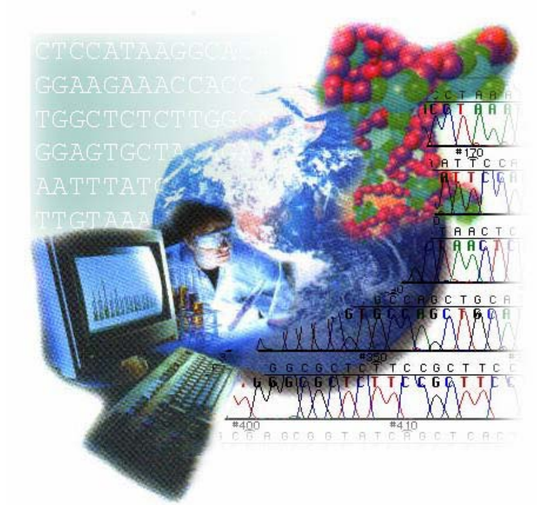
Bioinformatics Algorithms - Primer

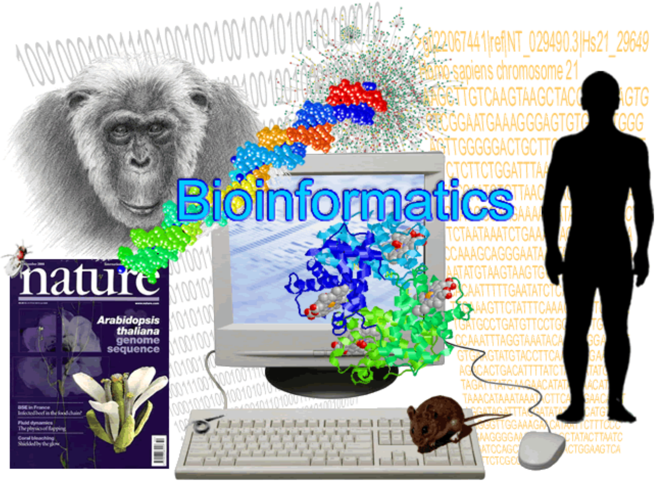
What is bioinformatics

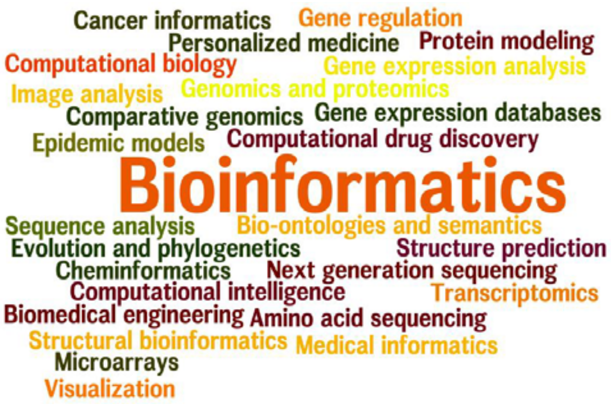
The systematic development and application of computing systems and computational solution techniques to the analysis of biological data obtained by experiments, modeling, database search and instrumentation



So, what is bioinformatics again?

Wikipedia: Using techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems usually on the molecular level.


Bioinformatics
gi|22067441|ref|NT_029490.3|Hs21_29649
Homo sapiens chromosome 21
nature
Arabidopsis thaliana genome sequence

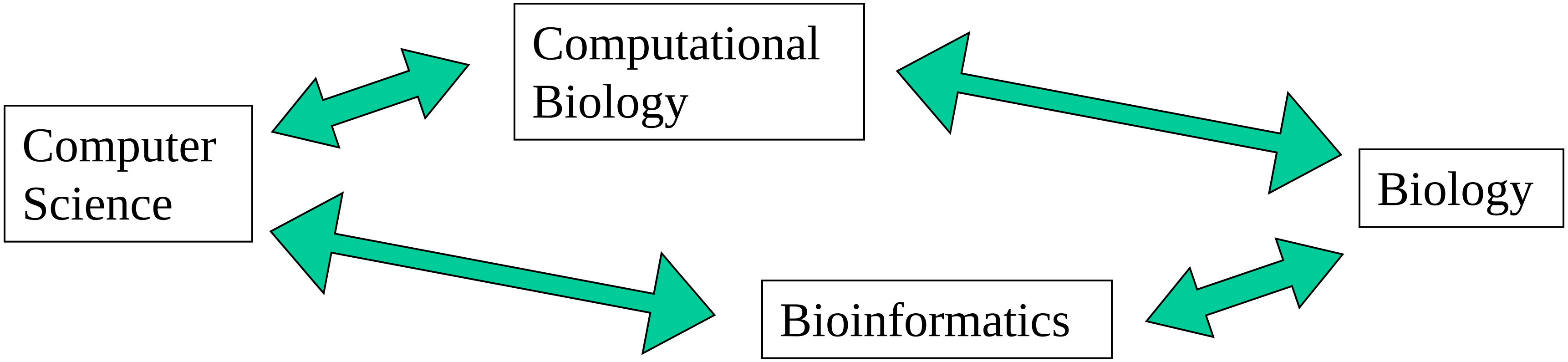


And what is Computational Biology?

The same thing?!

Some people consider computational biology closer to the computational sciences than to life sciences.

Schematically:



Computational Biology

Computer Science

Biology

Bioinformatics

Major research areas in bioinformatics

- Sequence Analysis (Alignment, sequencing, motif finding,…)
- Genome annotation
- Computational evolutionary biology
- Measuring biodiversity
- Gene expression
- Regulation
- Analysis of mutations
- Association studies
- Protein (RNA) structure prediction
- Comparative genomics
- Modeling biological systems
- Image Analysis
- Protein - protein (DNA) docking
- …

Why Bioinformatics in CS?

Bioinformatics is in a big part the application of a core technology of computer science (e.g. *algorithms*, artificial intelligence, databases) to problems arising from biology.

Bioinformatics is particularly exciting today because (1) the problems are large enough to motivate efficient algorithms, (2) the problems are accessible, fresh and interesting, (3) biology is increasingly becoming a computational science.

Developments in biology are coming astonishingly quickly, and with amazing possibilities.

Bioinformatics interest is increasing in both life science and computational science departments, not to mention industry.

Most problem ideas go from biology to CS: e.g. fragment assembly, sequence analysis, algorithms for phylogenic trees.

Some problem ideas go from CS to biology: e.g. sequencing by hybridization, DNA computing.

Example: Attack on the SARS Virus

The scientific reaction to the outbreak of the SARS virus after being first reported in Asia in February 2003 illustrates the critical roles that genomics and computation play in modern biology.

- DeRisi's analysis of microarray data reveals that the agent was a coronavirus in March 2003.
- The Michael Smith Genome Sciences Centre in Canada announce the sequencing of the SARS virus genome on April 12, 2003.
- Commercial SARS-specific microarrays become available in April 15, 2003.
- Gene predictions and analysis of SARS genome published in Science online May 1, 2003.
- Phylogenic analysis placing where SARS fits among the coronaviruses published in late May 2003.

We will study the computational problems of sequence assembly, gene prediction, microarray design/analysis, and phylogenic tree construction in this course.

Computer Scientists vs. Biologists

There are many different types of life scientists (biologists, ecologists, medical doctors, etc.), just as there are many different types of computational scientists (algorists, software engineers, statisticians, etc.).

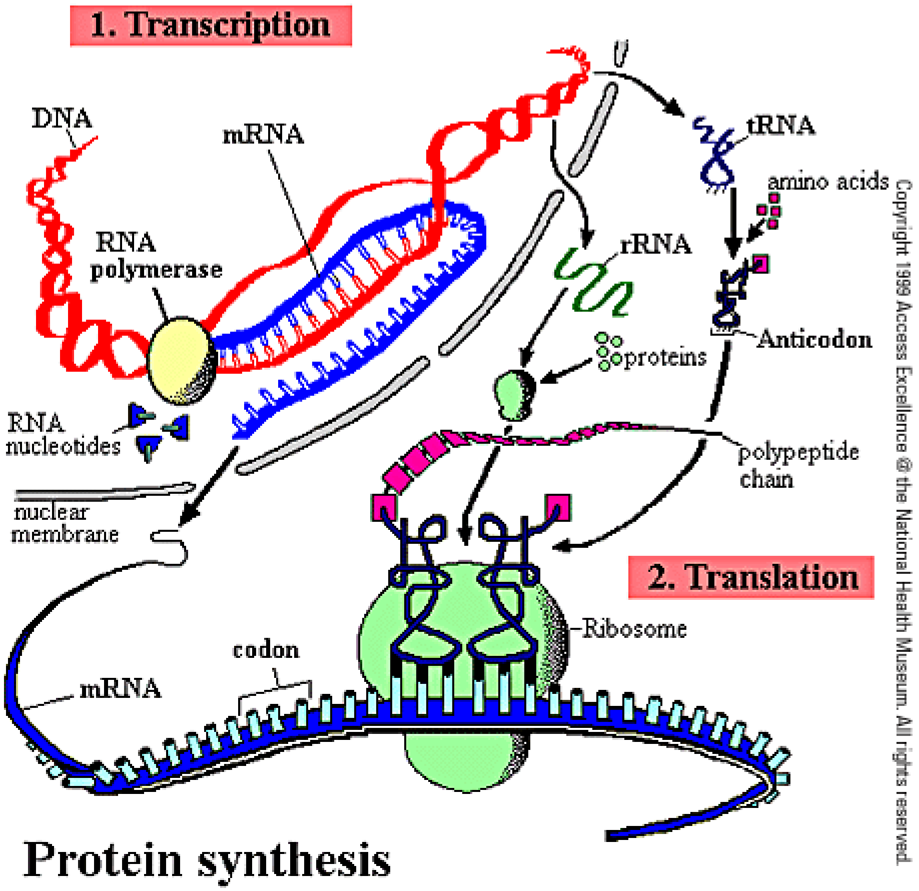
There are many fundamental cultural differences between computational and life scientists:

- *Almost nothing* is ever completely true or false in biology, where *everything* is either true or false in computer science / mathematics.
- Biologists strive to understand the very complicated, very messy natural world; computer scientists seek to build their own clean and organized virtual worlds.
- Biologists are *data* driven; while computer scientists are *algorithm* driven. One consequence is that quite often CS WWW pages have fancier graphics while Biology WWW pages have more content.
- Biologists are much more obsessed with being the first to discover something; computer scientists obsess with inventing something.
- Research biologists usually have to know more than computer scientists; computer scientists try to learn how to do more.

Computer Scientists vs. Biologists

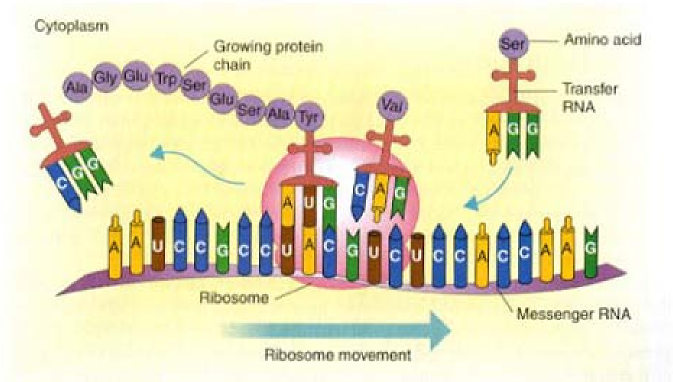
- Biologists are comfortable with the idea that all data has errors; computer scientists are not.
- Biologists live in stronger hierarchies than computer scientists: PI > postdocs > graduate students > lab assistants.
- The Platonic ideal of a biologist is running a big laboratory with many people. The Platonic ideal of a computer scientist is a hacker in garage.
- Biologists can get/spend infinitely more research money than computational scientists.
- Biotechnology/drug companies are largely science driven, while the computer industry is more engineering/marketing driven.
- Biologists seek to publish in prestigious journals like *Science* and *Nature*. Computer scientists seek to publish in prestigious refereed conference proceedings.
- One consequence is life science journals get refereed faster than computational science journals.
- Computer scientists can get interesting, high-paid jobs after a B.S. Biologists typically need to complete one or more postdocs...

Biology for Computer Scientists
Basic Functions



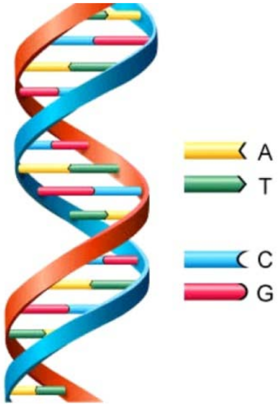
Copyright 1999 Access Excellence @ the National Health Museum. All rights reserved.

During translation, the genetic code in mRNA is read and converted into protein by means of the protein synthesizing machinery, which consists of ribosomes, tRNA, amino acids, and a number of enzymes.



Biology for Computer Scientists

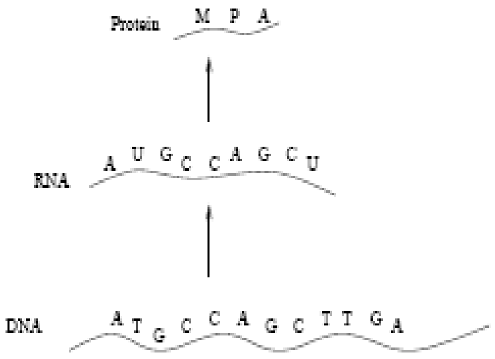
- DNA sequences can be thought of as strings of bases on a four-letter alphabet, {A, C, G, T}.
- Each base binds with its complement, A-T and C-G , so each sequence has a unique complementary sequence.
- The human genome is approximately 3 billion base-pairs long, and contains all the information necessary to make all the *proteins* which you are made of.
- Proteins are sequences of amino acids, and hence all proteins can be thought of as strings on a 20-letter alphabet.
- A *gene* is a DNA sequence which acts as a template for building a specific protein (and other molecules).
- Genes specify how to build proteins according to the *triplet code*, where each of the $4^3 = 64$ possible sequences or *codons* of three consecutive nucleotide bases map to one of the 20 different amino acids or the stop symbol.



Biology for Computer Scientists

RNA is an intermediate step in the translation process, and maps 1-to-1 with DNA.

The human genome contains about 30,000 protein coding genes, meaning that your body is made up of at least that many different components.

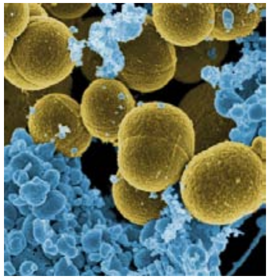


The “completed” human genome project seeks to *sequence* or read the entire set of DNA and protein strings.

But sequencing is just a first step towards understanding what the proteins do and how to manipulate them.

Only a small portion of the human genome consists of protein coding genes. The rest contains encodings of other types of genes, various binding/signaling sites, and less well understood “junk”.

Organisms



Living organisms differ greatly in complexity and organization.

Viruses are simplest organisms (~10,000 base pairs or bases long), which require a living host.

Prokaryotes are simplest free-living organisms, e.g. bacteria (~1,000,000 bp. long).

Eukaryotes have cells which contain internal structures such as a nucleus, e.g. yeast.

Multi-celled organisms involve cell specialization, requiring differential gene expression and inter-cellular signaling.

Historically, many biologists focused their careers on one model organism: E. Coli, yeast, drosophila, arabadopsis, zebrafish, sea urchins, mouse.

The advent of genomics has focused more attention on the similarity between organisms.

Evolution

Evolutionary change happens because of changes in genomes mainly due to *mutations* and *recombination*.

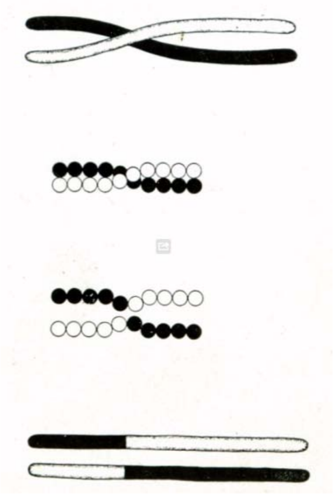
Mutations are rare events, sometimes single base changes, sometimes larger events.

Recombination is how your genome was constructed as a mixture of your two parents.

Through *natural selection*, favorable changes tend to accumulate in the genome.

Evolution motivates *homology* (similarity) search, because different species are assumed to have common ancestors.

Thus DNA/amino acid sequences for a given protein (e.g. hemoglobin) in two species or individuals should be more similar the closer the ancestry between them.



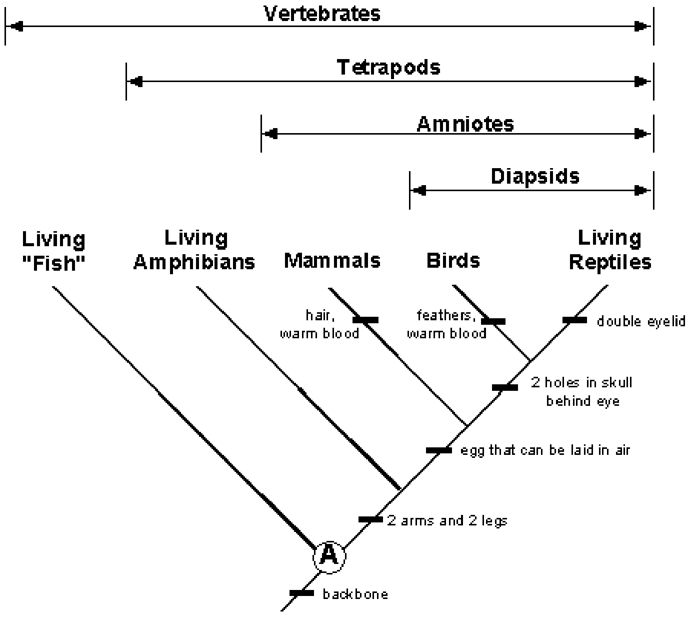
Evolution

The genetic variation between different people is surprisingly small, perhaps only 3 in 1000 base-pairs.

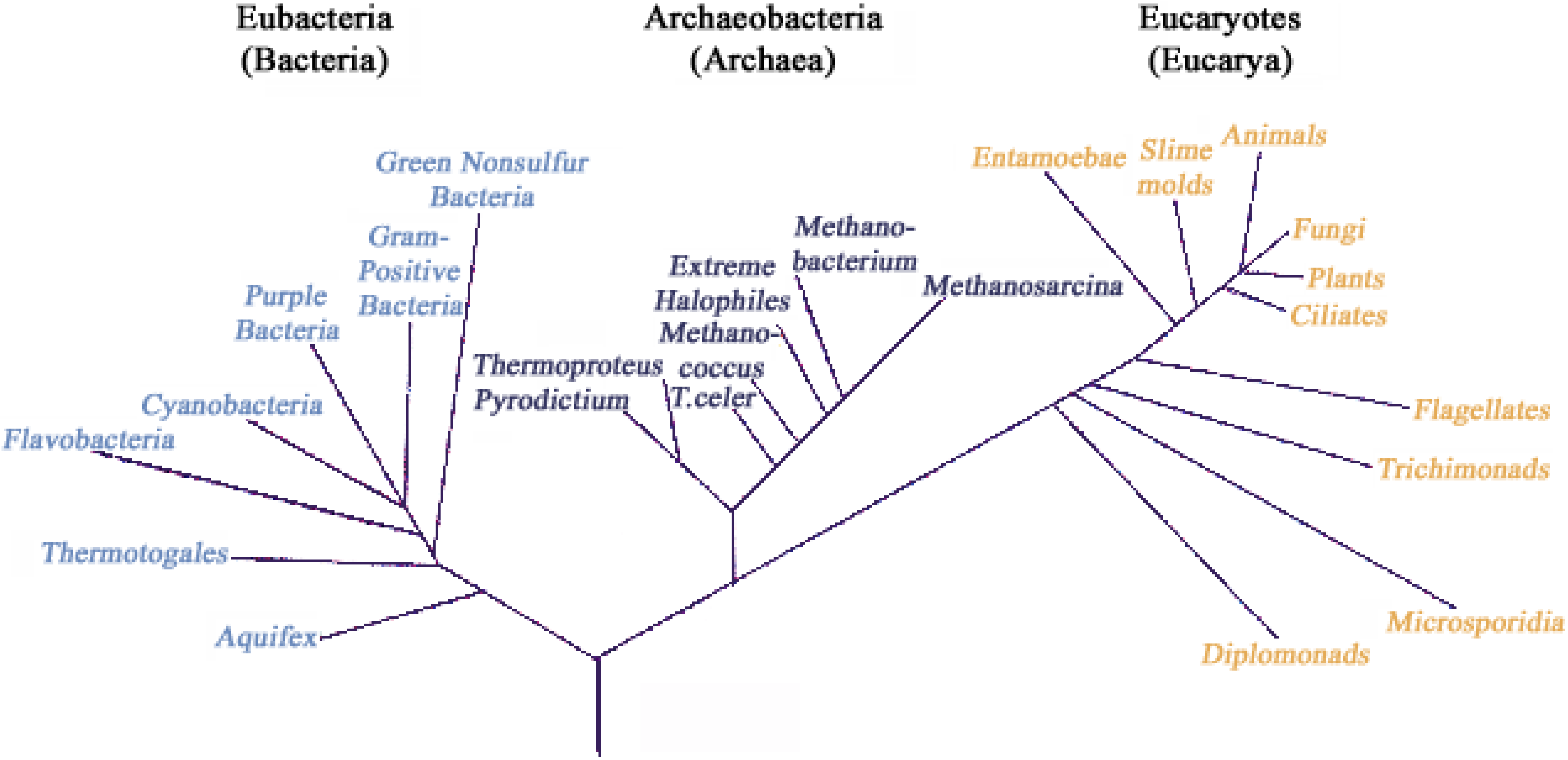
Homology searches can often detect similarities between extremely distant organisms (e.g. humans and yeast).

Phylogenic trees based on gene homologies have provided an independent confirmation of many phylogenies proposed by taxonomists. This is convincing evidence of the Theory of Evolution.

A host of interesting computational problems arise in trying to reconstruct evolutionary history.



Evolution – Tree based on 16S/18S rRNA gene



Biotechnologies

Amazing biotechnologies for manipulating DNA molecules have been developed, and are used as building blocks for even more powerful technologies.

These technologies are as amazing as the silicon etching/masking of VLSI fabrication.

DNA synthesis machines enable one to grow short DNA molecules of a specified sequence.

The *Polymerase chain reaction (PCR)* enables one to make large number of copies of a particular DNA sequence anywhere in solution given only the starting and ending sequences (primers).

Biotechnologies

PCR is one foundation of *DNA fingerprinting*, by turning a single molecule into billions.

Electrophoresis enables one to approximately measure the length of a DNA molecule, by measuring the time it takes to walk up an electrically charged Gel.

Since certain regions of the human genome have varying numbers of repeated characters, measuring their length by electrophoresis yields one method of DNA fingerprinting / identification.

DNA sequencing machines are built from both these technologies, and will be discussed when we talk about assembly.