Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W41–W44***
DOI: 10.1093/nar/gkh361

The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences

Michael Brudno, Rasmus Steinkamp[1] and Burkhard Morgenstern[1,*]

Department of Computer Science, Stanford University, Stanford, CA 94305, USA and [1]University of Göttingen, Institute of Microbiology and Genetics, Goldschmidtstrasse 1, 37077 Göttingen, Germany

Received February 9, 2004; Revised and Accepted March 1, 2004

ABSTRACT

Cross-species sequence comparison is a powerful approach to analyze functional sites in genomic sequences and many discoveries have been made based on genomic alignments. Herein, we present a WWW-based software system for multiple alignment of large genomic sequences. Our server utilizes the previously developed combination of CHAOS and DIALIGN to achieve both speed and alignment accuracy. CHAOS is a fast database search tool that creates a list of local sequence similarities. These are used by DIALIGN as anchor points to speed up the final alignment procedure. The resulting alignment is returned to the user in different formats together with a list of anchor points found by CHAOS. The CHAOS/DIALIGN software is freely available at http://dialign.gobics.de/chaos-dialign-submission.

INTRODUCTION

In recent years, cross-species sequence comparison has become a popular approach to genome sequence analysis. The idea is that functional parts of genomic sequences are evolutionarily more conserved than non-functional parts. Thus, islands of local sequence conservation usually correspond to biologically functional sites. This *phylogenetic footprinting* principle has been used by many researchers to detect novel functional elements in genomic sequences. Genomic sequence comparison has been used for gene prediction (1–5), to discover regulatory elements (6,7) and to study genomic duplications (8,9). Recently, multiple sequence comparison has been used to identifiy *signature sequences* of bacteria and viruses for rapid detection of pathogene microorganisms as part of the US biodefense program (10).

All these comparative studies rely on pair-wise or multiple *alignments* of genomic sequences; their accuracy is therefore limited by the accuracy of the underlying alignment tools, i.e. by their ability to correctly align functionally or evolutionarily related sites. Consequently, development of algorithms for genomic sequence alignment has become a highly active field in bioinformatics research; see (11,12) for a survey. DIALIGN (13,14) is a versatile tool for alignment of DNA and protein sequences that combines both global and local alignment features. It returns a global or a local alignment—or a mixture of both—depending on the extent of similarity among the input sequences. This is achieved by assembling pair-wise and multiple alignments from un-gapped local pair-wise *fragment alignments* or *fragments*. The ability to combine global and local alignment strategies is particularly useful if genomic sequences are to be compared where homologies may be separated by large stretches of non-conserved sequence. In the last few years, DIALIGN has therefore been used by numerous researchers to analyze genomic sequences. An independent study by Pollard *et al*. (15) evaluated the capability of alignment programs to detect conserved non-coding sites in genomic sequences. These investigators conclude that 'the distinct virtues of both global and local tools are currently incorporated in the output of only one tool, DIALIGN'. In their study, they found that 'DIALIGN can produce alignments with high coverage and sensitivity, as well as specificity to detect constrained sites'.

ANCHORED MULTIPLE ALIGNMENT

Initially, DIALIGN has been developed as a multi-purpose alignment tool. Though it produces genomic alignments of high quality, the original version of the program was far too slow to align sequences of hundreds of kilobases or even megabases in length. We therefore implemented an *anchored alignment* option, where user-specified anchor points can be used to reduce the alignment search space, thereby improving the program running time (16). To find suitable anchor points, we are using the recently developed software program CHAOS (17). CHAOS is a search tool for local alignment of genomic sequences. Based on the *trie* data structure, it identifies short local sequence similarities; the final output of the program is a *chain of local alignments*. The CHAOS algorithm is also used as part of the LAGAN

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

and Multi-LAGAN alignment tools (18). In a first step, our system applies CHAOS to identify chains of local similarities among the input sequences. In a second step, DIALIGN is used to accurately align the regions *between* those anchor points identified by CHAOS.

Our anchored-alignment approach can be applied for pair-wise as well as multiple alignment. For multiple alignment, CHAOS is run on all possible pairs of input sequences. The resulting local pair-wise similarities are then checked for *consistency* by DIALIGN and non-consistent ones are eliminated. This corresponds to the greedy approach that DIALIGN uses to construct multiple alignments; see (14). In a recent paper, we showed that this combined CHAOS/DIALIGN approach produces alignments that are very similar to the output of the original DIALIGN program while it is up to two orders of magnitude faster (17). In one instance, the combined program was even able to identify a regulatory site that could not be detected by DIALIGN alone.

THE CHAOS/DIALIGN WWW SERVER

We developed a WWW interface for the combined CHAOS/DIALIGN software at Göttingen bioinformatics compute server (GOBICS) (Figure 1). The input data is a single text file containing two or several genomic sequences in FASTA format. The maximum total length of the input sequences is currently 3 MB. The server runs CHAOS and DIALIGN on the

Figure 1. The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences. Input sequences are uploaded as a single multi-sequence file in FASTA format.

```
dog_il4        20565   AGAGCCTGGT CTGGAGCAAA GTTGATGTCT ACCTGTGCTT TCTTTAGCAG
hum_il4        21730   AGAGCCTGGT CTGGGGCAAA GTTGATGTCT ACCTGTGCTT TCTTTAGCAG
mus_il4        31390   ---------- ---GGGCCGA GCTGATGTCT ACCTGTGCTT TCTTTAGCAG

                       1111111111 1112222222 2222222222 2222222222 2222222222

dog_il4        20615   ATCAGATAta gGAG---TAC ACCAGTCGGG CATGAGCCTC TCCAGCTCTA
hum_il4        21780   ATCAGATAta gGAG---CAC ACCAGCCGGG CATGAGCCTC TCCAGCTCTA
mus_il4        31427   ATCAGATAta gccgcagCAC AGCAGTCGGG CATGAGCCTC TCCAACTCTA

                       2222222000 0222000333 3334444444 4444444444 4444444444

dog_il4        20662   AGGTGATGAT GACCAAGGCC AGTGTGGAGC CCTTGAacTG CAGCAGCTGG
hum_il4        21827   AGGTGATGAT GACCAAGGCC AGTGTGGAGC CCTTGAACTG CAGCAGCTGG
mus_il4        31477   AGGTGATGAT GACTAAAGCA AATGTGGAGC CCTTGAACTG CAGCAGCTGG

                       4444444444 4433666666 6666666666 6666662299 9999999999
```

Figure 2. Output alignment in DIALIGN format. Names of the aligned sequences are shown on the left. Numbers between names and sequences denote the position of the first residue in a line within the respective sequence. Capital letters denote aligned residues, i.e. residues involved in at least one of the fragments, the alignment consists of. Lower-case letters denote residues not belonging to any of these selected fragments. They are not considered to be aligned. Thus, if a lower-case letter is in the same column with other letters, this is pure chance; these residues are not considered to be homologous. Numbers below the alignment roughly reflect the degree of local similarity among the sequences. More precisely: they represent the sum of weight scores for those fragments that connect residues at the respective column. The numbers are normalized in such a way that every position gets a value between 0 and 9 and in *every* alignment, the region of maximum similarity is scored 9. Thus, these scores indicate *relative* rather than *absolute* similarity.

input sequences. For small input data, the resulting alignment is immediately shown on the computer screen. For larger sequence sets, the program output is stored at our server; the corresponding web addresses are sent to the user by email. Four different output files are created: (i) the output alignment in DIALIGN format as shown in Figure 2, (ii) the same alignment in FASTA format, (iii) a list of *fragments*, i.e. local segment pairs that are used as building blocks for the DIALIGN alignment and (iv) a list of anchor points identified by CHAOS.

Alignments in DIALIGN format contain additional information about the degree of local sequence similarity in the multiple alignment; see Figure 2. The program distinguishes between nucleotides that could be aligned and nucleotides with no statistically significant matches to the compared sequences. Upper-case and lower-case letters are used to indicate which nucleotides are considered to be aligned. This output format is designed for visual inspection of the returned alignments. The output in FASTA format contains essentially the same information but is more appropriate for further automatic analysis, since most sequence analysis programs accept FASTA-formatted files as input data. The list of returned fragments is annotated with some additional information that may be useful for more detailed analyses (Figure 3). This includes quality scores (so-called *weights*) of the fragments indicating the degree of local sequence similarity. In addition, calculated *overlap weights* are returned. Overlap weights reflect not only the similarity among two segments but also the degree of overlap with other segment pairs involving different pairs of sequences; see (13) for details. Finally, the fragment list states for each fragment if it was *consistent* with other fragments and could be included into the multiple alignment. The fragment list is also designed for automatized post-processing. It is easy to parse and contains more information than the resulting alignment alone. In addition to the fragment list, a list of anchor points created by CHAOS is

```
1) seq:  2  3  beg:  185955  178118 len:  90 wgt:  42.00 olw: 107.68 it: 1 cons
2) seq:  1  2  beg:  201612  185943 len:  90 wgt:  39.98 olw: 105.66 it: 1 cons
3) seq:  1  2  beg:   20700   21865 len:  90 wgt:  48.94 olw: 104.80 it: 1 cons
4) seq:  1  2  beg:  109548  109795 len:  84 wgt:  37.82 olw: 104.30 it: 1 cons
5) seq:  2  3  beg:   39483   53649 len:  90 wgt:  39.09 olw: 104.27 it: 1 cons
```

Figure 3. List of fragments (= aligned segment pairs) returned by the program. The list contains those fragments that are part of the respective optimal pairwise alignments in order of decreasing overlap weights. The list contains coordinates, weight scores and consistency information. For example, the first fragment involves sequences 2 and 3, starts at positions 185,955 and 178,118, respectively, within these sequences, is 90 nucleotides in length, has a weight score of 42.00, an overlap weight score of 107.68 and was found in the first iteration step of the alignment procedure; see (16). The fragment was consistent ('cons') in the multiple alignment procedure; i.e. it is included in the final multiple alignment.

returned. Our WWW server provides detailed online help regarding input and output formats.

AVAILABILITY

The CHAOS/DIALIGN server is located at Göttingen bioinformatics compute server (GOBICS): http://dialign.gobics.de/chaos-dialign-submission.

CHAOS and DIALIGN are also downloadable for local installation from the respective home pages at http://bibiserv.techfak.uni-bielefeld.de/dialign/ and http://www.stanford.edu/~brudno/chaos/

ACKNOWLEDGEMENTS

We would like to thank Serafim Batzoglou, Inna Dubchak and Chuong B. Do for their help. Two unknown referees made useful comments on the manuscript. The work was supported by Deutsche Forschungsgemeinschaft, project MO 1048/1-1.

REFERENCES

1. Bafna,V. and Huson,D.H. (2000) The conserved exon method for gene finding. *Bioinformatics*, **16**, 190–202.
2. Batzoglou,S., Pachter,L., Mesirov,J.P., Berger,B. and Lander,E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.
3. Korf,I., Flicek,P., Duan,D. and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**, S140–S148.
4. Wiehe,T., Gebauer-Jung,S., Mitchell-Olds,T. and Guigó,R. (2001) SGP-1: Prediction and validation of homologous genes based on sequence alignments. *Genome Res.*, **11**, 1574–1583.
5. Taher,L., Rinner,O., Gargh,S., Sczyrba,A., Brudno,M., Batzoglou,S. and Morgenstern,B. (2003) AGenDA: homology-based gene prediction. *Bioinformatics*, **19**, 1575–1577.
6. Loots,G.G., Locksley,R.M., Blankespoor,C.M., Wang,Z.E., Miller,W. Rubin,E.M. and Frazer,K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.
7. Göttgens,B., Barton,L., Gilbert,J., Bench,A., Sanchez,M., Bahn,S., Mistry,S., Grafham,D., McMurray,A., Vaudin,M. *et al.* (2000) Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.*, **18**, 181–186.
8. Prohaska,S. Fried,C., Flamm,C., Wagner,G.P. and Stadler,P.F. Surveying phylogenetic footprints in large gene clusters: applications to hox cluster duplications. *Mol. Evol. Phylog.*, in press.
9. Fried,C., Prohaska,S. and Stadler,P. (2003) Independent hox-cluster duplications in lampreys. *J. Exp. Zool. Part B*, **299**: 18–25.
10. Fitch,J., Gardner,S., Kuczmarski,T., Kurtz,S., Myers,R. Ott,L., Slezak,T., Vitalis,E., Zemla,A. and McCready,P. (2002) Rapid

development of nucleic acid diagnostics. *Proceedings of the IEEE*, **90**, 1708–1721.
11. Miller,W. (2001) Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics*, **17**, 391–397.
12. Chain,P., Kurtz,S., Ohlebusch,E. and Slezak,T. (2003) An applications-focused review of comparative genomics tools: capabilities, limitations, and future challenges. *Brief. Bioinform.*, **4**, 105–123.
13. 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.
14. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
15. Pollard,D.A., Bergman,C.M., Stoye,J., Celniker,S.E. and Eisen,M.B. (2004) Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, **5**, 6.
16. Morgenstern,B. Rinner,O., Abdeddaïm,S., Haase,D., Mayer,K., Dress,A. and Mewes,H.-W. (2002) Exon discovery by genomic sequence alignment. *Bioinformatics*, **18**, 777–787.
17. Brudno,M., Chapman,M., Gottgens,B., Batzoglou,S. and Morgenstern,B. (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**, 66.
18. Brudno,M., Do,C.B., Cooper,G.M., Kim,M.F., Davydov,E., Green, E.D., Sidow, A., Batzoglou, S. NISC Comparative Sequencing Program (2003) LAGAN and multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13**, 721–731.