6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

Lecture 1

Aims: An introduction to the course, and an introduction to molecular biology

Administrative Details:

-There were 4 handouts in class: a course information handout, the first problem set (due on Sept. 15 at 8 pm), the scribe policy, and a course survey. All of these are on the web page.

-The first precept was on Friday 9/5 at 11: pm. The precept notes are posted on the web page.

-There are 3 textbooks; they are on reserve at both MIT and at BU libraries. The course uses 3 books because no single book covers the all of the material in the class. No individual book covers both the algorithmic and the machine learning topics that will be addressed in the class. Also, because computational biology is a rapidly changing field, books quickly become outdated. The books are:

(1) Biological sequence analysis (Durbin, Eddy, Krogh, and Mitchison). The caveat regarding this book: it is about 10 years old, which is a long time in computational biology.

(2) An introduction to bioinformatics algorithms (Jones & Pevzner). As the title suggests, this book is a good resource for learning about bioinformatics algorithms.

(3) Pattern classification (Duda, Hart, and Stork). A machine learning book.

Why Computational Biology?

There are a number of reasons that it is appropriate and useful to apply computational approaches to the study of biological data.

-Many aspects of biology (such as sequence information) are fundamentally digital in nature. This means that they are well suited to computational modeling and analysis.

-New technologies (such as sequencing, and high-throughput experimental techniques like microarray, yeast two-hybrid, and ChIP-chip assays) are creating enormous and increasing amounts of data that can be analyzed and processed using computational techniques

-Running time & memory considerations are critical when dealing with huge datasets . An algorithm that works well on a small genome (for example, a bacteria) might be too time or space inefficient to be applied to 1000 mammalian genomes. Also, combinatorial questions dramatically increase algorithmic complexity.

-Biological datasets can be noisy, and filtering signal from noise is a computational problem.

-Machine learning approaches are useful to make inferences, classify biological features, & identify robust signals.

-It is possible to use computational approaches to find correlations in an unbiased way, and to come up with conclusions that transform biological knowledge. This approach is called data-driven discovery.

-Computational studies can suggest hypotheses, mechanisms, and theories to explain experimental observations. These hypotheses can then be tested experimentally.

-Computational approaches can be used not only to analyze existing data but also to motivate data collection and suggest useful experiments. Also, computational filtering can narrow the experimental search space to allow more focused and efficient experimental designs.

-Datasets can be combined using computational approaches, so that information collected across multiple experiments and using diverse experimental approaches can be brought to bear on questions of interest.

-Effective visualizations of biological data can facilitate discovery.

-Computational approaches can be used to simulate & model biological data.

Finding Functional Elements: A Computational Biology Question

We then discussed a specific question that computational biology can be used to address: how can one find functional elements in a genomic sequence? The slide that is filled with letters shows part of the sequence of the yeast genome. Given this sequence, we can ask:
-What are the genes that encode proteins? How can we find features (genes, regulatory motifs, and other functional elements) in the genomic sequence?

These questions could be addressed either experimentally or computationally. An experimental approach to the problem would be creating a knockout, and seeing if the fitness of the organism is affected. We could also address the question computationally by seeing whether the sequence is conserved across the genomes of multiple species. If the sequence is significantly conserved across evolutionary time, it's likely to perform an important function.

There are caveats to both of these approaches. Removing the element may not reveal its function—even if there is no apparent difference from the original, this could be simply because the right conditions have not been tested. Also, simply because an element is not conserved doesn't mean it isn't functional.

(Also, note that "functional element" is an ambiguous term. Certainly, there are many types of functional elements in the genome that are not protein-encoding. Intriguingly, 90-95% of the human genome is transcribed (used as a template to make RNA). It isn't known what the function of most of these transcribed regions are, or indeed if they are functional).

Course Outline:

The first half of the course will cover foundational material, the second half of the course will address open research questions. Each lecture will cover both biological problems and the computational techniques that can be used to study them. Some of the major topics that the course will address are:

1) Gene Finding
2) Sequence Alignment
3) Database Lookup: How can we effectively store and retrieve biological information?
4) Genome Assembly: How can we put together the small snippets of sequence produced by sequencing technologies into a complete genome.
5) Regulatory motif discovery: How can we find the sequence motifs that regulate gene expression?
6) Comparative genomics: How can we use the information contained in the similarities and differences between species' genomes to learn about biological function?
7) Evolutionary Theory: How can we infer the relationships among species using the information contained in their genomes?
8) Gene expression analysis
9) Cluster Discovery: How can we find emergent features in the dataset?
10) Gibbs Sampling: How can we link clusters to the regulators responsible for the co-regulation?
11) Protein Network Analysis: How can we construct and analyze networks that represent the relationships among proteins?
12) Metabolic Modeling
13) Emerging Network Properties

A Crash Course in Molecular Biology:

The final portion of the lecture was a crash course in molecular biology.

The Central Dogma of Molecular Biology

The central dogma of molecular biology specifies how information is stored and used in the cell. It states that DNA is transcribed into RNA, which is then translated into protein.

DNA

DNA is the molecule of heredity. Its structure reflects its ability to transmit information. DNA is a double helix, with a phosphate backbone and bases in the center. Since each complementary strand of the double helix contains sufficient information to reconstruct the other, there are 2 copies of the data, and a single strand can uniquely produce the other half. When the DNA is replicated, each strand is used as the template for a new, complementary strand. (Thus, after DNA is copied, each of the two resulting copies contain 1 old and 1 new strand).

The two strands of the double helix are related by complementary base-pairing, and connected by hydrogen bonds. DNA is composed of 4 nucleotides: A (adenine), C (cytosine), T (thymine), and G (guanine). A pairs only with T and C pairs only with G. A and T are connected by two hydrogen bonds, while C and G are connected by 3 bonds. Therefore, the A-T pairing is weaker than the C-G pairing. (For this reason, the genetic composition of bacteria that live in hot springs is ~80% G-C).

It is useful to have an abbreviation to refer to either of two bases because proteins may recognize DNA based on the characteristics of a particular subset of the bases. A transcription factor may be able to recognize either an A or a G in a specific position, for example, because of the shared biochemical and structural properties of adenine and guanine. Refer to the table in the lecture slides on “DNA: the four bases” to see the shared properties of pairs of bases and the abbreviations for each of the 6 different possible 2-base groupings.

The two DNA strands in the double helix are anti-parallel. DNA is not symmetrical: one end of a strand is the 5’ end, while the other is called the 3’ end. The 5’ end of one strand is adjacent to the 3’ end of the other. Replication occurs only in the 5’ to 3’ direction. For this reason, when DNA unzips for replication, it can be copied continuously in the 5’ to 3’ direction for one of the strands (the leading strand), but only piecewise in the 5' to 3' direction for the other strand (the lagging strand). Genes and other functional elements can be found on either strand of DNA. By convention, DNA sequences are always written in the 5’ to 3’ direction.

The protein coding portions of DNA are called exons, while introns are non-coding "intervening" regions found in genes. In prokaryotic cells, DNA is not localized to a compartment. In eukaryotic cells, DNA is stored in the nucleus of the cell in a highly compact form. Thus, DNA must be locally “unpacked” before it can be replicated or transcribed into RNA.

RNA

RNA is produced when DNA is transcribed. It’s structurally similar to DNA. However, in RNA, uracil (U) is substituted for T. Also, RNA contains ribose instead of deoxyribose (a difference of one oxygen). There are many different types of RNA:

- mRNA (messenger RNA) contains the information to make a protein and is translated into protein sequence. Immediately after the DNA sequence is transcribed, the new RNA molecule is known as pre-RNA. In eukaryotes, introns must be removed from this pre-RNA (also called primary transcript) to produce mature mRNA. Different regions of the primary transcript may be spliced out to lead to different protein products (alternative splicing).

· tRNA (transfer RNA) specifies codon-to-amino-acid translation. It contains a 3 base pair anti-codon complementary to a codon on the mRNA, and carries the amino acid corresponding to its anticodon attached to its 3' end.
· rRNA (ribosomal RBA) forms the core of the ribosome, the organelle responsible for the translation of mRNA to protein .
· snRNA (small nuclear RNA) is involved in splicing (removing introns from)pre- mRNA, as well as other functions.
Other functional kinds of RNA exist and are still being discovered. RNA molecules can have complex three-dimensional structures and perform diverse functions in the cell. In fact, there is an hypothesis (the "RNA world" hypothesis) that early life was entirely RNA-based. According to this hypothesis, RNA served as both the information repository and the functional workhorse in early organisms.

Protein
Protein is the molecule responsible for carrying out most of the tasks of the cell. Like RNA and DNA, proteins are polymers made from repetitive subunits. Proteins are built out of 20 different amino acid subunits. Each amino acid has special properties of size, charge, shape, and acidity. Since there are 20 amino acids and only 4 nucleotides, a sequence of 3 nucleotides is required to specify a single amino acid. In fact, each of the 64 possible 3-sequences of nucleotides (codon) uniquely specifies a particular amino acid (or is a stop codon that terminates protein translation). Since there are 64 possible codon sequences, the code is degenerate, and some amino acids are specified by multiple encodings. Most of the degeneracy occurs in the 3rd codon position. The three-dimensional shape, and thus the function, of a protein is determined by its sequence (however, solving the shape of a protein from its sequence is an unsolved problem in computational biology!)

Regulation: from molecules to life
Transcription is one of the steps at which protein levels can be regulated. The promoter region, a segment of DNA found upstream (past the 5' end) of genes, functions in transcriptional regulation. The promoter region contains motifs that are recognized by proteins called transcription factors. When bound, transcription factors can recruit RNA polymerase, leading to gene transcription. However , transcription factors can also participate in complex regulatory interactions. There can be multiple binding sites in a promotor, which can act as a logic gate for gene activation. Regulation in eukaryokes can be extremely complex, with gene expression affected not only by the nearby promoter region, but also by distant enhancers and repressors.

We can use probabilistic models to identify genes that are regulated by a given transcription factor. For example, given the set of motifs known to bind a given transcription factor, we can compute the probability that a candidate motif also binds the transcription factor (see the notes for precept #1). Comparative sequence analysis can also be used to identify regulatory motifs, since regulatory motifs show characteristic patterns of evolutionary conservation.

The lac operon in *E. coli* and other bacteria is an example of a simple regulatory circuit. In bacteria, genes with related functions are often located next to each other, controlled by the same regulatory region, and transcribed together; this group of genes is called an operon.
The lac operon functions in the metabolism of the sugar lactose, which can be used as an energy source. However, the bacteria prefer to use glucose as an energy source, so if there is glucose present in the environment the bacteria do not want to make the proteins that are encoded by the lac operon. Therefore, transcription of the lac operon is regulated by an

elegant circuit in which transcription occurs only if there is lactose but not glucose present in the environment.

Metabolism

Metabolism regulates the flow of mass and energy in order to keep an organism in a state of low entropy. Metabolic reactions can be grouped into two basic types. In catabolism, complex molecules are broken down to release energy. In anabolism, energy is used to assemble complex molecules.
Enzymes are a critical component of metabolic reactions. The vast majority of (but not all!) enzymes are proteins. Many biologically critical reactions have high activation energies, so that the uncatalyzed reaction would happen extremely slowly or not at all. Enzymes speed up these reactions, so that they can happen at a rate that is sustainable for the cell.

In living cells, reactions are organized into metabolic pathways. A reaction may have many steps, with the products of one step serving as the substrate for the next. Also, metabolic reactions often require an investment of energy (notably as a molecule called ATP), and energy released by one reaction may be captured by a later reaction in the pathway. Metabolic pathways are also important for the regulation of metabolic reactions—if any step is inhibited, subsequent steps may lack the substrate or the energy that they need to proceed. Often, regulatory checkpoints appear early in metabolic pathways, since if the reaction needs to be stopped, it is obviously better to stop it before much energy has been invested.

Systems Biology

Systems biology strives to explore and explain the behavior that emerges from the complex interactions among the components of a biological system. One interesting recent paper in systems biology is “Metabolic gene regulation in a dynamically changing environment” (Bennett et al., 2008). This work makes the assumption that yeast is a linear, time-invariant system, and runs a signal (glucose) through the system to observe the response. A periodic response to low-frequency fluctuations in glucose level is observed, but there is little response to high-frequency fluctuations in glucose level. Thus, this study finds that yeast acts as a low-pass filter for fluctuations in glucose level.

Synthetic Biology

Not only can we use computational approaches to model and analyze biological data collected from cells, but we can also design cells that implement specific logic circuits to carry out novel functions. The task of designing novel biological systems is known as synthetic biology.

A particularly notable success of synthetic biology is the improvement of artemesenin production. Artemesenin is a drug used to treat malaria. However, artemisinin was quite expensive to produce. Recently, a strain of yeast has been engineered to synthesize a precursor to artemisinic acid at half of the previous cost.