Allison D. Dupuy

Genomics and Bioinformatics MB&B 752a
Final Project
December 12, 2003

Topic 1: *Provide a historical summary of techniques for sequence alignment, from the early methods (NW, SW) through to the present. Discuss two ways in which these techniques, or their modern counterparts, are used in biology.*

Given the exponential increase in the quantity of data available about DNA, RNA, and proteins, computer alignment of molecular sequences has become one of the most powerful techniques in modern biology. One of the earliest methods of sequence alignment was introduced in 1970 by Needleman and Wunsch (NW). The NW algorithm introduced the key concept of dynamic programming and provided the first automatic method of global alignment (1). Ten years later, upon recognition of the fact that certain targeted, biologically significant regions of DNA or protein sequence can be aligned, Smith and Waterman introduced the concept of local alignment (2). In contrast to global alignment, local alignment searches for regions of local similarity and accommodates for the fact that insertions and deletions are attributed to evolutionary change, thereby relinquishing the requirement for analyzing the entire length of a sequence. Dynamic programming measures were expanded with the advent of the heuristic sequence database searching methods of FASTA, developed by Lipman and Pearson in 1985 (3), and BLAST, developed by Altschul et al in 1990 (4). Both of these methods use dynamic programming after performing specific preliminary calculations. While both of these methods consider alignments of certain fixed lengths (3,4), the main difference between the two is that FASTA considers exact matches (3) whereas BLAST uses a scoring function to measure similarity (4). In addition to basic BLAST, new, more sophisticated methods have emerged including BLAST2 (gapped BLAST), PSI-BLAST (Position-Specific Iterated BLAST), and PHI-BLAST (Pattern-Hit Initiated BLAST) (5). Depending on the task at hand, each of these methods provide the user with more targeted and detailed search options. In addition to global alignment and local alignment, the third and arguably the most widely used technique involves multiple sequence alignment (MSA), the first attempts of which were made in the mid-1980s. While a wide variety of different MSA algorithms exist, the most common methods include CLUSTALW and PILEUP, both of which are based on a progressive alignment algorithm proposed by Feng and Doolittle in 1987 (6). A second type of MSA technique involves the use of position-specific information for residues, insertions, and deletions at each position in the alignment. This includes profile, motif, and pattern analysis which surfaced in the late 1980s (5). Improvements were made in the early 1990s with the introduction of Hidden Markov models which describe a probability distribution to explain an infinite number of possible sequences (5).

The various sequence alignment techniques discussed above have a wide range of applications in modern biological research given their ability to discern functional, structural and evolutionary information. One simple application of multiple sequence alignment involves the identification of point mutations in specifically queried genes. This is particularly useful in human genomic analyses where, for example, a point mutation in a specific gene may be known to predispose certain individuals to a particular disease. Upon sampling the target gene from a statistically significant group of individuals in a disease-prevalent family, the sequences could be aligned and analyzed in order to identify the point mutation. A second and widely utilized application involves secondary structure prediction. Multiple sequence alignment methods can be used to compare a query protein sequence to sequences of known structure in order to deduce potential structure for specific regions of the protein of interest. Examples include transmembrane helix prediction algorithms such as TMHMM (7), coiled-coil prediction algorithms such as Paircoil (8), and intracellular targeting signal prediction algorithms such as PSORT (9).

Allison D. Dupuy
Genomics and Bioinformatics MB&B 752a
Final Project
December 12, 2003

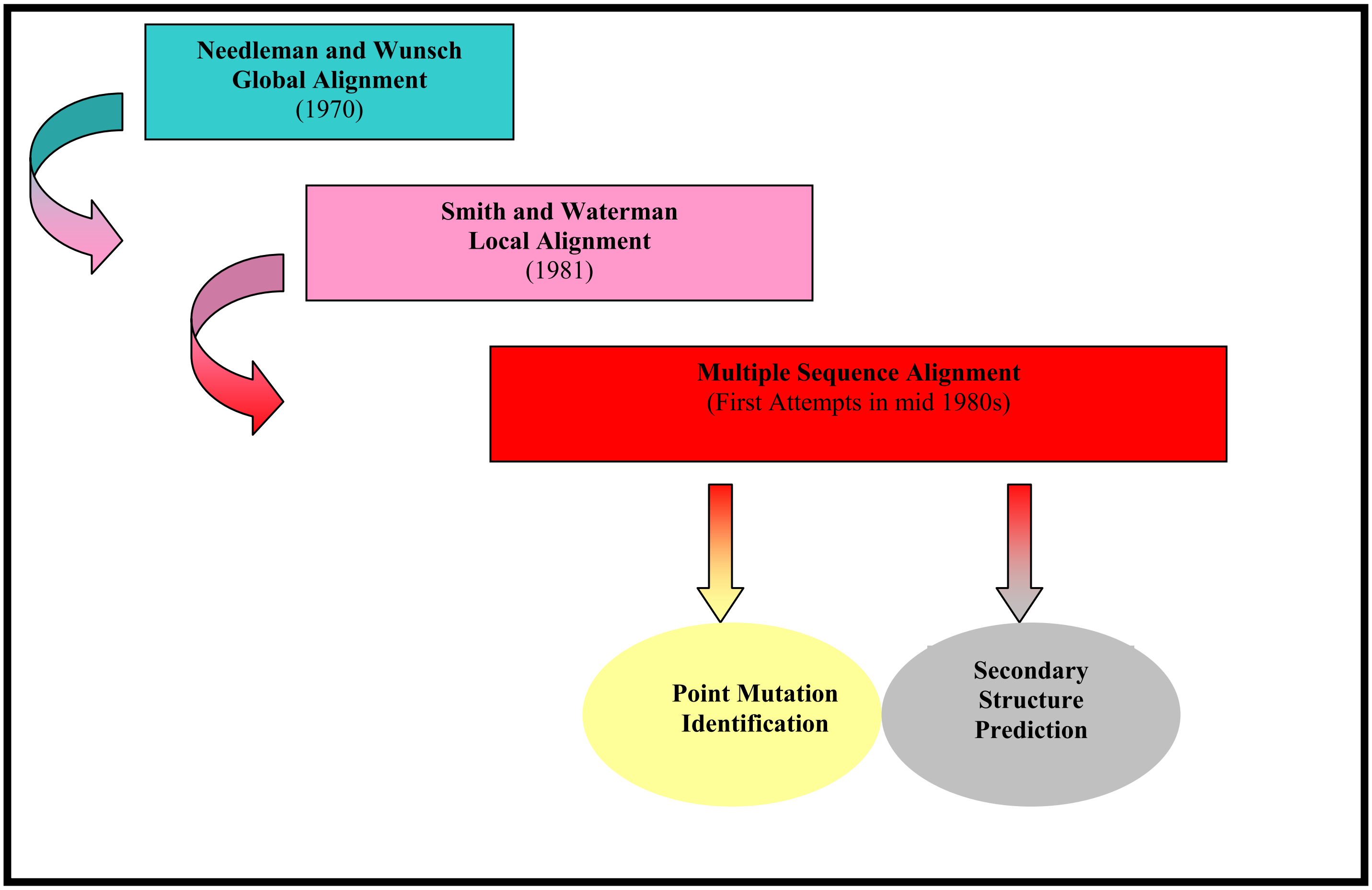

Figure 1: Development of Sequence Alignment Techniques. The three main branchings of sequence analysis techniques include the global alignment method of Needleman and Wunsch (NW), the local alignment of Smith and Waterman (SW) and multiple sequence alignment (MSA) methods. The time progression of the appearance of these techniques in biological research is such that NW appeared first and MSA is the most recent. The length of the bars encompassing each of the three main branches reflect the frequency of usage today. Two biologically relevant applications of MSA include point mutation identification and secondary structure prediction.

References:

(1) 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.
(2) Smith, T.F., Waterman, M.S., and Fitch, W.M. (1981) *Comparative biosequence metrics*. *J. Mol. Evol.* 18: 38–46.
(3) Pearson, W.R., and Lipman, D.L. (1988) *Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci.USA* 85: 2444–2448.
(4) Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *Basic local alignment search tool. J. Mol. Bio*., 215:403-410.
(5) Gerstein, M. (2002) *Bioinformatics Sequences*. <http://www.bioinfo.mbb.yale.edu>.

(6) Feng, D. and Doolittle, R. (1987) *Progressive sequence alignment as a prerequisite to correct polyogenetic trees. J. Mol. Evol.* 25:351-360.
(7) Krogh, A. et al. (2001) *TMHMM Server 2.0*. <http://www.cbs.dtu.dk/services/TMHMM-2.0/> (Last updated October 29, 2003)
(8) Berger, B. et al. (1995) *Paircoil Scoring Form*. <http://paircoil.lcs.mit.edu/cgi-bin/paircoil>.
(9) Nakai, K. *PSORT WWW Server*. <http://psort.nibb.ac.jp/> (Last updated April 16, 2003).

Allison D. Dupuy

Genomics and Bioinformatics MB&B 752a
Final Project
December 12, 2003

Topic 2: *Assume a new microbial genome (i.e. not yeast or E. coli) has been sequenced, and the open reading frames (ORFs) identified. You hope to perform computational functional genomics analyses on this new organism and must prepare a proposal for what you plan to do. Using examples of current techniques, discuss how and why you would approach each step of the wide-open task of learning more about this new organism, given known sequence and ORFs. These techniques could include mapping information from other organisms.*

A new microbe "*Newmoni microbium*" has been sequenced and the ORFs identified. As one of the fresh faces in the top Bioinformatics group in the country, I was recently assigned the wide-open task of learning more information about this new organism. With the energy and vigor of an eager second-year graduate student I accepted the task with open arms and devised a comprehensive plan to perform computational functional genomics analyses on this new organism, the specifics of which are detailed below.

As a first approach, I plan to extract as much information as possible via sequence analysis and mapping information from other sequenced organisms. This is an important step given that computational analysis of the genomic DNA will aid in the prediction of gene structure, gene clustering predictions will provide clues to function, and multiple genome alignment will aid in discerning valuable information concerning conserved regions of nucleotide/amino acid sequence, in addition to the evolutionary history of the microbe. The extensive sequence analysis will include sequence alignment, homology detection, single nucleotide polymorphism detection, genetic element discovery (including protein-coding and RNA genes, pseudogenes, promoters, enhancers, and repeats), secondary structure prediction (signal peptide predictions, transmembrane helix predictions, coiled-coil and helix-turn-helix motif predictions based on previously annotated homologs), and phylogenetic analysis (1). By making inferences about the function of various gene products in the new microbe based on the homology analyses, we will then attempt to reconstruct the metabolic pathways that reflect the organism's physiological makeup. It should be noted that, in general, roughly 30-40% of genes identified in whole genome sequencing projects encode proteins of unknown function (2). These unknown genes will serve as a rich target for additional biological experiments to determine the physiological and biochemical functions of *Newmoni microbium*.

The next step will involve systematic gene disruption and gene overexpression analyses of *Newmoni microbium* in order to functionally characterize the new microbe (3). These gene disruption and overexpression studies will allow us to analyze the phenotype of the organism under various conditions (e.g. changes in temperature, pH). One major assumption in this undertaking is that the new microbe is conducive to common molecular genetic manipulations. Incidentally, we will attempt to use such conventional procedures as single-step gene replacement strategies utilizing antibiotic resistance genes in order to construct deletion mutants.

The third component of our study will involve the analysis of gene (predicted) expression in *Newmoni microbium* using two-dimensional gel electrophoresis (3). By comparing the protein expression patterns under various physiological and environmental conditions we will be able to gain greater insight into the effects of these changes on the microbe. In addition, by combining two-dimensional gel electrophoresis with trypsin digestion and mass spectrometry, we will attempt to identify all of the microbe's proteins in addition to analyzing protein modifications under the varied environmental conditions.

The fourth component of our analysis will involve *in-situ* localization of predicted proteins in *Newmoni microbium* based on our sequence analysis using fluorescent tags (e.g. GFP-fusions) (4). This will allow for a systematic method of localizing all of the proteins in the microbe. These localization studies can be analyzed based on normal living conditions, changes in the environment and their effect of protein localization, in addition to comparison of the localization of similar proteins in both closely and distantly related microbes.

As a fifth and final mode of investigation, we will conduct genome-wide analysis of the expression profiles of *Newmoni microbium* using the most innovative microarray and chip technologies available (5). These experiments will serve as a second, more sophisticated method of obtaining valuable information regarding the proteome of *Newmoni microbium*, from which we hope to infer considerable functional information.

Allison D. Dupuy

Genomics and Bioinformatics MB&B 752a
Final Project
December 12, 2003

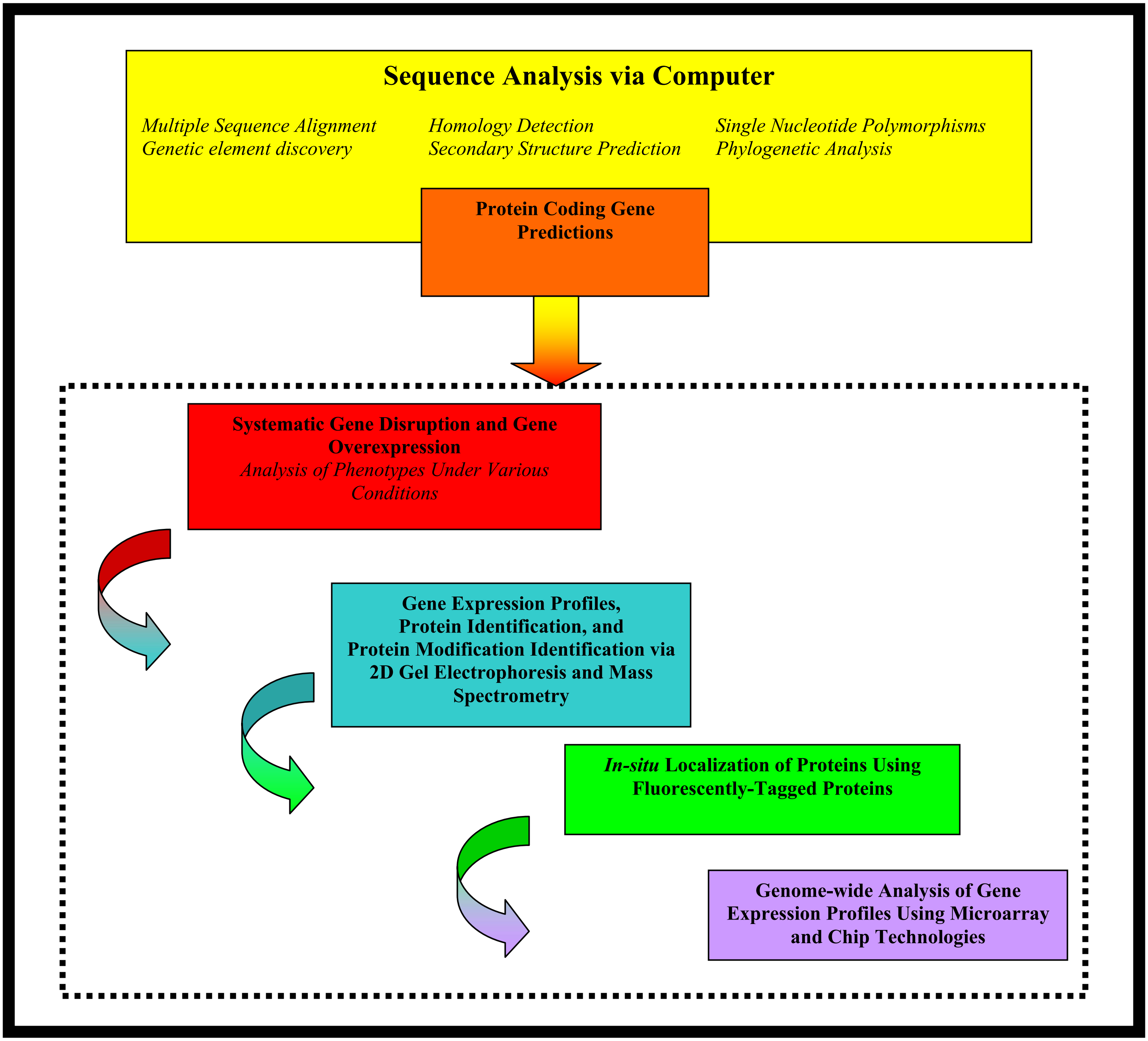
Sequence Analysis via Computer
Multiple Sequence Alignment
Genetic element discovery
Homology Detection
Secondary Structure Prediction
Single Nucleotide Polymorphisms
Phylogenetic Analysis
Protein Coding Gene Predictions
Systematic Gene Disruption and Gene Overexpression
Analysis of Phenotypes Under Various Conditions
Gene Expression Profiles, Protein Identification, and Protein Modification Identification via 2D Gel Electrophoresis and Mass Spectrometry
In-situ Localization of Proteins Using Fluorescently-Tagged Proteins
Genome-wide Analysis of Gene Expression Profiles Using Microarray and Chip Technologies

Figure 1: Computational Functional Genomics Analysis Proposal for *Newmoni microbium*. The project proposal for the computational functional genomics analysis of the microbe *Newmoni microbium* is composed of five main groupings. The arrows indicate the sequential flow of the analysis that will be undertaken.

References:

(1) Gerstein, M. (2002) *Bioinformatics Sequences*. <http://www.bioinfo.mbb.yale.edu>.
(2) Tatusov, R.L. *et al.* (1997) *A genomic perspective on protein families*. *Science* 278: 631-637.
(3) Gerstein, M. (2002) *Bioinformatics Introduction*. <http://www.bioinfo.mbb.yale.edu>.
(4) Gerstein, M. (2002) *Bioinformatics Datamining*. <http://www.bioinfo.mbb.yale.edu>.
(5) Gerstein, M. (2002) *Bioinformatics Microarrays*. <http://www.bioinfo.mbb.yale.edu>.