templates can be used), the system performs several operations:

- Builds a multiple sequence alignment of representative sequences for both the target and the template families. For the target, this step is performed using two iterations of hhblits (Remmert *et al*., 2012) with default parameters on the UniProt20 database. For templates, the alignments are derived from the HHSearch (Söding, 2005) alignment database. The output alignments are subsequently filtered using the hhfilter program from the HHSearch package;
- Aligns the input template sequence(s) with the sequences derived from the SEQRES and ATOM fields of the corresponding PDB entry(ies);
- Displays the sequence alignment of target, template(s) and their families in the interactive editor interface;
- Highlights residues with no coordinates in the template structure (by showing the residues in lower case);
- Graphically depicts sequence conservation for each column of the alignment as background shading;
- Displays secondary structure and solvent accessibility values. Values for the target are predicted using PSI-PRED (Jones, 1999) and ACCpro (Cheng *et al*., 2005), respectively. The values for the templates are calculated using DSSP (Kabsch and Sander, 1983) and POPS (Fraternali and Cavallo, 2002);
- Upon request, highlights potential errors in the alignment, such as: (i) insertions or deletions within secondary structure elements; (ii) cases where, a hydrophobic or charged residue

in the target is aligned to an exposed or buried residue in the template, respectively;

- Upon request, computes and displays the QMEAN (Benkert *et al*., 2011) global and local scores of the model implied by the current alignment (without modeling insertions);
- Performs all the above computations (including QMEAN) also for representative target homologs, giving an estimate of how the whole target family 'fits' to the selected template;
- Allows for editing operations such as residue shifts and insertions in either the target or the template. Importantly, the system automatically introduces the changes in all members of the family of the protein and recomputes all the data described above;
- Submits the alignment to Modeller for model building;
- Allows export of the alignment in FASTA or PIR format;
- Provides visualization of templates and model structures in Jmol (http://www.jmol.org/) also mapping the positions of insertions and deletions in the current target-template alignment on the template structures.

The editor interface is composed of several sections showing the above information and users can choose which sections to display. Options for saving and modifying the appearance of the results, including using different amino acid color schemes, are provided.

A snapshot of the tool is shown in Figure 1.

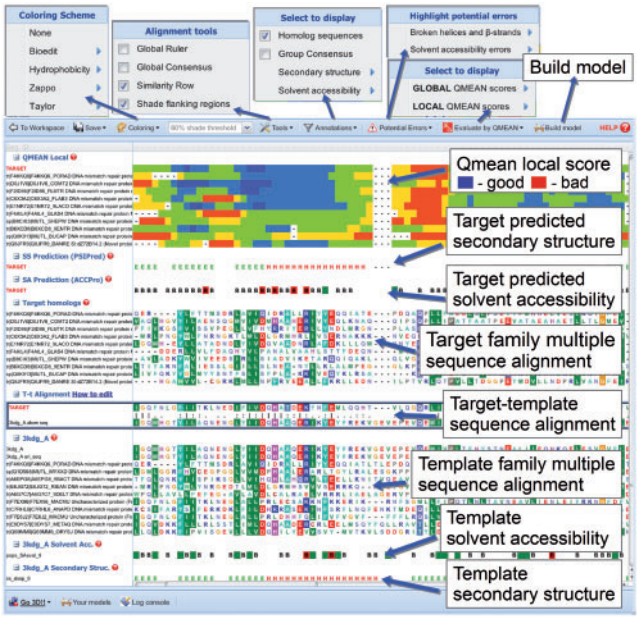

Fig. 1. The MODalign result page. The alignment can be edited in the target-template alignment section. When the alignment is modified, the output, such as coloring by sequence conservation or highlighting of potential errors, is modified in real time, while QMEAN can be recalculated on request.

3 IMPLEMENTATION

The server is built in Python (using Django web framework), HTML and JavaScript (using ExtJS library), and utilizes PyCogent (Knight *et al*., 2007).

4 CONCLUSIONS

While many useful automatic servers to build comparative models of proteins exist, it is still the case that a careful inspection of the sequence alignment makes a difference in the final model quality. Very often, papers reporting the results of a non-trivial comparative modeling experiment mention that a manual modification of the alignment was required before performing the model-building step.

In this article, we describe a tool that we believe will be of great help in these cases for both expert and novice users.

We also believe that the usability of its interface will make MODalign a useful tool for the inspection of an alignment also by scientists who do not want to exploit its editing capabilities. For example, it may be useful for inspecting their proteins of interest in the context of the alignment of their evolutionary families, secondary structure, solvent accessibility states and the likelihood of implied residue–residue interactions (as computed by QMEAN).

ACKNOWLEDGEMENTS

The authors would like to thank all members of the Biocomputing Group for fruitful discussions as well as members of Torsten Schwede's Structural Bioinformatics Group, in particular Valerio Mariani and Marco Biasini for their help with the integration of the QMEAN software.

Funding: KAUST Award No. KUK-I1-012-43 made by King Abdullah University of Science and Technology (KAUST), Fondazione Roma, the Italian Ministry of Health, Contract No. onc_ord 25/07, FIRB PROTEOMICA, and European Molecular Biology Organization (EMBO) long-term fellowship to J.K.

REFERENCES

Benkert,P. *et al*. (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, **27**, 343–350.

Cheng,J. *et al*. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, W72–W76.

Fraternali,F. and Cavallo,L. (2002) Parameter optimized surfaces (POPS): analysis of key interactions and conformational changes in the ribosome. *Nucleic Acids Res.*, **30**, 2950–2960.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Knight,R. *et al*. (2007) PyCogent: a toolkit for making sense from sequence. *Genome Biol.*, **8**, R171.

Remmert,M. *et al*. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Meth.*, doi:10.1038/nmeth.1818.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.