In its most basic form a sequence alignment is simply comparing two or more sequences by searching for character patterns and other similarities. Blasting some sequence is often the first step a researcher will take to characterize an unknown sequence or even a whole genome, but in 1970, when Needleman and Wunsch first introduced their algorithm for automated global alignment of sequences, there were still very few sequences to work with. Throughout the late seventies, determining nucleotide sequences was a torturous tedious process, but in 1977 there were two major technological breakthroughs, one pioneered by Maxam and Gilbert and the other by Sanger that opened the door for automated sequencing. These two developments represent the bedrock on which high through put sequencing is built.

Alignment techniques can be broken into two components. The first is construction of a scoring matrix, and the second is the actual algorithms used to compute a score. The simplest scoring matrix is a unitary matrix. If the character matches a one is assigned, and if not the matrix element is zero. In this most fundamental case, the alignment would simply be the path through this matrix. The two most commonly used scoring matrices today are PAM and BLOSUM. PAM or percent accepted mutation rate different substitution rates of amino acids were calculated based on alignments of protein sequences that were at least eight-five percent identical (Heinkoff, 1992). Rates for PAMs correlating to different evolutionary distances were then extrapolated from these original calculations. BLOSUM takes a slightly different approach. Instead of relying on extrapolated data, the various BLOSUMs are calculated directly. The heart of BLOSUM is BLOCKs. Instead of assuming that the evolutionary rate was homogenous across the entire protein, the amino acid substitution matrices were constructed from protein blocks (Heinkoff, 1992).

There are two basic flavors of alignments – global and local. In a global alignment, the goal is to find the best score across the entire query sequence, but in a local alignment, the goal is to find regions of high similarity. Needleman and Wunsch first described an automated method for finding global alignments by using dynamic programming to compute the sum matrix, and then looking in the last row or column to find the highest score (Needleman, 1970). Smith-Waterman modified this algorithm to find the highest score anywhere in the matrix, and thus pull out subsequences that had degrees of similarity (Smith 1981).

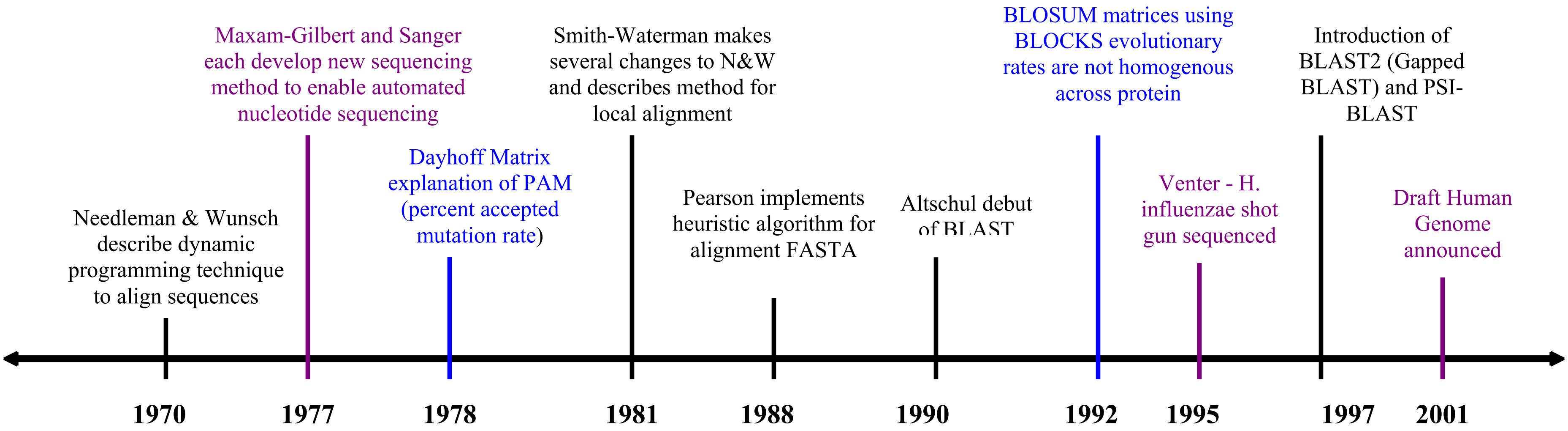
Dynamic programming algorithms are quadratic time, and therefore much too slow and computationally intense to be used for searching large databases. A number of heuristic algorithms have been developed in response to this problem. One of the first programs to attain widespread use was Pearson's FASTA. The guts of FASTA involved creating a hash table of short words from query and comparing them with the database (Pearson, 1988). Shortly after, the first version of BLAST came out. BLAST differed from FASTA in the notion of high scoring segment pairs. Whereas FASTA joined together word hits into diagonals to create the alignment, BLAST does compile a list of high scoring words or high scoring segment pairs which it then scans database for hits with, but BLAST then attempts to extend those hits using some threshold value (Altschul, 1990). Both Gapped BLAST and PSI-BLAST were introduced in 1997. Gapped BLAST allowed the introduction of gaps in aligning sequences, and PSI-BLAST uses a position specific scoring matrix iteratively to create optimal alignment (Altschul, 1997). In addition to sequence alignment, the same basic tools described can and have been applied to the somewhat more complex problem of structural alignment.

One can never truly know if the best scoring alignment actually corresponds to any biological relationship. It is assumed that similar sequence should mean similar protein, and perhaps similar function, but there are also many examples where proteins with a high degree of identity have vastly different functions, and there are also many examples where proteins with low degree of identity have similar functions. Blasting a sequence is often the first step in characterizing an unknown protein. Alignments play an integral role in scientific research, and with the advent of high throughput genomic sequencing, its importance in helping to make functional inferences will continue to grow. Alignments can also be used to provide information about evolutionary relationships. Programs such as Clustal use multiple alignments which can then be fed into a program like PAUP* to derive phylogenetic trees.

References

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990)
Basic local alignment search tool.
J. Mol. Biol. 215:403-10.
Altschul, SF, Madden, TL, Schaffer, AA, Zhang, J, Zhang, Z, Miller, W, and DJ Lipman (1997)
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
Nucleic Acids Res. 25(17):3389-402.
Henikoff, S. and Henikoff, J. (1992) Amino acid substitution matrices from protein blocks.
Proc. Natl. Acad. Sci. USA. 89(biochemistry): 10915 - 10919. 1992.
Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443-453.
W. R. Pearson and D. J. Lipman (1988) Improved Tools for Biological Sequence Comparison.
PNAS 85:2444- 2448.
Smith and Waterman (1981) Identification of common molecular subsequences. J. Mol. Biol. 147:195-197.

Timeline



1970 Needleman & Wunsch describe dynamic programming technique to align sequences

1977 Maxam-Gilbert and Sanger each develop new sequencing method to enable automated nucleotide sequencing

1978 Dayhoff Matrix explanation of PAM (percent accepted mutation rate)

1981 Smith-Waterman makes several changes to N&W and describes method for local alignment

1988 Pearson implements heuristic algorithm for alignment FASTA

1990 Altschul debut of BLAST

1992 BLOSUM matrices using BLOCKS evolutionary rates are not homogenous across protein

1995 Venter - H. influenzae shot gun sequenced

1997 Introduction of BLAST2 (Gapped BLAST) and PSI-BLAST

2001 Draft Human Genome announced

Legend

— Algorithms and Implementation

— Matrices and their Derivations

— Experimental Techniques and Breakthroughs

There are two sometimes competing, but eventually complementary, goals in genome wide analysis. On the one hand, the researcher is looking for the big picture. What is the overall organization of the genome? How does it compare with other known genomes? On the flip side, the question is what is the function of this one gene, this one protein? What protein(s) are homologous with other known genomes and what functional inferences can be made based on these homologies? This is a particularly daunting task given that even the smallest genome sequenced mycoplasma genitalium has almost five hundred genes. The biggest challenge of functional genomics is deciding how to balance these two questions. The following proposal addresses both sides.

Once the open reading frames have been identified, there are a number of things that can be done to further characterize the microbial genome as summarized in the flow chart on the following page. Since nucleotide sequence is not as highly conserved as amino acid sequence, one of the first steps would be to translate the nucleotide sequence to the amino acid sequence and use BLASTP or some similar program to search for matches within the database. However, microbial genomes often do not follow standard nucleotide to amino acid tables, so it would be important to use the appropriate amino acid codes to translate from the ORFs which would probably necessitate experimental verification.

I would particularly look for proteins that have multiple matches across a large evolutionary distance. If a protein is highly conserved, this is often a clue that it serves some critical function. For example, the glycolytic enzymes are ubiquitous across a wide range of species. This makes sense as an organism with defective glycolysis would be at a severe competitive disadvantage. In addition to Blast searches for conserved proteins, I would also look for conserved domains. As proteins are often modular in function, using a program such as Prosite to search for motifs can provide additional insight into possible functions. Leucine zippers or other DNA binding domain often characterize transcription factors. In addition to providing functional inferences, motifs and other conserved domains can also help to classify the protein. Proteins in a given family are more likely to share some functional similarity than those that are in different families. I would also be interested in examining the physical properties of the individual proteins. Programs such as ProtParam can predict a wide variety of detailed physical information including molecular weight, theoretical pI, grand average hydrophobicity, aliphatic index, etc. Based on this information deduction as to whether the protein is globular or fibrous, etc can be formed. In addition, I would use a program such as VAST or PredictProtein to make predictions about the secondary structure of the protein.

One of the most interesting lessons coming out of the flurry of new microbial genome sequence papers is how large the number of organism specific genes there are (Fraser, 2000). By comparing this microbial genome sequence specifically to already sequenced microbial genomes, it should be possible to begin to sort through those genes that are specific to this particular microbe and those that are general to most microbes. This is particularly exciting given the intimate interaction that microbes often have with their hosts (Cummings, 2002). Host adaptation maybe one convincing reason for the large amount of organism specific genes that are being founding in microbes (Wren, 2000). One of the big surprises of the *Treponema pallidum* (syphilis genome) is that fact that it obtains almost all the essential building blocks of life from its host including enzyme cofactors, carbohydrates, fatty acids, and even nucleotides. This was a beautiful explanation of why syphilis cannot be cultured to sheer frustration of scientists trying to study it. It also elucidates why five percent of its genes code for transport proteins (Pennisi).

This vast compilation of information would provide an excellent starting point to annotating the genome. Computational approaches can be paired with experimental approaches such as microarray experiments to learn more about the function of a particular genes or proteins. Perhaps, the most important aspect of a functional genomics approach is that it allows researchers to winnow the list of targets to study their system of interest.

References

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-10.

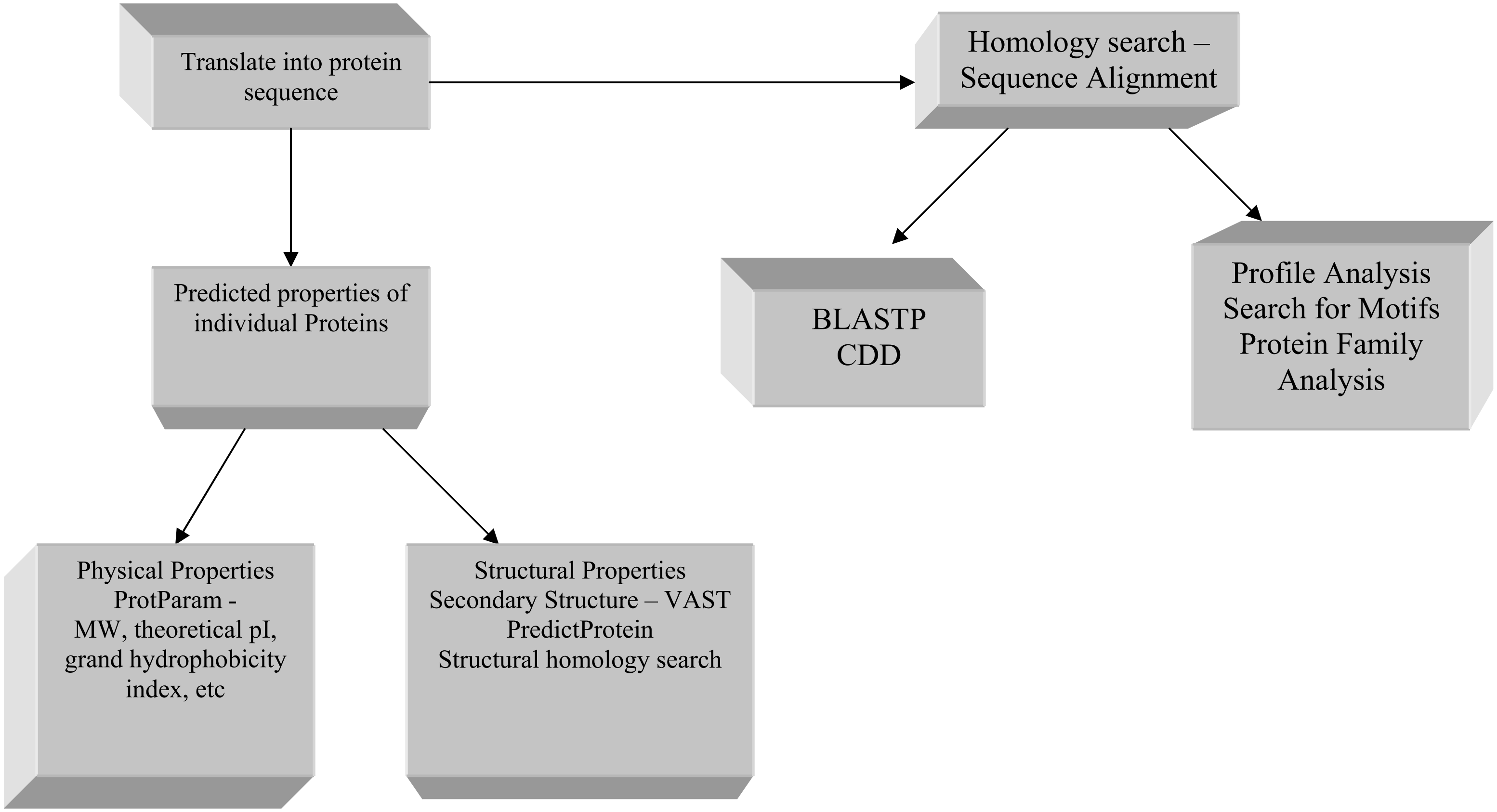
Cummings, Craig A. and Relman, David A. (2002) Microbial Forensics – "Cross Examining Pathogens." Science 296:1976-1979.

Fraser C., Eisen, JA., Eisen & Steven L.S. (2000) Microbial genome sequencing. Nature 406:799-803

Pennisi, E. SyphilisNews TIGR. Genome reveals wiles and weak points of syphilis. www.tigr.org/tdb/tdb.html.

Wren, BW. (2000) Microbial genome analysis: insights into virulence, host adapation and evolution. Nature 1:30-39.

Flowchart



Translate into protein sequence

Homology search – Sequence Alignment

BLASTP
CDD

Profile Analysis
Search for Motifs
Protein Family
Analysis

Predicted properties of individual Proteins

Physical Properties
ProtParam -
MW, theoretical pI,
grand hydrophobicity
index, etc

Structural Properties
Secondary Structure – VAST
PredictProtein
Structural homology search