Sarah Fiorentino
MBB 752
12/12/03

Sequence Alignment Techniques and Their Uses

Since rapid sequencing technology and whole genomes sequencing, the amount of sequence information has grown exponentially. With all of this data, it is possible to do comparisons where one can learn about the structure, function, and evolutionary relationships of different parts of organisms. Sequence alignment techniques have been developed to do comparisons. The goal of alignment is to obtain the optimal alignment of sequences. Pairwise alignment techniques, where two sequences are studied at a time, were developed first with multiple alignment techniques, where many sequences are compared at once, coming later (Fig. 1).

Dot plots were developed by W.M. Fitch (1969) as a way to visualize similarities and differences between two sequences. Regions of similarity appear as diagonal lines on the matrix. The Needleman-Wunsch (1970) algorithm globally aligns two sequences, meaning the path starts at one edge and runs continuously to the other edge. This method does not penalize the score of the alignment for the insertion of gaps into the sequence (Vingron, 2002). The Smith-Waterman (1981) algorithm was developed to detect local alignments. This method allows paths to begin and end inside the matrix (Schuler, 1998). In 1983, the concept of using "words" to look for local similarities was developed (Wilbur and Lipman, 1983). Next, modifications to the Smith-Waterman algorithm to detect the best nonintersecting, suboptimal, local alignment were added (Altschul and Erickson, 1986; Waterman and Eggert 1987). The first substitution matrix, PAM (Percent Accepted Mutation), was developed to measure evolutionary distances (Dayhoff, et al, 1978). Another substitution matrix, BLOSUM, was developed that compared sequences based on their maximum level of identity (Henikoff and Henikoff, 1992). In the mid-1980s, techniques to search similarities within databases were created. The first was FASTA, which used substitution matrices to match words (Lipman and Pearson, 1985). The next technique was BLAST (Basic Local Alignment Search Tool), which uses neighborhoods of words and Karlin-Altschul statistics, but does not allow gaps (Altschul, at al, 1990). BLAST was modified to BLAST 2.0, which allowed gaps, and PSI-BLAST, which creates and iteratively refines profiles (Altschul, et al, 1997).

The first multiple alignment method was creating profiles, which iteratively applied pairwise alignments with a fixed alignment of a subgroup, thus allowing the determination of conserved patterns and creation of hierarchical trees (Gribskov, et al, 1987). The next method developed was CLUSTAL, which also uses profiles, and has been modified over time (Higgins, et al, 1992; Higgins et al, 1996: Higgins, et al, 1997). Two other variations of profiles are MultAlin by Corpet (1988) and generalized profiles by Bucher and Karplus (1996). The Hidden Markov Model (HMM) is a powerful way to align and search sequences that "learns" the characteristic traits of the sequence sets (Krogh, et al, 1994).

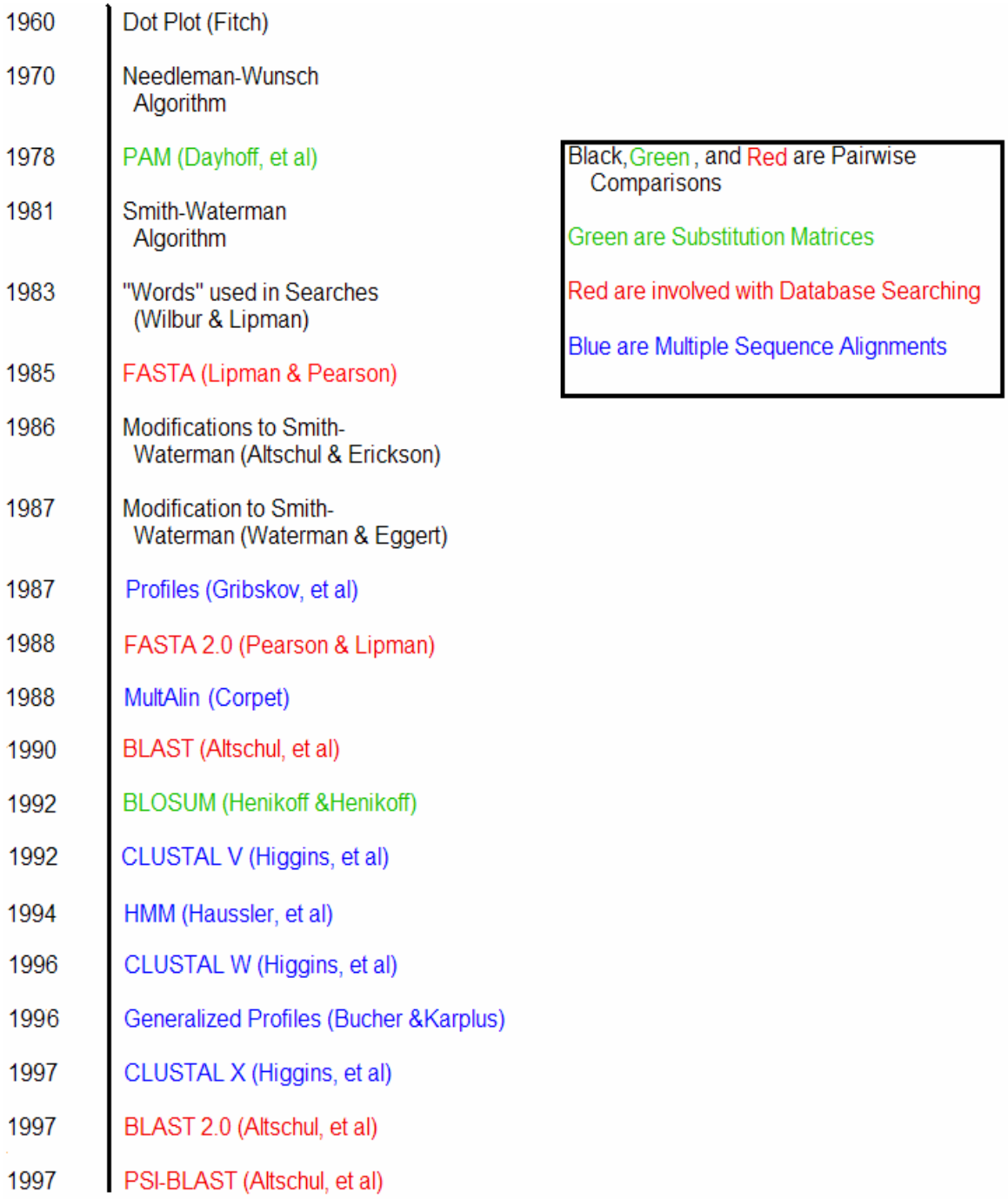
PAM matrices can be used to determine whether an amino acid substitution would be favored or avoided over time (Vingron, 2002). PAM matrices with lower numbers, e.g. PAM120, are closer in time than PAMs with higher numbers, and have a higher degree of similarity. Evolutionarily related proteins have biased amino acid frequencies that represent changes that have been accepted over time (Shuler, 1998).

Functions of unknown proteins can be determined by doing BLAST searches. Sequences that align best are most likely to have similar functions. If the function of the

best aligned protein is known, the function of the unknown protein can be inferred by "guilt by association", i.e. if it looks like it, it probably acts like it. Further biochemical experiments can then be performed to confirm the function.

Sarah Fiorentino
MBB 752
12/12/03

FIGURE 1. Timeline of Sequence Alignment Techniques



Sarah Fiorentino
MBB 752
12/12/03

References for "Sequence Alignment Techniques and Their Uses"

Altschul, S. F., and Erickson, B.W. (1986). Locally optimal subalignments using nonlinear similarity functions. *Bulletin of Mathematical Biology*. **48**: 633-660.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*. **215**: 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*. **25**: 3389-3402.

Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. *Computational Chemistry*. **20**: 3-23.

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

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In Dayhoff, M. O. (ed) *Altas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, D.C. pp. 345-352.

Fitch, W. M. (1969). Locating gaps in amino acid sequences to optimize the homology between two proteins. *Biochemical Genetics*. **3**: 99-108.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile Analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*. **84**: 4355-4358.

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*. **89**: 10915-10919.

Higgins, D. G., Bleasby, A. J., and Fuchs, R.(1992). CLUSTAL V: improved software for multiple sequence alignment. *Computer Applications in the Biosciences*. **8(2)**: 189-91.

Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods of Enzymology*. **266**: 383-402.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). Hidden Markov Models in computational biology: Applications to protein design. *Journal of Molecular Biology*. **235**: 1501-1531.

Lipman, D. J., and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*. **227**: 1435-1441.

Needleman, S. B. and Wunsch, C. (1970). A general method applicable the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. **48**: 443-453.

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*. **85**: 2444-2448.

Shuler, G. D., (1998). Sequence alignment and database searching. In Baxevanis, A. D., and Oullette, B. F. F. (eds) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Inc. New York. pp. 145-171.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*. **147**: 195-197.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin F., and Higgins, D. G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*. **25(24)**: 4876-82.

Vingron, M. (2002). Sequence Analysis. In Lengauer, T. (ed) *Bioinformatics – From Genomes to Drugs, Volume 1: Basic Technologies*. Wiley-WCH, Weinheim. pp. 27-58.

Waterman, M. S., and Eggert, M. (1987). A new algorithm for best subsequence alignments with applications to tRNA-rRNA comparisons. *Journal of Molecular Biology*. **197**: 723-728.

Wilbur, W. J., and Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences of the United States of America*. **80**: 726-730.

Sarah Fiorentino
MBB 752
12/12/03

Computational Functional Genomics Analyzes with a New Genome

First, I would perform various microarray experiments to obtain expression data that can later be combined with other data to predict protein function. I would then translate the ORFs into the amino acid sequence. Amino acid sequences are better for comparisons because are not as degenerate as the nucleotide sequence, and so would be more sensitive for doing alignments. Because this is a microbe, I am assuming it would be a bacterium; therefore I would not need to worry about splicing out introns.

I would then query PROSITE with the protein sequences to see if my genome contains any known patterns, such as localization tags or transmembrane domains (Wishart, 2001). The patterns could be combined with expression data and clustered to find any interactions among the proteins. Proteins that interact most likely are involved in the same cellular processes and thus have similar functions. A similar comparison could also be performed with expression data and proteins with localization tags. Specific genes have characteristic expression levels in different areas in the cell, such as in the cytoplasm, periplasm, or along the cell membrane. When these expression levels are combined with localization tags, prediction of protein subcellular localization can be made. If I know where something localizes, I can then make predictions about its function. A protein's function is often associated with where it is in the cell, for example, transmembrane domains in cell membranes will probably be associated with transport or motility.

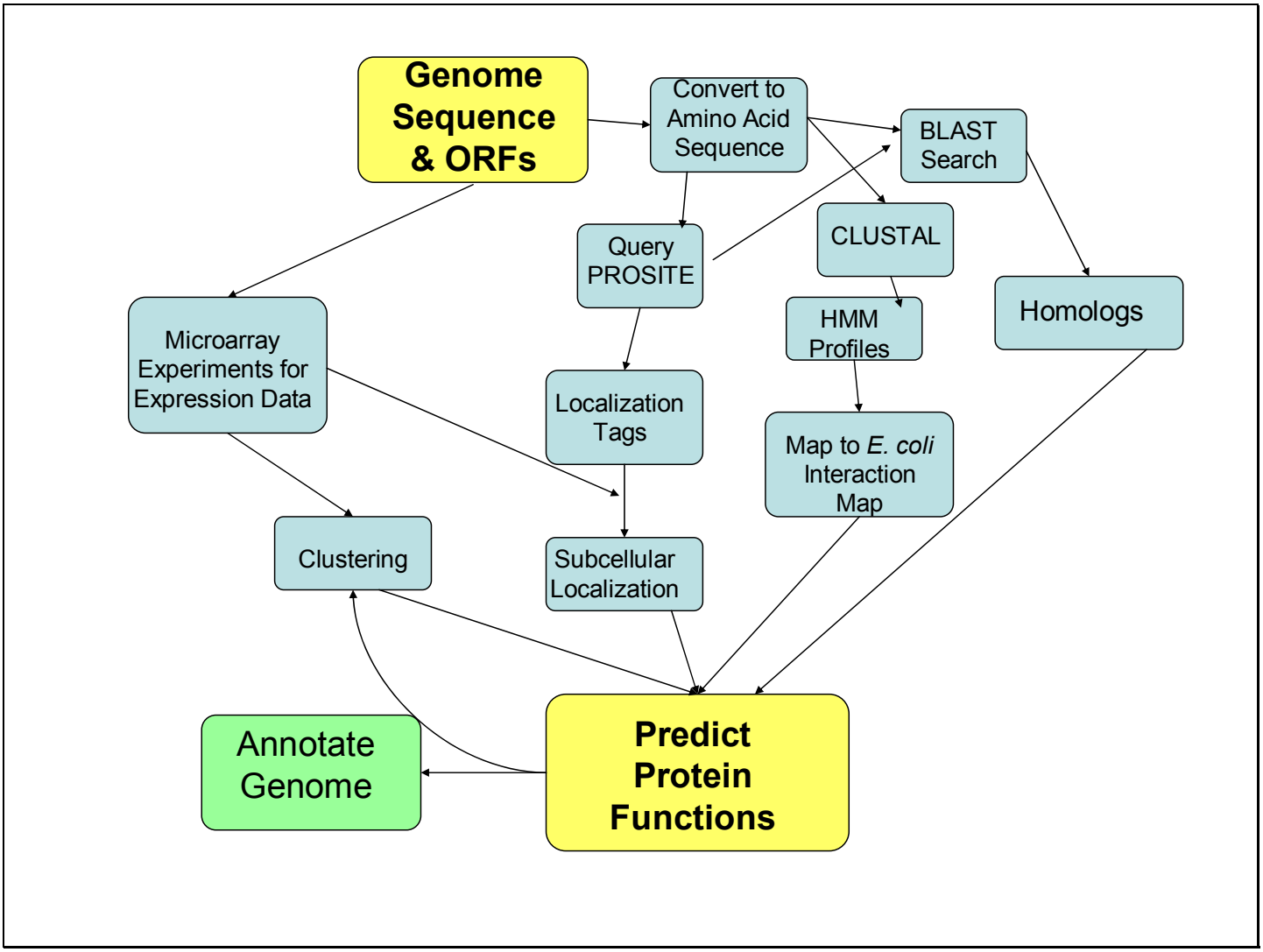
I could also take the protein sequences and divide them into segments of less than 300 amino acids, however making sure to keep motifs and patterns intact, and do a BLAST search. The BLAST search would produce proteins of, hopefully, known functions that align well with the unknown proteins. These proteins could be homologs to the unknown proteins, thus allowing me to infer the unknown proteins' functions, because proteins with similar sequences often have similar functions. I could do a PSI-BLAST search to find any orthologs or paralogs to the unknown proteins, which would allow me to know more about functions the proteins and their evolutionary history (Dunbrack, Jr., 2002).

Interactions between proteins could also be determined by mapping the unknown proteins to a known protein interaction map, e.g. *E. coli* interaction map. Pairwise comparisons with BLAST are first performed and then CLUSTAL would be used for multiple alignments (Schachter, 2002). From this data, HMM profiles would be made and compared to look for values below a certain threshold, which would define homology between the unknown proteins and *E. coli* (Schachter, 2002). Again homologies and similar interactions imply similar functions.

With all of the predictions for protein functions, the genome could be putatively annotated. The annotation would be verified as biochemical experiments are performed to verify the protein function predictions.

Sarah Fiorentino
MBB 752
12/12/03

Figure 1. Outline of Plan



Sarah Fiorentino
MBB 752
12/12/03

References for Computational Functional Genomics Analyzes with a New Genome

Dunbrack, Jr., R. L.. (2002). Homology in Modeling in Biology and Medicine. In Lengauer, T. (ed) *Bioinformatics – From Genomes to Drugs, Volume 1: Basic Technologies*. Wiley-WCH, Weinheim. pp. 145-235.

Schachter, V. (2002). Construction and Prediction of Protein-Protein Interaction Maps. In Mewes, H.-W., Seidel, H., and Weiss, B. (eds) *Bioinformatics and Genome Analysis*. Springer, Berlin. pp. 191-220.

Wishart, D. S. (2001). Tools for Protein Technologies. In Rehm, H.J., Reed, G., Puhler, A., and Stadler, P. (eds) *Biotechnology, Volume 5b* (2^{nd} ed). Wiley-VCH, Weinheim. pp. 325-344.