

An alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs

Jian Qiu¹

Keywords: alignment accuracy, secondary structure, relative solvent accessibility, structure-dependent gap penalties, structurally aligned protein pairs

1 Introduction and Method.

The growth rate of experimentally determined protein structures falls far behind that of protein sequences. Protein structure prediction with theoretical methods is of both intellectual interest and practical importance. If a protein structure exists in the PDB that shares significant structural similarity with the target protein sequence, a structural model of the target sequence can be constructed using its structural homologue as a template. Template-based modeling of protein structures generally involves three steps: template identification, generation of an alignment between the target sequence and the template structure, and model construction based on the template structure and the alignment. The accuracy of the alignment generated in the second step significantly affects the quality of the final model.

In this study, we propose an alignment algorithm SSALN that learns substitution matrices and structure-dependent gap penalties from a database of structurally aligned protein pairs. The substitution scoring matrix is a linear combination of three component matrices: 1) amino acid types substitution matrix; 2) amino acid type of a target position *vs.* the secondary structure and solvent accessibility information of a template position; 3) the predicted secondary structure and solvent accessibility information of a target position *vs.* the actual secondary structure and solvent accessibility information of a template position. Secondary structure and solvent accessibility information of a position are also used to derive position-specific gap penalties. Secondary structure and solvent accessibility information of a target sequence are predicted with program SABLE [1], and those of a template structure are computed with program DSSP [2].

2 Results.

In a test set of CASP5 targets, our method SSALN outperforms sequence alignment methods such as a Smith-Waterman algorithm with BLOSUM50 and PSI_BLAST (table 1). SSALN is also compared with several threading methods and sequence alignment methods on the ProSup benchmark (table 2) [3]. SSALN has the highest alignment accuracy among the methods compared. In CASP6 (<http://predictioncenter2.llnl.gov/casp/casp6/public/cgi-bin/results.cgi>), LOOPP server (<http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm>) predicted a model ranked the best for target T0280_1 based on an SSALN alignment. All the structure-dependent substitution matrices and gap penalties can be found at <http://www.cs.cornell.edu/~jianq/research.htm>.

¹ Department of Computer Science, 4130 Upson Hall, Cornell University, Ithaca, NY 14850, USA E-mail: jiangq@cs.cornell.edu

3 Figures and tables.

Method	Number of correctly aligned positions	Fraction of correctly aligned positions(%)	Average shift
Blosum50	5315	25.3	23.0
PSI_BLAST	7096	32.4	12.1
SSALN	10416	46.7	6.5

Table 1: the alignment accuracy of the CASP5 test set

Method	Sequence ^a	Threading ^a	PSI_BLAST ^b	Stroma ^b	SPARKS ^b	SSALN ^c
$\sigma_0(\%)$	34.1	48.0	35.6	36.1	57.2	58.3

Table 2: Fraction of Correctly Aligned Positions for ProSup Benchmark

a: Results from [3].

b: Results from [4].

c. This study.

4 References and bibliography.

- [1] Adamczak, R., Porollo, A., and Meller, J. 2004. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* **56**, 753-67.
- [2] Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637.
- [3] Domingues, F. S., Lackner, P., Andreeva, A., and Sippl, M. J. 2000. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J Mol Biol* **297**, 1003-13.
- [4] Zhou, H., and Zhou, Y. 2004. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* **55**, 1005-13.