Sequence Alignment in Bioinformatics -
A Historical Perspective

Sequence alignment in bioinformatics is a field of research focused on developing tools for comparing and finding similar sequences of amino acids or DNA base pairs with the aid of computers. The sequence similarity is used to assess gene and protein homology, classify genes and proteins, predict biological function, secondary and tertiary protein structure, detect point mutations, construct evolutionary trees, etc.

There are two main areas of sequence alignment: pairwise sequence alignment and multiple sequence alignment (see figure).

Pairwise Sequence Alignment

Pairwise sequence alignment is concerned with comparing two DNA or aminoacid sequences – finding the global and local "optimum alignment" of the two sequences. Based on differences between the two sequences, one can calculate the "cost" of aligning the two sequences by using replacements, deletions and insertions, and assign a similarity score. The problem has tractable solutions by means of dynamic programming and Hidden Markov Models and is the basis of popular heuristic search methods such as FASTA or BLAST.

Needleman and Wunsch (1970) were the first to present a dynamic programming algorithm that could find the **global alignment** between two aminoacid sequences. Smith and Waterman (1981) introduced a new algorithm with a different method of scoring similarity aimed at finding optimum **local alignment** sub-sequences, at the expense of the global score. Global algorithms are generally not sensitive for highly diverged sequences with some localized similarities within them.

A particular application of pairwise sequence alignment is *quickly* searching large DNA and protein databases for matches to a query sequence. Popular heuristic algorithms, such as those from the **FASTA** (Pearson and Lipman 1985, 1988) or **BLAST** (Altschul et al 1990, 1997) families, are much faster than algorithms based on dynamic programming.

Multiple Sequence Alignment (MSA)

Multiple sequence alignment aims to find similarities between many sequences. MSA is hard and less tractable than pairwise alignment. Dynamic programming is impractical for a large number of sequences.

The most successful MSA solutions are heuristic algorithms with approximate approaches, such as the **CLUSTAL** family of programs created by Higgins, which use a progressive algorithm (Feng and Doolittle 1987): CLUSTAL (1988), ClustalV (1992), ClustalW (1994), ClustalX (1998). Profile Hidden Markov Models (**HMMs**) provide another successful solution to the problem of MSA. They were introduced by Krogh and colleagues in 1994.

Substitution matrices

Both pairwise and multiple sequence alignment algorithms use substitution matrices to score the sequence alignment. In substitution matrices each possible residue substitution is given a score reflecting the probability of such a change. There are two popular protein substitution matrix models: **P**ercent **A**ccepted **M**utation (**PAM** - Dayhoff 1978) and **Blo**cks **Su**bstitution **M**atrix (**BLOSUM** - Henikoff and Henikoff 1992).

Two Sample Applications

Sequence alignment algorithms are often used to characterize newly sequenced genes or gene products. For example, the sequenced genome of the SARS virus was investigated by using BLAST, FASTA, Pfam, and ClustalX to find proteins with sequences similar to those expected to be produced by the SARS virus ORFs (Mara 2003, Rota 2003). Biological function and structure was then predicted for the SARS proteins based on the information available for the homologous proteins.

Another application of sequence alignment tools is the study of phylogenetics. Phylogenetics is a field of molecular evolution that correlates mutations in DNA and protein sequences with evolutionary divergence. Molecular distances of evolution between species can be calculated using various metrics based on DNA or protein sequence difference. The smaller the number of differences in the DNA and/or protein sequences of similar genes from two related organisms, the less they have evolutionarily diverged from each other.

References

Needleman, S.B. & Wunsch, C. D. (1970), "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins", *J.Mol. Biol*., vol. 48, pp. 443-453.

Smith, T. F. & Waterman, M. S. (1981), "Identification of common molecular subsequences", *J. Mol. Biol.*, vol. 147, pp 195-197.

Dayhoff, M.O. et al. (1978), "A model of evolutionary change in protein", *Atlas of Protein Sequence and Structure*. Vol. 5, Suppl. 3 National Biomedical Reserach Foundation, Washington D.C. U.S.A, pp 345-352.

Henikoff, S. & Henikoff J.G. (1992), "Amino acid substitution matrices from protein blocks", *PNAS* , vol. 89, pp 10915-10919

Lipman, D.J. & Pearson, W.R. (1985), "Rapid and sensitive protein similarity searches", *Science*, vol. 227, no. 4693, pp. 1435-41

Pearson, W.R. & Lipman, D.J. (1988), "Improved tools for biological sequence comparison", *Proceedings of the National Academy of Sciences USA***,** vol 85, pp. 2444-2448.

Altschul, S.F. et al (1990), "Basic local alignment search tool", *Journal of Molecular Biology***,** vol. 215, pp. 403-410. The 1990 BLAST paper was the most highly cited paper published in the 1990s (cited more than 10,000 times, August 2000, data from www.sciencewatch.com).

Altschul, S.F. et al (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389-402.

Karlin, S. & Altschul, S.F. (1990) "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." Proc. Natl. Acad. Sci. USA, vol. 87, pp. 2264-2268

Higgins, D.G. & Sharp P.M. (1988), “CLUSTAL: a package for performing multiple sequence alignment on a microcomputer”, *Gene*, vol. 73, pp 237-44.

Higgins, D.G et al. (1992), “ClustalV—improved software for multiple sequence alignment”, *Comput. Appl. Biosci*., vol. 8, pp. 189-91. It is the most-cited paper over the past decade in the field of Computer Science, with 1,886 citations to date. (February 2003, data from http://www.in-cites.com/scientists/DesHiggins.htm)

Higgins, D.G et al. (1994), “ClustalW—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”, *Nucl. Acid Res*., vol. 22, no. 22, pp 4673-80. It is the most-cited paper in the past decade in the field of Biology & Biochemistry, with 8,764 citations to date. (February 2003, data from http://www.in-cites.com/scientists/DesHiggins.htm)

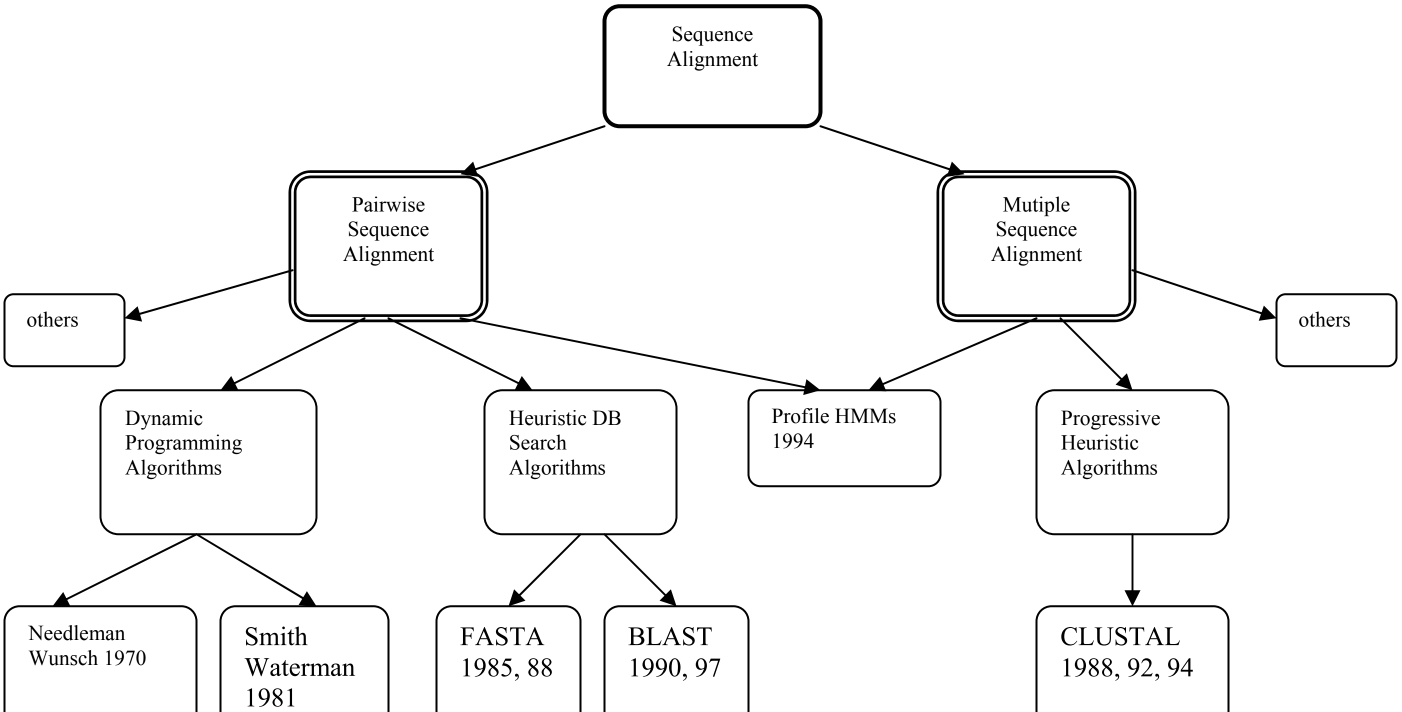
Feng D. & Doolittle R. F (1987), “Progressive sequence alignment as a prerequisite to correct phylogenetic trees”, *J. Mol. Evol.*, vol. 60, pp 351-360.

Krogh, A. et al. (1994), “Hidden Markov models in computational biology: applications to protein modeling”, *J. Mol. Biol.*, vol. 235, pp. 1501-1531.

Bateman A. et al. (2000), “The Pfam protein families database”, *Nucleic Acids Res*., vol. 28, no. 1, pp. 263-266.

Mara M.A. et al (2003), “The Genome Sequence of the SARS-Associated Coronavirus”, *Science*, vol. 300, no. 5624, pp. 1399-1404.

Rota, P.A. et al (2003), “Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome”, *Science*, vol. 300, no. 5624, 1394-1399.


Sequence Alignment
Pairwise Sequence Alignment
Mutiple Sequence Alignment
others
others
Dynamic Programming Algorithms
Heuristic DB Search Algorithms
Profile HMMs 1994
Progressive Heuristic Algorithms
Needleman Wunsch 1970
Smith Waterman 1981
FASTA 1985, 88
BLAST 1990, 97
CLUSTAL 1988, 92, 94

Proposal for the Annotation of the *Bioinformaticus Maximus* Genome

A major epidemic caused by *Bioinformaticus Maximus* was recently reported on the Yale University Campus. In record time, the virus was isolated, its genome was sequenced and the ORFs were identified. The next step in the process of learning about the virus involves annotating the proteins encoded by the ORFs.

I propose the use of bioinformatics techniques to identify for each ORF a set of homologous proteins from other organisms. Structure and function can then be predicted for the *Bioinformaticus Maximus* proteins based on the homologous proteins' annotation. I propose a gradual approach to identify the homologous protein(s) for each ORF based on protein sequence and domain similarity. Increasingly sensitive algorithms will be used to narrow down the field of possible matches. In this way, a balance between speed and sensitivity will be attained. The search will focus on querying protein databases rather than nucleotide databases because amino acid searches are more sensitive. Similar techniques were recently used to successfully characterize the SARS virus genome (Marra 2003, Rota 2003), and are in line with the TIGR methodology for annotating new microbial genomes.

In the first stage (see figure), each ORF sequence will be queried against a comprehensive database containing all the known protein sequences. Such a database would be a non-repeating union of protein databases such as GenBank, Swiss-Prot and TIGR's CMR. A fast running heuristic search algorithm, such as one from the BLAST (Altschul et al 1997) or FASTA (Pearson and Lipman 1988) families, will be used. For example, the tblastx algorithm could be used to compare the six-frame translation of each ORF sequence against the protein database. If the ORF sequence is longer than1000 bp, it will be divided into smaller sub-sequences. For each ORF sequence, a set of global matches above a certain score will be obtained.

In the second stage, a slower but more sensitive algorithm, such as the dynamic programming based Smith-Waterman (1981) algorithm, will be run for each ORF sequence against the set of matches found in the first stage. For each ORF, a new subset of matches will be obtained, containing the protein sequences with the highest local (and global, from the first state) sequence match scores.

In the third stage, a Hidden Markov Model (HMM) based algorithm against the Pfam database (Bateman 2000) will be used to determine for each ORF sequence the protein or protein families that have the most similar sequence and domain. In the fourth stage, the resulting sets from the second and third stages will be intersected, to obtain the most similar homologous protein(s) with respect to global alignment, local alignment and domain.

The degree of similarity will then be used to confer annotation to the *Bioinformaticus Maximus* proteins. For example, a homologous protein obtained in the last stage will determine a high probability of homology and certainty in predicted annotation. If no protein is obtained in a stage, the protein with the highest score from the previous stage will be selected. The certainty in the annotation level will be decreased. It is expected that some ORFs will return no significant matches in any of the query stages, indicating that no meaningful homolog could be found.

A final, manual curation, stage, will be required to ensure that the annotation is meaningful. Intermediate query results from each of the previous stages (I through IV) will be available for review. This procedure should allow for a fairly quick and cost-effective annotation of most of the ORFs of the *Bioinformaticus Maximus* genome. The knowledge could be used to understand its virulence and make sure that no antidote is found.

References

Mara M.A. et al (2003), “The Genome Sequence of the SARS-Associated Coronavirus”, *Science*, vol. 300, no. 5624, pp. 1399-1404.

Rota, P.A. et al (2003), “Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome”, *Science*, vol. 300, no. 5624, 1394-1399.

TIGR – The Institute for Genomic Research. Additional information about the Annotation Engine available at:
http://www.tigr.org/edutrain/training/annotation_engine.shtml

Smith, T. F. & Waterman, M. S. (1981), "Identification of common molecular subsequences", *J. Mol. Biol.*, vol. 147, pp 195-197.

Lipman, D.J. & Pearson, W.R. (1985), “Rapid and sensitive protein similarity searches”, *Science*, vol. 227, no. 4693, pp. 1435-41

Pearson, W.R. & Lipman, D.J. (1988), “Improved tools for biological sequence comparison”, *Proceedings of the National Academy of Sciences USA***,** vol 85, pp. 2444-2448.

Altschul, S.F. et al (1997), “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”, *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389-402.

Bateman A. et al. (2000), “The Pfam protein families database”, *Nucleic Acids Res.*, vol. 28, no. 1, pp. 263-6.

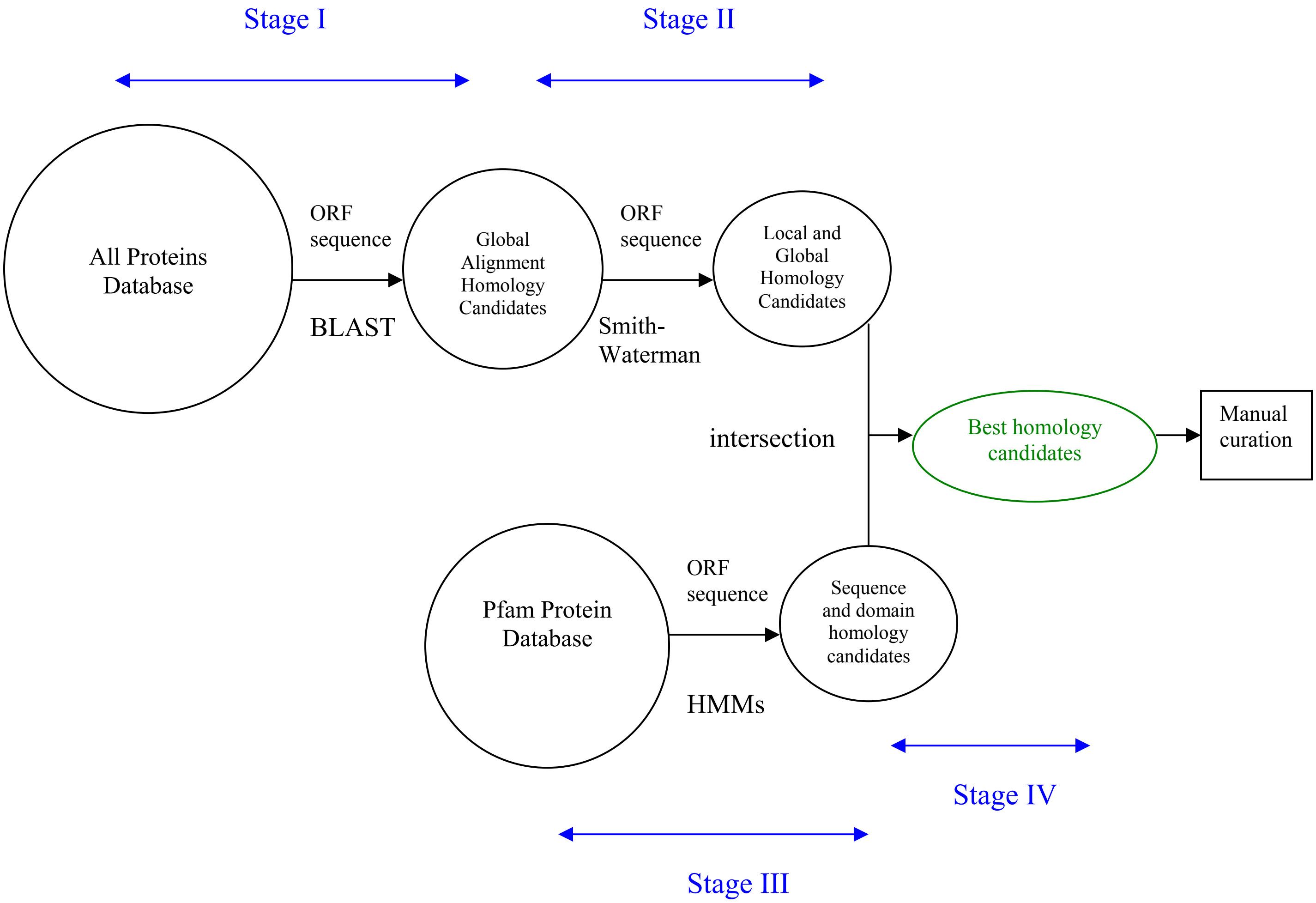


Fig. For each ORF sequence, the best homologous protein candidate is searched by means of sequence and domain similarity.