TOPIC 1.

An examination of the history of sequence alignment techniques reveals that their evolution is a logical progression. The Needleman-Wunsch (NW) method was published in 1970 as the first computer adaptable means by which similarities between amino acid sequences of two proteins could be identified and their significance assessed, i.e. whether the similarities could occur by chance or whether they imply an evolutionary relationship. The central tenet of this method is the maximum match value which results from the pathway through a two-dimensional matrix of all possible pairs that produces the largest sum, given that matches and mismatches are assigned numerical values (e.g. 1 and 0, respectively).[1] This initial approach was rational as the goal was to determine homology between sequences which was plausibly conceived as referring to the sequences as wholes. Ten years later, this global alignment perspective was transformed into the Smith-Waterman (SW) method. It should be noted that during that ten year period, Waterman introduced a metric that allowed for deletions and insertions of arbitrary lengths which correspond to the mutations necessary to produce similarity between sequences. The SW method differs conceptually from the NW method, for it focuses on similarities between *segments* of sequences. Because it assesses total homology based on similarity between parts of sequences while allowing for internal deletions and insertions, the approach is a local alignment technique. Figure 1 below illustrates the fundamental difference between global (NW) and local (SW) alignment methods using Google search techniques as an elementary analogy. The words are identical in both searches, representing a given sequence. However, the use of quotes means that Google must find an exact match in that specific word order which is akin to finding similarities between the entirety of sequences. The other search identifies individual words corresponding to segments of a sequence as used in local alignment.

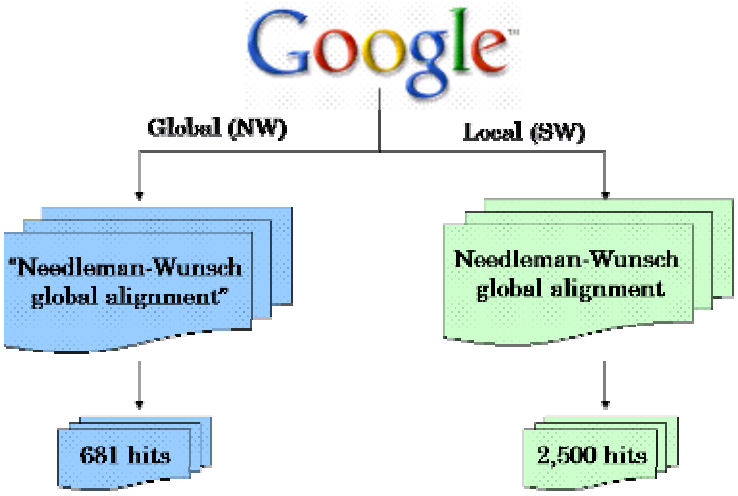

Figure A. The global versus local Google search technique.

[1] This description pertains to the simplest implementation of the NW method. Penalty factors may be used for allowing gaps. Additionally, an n-dimensional matrix may be used to compare n amino acid sequences.

The implementation of NW and SW methods in completely automated search tools and their rapid extension to multiple alignments was a function of the simultaneous advancements in computing technologies during the mid 1980s. In 1985, FASTA emerged utilizing aspects of both local and global alignment with to empower database searching. FASTA identifies regions of identity, uses a scoring matrix, e.g. PAM250, to determine the best segments, and joins those segments allowing for gaps to form a single alignment, the latter step being global alignment.[2] Shortly thereafter, BLAST and BLAST2 surfaced utilizing solely local alignment to drive database queries. The algorithm scores the segment matches with the BLOSUM62 matrix set as the default and uses score statistics to evaluate the similarity. PSI-BLAST creates a position-specific scoring matrix (PSSM) by assigning a score for each position in the alignment from a traditional BLAST search. Iterative BLAST searches are done which result in increased sensitivity.

These modern techniques have expanded the knowledge base of biology. The ability to assess relationships between sequences within an organism or between organisms has enabled biologists to infer secondary structure of proteins more systematically. Extensive information about protein structure will provide valuable insight for the pharmaceutical industry about the discovery and design of effective and specific molecules. For the industry, analysis of insertions/deletions may assist in the development of tailored drugs. Additionally, alternative splicing presents a hurdle to finding genes and these techniques can be applied to detect homology within variants of protein sequences and link them to a given gene – a backwards approach to locating genes.

[2] Pearson WR and Lipman DJ. "Improved Tools for Biological Sequence Comparison." Proceedings of the National Academy of Sciences. 85(8):, 2444-2448.

TOPIC 2.

In order to perform computational functional genomics analyses given a sequence and ORFs, one must first conceptualize this intensive, yet manageable project. A systems biology approach seems appropriate which would require that (1) various subsystems of a total system be defined; (2) elements of the subsystems be perturbed by systematic genetic (digital) and environmental changes; and (3) data from each subsystem be quantified and integrated to compile a model of the entire system.[3] The approach is founded on the intuitive understanding that a whole is the sum of its parts; therefore, one can learn about an entire organism through comprehensive examination of specific aspects of that organism at various levels.

There are various computational issues that must also be addressed for interpreting expression profiles from GeneChips, and SAGE.[4] The data should be interpreted into overall populations of a subcellular compartment (e.g. the mitochondria) instead of individual genes, assigning them into functional categories to reduce noise and to give a more robust and global answer.[5] Internal analysis, particularly a top-down partitioning, should be done in an iterated manner. Top-down is chosen because it does not assume a tree structure which makes it more flexible. Function should be defined to adequately narrow the scope of the task, so it should be measured by biochemical activity, e.g. enzymes. The data should be arranged in a decision tree such as the oversimplified one below.

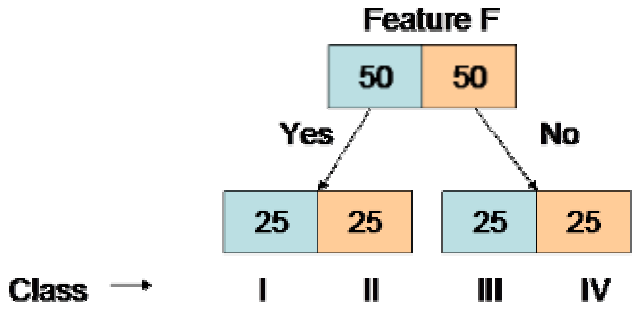

Figure B. Decision tree for a specific biochemical function where class categorizes the protein by subcellular location..

The expression and sequence data can be combined to predict subcellular localization. The use of Bayes' Rule and the probabilities described in the decision tree used in the supervised training set would provide a better and standardized estimate of relationships and, consequently, function. Since proteins may be multifunctional, external mapping to a model organism should be done to find homologies and possibly extrapolate any known functions of the homologies to the unknown microbe's proteins.

---

[3] Hood L and Gala D. "The digital code of DNA." Nature. 23 January 2003. Vol. 421. 444-447.

[4] For the sake of simplicity, cDNA microarrays are not considered because they do not yield "absolute" expression levels. Based on Gerstein M and Jansen R. "The current excitement in bioinformatics, analysis of whole-genome expression data: How does it relate to protein structure and function?" Current Opinion in Structural Biology 2000, 10:574~V584

[5] Greenbaum D et al. "Interrelating Different Types of Genomic Data, from Proteome to Secretome: 'Oming in on Function." Genome Research.

Referring back to the systems biology approach, the information about subcellular localization must be studied in the context of the various cell types that exist within the organism. Those types will likely have different expression profiles depending on the state of the cells influenced by external factors which can be examined by systematic perturbations. Relationships between the numerous 'omes, specifically the proteome, functome and translatome, can be explored. The destination of proteins of unknown localization, such as those proteins with multiple functions, can be determined computationally by using the known relationship between gene expression level and subcellular location.[6]

Thus, the described approach would study the entire organism in terms of its individual parts whose 'omes can be isolated and classified. Expansive laboratory techniques such as the GeneChip and fast computational analyses combined enable the integration of the subsystems to assemble information about the organism as whole.

[6] Ibid