Understanding Bioinformatics

by Marketa Zvelebil and Jeremy Baum

Last updated: January 16, 2013

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 25-26.
  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:
    - Amino acids 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 **α-helices** and **β-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 in 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 oligometric 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.

Part Five: Secondary Structures

Chapters 11 and 12 are about methods for predicting the secondary structures of biological molecules. Chapter 11 introduces available methods including implementation issues and how to interpret the result obtained by the methods. It also contains sections on "specialized predictions", such as the prediction of secondary structures of RNA and transmembrane proteins. Chapter 12 gives in-depth descriptions of methods used for secondary structure predictions. It also gives the underlying principles of some methods including neural networks and hidden Markov model techniques.

Chapter Eleven: Obtaining Secondary Structure from Sequence

Read the following sections of the chapter:

- Introduction: pages 411-413.

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

- 11.9 RNA Secondary Structure Prediction: pages 455 – 458.

--- --- --- --- --- --- --- --- --- --- End of Part Five --- --- --- --- --- --- --- ---

Part Two: Sequence Alignments

Chapters Four, Five, and Six 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. Chapters 5 and 6 focus on methods used for pairwise and multiple sequence alignment, pattern searching, and database searching. We covered Chapter 4 in Biology/CS 123A. We are now going to cover Section 2 of Chapter 6. The section is mainly on Hidden Markov Models (HMMs) and more specifically, on profile HMMs that define sequence profiles from protein families.

- 6.2 Profile Hidden Markov Models: pages 179-193.
  - Introduction: pages 179-180.
    - Go over Flow Diagram 6.2 (the yellow part on the right and bottom parts of the flowchart).
    - Make sure you fully understand all four paragraphs and do not at all worry about the comparisons made to PSI-BLAST and to PSSM parameters.
  - The basic structure of HMMs used in sequence alignment to profiles: pages 180-185.
    - This is the most important subsection of Section 6.2. Make sure you fully understand everything up to and including the first paragraph of page 183 (until …to the overall amino acid composition). Skim through the rest of the subsection, including Figure 6.8, Figure 6.9, and Figure 6.10.
    - Make sure you fully understand Figure 6.6 and Figure 6.7.
  - Estimating HMM parameters using aligned sequences: pages 185-187.
    - This is a very important subsection. Understand every single paragraph. Do not be intimidated by the notation. If you prefer, stick to the notation I have in my lecture notes.

  - Scoring a sequence against a profile HMM: The most probable path and the sum over all paths: pages 187-191.
    - Make sure you fully understand every single sentence of the first two paragraphs from the bottom of page 187 to the top of page 188.
    - Continue reading the next 2 paragraphs on page 188 without paying too much attention to the analogies made to the dynamic programming method for pairwise alignment of Section 5.2.
    - Skim through the rest of the section, unless of course, you are a computer science major who would like to implement a hidden Markov Model. Then the rest of the section is very important.
    - Also note that the rest of the chapter is equivalent to what I have in my lecture notes. These are the first two points of Rabiner: Evaluation and Decoding.
  - Estimating HMM parameters using unaligned sequences: pages 191-193.
    - Here too, skim through the section unless you are a computer science major who would like to implement a Hidden Markov Model.
    - Note that this section is mainly about the Baum-Welch algorithm which is found in my lecture notes under Rabiner's third point of Learning.

Part Four: Genome Characteristics

Genome Characteristics deals with the analysis required to interpret raw genome sequence data. The genome encodes all the machinery and information for any living organism, from the most simple to the most complex. To understand how an organism works, it is crucial to know how each part of the genome works. Chapter 9 is about techniques used to locate, predict, and annotate genes in both, prokaryotes as well as eukaryotes. Chapter 10 mainly describes various techniques for gene prediction and genome annotation.

Chapter Nine: Revealing Genome Features

Read the following sections of the chapter:

- Introduction: pages 317-318.
  Read the three paragraphs and note that the chapter is limited to identifying genes and their control regions and to obtaining sequences of related proteins, as mentioned in the beginning of the second paragraph.

- 9.1 Preliminary Examination of Genome Sequence: pages 318 to 322.
  - Introduction: pages 318-319.
    The most common method for gene prediction is to locate open reading frames (ORFs). This is easy for prokaryotes, challenging for eukaryotes due to introns and alternative splicing.
  - Whole genome sequences can be split up to simplify gene searches: page 319. Understand Flow Diagram 9.1 and the reason behind splitting genomes.
  - Structural RNA genes and repeat sequences can be excluded from futher analysis: pages 319-321. Understand Figure 9.1, parts A and B.

- Box 9.1. Expressed sequence tags: page 321.
  Read the box as ESTs are very important. An additional source for reading and learning about ESTs can be found at:
  http://www.ncbi.nlm.nih.gov/About/primer/est.html
- Homology can be used to identify genes in both prokaryotic and eukaryotic genomes: page 322.
  - Go over Figure 9.2.
  - Once again, the importance of ESTs is highlighted here.

- 9.2 Gene Prediction in Prokaryotic Genomes: pages 322-323.
  - Introduction: pages 322-323.
    Make sure you fully understand Figure 9.3.

- 9.3 Gene Prediction in Eukaryotic Genomes: pages 323-337.
  - Introduction: pages 323-324.
    Go over Flow Diagram 9.3 and Figure 9.4.
  - Programs for predicting exons and introns use a variety of approaches: pages 324-325.
    Note that predicting the first and last exons are generally more challenging than the detecting the internal exons.
  - Gene predictions must preserve the correct reading frame: pages 325-327.
    - Skip Table 9.2.
    - Go over Figure 9.5 and Table 9.4 and understand every single detail represented in Figure 9.5.
    - Skim over Figure 9.6 and Figure 9.7.
  - Box 9.2. Two different gene sequences for prediction: pages 326-327.
    Note that ALDH10 was not included in the training sets of most gene predictors. See last sentence.
  - Some programs search for exons using only the query sequence and a model for exons: pages 327-332.
    - Make sure to understand what is meant by "MZEF employs statistical analysis of sequences patterns that enables exons to be discriminated from the rest of the DNA sequence".
    - What are some of these sequence patterns?
    - Make sure you understand all the different "Scores" of Table 9.3.
    - Skim through Table 9.4 and Table 9.5.
    - Make sure you thoroughly understand Figure 9.8.
  - Box 9.3. Arabidopsis thalania: a model plant: page 330.
  - Some programs search for genes using only the query sequence and a gene model: pages 332-334.
    - Here too, skim through Table 9.4 and Table 9.5.
    - Here too, go over Figure 9.8.
    - Skim through Figure 9.9.
  - Genes can be predicted using a gene model and sequence similarity: pages 334-336.

        - Make sure you understand very well the first paragraph.
        - The rest of the section briefly introduces a few gene predictors: GenScan, GrailEXP, AAT, GeneBuilder, GeneWalker. Just skim through the paragraphs and carefully read the last paragraph that shows that we are far from having very reliable gene predictors.
    - Genomes of related organisms can be used to improve gene prediction: pages 336-337.
      Read the whole section, but especially the last paragraph that explains how to treat the packages and examples illustrated throughout the textbook.
    - Box 9.4. FGENESH and the rice genome: page 335.
    - Box 9.5. Transposons and repeated elements: page 337.

- 9.4 Splice Site Detection: pages 337-338.
    - Introduction: pages 337-338.
      Skim through this section.
    - Splice sites can be detected independently by specialized programs: page 338.
      Skim through this section.

- 9.5 Prediction of Promoter Regions: pages 338-342.
    - Introduction: pages 338-339.
    - Prokaryotic promoter regions contain relatively well-defined motifs: pages 339-340.
      Concentrate on the 4 conserved patterns that can be used to find promoter regions in most prokaryotes.
    - Eukaryotic promoter regions are typically more complex than prokaryotic promoters: page 340.
      A very important small section. Make sure to understand all of it.
    - A variety of promoter-prediction methods are available online: pages 340-341.
      The section briefly introduces a few promoter predictors: FunSiteP, NNNP, TSSG, TSSW, CorePromoter, Promoter 2.0, ProScan, PromoterInspector. Just skim through the section.
    - Promoter prediction results are not very clear-cut: pages 341-342.
      Skim through the section. The bottom line: never trust one package.

- 9.6 Confirming Predictions: pages 342-346.
    - Introduction: page 342.
      Skim through it.
    - There are various methods for calculating the accuracy of gene-prediction programs: pages 342-343.
      These are criteria used to compare gene-predictors. Read the whole section and understand Figure 9.12.
    - Translating predicted exons can confirm the correctness of the prediction: page 343.
      By now the first sentence should be clear. Go over and understand Figure 9.13 where blastx was used.
    - Constructing the protein and identifying homologs: pages 343-346.

Read the section paying special attention to Figure 9.15 and Table 9.7. Note that the table summarizes well the steps to be taken to predict eukaryotic genes.

- 9.7 Genome Annotation: pages 346-352.
  - Introduction: pages 346-347.
    Carefully read this section and understand every single word.
  - Genome annotation is the final step in genome analysis: pages 347-348.
    Trying to place a gene in a pathway is part of genome annotation and is a very challenging task. Read the section and skim through Figure 9.16.
  - Gene ontology provides a standard vocabulary for gene annotation: pages 348-352.
    - This section highlights the importance of gene ontology.
    - It also gives the names of important sites where one can find more information on genes such as the UCSC Genome Browser and OMIM.
  - Box 9.6 The Sjögren-Larsson syndrome: page 351.
    Skim through the section.

- 9.8 Large Genome Comparisons: pages: 353-354.
  - Introduction: pages 353-354.
    - Read the section.
    - Skim through Figure 9.20 and Figure 9.21.

- Summary: page 354.
  Read the short summary and understand every single word.

--- --- --- --- --- --- --- --- --- --- End of Part Four --- --- --- --- --- --- --- ---