CSE 527: Computational Molecular Biology September 29, 2004

Lecture 1

Lecturer: Larry Ruzzo *Scribe: Mukund Narasimhan*

1 Administratrivia

- Class Web Site : http://www.cs.washington.edu/education/courses/527/04au.
- Class Syllabus : http://www.cs.washington.edu/education/courses/527/04au/syl.pdf.
- Related Classes
  - Genome 540/541 (Introduction to Computational Molecular Biology).
  - Stat/Biostat 578 (Statistical Analysis of Microarrays).
  - CSE 590CB (Reading/Research in Computational Biology).
  - Genome 521 (Combi Seminar)
- Homework #1 : Find and read a good primer on either *Biology for Computer Scientists* or *Computer Science for Biologists*. Send email with a citation (or link) critiquing it.

2 Introduction/Motivation

Moore's Law predicts a doubling of the number of transistors every 18 months. GenBank has had similar exponential growth. Contrary to reports that the *human genome has been finished*, gathering this data is only the beginning. The real problem is mining this vast database for useful information, and that has only just begun. This explosive growth in biological data is revolutionizing biology and medicine, and "pre-genomic lab techniques are obsolete" in the sense that mathematicas and computational techinques are an essential component of analysis in the post-genomic lab.

3 Biology Review

Genetics is the study of heredity. The *Genome* consists of the hereditary information present in every cell. This information in encoded in DNA molecules, which are long sequences of nucleotides, A (Adenine), C (Cytosine), T (Thymine), G (Guanine). Humans have about 3×10^9 nucleotides. The problems of extracting and interpreting the genomic information, applying this information to the genetics of disease, and better understanding evolution are part of the problems of the genome project.

3.1 DNA

While DNA was discovered in 1869, its role as the carrier of genetic information was realized much later. The double helix structure of DNA was discovered by Watson & Crick in 1953. The two strands which form the double helix are each formed by a chain of nucleotides. The nucleotide pairs A/T and C/G are complementary and always bind to each other. The genetic information is encoded in this linear ordering of nucleotides.

A gene, classically, is an abstract heritable attribute existing in variant forms (alleles). Generally, it is taken to mean a part of the genetic code sufficient to define one protein. Mendel studied transmission of genetic information in pea plants, and his studies led him to conclude that

- Each individual has two copies of each gene.
- Each parent contributes one (randomly).
- Independent Assortment takes place.

3.2 Cells

Cells are a bunch of chemicals in a sac, a fatty layer called the *plasma membrane*. *Prokaryotic* cells have no recognizable nucleus, have very little substructure, and are more homogenous. *Eukaryotic* cells have their genetic material stored in a separate nucleus, and have other organelles for specialized functions. While Prokaryotic cells are present in unicellular organisms like bacteria, Eukaryotic cells are present in all multicellular organisms and many single celled ones like yeast. The genetic material is organized into chromosomes, which is a pair of DNA molecules (along with a protein wrapper). Most prokaryotes have just one chromosome. In Eukaryotes, all cells have the same number of chromosomes (8 in fruit flies, 46 in humans and bats, 84 in rhinoceros etc). A diploid cell has homologous pairs of chromosomes, one maternal and one parternal except for sex chromosomes. A haploid cell has only one copy of each chromosome. Most "higher" Eukaryotic cells are diploid.

When cell division occurs through *Mitosis*, each chromosome is duplicated, and one copy goes to each daughter cell. When cell division occurs through *Meiosis*, two cell divisions create four haploid cells. During Recombination/Crossover material is exchanged between paternal and maternal copies, and then through fertilization sperm and egg cells combine to form a diploid zygote.

3.3 Proteins

A protein is a chain of amino acids, of which there are 20 types. Proteins are the major functional elements in cells. The function of proteins is determined by the 3D structure into which the protein folds. Proteins make up, e.g.,

- Cellular structure.
- Enzymes (to catalyze chemical reactions).
- Receptors (for hormones, odorants and other signaling molecules).
- Transcription factors.

The functionality of the protein is determined by the 3D-structure into which it folds. However, determining the 3D-structure of the protein from the amino acid sequence is a major open problem.

3.4 The Central Dogma

The Central Dogma asserts that DNA is encoded into messenger RNA which then migrates to the ribosomes which reads the RNA and makes proteins through the triplet code (codons). The process of going from DNA to mRNA is called transcription, and the process of going from mRNA to protein is called translation. Going from mRAN to DNA is called reverse transcription and is what retro-viruses do in order to get their genetic material incorporated into that of the cell. Many functionally different types of RNA exist including mRNA, tRNA and rRNA.

The process of translation which is initiated in a region of the DNA called the promoter which is near the 5' end. A's become U's, T's become A's, C's become G's and G's become C's in the mRNA. The U is another nucleotide that basically holds the same position that a T usually would. Each codon (a sequence of 3 nucleotides) codes for an amino acid or a special start/stop value. Three pairs and four nucleotides allow for $4^3 = 64$ different values, and there are 20 different nucleotides. Hence there are several different codes for the same amino acid. In most Eukaryotes, after transcription, sequences known as introns are spliced out while exons are spliced together. Complex control of transcription rate is achieved by having proteins that bind to DNA (sometimes far from the site where transcription actually starts).