# Chapter 1

# Introduction

## Preliminaries

> "*The spread, both in width and depth, of the multifarious branches of knowledge during the last hundred odd years has confronted us with a queer dilemma. We feel clearly that we are only now beginning to acquire reliable material for welding together the sum total of all that is known into a whole; but, on the other hand, it has become next to impossible for a single mind fully to command more than a small specialized portion of it.*
>
> "*I can see no further escape from this dilemma (lest our true aim be lost for ever) than that some of us should venture to embark on a synthesis of facts and theories, albeit with second-hand and incomplete knowledge of some of them – and at the risk of making fools of ourselves.*
>
> "*So much for my apology.* "

—Erwin Schrödinger, **What is Life?**

## 1.1 Information

**Bud Mishra**
Room 801, Warren Weaver Hall
Courant Institute
Tel #: 212.998.3464
e-mail: `mishra@nyu.edu`
URL:`http://cs.nyu.edu/cs/faculty/~mishra/index.html`

## 1.2 Computational Biology X

The goal of the first lecture was to design a syllabus for the course in accordance with the interest and aptitude of the students of the class. Since it was unclear what the composition of the class would be, the syllabus that we will come up with remains somewhat of a mystery. Only to the extent that I can influence this class in designing a meaningful syllabus, I offer a big picture of the subject and a set of topics currently of active research. Here are my choice of topics:

**Human Genome Project:** Read 3 billion base pairs in 46 human chromosomes.

**Single Nucleotide Polymorphisms:** Catalog the single base pair variations occurring about 1 in 800 base pairs of human genome over the entire population.

**Gene Hunting:** Identify all (about 100,000) the genes in human genome.

Particularly interesting are the ones involved in cancer—about 100 *oncogenes*.

**Linkage Analysis:** Relate genes to phenotypes (externally observable traits) by analyzing genomes in a family or over a population.

**Functional Genomics:** Understand how an interactive network of genes affect a chain of metabolic pathways to ultimately determine the phenotypes.

**Cell Informatics:** Interaction between proteins (membrane and soluble ones) to determine the dynamics of a cell.

Interaction among a heterogenous population of cells.

**Rational Drug Design:** Design of drugs and delivery systems to modify the dynamics of the cells.

**Phylogenomics:** Relates genes within and across species to understand their evolutionary relationship.

**DNA Computers:** Build highly parallel computers using DNA strands as data encoding elements.

**DNA Nanorobots:** Build highly parallel and cooperative nanorobots with actuators and sensors (as well as distributed controllers) all built out of DNA and amino acid sequences.

### 1.2.1 Biological Problems

**Pairwise and Multiple Sequence Alignment:**
(Dynamic programming, similarity matrices, competitive heuristics)

**Fragment and Map Assembly:**
(Interval Tree and Other Graph Theoretic Approaches, Bayesian Approaches)

**Sequence Feature Extraction:**
(Data Mining, Bioinformatics— Bayesian Inference, HMM (Hidden Markov Models), Neural Networks, Genetic Algorithms)

**Phylogenomics:**
(Phylogenetic Tree Construction)

**RNA Secondary Structure Prediction:**
(??)

**Proteonomics:**

- Protein Homology Modeling
- Protein Threading
- Protein Molecular Dynamics
- Protein *ab initio* Structure Prediction

### 1.2.2 Computational Tools

**Bayesian Statistics:**

- Hidden Markov Models (HMM)
- Expectation Maximization (EM)
- Monte Carlo Methods
- Neural Networks
- Genetic Algorithms
- Bounded/Constrained Search
- Simulated Annealing

**Combinatorial Approaches:**

- Stringology
- Interval Graphs
- Tree Algorithms
- Dynamic Programming

**Information Theoretic Approaches:**

- Entropy Maximization
- Competitive Methods (Universal Schemes)
- Stochastic Control

### 1.2.3 Technology

**Cloning:**

- *In vivo* methods
- *PCR* (Polymeric Chain Reaction)

**Mapping:**

- Fingerprints
- Multiple Complete Digestion
- Optical Mapping
- Radiation Hybridization

**Sequencing:**

- Sanger Sequencing
- Sequencing by Hybridization
- MALDI-TOF (Mass Spectrometry)
- Other Single molecule methods

**Probing:**

- In situ Hybridization
- Gene Chips
- Southern Blotting

## 1.3 State-of-the-Arts

What can be accomplished in Genomics can be inferred by taking a closer look at various completed and ongoing genomic projects. Many of the completed genomic projects deal with microbes (2–3 Mb genome size).

These organisms had been selected for different reasons. Since *E. coli* is one of the best characterized organism both genetically and biochemically, it was a natural choice. *B. subtilis* was chosen because it is Gram positive, as opposed to *E. coli*, which is Gram negative. Also, *B. subtilis* goes through differentiation during the sporulation process. *S. cerevisae* is eukaryotic, but can be handled much the same way as any other microorganism. Since, it has chromosomes in a nucleus and undergoes meiotic and mitotic processes, its genome is likely to tell us a lot about other eukaryotes. The nematode worm, *C. elegans*, is a simple multicellular organism with about 2000 cells, and biologists have already mapped the line of descent from zygote for each of these cells. So the genome sequence of this organism is of considerable interest to developmental biologists. It will then be possible to see how and which genes are expressed and when and where the different cell lineages branch off during the differentiation.

1. *Haemophilus influenzae*: First organism to have its genome completely sequenced. (Fleischmann, TIGR, 1995). A genome of 1.8Mb encoding 1743 genes.

2. *Escherichia coli*: A gram negative bacterium. K12 is the common strain and not virulent. However, the strain 0157 has been implicated in serious virulent outbreaks. (Blattner, Wisconsin). A genome of 4.6Mb encoding 4300 genes.

3. *Bacillus subtilis*: A gram positive bacterium. (Kunst, 46 laboratories involving 160 people). A genome of 4.2 Mb encoding 4100 genes.

4. *Synechocystis PCC6803*: A cyanobacterium. (Kaneko)

5. *Mycobacterium tuberculosis*: The causative agent of tuberculosis.

6. *Treponema pallidum*: The spirocehete causing syphillis.

7. *Borrelia burgodorferi*: The spirocehete causing Lyme disease.

8. *Deinococcus radiodurans*: An organism capable of withstanding unusually high degree of UV radiation.

9. *Aquifex aeolicus*: A marine hypothermophile capable of growth at $95°C$. It represents the deepest lineage within the bacterial domain and may hold clues to primitive prokyrotic life-forms.

10. *Caenorhabdeitis elegans*: The nematode worm—first animal genome to be completely sequenced.

11. *Saccharomyces cervisiae*: The brewer's yeast, an eukaryote. A genome of 12 Mb with 5900 genes. (96 labs with 640 people in 6 years).

12. *Arabidopsis*: A plant—ongoing.

13. *Oryza sativa*: Rice—ongoing.

14. *Homo sapiens*: About 3% completed.

The information gleaned from these genome projects have already elucidated several biochemical processes of fundamental importance in understanding life itself. Here are some examples:

1. The genome of *D. radiodurans* indicated that it is the first non-photosynthetic organism to possess the light-sensing protein phytochrome, which regulates synthesis of pigments against radiation. It also has a new type of RecA protein that helps it in DNA-repair. The related metabolic pathway repairs DNA breakages (more than 150 breakages at a time), caused by UV radiation. It was previously known that this octaploid organism has enough redundancy to provide sufficient information to assist the repair.

2. Study of *E. coli* and *B. Subtilis* genomes shed more light on bacterial DNA restriction-modification systems (enzyme systems that protect bacteria from viral DNA by cutting it up to small pieces while leaving its own DNA unharmed

by a methylation process). These bacteria seemed to have 3 to 4 such restriction-modification systems, while more pathogenic variety seem to have even many more. (*H. influenza*, 7, *N. gonorrhoeae*, 18, *H. pylori*, 23).

3. New metabolic pathway in *E. coli* for sugar acid idonate was discovered.

4. Comparison of pathogenic *E. coli* (0157) strain with the non-pathogenic strain (K12) showed that 0157 has a much bigger genome (about 1.2 Mb more, representing a 20% increase). In the extra genetic material the organism seem to code for virulence factor-bearing prophage. Also, there seem to be "pathogenicity island" in that region that allows the bacteria to secrete toxic protein into the host cells. More surprisingly, the extra DNA in 0157 (the pathogenic strain) is shared by *Y. pestis*, the agent involved in *plague*!

5. Many bacteria were found to have metabolic pathways for the inter-conversion of five- and six-carbon sugars. Previously, it was thought that this was only possible in methanogenic bacteria.

6. Many new protein paralogs of unknown function have been discovered.

7. One can now estimate metabolic capability of an organism easily. For instance, one can estimate the nutrients that can be taken up by a cell. This knowledge is rather important in choosing bacteria for bioremediation.

8. Comparative genomics allows the researchers to infer horizontal gene-transfer between related gram-negative bacteria.

9. It helps the biotechnologists to construct expression vectors, gene knockouts, reporter-gene, etc.