BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 3 2003, pages 425–426
DOI: 10.1093/bioinformatics/btf882

AltAVisT: Comparing alternative multiple sequence alignments

Burkhard Morgenstern[1,*], Sachin Goel[1], Alexander Sczyrba[2] and Andreas Dress[3]

[1]International Graduate School in Bioinformatics and Genome Research, [2]Faculty of Technology, Research Group in Practical Computer Science and [3]Department of Mathematics, University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany

Received on July 18, 2002; revised on September 18, 2002; accepted on September 28, 2002

ABSTRACT

Summary: We introduce a WWW-based tool that is able to compare two alternative multiple alignments of a given sequence set. Regions where both alignments coincide are color-coded to visualize the local agreement between the two alignments and to identify those regions that can be considered to be reliably aligned.

Availibility: http://bibiserv.techfak.uni-bielefeld.de/altavist/

Contact: burkhard@TechFak.Uni-Bielefeld.DE

Sequence alignment is the most fundamental tool for sequence data analysis in molecular biololgy. Practically all methods of computational sequence analysis rely in one way or the other on sequence comparison, so their results depend on the quality of the underlying alignments. Pairwise and multiple alignment therefore continue to be among the most active areas of bioinformatics research. There are two major challenges in the context of sequence alignment: (a) it can be hard to distinguish weak local homologies from random similarities and (b) alignment programs can only detect those homologies that appear in the *same relative order* in the input sequences. The latter problem is inherent in sequence alignment and means that, for many data sets, correct alignment of one homologous region necessarily prevents other homologies from being properly aligned.

No single alignment procedure can be expected to construct biologically reasonable alignments in all possible situations. The reason for this is that every alignment program tries (explicitly or implicitly) to find optimal alignments according to some relatively simple mathematical scoring function. Yet it cannot be expected that any given scoring function will, under all conditions, be in accordance with biology giving the *mathematically* highest score to the *biologically* correct alignments. Consequently, human intervention is often necessary to check the results of automated alignment procedures and to obtain meaningful alignments.

A popular way of testing the (local) reliability of pairwise or multiple alignments is to construct *alternative* alignments of the same sequences using different alignment methods. Notredame *et al.* (2000) used this idea systematically and developed a software tool that integrates results from different multi-alignment methods into one single output alignment.

For multiple alignment, a variety of software programs are now available that rely on very different objective functions and optimization techniques. The results of these methods can therefore be quite diverse, see Notredame (2002) for an excellent review of the state-of-the-art multi-alignment algorithms and Thompson *et al.* (1999b) and Lassmann and Sonnhammer (2002) for systematic evaluation of the corresponding software tools.

If two alignments have been constructed by different methods, those regions where both alignments *coincide* are generally considered to be more reliable than regions where the two methods disagree. However, manually comparing different multiple alignments is a tedious task. Herein, we introduce *AltAVisT* (*Alt*ernative *A*lignment *Vis*ualization *T*ool), a WWW-based tool that compares two different multiple alignments of a given data set and highlights regions where both alignments coincide. Two input options are available:

(1) It is possible to enter a family of *sequences*. In this case, our program runs DIALIGN (Morgenstern, 1999) and CLUSTAL W (Thompson *et al.*, 1994) on the input sequences and compares the resulting alignments to each other. These two methods are currently among the most popular multi-alignment tools. They rely on fundamentally different algorithmical approaches, so agreement between them should indicate (local) correctness of the alignments.

*To whom correspondence should be addressed.

Fig. 1. AltAVisT applied to a small test sequence set. The first alignment has been produced by DIALIGN, the second one by CLUSTAL. For each column in the first alignment, those residue pairs are cololred that also appear in one common column in the second alignment. Different colors are used to distinguish groups of residues where the alignment coincides *within* groups but not *between* different groups. For example, the two *M*s in column four in the DIALIGN alignment appear in a common column in the CLUSTAL alignment (column 21), they are therefore colored. The same holds true for the two *C*s in the same column of the DIALIGN alignment; they also appear in one column in the CLUSTAL alignment (column four). However, the *M*s and *C*s belong to different columns in the CLUSTAL alignment so different colors are used. All lower-case residues in the DIALIGN alignment are printed in black because they are not considered aligned by DIALIGN. In the second alignment, all residues have the same color as in the first alignment so the two alignments can be easily compared. This may imply, however, that residues in a column of the second alignment are in the same color even though they are not aligned together in the first alignment, see for example column 21 in the second alignment.

(2) It is possible to enter two different *pre-calculated* alignments of a sequence family set that may have been produced by any method; this way the user can compare the output of arbitrary alignment methods.

With the second option, it is possible to distinguish between upper-case and lower-case letters in the input alignments and to consider only upper-case letters for the alignment comparison. This can be used in situations where only subareas of an alignment are of interest; it corresponds to the output of DIALIGN where lower-case letters are not considered to be aligned. With either option, those residue pairs that are aligned to each other in *both* alignments are colored. Different colors are used to distinguish groups of residues where the alignments coincide *within* groups but not *between* different groups; see Figure 1 for an example. In other words, considering alignments as *consistent equivalence relations* as outlined in Morgenstern *et al.* (1996) and Abdeddaïm and Morgenstern (2001), residue pairs in the same column having the same color belong to the set-theoretical intersection of the two respective equivalence relations.

Our tool can not only be used to find (locally) reliable alignments but also to evaluate alignment programs by comparing their results to reference alignments that are seen as a *standard of truth*. There is now a high-quality benchmark data base called *BAliBASE* that has been designed to evaluate multiple alignment methods (Thompson *et al.*, 1999a); other benchmark data have been compiled by Lassmann and Sonnhammer (2002). The authors of BAliBASE also provide software that compares arbitrary alignments of their benchmark data to the corresponding reference alignments and determines the overal degree of agreement between these two alignments. However, for the development of alignment methods, it can be interesting to know not only the overal performance of a method but also to see where exactly the produced alignments are in agreement with biologically correct reference alignments. Our method can be used for this purpose and should therefore be useful for the further development of pairwise and multiple alignment methods.

REFERENCES

Abdeddaïm,S. and Morgenstern,B. (2001) Speeding up the DIALIGN multiple alignment program by using the greedy alignment of biological sequences library' (GABIOS-LIB). *Lecture Notes in Computer Science*, **2066**, 1–11.

Lassmann,T. and Sonnhammer,E.L.L. (2002) Quality assessment of multiple alignment programs. *FEBS Letters*, **529**, 126–130.

Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.

Morgenstern,B., Dress,A.W.M. and Werner,T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison.. *Proc. Natl Acad. Sci. USA*, **93**, 12098–12103.

Notredame,C. (2002) Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, **3**, 131–144.

Notredame,C., Higgins,D. and Heringa,J. (2000) T-Coffee: a novel algorithm for multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Thompson,J.D., Plewniak,F. and Poch,O. (1999a) BAliBASE: a benchmark alignment database for the evaluation of multiple sequence alignment programs. *Bioinformatics*, **15**, 87–88.

Thompson,J.D., Plewniak,F. and Poch,O. (1999b) A comprehensive comparison of protein sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.