# Giới thiệu về tin sinh học

Hồ Tú Bảo
Viện Công nghệ Thông tin, TTKHTN&CNQG
Viện Khoa học và Công nghệ Tiên tiến Nhật bản (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

- **Khái niệm cơ bản của sinh học**
  (http://www.ebi.ac.uk/microarray/biology_intro.html#Genomes)
  - Phân tử trong sự sống
  - Gene và gene học
- **Tin sinh học là gì?**
- **Về một vài bài toán trong tin sinh học**



# "Sống", Tạ Quang Bửu (1948)

"…Một đêm tháng 10 năm 1910, một tế bào haploid (cùng một gamète với 24 chromosome) của cha tôi gặp một tế bào (cùng một gamète với 24 chromosome) của mẹ tôi.

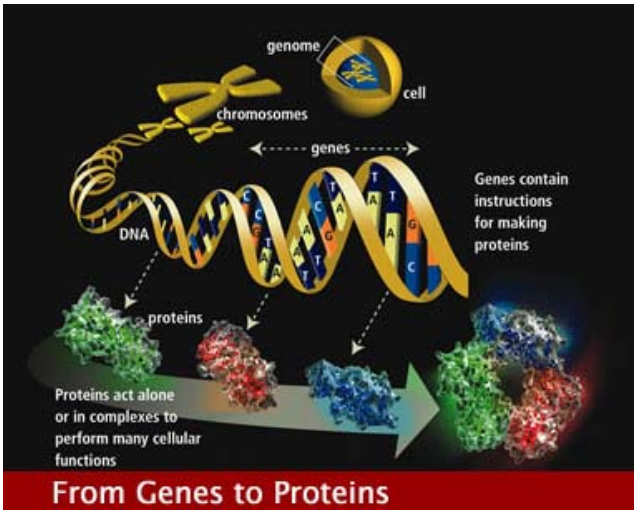
Hai tế bào ấy phối hợp với nhau thành một tế bào trứng với hai lần 24 chromosome. Tế bào này chẻ đôi sinh ra hai tế bào nữa, rồi hai sinh ra bốn, bốn sinh ra tám, v,v… thành một khối tế bào. Khối tế bào này là tôi.

Chín tháng sau tôi ra đời với những đặc điểm này: da đen, mắt hoe, chân ngắn như ông nội tôi; mồm rộng, vai ngang, tai nhỏ như bà ngoại tôi. Ngoài ra trong thân thể có chỗ thì giống ông ngoại, có chỗ giống bà nội tôi. Còn tính lười đặc biệt của tôi thì xem gia phả đến bậc ông cố nội ngoại cũng không thấy tông tích. Có lẽ phải lên xa nữa.

Ba năm sau, cũng theo một loạt biến cố như trên, em tôi ra đời. Em tôi thì mồm rộng, da trắng, mắt hoe, chân dài. Những đặc điểm của nó cũng là những đặc điểm của hai gia đình chúng tôi, nhưng phân phối lại cách khác."

# Basic genetics  Gene học cơ sở

- Phần lớn của 100 tỷ **tế bào** (cell) trong cơ thể con người có sự sao chép của toàn bộ **hệ gene** (human genome), là toàn bộ thông tin di truyền cần thiết để tạo ra cơ thể sống.
- Hạt nhân tế bào (cell nucleus) chứa **DNA** gói trong các cặp **nhiễm sắc thể** (chromosomes).
- DNA chứa **gene**, là mã của cơ thể và điều khiển mọi khía cạnh về phát triển và kế thừa của tế bào.
- **Protein**, tạo ra từ amino acids, là các thành phần thiết yếu của mọi cơ quan (organs) và hoạt động hóa học.



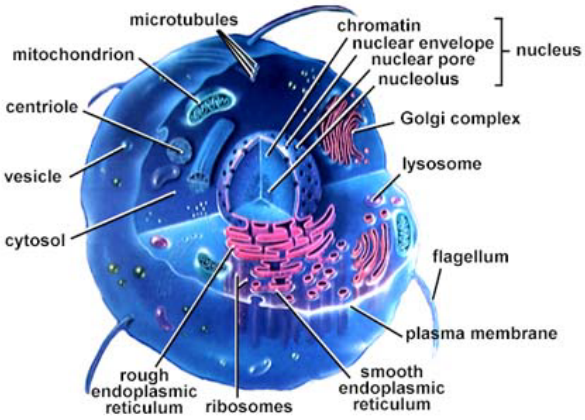



# Sinh vật và tế bào  (1/2)

- Mọi sinh vật đều gồm các **tế bào** **(cells)**. Mỗi tế bào là một hệ thống phức tạp gồm nhiều khối tạo dựng (building blocks) khác nhau bọc bởi các **màng** **(membrane)**.
- Có khoảng $6x10^{13}$ tế bào trong cơ thể người, với khoảng 320 kiểu khác nhau, như tế bào da, cơ bắp, não (neurons), etc. Tế bào có kích thước khác nhau: hồng cầu có đường kính chừng 0.005 mm còn neuron dài chừng 1 mét.
- Hai kiểu sinh vật và tương ứng hai kiểu tế bào, là kết quả của những con đường tiến hóa khác nhau.
  - **Nhân chuẩn** **(Eukaryotes):** cỏ, hoa, lúa mì, giun, ruồi, chuột, chó, mèo, người, nấm, men bia, etc.
  - **Nhân sơ** **(Prokaryotes):** bacteria

# Sinh vật và tế bào (2/2)

- Mỗi tế bào nhân chuẩn đều gồm một **nucleus (nhân),** được tách khỏi phần còn lại của tế bào bởi một màng ngăn.
- Một đặc tính cơ bản của mọi tế bào sống là khả năng phát triển (to grow) trong một môi trường thích hợp và trải qua sự phân chia tế bào (cell division).
- Sự phân chia tế bào và biệt lập tế bào cần được kiểm soát. Khi tế bào phát triển không được kiểm soát có thể tạo thành các u (tumours) và **ung thư**.






# Molecules of life
# Phân tử của sự sống

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

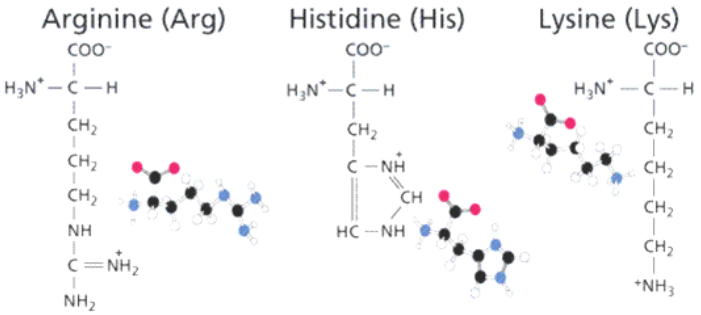
} Biological macromolecules

# Small molecules Tiểu phân tử

- Có thể có các vai trò độc lập hoặc có thể là các khối tạo dựng của các đại phân tử (macromolecules). Thí dụ như phân tử nước, đường, acids béo (fatty), amino acids và đơn phân tử (nucleotides).

- Có 20 loại **amino acids** khác nhau, là các khối tạo dựng của proteins, mỗi loại được ký hiệu bởi một chữ cái Latin.

*A. Amino acids with electrically charged side chains: Positive*





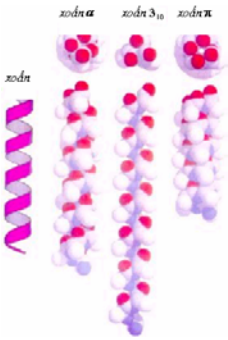

# Proteins

**Protein** là một đại phân tử tạo thành từ một hay nhiều dãy amono acids theo một thứ tự đặc biệt; thứ tự này được xác định bởi dãy cơ sở (bazơ) các nucleotides trong gene mã hóa cho protein. Các proteins cần thiết cho cấu trúc, chức năng và điều chỉnh tế bào, mô và tổ chức, mỗi protein có một vai trò đặc biệt. Vài thí dụ về proteins là:

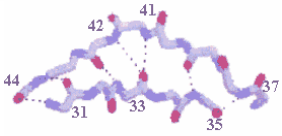
- **Protein cấu trúc** **(Structural proteins)**, có thể coi như các khối tạo dựng cơ sở của sinh vật.
- **Enzymes**, thực hiện (xúc tác) một số lớn các phản ứng sinh hóa học (biochemical reactions). Cùng với các phản ứng này và các **đường chuyển hóa** **(pathway)** chúng tạo ra sự **trao đổi chất** **(metabolism)**.
- **Protein màng** **(transmembrane proteins)**: chìa khóa của sự duy trì môi trường tế bào (cellular environment), điều hòa dung tích tế bào, etc.
- Hormones, antibodies, etc.

# Protein structures  Cấu trúc protein

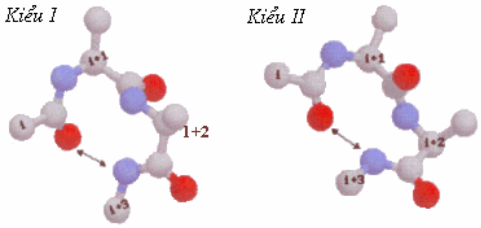
- **Cấu trúc bậc một (primary structure):** Các dãy của 20 loại amino acids khác nhau, nối với nhau theo một thứ tự tuyến tính bất kỳ (poly-peptide chains). Độ dài của phân tử protein có thể thay đổi từ vài đến nhiều ngàn amino-acids.
- **Cấu trúc bậc hai (secondary structure):** Là sự xoắn gấp (folding) của dãy các amino acids. Có hai loại cấu trúc thường thấy trong các dãy xoắn gấp: **alpha-helices (xoắn α)** và **beta-strands (dải β)**. Chúng được hợp với nhau một cách đặc trưng bởi các cấu trúc kém thông thường hơn **(loops, vòng).**



Cấu có lặp β song song




Cấu có lặp β không song song




*Hai kiểu cấu trúc vòng*



# Protein structures  Cấu trúc protein

- **Cấu trúc bậc ba (tertiary structure):** Do xoắn gấp, nhiều phần của dãy phân tử protein có sự tiếp xúc (contact) với nhau, tạo ra nhiều lực hút và lực đẩy giữa chúng, tạo cho phân tử có được một cấu trúc 3D tương đối bền vững và cố định.
- **Cấu trúc bậc bốn (quaternary structure):** Một protein có thể được tạo ra từ nhiều hơn một dãy amino-acids, và khi này nó được gọi là có *cấu trúc bậc bốn*. Thí dụ như haemoglobin được tạo ra từ bốn dãy trong đó mỗi dãy có khả năng bó lại (binding) một phân tử iron.

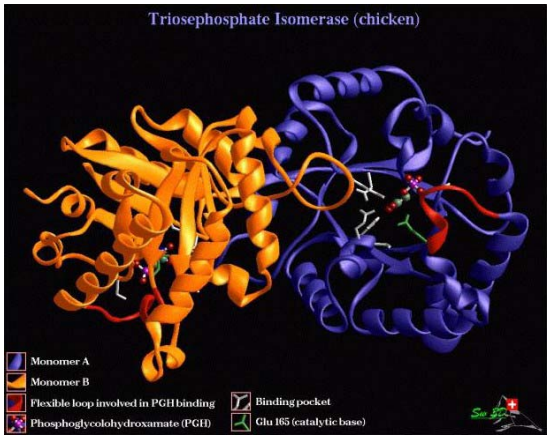
*Siêu cấu trúc bậc hai (xoắn-cuộn-xoắn)*

*Cấu trúc bậc ba của protein*

*Cấu trúc bậc bốn của protein*

# Proteins

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




Kích thước một protein có thể từ 3 đến 10 nanometers (nm), i.e., 3 đến 10 x tỷ mét ($10^{-9}$ m), và tìm ra cấu trúc của chúng là bài toán khó và tốn kém (cần khoảng €50,000 - €200,000 để tìm ra một cấu trúc mới).



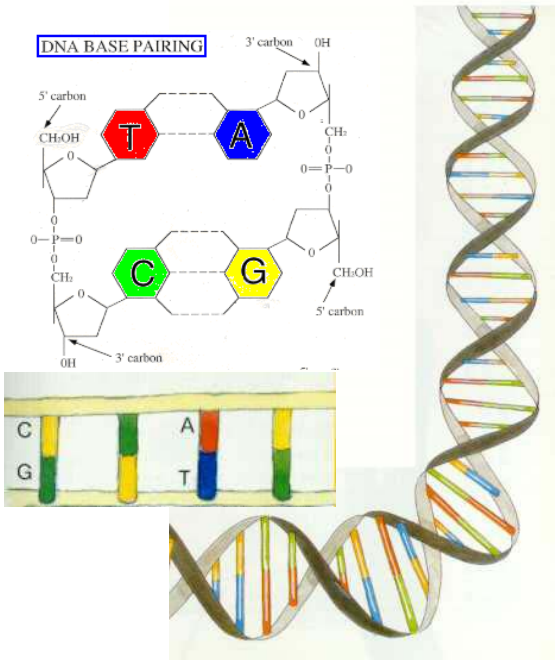
# DNA (Deoxyribonucleic acid)

- DNA là phân tử mang thông tin chủ yếu trong một tế bào. DNA có thể là xoắn đơn (single) hay xoắn kép (double)
- Phân tử DNA xoắn đơn là một dãy các đơn phân tử (**nucleotides**), còn gọi là **đa đơn phân tử (polynucleotide).**
- Bốn đơn phân tử khác nhau chia thành hai nhóm, gọi là bazơ **(bases):**
  - nhóm purines gồm adenosine (**A**) và guanine (**G**);
  - nhóm pyrimidines gồm cytosine (**C**) và thymine (T).
- Các đơn phân tử khác nhau có thể được nối với nhau theo mọi thứ tự dưới dạng đa đơn phân tử, như

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

# DNA (Deoxyribonucleic acid)

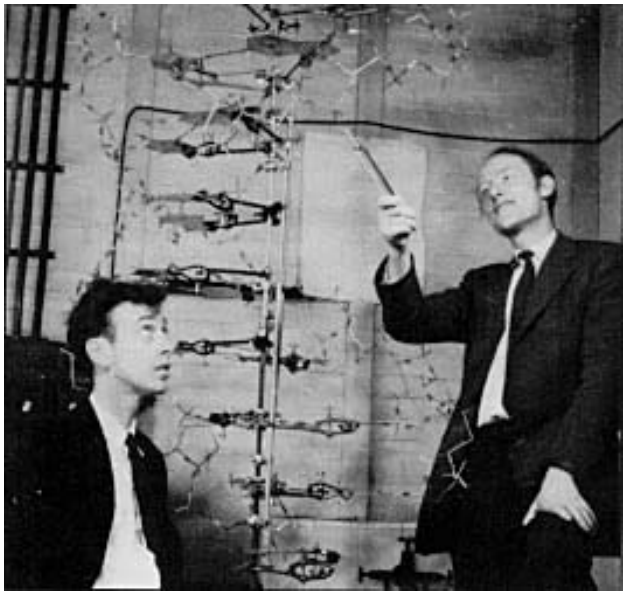
- Các cặp đơn phân tử đặc biệt có thể tạo nên các liên kết yếu (weak bonds) giữa chúng: **A liên kết với T, C liên kết với G.** Các cặp **A-T** và **G-C** gọi là các **cặp cơ sở (base-pairs, bp)**
- Khi hai dãy đa đơn phân tử liên kết với nhau, chúng thường dính vào nhau, gọi là các **DNA xoắn kép (double helix)**.
- Hai dải như vậy gọi là **liên kết** với nhau (**complementary**), và mỗi dải có thể thu được từ dải kia bằng cách thay tương hỗ A với T, C với G, và đổi hướng của phân tử theo chiều ngược lại.



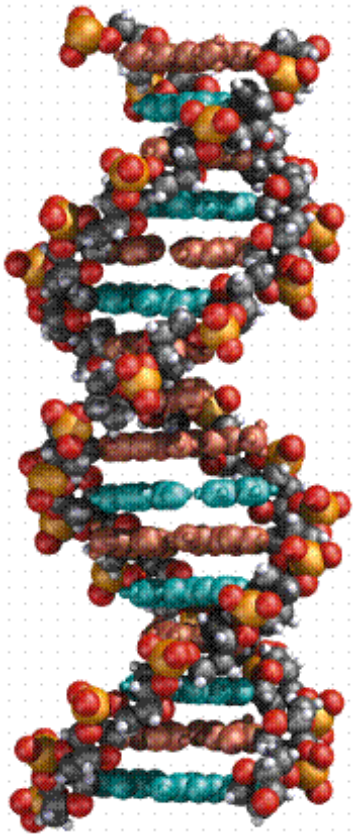

T-T-G-A-C-T-A-T-C-C-A-G-A-T-C
A-A-C-T-G-A-T-A-G-G-T-C-T-A-G



# DNA



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



# RNA (ribonucleic acid)

- RNA được tạo thành từ đơn phân tử như DNA. Tuy nhiên, RNA dùng **U (uracil)** thay vì T (pyrimidine thymine) là thành phần không có trong DNA (chỉ có dải đơn).
- RNA có nhiều chức năng trong tế bào, như mRNA và tRNA là các kiểu chức năng khác nhau của RNA, cần thiết trong sự tổng hợp protein.
- RNA có thể liên kết với một dải đơn của một phân tử DNA, bằng cách thay T bằng U, và các phân tử kiểu này có vai trò quan trọng trong các quá trình sống và công nghệ sinh học.

```
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 (Gene và các hệ gene)

1. **Chromosomes, genomes and sequencing** (Nhiễm sắc thể, hệ gene, và sắp dãy)
2. **Genes and protein synthesis** (gene và tổng hợp protein)
3. **Gene prediction** (đoán nhận gene)
4. **Genome similarity and SNPs** (sự giống nhau giữa các hệ gene và SNP)

## Chromosomes, genomes and sequencing
## Nhiễm sắc thể, hệ gene, và sắp dãy

- Nhiễm sắc thể **(chromosome)**: Một hay một vài phân tử DNA xoắn kép dài có tổ chức.
- Con người có 24 cặp nhiễm sắc thể.
- Chromasomal và mitochondrial DNA tạo nên **hệ gene (genome)** của sinh vật. Mọi sinh vật đều có hệ gene, và người ta tin rằng hệ gene mã hóa hầu hết thông tin di truyền của sinh vật.
- Mọi tế bào của một sinh vật đều chứa các hệ gene như nhau (identical genomes), với rất ít ngoại lệ, là kết quả cuả sự tái tạo DNA (DNA replication) khi tế bào phân chia.



## Chromosomes, genomes and sequencing
## Nhiễm sắc thể, hệ gene, và sắp dãy

- Xác định dãy bốn chữ cái của một phân tử DNA cho trước gọi là sắp dãy DNA **(DNA sequencing).**
  - Bộ gene của một vi khuẩn (a bacterium) được sắp dãy toàn bộ năm 1995. Bộ gene của (yeast) gđược sắp dãy năm 1997, giun (worm) năm 1999, ruồi (fly) năm 2000, và cỏ dại (weed) năm 2001.
  - Việc sắp dãy toàn bộ hệ gene con người được hoàn thành năm 2003, được biết như hệ gene người **(human genome).**
- Các hệ gene đều chứa gene, và phần lớn chúng mã hóa proteins.

# Genes và sự tổng hợp protein

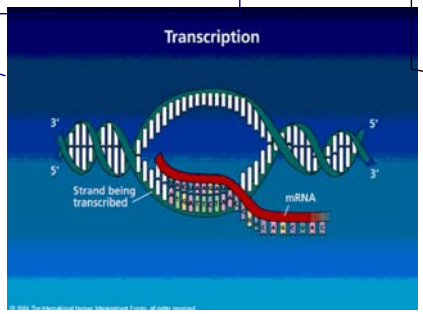
- **Genes** là các đoạn đặc biệt của DNA có chức năng điều khiển cấu trúc và hoạt động của tế bào; là đơn vị chức năng của sự di truyền.
- Để hiểu rõ hơn về gene, ta cần mô tả cơ chế tạo ra proteins dựa trên thông tin được mã hóa trong genes. Quá trình này được gọi là **sự tổng hợp proteins**, và gồm ba giai đoạn chính:

  1. Transcription (phiên mã)
  2. Splicing (ghép mã)
  3. Translation (dịch mã).

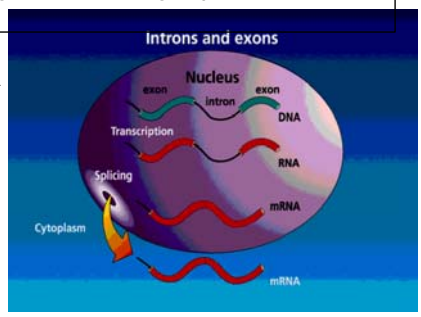


# Tổng hợp protein

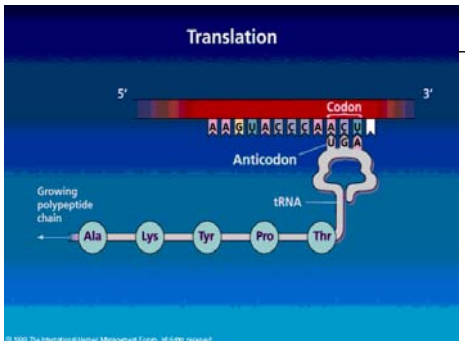
Một đoạn phân tử DNA được sao chép vào mRNA bổ sung (phiên mã)



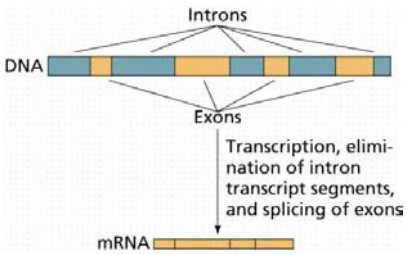

Bỏ đi vài mẩu của pre mRNA, gọi là **introns**, phần còn lại, gọi là **exons**, sẽ được nối với nhau. Số lượng và kích thước các introns và exons khác nhau rất đáng kể các genes cũng như giữa các chủng loại.




Sự dịch mã là một quá trình phức tạp và nhiều chi tiết chưa được biết.




Tạo proteins bằng cách nối các amino acids theo thứ tự được mã hóa trong mRNA. Thứ tự của amino acids được xác định bởi 3 đơn phân tử kề nhau trong DNA, gọi là **bộ ba hoặc mã di truyền (triplet or genetic code)**. Mỗi bộ ba được gọi là **codon** và mã cho một amino acid.



# Bài toán đoán nhận gene
# Gene prediction problem

- **Gene prediction**: Cho một dãy DNA, hãy nói gene ở đâu trong dãy này?

| Sinh vật | Số genes đã được đoán nhận | Phần của hệ gene mã hóa proteins (exons) |
|---|---|---|
| E.Coli (bacteria) | 5000 | 90% |
| Yeast (men) | 6000 | 70% |
| Worm (giun) | 18,000 | 27% |
| Fly (ruồi) | 14,000 | 20% |
| Weed | 25,500 | 20% |
| Human | 30,000 | < 5% |



# Sự tương tự của hệ gene và SNFs
# Genome similarity and SNPs

- Mọi hệ gene của người được xem là **tương đương đến 99.9%** và trung bình giữa các hệ genes của hai cá thể khác nhau cứ một nghìn đơn phân tử chỉ có một khác nhau.
- Sự biến dạng trong các phần không mã hóa của hệ gene được phân tích để để tạo ra các dạng (patterns) tin cậy để phân biệt các ca thể.
- Các biến dạng đặc biệt quan trọng trong hệ gene là **đa đẳng đơn phân tử** (**single nucleotide polymorphisms** (**SNP**), có thể xuất hiện trong các phần được mã hóa hay không mã hóa trong hệ gene. SNPs là các biến dạng dãy DNA xuất hiện khi các cơ sở đơn (A,C,G, or T) được đan xen sao cho các cá thể khác nhau có các chữ cái khác nhau tại các vị trí này.

# Functional genomics (Gene học chức năng)

Gene học chức năng **(functional genomics)** có thể được định nghĩa nôm na như việc dùng tri thức tiêu biểu về hệ gene để tìm hiểu về genes, về các chức năng sản xuất và sự tương tác của chúng, và quan trọng hơn là vì sao điều này làm cho các sinh vật hoạt động.

- **Gene functions** (Chức năng gene)
- **Protein abundance in a cell** (Sự dư thừa protein trong tế bào)
- **Gene regulation and networks** (Điều khiển gene và mạng gene)



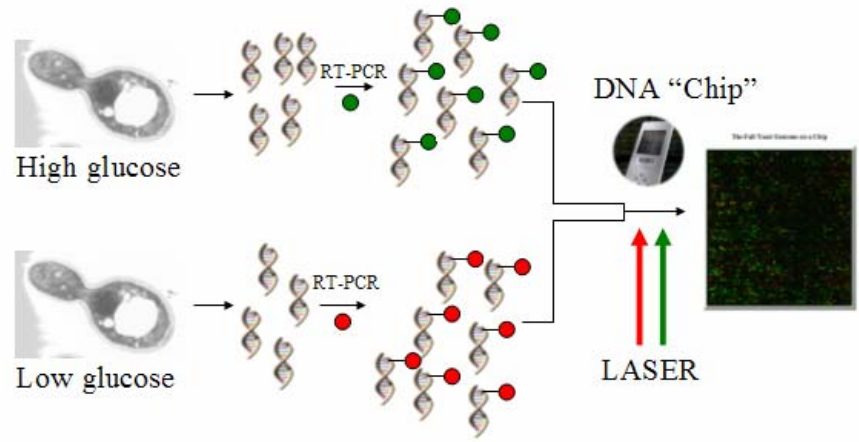
# Functional genomics Gene học chức năng

- Dường như có một hệ hạn chế các genes **(a limited universe of genes)** và proteins tương ứng của chúng. Từ quan điểm chức năng, rất nhiều trong chúng có trong phần lớn hoặc toàn bộ hệ các genes.
- Sự dư thừa protein **(protein abundance)** có thể phụ thuộc vào nhiều yếu tố như liệu gene tương ứng có được **thể hiện** (expressed) (i.e., được sao chép tích cực) hay không, được thể hiện nhanh và mạnh thế nào, được nối ghép, dịch chuyển, và thay đổi nhanh thế nào, etc.
- Thể hiện gene **(gene expression)** là quá trình qua đó thông tin mã hóa trong một gene được truyền vào cấu trúc đang có trong tế bào và điều khiển tế bào (hoặc proteins hoặc RNAs).

Một câu hỏi quan trọng và lý thú khác trong sinh học là sự thể hiện gene được "bật" và "tắt" thế nào, tức là các genes được điều chỉnh thế nào.

# Microarrays and gene expression databases

**Công nghệ microarray** sử dụng nguồn tạo bởi các đề tài về hệ gene  và các nỗ lực về dãy để trả lời câu hỏi các genes nào được thể hiện trong một kiểu tế bào đặc biệt của một sinh vật, ở một thời điểm đặc biệt, trong những điều kiện đặc biệt.






# Outline

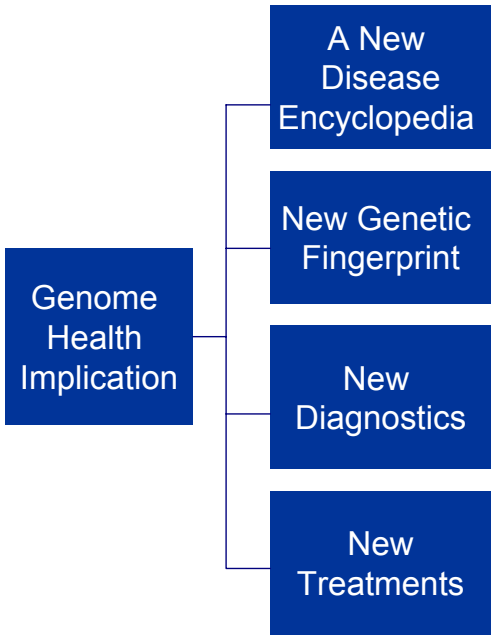
- Khái niệm cơ bản của sinh học
- Sinh tin học là gì?
- Về một vài bài toán trong sinh tin học


- 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
# Dự án về hệ gene người

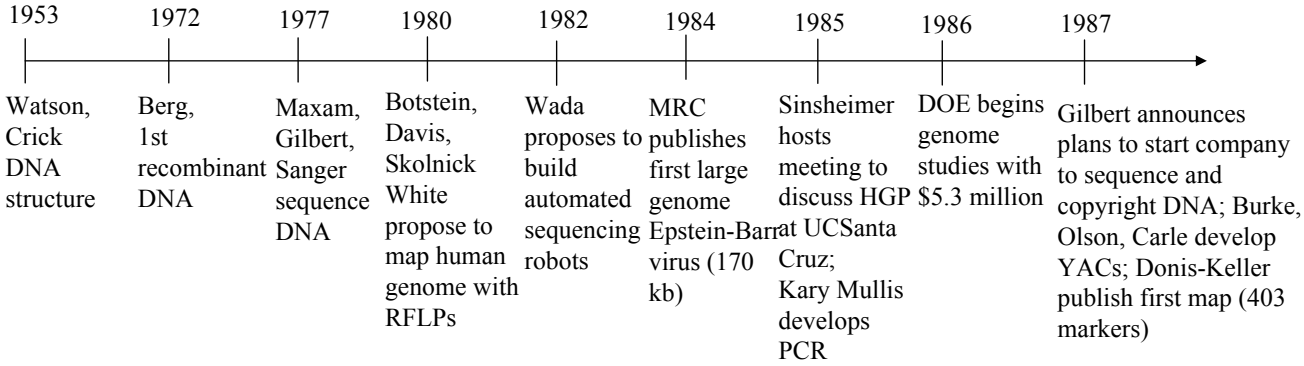
**Mục tiêu** (15 năm từ 1990)

- **Nhận biết** (identify) toàn bộ chừng 30,000 genes trong DNA của con người.
- **Xác định** (determine) các dãy của 3 tỷ cặp cơ sở tạo nên DNA của con người.
- **Lưu trữ** (store) thông tin này trongcác cơ sở dữ liệu.
- **Hoàn thiện** (improve) các công cụ phân tích dữ liệu.
- **Chuyển giao** (transfer) các công nghệ liên quan đến các doanh nghiệp tư nhân.
- **Đề cập** (address) các vấn đề về đạo đức, luật lệ, và xã hội (ELSI) có thể nảy sinh từ đề tài.






# History of the Human Genome Project
# Lịch sử của dự án hệ gene người



# History of the Human Genome Project
# Lịch sử của dự án hệ gene người (tiếp)

| 1987 (cont) | 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 1995 | 1996 |
|---|---|---|---|---|---|---|---|---|
| Hood produces first automated sequencer; Dupont devolops fluorescent dideoxy-nucleotides | NIH supports the HGP; Watson heads the project and allocates part of the budget to study social and ethical issues | Hood, Olson, Botstein Cantor propose using STS's to map the human genome | Proposal to sequence 20 Mb in model organism by 2005; Lipman, Myers publish the BLAST algorithm | Venter announces strategy to sequence ESTs. He plans to patent partial cDNAs; Uberbacher develops GRAIL, a gene finding program | Simon develops BACs; US and French teams publish first physical maps of chromosomes; first genetic maps of mouse and human genome published | Collins is named director of NCHGR; revise plan to complete seq of human genome by 2005 | Venter publishes first sequence of free-living organism: H. influenzae (1.8 Mb); Brown publishes on DNA arrays | Yeast genome is sequenced (S. cerevisiae) |



# History of the Human Genome Project
# Lịch sử của dự án hệ gene người (tiếp)

| 1997 | 1998 | 1999 | 2000 | 2001 | 2003 |
|---|---|---|---|---|---|
| Blattner, Plunket complete E. coli sequence; a capillary sequencing machine is introduced. | SNP project is initiated; rice genome project is started; Venter creates new company called Celera and proposes to sequence HG within 3 years; C. elegans genome completed | NIH proposes to sequence mouse genome in 3 years; first sequence of chromosome 22 is announced | Celera and others publish Drosphila sequence (180 Mb); human chromosome 21 is completely sequenced; proposal to sequence puffer fish; Arabadopsis sequence is completed | Celera publishes human sequence in Science; the HGP consortium publishes the human sequence in Nature | Completely sequenced human genome. |

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

# What is bioinformatics?
# Tin sinh học là gì?

- **Bio**: Sinh học phân tử (Molecular Biology)
- **Informatics**: Khoa học tính toán
- **Bioinformatics**: Giải quyết các bài toán sinh học bằng việc sử dụng các phương pháp của khoa học tính toán.

  **Synonyms**: Computational biology, Computational molecular biology, Biocomputing



# Thay đổi trong sinh học
# Paradigm shift in biology

Một kiểu thức mới đang xuất hiện là tất cả các 'genes' sẽ sớm được biết hết (theo nghĩa có trong các cơ sở dữ liệu điện tử), và nghĩa là điểm bắt đầu của một khảo sát sinh học sẽ là lý thuyết. Mỗi nhà khoa học sẽ khởi đầu bằng một **ước đoán lý thuyết**, **rồi mới chuyển qua làm thí nghiệm** để theo hoặc kiểm tra giả thuyết.

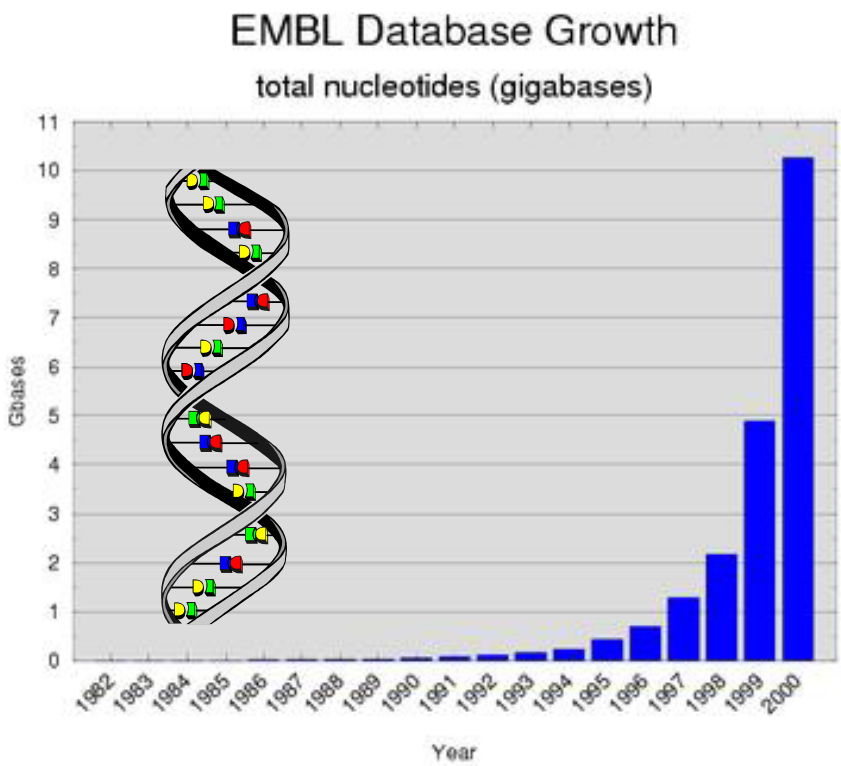
Để dùng dòng chảy tri thức trên các mạng toàn cầu, các nhà sinh học không những phải biết **dùng máy tính**, mà còn phải **thay đổi cách tiếp cận của mình** đối với bài toán hiểu sự sống.

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.

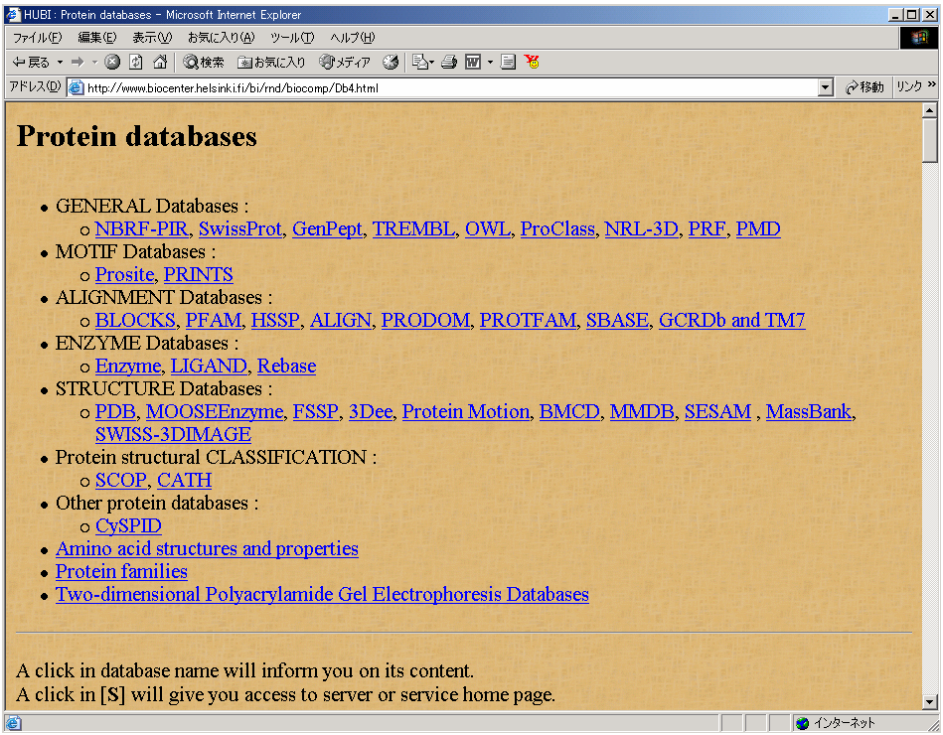
# Base Pairs in GenBank




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



# Public databases



# Mở rộng các khái niệm của Tin sinh học

- **Gene học (genomics)**
  - Gene học chức năng
  - Gene học cấu trúc
- **Protein học (Proteomics):** Phân tích proteins của một sinh vật ở nhiều mức (large scale)
- **Gene dược học (Pharmacogenomics):** Phát triển các thuốc mới nhằm đến các bệnh đặc biệt
- **Microarray (genome chip):** DNA chip, protein chip

Xác định và đặc trưng chức năng của genes.

Nghiên cứu thể hiện gene ở mọi mức của protein bởi đồng nhất và đặt trưng proteins có trong các mẫu sinh học.

Dùng thông tin về gene để dự đoán sự an toàn, độc tính và/hoặc hiệu quả của thuốc với người bệnh hoặc nhóm người bệnh.

Một công nghệ mới nhằm đưa toàn bộ hệ gene trên một chip sao cho các nghiên cứu viên có một bức tranh tốt hơn về tương tác đồng thời của hàng ngàn genes



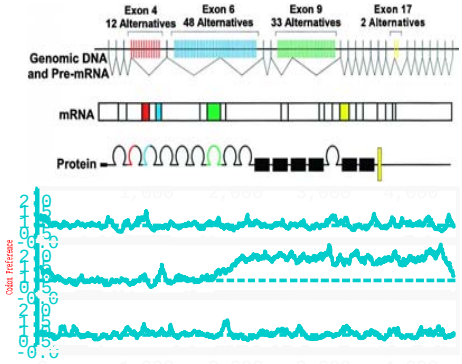
# Problems in Bioinformatics

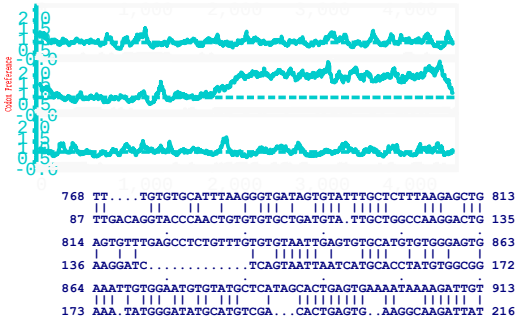
**Phân tích cấu trúc**
- So sánh cấu trúc protein
- Dự đoán cấu trúc protein
- Mô hình hóa cấu trúc RNA

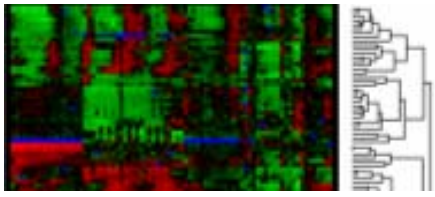
**Phân tích đường chuyển hóa**
- Đường trao đổi chất (metabolic pathway)
- Mạng điều tiết (regulatory networks)

**Phân tích dãy**
- Sắp dãy (sequence alignment)
- Dự đoán chức năng và cấu trúc
- Tìm gene (Gene finding)

**Phân tích thể hiện**
- Phân tích thể hiện gene
- Phân nhóm gene

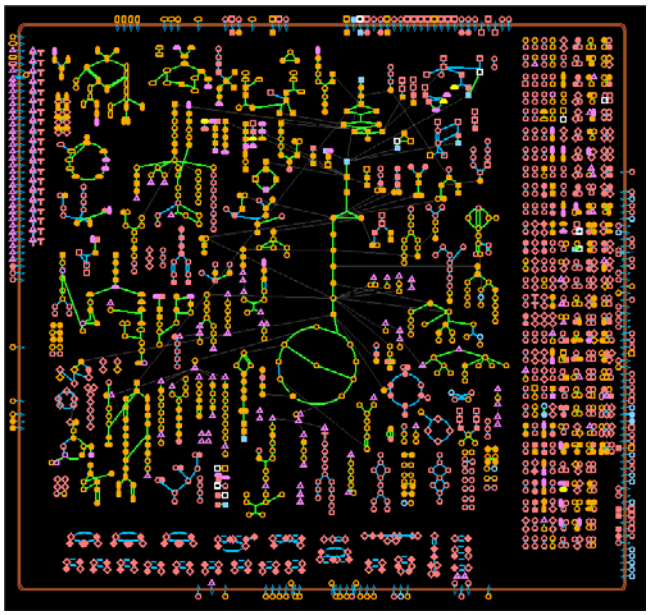


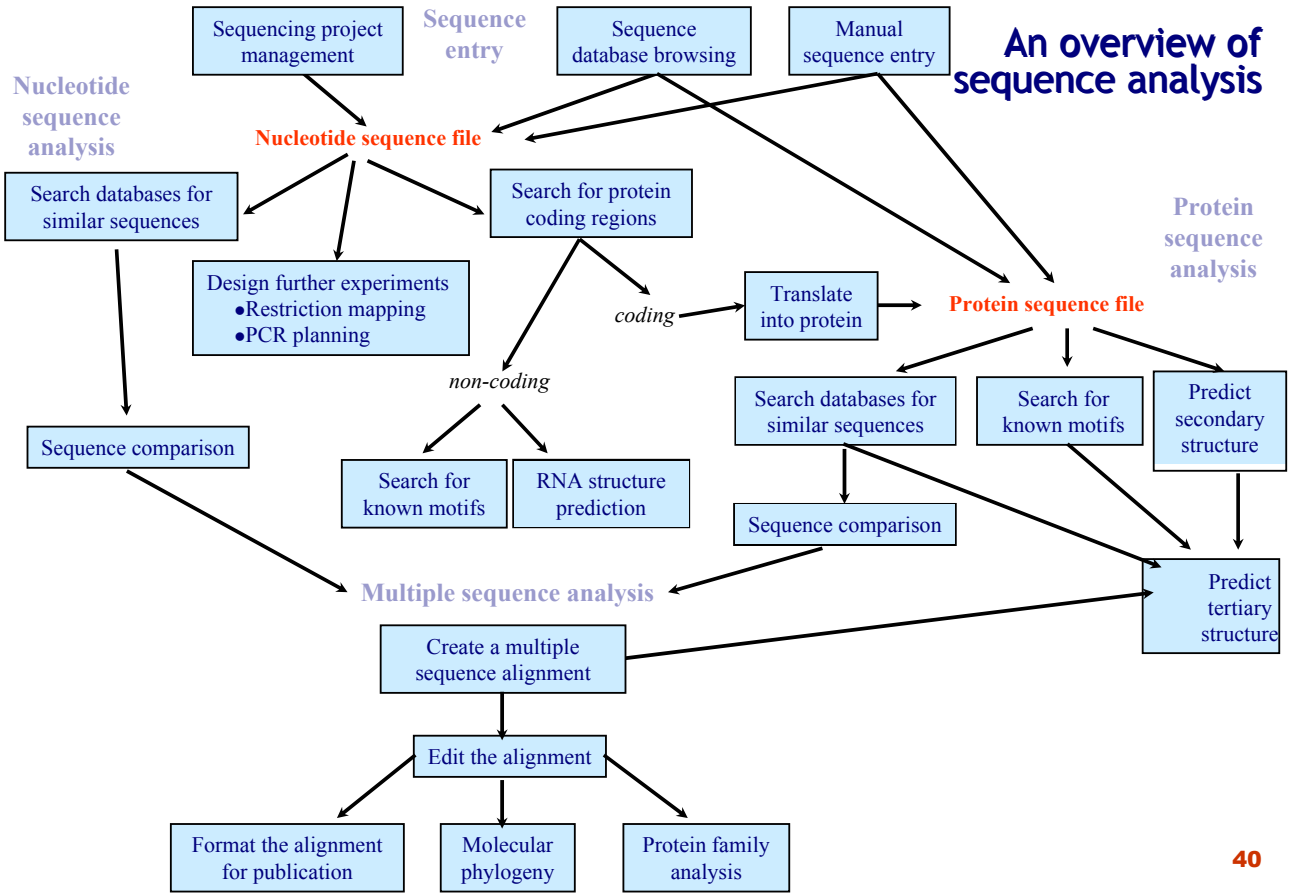



# Pathway analysis

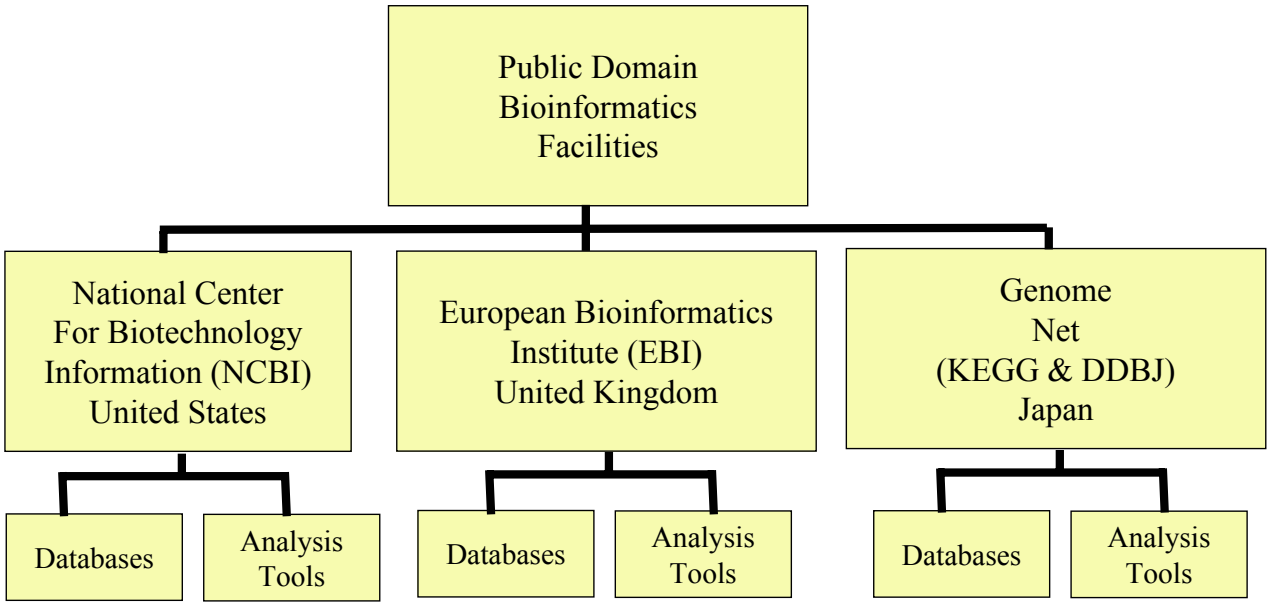
- Mỗi **phản ứng hóa học** hoán chuyển (interconverts) các thành phần hóa học
- Một **enzyme** là một protein có chức năng thúc đẩy các phản ứng hóa học
- Một **đường chuyển hóa** (**pathway**) là một tập các phản ứng hóa học được nối với nhau.







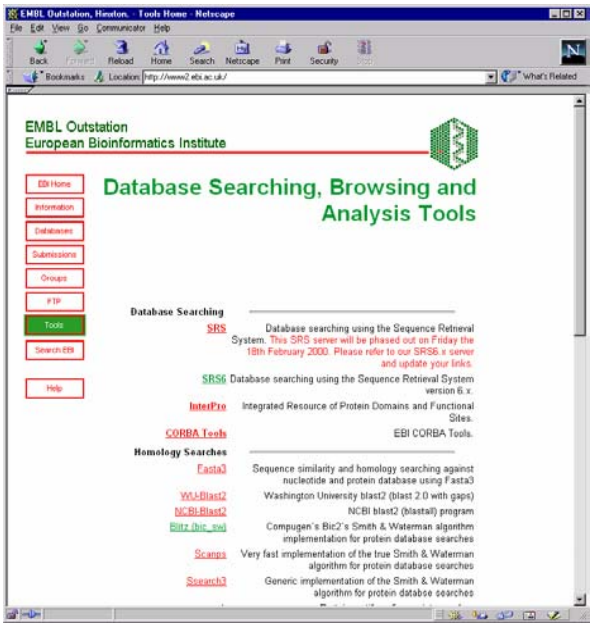
# Primary public domain bioinformatics servers






# Analysis Tools  Công cụ phân tích

EBI lo các versions để tìm các cơ sở dữ liệu trong các lĩnh vực công cộng chủ yếu và các công cụ phân tích như FASTA, CLUSTALW, BLAST, và các cài đặt của Smith & Waterman.



# Challenges in Bioinformatics

- Tin sinh học đòi hỏi:
  - Truy nhập vào được nhiều nguồn phân tán (Access to multiple distributed resources)
  - Cần thông tin được cập nhật (Needs information to be up-to-date)
  - Dư thừa dữ liệu tối thiểu (Minimal data redundancy)
  - Các ứng dụng ổn định (Robust applications)
  - Các ứng dụng mở rộng được (Extendable applications)
    - Monolithic App. vs. Components
  - Các phần mềm chuyển tải được (Portable software)



# Challenges in Bioinformatics

- Bùng nổ thông tin
  - Cần phân tích được nhanh, tự động để xử lý được lượng thông tin lớn
  - Cần tích hợp được nhiều kiểu thông tin khác nhau (sequences, literature, annotations, protein levels, RNA levels etc…)
  - Cần các phần mềm "thông minh hơn" để nhận biết được các quan hệ quan trọng trong các tập dữ liệu rất lớn.
- Thiếu các "nhà tin sinh học" ("bioinformaticians")
  - Phần mềm cần dễ truy nhập, dễ dùng và dễ hiểu hơn
  - Nhà sinh học cần học phần mềm, thấy hạn chế của chúng, và cách giải thích kết quả của chúng.

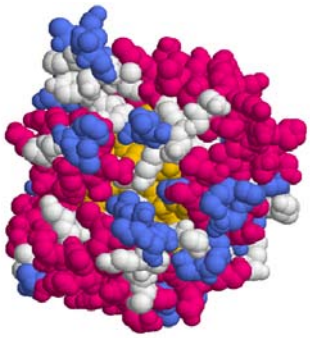
# Outline

- Khái niệm cơ bản của sinh học
- Sinh tin học là gì?
- Về một vài bài toán trong sinh tin học



# Bài toán đoán nhận cấu trúc protein

- Có khoảng 15,000 cấu trúc protein trong các cơ sở dữ liệu công cộng, và trong số này rất nhiều cấu trúc giống nhau. Con người mới biết chừng 1,500 cấu trúc protein khác nhau.
- Dự đoán cấu trúc protein từ các dãy amino-acid là một trong các bài toán quan trọng nhất của tin sinh học, và con người còn đang cách lời giải rất xa.



# Đối sánh dãy (string matching)

## (Approximate) String Matching

**Input:** Text T , Pattern P

**Question(s):**

P xuất hiện trong T?

Tìm một xuất hiện của P trong T.

Tìm mọi xuất hiện của P trong T.

Tính số xuất hiện của P trong T.

Tìm dãy con dài nhất của P trong T.

Tìm dãy con gần nhất của P trong T.

Xác định các lặp trực tiếp của P trong T.

**và nhiều biến dạng khác**

**Applications:**

Liệu P đã có trong cơ sở dữ liệu T?

Xác định vị trí của P trong T.

Liệu có thể dùng P như một nguyên tố của T?

P có tương đồng với gì đó trong T?

P có bị hỏng bởi T?

Liệu *prefix*(P) = *suffix*(T)?

Xác định các lặp sau trước (tandem) của P trong T.



# Đối sánh dãy String matching

| | |
|---|---|
| **Input:** | Text **T**; Pattern **P** |
| **Output:** | Mọi xuất hiện của **P** trong **T**. |

**Chiến lược trượt window:**

Khởi tạo một window từ đầu của T;

While (window còn trong T) do

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

Shift: dịch window về bên phải (một vị trí)

endwhile;

# Đối sánh dãy String matching

ATAQAANANASPVANAGVERANANESISITALVDANANANANAS

AAAAAA ANANAS ANANAS ANANAS

AN AN ANANAS



# Sắp thẳng dãy từng cặp
# Pairwise Sequence Alignment

- **Input**
  - Hai dãy chữ cái
  - Một cách cho điểm
- **Output**
  - Cách sắp thẳng dãy tối ưu

- Bài toán cơ bản nhất của tin sinh học
- Các dãy được sắp thẳng ⇒ có dùng cấu trúc hoặc chức năng
- Cho nhiều gợi ý nếu cấu trúc và chức năng của một trong các dãy được sắp thẳng đã biết

ATTGCGC → ATTGCGC

ATTCCGC → ATCCGC

→ ATTGCGC
AT-CCGC

→ ATTGCGC
ATC-CGC

→ ATTGCGC
ATCCG-C

# HMM in sequence alignment
# HMM trong bài toán sắp dãy

- Các trạng thái của HMM sẽ được chia thành các loại: đối sánh (*match*), thêm vào (*insert)* và xóa (*delete*).
- Bảng chữ cái M bao gồm hai mươi amino acids với một ký hiệu câm $\delta$ (dummy symbol) biểu diễn cho "delete". Trạng thái xóa chỉ cho ra $\delta$ (output $\delta$).
- Mỗi trạng thái "đối sánh" và "thêm vào"có phân bố riêng trên 20 amino acids, và ký tự $\delta$ không được tryuền đi.
- Các dãy được sắp dãy sẽ được dùng như dữ liệu huấn luyện, để học các tham số của mô hình
- Với mỗi dãy, thuật toán Viterbi được dùng để xác định một đường (path) khả dĩ nhất để tạo ra dãy.



# HMM in sequence alignment
# HMM trong bài toán sắp dãy

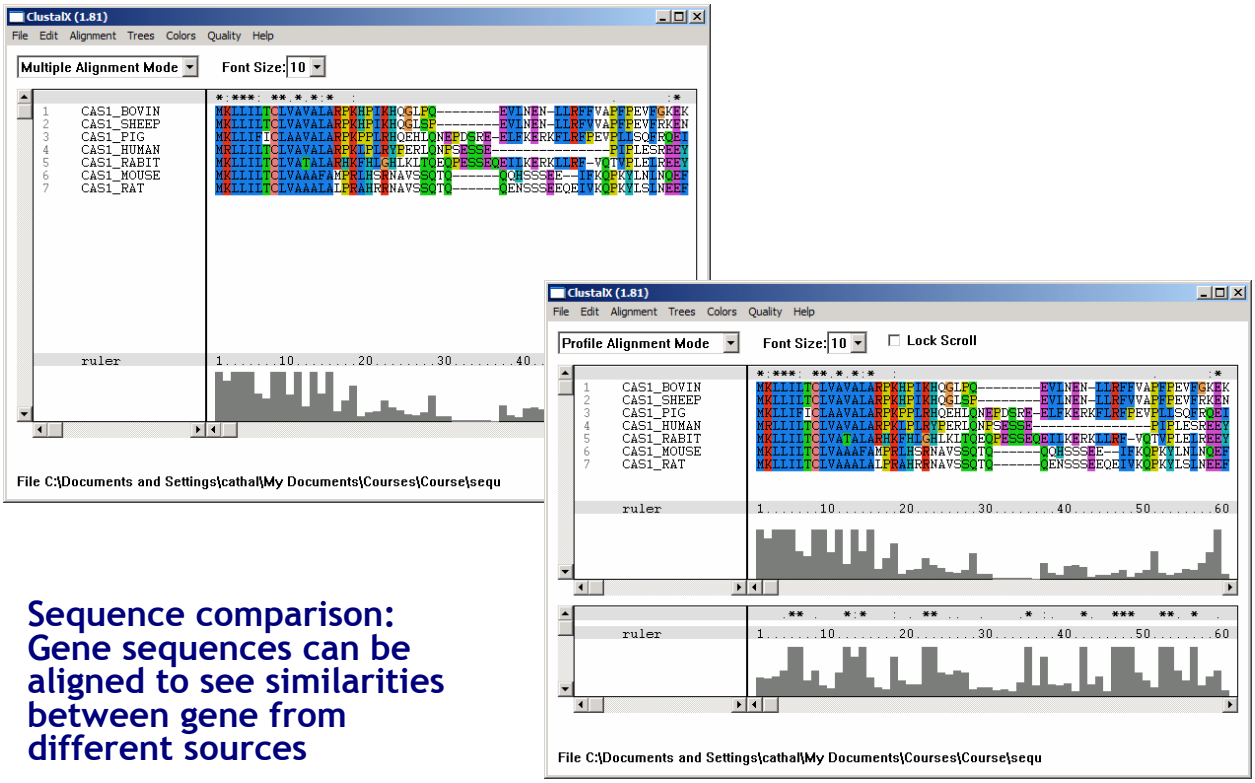
- Xét các dãy
  - CAEFDDH
  - CDAEFPDDH
- Giả sử mô hình có độ dài 10 và những đường khả dĩ (likely) nhất trong mô hình là
  - $m_0m_1m_2m_3m_4d_5d_6m_7m_8m_9m_{10}$
  - $m_0m_1i_1m_2m_3m_4d_5m_6m_7m_8m_9m_{10}$
- Phép sắp hàng được tìm ra bởi sắp các vị trí vốn được sinh ra mởi cùng một trạng thái đối sánh. Kết qủa là pháp sắp dãy sau
  - C–AEF –DDH
  - CDAEFPDDH

# Sắp dãy từng cặp và sắp dãy bội
# Pairwise vs Multiple Sequences

- Các cặp dãy được sắp một cách tiêu biểu do dùng các thuật toán vét cạn bởi quy hoạch động.
  - Độ phức tạp của các phương pháp vét cạn là $O(2^n m^n)$
  - *n = số các dãy*
- Sắp dãy bội xử dụng các phương pháp heuristic

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





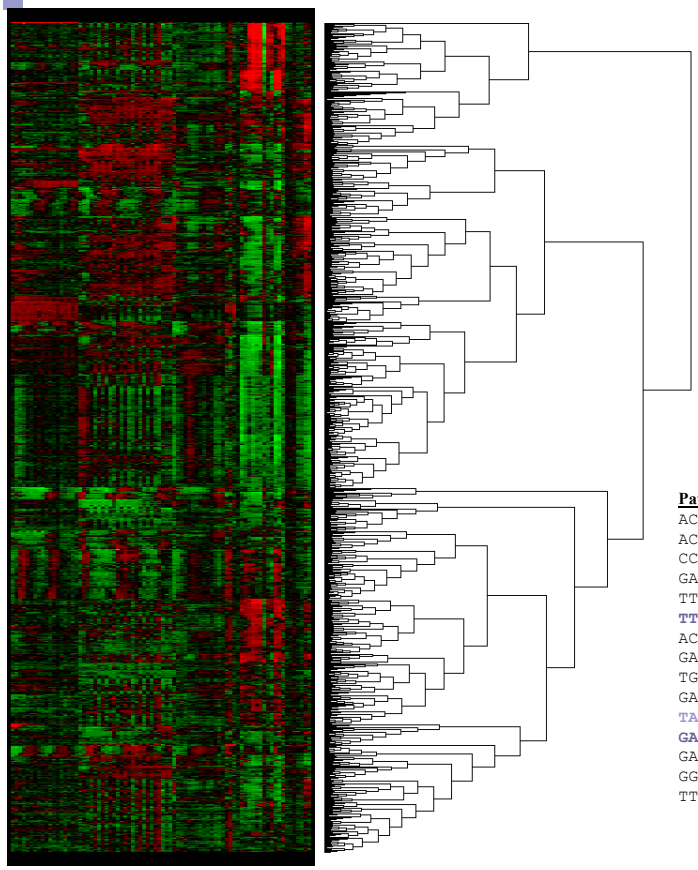

**Sequence comparison: Gene sequences can be aligned to see similarities between gene from different sources**

# Đoán nhận gene  Gene prediction

Là bài toán quan trọng của tin sinh học và hiện có nhiều thuật toán cho đoán nhận gene dựa trên các gene đã biết như dữ liệu huấn luyện. Một kỹ thuật toán nhận gene phổ biến là Hidden Markov Models (HMMs).

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

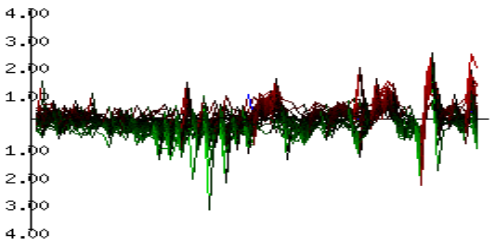


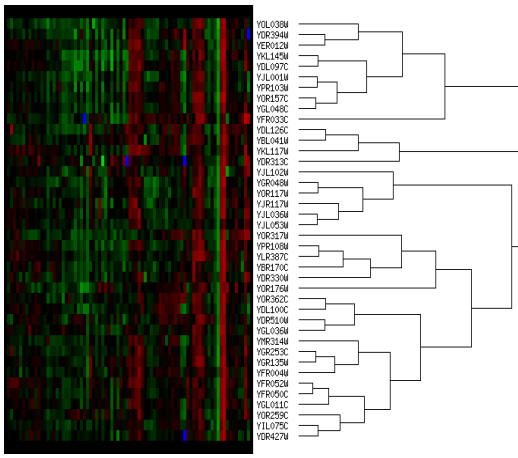



# Gene clustering and some discovered patterns

| 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



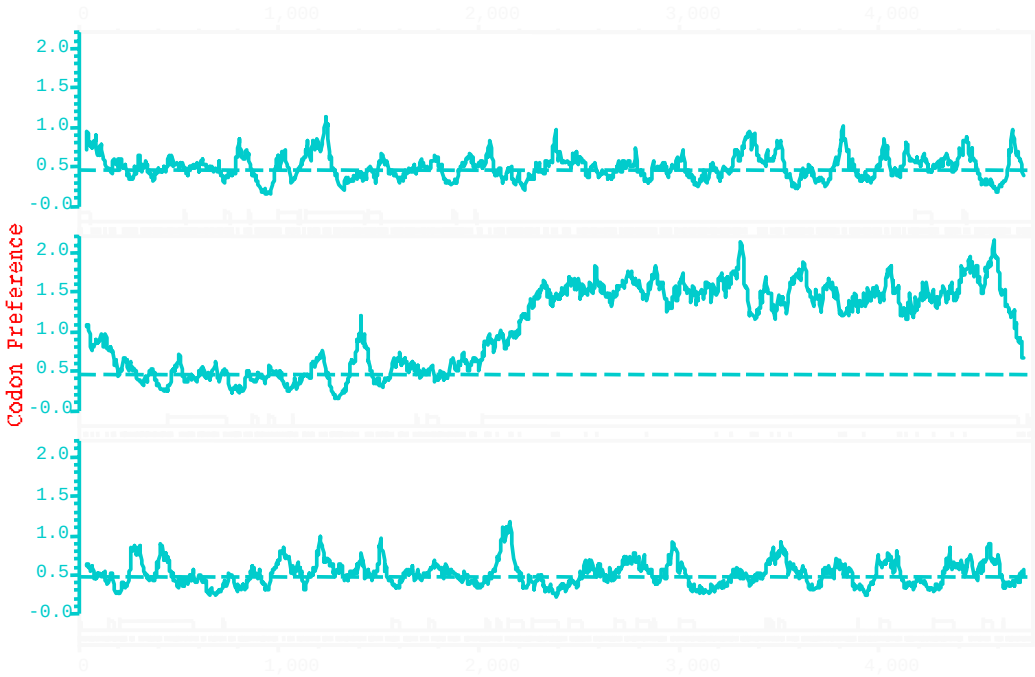




| ORF | Gene | Description | | |
|---|---|---|---|---|
| YBL041W | PRE7 | 20S proteasome subunit(beta6) | | |
| YBR170C | NPL4 | nuclear protein localization factor and ER translocation component | | |
| YDL126C | CDC48 | microsomal protein of CDC48/PAS1/SEC18 family of ATPases | | |
| YDL100C | | similarity 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:

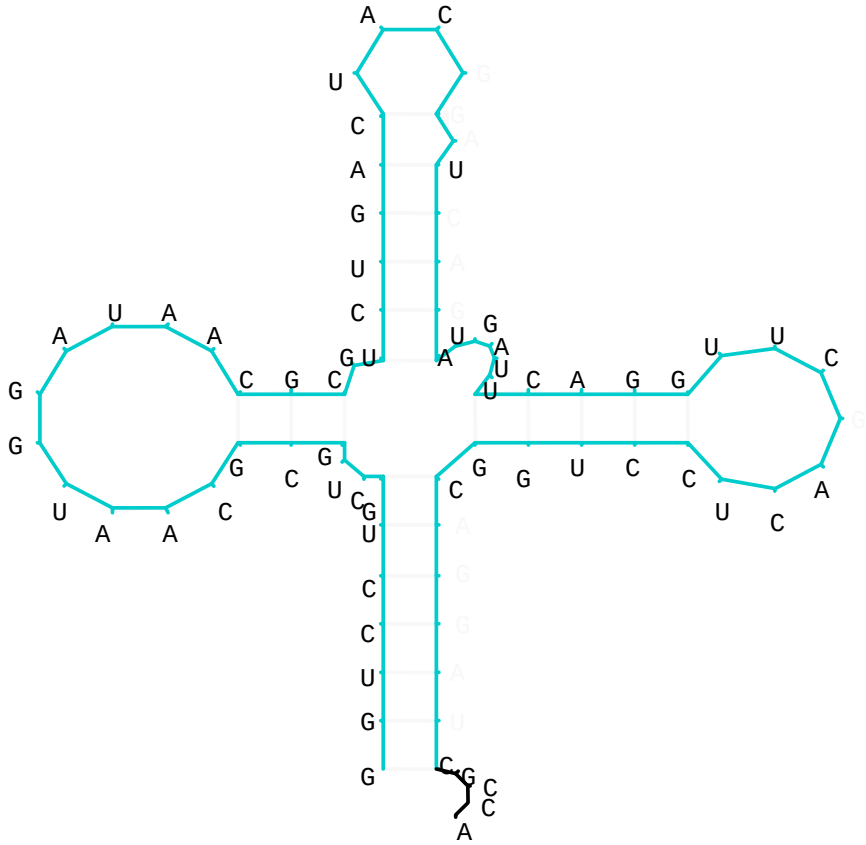
## Computer program can be used to recognise the protein coding regions in DNA




Plot created using codon preference (GCG)

# RNA structure prediction

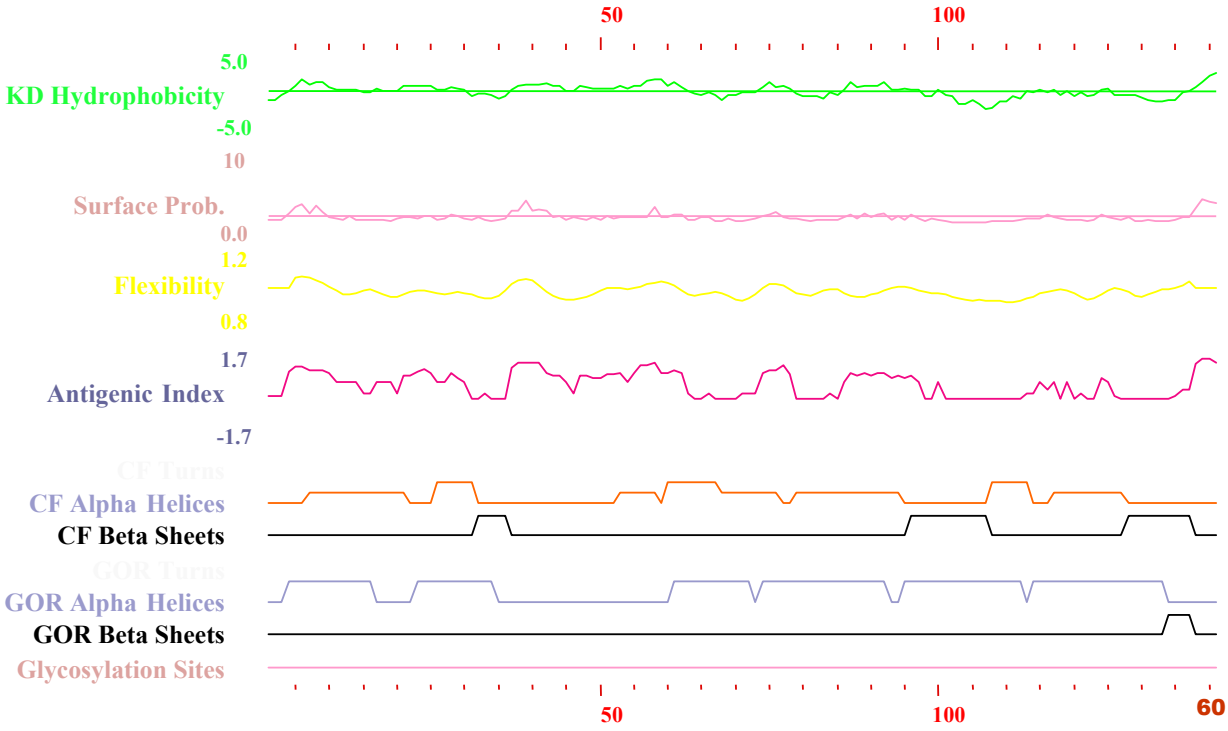
Structural features of RNA can be predicted





# 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

- Đề cập một số khái niệm cơ bản trong sinh học và tin sinh học, và những bài toán chính của tin sinh học.
- Tin sinh học là một lĩnh vực quan trọng, đầy thách thức.
- Tin sinh học liên quan chặt với data mining and machine learning.
- Ta cần đi con đường nào?

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