INFORMATION ABOUT SECONDARY STRUCTURE IMPROVES QUALITY OF PROTEIN ALIGNMENT

Litvinov I.I.*[1], Mironov A.A.[2], Finkelstein A.V.[3], Roytberg M.A.*[1]

[1] Institute of Mathematical Problems in Biology RAS, Puschino, Russia; [2] Moscow State University, Department of Biotechnology and Bioinformatics, Moscow, Russia; [3] Institute of Protein Research RAS, Puschino, Russia
* Corresponding authors: e-mail: litvinov@mail.ru, Roytberg@impb.psn.ru

Keywords: *protein sequence alignment, secondary structure prediction*

Summary

Motivation: The Smith-Waterman (SW) alignment algorithm is known as the most accurate algorithm for pair wise alignment of amino acid sequences.. It means, that SW alignments are more similar to alignments of corresponding 3D-structures, than FASTA, BLAST, etc. alignments. But even SW algorithm is unable to restore alignment of proteins' 3D-structures if the sequence identity is less than 30 % ("twilight zone"). Our goal is to design a new alignment method, which is significantly more accurate, than SW algorithm.

Results: We propose to modify SW alignment score to take into account protein secondary structure. We give bonus for alignment of residues belonging to the regions of same secondary structure type. We have shown that alignments maximizing the improved score are much more accurate, than SW alignments (57 % accuracy vs. 31 % for the twilight zone sequence identity; both experimentally determined and theoretically predicted secondary structure can be used). The dynamic programming algorithm to find the optimal secondary structure alignment was designed and implemented as C++ program STRUSWER.

Availability: Program STRUSWER is available on request from the authors.

Introduction

Alignment of amino acid sequences is the core of many modern bioinformatics methods, e.g. homology based 3D-modeling of proteins, database search, etc. The algorithmically produced protein alignments usually differ from the "structure" alignments, i.e. alignments obtained by superposition of 3D-structures. The latter can be considered as "golden standard" ones because of two reasons. First, protein 3D-structures is much more conservative, than their amino acid sequences, and therefore structure alignments are more close to evolutionary based alignments. Second, for homology based 3D-modeling of proteins and many other applications, it is important to restore 3D-structure alignments.

It is known, that Smith-Waterman (SW) alignment algorithm [6] creates alignments which are most similar to the structure alignments compared to alignments produced by other popular alignment algorithms, e.g. BLAST [1, 2], FASTA [5], etc. But even SW algorithm is unable to restore alignment of proteins' 3D-structures if the sequence identity is less than 30 % [7].

To improve the quality of SW algorithm we propose to take into account proteins' secondary structure. Both experimentally determined and theoretically predicted structures can be used. The predictions can be done both in "strict form", i.e. ascribing each residue with mark H (for helix); E (for beta-strand) and L (loop), and in "propensity form", i.e. ascribing each residue with probabiliities to belong to a helix, strand or loop. The program STRUSWER, which implements the approach have shown up significantly better quality than standard implementation of SW algorithm.

Materials and Methods

Structure alignments. We use manually verified structure alignments from the BAliBase [8] protein structure database, as a source of "golden standard" alignments. BAliBase contains manually curated multiple alignments, initially based on 3D protein structures. We have used alignments from BAliBase Reference 1, the sequence identity level for the Reference is mainly 10–50 %. The test set was consisted of all protein pairs meeting following condition: both proteins belong to the same multiple alignment of BAliBase's Reference 1 and their 3D-structures are known.

Secondary structure. Experimentally determined structure was obtained from database DSSP [4]. The 8-state DSSP description of a structure was converted to 3-state structure form using following rules:

(H,G,I)->H; (E,B)->E; (T,S)->L,

where H = alpha helix; B = residue in isolated beta-bridge; E = extended strand, participates in beta ladder; G = 3-helix (3/10 helix) ; I = 5 helix (pi helix); T = hydrogen bonded turn ; S = bend.

To predict secondary structures we have used the PsiPred [3] program in its full (using homology search technique), and restricted (using only amino acid itself, AKA "PsiPred single") versions. We have used two forms of secondary structure representation. First ("strict form"): each residue R_i is ascribed with a mark $T(R_i)$; the possible values for T_i :are 'H' (for a helix); 'E' (for a beta-strand) and 'L' (for a loop). This form is also used to represent the experimentally determined secondary structures. Second, ("propensity form"): each residue. R_i is ascribed with a probabilities $P(R_{i,}, T)$; that R_i belongs to a secondary structure of type T, i.e. a helix, a strand or a loop. Below we use following abbreviations for the above types of secondary structure assignments: a) Exp – Experimentally determined structure; b) Psi – the structure, predicted by the full version of PsiPred; strict representation of secondary structure is used; c) PsiPro – the structure, predicted by the full version of PsiPred; propensity representation of secondary structure is used; d) PPS – the structure, predicted by the restricted version of PsiPred; strict representation of secondary structure is used; e) PPSPro – the structure, predicted by the restricted version of PsiPred; propensity representation of secondary structure is used.

Alignment score and alignment algorithm. Original Smith-Waterman alignment score [8] was modified as follows. We define score $W[a_i, b_j]$ of matching of residues a_i from the sequence A and b_j from the sequence B is defined as:

$$W(a_i, b_j) = M(a_i, b_j) + SBON*Q(a_i, b_j),$$

where $M(a_i, b_j)$ is a substitution score given by substitution matrix (e.g. blosum62), SBON is a weighting parameter and $S(a_i, b_j)$ reflects similarity of secondary structure marks. If the strict form of secondary structure description is used, then

$$Q(a_i, b_j) = 1 \text{ if } (T(a_i) = T(b_j) = \text{'H'}) \text{ or } (T(a_i) = T(b_j) = \text{'E'})$$
$$= 0 \text{ otherwise}$$

If propensity form is in use,

$$Q(a_i, b_j) = Hp1(a_i)*Hp2(b_j)+Ep1(a_i)*Ep2(b_j),$$

where Hp1, Hp2 are helix probabilities for 1st and 2nd sequences respectively; Ep1, Ep2 are strand probabilities for 1st and 2nd sequences. The "structure SW" score of alignment differs from the SW one only in one point: it uses the value $W(a_i, b_j)$ where SW score uses substitution score $M(a_i, b_j)$. Analogously, the STRUSWER algorithm, that finds an optimal alignment with respect to structure SW score can be obtained from the SW algorithm by calculating $W(a_i, b_j)$, where SW-algorithm calculates $M(a_i, b_j)$.

Alignment quality estimations. To compare algorithmic alignments with golden standard ones and to estimate quality of algorithmic alignment we use two characteristics, *accuracy* and *confidence.*

Alignment accuracy is the number I of positions *Identically* superimposed in algorithmic and golden standard alignment (GS) divided by the total number G of positions in the GS alignment.

$$Acc = I / G.$$

Alignment confidence is the number I of positions *Identically* superimposed in algorithmic and golden standard alignment divided by the total number A of aligned positions in the Algorithmic alignment

$$Conf = I/A.$$

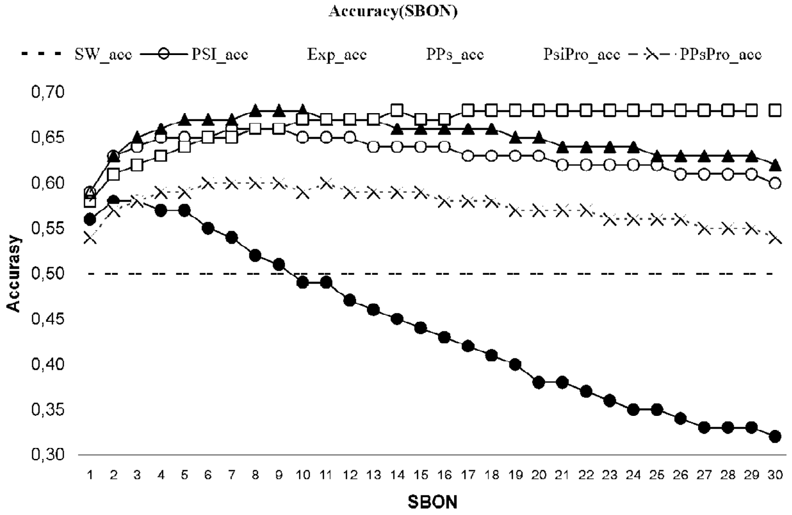

Fig. SBON parameter optimization. Methods abbreviations and description see in "Secondary structure" chapter. Each point represents average accuracy (given by Y-axes) of alignments obtained with a given SBON value (X-axes) and type of secondary structure assignment.

Results and Discussion

For every pair of proteins from the test set, we have constructed six alignments: the Smith-Waterman alignment (SW); and STRUSWER alignments with five types of secondary structure assignment (see "*Secondary structure*" above). In all cases we have used standard SW parameters, i.e. substitution matrix BLOSUM62, Gap Opening Penalty (GOP) 11, Gap Elongation Penalty (GEP) 1. To find optimal value of the parameter SBON we have tried all values from 1 to 30; and calculated alignment accuracy for all obtained alignments. The results are given in Fig. Table shows the average values of alignment accuracy and confidence, obtained for optimal values of SBON. The data are shown both for all data set, and for twilight zone only. One can see, that both experimental structures, and predicted by full version of PsiPred, allow significantly improve alignment accuracy without sacrificing with alignment confidence.

Table. Accuracy and confidence of various alignment methods (see notation in "*Secondary structure*"). The data are given for full test set (576 protein pairs) and for twilight zone only (protein pairs with sequence identity below 30 %, 368 pairs)

Method SBON	*SW* -	*Exp* 10	*PsiPro* 14	*Psi* 8	*PPsPro* 9	*PPs* 2
Full set (576 pairs) Acc	0,5	0,68	0,68	0,66	0,6	0,58
Full set (576 pairs) Conf	0,64	0,7	0,69	0,68	0,61	0,62
ID <30 (368 pairs) Acc	0,31	0,57	0,57	0,55	0,46	0,43
ID <30 (368 pairs) Conf	0,51	0,6	0,58	0,57	0,47	0,48

Aknowledgements

The work has been supported by the Russian Foundation for Basic Research (project Nos. 03-04-49469, 02-07-90412) and by grants from the RF Ministry for Industry, Science, and Technology (20/2002, 5/2003) and NWO.

References

1. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool // Mol. Biol. 1990. V. 215. P. 403–410.
2. Altschul S.F., Madden T.L., Schäffer 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 Acids Res. 1997. V. 25. P. 3389–3402.
3. Jones D.T. Protein secondary structure prediction based on position-specific scoring matrices // J. Mol. Biol. 1999. V. 292. P. 195–202.
4. KabaschW., Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features // Biopolymers. 1983. V. 2(12). P. 2577–637.
5. Pearson W.R. Effective protein sequence comparison // Methods Enzymol. 1996. V. 266. P. 227–58.
6. Smith T.F., Waterman M.S. Identification of common molecular subsequences // J. Mol. Biol. 1981. V. 147. P. 195–197.
7. Sunyaev S.R., Bogopolsky G.A., Oleynikova N.V., Vlasov P.K., Finkelstein A.V., Roytberg M.A. From analysis of protein structural alignments toward a novel approach to align protein sequences // PROTEINS: Structure, Function, and Genetics. 2004. V. 54(3). P. 569–582.
8. Thompson J., Plewniak F., Poch O. BAliBASE: A benchmark alignments database for the evaluation of multiple sequence alignment programs // Bioinformatics. 1999. V. 15. P. 87–88.