BIOINFORMATICS

Vol. 14 no. 3 1998
Pages 290–294

DIALIGN: Finding local similarities by multiple sequence alignment

Burkhard Morgenstern[1], Kornelie Frech[2], Andreas Dress[3] and Thomas Werner[2,4]

GSF–National Research Center for Environment and Health, [1]Institute of Biomathematics and Biometry, [2]Institute of Mammalian Genetics, Ingolstädter Landstr. 1, 85764 Neuherberg and [3]Research Center for Interdisciplinary Studies on Structure Formation (FSPM), Universität Bielefeld, 33501 Bielefeld, Postfach 100131, Germany

Received on October 17, 1997; revised on December 2, 1997; accepted on December 8, 1997

Abstract

Motivation: *DIALIGN is a new method for pairwise as well as multiple alignment of nucleic acid and protein sequences. While standard alignment programs rely on comparing single residues and imposing gap penalties, DIALIGN constructs alignments by comparing whole segments of the sequences. No gap penalty is employed. This point of view is especially adequate if sequences are not globally related, but share only local similarities, as is the case in genomic DNA sequences and in many protein families.*

Results: *Using four different data sets, we show that DIALIGN is able correctly to align conserved motifs in protein sequences. Alignments produced by DIALIGN are compared systematically to the results of five other alignment programs.*

Availability: *DIALIGN is available to the scientific community free of charge for non-commercial use. Executables for various UNIX platforms including LINUX can be downloaded at http://www.gsf.de/biodv/dialign.html*

Contact: *{werner,morgenstern}@gsf.de*

Introduction

Alignment of nucleic or amino acid sequences is one of the most important tools of sequence analysis in molecular biology. Consequently, an important challenge for computational biology is to design algorithms capable of automatically finding 'biologically correct' alignments, i.e. alignments which correlate the functionally, structurally or evolutionarily related parts of sequences in question. The two major prerequisites involved are: (i) a scoring scheme that allows assignment of a distinct score to every possible alignment of a given set of sequences and (ii) a suitable algorithm capable of finding optimal, or at least reasonable suboptimal, alignments according to this scoring scheme.

Since the early 1970s, most alignment algorithms have employed versions of a scoring scheme proposed by Needleman and Wunsch (1970). Given a similarity matrix, e.g. PAM (Dayhoff *et al.*, 1978) or BLOSUM (Henikoff and Henikoff, 1994), the overall similarity score of a pairwise alignment is defined by the sum of all similarity values of the aligned residue pairs minus a so-called gap penalty for every gap introduced into the alignment. Needleman and Wunsch have proposed a dynamic programming algorithm which is able to find optimal alignments according to this scoring scheme.

Since then, the alignment problem has been widely considered as being solved for pairwise alignments and most efforts focused on improving the algorithm to find optimal or reasonably good suboptimal multiple alignments according to the Needleman–Wunsch scoring scheme (Feng and Doolittle, 1987; Carrillo and Lipman, 1988; Thompson *et al.*, 1994; Tönges *et al.*, 1996; Abdeddaïm, 1997; Stoye *et al.*, 1997). In addition, considerable efforts have been made to define appropriate parameter settings, especially for the gap penalty, a crucial determinant of the final alignment (Fitch and Smith, 1983; Vingron and Waterman, 1994).

The Needleman–Wunsch algorithm produces reasonable, i.e. 'biologically correct' (or at least, acceptable) alignments if sequences are closely related and only a small number of gaps have to be inserted during the alignment procedure. However, the scoring scheme based on single matches and gap penalties cannot be appropriate if the sequences share only local similarity which might be caused by genetic processes like recombination or exon shuffling events.

Smith and Waterman (1981) have developed a 'local' version of the Needleman–Wunsch method which can be successfully applied if two sequences share one single region of high similarity and are not related outside of this region. The situation is more difficult if sequences share several regions of local similarity which are separated by unrelated regions, e.g. by introns for genomic DNA or loops for proteins. Recently, we have proposed a novel alignment algorithm which is especially suited to detect local similarities even if these similarities are separated by long or short unrelated parts of the sequences (Morgenstern *et al.*, 1996) and which, as dis-

© Oxford University Press

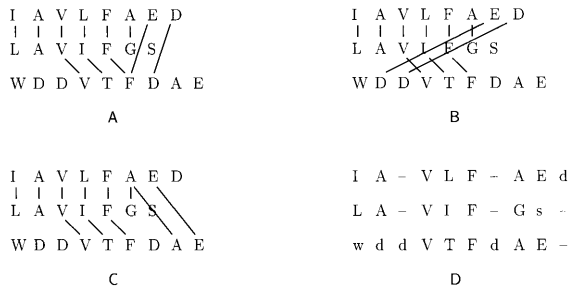

Fig. 1. Non-consistent and consistent collections of diagonals (segment pairs). (**A**) and (**B**) represent non-consistent collections of diagonals. In (A), the 'F' in the third sequence is assigned simultaneously to two different residues of the first sequence. In (B), there is a 'cross-over' assignment of residues. By contrast, (**C**) is a consistent collection of diagonals. It is possible to introduce gaps into the sequences such that residues connected by diagonals are in the same column of the resulting alignment (**D**). Residues not involved in any of the three diagonals are printed in lower-case letters. They are not considered to be aligned.

cussed below, reflects in a rather direct way the basic principles of sequence evolution as seen today. Here, we present the algorithm in general terms and describe the implementation into a program called DIALIGN 1.0.

Algorithm

The basic idea of our algorithm is to build sequence alignments by comparison of whole segments (i.e. uninterrupted stretches of residues) of the sequences rather than by comparison of single residues. Accordingly, alignments are composed from gap-free pairs of segments of equal length. Such pairs of segments are referred to as diagonals since they would form diagonals in a dot-matrix comparison of two sequences. Diagonals of various length are considered simultaneously and mismatches are allowed within diagonals.

A pairwise as well as a multiple alignment comprises a suitable collection of diagonals meeting a certain consistency criterion [a mathematical definition of consistency is given in Morgenstern *et al.* (1996)]. In short, a collection of diagonals is called consistent if there is no conflicting double or cross-over assignment of residues (see Figure 1). We assign a so-called weight to every possible diagonal, and then try to find a consistent collection of diagonals with maximal sum of weights. Gaps are not considered in the calculation of the alignment score. An optimal alignment, i.e. a collection of diagonals with maximum sum of weights, can be found by a modification of the standard dynamic programming scheme which is feasible at least for pairwise alignments.

The weight function for diagonals is based on probabilistic considerations [for a mathematical definition, see Morgenstern *et al.* (1996)]. To reduce the 'noise' of small random diagonals, a threshold T is used as a lower cut-off criterion for diagonals to be taken into consideration, which can be specified by the user.

Multiple alignments are constructed as follows. In a first step, all optimal pairwise alignments are formed. The diagonals incorporated into these alignments are sorted (i) according to their weight scores and (ii) according to the degree of overlap with other diagonals in order to emphasize motifs occurring in more than two sequences (so-called overlap weights). The resulting list of diagonals is then used to assemble a multiple alignment in a greedy manner: the diagonal with the highest weight is the first one to be selected for the alignment. Then, the next diagonal from the list is checked for consistency and added to the alignment if consistent. The algorithm proceeds in this way until the whole list of diagonals has been processed. Once a diagonal is selected, it becomes part of the alignment and cannot be removed at any later stage.

The process of performing pairwise alignments, sorting diagonals, and incorporating them greedily into a growing multiple alignment is repeated iteratively until no additional diagonals can be found. [A similar greedy approach was proposed independently in Abdeddaïm (1997).]

In a final step, the program introduces gaps into the sequences until all residues connected by the selected diagonals are properly arranged. In the output, these residues are printed in upper-case letters, whereas residues not involved in any of the selected diagonals are printed in lower-case letters. They are not considered to be aligned (see Figure 1D). If sequences are only locally related, DIALIGN does not attempt to generate a global alignment of sequences and will only align residues connected by selected diagonals.

Results

To test our method and to compare it to other methods, we have employed four different data sets: (i) a set of 30 helix–turn–helix proteins used in Lawrence *et al.* (1993) as test material for their Gibbs sampling method; (ii) a set of 16 acetyltransferase proteins as described in Neuwald *et al.* (1994); (iii) a set of nine protein sequences of the basic helix–loop–helix (bHLH) family of transcription factors as described by Atchley and Fitch (1997) (accession numbers: P41894, Q02575, P17106, A55438, U10638, P13902, Q04635, U11444, A48085); (iv) a set of 12 RH proteins (McClure *et al.*, 1994).

Table 1. Comparison of alignment methods using four different sets of protein sequences. The table contains the numbers of correctly aligned domains. In many instances, there are several groups of sequences where a domain was correctly aligned within these groups but not between groups. The table reports the number of sequences for each of these correctly aligned groups; e.g. with $T = 0$, DIALIGN correctly aligned the first domain of the transferase sequences within a group of 12 sequences and within another group of two sequences, but the domain could not be correctly aligned between these two groups. A domain is considered to be correctly aligned if at least 75% of the residues are correctly aligned

Data set	HTH	Transferase		bHLH		RH			
Number of sequences	30	16		9		12			
Conserved domain		I	II	I	II	I	II	III	IV
DIALIGN ($T = 0$)	6,6,3,2,2	12,2	9	7	3,2,2	11	9	6,2,2	12
DIALIGN ($T = 10$)	19,2,2	16	13,2	9	9	8,2	6,2	7	8,2
CLUSTAL W	5,3,2,2,2	13	12	3,2	3,2	11	6,2	6	8,3
MULTALIN	6,5,4,2,2,2	8,3,2	7,5,2	5	4	7,3	6,2	5,2,2	5,4,3
MAP	6,5,4,3,2,2	7,2,2,2,2	4,4,2	5,3	4,3	6,3	6,2	6,2	3,2,2
PIMA	5,4,3,3,2	10,3,2	8,3,2	2	2	10	8	7,3	3,3,2,2
MATCH-BOX	3	0	0	0	8	5	0	3	0

In each data set, sequences contain one or several conserved domains as described in the literature. We tested various alignment programs with regard to their ability to align these domains correctly: DIALIGN (this study), CLUSTAL W (Thompson *et al.*, 1994), MULTALIN (Corpet, 1988), MAP (Huang, 1994), PIMA (Smith and Smith, 1992) and MATCH-BOX (Depiereux and Feytmans, 1992). CLUSTAL W, MULTALIN and MAP are global progressive alignment methods; PIMA and MATCH-BOX are local methods.

All programs were applied with default parameters. In addition, we used DIALIGN with a threshold $T = 10$ in order to study the influence of this parameter on the resulting alignments.

The results of this comparison are summarized in Table 1 and one example is given in detail in Figure 2. For all test examples, DIALIGN was among the best-scoring programs. However, in all but one example, the best results were not obtained with the default threshold $T = 0$, but with $T = 10$. It seems that this threshold improves the resulting alignments if sequences share significant local similarities occurring at different positions within the sequences. This situation occurs in helix–turn–helix, acetyltransferase and helix–loop–helix motifs. In these examples, DIALIGN yields the best results with $T = 10$.

Future efforts should be made in order to study the influence of the parameter T in more detail and to improve the weighting scheme further.

Discussion

The alignment algorithm described here differs fundamentally from standard algorithms by its way of scoring the quality of alignments. Unlike alignment methods relying on the sum of individual similarity values and on gap penalties as optimization criteria, we focus on comparing complete segments of sequences. Therefore, DIALIGN is able to locate small conserved regions that cannot be detected by standard alignment programs.

If sequences share only limited regions of similarity, DIALIGN aligns these regions and ignores the unrelated parts of the sequences. However, unlike pure motif search programs (Henikoff and Henikoff, 1994; Neuwald *et al.*, 1995, 1997), DIALIGN will return a global alignment if detectable similarity extends over the full range of the sequences.

The present implementation of DIALIGN uses a rather simple weighting scheme to assess the quality of diagonals. However, this specific weighting scheme is not essential for our algorithm. Different weighting schemes should be tested in order to improve the performance of the algorithm further.

The basic concept of segment comparison is also in agreement with some of the most fundamental principles of sequence evolution that are now generally accepted. The driving force in most cases appears to be exchange of whole segments of sequences by recombination (Mushegian and Koonin, 1996) or transposition (Plasterk, 1993) which also includes mechanisms of gene conversion (Gangloff *et al.*, 1996). Point mutations add the fine tuning of sequences, while insertions or deletions of single nucleotides are relatively rare events in functional genomic sequences as compared to insertions of longer sequence elements (e.g. retrotransposons; Batzer *et al.*, 1996). All of these mechanisms are accounted for in DIALIGN: high-scoring diagonals or sets of diagonals correspond to shuffled sequence regions, mismatches within the diagonals represent point mutations and insertions or deletions within conserved regions can be accommodated by splitting diagonals into smaller subdiagonals.

Recombinations cause abrupt termination of biological homology. Even where standard alignment methods are able to align isolated homologies correctly, they tend to extend the

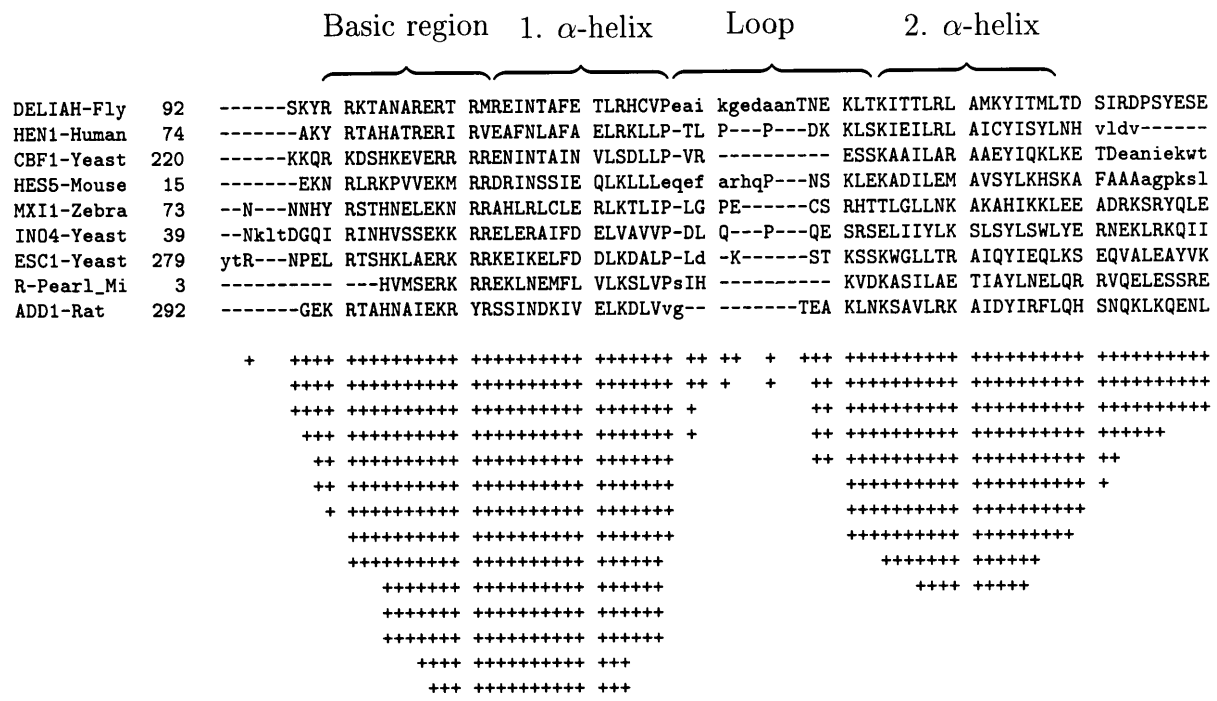

Fig. 2. Part of an alignment of nine basic helix–loop–helix proteins as constructed by DIALIGN 1.0 with threshold $T = 10$. For each position in the alignment, the number of plus signs represents the sum of the weights of all diagonals involving residues at this positi on. The two regions with the highest number of plus signs in the entire alignment correspond precisely to the two parts of the functional domain of the sequences: (i) DNA-binding basic region and first α-helix and (ii) second α-helix. The two parts of the motif are separated by a loop region. Numbers on the left-hand side of the alignment refer to the left-most residue in a line and denote their positions within the sequence. Lower-case letters denote residues not belonging to any of the selected diagonals. They are not considered to be aligned.

alignment beyond the homologous regions since they try to maximize the total sum of individual similarity scores and minimize the number of gaps. By contrast, DIALIGN tries to find local similarities among sequences and restricts the alignment to segments of the sequences that are more similar to each other than can be expected by chance alone.

The new concept implemented in DIALIGN has already proved to be useful for many users who downloaded the program from our WWW server. We hope that this paper fulfills the frequent demand for a more detailed description and discussion of the biological motivation of the algorithm to complement the mathematical principles detailed in the first publication (Morgenstern *et al.*, 1996). We believe that the potential of the concepts on which DIALIGN is based deserves to be developed further and will prove to be a valuable addition to the current collection of alignment algorithms.

Acknowledgements

We would like to thank Korbinian Grote, Kerstin Quandt and Kurt Wollenberg for critically reading the manuscript, and William R.Atchley for many helpful discussions. Part of this work was supported by EU grant BI04-CT95-0226 (TRADAT).

References

Abdeddaïm,S. (1997) Incremental computation of transitive closure and greedy alignment. In Apostolico,A. and Hein,J. (eds), *Proceedings of the 8th Annual Symposium on Combinatorial Pattern Matching. Lecture Notes Comput. Sci.*, **1264**, 167–179.

Atchley,W.R. and Fitch,W.M. (1997) A natural classification of the basic helix-loop-helix class of transcription factors. *Proc. Natl Acad. Sci. USA*, **94**, 5172–5176.

Batzer,M.A. *et al.* (1996) Genetic variation of recent Alu insertions in human populations. *J. Mol. Evol.*, **42**, 22–29.

Carrillo,H. and Lipman,D. (1988) The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.*, **48**, 1073–1082.

Corpet,F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.*, **16**, 10881–10890.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. *Atlas Protein Seq. Struct.*, **6**, 345–362.

Depiereux,E., Baudoux,G., Briffeuil,P., Reginster,I., De Boll,X., Vinals,C. and Feytmans,E. (1996) Match-Box-server: a multiple sequence alignment tool placing emphasis on reliability. *Comput. Applic. Biosci.*, **13**, 249–256.

Gangloff,S., Zou,H. and Rothstein,R. (1996) Gene conversion plays the major role in controlling the stability of large tandem repeats in yeast. *EMBO J.*, **15**, 1715–1725.

Feng,D.F. and Doolittle,R.G. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.

Fitch,W.M and Smith,T.F. (1983) Optimal sequence alignments. *Proc. Natl Acad. Sci. USA*, **80**, 1382–1386.

Henikoff,S. and Henikoff,J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics*, **19**, 97–107.

Huang,X. (1994) On global sequence alignment. *Comput. Applic. Biosci.*, **10**, 227–235.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993) Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–213.

McClure,M.A., Vasi,T.K. and Fitch,W.M. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.*, **11**, 571–592.

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.

Mushegian,A.R. and Koonin,E.V. (1996) Sequence analysis of eukaryotic developmental proteins: Ancient and novel domains. *Genetics*, **144**, 817–828.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Neuwald,A.F. and Green,P. (1994) Detecting patterns in protein sequences. *J. Mol. Biol.*, **239**, 698–712.

Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618–1632.

Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Res.*, **25**, 1665–1677.

Plasterk,R.H.A. (1993) Molecular mechanisms of transposition and its control. *Cell*, **74**, 781–786.

Smith,R.F. and Smith,T.F. (1992) Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modeling. *Protein Eng.*, **5**, 35–41.

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.

Stoye,J., Perrey,S.W. and Dress,A. (1997) Improving the divide-and-conquer approach of sum-of-pairs multiple sequence alignment. *Appl. Math. Lett.*, **10**, 67–63.

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.

Tönges,U., Perrey,S.W., Stoye,J. and Dress,A. (1996) A general method for fast multiple sequence alignment. *Gene*, **172**, GC33–41.

Vingron,M. and Waterman,M.S. (1994) Sequence alignment and penalty choice—review of concepts, case studies and implications. *J. Mol. Biol.*, **235**, 1–12.