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

Michael Brudno, Rasmus Steinkamp and Burkhard Morgenstern

Department of Computer Science, Stanford University, Stanford, CA 94305, USA and 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 identify signature sequences of bacteria and viruses for rapid detection of pathogenic 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
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 pairwise 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

![Image of CHAOS/DIALIGN job submission window](http://dialign.gobics.de/chaos-dialign-submission)

**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.
same alignment in FASTA format, (iii) a list of fragments is immediately shown on the computer screen. For larger input sequences. For small input data, the resulting alignment is more appropriate for further automization.

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 not belonging to any of these selected fragments. They 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 homologous. 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 automated 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 returned. Our WWW server provides detailed online help regarding input and output formats.

### 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


