

# Understanding Bioinformatics

by Marketa Zvelebil and Jeremy Baum

Last updated: May 1, 2009

## Textbook Reading Guidelines

### Preface:

Read the whole preface, and especially:

- For the students with Life Science background:
  - Page v, third paragraph: “to perform a proper analysis ... which it is based”.
- For the students with Information Technology background:
  - Page vi, first paragraph: “many postgraduate students ... biomedical science”.

### Part One: Background Basics

Chapters One, Two and Three provide introductory key knowledge that will be assumed throughout the remainder of the book.

#### Chapter One: The Nucleic Acid World

This chapter is an excellent review of DNA, RNA, proteins and the central dogma of molecular biology for biology majors. Most of these concepts are new to CS majors unless they have taken a molecular biology course in the recent past. The chapter covers the basics of gene structure, and gene expression and gives a brief introduction to molecular evolution. We shall cover most of the important concepts of this chapter in class. Please make sure that you know all the terms from “Biology Terms” that are found in this chapter. Definitely read the whole chapter.

#### Chapter Two: Protein Structure

Read the following sections of the chapter:

- Introduction: pages 25-27.  
The right side of Mind Map 2.1 is more important than the left side.
- 2.1 Primary and Secondary Structure: pages 26-35.
  - Introduction: page 26
  - Protein structure can be considered on several different levels: pages 26-27
  - Amino acids are the building block of proteins: pages 27-28.  
In bioinformatics, the one-letter code is usually used. Just know which letters represent amino acids [Example: [A..Y] – [BJOUX]].
  - The differing chemical and physical properties of amino acids are due to their side chains: pages 28-29.  
Important points:
    - Proteins can be classified into overlapping groups that share common physical and chemical properties.

- The difference between **hydrophobic** and **hydrophilic**.
- Only a few sequence modifications are needed to destabilize the 3-D conformation of a protein.
- Amino acids are covalently linked together in the protein chain by peptide bonds: pages 29-33.  
Just know that protein sequences are written from the N terminus to the C terminus, from left to right.
- Secondary structure of proteins is made up of  **$\alpha$ -helices** and  **$\beta$ -strands**: pages 33-35.  
Most proteins contain one or more stretches of amino acids that take on a characteristic structure in 3-D space. The most common of these are the alpha helix and the beta sheet conformations.
- 2.2 Implications for Bioinformatics: pages 37-39.
  - Introduction: page 37.
  - Evolution has aided sequence analysis: page 38.  
A very important paragraph. Make sure to understand “**homologous proteins**”.
  - Visualization and computer manipulation of protein structures: pages 38-39.  
Just familiarize yourself with the 6 representations of Figure 2.13 on page 39.
- 2.3. Proteins Fold to Form Compact Structures: pages 40-43.
  - Introduction: pages 40-41.  
Clearly understand the different biological functions of proteins; for example as **enzymes** (as depicted in Figure 2.14) and what is meant by protein **domain**.
  - The tertiary structure of a protein is defined by the path of polypeptide chain: page 41.  
An important section that explains the role of protein domain, protein folding, and structure.
  - Many proteins are formed of multiple subunits: pages 42-43.  
Make sure you go over oligomeric proteins, monomers (subunits), and quaternary structure.
- Summary: page 43.  
The first paragraph is a summary of protein structure and function. The second paragraph is about one of the most important goals of bioinformatics: to predict and analyze the structure of proteins and the relationship of the structure to the function. This leads to performing sequence alignments which is the main topic of Chapters 4 to 8.

## **Chapter Three: Dealing with Databases**

Read the following sections of the chapter:

- Introduction: pages 45-46.
- 3.1 The Structure of Databases: pages 46-52,
  - Introduction: pages 46-48.  
Skim through the section paying more attention to Flow Diagram 3.1 and Figure 3.1.
  - Flat-file databases store data as text files: pages 48-49.
  - Relational databases are widely used for storing biological information: pages 49-50.  
Skim through the section paying more attention to the last paragraph on page 49.
  - XML has the flexibility to define bespoke data classifications: pages 50-51.  
Skim through the section paying more attention to the first paragraph.
  - Many other database structures are used for biological data: pages 51-52.  
Skim through the section.
  - Databases can be accessed locally or online and often link to each other: page 52.  
Skim through the section.
- 3.2 Types of Databases: pages 52-55.
  - Introduction: pages 52-53.  
Make sure to go over Flow Diagram 3.2 and that you understand what is meant by **annotation** on page 53.
  - There's more to databases than just data: page 53.  
Skim through the section.
  - Primary and derived data: pages 53-54.  
Make sure you understand the difference between primary and derived data.
  - How we define and connect things is important: Ontologies: pages 54-55.  
Make sure to go over the definition of **Gene Ontology**.
- 3.3 Looking for Databases: pages 55-61.
  - Introduction: page 55.
  - Sequence databases: pages 55-58.  
Read the section paying more attention to:
    - The 3 types of DNA sequences on page 56: raw data, **cDNA**, and **EST**.
    - Figure 3.7. Understand the different components of the GenBank DNA sequence file. Know what is meant by each type.
  - Microarray databases: page 58.  
Skim through the section.
  - Protein interaction databases: pages 58-59.  
Skim through the section.
  - Structural databases: pages 59-61.  
Skim through the section paying attention to Figure 3.8.

- 3.4 Data Quality: pages 61-66.
  - Introduction: pages 61-62.  
Go over Flow Diagram 3.4.  
Make sure you realize that there are 2 methods for checking the accuracy of data analysis: computer-based analysis and manual curating.
  - Nonredundancy is especially important for some applications of sequence databases: pages 62-63.  
Make sure you understand what is meant by a **nonredundant database**.
  - Automated methods can be used to check for data consistency: pages 63-64.  
Skim through the section.
  - Initial analysis and annotation is usually automated: pages 64-65.  
Skim through the section paying more attention to the last paragraph.
  - Human intervention is often required to produce the highest quality annotation: page 65.  
Read the section making sure you understand the importance of manual annotation.
  - The importance of updating databases and entry identifier and version numbers: pages 65-66.  
Read the section making sure to understand why it is important to regularly update and correct existing data in databases, and not just include new entries.
  
- Summary: page 66.  
Make sure you go over all 4 paragraphs.

----- **End of Part One** -----

## **Part Two: Sequence Alignments**

Chapters Three, Four and Five deal with a variety of analyses of sequences, all relating to identifying similarities. Chapter 4 is a practical introduction to sequence alignments with examples on different analyses and demonstrations of some potential problems as well as successful results.

### **Chapter Four: Producing and Analyzing Sequence Alignments**

Read the following sections of the chapter:

- Introduction: pages 71-72.  
A very good introduction on the importance of the identification of similar sequences.  
Make sure you thoroughly understand all 4 paragraphs.
  
- 4.1 Principles of Sequence Alignment: pages 72-76.
  - Introduction: pages 72-73.
  - Alignment is the task of locating regions of two or more sequences to maximize their similarity: pages 73-74.  
Make sure to go over the 3 alignments they have on both pages.
  - Go over Box 4.1 Genes and pseudogenes: page 73.
  - Alignment can reveal homology between sequences: pages 74-75.

- Go over Box 4.2 Convergent and divergent evolution: page 75.
- It is easier to detect homology when comparing protein sequences than when comparing nucleic acid sequences: pages 75-76.
- 4.2 Scoring Alignments: pages 76-81.
  - The quality of an alignment is measured by giving it a quantitative score: page 76.
  - The simplest way of quantifying similarity between two sequences is percentage identity: pages 76-77.  
Make sure you thoroughly understand Figure 4.1.
  - The dot-plot gives a visual assessment of similarity based on identity: pages 77-79.  
Make sure you thoroughly understand Figure 4.2 (and how to go from Figure 4.2.A to Figure 4.2.B) and also Figure 4.3.  
Skim through Box 4.3 and Box 4.4.
  - Genuine matches do not have to be identical: pages 79-81.  
This is a very important section. Make sure you understand it.
  - There is a minimum percentage identity that can be accepted as significant: page 81.
  - There are many different ways of scoring an alignment: page 81.
- 4.3 Substitution Matrices: pages 81-85.
  - Substitution matrices are used to align individual scores to aligned sequence positions: pages 81-82.  
Make sure you go over the alignment of page 82 and you understand how they got the score of 52.
  - The PAM substitution matrices use substitution frequencies derived from sets of closely related protein sequences: pages 82-84.  
Make sure to go over Figure 4.4 and to understand the significance of the substitutions represented by different colors in the matrices. All the substitution matrices we will use in the course have the same flavor – are similar to the ones presented in Figure 4.4.
  - The BLOSUM substitution matrices use mutation data from highly conserved local regions of sequence: page 84.
  - The choice of substitution matrix depends on the problem to be solved: pages 84-85.  
Make sure you thoroughly understand this section.
- 4.4 Inserting Gaps: pages 85-87.
  - Gaps inserted in a sequence to maximize similarity with another require a scoring a scoring penalty: pages 84-86.  
Read the whole section making sure to understand the difference between the two alignments of Figure 4.5.  
Skim through Box 4.5.
  - Dynamic programming algorithms can determine the optimal introduction of gaps: pages 86-87.

- 4.5 Types of Alignment: pages 87-93.
  - Different kinds of alignments are useful in different circumstances: pages 87-89.  
Make sure you understand the fundamental difference between global and local alignments.  
Thoroughly understand Figure 4.7. Notice how the global alignment fails to align the functionally important residues.
  - Multiple sequence alignments enable the simultaneous comparison of a set of similar sequences: page 90.
  - Multiple alignments can be constructed by several different techniques: pages 90-91.
  - Multiple alignments can improve the accuracy of alignment for sequences of low similarity: pages 91-92.  
Make sure you thoroughly understand Figure 4.10. Convince yourself that sometimes a multiple sequence alignment is preferred over pairwise alignment.
  - ClustalW can make global multiple alignments of both DNA and protein sequences: page 92.
  - Multiple alignments can be made by combining a series of local alignments: pages 92-93.
  - Alignment can be improved by incorporating additional information: page 93.
  
- 4.6 Searching Databases: pages 93-97.
  - Introduction: pages 93-94.
  - Fast yet accurate search algorithms have been developed: pages 94-95.
  - FASTA is a fast database-search method based on matching short identical segments: page 95.  
Skim through the section.
  - BLAST is based on finding very similar short segments: page 95.
  - Different versions of BLAST and FASTA are used for different problems: pages 95-96.  
Concentrate on BLAST. Read the right side of Table 4.1.
  - PSI-BLAST enables profile-based database searches: pages 96-97.
  - SSEARCH is a rigorous alignment method: page 97.
  
- 4.7 Searching with Nucleic Acid or Protein Sequences: pages 97-103.
  - DNA or RNA sequences can be used either directly or after translation: page 97.
  - The quality of a database match has to be tested to ensure that it could not have arisen by chance: pages 97-98.
  - Choosing an appropriate E-value threshold helps to limit a database search: pages 98-100.
  - Low-complexity regions can complicate homology searches: pages 100-102.
  - Skim through Box 4.6.
  - Different databases can be used to solve particular problems: pages 102-103.

- 4.8 Protein Sequence Motifs or Patterns: pages 103-107.
  - Introduction: pages 103-104.  
Make sure you understand the definition of a motif and realize that the term is mainly used with protein sequences.
  - Creation of pattern databases requires expert knowledge: pages 104-105.
  - The BLOCKS database contains automatically compiled short blocks of conserved multiply aligned protein sequences: pages 105-107.  
Make sure you understand how BLOCKS was created.
  
- 4.9 Searching Using Motifs and Patterns: pages 107-109.
  - The PROSITE database can be searched for protein motifs and patterns: pages 107-108.  
Skim through the section.
  - The pattern-based program PHI-BLAST searches for both homology and matching motifs: page 108.  
Skim through the section.
  - Patterns can be generated from multiple sequences using PRATT: page 108.  
Skim through the section.
  - The PRINTS database consists of fingerprints representing sets of conserved motifs that describe a protein family: page 109.  
Skim through the section.
  - The Pfam database defines profiles of protein families: page 109.  
Skim through the section.
  
- 4.10 Patterns and Protein Function: pages 109-111.
  - Searches can be made for particular functional sites in proteins: pages 109-110.  
Skim through the section.
  - Sequence comparison is not the only way of analyzing protein sequences: pages 110-111.  
Skim through the section.
  - Skim through Box 4.7.
  
- Summary: pages 111-112.  
Carefully read all 5 paragraphs.

----- **End of Part Two** -----

## **Part Three: Evolutionary Process**

Chapters 7 and 8 deal with phylogenetic tree construction. Chapter 7 is about the basic ideas involved in reconstructing the evolutionary history of gene and protein sequences and showing how the methods can be applied to various scientific problems. We will only cover Chapter 7.

### **Chapter Seven: Recovering Evolutionary History**

Read the following sections of the chapter:

- Introduction: pages 223-225.

A good introduction on the importance of phylogenetic trees. Read it very carefully and try to understand every paragraph.

- 7.1 The Structure and Interpretation of the Phylogenetic Trees: 225-235.
  - Introduction: page 225.
  - Phylogenetic trees reconstruct evolutionary relationships: pages 225-230.  
Make sure to read the whole section. It is relatively long but gives a lot of information on very important concepts that will be seen later.  
Make sure to go over Figure 7.1 and Figure 7.2 and understand the different tree topologies.  
Skim through Box 7.1 on page 228.
  - Tree topology can be described in several ways: pages 230-232.  
Go over Figure 7.4 and skim through the whole section.
  - Consensus and condensed trees report the results of comparing tree topologies: pages 232-235.  
Go over the first two paragraphs and Figure 7.5. Skip the rest of the section.  
If you want to read more about bootstrapping, please go over Box 8.4 on page 310 and skip the last paragraph.
- 7.2 Molecular Evolution and its Consequences: pages 235-248.
  - Introduction: pages 235-236.  
Go over this short section.
  - Most related sequences have many positions that have mutated several times: page 236.  
The first paragraph is important since it is about assumptions that are made by tree construction packages: constant rate of mutation and single mutation per site (and not overlapping mutations). Skim through the rest of the section.
  - The rate of accepted mutation is usually not the same for all types of base substitution: pages 236-238.  
Make sure to understand the first half of the first paragraph.  
Make sure to go over the part that discusses transitions and transversions in the second paragraph.  
Go over Figure 7.8(A) on page 238.  
Skim through the rest of the section.
  - Different codon positions have different mutation rates: pages 238-239.  
Skim through the section including Box 7.2.
  - Only orthologous genes should be used to construct species phylogenetic

trees: pages 239-247.

Read the first three paragraphs very carefully.

Make sure you understand Figure 7.10 on page 242.

Skim through the rest of the section including Box 7.3, except for the parts on homoplasy, and horizontal gene transfer. Make sure to go over Figure 7.15 on page 246.

- Major changes affecting large regions of the genome are surprisingly common: pages 247-248.  
Carefully read the section.
- 7.3 Phylogenetic Tree Reconstruction: pages 248-263.
  - Introduction: pages 248.
  - Small ribosomal rRNA sequences are well suited to reconstructing the evolution of species: page 249.  
Skim through the section.
  - The choice of the method for tree reconstruction depends to some extent on the size and quality of the dataset: pages 249-251.  
Read the section and if you want to know more about UPGMA, please go over Section 8.2 on pages 278 and 279.
  - A model of evolution must be chosen to use with the method: pages 251-255.  
Skim through the section.
  - All phylogenetic analyses must start with an accurate multiple alignment: page 255.  
Skim through the section.
  - Phylogenetic analyses of a small dataset of 16S RNA sequence data: pages 255-259.  
Skim through the section.
  - Building a gene tree for a family of enzymes can help to identify how enzymatic functions evolved: pages 259-263.  
Skim through the section.
- Summary: page 264.  
Carefully read all three paragraphs of the short summary.

----- **End of Part Three** -----