International Scholars Journals

Frontiers of Agriculture and Food Technology ISSN 7295-2849 Vol. 9 (10), pp. 001-005, October, 2019. Available online at www.internationalscholarsjournals.org © International Scholars Journals



*Review*

# Web-based bioinformatic resources for protein and nucleic acids sequence alignment

**Kamel A. Abd-Elsalam**

Molecular Markers Lab., Plant Pathology Research Institute, Agricultural Research Center, Orman 12619, Giza, Egypt.
E-mail: kaabdelsalam@msn.com.

Accepted 14 May, 2019

**DNA sequencing is the deciphering of hereditary information. It is an indispensable prerequisite for many biotechnical applications and technologies and the continual acquisition of genomic information is very important. This opens the door not only for further research and better understanding of the architectural plan of life, but also for future clinical diagnosis based on the genetic data of individuals. Bioinformatics can be broadly defined as the creation and development of advanced information and computational techniques for problems in biology. More narrowly, bioinformatics is the set of computing techniques used to manage and extract useful information from the DNA/RNA/protein sequence data being generated (at high volumes) by automated techniques (e.g. DNA sequencers, DNA microarrays) and stored in large public databases (e.g. GenBank, Protein DataBank). Certain method for analyzing genetic/protein data has been found to be extremely computationally intensive, providing motivation for the use of powerful computers. The advent of the Internet and the World Wide Web has substantially increased the availability of information and computational resources available to experimental biologists. This review will describe the current on-line resources available, including protein and nucleic acids sequence alignment.**

**Key words:** Sequence alignment, DNA, Protein, ClustalW, FASTA.

## INTRODUCTION

Bioinformatics is the application of Information technology to store, organize and analyze the vast amount of biological data which is available in the form of sequences and structures of proteins (the building blocks of organisms) and nucleic acids (the information carrier). The biological information of nucleic acids is available as sequences while the data of proteins is available as sequences and structures. Sequences are represented in single dimension where as the structure contains the three dimensional data of sequences. Sequence alignment is by far the most common task in bioinformatics. Procedures relying on sequence comparison are diverse and range from database searches (Altschul et al., 1990) to secondary structure prediction (Rost et al., 1994). Usually sequences either protein or DNA come in families. Sequences in a family

have diverged from each other in their primary sequence during evolution, having separated either by a duplication in the genome or by speciation giving rise to corresponding sequences in related organisms. In either case they normally retain a similar function. If you have already a set of sequences belonging to the same family is available, one can perform a database search for more members using pairwise alignments with one of the known family members as the query sequence (e.g. BLAST). However pairwise alignments with any one of the members may not find sequences distantly related to the ones you already have. An alternative approach is to use statistical features of the whole set of sequences in the search. Such features can be captured by a multiple sequence alignment. This review summarizes and extends the below-mentioned on-line tools, which are publicly available, in the context of the analysis procedure for sequence alignment, and gives an overview of the most versatile and efficient websites.

## SEQUENCE ALIGNMENT METHODS

Alignment provides a powerful tool to compare related sequences, and the alignment of two residues could reflect a common evolutionary origin, or could represent common structural and/or catalytic roles, not always reflecting an evolutionary process. Deletions, insertions and single residue substitutions are generally emphasized by alignments. Deletions or insertions are represented by null characters, added to one of the sequences, which will be aligned with letters in the other sequence (Rehm, 2001). There are various forms of sequence alignment. Alignments can be made between sequences of the same type (for example, between the primary structures of proteins) or between sequences of different type (for example, alignment of a DNA sequence to a protein sequence or of a protein to a three-dimensional structure). *Pairwise alignment* involves only two sequences, whereas *multiple sequence alignment* involves more than two sequences (although the term sometimes encompasses pairwise alignment also). *Global* alignment aligns whole sequences, whereas *local* alignment aligns only parts of sequences.

Database searches to extract homologous sequences are at the heart of sequence analysis, hence a variety of methods have been developed and applied in widely available packages or as network servers. In general, there is a trade-off between speed and sensitivity of the algorithms. The quick word search program FASTA (Pearson and Lipman, 1988) and the more recent and even faster BLAST (Altschul et al., 1990) are now the workhorses of database searching.

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. 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 (Gaskell, 2000). Sequence aliginment search tools may be divided into two groups, illustrated below.

## PAIRWISE SEQUENCE ALIGNMENT (PSA)

Pairwise and multiple alignment therefore continue to be among the most active areas of bioinformatics research. Pairwise sequence alignment given two DNA or protein sequences, find the best match between them. In such a match, there is a penalty for opening gaps or extending gaps for each of the sequences and for nucleotide/amino acids that are different. The best match is the one with the minimum sum of such penalties. Pairwise comparison provides computer tools to directly compare two sequences, either nucleic acid or peptide. They are the starting points for all kinds of sequence analysis. These tools are very useful when verifying sequence data, cloning projects, PCR analysis, and many more.

Most sequence alignment methods seek to optimize the criterion of similarity. There are two modes of assessing this similarity, local and global. Local methods try to determine if subsegments of one sequence (A) are present in another (B). These methods have their greatest utility in data base searching and retrieval (e.g. BLAST, Altschul et al., 1990). Although they may be of utility in detecting sequences with a certain degree of similarity that may or may not be homologous, in phylogenetic analysis it is assumed that the sequences being compared are orthologous. Global methods make comparisons over the entire lengths of the sequences; in other words, each element of sequence A is compared with each element in sequence B. Global comparison is the principal method of alignment for phylogenetic analysis. Several pairwise sequence alignment programs available on the Web (Table 1).

## MULTIPLE SEQUENCE ALIGNMENT (MSA)

Comparison of multiple sequences can reveal gene functions that are not evident from simple sequence homologies. As a result of genome sequencing projects, new sequences are often found to be similar to several un-characterized sequences, defining whole families of novel genes with no informative BLAST or FASTA similarities. However, such a family enables the application of efficient alternative similarity search methods. Software packages are available that derive profiles from a multiple sequence alignment. Profiles incorporate position-specific scoring information that is

**Table 1.** Parwise sequence alignment links.

| Server/Topic | Web | Reference |
|---|---|---|
| **1-BLAST 2 SEQUENCES:**This tool produces the alignment of two given sequences using BLAST engine for local alignment. The stand-alone executable for blasting two sequences (bl2seq) can be retrieved from NCBI ftp site | http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html<br>http://genopole.toulouse.inra.fr/blast/wblast2.html<br>http://embnet.cifn.unam.mx/blast/wblast2_cs.html<br>http://humangen.med.ub.es/tools/blast%202%20sequencias.htm<br>http://210.99.102.240/blast/wblast2.html | Tatusova and Madden, 1999. |
| **2-lalign and prss:** Two-sequence alignment server | http://www.isrec.isb-sib.ch/experiment/ALIGN_form.html | Pearson, 1990. |
| **3-ALIGN:** Query using sequence data | http://xylian.igh.cnrs.fr/bin/align-guess.cgi<br>http://genome.cs.mtu.edu/align/align.html | |
| **4-LAGAN:** a system for rapid global alignment of two homologous genomic sequences | http://lagan.stanford.edu | Brudno et al., 2003. |
| **5-Spidey:** is an mRNA-to-genomic alignment program. | http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/ | |
| **6-LALIGN:** find multiple matching subsegments in two sequences | http://www.ch.embnet.org/software/LALIGN_form.html | Huang and Miller, 1991. |
| **7-Parwise FLAG:** Fast Local Alignment for Gigabases; *Industrial Technology Research Institute, Taiwan*) performs local alignment for two different DNA sequences with newly featured Alignment Plot. | http://bioinformatics.itri.org.tw/prflag/prflag.php | Eisen et al., 2000. |
| **8-BCM Search Launcher:** Pairwise Sequence Alignment | http://searchlauncher.bcm.tmc.edu/seq-search/alignment.html | Smith et al., 1996. |
| **9-PipMaker:** computes alignments of similar regions in two DNA sequences. The resulting alignments are summarized with a ``percent identity plot'', or ``pip'' for short. All pairwise alignments with the first sequence are computed and then returned as interleaved pips. | http://bio.cse.psu.edu/pipmaker/ | Schwartz et al. , 2000. |
| **10-SIM:** finds k best non-intersecting alignments between two sequences or within a sequence using dynamic programming techniques. The alignments are reported in order of decreasing similarity score and share no aligned pairs. SIM requires space proportional to the sum of the input sequence lengths and the output alignment lengths, so it accommodates 100,000-base sequences on a workstation. | http://us.expasy.org/tools/sim-prot.html | Huang and Miller, 1991. |
| **11-Lalign:** server compares two sequences using lalign, which is Pearsons implementation of the 'sim'- algorithm of Huang and Miller. The program returns the best *n* local alignments that do not share aligned symbol pairs. | http://enterprise.molbiol.ox.ac.uk/cgi-bin/lalign.cgi<br>http://fasta.bioch.virginia.edu/fasta_www/lalign.htm | Huang and Miller, 1991. |
| **12-BALSA:** Bayesian Algorithm for Local Sequence Alignment | http://bayesweb.wadsworth.org/cgi-bin/BALSA_file.pl | Webb et al., 2002. |
| **13-Fasta:** Provides sequence similarity and homology searching against nucleotide and protein databases using the Fasta programs. Fasta can be very specific when identifying long regions of low similarity especially for highly diverged sequences. | http://www.ebi.ac.uk/fasta33/ | Pearson, 1990 |
| **14-PRSS3**: evaluates the significance of a protein sequence alignment | http://www.ch.embnet.org/software/PRSS_form.html | Pearson and Lipman, 1988. |
| **15-The Wise2:** form compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors. | http://www.ebi.ac.uk/Wise2/<br>http://bioweb.pasteur.fr/seqanal/interfaces/wise2.html | |
| **16-SCAN2:**: program for aligning two multimegabyte-size sequences | http://www.softberry.com/berry.phtml?topic=scanh&prg=SCAN2 | |

derived from the frequency with which a given residue is seen in an aligned column. Because sequence families preferentially conserve certain critical residues and motifs, this information should allow more sensitive database searches. Most new profile software is based on statistical HMMs. Much more comprehensive reviews of the literature on profile HMM methods are available elsewhere (Eddy 1996; Baldi and Brunak, 1998; Durbin

**Table 2.** Multiple sequence alignment links.

| Server/Topic | Web | Reference |
|---|---|---|
| **1-Clustal W:** is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. | http://www.ebi.ac.uk/clustalw/<br>http://www.ch.embnet.org/software/ClustalW.html<br>http://bioweb.pasteur.fr/seqanal/interfaces/clustalw.html<br>http://transfac.gbf.de/programs/clustalw/clustalw.html<br>http://www.clustalw.genome.ad.jp/<br>http://pbil.ibcp.fr/<br>http://www.bionavigator.com | Thompson et al., 1994 |
| **2-T-COFFEE:** This program is more accurate than ClustalW for sequences with less than 30% identity, but it is slower | http://www.ch.embnet.org/software/TCoffee.html<br>http://igs-server.cnrs-mrs.fr/Tcoffee/<br>http://www.es.embnet.org/Services/MolBio/t-coffee/<br>http://www.cmbi.kun.nl/bioinf/tools/T_COFFEE/ | Notredame et al., 2000 |
| **3-Multi-LAGAN:** a system for multiple global alignment of genomic sequences. | http://lagan.stanford.edu | Brudno et al. ,2003 |
| **4-MAP:** Multiple Alignment Program | http://genome.cs.mtu.edu/map.html | Huang, 1994 |
| **5-MultiAlin:** Multiple sequence alignment | http://prodes.toulouse.inra.fr/multalin/multalin.html | Corpet, 1988 |
| **6-MATCH-BOX:** The Match-Box software proposes protein sequence multiple alignment tools based on strict statistical criteria. The method circumvents the gap penalty requirement: in the Match-Box method, gaps are the result of the alignment and not a governing parameter of the matching procedure. | http://www.ch.embnet.org/software/BOX_form.html<br>www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html | Depiereux et al., 1997 |
| **7-MUSCA:** Multiple Sequence Alignment | http://cbcsrv.watson.ibm.com/Tmsa.html | Parida et al., 1998 |
| **8-MAFFT:** is a fast and highly accurate method for multiple sequence alignment. | http://bioinformatics.uams.edu/mafft/ | Kuma and Miyata, 2002 |
| **9-AMAS:** Analyze Multiply Aligned Sequences | http://barton.ebi.ac.uk/servers/amas_server.html | |
| **10-AltAVisT:** *(Al*ternative Alignment *Vi*sualization Tool), a WWW-based tool that compares two different multiple alignments of a given data set and highlights regions where both alignments coincide. | http://bibiserv.techfak.uni-bielefeld.de/altavist/ | Morgenstern et al., 2003 |
| **11-DIALIGN 2:** DNA and protein sequence alignment based on segment-to-segment comparison | http://bibiserv.techfak.uni-bielefeld.de/dialign/<br>http://kun.homelinux.com/Pise/5.a/dialign2.html<br>http://bioweb.pasteur.fr/intro-uk.html<br>www.hgmp.mrc.ac.uk/ | Morgenstern, 1999 |

et al., 1998; Krogh, 1998). There are a numerous web-based resources for multiple sequence alignment (Table 2).

## OUTLOOK

The growth in output of DNA sequence data has run in parallel with the well known exponential growth rate of computing power, as well as with the advent and exploitation of the Internet and the World Wide Web (WWW). These circumstances have encouraged the development of biological databases, and nucleotide and protein analysis tools, so that a vast array of tools and databases is now available. I have focused on a number of different sequence alignment tools available as services over 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. Also, these tools are very useful when verifying sequence data, cloning projects, PCR analysis, and many more. Sequence alignment plays a central role in the bioinformatics research no matter whether it is realized or not.

## REFERENCES


Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Baldi P, Brunak S (1998). Bioinformatics: The Machine Learning Approach, MIT Press.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S (2003). NISC Comparative Sequencing Program LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DANN. Genome Res. 13:721-731.

Corpet F (1988). Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 16: 10881-10890.

Depiereux E, Baudoux G, Briffeuil P, Reginster I, De Bolle X, Vinals C, Feytmans E (1997). Match-Box_server: a multiple sequence alignment tool placing emphasis on reliability. Comput. Appl. Biosci. 13:249-256.

Durbin R, Eddy SA, Krogh A, Mitchison G (1998). Biological sequence analysis probabilistic models of proteins and nucleic acids.

Cambridge University Press.
Eddy SR (1996). Hidden Markov Models. Curr. Opin. Struct. Biol. 6: 361-365.
Eisen JA, Heidelberg JF, WhiteO, Salzberg SL (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. 1:1-19.
Gaskell JG (2000). Multiple Sequence Alignment Tools on the Web. BioTechniques 29: 60-62.
Huang X (1994). On Global Sequence Alignment. Comput. Appl. Biosci. 10: 227-235.
Huang X, Miller W (1991). A time-efficient, linear-space local similarity algorithm. Adv. Appl. Math. 12:337-357.
Krogh A (1998). Computational Methods in Molecular Biology (Salzberg S, Searls D, Kasif S, eds), pp. 45-63, Elsevier Science.
Kuma K, Miyata T (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acid Res. 30:3059-3066.
Morgenstern B (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211-218.
Morgenstern B, Goel S, Sczyrba A, Dress A (2003). AltAVisT: Comparing alternative multiple alignments. Bioinformatics 19:425-426.
Notredame C, Higgins D, Heringa J (2000). T-Coffee: A novel method for multiple sequence alignments. J. Mol. Biol. 302: 205-217. Parida L, Floratos A Rigoutsos I (1998). MUSCA: An Algorithm for Constrained Alignment of Multiple Data Sequences. Proceedings 9th Workshop on Genome Informatics, Tokyo, Japan. December 1998.
Pearson WR (1990).Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 183:63-98.
Pearson WR, Lipman DJ (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85:2444-2448,
Rehm BHA (2001). Bioinformatic tools for DNA/protein sequence analysis,functional assignment of genes and protein classification. Appl. Microbiol. Biotechnol. 57:579–592.
Rost B, Sander C, Schneider R (1994). PHD - anautomatic server for protein secondary structure prediction. Comput. Appl. Biosci. 10: 53-60.
Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000). PipMakerA Web Server for Aligning Two Genomic DNA Sequences. Genome Res. 10:577-586.
Smith RF, Wiese BA, Wojzynski MK, Davison DB, Worley KC (1996). BCM Search Launcher--An Integrated Interface to Molecular Biology Data Base Search and Analysis Services Available on the World Wide Web. Genome Res. 6:454-62.
Tatusova TA, Madden TL (1999). Blast 2 sequences - a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. 174:247-250.
Thompson JD, Higgins DG, Gibson TJ (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. 11:4673-4680.
Webb BM, Liu JS, Lawrence CE (2002). BALSA: Bayesian Algorithm for Local Sequence Alignment. Nucleic Acids Res. 30:1268-1277.