OPTIMIZATION OF ACCURACY AND CONFIDENCE FOR ALIGNMENT ALGORITHMS EXPLOITING DATA ON SECONDARY STRUCTURE

***Litvinov I.I.*[*1, 3], *Finkelshtein A.V.*[2], *Roytberg M.A.*[*1, 3]**
[1] Institute of Mathematical Problems in Biology, RAS, Pushchino, Moscow Region, 142290, Russia;
[2] Institute of Protein Research, RAS, Pushchino, Moscow Region, 142290, Russia; [3] Pushchino State University, Pushchino, Moscow Region, 142290, Russia
[*] Corresponding authors: e-mail: mroytberg@mail.ru

Key words: protein sequence alignment, secondary structure prediction, alignment quality, accuracy, confidence

SUMMARY

Motivation: The quality of protein sequences alignment is a similarity between the alignment and the "golden standard" alignment reflecting the evolutionary history. The quality of algorithmically obtained alignment is crucial for many bioinformatics tasks.

Two main measures of alignment quality are *accuracy*, i.e. part of correctly restored positions of the golden standard alignment, and *confidence*, i.e. part of positions of the algorithmic alignments that belong to the golden standard alignment. The measures often are contradictory, i.e. the parameters optimizing one of the measures can result in low values of another.

Results: We have performed detailed investigation of accuracy and confidence of alignments obtained by different methods with different values of parameters. It was shown that the methods exploiting information about the secondary structure admit the simultaneous optimization of alignment accuracy and confidence with the same parameters values. This contrasts with the behavior of alignment accuracy/confidence for classic Smith-Waterman method.

INTRODUCTION

Pair-wise alignment of amino acid sequences is a core of many bioinformatics methods. The ideal goal of all alignment algorithms is to find a biologically correct alignment reflecting the evolutionary history of homologous proteins (Sunyaev *et al*., 2004); i.e. aligned positions have to correspond to the same position of their common ancestor. The "quality" of an algorithmic alignment of amino acid sequences (i.e., its similarity to the biologically correct alignment) is critical for many applications, e.g. homology modeling, database homology search, protein domains analysis, etc. Biologically correct alignment is unknown, thus to measure the alignment quality one has to use an approximation of the biologically correct alignment as the "golden standard". Since the tertiary structure of proteins is much more conservative than their sequences, we use the alignments obtained by superimposing the protein spatial structures as a "golden standard". Alignment quality can be described by two complementary measures: *accuracy* (a number of identically aligned positions in algorithmic and reference alignment divided by total number of positions aligned in algorithmic alignment) and *confidence* (a number of identically aligned positions in the algorithmic and reference alignment divided by total number of positions aligned in the reference alignment).

The quality of algorithmic alignments crucially depends on the similarity of the sequences to be compared. For instance, the accuracy of the Smith-Waterman (SW) algorithm is 84 % when the protein identity (i.e., the portion of identical positions in two proteins) is no less than 30 %; and if the identity is below 30 %, the alignment accuracy is about 30 % (Sunyaev *et al*., 2004). The rapid approximate alignment algorithms, such as BLAST and FASTA, are even less accurate. To improve the accuracy of algorithmic alignments one can use combined methods taking into account both sequences and the (predicted) secondary structures. E.g. we have proposed the method STRUSWER (Litvinov *et al*., 2006) algorithm, which utilizes an additional bonus for matching identical elements of secondary structures; secondary structures can be determined experimentally or theoretically. Another method of this type is the Wallqvist-Fukunishi-Murphy-Fadel-Levy algorithm (WFMFL) (Wallqvist *et al*., 2000).

The optimal values of parameters depending on the protein sequence identity were found for all above algorithms. However, the two measures of alignment quality usually lead to different values of parameters. E.g. it is common knowledge that the alignment confidence is more essential for the database search, but the parameter values optimizing the confidence results in very low values of the accuracy.

The aim of the presented work was detailed investigation of the dependence of alignment quality on the algorithm parameters. We show that unlike the classic alignment algorithms (Smith-Waterman, etc) the secondary structure based methods allow simultaneous optimization of accuracy and confidence.

MATERIALS AND METHODS

Secondary structure. To predict the secondary structure we have used the PSIPRED program (Jones, 1999). The data presented below were obtained with the full version (prediction based on preliminary homology search) and the deterministic representation of the prediction (each residue is assigned with one of three letters: H (helix), E (beta) and L (loop). The other modes of the PSIPRED program as well as usage of experimentally obtained secondary structures from the DSSP database lead to the similar results.

Golden standard alignments. As a golden standard, we used manually verified structure alignments from the BAliBase (Bahr *et al*., 2001) protein structure database, as a source of "golden standard" alignments. 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.

Evaluation of the alignment quality. To compare two alignments (algorithmic and golden standard ones) and to estimate the agreement between them, we used two measures, accuracy and confidence. The alignment accuracy (Acc) was defined as a ratio of the number of positions (I) aligned similarly in the reference and algorithmic alignments to the number of aligned positions in the reference alignment (G): Acc = I/G.

The alignment confidence (Conf) was defined as a ratio of the number of positions aligned similarly in the reference and algorithmic alignments to the number of aligned positions in the algorithmic alignment (A): Conf = I/A.

Alignment algorithms utilizing the secondary structure data. We have tested two such algorithms, our algorithm STRUSWER (Litvinov *et al*., 2006) and the algorithm of Wallqvist-Fukunishi-Murphy-Fadel-Levy (WFMFL) (Wallqvist *et al*., 2000). STRUSWER is a modification of the Smith-Waterman (SW) algorithm. The only difference is that the score of the matching of i-th amino acid residue of one sequence with the jth residue of the other involves an extra summand SBON*SS[i, j], where SBON is a parameter of the algorithm and SS[i, j] = 1 if the residues are assigned with the same secondary structure type and the type is H or E; otherwise and SS[i, j] = 0. SW algorithm corresponds to SBON = 0.

The WFMFL algorithm modifies the Smith-Waterman algorithm in a similar way, but the extra summand is determined by the predefined 3 × 3 matrix depending on secondary structure types of compared residue (Wallqvist *et al*., 2000).

Optimization of the parameters of the program. Three algorithms were run for each pair of proteins from BAliBASE set and for each set of parameters: (1) SW algorithm (secondary structure disregarded, i.e. SBON = 0); (2) STRUSWER algorithm with the secondary structure predicted using the PSIPRED program; (3) WFMFL alignment with the secondary structure predicted using the PSIPRED program. Each algorithm was implemented with different values of parameters; the following integer values of parameters were checked: Gap Opening Penalty (GOP): from 4 to 20, Gap elongation penalty (GEP) from 1 to 7; SBON: from 1 to 30, GOP from 4 to 20, and GEP from 1 to 7. Thus, for each protein pair we have constructed 17 × 7 of SW and WFMFL alignments and 30 × 17 × 7 STRUSWER alignments (parameter SBON is applicable only for STRUSWER). Each of the algorithmic alignments was compared with the corresponding golden standard alignment, to obtain its accuracy and confidence. Finally, the results obtained for all protein pairs were averaged to yield average values <Acc> = <I/G> and <Conf> = <I/A> for a given algorithm and a set of parameters.

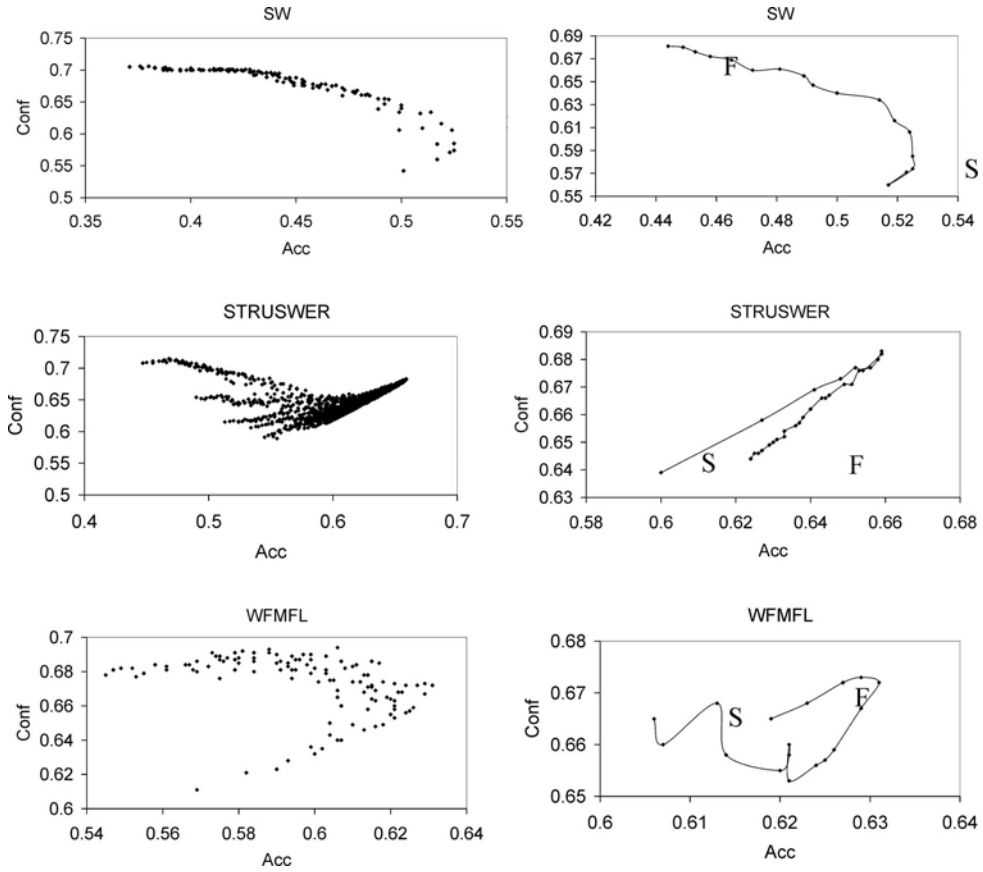

Figure 1. Full dataset. Accuracy/Confidence scatter-plots (left) and "trajectory" plots (right) for each method. Each point of the scatter-plot corresponds to a set of parameters GOP, GEP and SBON (the last for STRUSWER only). Each point on "trajectory" plot corresponds to a value of main parameter (SBON for STRUSWER; GOP for SW and WFMFL). Other parameters are chosen to optimize the average accuracy. Start (the lowest value of the main parameter) and finish (the greatest value) are marked with "S" and "F" respectively. X-axes presents the average accuracy and Y-axes presents the average confidence for a parameter set (see Materials and Methods).

RESULTS AND DISCUSSION

In our previous work (Litvinov *et al*., 2006) we have shown that methods using information about secondary structure provide essentially more accurate alignments than the Smith-Waterman algorithm. This advantage is the more valuable the lower is the sequence identity (see Fig. 2). Here we present the detailed investigation of the dependence of accuracy and confidence of alignments on the parameters of alignment algorithms. The most striking result is possibility to achieve both maximal accuracy and almost maximal confidence for the same values of STRUSWER parameters (see Fig. 1). The maximal value of confidence (0.7) corresponds to "strong" values GOP = 17; GEP = 6, SBON = 1 but it provides very low accuracy (0.47). Fortunately, "weak" parameters providing maximal value of accuracy (SBON = 8; GOP = 9; GEP = 1) correspond to the almost maximal value of the confidence (0.683 compared to 0.707). Fig. 1 (middle-right) shows how the accuracy depends on the SBON parameter. The method WFMFL (Wallqvist *et al*., 2000) demonstrate the similar behavior related to parameters GOP/GEP (see Fig. 1, bottom).

The optimal values of parameters (leading to acc= 0.63; conf = 0.67) are those maximizing accuracy and they coincide with the values recommended in (Wallqvist *et al*., 2000). In contrast the Smith-Waterman method that has only two parameters, does not allow simultaneous optimization of accuracy and confidence (see Fig. 1, top). The behavior of the algorithms' accuracy/confidence is essentially the same if we restrict ourselves with low-homology protein pairs (see Fig. 2). The optimal parameter values are almost the same.

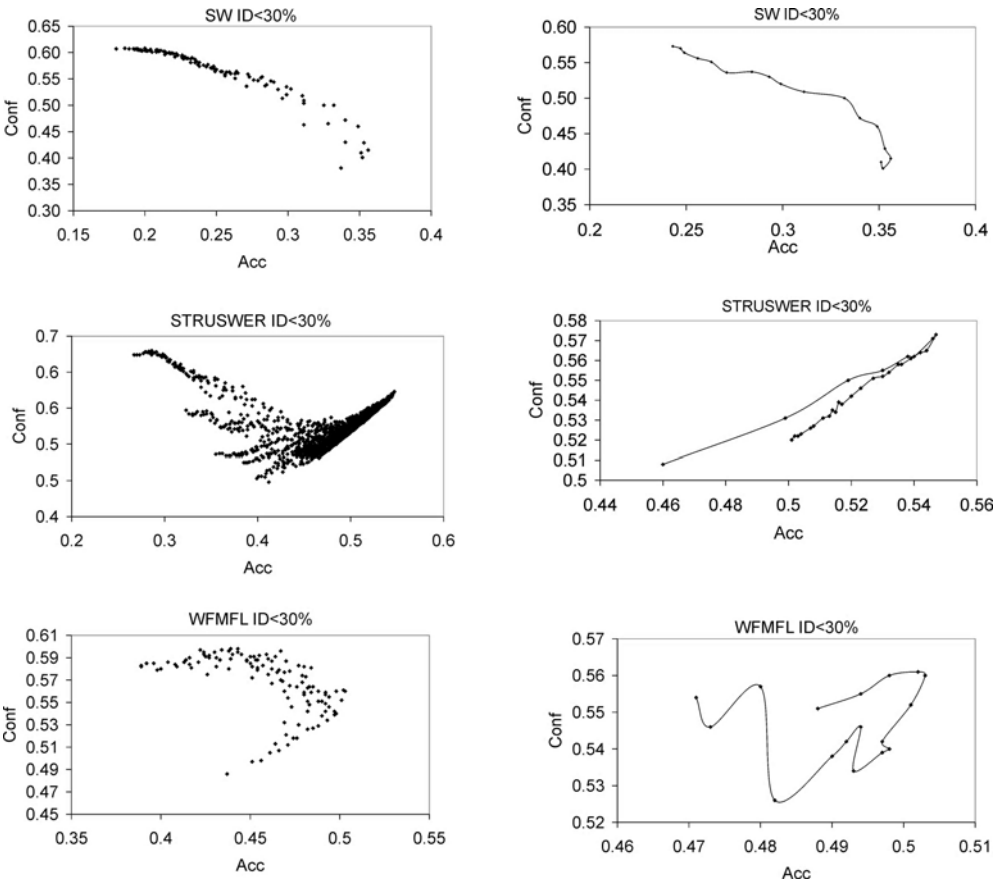

Figure 2. Accuracy/Confidence plots (see Fig. 1) for proteins with sequence similarity less than 30 %.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (project Nos 03-04-49469, 02-07-90412), by grant from the RF Ministry for Industry, Science, and Technology (20/2002, 5/2003), NWO, ECO-NET, and by the program of RF Ministry of Science and Education (contract No. 02.434.11.1008).

REFERENCES

Bahr A., Thompson J.D., Thierry J.-C., Poch O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucl. Acids Res.*, **29**, 323–326.

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

Litvinov I.I., Lobanov M. Yu., Mironov A.A., Finkelstein A.V., Roytberg M.A. (2006) Information about secondary structure improves quality of protein alignment. *Mol. Biol*. In press.

Sunyaev S.R., Bogopolsky G.A., Oleynikova N.V., Vlasov P.K., Finkelstein A.V., Roytberg M.A. (2004) From analysis of protein structural alignments toward a novel approach to align protein sequences. *Proteins*, **54**, 569–582.

Wallqvist A., Fukunishi Y., Murphy L.R., Fadel A., Levy R.M. (2000) Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics*, **16**, 988–1002.