An Introduction to Bioinformatics

Ho Tu Bao
Japan Advanced Institute of Science and Technology (JAIST)

"The two technologies that will shape the next century are biotechnology and information technology"

Bill Gates

"The two technologies that will have the greatest impact on each other in the new millennium are biotechnology and information technology"

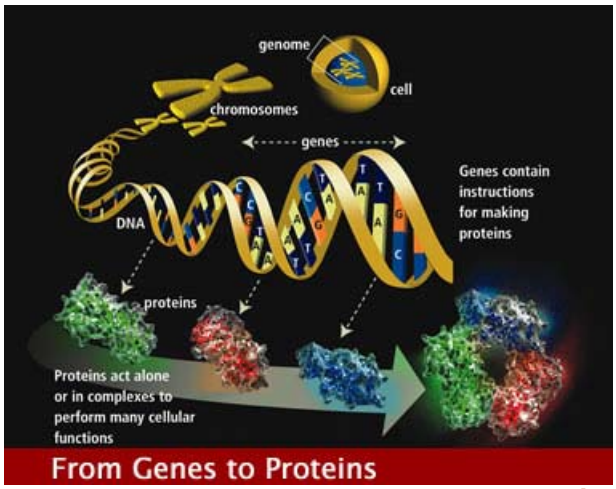
Martina McGloughlin

Outline

- **Elements of biology**
  (http://www.ebi.ac.uk/microarray/biology_intro.html#Genomes)
  - Molecules of life
  - Genes and genome
- What is bioinformatics?
- About some problems in bioinformatics

Basic molecular biology

- Most of 100 billion **cells** in the human body contains a copy of the entire human **genome** (all the genetic information necessary to build a human being).
- The cell nucleus contains six feet of **DNA** packed into 23 pairs of **chromosomes**. We each inherit one set of 23 chromosomes from our mother and another set from our father. DNA contains the code for the body **(genes)** governing all aspects of cell growth and inheritance.
- **Protein**, made up amino acids, are essential components of all organs and chemical activities.

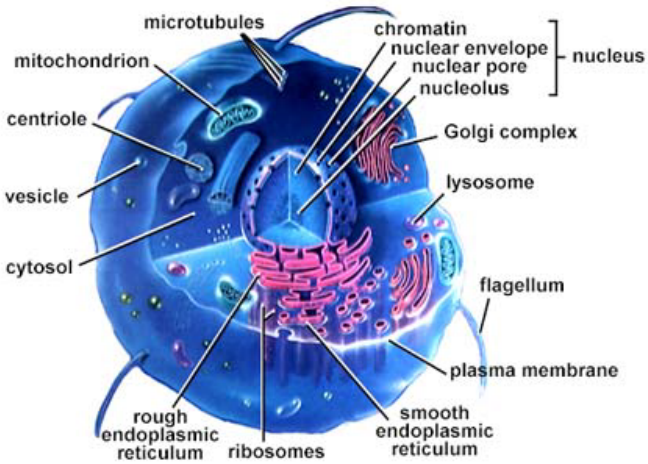


Organisms and cells (1/2)

- All organisms consist of small **cells**. Each cell is a complex system consisting of many different building blocks enclosed in **membrane**.
- There are estimated about $6x10^{13}$ cells in a human body, of about 320 different types, e.g., skin cells, muscle cells, brain cells (neurons), etc. The cell sizes may vary, e.g., a human red blood cell is about 5 microns (0.005 mm) in diameter, while some neurons are about 1 m long.
- Two types of organisms and two types of cells respectively, resulted by different evolutionary paths.
  - **Eukaryotes** (grass, flowers, weeds, worms, flies, mice, cats, dogs, humans, mushrooms and yeast, etc.)
  - **Prokaryotes** (bacteria)

Organisms and cells (2/2)

- A eukaryotic cell has a **nucleus**, which is separated from the rest of the cell by a membrane.
- An essential feature of most living cells is their ability to grow in an appropriate environment and to undergo cell division.
- Cell division and differentiation need to be controlled. **Cancerous cells** grow without control and can go on to form tumours.



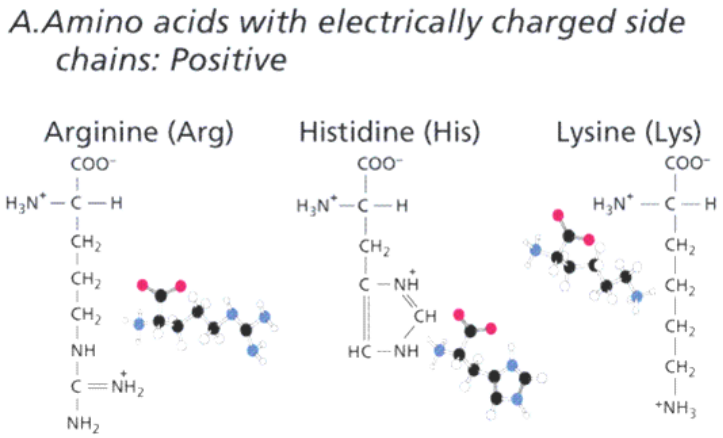
Molecules of life

1. Small molecules
2. Proteins
3. DNA
4. RNA

} Biological macromolecules

Small molecules

- Can be the building blocks of the macromolecules or they can have independent roles, e.g., sugars, fatty acids, amino acids and nucleotides.
- There are 20 different **amino acid molecules**, which are the building blocks for proteins, each is denoted by a letter in Latin alphabet.



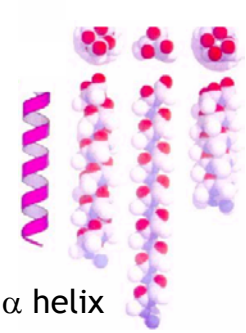
Proteins

Protein is a molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene coding for the protein. Proteins are required for the structure, functions and regulation of cells, tissues and organs, each protein having a specific role. Examples of proteins are:

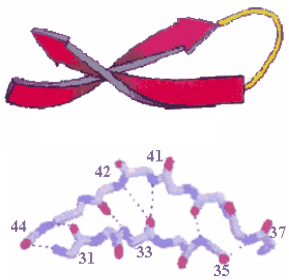
- **Structural proteins**, which can be thought of as the organism's basic building blocks.
- **Enzymes**, which perform (catalyse) a multitude of biochemical reactions. Together these reactions and the pathways they make up is called **metabolism.**
- **Transmembrane** proteins are key in maintenance of the cellular environment, regulating cell volume, etc.
- **Hormones, antibodies, etc.**

Protein structures

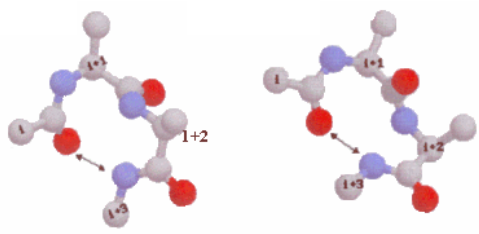
- **Primary structure**: Proteins are chains of 20 different types of amino acids, which in principle can be joined together in any linear order (poly-peptide chains). The length of the protein molecule can vary from few to many thousands of amino-acids.
- **Secondary structure**: Although the primary structure of a protein is linear, the molecule is not straight, and the sequence of the amino acids affects the folding. There are two common substructures often seen within folded chains: **alpha-helices** and **beta-strands**. They are typically joined by less regular structures **(loops).**



α helix



β strand



Two kinds of loops

Protein structures

- **Tertiary structure**: Because of folding, parts of a protein molecule chain come into contact with each other and various attractive or repulsive forces (hydrogen bonds, disulfide bridges, etc.) between such parts cause the molecule to adopt a fixed relatively stable 3D structure.
- **Quaternary structure**: A protein may be formed from more than one chain of amino-acids, in which case it is said to have *quaternary structure*. For example haemoglobin, is made up of four chains each of which is capable of binding an iron molecule.

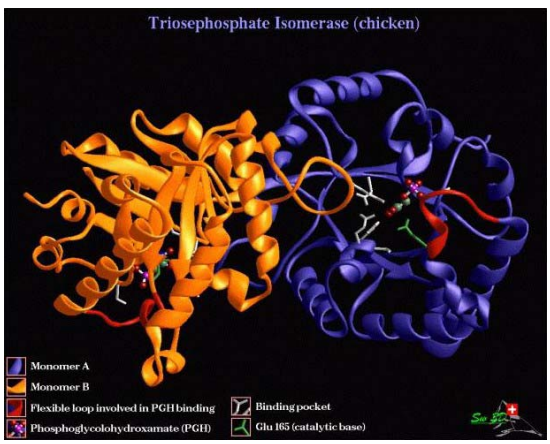
Helix-strand-helix

Tertiary structure

Quaternary structure

Protein structures

The images below shows the structure of triosephosphate isomerase visualised by RasMol software package, a 3D viewer for MSD structures



a characteristic protein size varies from about 3 to 10 nanometers (nm), i.e., 3 to $10x10^{-9}$ m, and solving (i.e., discovering) their structure is a difficult and expensive exercise (approximately €50,000 - €200,000 per novel structure)

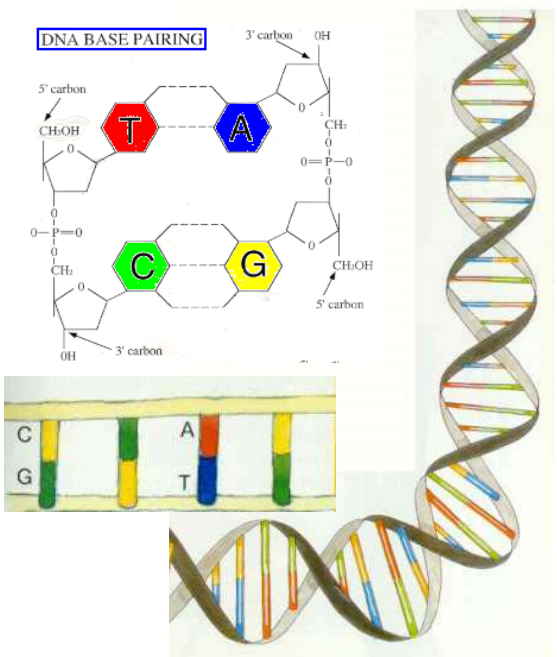
DNA (Deoxyribonucleic acid)

- DNA is the main information carrier molecule in a cell. DNA may be **single** or **double** stranded.
- A **single stranded DNA** molecule, also called a **polynucleotide**, is a chain of small molecules, called **nucleotides**.
- Four different nucleotides grouped into two types, purines: adenosine (**A**) and guanine (**G**) and pyrimidines: cytosine (**C**) and thymine (**T**), referred to as **bases**.
- Different nucleotides can be linked together in any order to form a polynucleotide, e.g.,

A-G-T-C-C-A-A-G-C-T-T

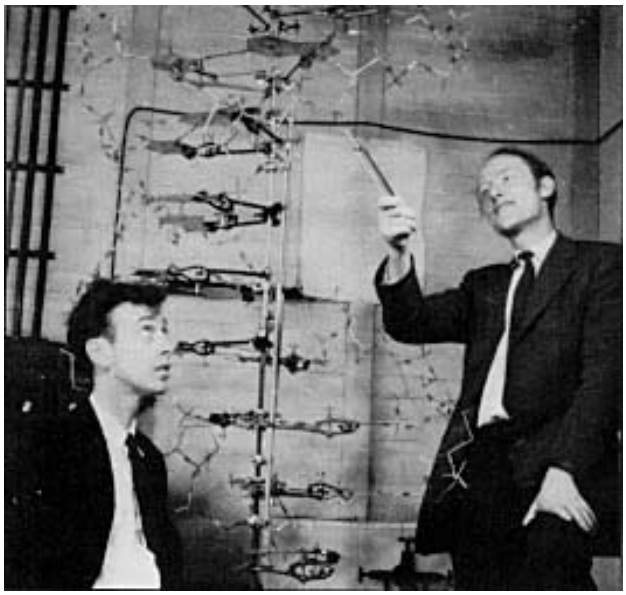
DNA (Deoxyribonucleic acid)

- Specific pairs of nucleotides can form weak bonds (liên kết) between them: **A binds to T, C binds to G.** The A-T and G-C pairs are called **base-pairs** (bp)
- When two longer complementary polynucleotide chains meet, they tend to stick together, known as a the **DNA double helix.**
- Two such strands are termed **complementary**, if one can be obtained from the other by mutually exchanging A with T and C with G, and changing the direction of the molecule to the opposite.

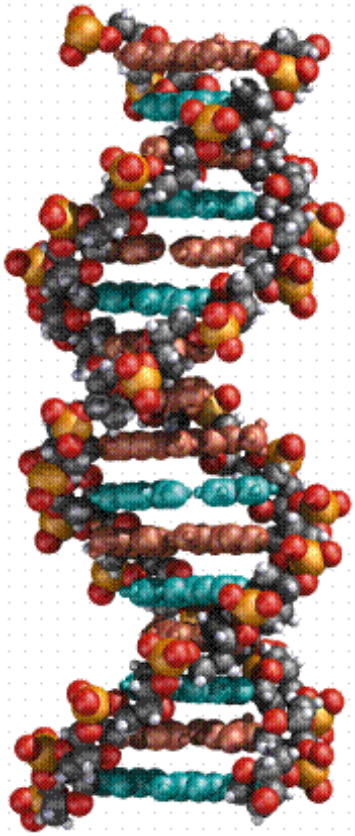


T-T-G-A-C-T-A-T-C-C-A-G-A-T-C
A-A-C-T-G-A-T-A-G-G-T-C-T-A-G

DNA



This structure was first figured out in 1953 in Cambridge by Watson and Crick



RNA (ribonucleic acid)

- **RNA** like DNA is constructed from nucleotides. But instead of the **T** (pyrimidine thymine), it has an alternative **U** (uracil), which is not found in DNA (only single strands).
- RNA has various functions in a cell, e.g., **mRNA** and **tRNA** are functionally different types of RNA which are both required for protein synthesis.
- RNA can bind complementary to a single strand of a DNA molecule, even though T is replaced by U, so molecules like this play an important role in life processes and in biotechnology

```
C-G-A-T-T-G-C-A-A-C-G-A-T-G-C   DNA
| | | | | | | | | | | | | | |
G-C-U-A-A-C-G-U-U-G-C-U-A-C-G   RNA
```

Genes and genomes

1. **Chromosomes, genomes and sequencing**
2. **Genes and protein synthesis**
3. **Gene prediction**
4. **Genome similarity and SNPs**

Chromosomes, genomes and sequencing

- **Chromosome**: one or several long double stranded DNA molecules organised.
- A human has 23 pairs of chromosomes.
- Chromasomal and mitochondrial DNA forms the **genome** of the organism. All organisms have genomes and they are believed to **encode almost all the hereditary information** of the organism.
- All cells in an organism contain **identical genomes** (with few rather special exceptions), as the result of DNA replication at each cell division.

Chromosomes, genomes and sequencing

- Determining the four letter sequence for a given a DNA molecule is known as the **DNA sequencing.**
  - Full genome for a bacterium was sequenced in 1995. The yeast genome was sequenced in 1997, worm in 1999, fly in 2000, and weed at 2001.
  - All of the human genome was completed in 2003.
- Genomes contain genes, most of which encode proteins.

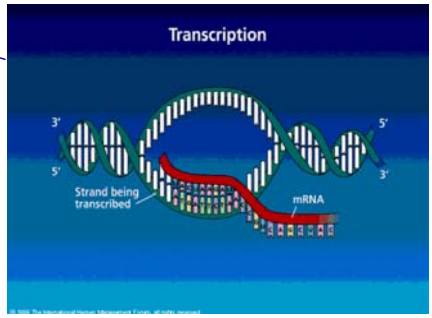
Genes and protein synthesis

- **Genes** are specific segments of DNA that control cell structure and function; the functional units of inheritance

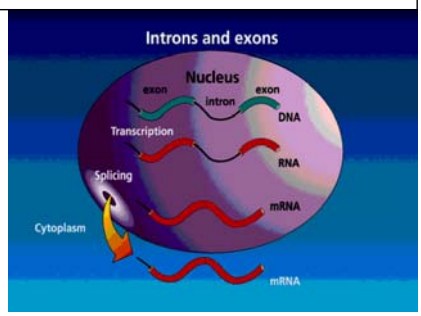
  (A gene is a **unit of inheritance**; a working subunit of DNA)
- To better understand it we need to describe the molecular machinery making proteins based on the information encoded in genes. This process is called **protein synthesis** and has three essential stages:
  - transcription
  - splicing
  - translation

Protein synthesis

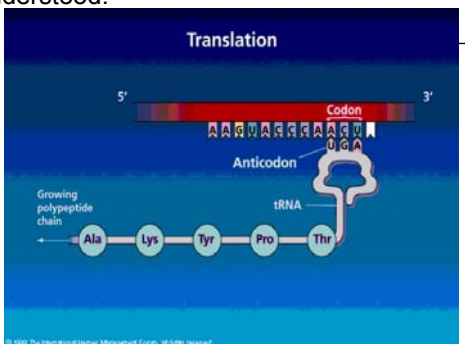
one strand of DNA molecule is copied into a complementary mRNA



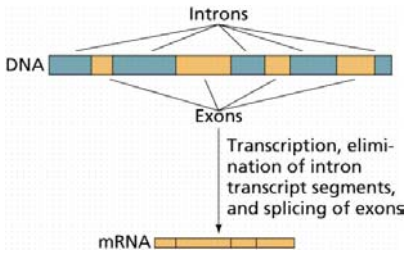
removes some stretches of the pre mRNA, called **introns**, the remaining sections called **exons** are then joined together. The number and size of introns and exons differs considerably between genes and also between species.



The **translation is a complex process** and not all the details are understood.



making proteins by joining together amino acids in order encoded in the mRNA. The order of the amino acids is determined by 3 adjacent nucleotides in the DNA. This is known as the **triplet** or **genetic code**. Each triplet is called a **codon** and codes for one amino acid.



Gene prediction problem

- **Gene prediction**: It is an interesting question: given the genomic DNA sequence, can we tell where the genes are?

Organism	The number of predicted genes	Part of the genome that encodes proteins (exons)
E.Coli (bacteria)	5000	90%
Yeast	6000	70%
Worm	18,000	27%
Fly	14,000	20%
Weed	25,500	20%
Human	30,000	< 5%

Genome similarity and SNPs

- All human genomes are deemed to be roughly **99.9% equivalent** and on average one in a thousand nucleotides are different in genomes of two different individuals.
- Variations in non-coding parts of the genome are analysed to produce patterns that can reliably distinguish individuals
- Particularly important variations in individual genomes are the **single nucleotide polymorphisms (SNP),** which can occur both in coding and non-coding parts of the genome. SNPs are DNA sequence variations which occur when a single base (A,C,G, or T) is altered so that different individuals may have different letters in these positions.

Functional genomics

- **Gene functions**
- **Protein abundance in a cell**
- **Gene regulation and networks**

Functional genomics can be roughly defined as using the emerging knowledge about genomes to understand the gene and their product functions and interactions, and most importantly of all, how all this makes organisms to function the way they do.

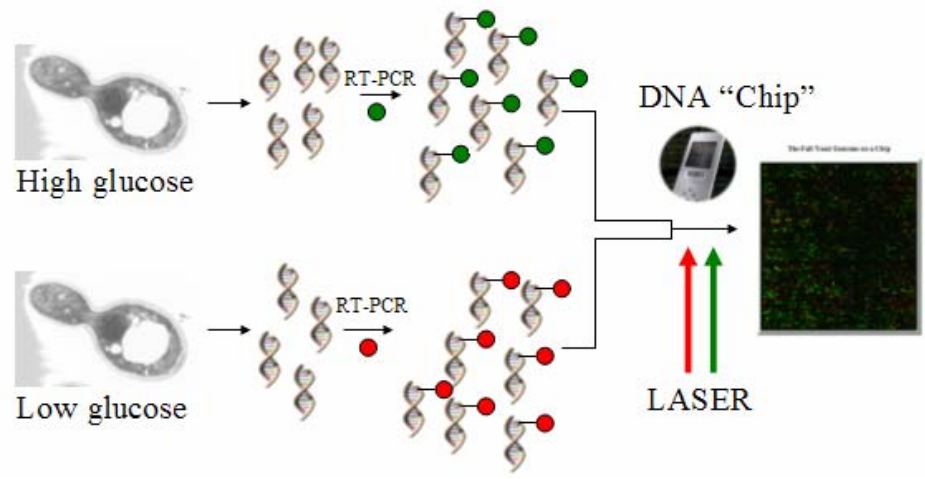
Functional genomics

- There is likely to be a **limited universe of genes** and their respective proteins, from the functional point of view, many of which are present in most or all genomes.
- The **protein abundance** may depend on many factors such as whether the respective gene is **expressed** (i.e., is actively transcribed) or not, how intensively (how fast) it is expressed, whether and how fast it is spliced, translated and modified, etc.
- Another important and interesting question in biology is how **gene expression** is switched on and off, i.e., how genes are regulated

(Gene expression = the process by which a gene's coded information is translated into the structures present and operating in the cell (either proteins or RNAs)).

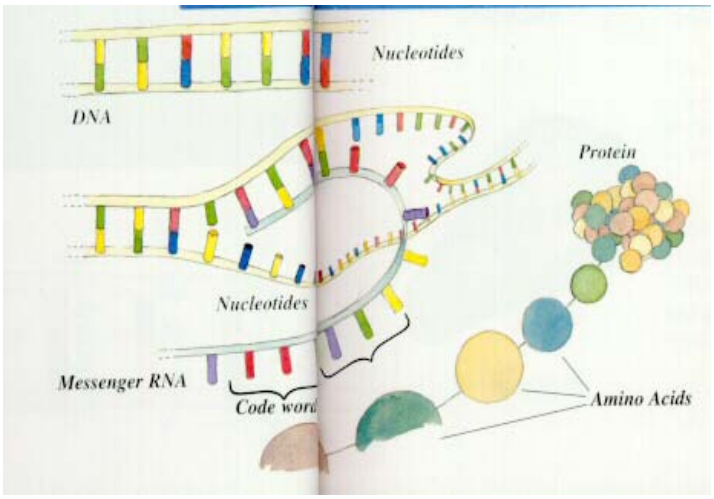
Microarrays and gene expression databases

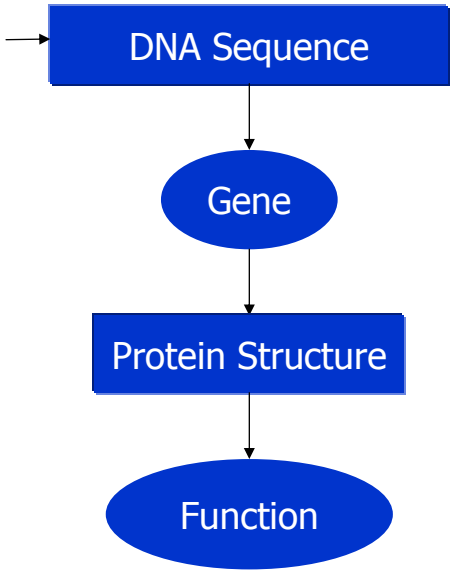
- **Microarray** technology makes use of the sequence resources created by the genome projects and other sequencing efforts to answer the question, what genes are expressed in a particular cell type of an organism, at a particular time, under particular conditions.



Molecular Biology: Flow of Information

DNA → RNA → Protein → Function





Overall Goals

Outline

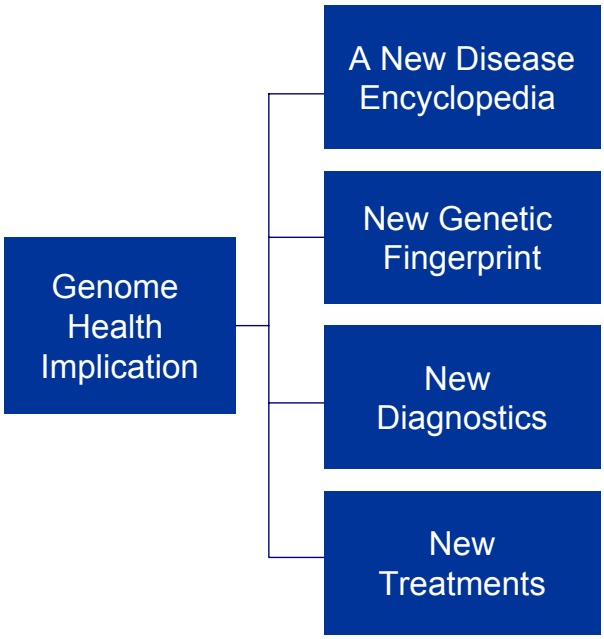
- Elements of biology
- **What is bioinformatics?**
- About some problems in bioinformatics

- Bioinformatics: the machine learning approach, Pierre Baldi, Soren Brunak, MIT Press 2001
- Bioinformatics basics: applications in biological sciences and medicine, Hooman H. Rashidi and Lukas K. Buehler, CRC Press, 2002

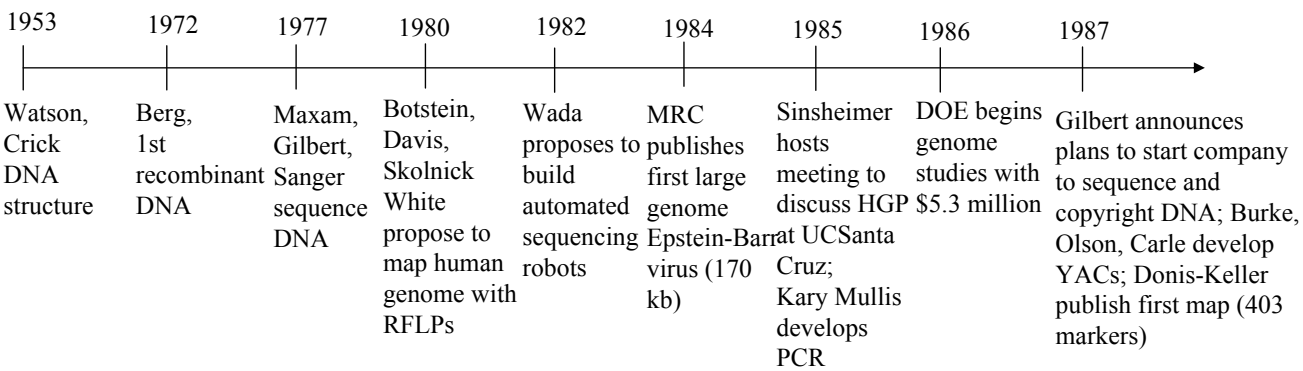
Human Genome Project

Goal (15 years since 1990)

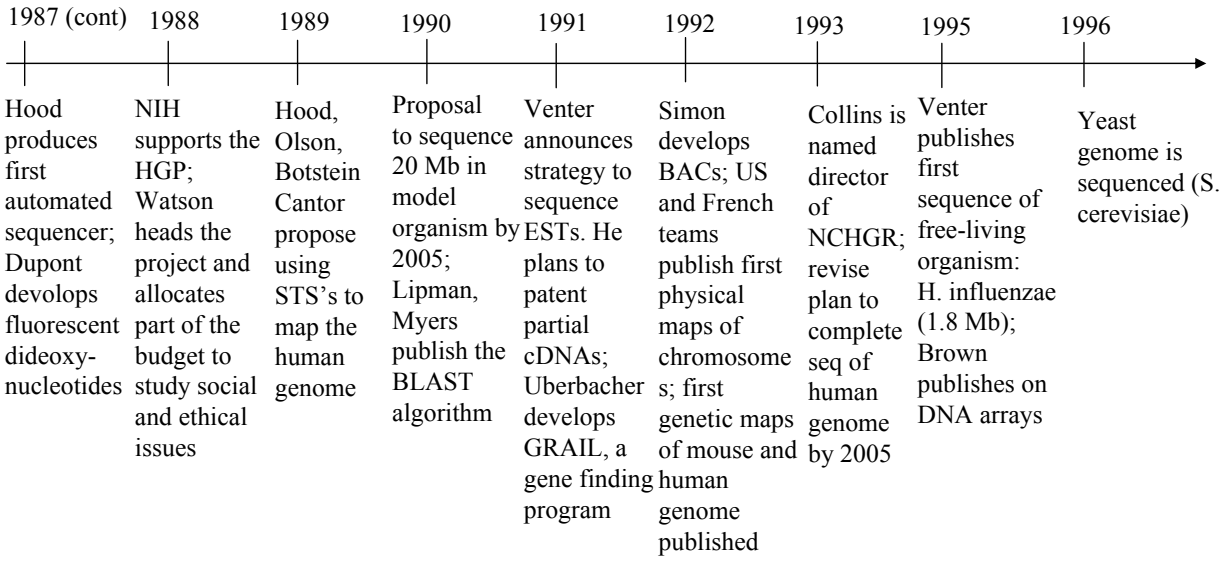
- **identify** all the approximately 30,000 genes in human DNA,
- **determine** the sequences of the 3 billion chemical base pairs that make up human DNA,
- **store** this information in databases,
- **improve** tools for data analysis,
- **transfer** related technologies to the private sector, and
- **address** the ethical, legal, and social issues (ELSI) that may arise from the project.



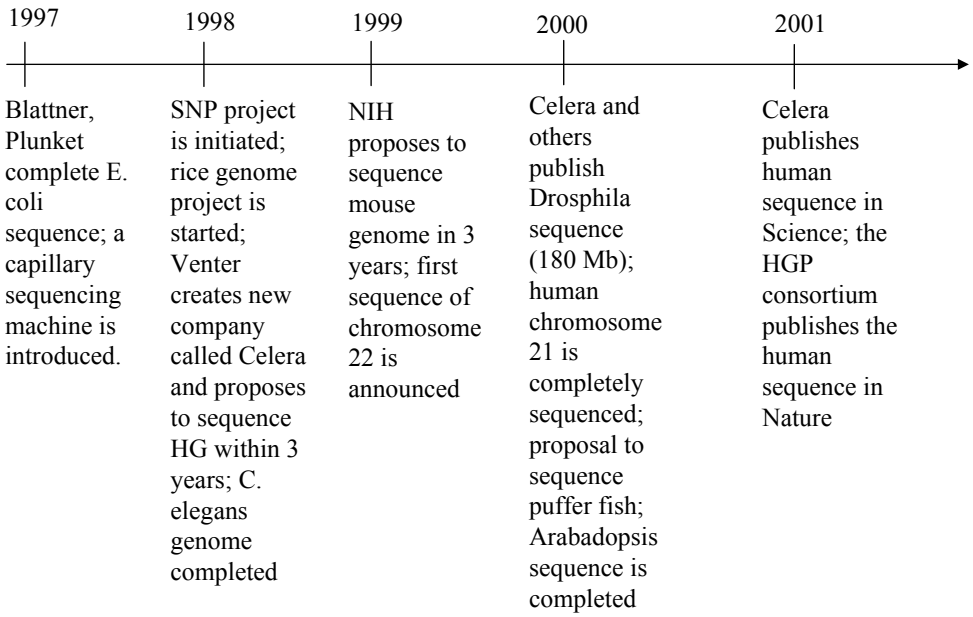
History of the Human Genome Project



History of the Human Genome Project (continued)



History of the Human Genome Project (continued)



http://www.d-trends.com/Bioinformatics/timeline.html

What is bioinformatics?

- **Bio**: Molecular Biology
- **Informatics**: Computer Science
- **Bioinformatics**: Solving problems arising from biology using methodology from computer science.

  Synonyms: Computational biology, Computational molecular biology, Biocomputing

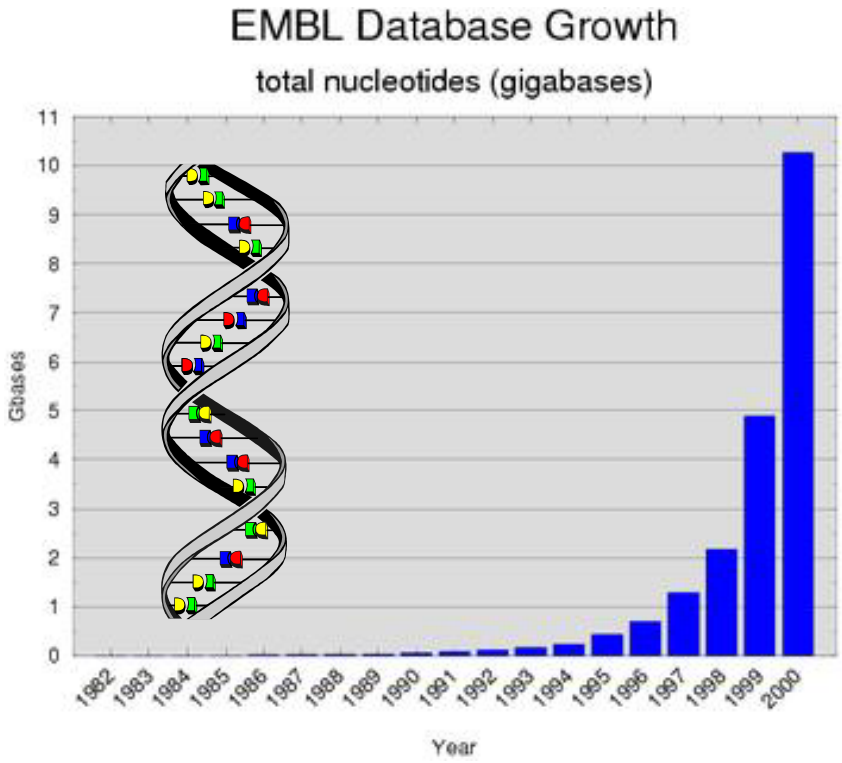
Paradigm Shift in Biology

The new paradigm, now emerging, is that all the 'genes' will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a **theoretical conjecture**, only **then turning to experiment** to follow or test that hypothesis.

To use [the] flood of knowledge, which will pour across the computer networks of the world, biologists not only must become **computer literate**, but also **change their approach** to the problem of understanding life.

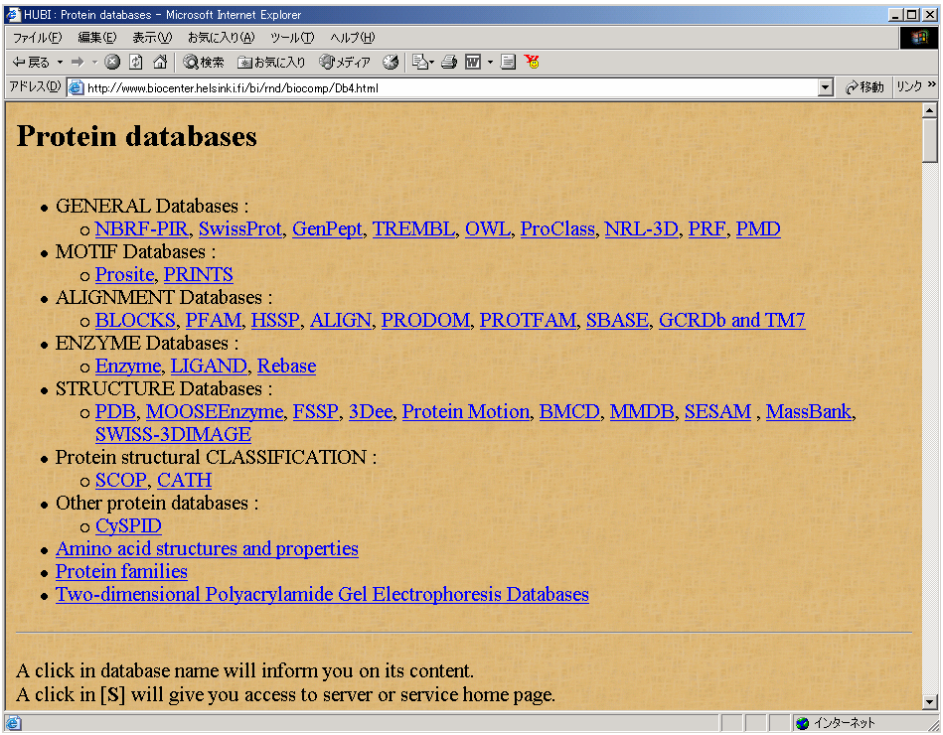
Walter Gilbert. 1991. Towards a paradigm shift in biology. *Nature*, 349:99.

Base Pairs in GenBank



10,267,507,282 bases in 9,092,760 records.

Public databases



Extension of Bioinformatics Concept

- **Genomics**
  - Functional genomics
  - Structural genomics
- **Proteomics:** large scale analysis of the proteins of an organism
- **Pharmacogenomics:** developing new drugs that will target a particular disease
- **Microarray (genome chip):** DNA chip, protein chip

The identification and functional characterization of genes.

The study of gene expression at the protein level, by the identification and characterization of proteins present in a biological sample.

The use of genetic information to predict the safety, toxicity and/or efficacy of drugs in individual patients or groups of patients.

a new technology aims to monitor the whole genome on a single chip so that researchers can have a better picture of the interactions among thousands of genes simultaneously

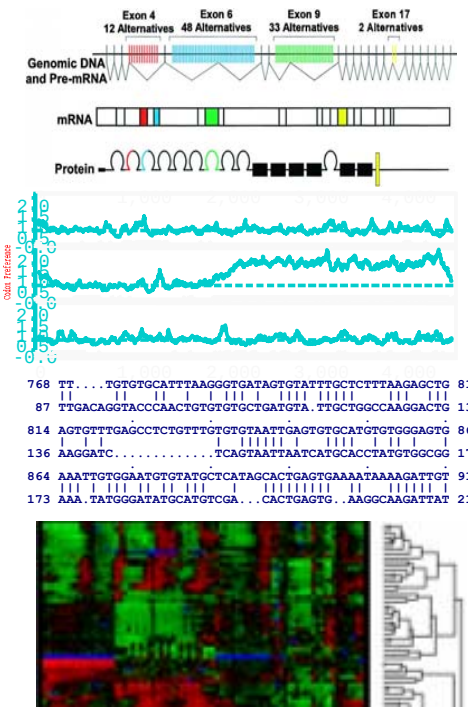
Problems in Bioinformatics

Structure analysis
- Protein structure comparison
- Protein structure prediction
- RNA structure modeling

Pathway analysis
- Metabolic pathway
- Regulatory networks

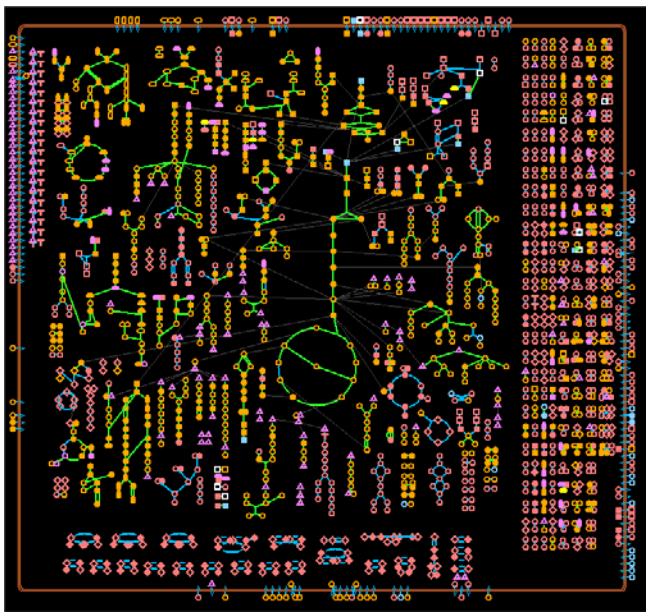
Sequence analysis
- Sequence alignment
- Structure and function prediction
- Gene finding

Expression analysis
- Gene expression analysis
- Gene clustering

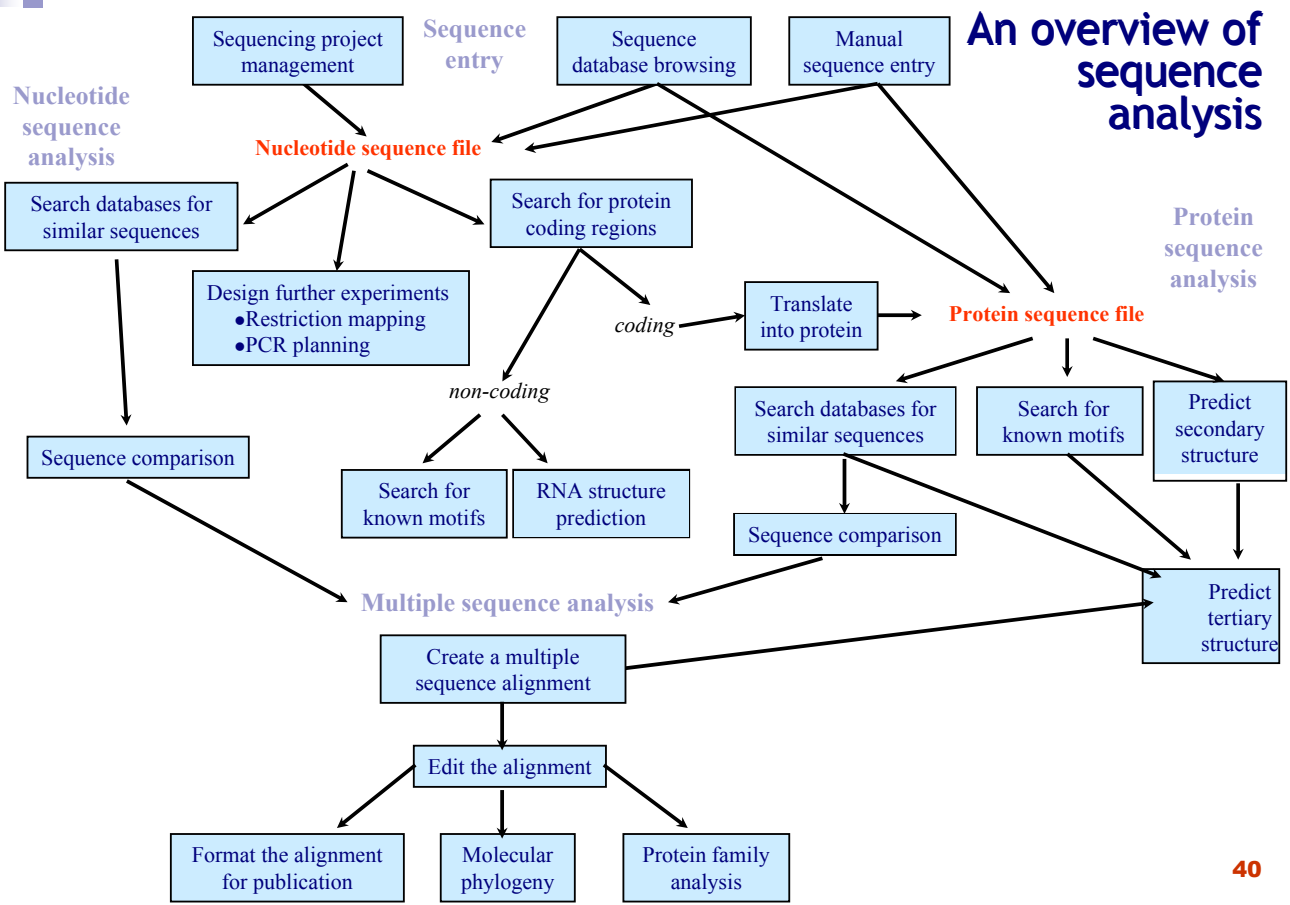


Pathway analysis

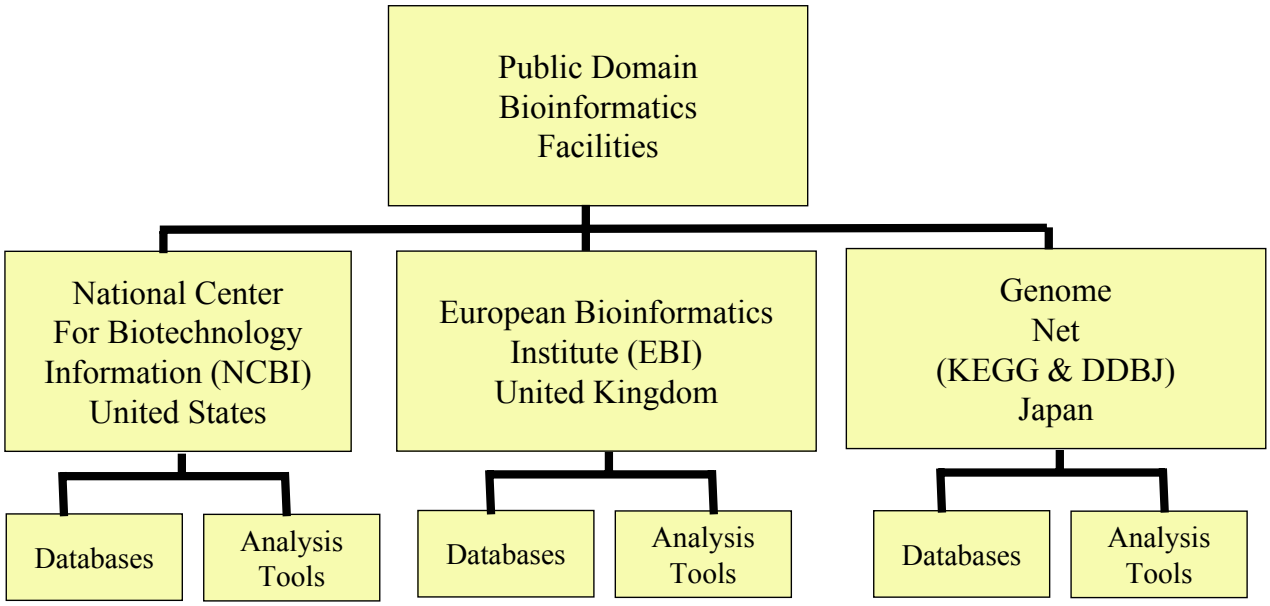
- A chemical **reaction** interconverts chemical compounds
- An **enzyme** is a protein that accelerates chemical reactions
- A **pathway** is a linked set of reactions



An overview of sequence analysis

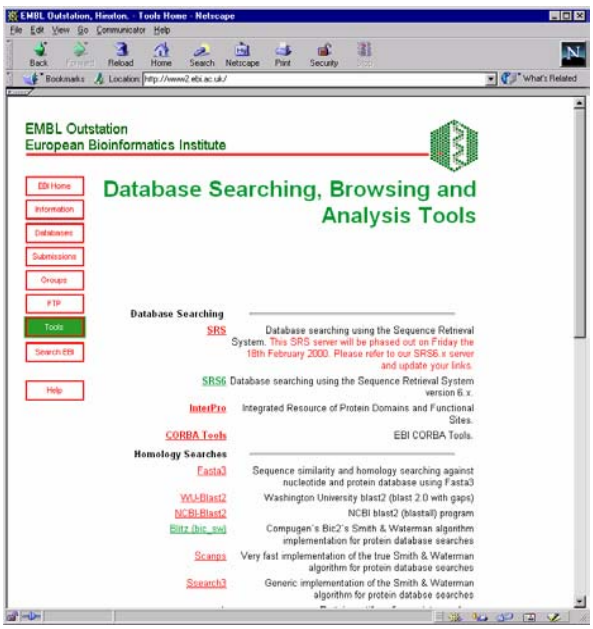


Primary public domain bioinformatics servers



Analysis Tools

The EBI maintains versions of major public domain sequence database searching and analysis tools, e.g. FASTA, CLUSTALW, BLAST, and Smith & Waterman implementations.



Challenges in Bioinformatics

- **Bioinformatics requires:**
  - Access to multiple distributed resources
  - Needs information to be up-to-date
  - Minimal data redundancy
  - Robust applications
  - Extendable applications
    - Monolithic App. vs. Components
  - Portable software

Challenges in Bioinformatics

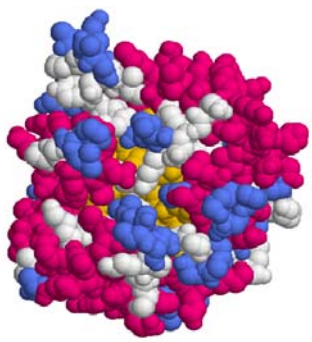
- **Explosion of information**
  - Need for faster, automated analysis to process large amounts of data
  - Need for integration between different types of information (sequences, literature, annotations, protein levels, RNA levels etc…)
  - Need for "smarter" software to identify interesting relationships in very large data sets
- **Lack of "bioinformaticians"**
  - Software needs to be easier to access, use and understand
  - Biologists need to learn about the software, its limitations, and how to interpret its results

Outline

- Some basic concepts
- What is bioinformatics?
- **About some problems in bioinformatics**

Protein prediction task

- There are roughly 15,000 protein structures deposited in public databases, though many of them are very similar to each other. There are about 1,500 different representative protein **structures known.**
- **Predicting protein structure** from the amino-acid sequence is one of the most important problems of computational biology (bioinformatics) and is far from being solved.



String matching

(Approximate) String Matching

Input: Text T , Pattern P

Question(s):

- Does P occur in T?
- Find one occurrence of P in T.
- Find all occurrences of P in T.
- Count # of occurrences of P in T.
- Find longest substring of P in T.
- Find closest substring of P in T.
- Locate direct repeats of P in T.
- *Many More variants*

Applications:

- Is P already in the database T?
- Locate P in T.
- Can P be used as a primer for T?
- Is P homologous to anything in T?
- Has P been contaminated by T?
- Is *prefix*(P) = *suffix*(T)?
- Locate tandem repeats of P in T.

String matching

Input:	Text **T**; Pattern **P**
Output:	All occurrences of **P** in **T**.

Sliding Window Strategy:

Initialize window on T;

While (window within T) do

Scan: if (window = P) then report it;

Shift: shift window to right (by one position)

endwhile;

String matching

Pairwise Sequence Alignment

- **Input**
  - Two sequences of letters
  - A scoring scheme
- **Output**
  - Optimal alignment

- ❖ Most fundamental bioinformatic problem
- ❖ Aligned sequences ⇒ same structure/function
- ❖ Yield insight if the structure/function of one of the aligned sequences is known

ATTGCGC → ATTGCGC

ATCCGC → ATCCGC

ATTGCGC
→ AT-CCGC

ATTGCGC
→ ATC-CGC

ATTGCGC
→ ATCCG-C

HMM in sequence alignment

- The states of HMM will be divided into *match* states, *insert* states and *delete* states.
- The alphabet M consists of twenty amino acids together with one dummy symbol δ representing "delete". *Delete* states output δ only.
- Each insert and match state has its own distribution over the 20 amino acids, and does not emit the symbol δ.
- The sequences to be aligned are used as the training data, to train the parameters of the model.
- For each sequence, the Viterbi algorithm is then used to determine a path most likely to have produced that sequence.

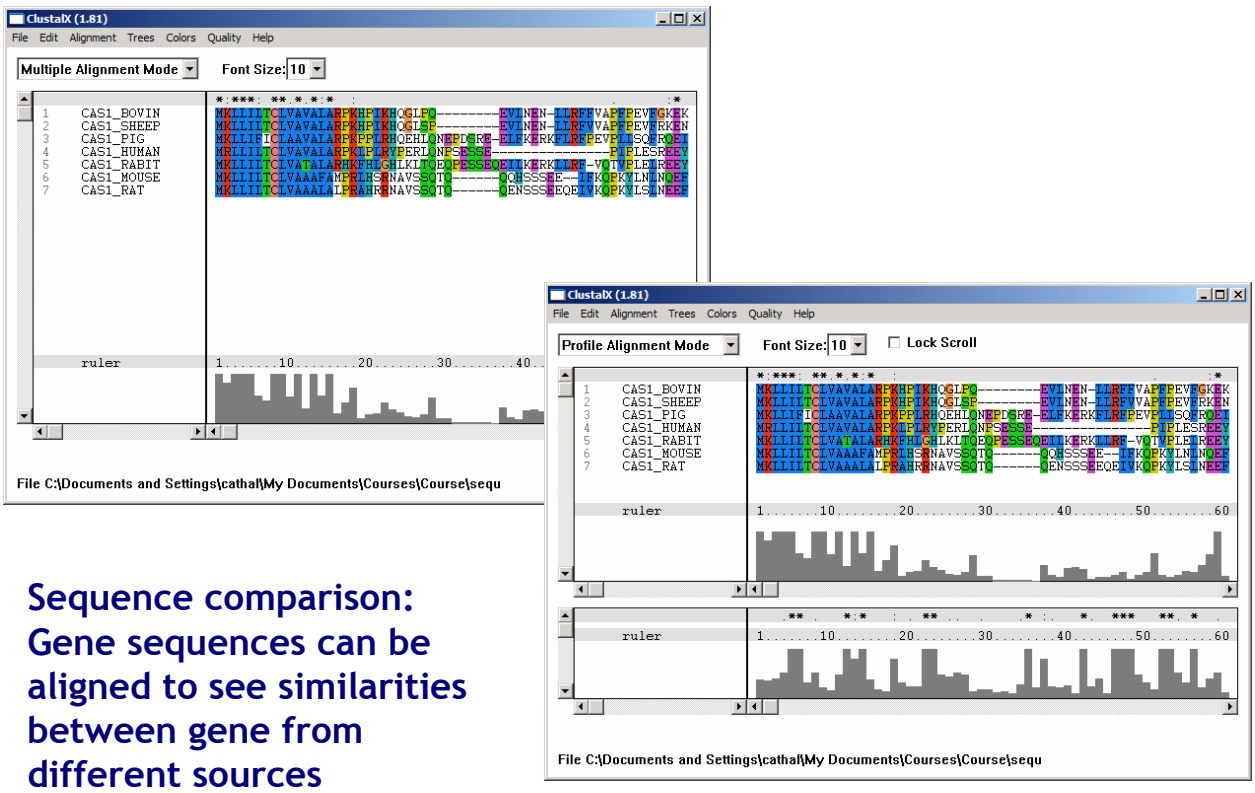
HMM in sequence alignment

- Consider the sequences
  - CAEFDDH
  - CDAEFPDDH
- Suppose the model has length 10 and their most likely paths through the model are
  - $m_0m_1m_2m_3m_4d_5d_6m_7m_8m_9m_{10}$
  - $m_0m_1i_1m_2m_3m_4d_5m_6m_7m_8m_9m_{10}$
- The alignment induced is found by aligning positions that were generated by the same match state. This leads to the alignment
  - C–AEF –DDH
  - CDAEFPDDH

Pairwise vs Multiple Sequences

- Pairs of sequences typically aligned using exhaustive algorithms (dynamic programming)
  - complexity of exhaustive methods is $O(2^n m^n)$
    n = number of sequences
- Multiple sequence alignment using heuristic methods

```
#Rat        ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGT
#Mouse      ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGT
#Rabbit     ATGGTGCATCTGTCCAGT---GAGGAGAAGTCTGC
#Human      ATGGTGCACCTGACTCCT---GAGGAGAAGTCTGC
#Oppossum   ATGGTGCACTTGACTTTT---GAGGAGAAGAACTG
#Chicken    ATGGTGCACTGGACTGCT---GAGGAGAAGCAGCT
#Frog       ---ATGGGTTTGACAGCACATGATCGT---CAGCT
```

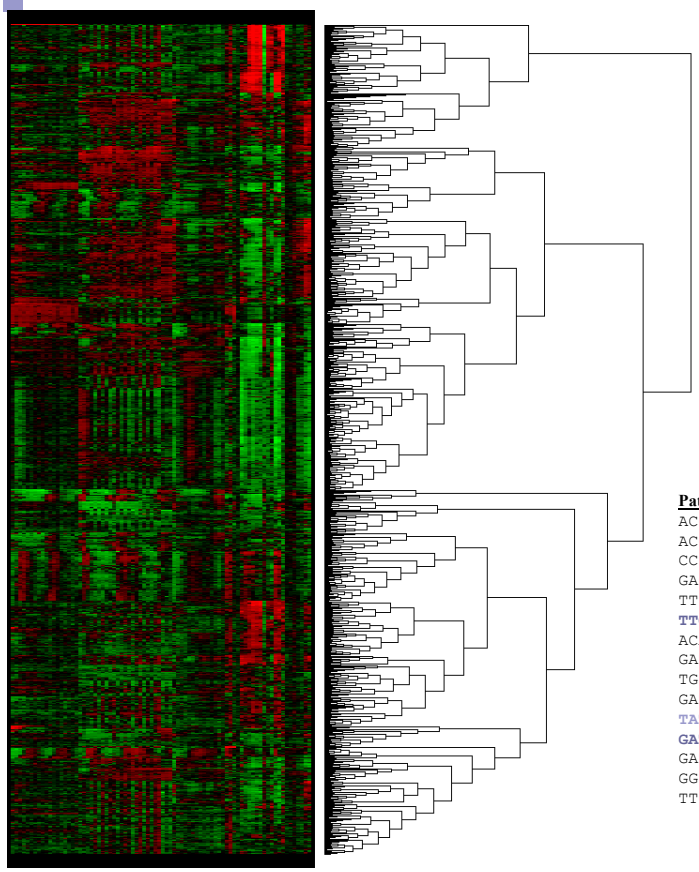



Sequence comparison: Gene sequences can be aligned to see similarities between gene from different sources

Gene Prediction

Gene prediction is an important problem for computational biology and there are various algorithms that do gene prediction using known genes as a training data set. A popular algorithmic technique used in gene prediction are hidden Markov models (HMMs).

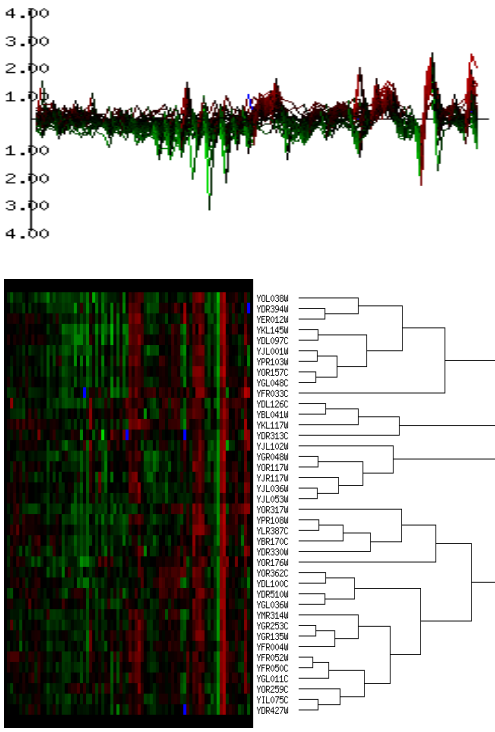
(given the genomic DNA sequence, can we tell where the genes are?)



Gene clustering and some discovered patterns

Pattern	Probability	Cluster	No.	Total
ACGCG	6.41E-39	96	75	1088
ACGCGT	5.23E-38	94	52	387
CCTCGACTAA	5.43E-38	27	18	23
GACGCG	7.89E-31	86	40	284
TTTCGAAACTTACAAAAAT	2.08E-29	26	14	18
TTCTTGTCAAAAAGC	2.08E-29	26	14	18
ACATACTATTGTTAAT	3.81E-28	22	13	18
GATGAGATG	5.60E-28	68	24	83
TGTTTATATTGATGGA	1.90E-27	24	13	18
GATGGATTTCTTGTCAAAA	5.04E-27	18	12	18
TATAAATAGAGC	1.51E-26	27	13	18
GATTTCTTGTCAAA	3.40E-26	20	12	18
GATGGATTTCTTG	3.40E-26	20	12	18
GGTGGCAA	4.18E-26	40	20	96
TTCTTGTCAAAAAGCA	5.10E-26	29	13	18

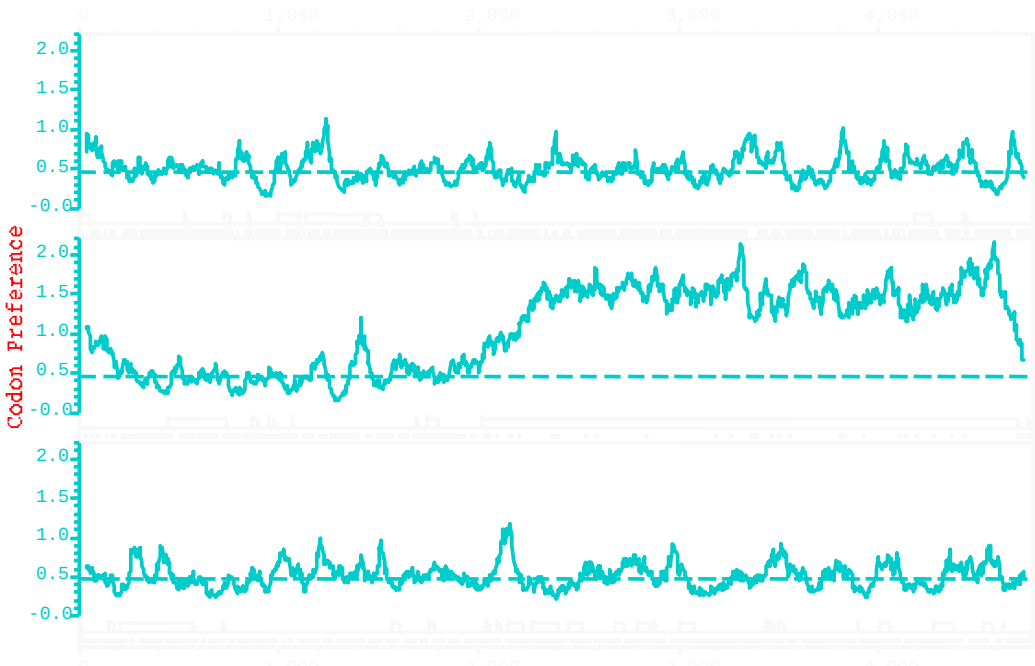
The "GGTGGCAA" Cluster



ORF	Gene	Description		
YBL041W	PRE7	20S proteasome subunit(beta6)		
YBR170C	NPL4	nuclear protein localization factor and ER translocation component		
YDL126C	CDC48	microsomal protein of CDC48/PAS1/SEC18 family of ATPases		
YDL100C		similarity to E.coli arsenical pump-driving ATPase		
YDL097C	RPN6	subunit of the regulatory particle of the proteasome		
YDR313C	PIB	phosphatidylinositol(3)-phosphate binding protein		
YDR330W		similarity to hypothetical S. pombe protein		
YDR394W	RPT3	26S proteasome regulatory subunit		
YDR427W	RPN9	subunit of the regulatory particle of the proteasome		
YDR510W	SMT3	ubiquitin-like protein		
YER012W	PRE1	20S proteasome subunit C11(beta4)		
YFR004W	RPN11	26S proteasome regulatory subunit		
YFR033C	QCR6	ubiquinol--cytochrome-c reductase 17K protein		
YFR050C	PRE4	20S proteasome subunit(beta7)		
YFR052W	RPN12	26S proteasome regulatory subunit		
YGL048C	RPT6	26S proteasome regulatory subunit		
YGL036W	MTC2	Mtf1 Two hybrid Clone 2		
YGL011C	SCL1	20S proteasome subunit YC7ALPHA/Y8 (alpha1)		
YGR048W	UFD1	ubiquitin fusion degradation protein		
YGR135W	PRE9	20S proteasome subunit Y13 (alpha3)		
YGR253C	PUP2	20S proteasome subunit(alpha5)		
YIL075C	RPN2	26S proteasome regulatory subunit		
YJL102W	MEF2	translation elongation factor, mitochondrial		
YJL053W	PEP8	vacuolar protein sorting/targeting protein		
YJL036W		weak similarity to Mvp1p		
YJL001W	PRE3	20S proteasome subunit (beta1)		
YJR117W	STE24	zinc metallo-protease		
YKL145W	RPT1	26S proteasome regulatory subunit		
YKL117W	SBA1	Hsp90 (Ninety) Associated Co-chaperone		
YLR387C		similarity to YBR267w		
YMR314W	PRE5	20S proteasome subunit(alpha6)		
YOL038W	PRE6	20S proteasome subunit (alpha4)		
YOR117W	RPT5	26S proteasome regulatory subunit		
YOR157C	PUP1	20S proteasome subunit (beta2)		
YOR176W	HEM15	ferrochelatase precursor		
YOR259C	RPT4	26S proteasome regulatory subunit		
YOR317W	FAA1	long-chain-fatty-acid--CoA ligase		
YOR362C	PRE10	20S proteasome subunit C1 (alpha7)		
YPR103W	PRE2	20S proteasome subunit (beta5)		
YPR108W	RPN7	subunit of the regulatory particle of the proteasome		

Gene discovery:

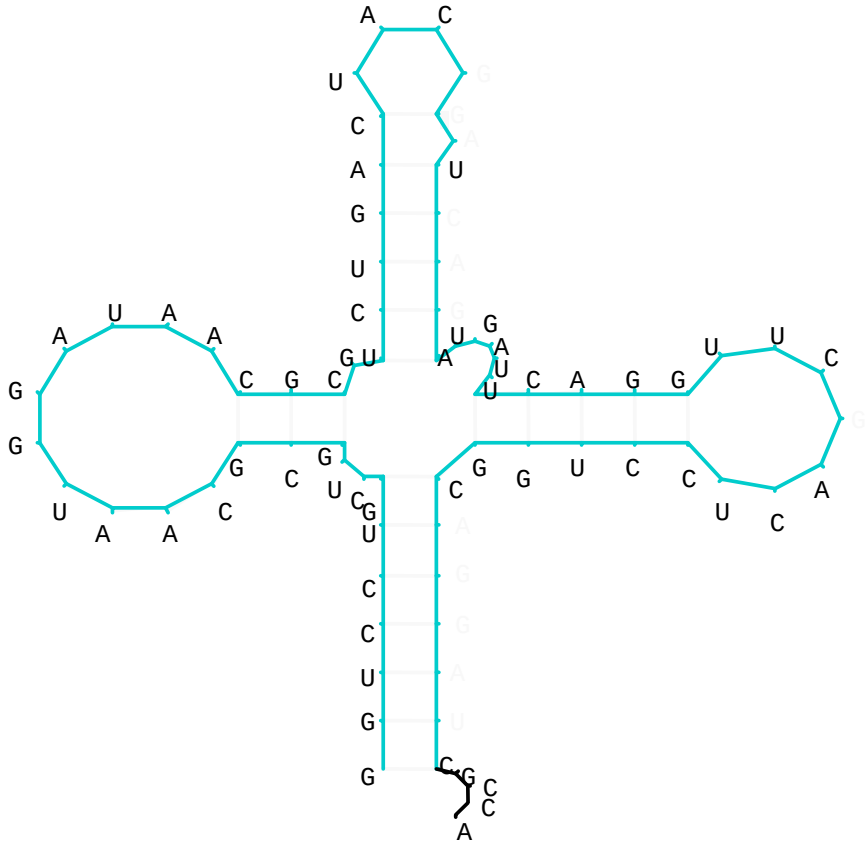
Computer program can be used to recognise the protein coding regions in DNA



Plot created using codon preference (GCG)

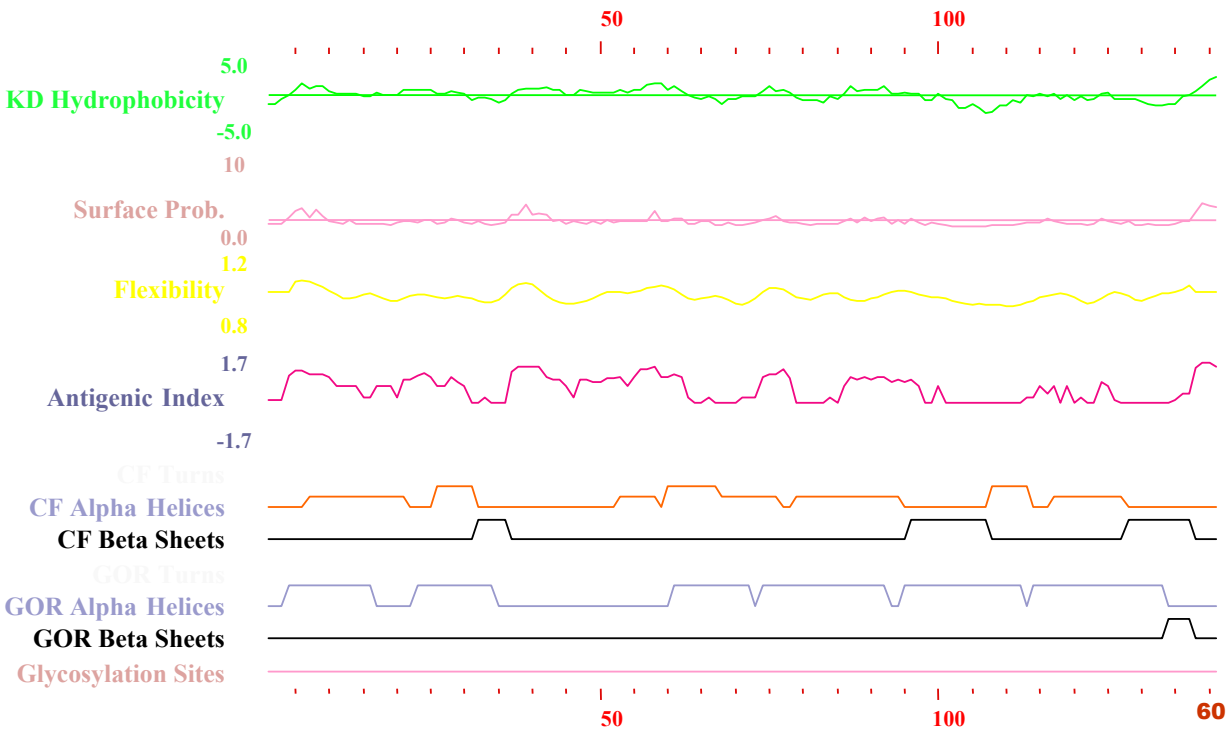
RNA structure prediction

Structural features of RNA can be predicted



Protein structure prediction:

Particular structural features can be recognised in protein sequences



Machine learning tools for bioinformatics

- Neural Networks
  - Sequence Encoding and Output Interpretation
  - Prediction of Protein Secondary Structure
  - Prediction of Signal Peptides and Their Cleavage Sites
  - Applications for DNA and RNA Nucleotide Sequences
- Hidden Markov Models
  - Protein Applications
  - DNA and RNA Applications
- Probabilistic Graph Models
- Probabilistic Models of Evolution
- Stochastic Grammars and Linguistics

(Bioinformatics: the machine learning approach, Pierre Baldi, Soren Brunak, MIT Press)

Summary

- Addressed the basic concepts in biology and bioinformatics, and key problems of bioinformatics
- Bioinformatics is an important field, and very challenging.
- Strongly related to data mining and machine learning.
- Which line of research could we follow?

Darwin: It's not the strongest, nor the most intelligent, but the species most adaptable to change has the best chance of survival.