DIALIGN-TX and multiple protein alignment using secondary structure information at GOBICS

Amarendran R. Subramanian[1], Suvrat Hiran[2,3], Rasmus Steinkamp[2], Peter Meinicke[2], Eduardo Corel[2] and Burkhard Morgenstern[2,*]

[1]Wilhelm-Schickard-Institut für Informatik, University of Tübingen, Sand 13, 72076 Tübingen, [2]Institute of Microbiology and Genetics, University of Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany and [3]Department of Mathematics, Indian Institute of Technology, Kharagpur, 721 302, India

Received February 12, 2010; Revised May 4, 2010; Accepted May 9, 2010

ABSTRACT

We introduce web interfaces for two recent extensions of the multiple-alignment program DIALIGN. DIALIGN-TX combines the greedy heuristic previously used in DIALIGN with a more traditional 'progressive' approach for improved performance on locally and globally related sequence sets. In addition, we offer a version of DIALIGN that uses predicted protein secondary structures together with primary sequence information to construct multiple protein alignments. Both programs are available through 'Göttingen Bioinformatics Compute Server' (GOBICS).

INTRODUCTION

Multiple sequence alignment (MSA) is the basis of almost all methods for sequence analysis in bioinformatics. Thus, the results of these methods crucially depend on the underlying alignments. A striking example is a recent study by Wong *et al*. (1). These authors demonstrated that uncertainties in multiple alignments drastically influence the output of standard phylogeny programs. Development and evaluation of MSA methods is therefore a central field of research in bioinformatics since the mid-1980s. Recent reviews on MSA methods are given, for example, by Edgar and Batzoglou (2), Morrison (3) or Kemena and Notredame (4).

DIALIGN

Since its first release in 1996, *DIALIGN* is a widely used software tool for multiple alignment of DNA, RNA and protein sequences (5,6). It differs in various aspects from other MSA algorithms. DIALIGN tries to align only those parts of the sequences to each other that exhibit some statistically relevant degree of similarity. Non-related parts of the sequences remain unaligned. This way, the method combines local and global alignment features. It returns global alignments where sequences are homologous over their entire length, but local alignments where only local homologies are detectable. DIALIGN constructs alignments based on gap-free local alignments, so-called fragments for which a scoring function is defined based on the probability of their random occurrence. Multiple alignments are constructed in a greedy way by incorporating fragments that are mutually consistent, i.e. fragments that fit into one single output MSA (7).

As most MSA methods, the standard version of DIALIGN is fully automated and works without human intervention. In addition, however, DIALIGN has an option for 'anchored alignment' where MSAs are produced in a 'semi-automatic' way (8,9). With this option, the program can be 'forced' to align user-defined positions of the sequences to each other, and the remainder of the sequences is aligned automatically. Anchored alignment can also be used to speed-up the alignment procedure where long genomic sequences are to be aligned (10,11) or to study the behaviour of alignment methods in detail (12).

Numerous studies have shown that DIALIGN is superior to other MSA tools if locally related sequence sets are aligned, but on globally related sequences with weak primary-sequence similarity, it is often outperformed by global methods such as 'CLUSTAL W' (13), 'MUSCLE' (14,15), 'MAFFT' (16) or 'PROBCONS' (17). Since the first release of the DIALIGN, various alternative optimization algorithms have been applied to the fragment-based alignment approach in order to improve its performance (18,19), but recent results indicate that the relative weakness of DIALIGN on global homologies is due to the underlying objective function and not so much on the greedy optimization algorithm (12).

[*] To whom correspondence should be addressed. Tel: +49 551 39 14628; Fax: +49 551 39 14929; Email: bmorgen@gwdg.de

DIALIGN-TX

DIALIGN-T is a complete re-implementation of DIALIGN developed by the first author of this article (20). In the first step, it performs all possible pairwise alignments of the input sequences in the sense of DIALIGN (21,22). For multiple alignment, however, DIALIGN-T uses a number of heuristics to prevent the algorithm from aligning spurious, isolated random similarities that might destroy a biologically more meaningful global alignment. For example, in the greedy algorithm for MSA, DIALIGN-T considers not only the local degree of similarity in a fragment, but also its context. Fragments that are part of a high-scoring pairwise alignment are preferred compared to isolated fragments. Also, low-scoring regions are removed from long fragments to counterbalance the bias of DIALIGN in favour of high-scoring fragments and to support groups of lower scoring fragments. Together with some other heuristics, this led to a considerable improvement of the performance compared with the original implementation of DIALIGN.

These ideas were taken a step further in the latest release of the program, 'DIALIGN-TX' (23). Here, the traditional progressive approach to multiple alignment (24–26) is adapted to the fragment-based alignment as used in DIALIGN. First a guide tree is calculated based on pairwise fragment alignments. Then pairwise alignments of sequences and groups of previously aligned sequences are performed going from the leaves to the root of the guide tree. In traditional progressive alignment methods, such groups of already aligned sequences are represented as 'profiles' and aligned by 'profile alignment'. This is not possible in DIALIGN, where an alignment is seen as a consistent set of fragments and only parts of the sequences may be aligned. To align two groups G_1 and G_2 of previously aligned sequences to each other, DIALIGN-TX selects a set of fragments each of which aligns a sequence from G_1 with a sequence from G_2. A vertex-cover algorithm by Clarkson (27) is used to remove inconsistencies and to select high-scoring sets of consistent fragments.

DIALIGN USING PROTEIN SECONDARY STRUCTURE INFORMATION

As most methods for multiple protein alignment, DIALIGN and DIALIGN-TX are based on primary structure information alone. However, attempts have been made in the past to use predicted secondary structures in alignment algorithms (28,29). We implemented a software pipeline that takes predicted protein secondary structures as additional input information for DIALIGN.

(1) In the first step, the standard version of DIALIGN is run to obtain pairwise alignments in the sense of the fragment-based alignment approach. That is, an optimal chain of fragments is calculated for each pair of input sequences.
(2) Next, we run PSIPRED (30) on the individual sequences to predict their secondary structures. PSIPRED is one of the most accurate *de novo* predictors for protein secondary structures (31). It assigns one of three different states—'helix' (H), strand (E) or 'coil' (C)—to every position of the sequences.
(3) We defined a modified weight function w' on the set of fragments that takes both primary and secondary structure into account. Based on the secondary structures predicted by PSIPRED, a new weight score $w'(f)$ of a fragment f is defined as

$$w'(f) = e^{\gamma s(f)} w(f)$$

where $w(f)$ is the original, primary sequence-based fragment weight as used in DIALIGN (6). $s(f)$ is a measure of similarity at the secondary-structure level for fragments and is defined as

$$s(f) = \alpha_H m_H + \alpha_E m_E + \alpha_C m_C + \beta_H p_H + \beta_E p_E + \beta_C p_C + \delta Sov(f).$$

Here, m_x is the proportion of matching states x, and p_x the proportion of predicted states x, where x can be H, E or C, as predicted by the PSIPRED program. Optimal values for the parameters α, β, γ and δ have been identified using a least squares support vector machine (32).

(1) The measure $Sov(f)$ of the similarity of predicted secondary structures for the segments composing the fragment f has been defined by Kim and Xie (29) on the basis of the original *Sov* score (33).
(2) For multiple alignment, we use the greedy algorithm implemented in DIALIGN, but fragments are ranked according to their sequence structure-based weights w' instead of the sequence-based weights w. Technically, this is done by defining the fragments contained in the respective pairwise DIALIGN alignments as 'anchor points' using the modified scores $w'(f)$ as weights that determine the priority of fragments in the greedy algorithm.

We evaluated our secondary structure-based MSA approach using the current release of 'BAliBASE 3' (34). Table 1 shows that, 'on average', the performance of DIALIGN using secondary structure information is similar to the performance of the program with primary-sequence information alone. For many data sets, however, we observed great differences in the resulting alignments. In some cases, the structure-based alignments were far better than the original ones, while in other cases it was the other way around. For some sequence sets, our secondary structure approach achieved an improvement of 29.7 percentage points in the sum-of-pairs (SP) score (or a relative improvement of 62%, respectively) compared to the purely sequence-based alignment. Therefore, we believe that our secondary structure-based alignments may contain valuable information that is not available in sequence-based MSAs and could therefore be a useful addition to sequence-based alignments.

Table 1. Performance of DIALIGN 2.2 with primary sequence information alone, our secondary structure-based alignment (DIALIGN SEC) and DIALIGN-TX on BAliBASE 3 under the 'sum-of-pairs' scoring scheme

	RV11	RV12	RV20	RV30	RV40	RV50
DIALIGN 2.2	50.7	86.2	86.9	71.0	82.3	79.8
DIALIGN SEC	49.8	84.5	86.6	74.7	83.1	81.8
DIALIGN-TX	51.5	89.1	87.8	76.1	83.6	82.2

On average, the performance of our secondary-structure based alignment is similar to the original version of the program. However, for some data sets in BAliBASE, there are great differences between the sequence-based and sequence-structure-based alignments. DIALIGN-TX clearly outperforms the previous release of DIALIGN with and without structure information. More detailed test results and a comparison to other methods are given in (23).

WWW SERVER AT GOBICS

To make the new versions of DIALIGN easily available to the research community, we set up WWW interfaces for them at 'Göttingen Bioinformatics Compute Server' (GOBICS). DIALIGN-TX is available at http://dialign-tx.gobics.de/submission.

Various parameter values can be selected by the user. For exclusion of low-scoring regions in long fragments, the minimum fragment length T from which low-scoring sub-fragments are excluded can be specified, as well as the length L of low-scoring regions that are excluded from alignment. That is, if a fragment f of length $\geq T$ contains a sub-fragment of length L, this sub-fragment is removed and f is split into the two remaining sub-fragments. Also, there are options to increase the program speed, possibly at the expense of sensitivity. For DNA alignment, there are several options to translate DNA fragments into peptide fragments according to the genetic code and to consider open reading frames for alignment.

The downloadable program versions contains more options and adjustable parameters which are explained in the user guide. Also, the downloadable program now comes with an 'anchored-alignment' option.

DIALIGN with secondary-structure information is available at: http://dialign-sec.gobics.de/submission.

FUNDING

Deutsche Forschungsgemeinschaft (grants MO 1048/1-1 and MO 1048/6-1 to B.M., in part). Funding for open access charge: Annual budget of department of bioinformatics, University of Göttingen.

Conflict of interest statement. None declared.

REFERENCES

1. Wong,K.M., Suchard,M.A. and Huelsenbeck,J.P. (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.
2. Edgar,R.C. and Batzoglou,S. (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**, 368–373.
3. Morrison,D.A. (2006) Multiple sequence alignment for phylogenetic purposes. *Aust. Syst. Bot.*, **19**, 479–539.
4. Kemena,C. and Notredame,C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, Vol. 25, pp. 2455–2465.
5. Morgenstern,B., Dress,A. and Werner,T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl Acad. Sci. USA*, **93**, 12098–12103.
6. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
7. Abdeddaïm,S. and Morgenstern,B. (2001) Speeding up the DIALIGN multiple alignment program by using the 'greedy alignment of biological sequences library' (GABIOS-LIB). *Lect. Notes Comput. Sci.*, **2066**, 1–11.
8. Morgenstern,B., Werner,N., Prohaska,S.J., Steinkamp,R., Schneider,I., Subramanian,A.R., Stadler,P.F. and Weyer-Menkhoff,J. (2005) Multiple sequence alignment with user-defined constraints at GOBICS. *Bioinformatics*, **21**, 1271–1273.
9. Morgenstern,B., Prohaska,S.J., Pöhler,D. and Stadler,P.F. (2006) Multiple sequence alignment with user-defined anchor points. *Algorithms for Molecular Biology*, **1**, 6.
10. Brudno,M., Steinkamp,R. and Morgenstern,B. (2004) The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences. *Nucleic Acids Res.*, **32**, W41–W44.
11. Pöhler,D., Werner,N., Steinkamp,R. and Morgenstern,B. (2005) Multiple alignment of genomic sequences using CHAOS, DIALIGN and ABC. *Nucleic Acids Res.*, **33**, W532–W534.
12. Corel,E., Pitschi,F. and Morgenstern,B. (2010) A min-cut algorithm for the consistency problem in multiple sequence alignment. *Bioinformatics.*, **26**, 1015–1021.
13. 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.
14. Edgar,R.C. (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
15. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
16. Katoh,K., Kuma,K., Toh,H. and Miyata,T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
17. Do,C.B., Mahabhashyam,M.S.P., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
18. Lenhof,H.-P., Morgenstern,B. and Reinert,K. (1999) An exact solution for the segment-to-segment multiple sequence alignment problem. *Bioinformatics*, **15**, 203–210.
19. Kececioglu,J.D., Lenhof,H.-P., Mehlhorn,K., Mutzel,P., Reinert,K. and Vingron,M. (2000) A polyhedral approach to sequence alignment problems. *Discrete App. Math.*, **104**, 143–186.
20. Subramanian,A.R., Weyer-Menkhoff,J., Kaufmann,M. and Morgenstern,B. (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.
21. Morgenstern,B. (2000) A space-efficient algorithm for aligning large genomic sequences. *Bioinformatics*, **16**, 948–949.
22. Morgenstern,B. (2002) A simple and space-efficient fragment-chaining algorithm for alignment of DNA and protein sequences. *Appl. Math. Lett.*, **15**, 11–16.
23. Subramanian,A.R., Kaufmann,M. and Morgenstern,B. (2008) DIALIGN-TX: greedy and progressive approaches for the segment-based multiple sequence alignment. *Algorithms Mol. Biol.*, **3**, 6.
24. Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.
25. Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL - a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.
26. Taylor,W.R. (1988) A flexible method to align large numbers of biological sequences. *J. Mol. Evol.*, **28**, 161–169.

27. Clarkson,K.L. (1983) A modification of the greedy algorithm for vertex cover. *Inf. Process. Lett.*, **16**, 23–25.
28. Heringa,J. (1999) Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comput. Chem.*, **23**, 341–364.
29. Kim,N.-K. and Xie,J. (2006) Protein multiple alignment incorporating primary and secondary structure information. *J. Comput. Biol.*, **13**, 75–88.
30. Jones,D.T. (2004) Protein secondary structure prediction based on position-specific scoring matrices. *Nucleic Acids Res.*, **32**, W41–W44.
31. Montgomerie,S., Sundararaj,S., Gallin,W.J. and Wishart,D.S. (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics*, **7**, 301.
32. Suykens,J.A.K. and Vandewalle,J. (1999) Least squares support vector machine classifiers. *Neural Process. Lett.*, **9(3)**, 293–300.
33. Zemla,A., Venclovas,C., Fidelis,K. and Rost,B. (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**.
34. Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.