CSE 527	Professor Larry Ruzzo	Lecture 1
Class Notes – September 27, 2006		by: Francisco M. Cruz Morales

Administrative Information

The class website at http://www.cs.washington.edu/education/courses/527/06au/ is the main source of information for the course. Additional information provided in class:

- Course homework centered on reading assignments. The required text is *Biological Sequence Analysis* by Durbin et al.
- No formal exam, grading based on homework –related to readings – and final project.
- Material on the course will cover mostly Sequence Analysis, with some later focus on microarrays. We will discuss enough bio to motivate the problems and discussion.

Background and Motivation

As Moore's Law has steadily marked the pace of integration of computing power in new chips, so has the growth of data (nucleotides) in the GenBank. It is not simply that endeavors like the Human Genome Project generate large amounts of raw data; it has become evident that these bits of information are interrelated, conforming complex system maps, slightly resembling combinatorial logic problems.

This flood of data brings the challenging task of handling this amount of information. Computational and mathematical techniques become crucial to biological and medical research in the post-genomical analysis.

Introduction to Molecular Biology Nomenclature

Genetics

The *genome* is defined as the hereditary information present in every cell, and the *DNA double helix*, each one composed of a chain of nucleotides, is its exposed structure. While DNA was discovered in 1869, it was not until much later that we learned of its role as carrier of genetic information.

Genetics is the study of heredity, classically referred as Mendelian genetics. From this view, each individual's cells carry a randomly selected copy of a *gene* – an heritable attribute present in variant forms, or *alleles* – from each parent.

Cells

In its simplest expression, a *cell* is simply an assortment of chemicals inside a sac, a fatty layer called the *plasma membrane*. According to its complexity, cells are taxonomically divided into *Prokaryotes* – without nucleus or more than a few simple structures, although most of them have one *chromosome* – and *Eukaryotes* – carrying their genetic material in the nucleus, and presenting other *organelles* performing specialized functions.

Chromosomes are pairs of complementary DNA molecules with a protein wrapper. As mentioned above, most prokaryotes have only one chromosome. In eukaryotes, the number of chromosomes varies among species, but is constant within a species.

Diploid cells – like most of the eukaryotes' – have two copies of each gene from homologous pairs of chromosomes (one from each parent.) *Haploid* cells have only one copy of each chromosome.

Cell multiplication happens following two different methods: with *mitosis*, each chromosome is duplicated, and a copy goes to each daughter cell. In the case of *meiosis*, two subsequent divisions form four haploid gametes. In this process we observe that recombination and crossover of segments remove some randomness from the hereditary process.

Proteins

Proteins are chains composed of elements from a set of 20 amino acids, and are the major functional elements in the cell: provide *structure*, catalyze chemical reactors (*enzymes*), act as *receptors*, factors binding to DNA to regulate *transcription*, …
Such a chain is not a simple linear structure: a fundamental feature of the protein is its three-dimensional structure. Its shape is coded within the sequence, although it is still unclear how. Nevertheless, this *folding structure* is a crucial problem, much more critical than *protein sequencing*, which has been more widely studied.

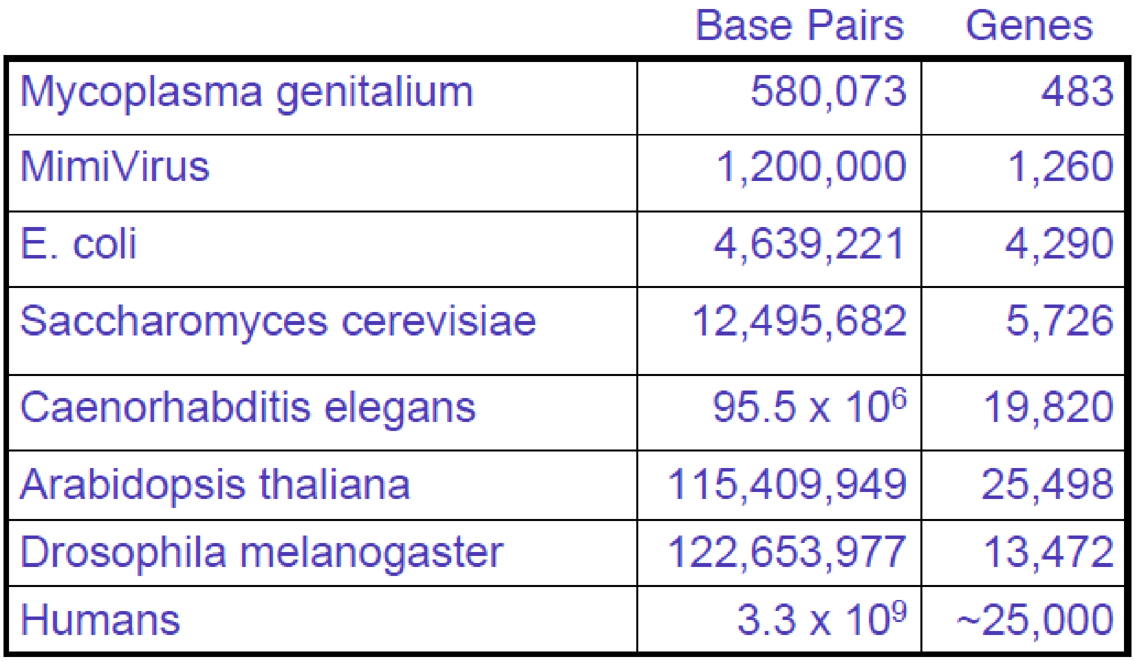
The Central Dogma

One-Directional Flow: DNA –> mRNA –> Protein.

- DNA is transcribed to mRNA
- mRNA translated into proteins
- Amino acids encoded as triplets (*codons*) of nucleotides (see table in slides)

The encoding of proteins is such a critical process that may have developed a certain level of redundancy to increase its reliability: with four nucleotides in sets of three, there can be up to 4^3 = 64 amino acids, but as we said above, only 20 are available.
As DNA never leaves the nucleus – the cytoplasm is a hostile environment – mRNA plays the role of carrier of the genetic information outside, where the proteins are created. mRNA is produced from DNA in a process called *transcription*: it is managed by a protein molecule called *RNA polymerase*, which guides the location of the nucleotide that is produced at each step of the transcription. Later, outside of the nucleus, in the process called *translation*, ribosomes run through the mRNA building the protein.

Genome Sizes



	Base Pairs	Genes
Mycoplasma genitalium	580,073	483
MimiVirus	1,200,000	1,260
E. coli	4,639,221	4,290
Saccharomyces cerevisiae	12,495,682	5,726
Caenorhabditis elegans	95.5×10^6	19,820
Arabidopsis thaliana	115,409,949	25,498
Drosophila melanogaster	122,653,977	13,472
Humans	3.3×10^9	~25,000

It is illustrating to notice a few surprising facts from the above table:

- Organisms that are more anatomically and functionally complex than others may have similar number of base pairs or genes.
- Some virus present greater complexity than some bacteria.

Research has shown other interesting factors:

- Splicing of proteins in different ways may compensate for the lower number of genes that human present
- Identified functions in regions of the genome previously considered *junk*
- Presence of identical patterns of non-coding regions that are preserved across species (as in some regions common for all vertebrates.)