

Internet On-Ramp

Multiple Sequence Alignment Tools on the Web

Giles J. Gaskell, eBioinformatics, Sydney, Australia (Giles.Gaskell@eBioinformatics.com)

The immense number of nucleotide and protein sequences that can be accessed through public databases on the Internet is an invaluable resource to scientists working in the fields of molecular biology, protein chemistry and molecular diagnostics. Tools such as BLAST (1) (also discussed in the March 2000 Internet On-Ramp) and FASTA (7) are used mainly to search for database sequences that are similar or potentially homologous to query sequences supplied by the investigator. However, in many circumstances, this is only the first stage of a more complex series of sequence analyses. Multiple sequence alignment analyses are fundamental in allowing investigators to draw conclusions about the similarity, function, structure and the potential evolutionary relationships between sequences. This column will focus on a number of different multiple sequence alignment tools available as services over the Web.

ClustalW

ClustalW (10), one of the most popular multiple nucleotide or protein sequence alignment programs, uses a progressive alignment approach (4). Progressive multiple alignments are created by first aligning the most similar of a set of sequences and then incorporating less similar sequences successively into the alignment. ClustalW aligns sequences using a global alignment algorithm (6) that generates an alignment over the entire length of the sequences. A neighbor-joining method (8) is first used to construct a guide tree. This determines the order in which the sequences are incorporated into the alignment. A comparison of multiple sequence alignment programs revealed that ClustalW performs well when aligning equidistant sequences of a similar length and when aligning small to large families of similar sequences, in which a few divergent sequences are also included in the alignment (11). In both cases, the performance of ClustalW was maintained from short sequences of less than 100 residues to those of over 400 residues.

ClustalW is available on the Internet through several locations: the European Bioinformatics Institute (EBI) (<http://www2.ebi.ac.uk/clustalw/>), the Baylor College of Medicine (<http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html>), the GenomeNet server in Japan (<http://www.clustalw.genome.ad.jp/>), the Institut Pasteur Software Environment (PISE) in France (<http://bioweb.pasteur.fr/intro-uk.html>), the Pôle Bio-Informatique Lyonnais (<http://pbil.ibcp.fr/>) also in France, the

Transfac group at the GBF in Germany (<http://transfac.gbf.de/programs.html>) and the Web-based sequence analysis service provider, BioNavigator™ (<http://www.bionavigator.com>). These servers allow investigators to cut and paste their sequences into forms on their Web sites and set various parameters, such as penalty values associated with the insertion of gaps into the sequences, to optimize the overall alignment. After submitting the job, the server computes the alignment, which is then displayed on a new page. Note that most ClustalW Web services that present the alignment directly to the investigator often time out on long alignments and lose the results. These instances occur when many and/or long sequences are submitted to these servers. Therefore, they are only suitable for aligning small numbers of short sequences. Services that get around this limitation are those that send their results by e-mail and those that store the results on the server, such as BioNavigator.

PileUp

Like ClustalW, PileUp uses both a progressive approach and the global alignment algorithm (6) when it aligns sequences. However, the order in which the sequences are incorporated into the alignment is determined by a guide tree that is constructed using the UPGMA method (9). PileUp is usually distributed as part of the Wisconsin Package from the Genetics Computer Group (GCG) (<http://www.gcg.com>), which is licensed to numerous bioinformatics services on the Internet, including BioNavigator, NIH Helix Systems (<http://helix.nih.gov/newhelix.html>), the Human Genome Mapping Project, UK (HGMP) (<http://www.hgmp.mrc.ac.uk/>) and the Biotechnology Computing Facility at the University of Arizona (<http://bcf.arl.arizona.edu/gcg.html>). The Wisconsin Package also includes tools for improving the appearance of multiple sequence alignments, which will be discussed later in this paper.

Dialign

Dialign (5) is a multiple sequence alignment program that takes an approach that is different from ClustalW and PileUp. Dialign uses a local alignment algorithm that compares sequence segments with each other to generate a multiple alignment. Sequences to be aligned are first broken down into gap-free segments of high similarity. These segments are then built up into a multiple alignment using an iterative procedure. In the same study comparing multiple sequence alignment programs (11), Dialign performed well when aligning sequences in which large terminal extensions or large insertions were present. In contrast to ClustalW, this program is useful for finding blocks of highly conserved regions within a set of sequences. Dialign is available on the Internet from Genomatix (<http://genomatix.gsf.de/>), Bielefeld University's Faculty of Technology (<http://bibiserv.techfak.uni-bielefeld.de/>), PISE and the HGMP.

Internet On-Ramp

Match-box

Match-box (3) is a multiple sequence alignment tool that is designed to align blocks of conserved or similar regions within sets of protein sequences. The Match-box server is located at the University of Namur in Belgium (http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html). Its method of alignment does not take into account gap penalty values, and the program requires that only a few parameters be set. A study that compared Match-box to six other multiple sequence alignment programs (2) revealed that Match-box was able to predict, with a high degree of reliability, structurally conserved regions between sets of protein sequences of known structure. Match-box scores each column in an alignment block to indicate the reliability of the aligned sequences within that block. This feature is not available with many other programs.

A comprehensive listing of multiple sequence analysis tools available from the Internet can be found at the VSNS BioComputing Division at Bielefeld University's Faculty of Technology (<http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/welcome.html>).

Tools for Improving the Appearance of Multiple Sequence Alignments

So far, tools that generate multiple sequence alignments have been discussed. However, many of these programs display their results in their own formats, which are often difficult to read and may not be suitable for publication. This section will discuss a few tools that take multiple sequence alignments and reformat them into publication-quality figures. Note that these programs do not actually generate the multiple sequence alignments themselves.

BoxShade (written by Kay Hofmann and Michael Baron) is a tool that reformats multiple sequence alignments. It takes an alignment and shades or colors the nucleotides or amino acids of each sequence according to their degree of similarity in the alignment. This shading by similarity is either based on an identity threshold for residues in each column of the alignment or according to a specific sequence. The program also provides the option of displaying a consensus sequence beneath the alignment. The resulting file can be saved in a variety of different formats that can be imported into common word processors, image editors or desktop publishing packages. BoxShade is available on the Web through the Swiss Node of the European Molecular Biology network (http://www.ch.embnet.org/software/BOX_form.html) and PISE. Both servers allow multiple alignments to be pasted or uploaded into the program and are capable of shading alignments in varying tones of gray (depending on the level of similarity).

Two other programs that format multiple alignments are Pretty and PrettyBox from the Wisconsin Package (GCG). Pretty generates and displays consensus sequences in an alignment, but it does not shade residues according to lev-

els of similarity. PrettyBox works essentially the same way as Pretty, though the program also provides shading features that make it function similarly to BoxShade. However, unlike BoxShade, PrettyBox gives the investigator additional control over the shading thresholds for different levels of similarity. Integrated bioinformatics Web sites can allow the investigator to align multiple sequences using a program such as ClustalW and to improve the appearance of the resulting alignment using PrettyBox or BoxShade. Both functions can be accomplished with relative ease in contrast to several other stand-alone multiple sequence alignment services that require several manual steps between different Web sites to achieve the same result.

The Internet is a significant resource to the bioscientist and is one of the easiest means of gaining access to a wide range of powerful bioinformatics tools for the analysis of sequence data. While there are several multiple sequence alignment programs available on the Web, it is important to realize that the results they generate can be used in other analytical tools, such as those designed for molecular phylogenetic or protein molecular modeling studies. Integrated Web servers such as BioNavigator greatly facilitate the ability to conduct such studies.

ACKNOWLEDGMENT

I greatly appreciate the advice and help of Dr. Bruno Gaëta, Senior Scientist at eBioinformatics, in the preparation of this manuscript.

REFERENCES

1. Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
2. Briffeuil, P., G. Baudoux, C. Lambert, X. De Bolle, C. Vinals, E. Feytmans and E. Depiereux. 1998. Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions. *Bioinformatics* 14:357-366.
3. Depiereux, E., G. Baudoux, P. Briffeuil, I. Reginster, X. De Bolle, C. Vinals and E. Feytmans. 1997. Match-Box_server: a multiple sequence alignment tool placing emphasis on reliability. *Cabios* 13:249-256.
4. Feng, D.F. and R.F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 25:351-360.
5. Morgenstern, B., A. Dress and T. Werner. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA* 93:12098-12103.
6. Needleman, S.B. and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443-453.
7. Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85:2444-2448.
8. Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.
9. Sneath, P.H. and R.R. Sokal. 1973. *Numerical Taxonomy*. Freeman, San Francisco, CA.
10. Thompson, J.D., D.G. Higgins and T.J. Gibson. 1994. ClustalW: 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.
11. Thompson, J.D., F. Plewniak and O. Poch. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 27:2682-2690.