# Quantitative Biology

**MiniTAGp, Annecy 16/09/2011**

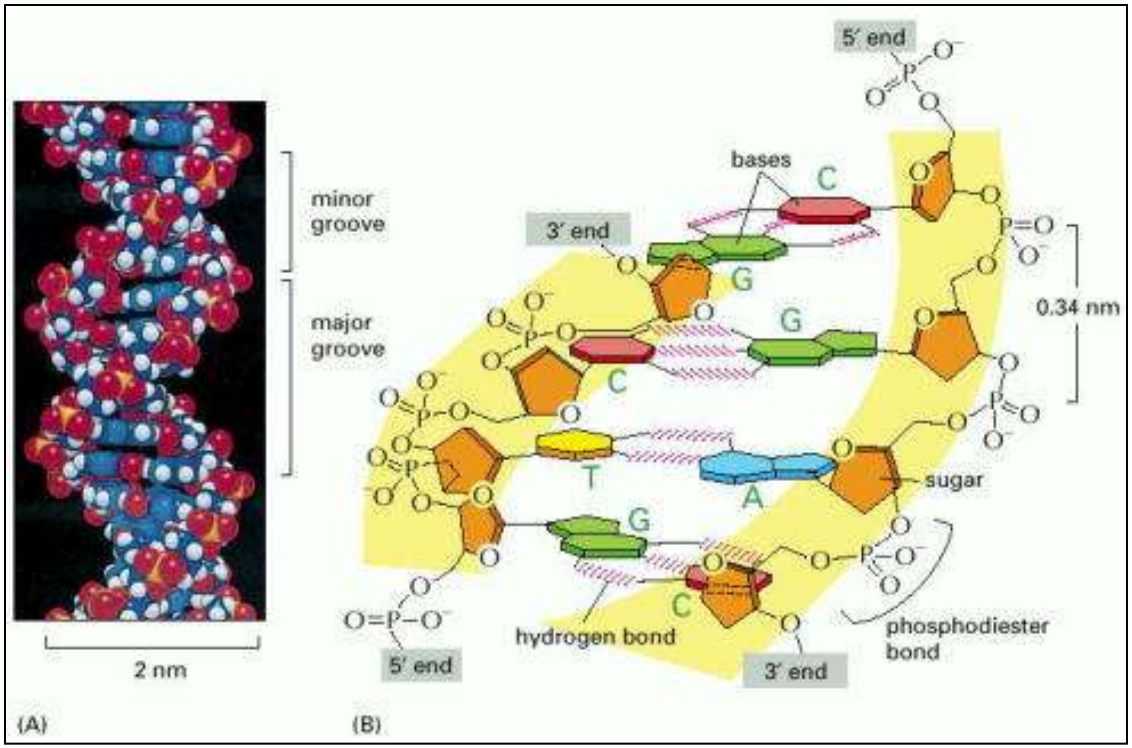
*Michele Caselle – University of Torino and INFN*
*caselle@to.infn.it*

# Plan of the lecture

1. Introduction: DNA, genes and proteins
2. The last ten years: The “genomic revolution”
3. New tools and ideas:
   Computational Biology and Systems biology
4. Example 1: Evolutionary models
5. Example 2: Gene Regulation
6. Example 3: Chemotaxis

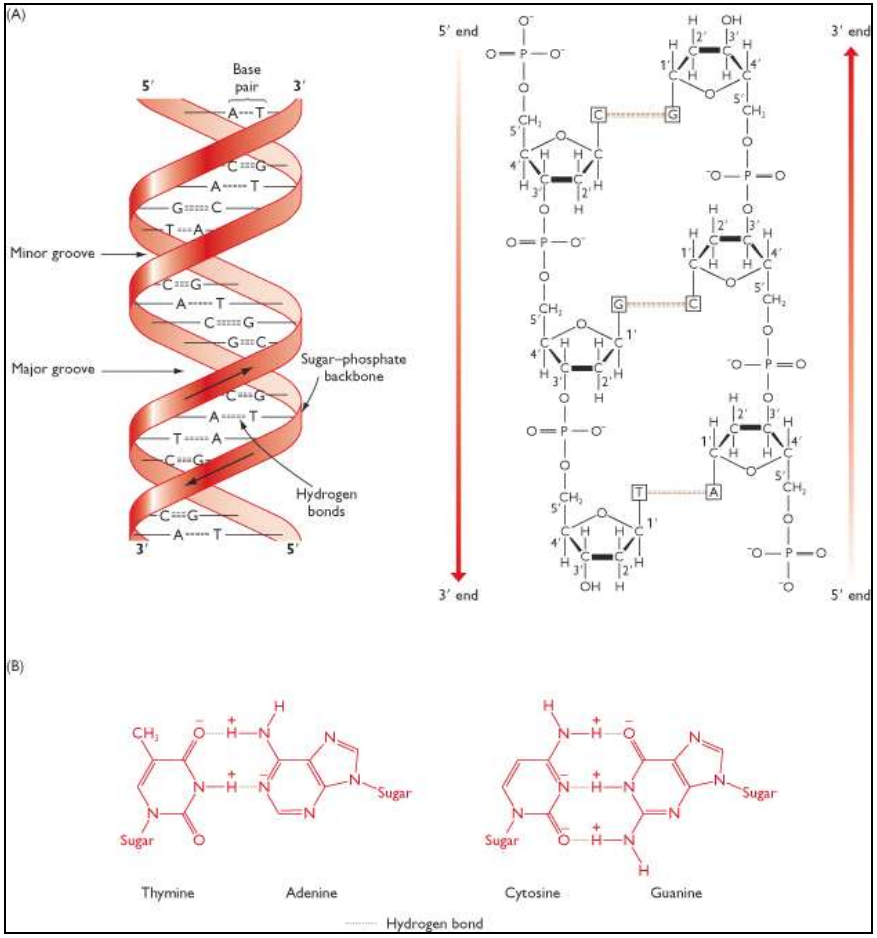
# DNA

- Genomic information is encoded in the **DNA** chain.
- In the human case the genome is composed by 3x10^9 base pairs which may take four possible values: **A,C,G,T**



# DNA

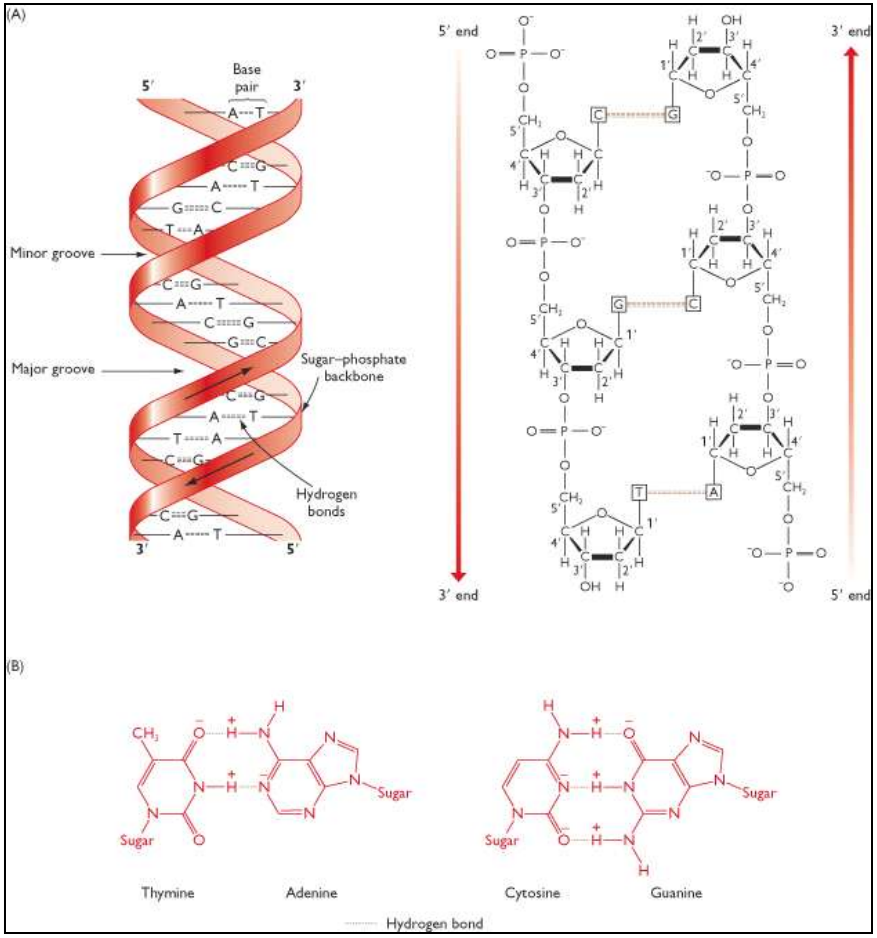
The main property of the **DNA** chain is base pairing: (A,T) and (C,G). This allows both DNA replication and the use of the chain as a template for protein production.



# DNA

The **DNA** chain has a well defined "direction". The "beginning" of a strand of a DNA molecule is defined as 5′, the "end" is defined as 3′. (5′ and 3′ refer to the position of the bases relative to the sugar molecule in the DNA backbone).

The two strands in a double helix run in opposite directions.



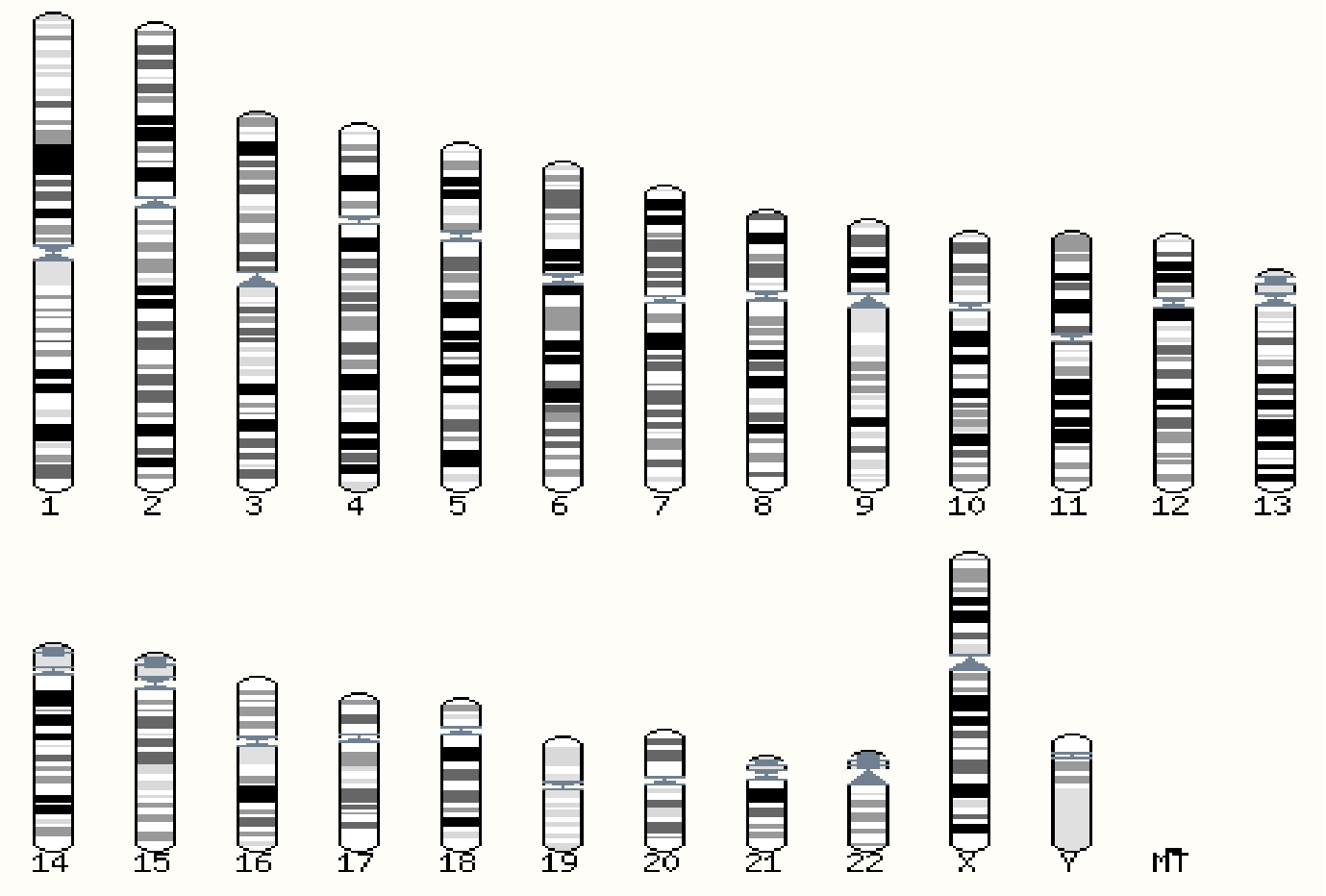
# Genome Organization I

An organism's total DNA content is known as its **Genome**.

- The human genome consists of 22 pairs of autosomal **chromosomes** and two sex chromosomes X and Y.
- **Genes** are sequences of bases that encode informations for proteins and some RNA molecules (such as ribosomal RNAs, Transfer RNAs, miRNAs). They can range in size from less than 100 bp (base pairs) to several millions of bp.
- The portion of the genome coding for proteins decreases as the complexity of the organism increases. It is very high in procaryotes and yeast but very low in mammalian.

  97% of the human genome is **non-coding**!!
- Most of this non-coding DNA is involved in the **regulation of gene expression**

# Human Genome



# Genome Sizes (Mb)

## *Procaryotes:*

| | |
|---|---|
| **Mycoplasma Genitalium** | **0,58** |
| **Escherichia Coli** | **4,64** |

## *Eucaryotes:*

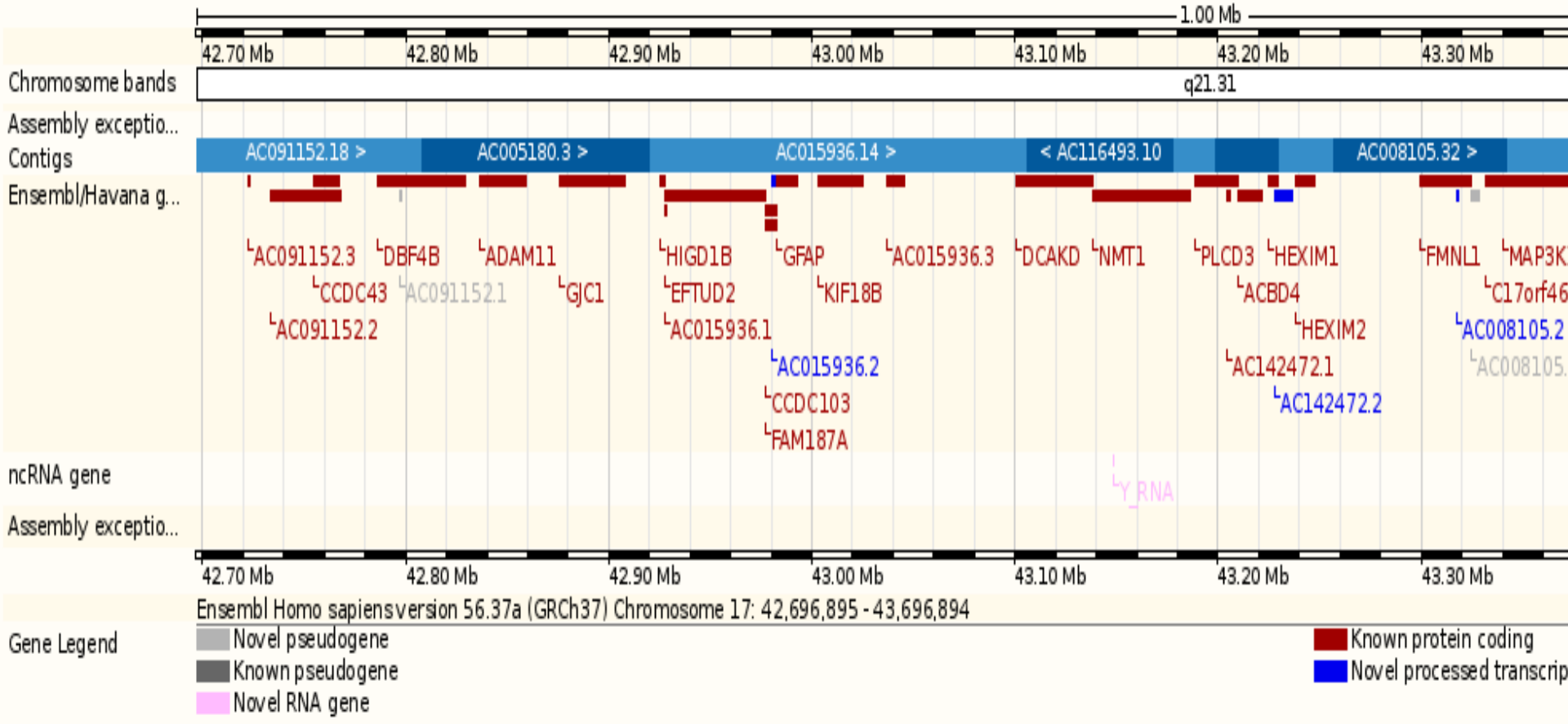
| | |
|---|---|
| **Saccaromices cerevisiae** | **12** |
| **Arabidopsis thaliana** | **100** |
| **Drosophila Melanogaster** | **140** |
| **Caenorabditis Elegans** | **100** |
| **Homo Sapiens** | **3000** |

# Gene structure

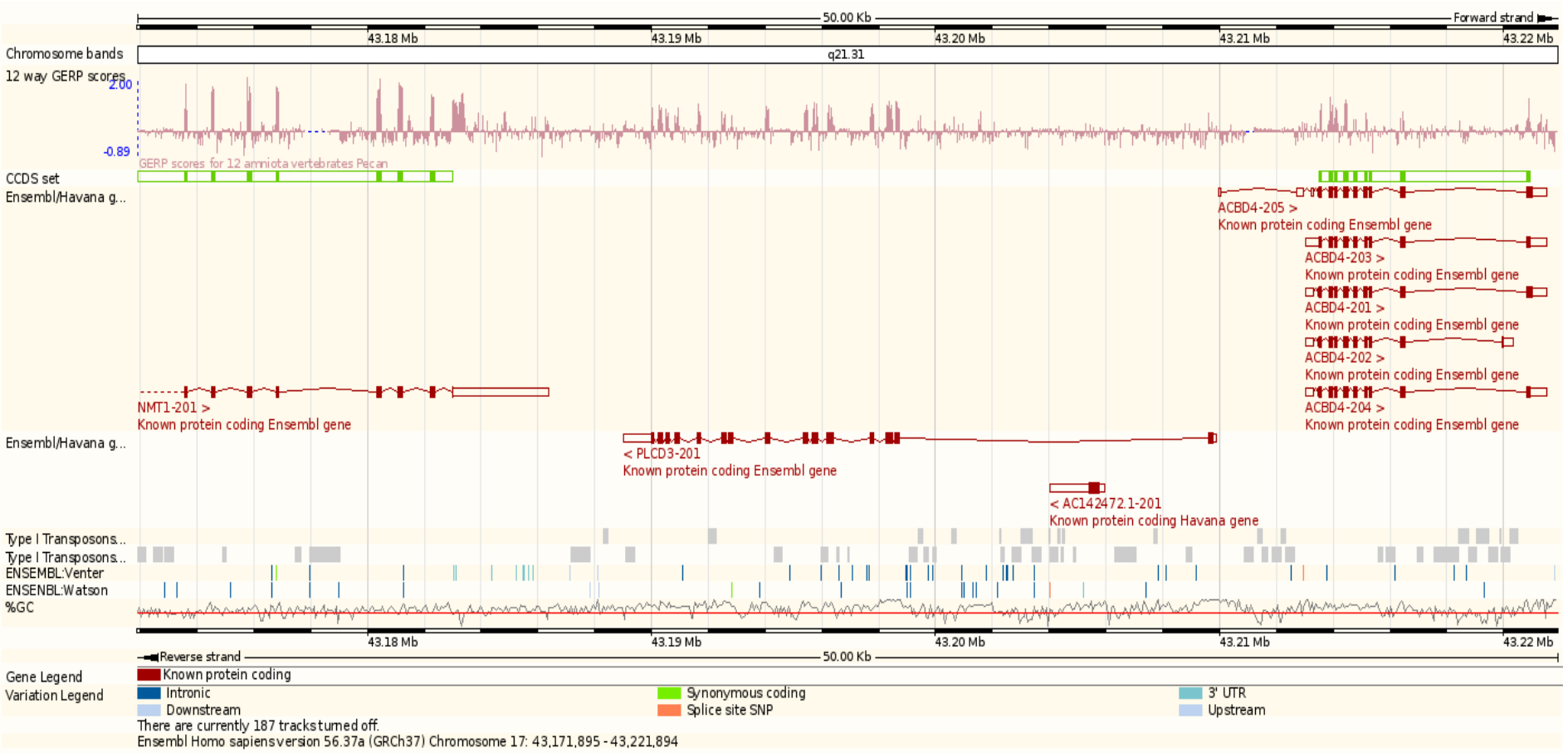
A typical human gene has a very complex internal structure. It is composed by coding blocks (**exons**) separated by long non-coding sequences (**introns**). Exons are glued together during the mRNA maturation (**splicing process**). They can be glued in many different ways thus giving, upon translation several different proteins (**alternative splicing**)

At the beginning and at the end of the mRNA there are two untranslated regions: (**5'UTR**) and (**3'UTR**) which are important for controlling functions and activities of the genes.

# Ensembl Genome Browser



# Zoom !


50.00 Kb
Forward strand
43.18 Mb
43.19 Mb
43.20 Mb
43.21 Mb
43.22 Mb
Chromosome bands
q21.31
12 way GERP scores
2.00
-0.89
GERP scores for 12 amniota vertebrates Pecan
CCDS set
Ensembl/Havana g...
ACBD4-205 >
Known protein coding Ensembl gene
ACBD4-203 >
Known protein coding Ensembl gene
ACBD4-201 >
Known protein coding Ensembl gene
ACBD4-202 >
Known protein coding Ensembl gene
ACBD4-204 >
Known protein coding Ensembl gene
NMT1-201 >
Known protein coding Ensembl gene
Ensembl/Havana g...
< PLCD3-201
Known protein coding Ensembl gene
< AC142472.1-201
Known protein coding Havana gene
Type I Transposons...
Type I Transposons...
ENSEMBL:Venter
ENSEMBL:Watson
%GC
43.18 Mb
43.19 Mb
43.20 Mb
43.21 Mb
43.22 Mb
Reverse strand
50.00 Kb
Gene Legend
Known protein coding
Variation Legend
Intronic
Synonymous coding
3' UTR
Downstream
Splice site SNP
Upstream
There are currently 187 tracks turned off.
Ensembl Homo sapiens version 56.37a (GRCh37) Chromosome 17: 43,171,895 - 43,221,894

# Proteins

Most of the functions in the cell are performed by **proteins** which are composed by 20 different types of elementary constituents: the **aminoacids**.
Proteins synthesis from the DNA template ("gene expression") occurs in three main steps

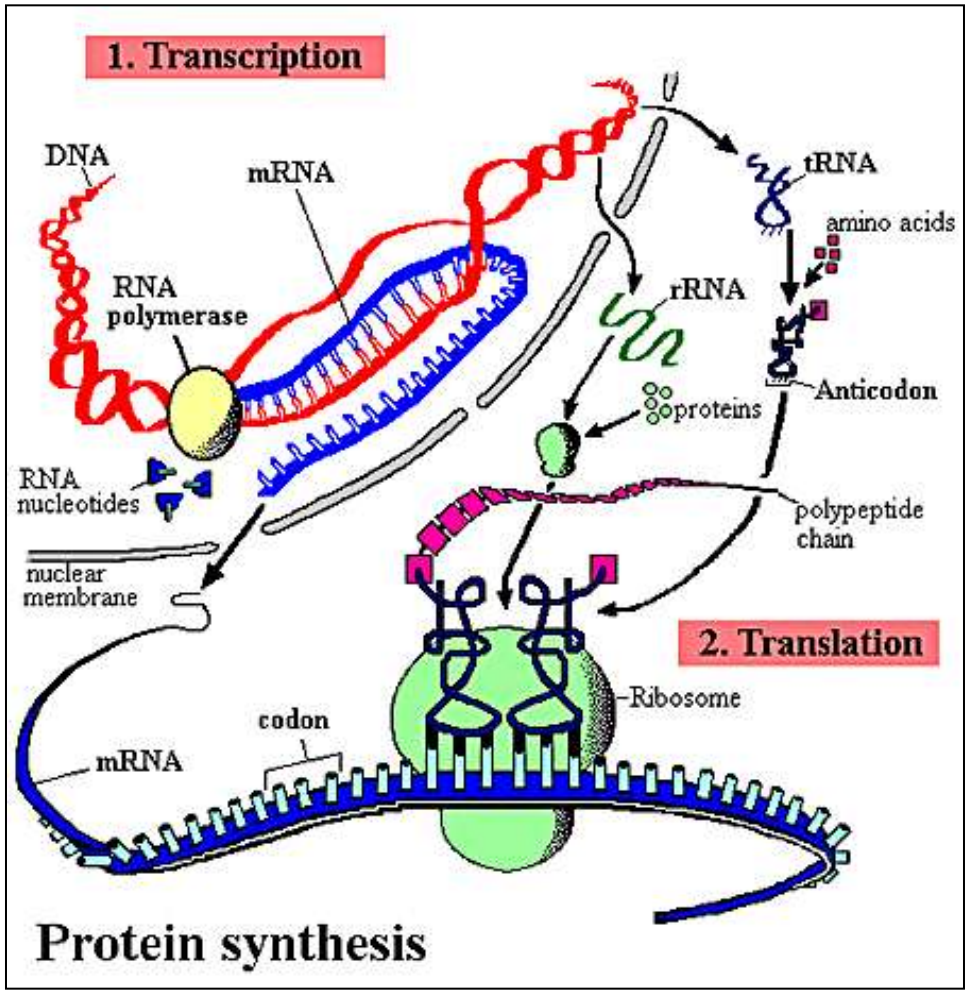
# Gene expression

Gene expression in Eukaryotes involves three main steps:

Transcription (from DNA to mRNA)

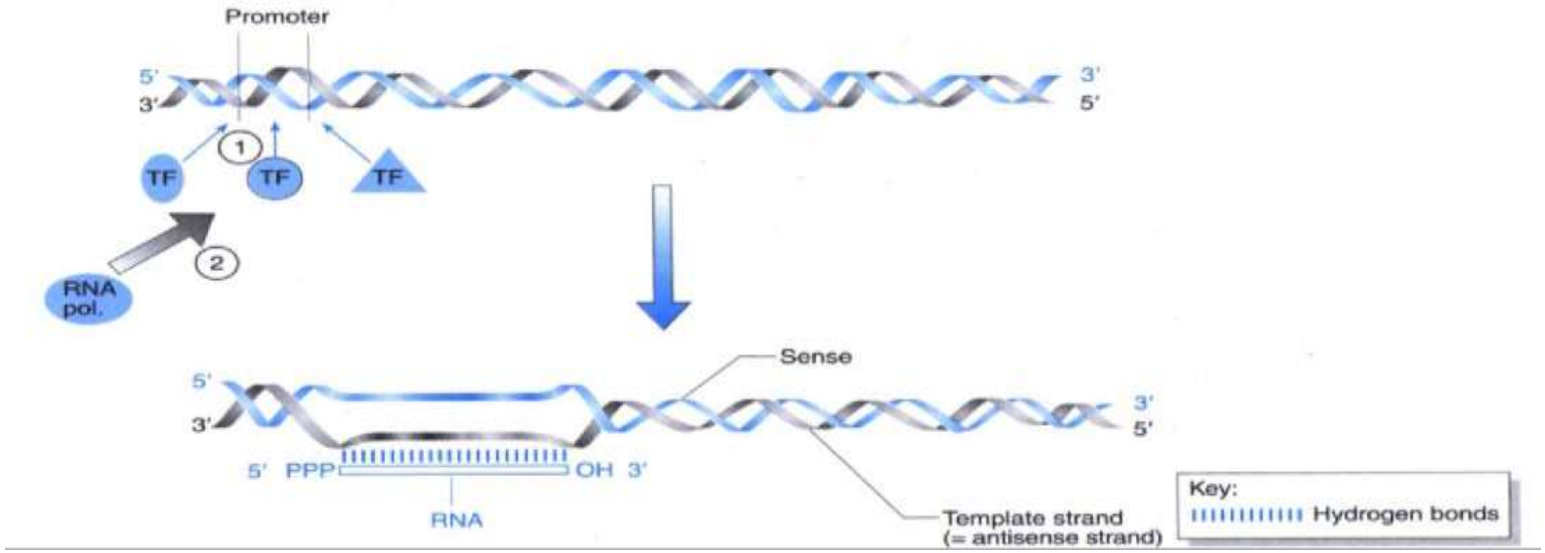
Splicing (mRNA maturation)

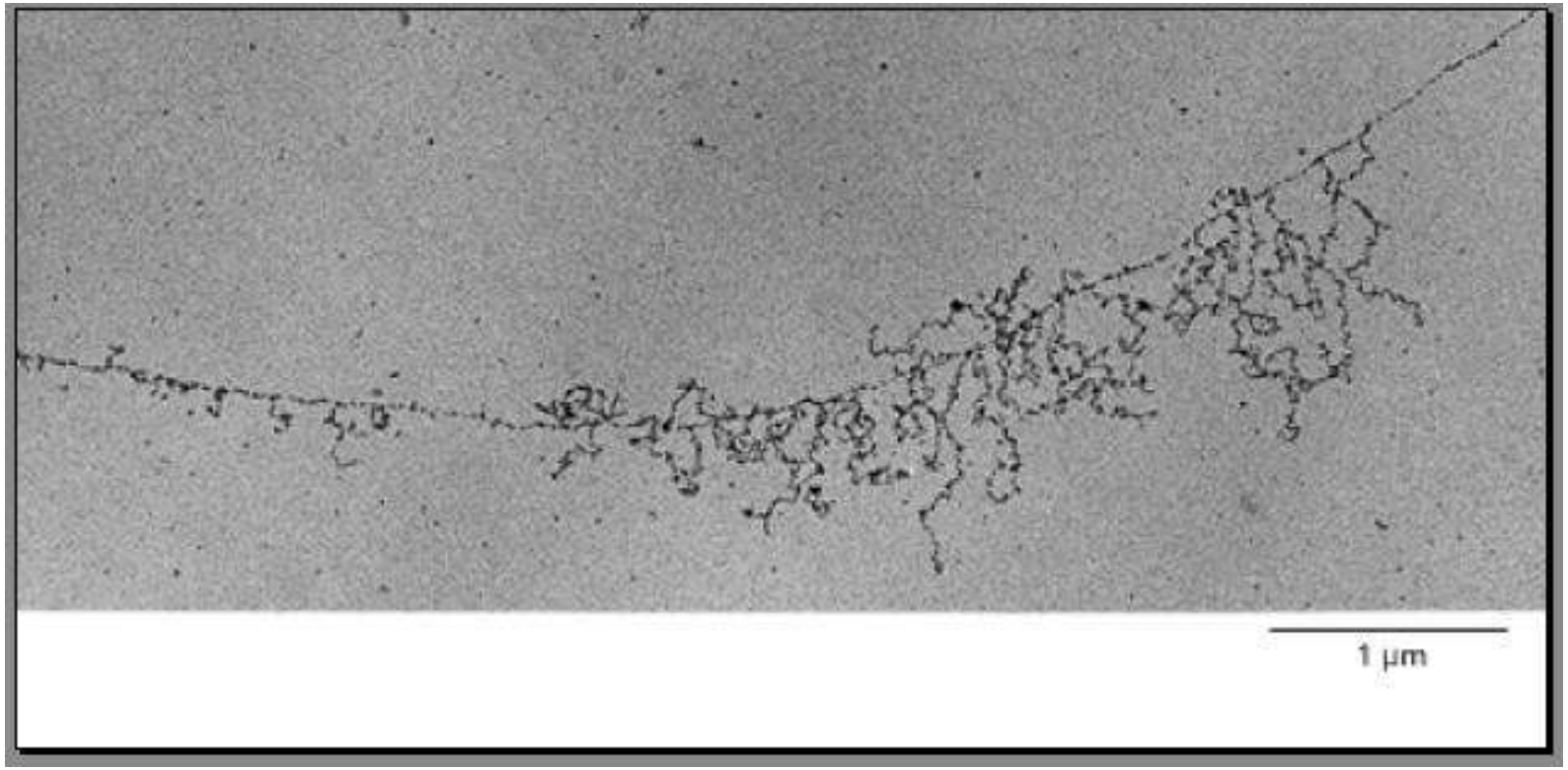
Translation (from mature mRNA to Proteins)



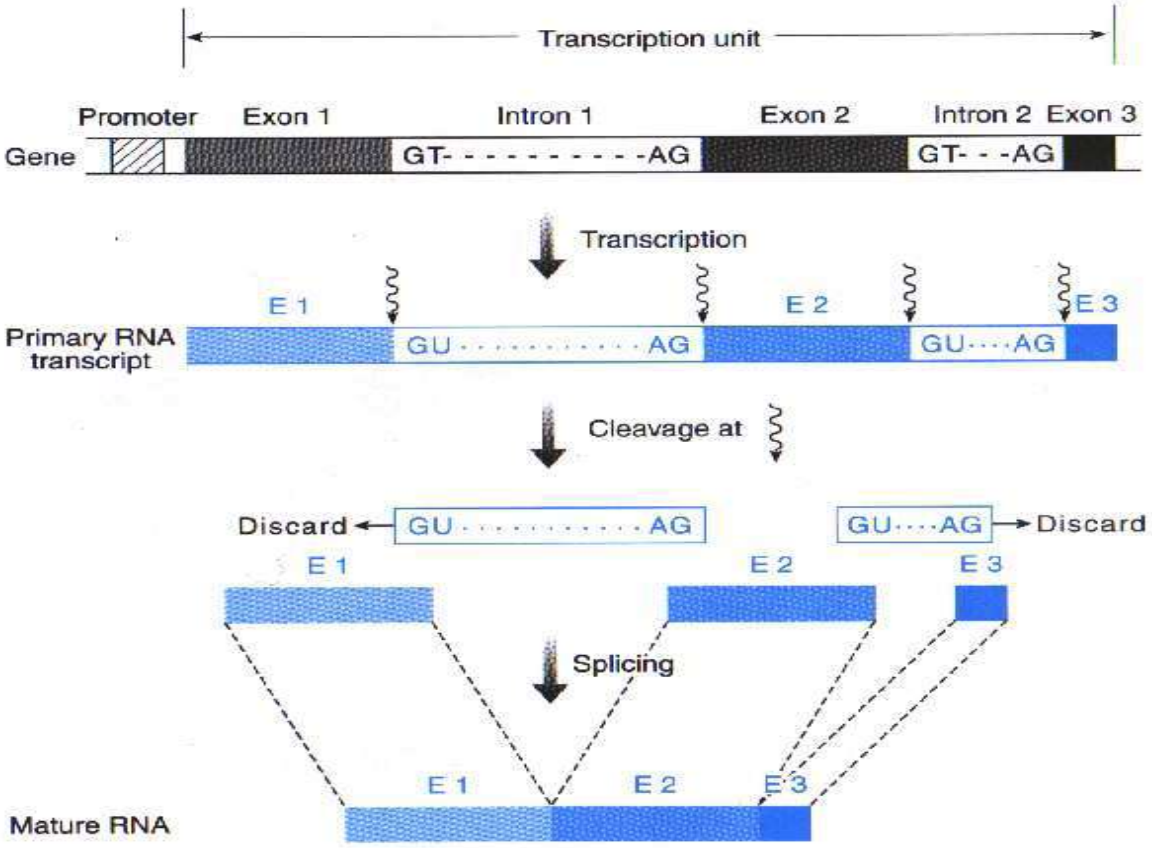
# Transcription

During transcription genetic information in DNA is copied into messenger RNA (**mRNA**).
Transcription begins at the start of the gene in 5′ (the promoter region) and continues until the end of the gene in 3′.
The mRNA sequence is complementary to the DNA template strand it transcribes (except uracile bases that replace the thymine ones)




1 µm

# Splicing


Transcription unit
Promoter
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Gene
GT- - - - - - - - - -AG
GT- - -AG
Transcription
E 1
E 2
E 3
Primary RNA transcript
GU - - - - - - - - - - AG
GU - - - - AG
Cleavage at
Discard
GU - - - - - - - - - - AG
GU - - - - AG
Discard
E 1
E 2
E 3
Splicing
E 1
E 2
E 3
Mature RNA

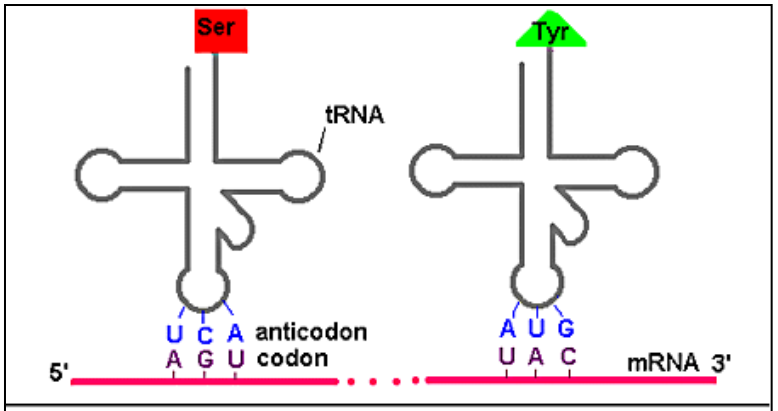
# Translation

During **translation** the mature mRNA is used as a template to synthetize a protein.
- Translation takes place outside the nucleus, in the cytoplasm.
- Proteins are made of aminoacids (20 of them). Three nucleotides (a codon) specify an aminoacid.
- Since there are only 20 aminoacids and 4^3 =64 codons, several codons specify the same aminoacid. The genetic code is degenerate. There are also Start and Stop codons.

| First Position | Second Position of Codon: T | Second Position of Codon: C | Second Position of Codon: A | Second Position of Codon: G | Third Position |
|---|---|---|---|---|---|
| T | TTT Phe [F] | TCT Ser [S] | TAT Tyr [Y] | TGT Cys [C] | T |
| | TTC Phe [F] | TCC Ser [S] | TAC Tyr [Y] | TGC Cys [C] | C |
| | TTA Leu [L] | TCA Ser [S] | TAA *Ter* [end] | TGA *Ter* [end] | A |
| | TTG Leu [L] | TCG Ser [S] | TAG *Ter* [end] | TGG Trp [W] | G |
| C | CTT Leu [L] | CCT Pro [P] | CAT His [H] | CGT Arg [R] | T |
| | CTC Leu [L] | CCC Pro [P] | CAC His [H] | CGC Arg [R] | C |
| | CTA Leu [L] | CCA Pro [P] | CAA Gln [Q] | CGA Arg [R] | A |
| | CTG Leu [L] | CCG Pro [P] | CAG Gln [Q] | CGG Arg [R] | G |
| A | ATT Ile [I] | ACT Thr [T] | AAT Asn [N] | AGT Ser [S] | T |
| | ATC Ile [I] | ACC Thr [T] | AAC Asn [N] | AGC Ser [S] | C |
| | ATA Ile [I] | ACA Thr [T] | AAA Lys [K] | AGA Arg [R] | A |
| | ATG Met [M] | ACG Thr [T] | AAG Lys [K] | AGG Arg [R] | G |
| G | GTT Val [V] | GCT Ala [A] | GAT Asp [D] | GGT Gly [G] | T |
| | GTC Val [V] | GCC Ala [A] | GAC Asp [D] | GGC Gly [G] | C |
| | GTA Val [V] | GCA Ala [A] | GAA Glu [E] | GGA Gly [G] | A |
| | GTG Val [V] | GCG Ala [A] | GAG Glu [E] | GGG Gly [G] | G |

# Genetic Code

**Genetic code** is the rule which allows to translate the 4 symbols alphabet of **DNA** to the 20 symbols one of **proteins.**



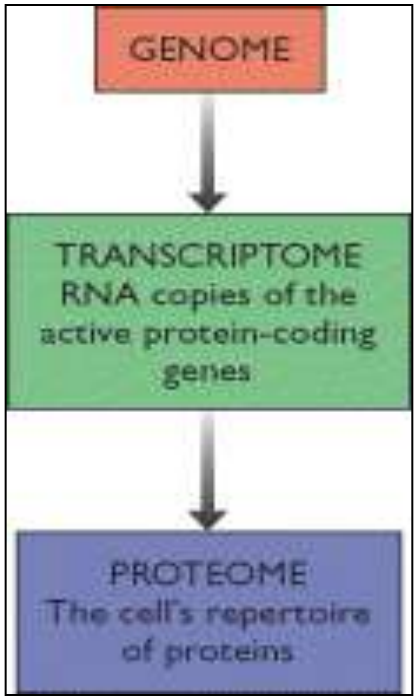

2nd base in codon

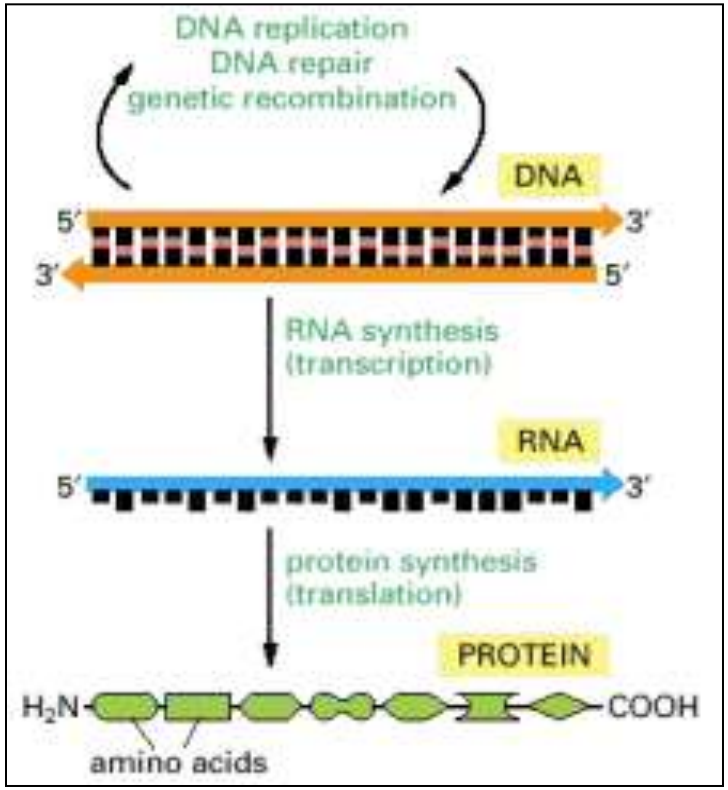
| 1st base in codon | U | C | A | G | 3rd base in codon |
|---|---|---|---|---|---|
| U | Phe Phe Leu Leu | Ser Ser Ser Ser | Tyr Tyr STOP STOP | Cys Cys STOP Trp | U C A G |
| C | Leu Leu Leu Leu | Pro Pro Pro Pro | His His Gln Gln | Arg Arg Arg Arg | U C A G |
| A | Ile Ile Ile Met | Thr Thr Thr Thr | Asn Asn Lys Lys | Ser Ser Arg Arg | U C A G |
| G | Val Val Val Val | Ala Ala Ala Ala | Asp Asp Glu Glu | Gly Gly Gly Gly | U C A G |

The Genetic Code

# Information flow in the cell

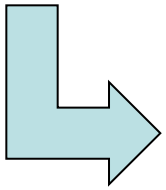
**"Central Dogma"** of molecular biology



# The Genomic Revolution

Started at the end of ’90, triggered by



Impressive technological improvements:

high-throughput experiments

- massive sequencing projects
- microarray
- proteomics
- world wide SNP studies

## A central role in this revolution was played by physics.

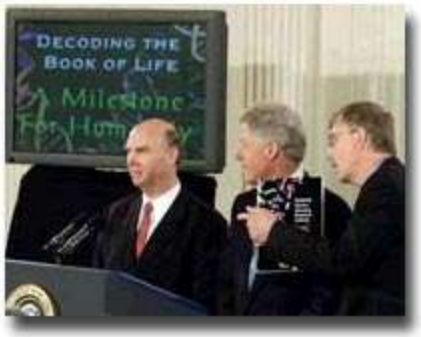
Both on the experimental side:
- nanotechnolgy
- microfluidics

And on the theoretical side:
- new inference methods
- modeling of complex systems
- network theory
- alignment tools

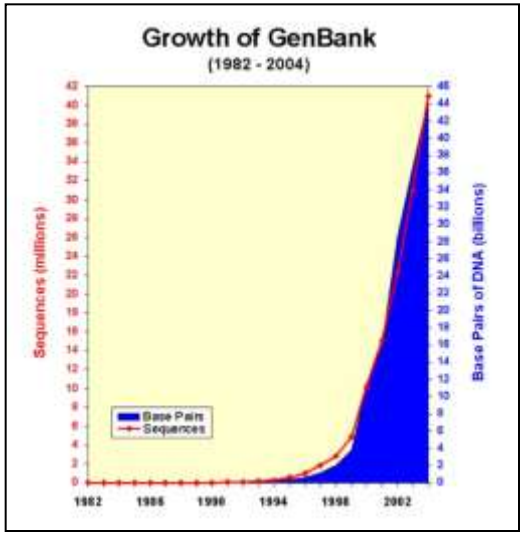
# Genomic Revolution: *sequences*

- Automatic sequencing of DNA
- Open access information: GenBank
- Sequencing projects for
  thousand of different organisms
  (and individuals)

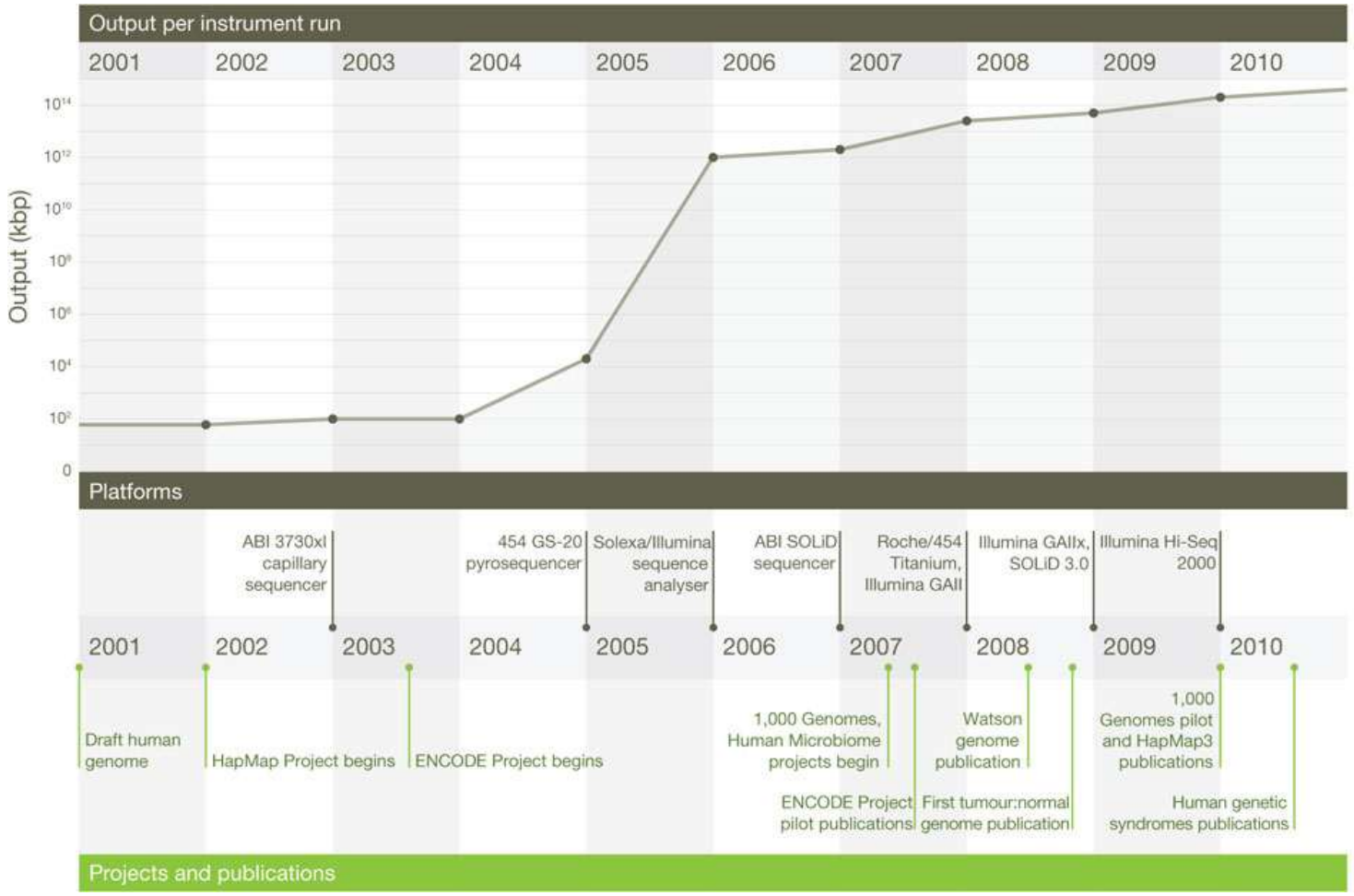
> *homo_sapiens*
ACTTTTTTACCCTCGTGTGTTGC
AGACTTTTTGCCACTTTTAAAAC
GCTGACAATTCGACCCTTTCCAA
GTGCAAAAAGTGCCAAGATTTA
CGATAAAATTCCCCCGAGAGAC
GTGTGCA……



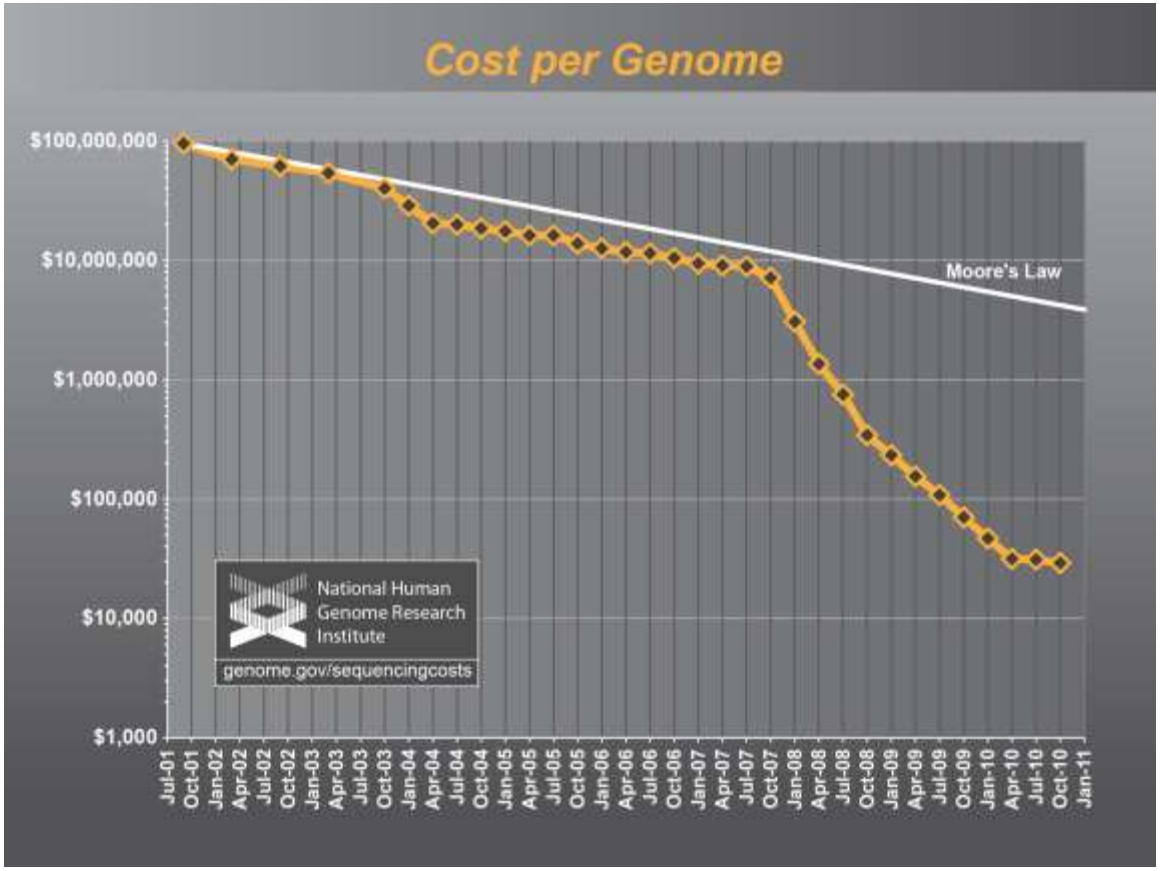


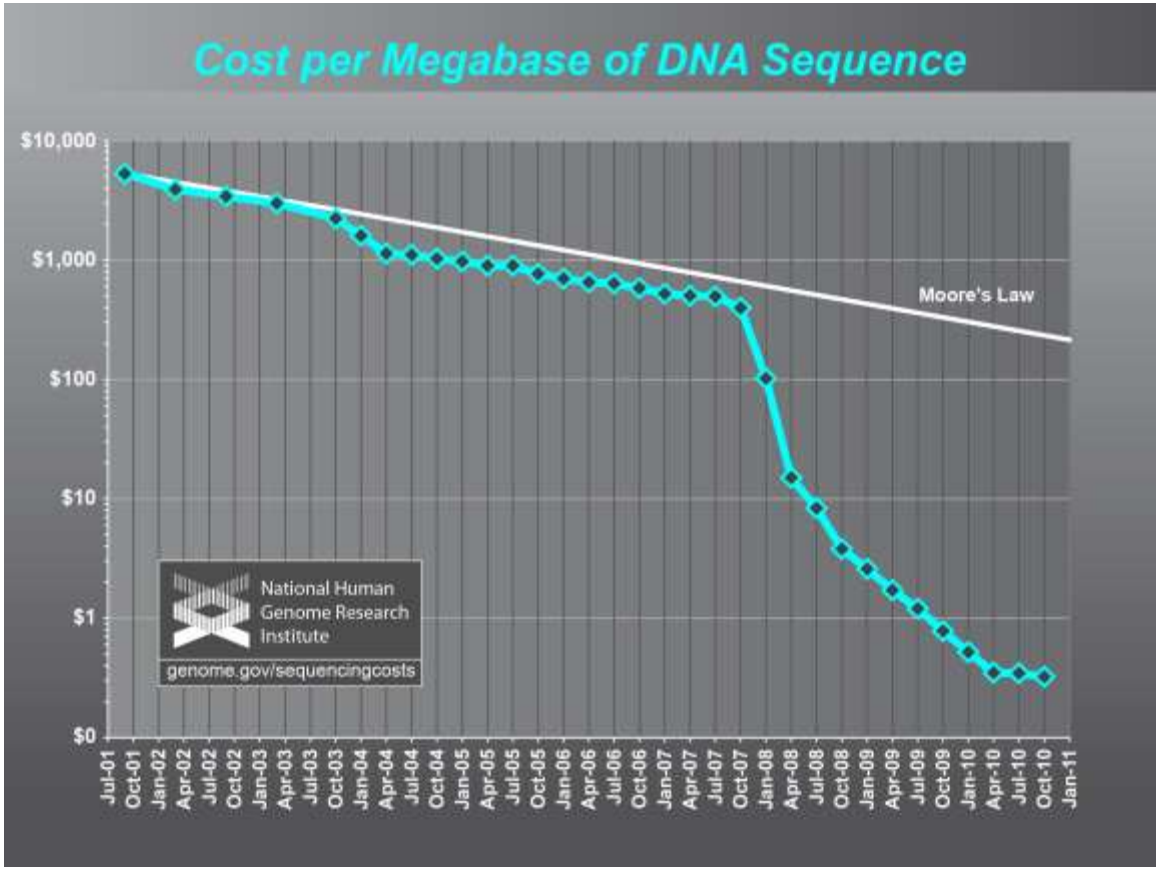


# Changes in instrument capacity over the past decade



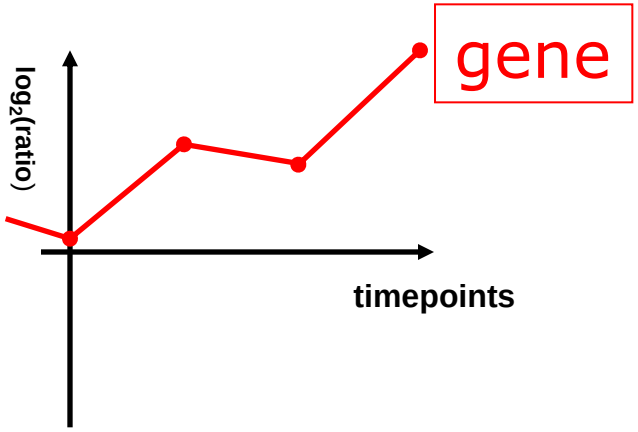

Timing of the major sequencing projects

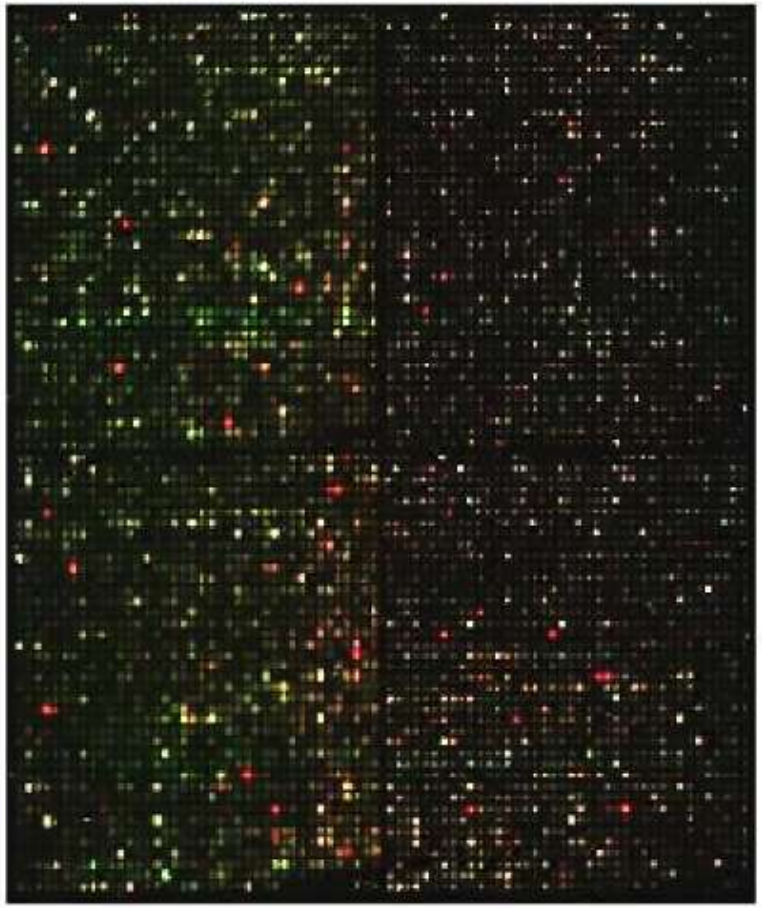

Cost per Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
Jul-01
Oct-01
Jan-02
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11


Cost per Megabase of DNA Sequence
$10,000
$1,000
$100
$10
$1
$0
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
Jul-01
Oct-01
Jan-02
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11

# Transcriptomics: *microarray*

- A typical microarray experiment measures the expression level (amount of mRNA in the cell) of thousands of genes in a single run .

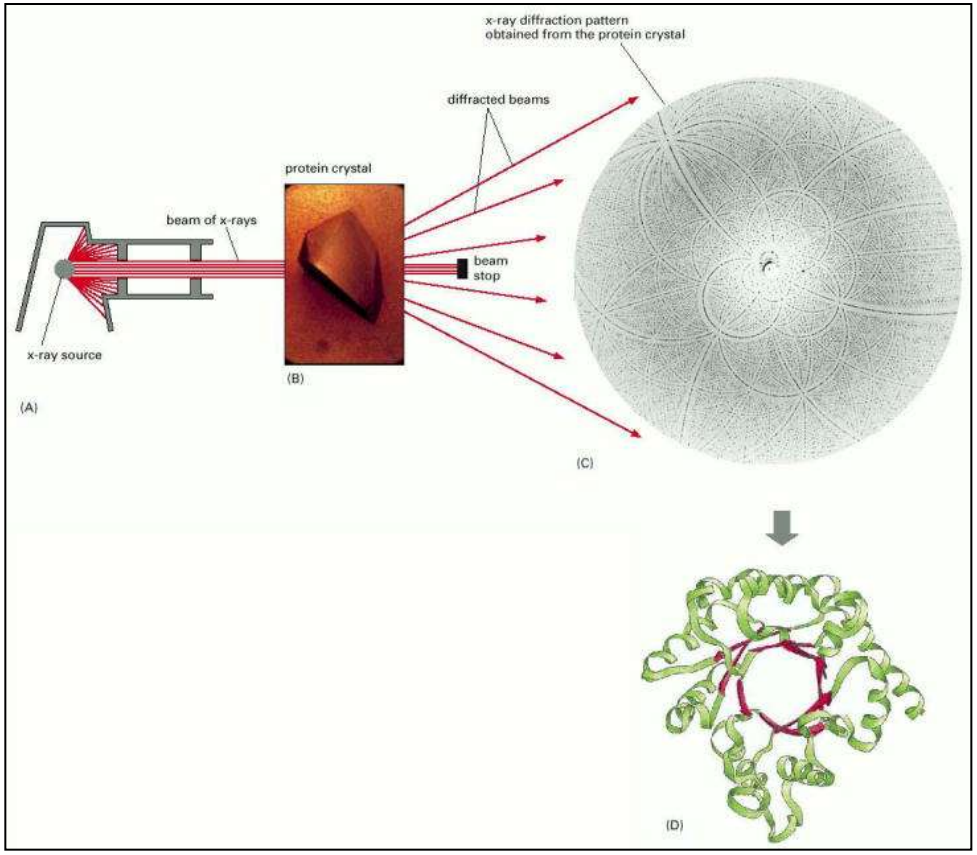


# Proteomics:

- Systematic study of 3D protein structure using X-ray spectroscopy

- Sistematic study of protein interactions.



# New questions, new ideas

- How is it organized the Genome?
- How many genes do we have?
- Which is the role of **non coding DNA**?
- How different are humans and chimps ?
- Where is it hidden the impressive complexity of multicellular organisms?

# New Theoretical Tools:

# Systems biology and Computational Biology

# Computational Biology

With the terms "Computational Biology" or "Bioinformatics"
one usually refers to all the data mining tool based
on methods and ideas coming from
mathematics / physics / statistics / computer-science .

Genomic data (both sequences and annotations)
Can be easily downloaded from huge "open access" data banks.

These data contain a lot of hidden information.
In general only a fraction of it has been recognized and published
by the authors of the experiments.

Relevant original results can be obtained with no need of new
costly experiments but simply using in a clever way existing data.

# Systems Biology

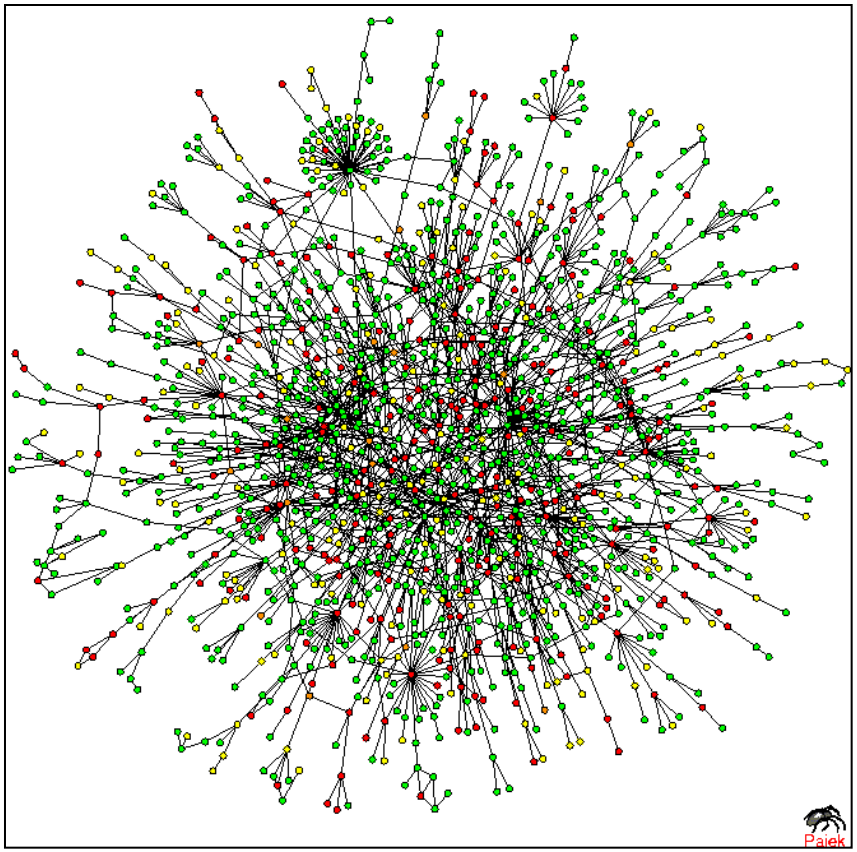
Network theory: Complex functions, must be described at the network level and not at the level of single genes, proteins or neurons.

Modeling: These networks can be decomposed in elementary circuits. ("network motifs") which may be modeled using differential or stochastic equations.

Ontologies: biological (and medical) information must be organized in a quantitative and standardized way

# Modern Genomics: *networks*

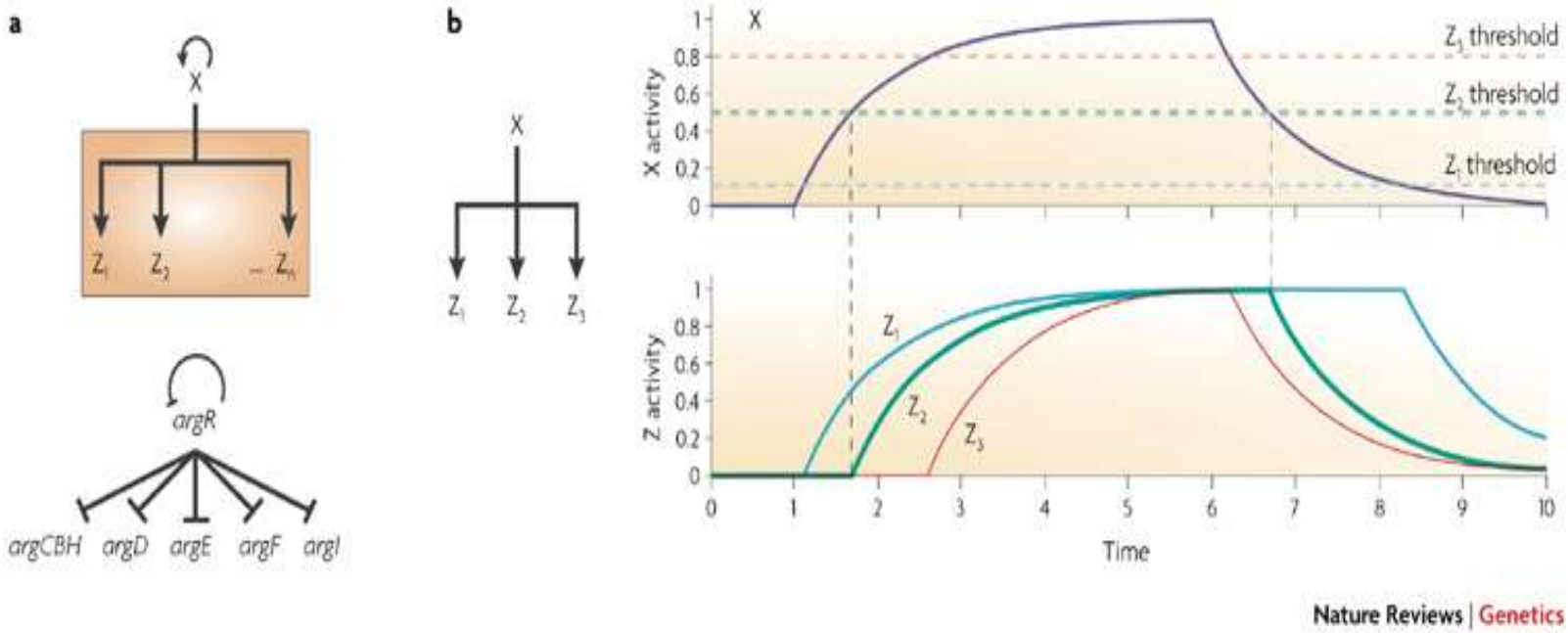
- genes and proteins of a given organism are organized in networks .
- Cells react to external stimuli in a "global" way.




H.Jeong et al.
Nature, 411 (2001) 41

# Network motifs

Example: SIM (Single Input Module) (a) experimental realization: arginine biosynthesis b) Circuit behaviour: different genes are activated at different times as a function of their different activation threshold as the concentration of X (master regulator) changes in time R.Milo et al. Science 298 (2002) 824



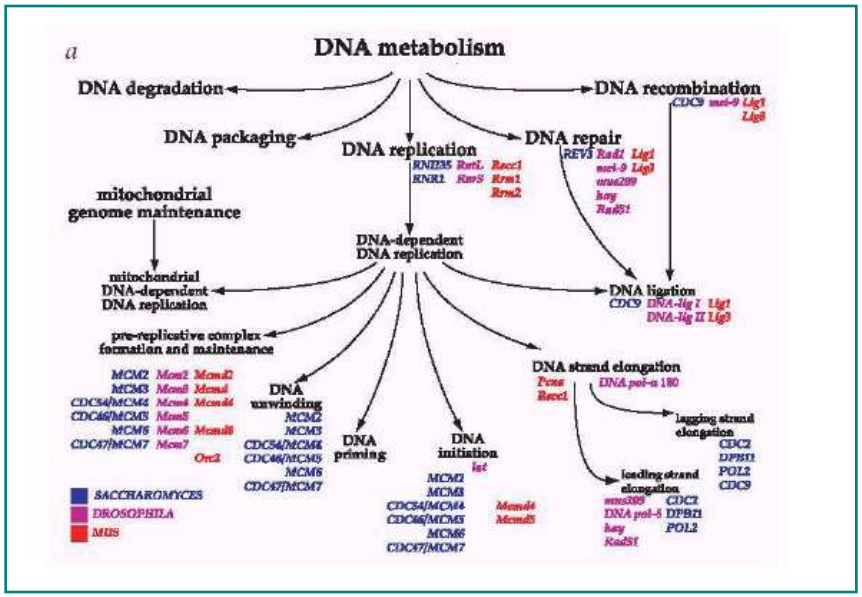
# Modern Genomics: *Gene Ontology*

- **Gene Ontology** is an example of standardization of biological data.

- The goal is the construction of a controlled vocabulary to describe:
  - Molecular function
  - Biological process
  - Cellular component

  of a given gene.

- The ontologies are organized as hierarchical networks (Directed acyclic graphs)

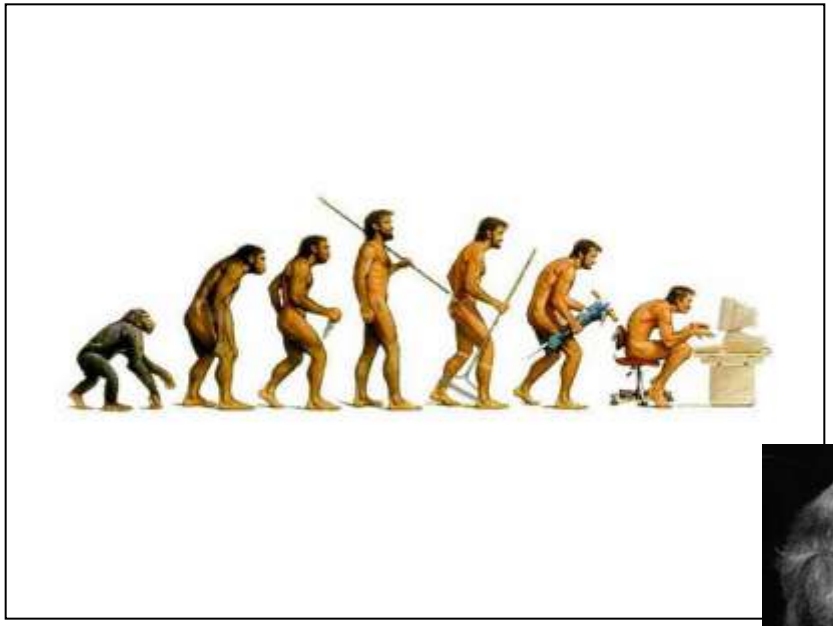


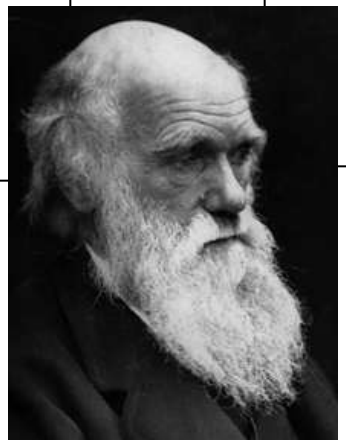

The G.O. Consortium
Nature Genet. 25 (2000) 25

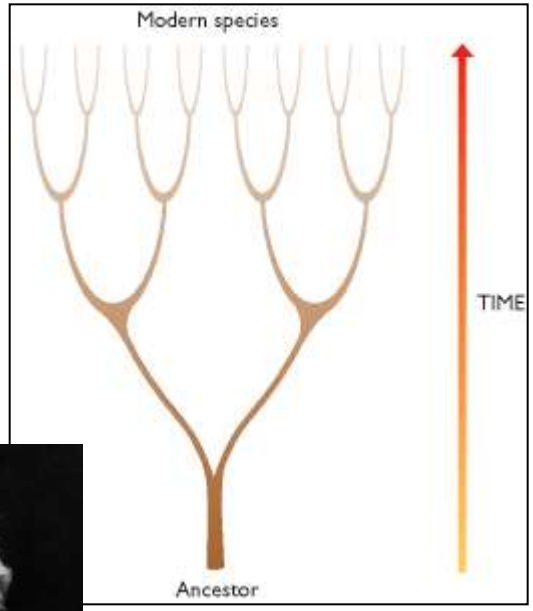
# Three examples of applications

- Evolutionay models
- Gene Regulation
- Chemotaxis

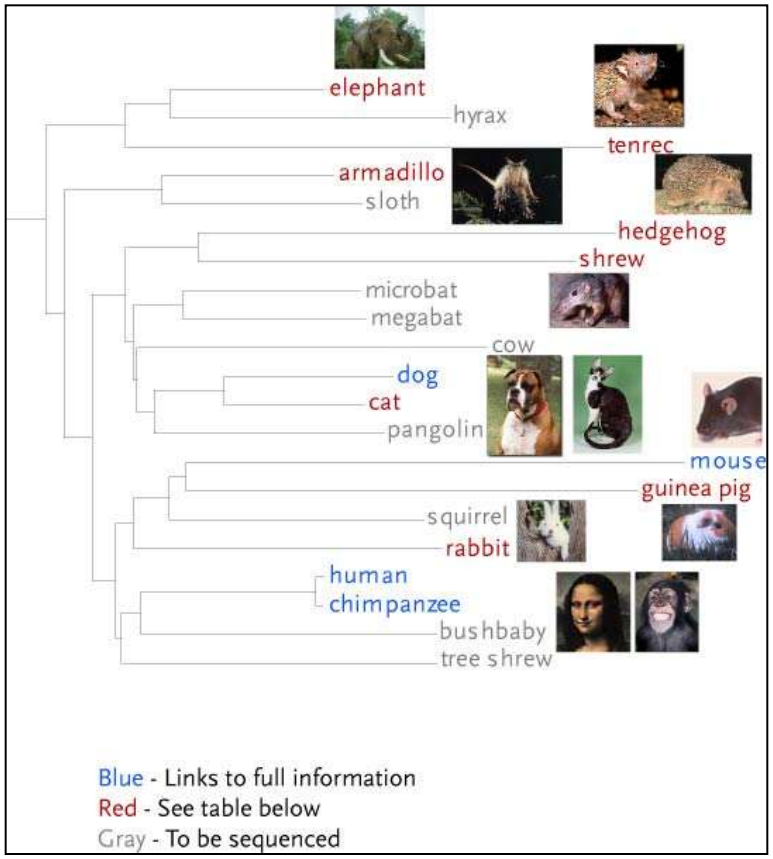
# Evolutionary models



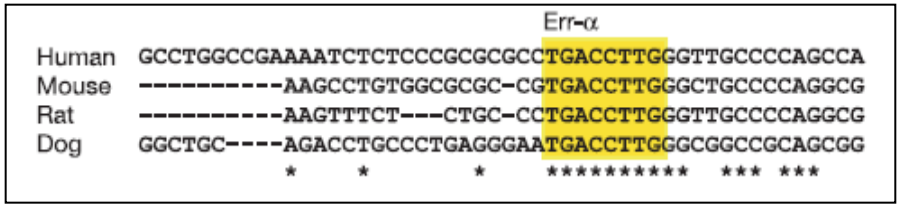




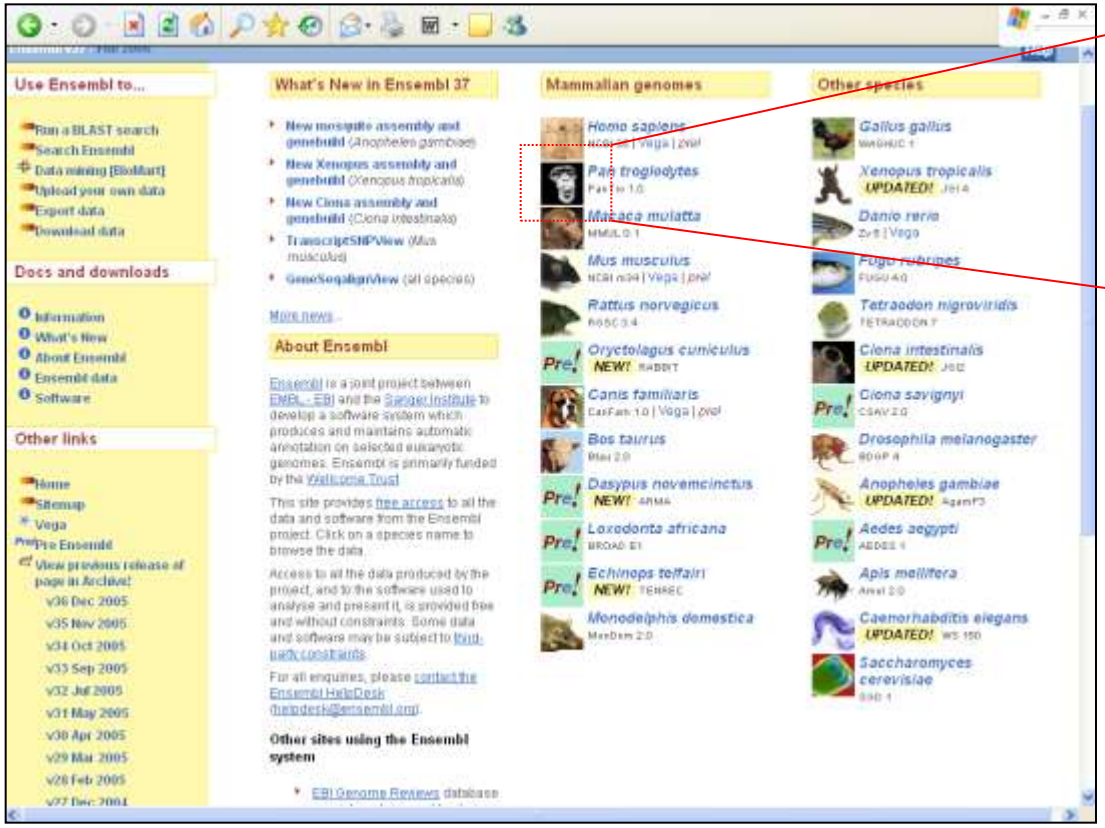
# Taxonomic versus Genomic trees




Genomic trees may be obtained using alignment algorithms. They are impressively similar to the taxonomic trees. This is a highly non trivial test of Evolution theory.



# Human and Chimps





96% of the human genome coincides with the chimp's one! Most of the differences are non-coding!

# Evolution and gene regulation

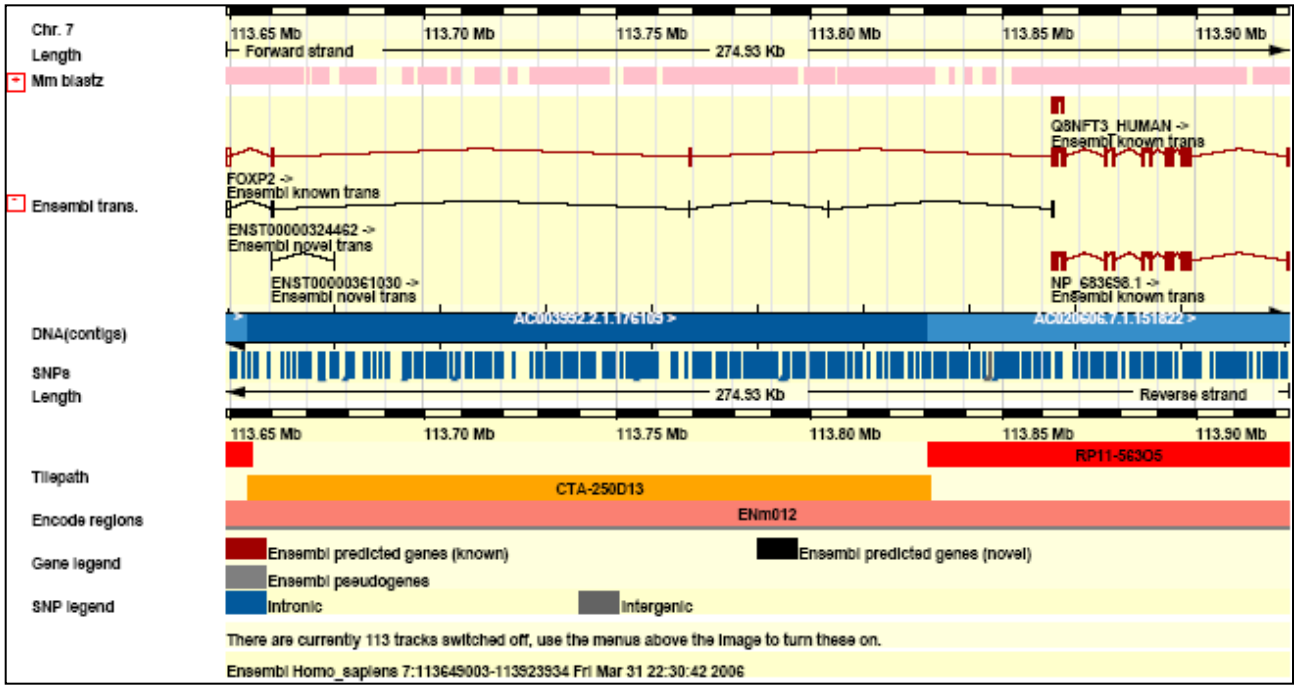
- Goal: use evolutionary conservation to identify functionally important regions of the genome. Different regions show different levels of conservation

"Ultraconserved regions" have been protected against mutations for hundreds of millions of years. They are likely to be crucially important regulatory regions.

One of these appears to be mutated in the human gene FOXP2.

# FOXP2 !!

Mutations (SNPs) in the FOXP2 gene are associated to deep alterations in speaking ability.

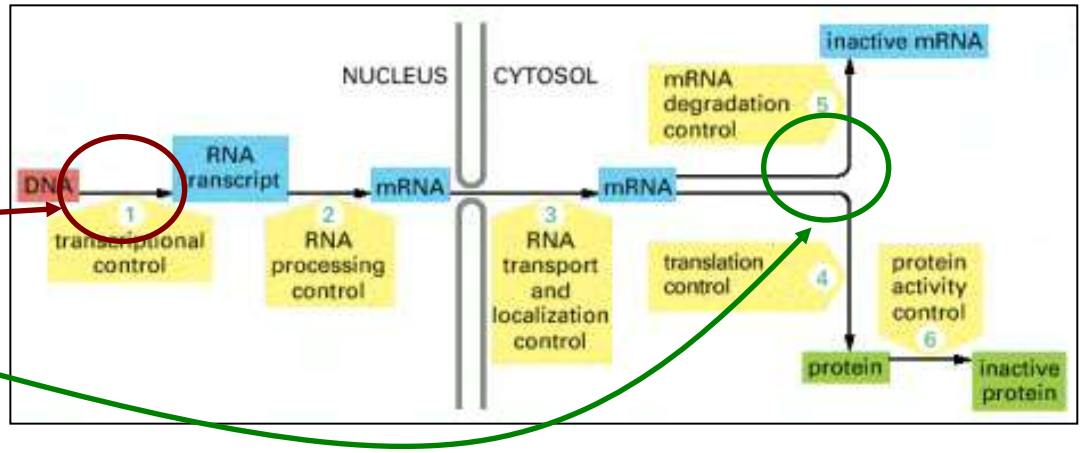


# Gene Regulation

Gene expression is tightly regulated. All cells in the body carry the full set of genes, but only express about 20% of them at any particular time. Different proteins are expressed in different cells (neurons, muscle cells....) according to the different functions of the cell.

Among the various regulatory steps the most important ones are:

- transcriptional control, by **Transcription Factors.**
- post-transcriptional control, by **microRNAs.**

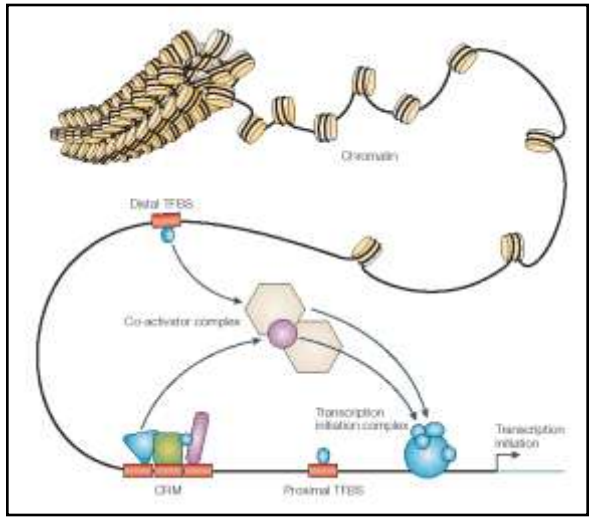



Alberts, *Molecular Biology of the Cell*

# Transcription Factors and miRNAs

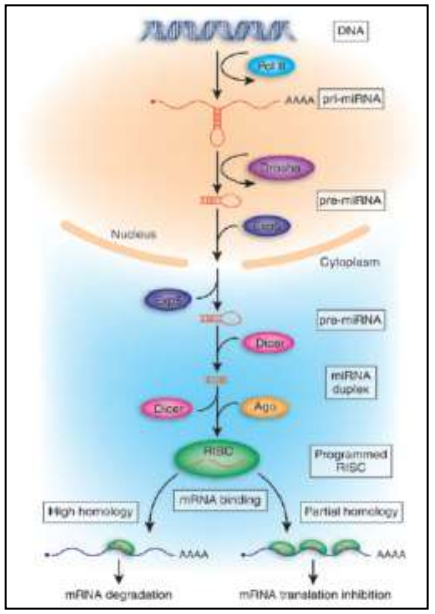
- **Regulation of gene expression** mainly mediated by:

**Transcription Factors (TFs)**: proteins binding to specific recognition **motifs (TFBSs)** usually short (5-10 bp) and located **upstream** of the coding region of the regulated gene.

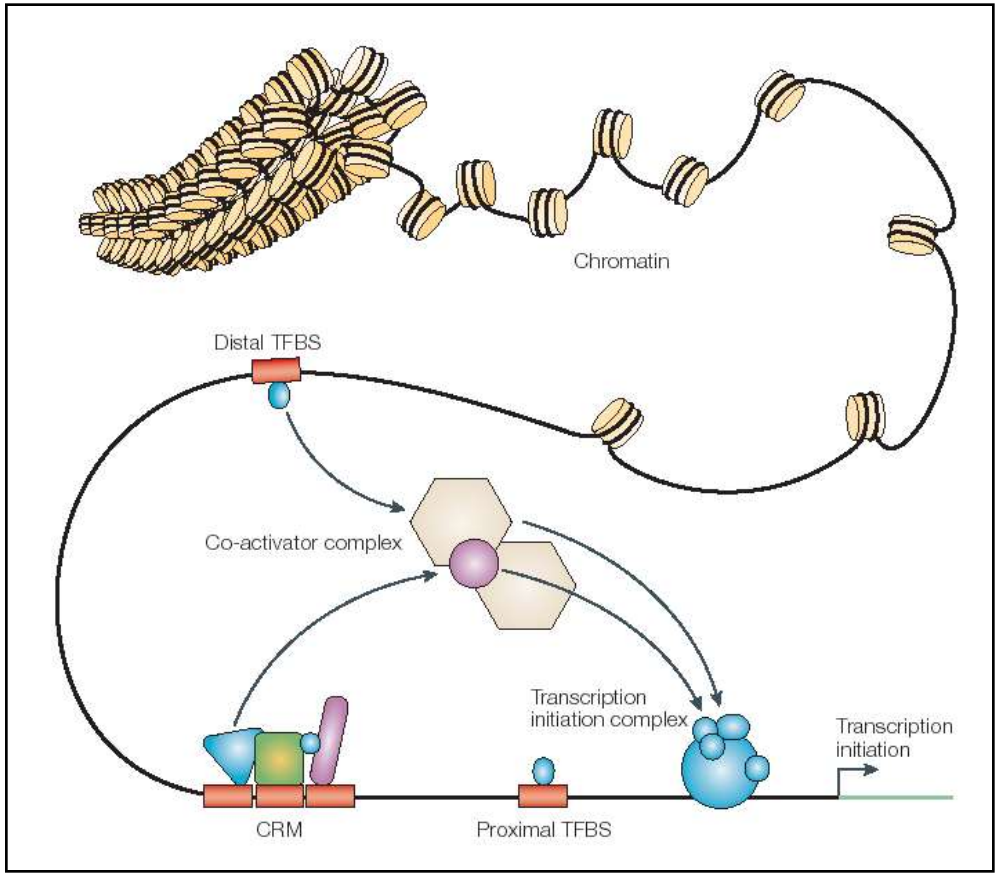



*Wassermann*, Nat. Rev. Genetics

**MicroRNAs (miRNAs)** are a family of small RNAs (typically **21 - 25** nucleotide long) that **negatively regulate gene expression at the posttranscriptional level**, (usually) thanks to the "seed" region in 3'-UTR regions.



# Transcription Factors



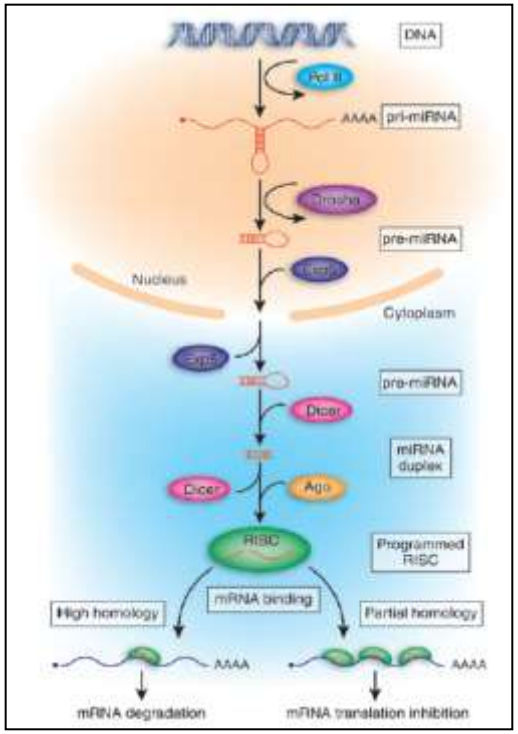
# MicroRNA biogenesis

**MicroRNAs (miRNAs)** are a family of small RNAs (typically **21 - 25** nucleotide long) that negatively regulate gene expression at the post-transcriptional level.

MiRNAs derive from larger precursors transcribed from genomic DNA

- MiRNA transcripts (pri-miRNA) are processed into ~100 nucleotide precursors (pre-miRNA) by Drosha.
- cleavage of the precursors generate 21 - 25 nucleotide mature miRNAs in cytoplasm.
- mature miRNAs couple with a special protein complex called RNA-Induced Silencing Complex (RISC).

miRNAs are able to negatively affect the expression of a "target" gene via mRNA cleavage or translational repression, after **antisense complementary basepair** matching to specific target sequences in the 3'-UTR of the regulated genes (the "**seeds**").




*He L. , Hannon GJ.* Nature Review Genetics 5, 522 - 531 (2004)

# MicroRNA: functions

Members of the miRNA family were initially discovered as small temporal RNAs that regulate developmental transitions in Caenorhabditis Elegans (lin-4). (Chalfieet al. 1981; Lee et al. 1993) but considered only as a peculiarity of worms. In 2002-2003 it was suddenly realized that miRNA exist in all higher Eukaryotes in several copies and that they play an essential role in development and differentiation of tissues.

The functions in which miRNAs are involved are extremely wide and, in animals, they include: developmental timing, pattern formation and embryogenesis, differentiation and organogenesis, growth control and cell death.

# MicroRNA: evolution

MiRNAs also show interesting evolutionary properties between different species. Up to one third of the miRNAs discovered in C. elegans have an orthologous in human.

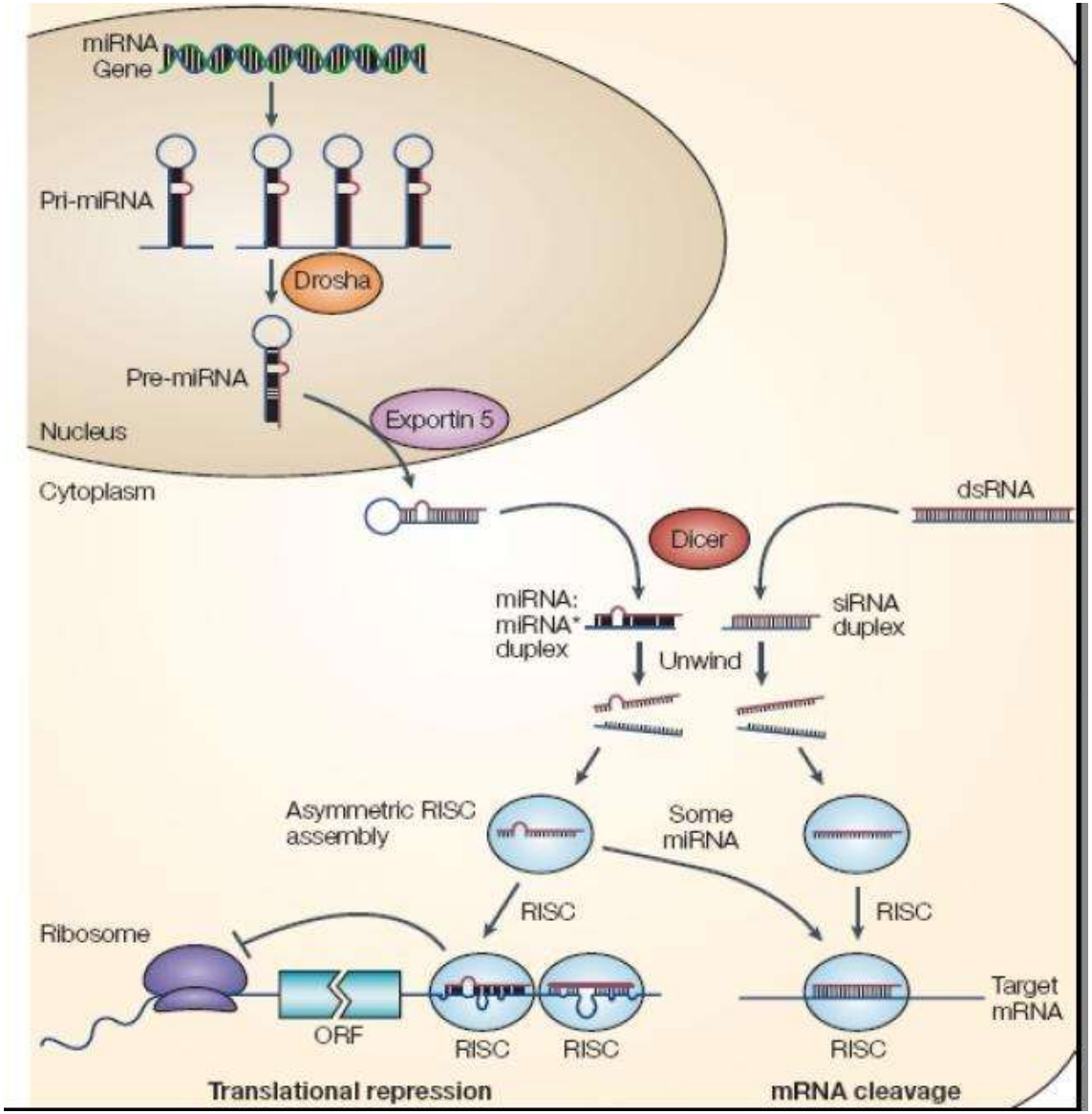
Tracing back this evolutionary pattern it is possible to guess that miRNA appeared as a new regulatory mechanism about 500 Myears ago. It is interesting to observe that this time scale almost coincides with the impressive explosion of new species in the Cambrian age and with the almost simultaneous appearance of retrotransposons in Eukaryotes.

# MicroRNAs as regulatory genes

MiRNAs expression is regulated by the same TF which regulate all the other genes

Regulation by miRNAs is a combinatorial process. Each miRNA is expected to control from one to hundreds of targets while a given mRNA can be under control of many different miRNAs. Usually miRNA binding sites are overrepresented in the 3'-utr sequence of target genes.

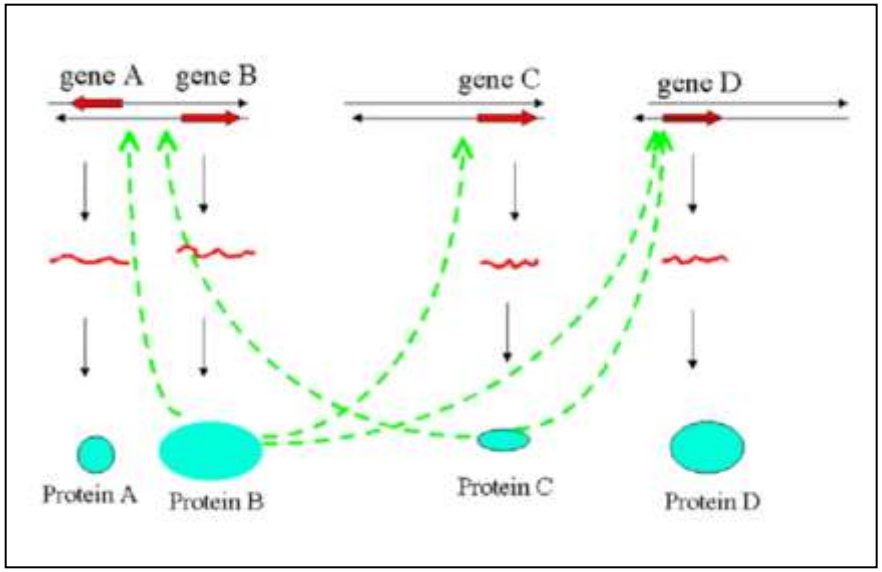
Transcription Factors and miRNAs share a very similar behaviour. The main difference between the two is that while TF act as a sort of on/off switch, it seems that the miRNA role is to fine tune the gene expression.

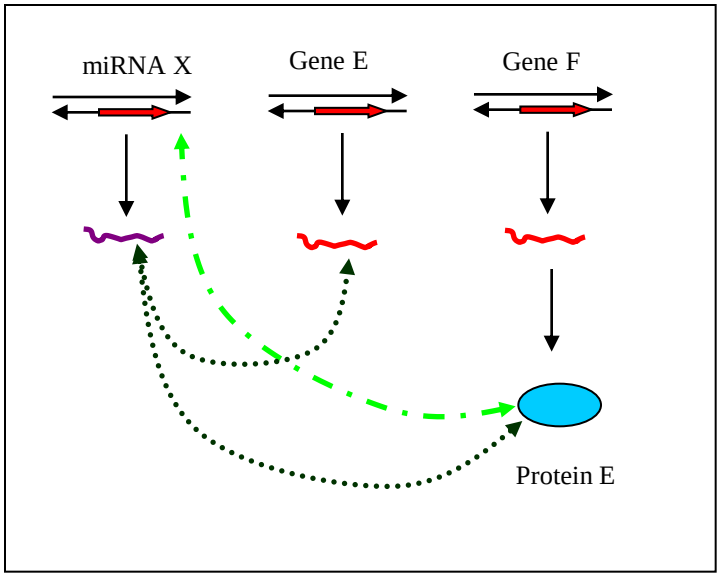

miRNA Gene
Pri-miRNA
Drosha
Pre-miRNA
Nucleus
Exportin 5
Cytoplasm
dsRNA
Dicer
miRNA: miRNA* duplex
siRNA duplex
Unwind
Asymmetric RISC assembly
Some miRNA
RISC
RISC
Ribosome
ORF
RISC
RISC
RISC
Target mRNA
Translational repression
mRNA cleavage

# Regulatory Networks 1

**Key 1 -->** **TFs** are themselves proteins produced by other genes, and they act in a combinatorial way, resulting in a complex network of interactions between genes and their products.
**--> Transcriptional Network**

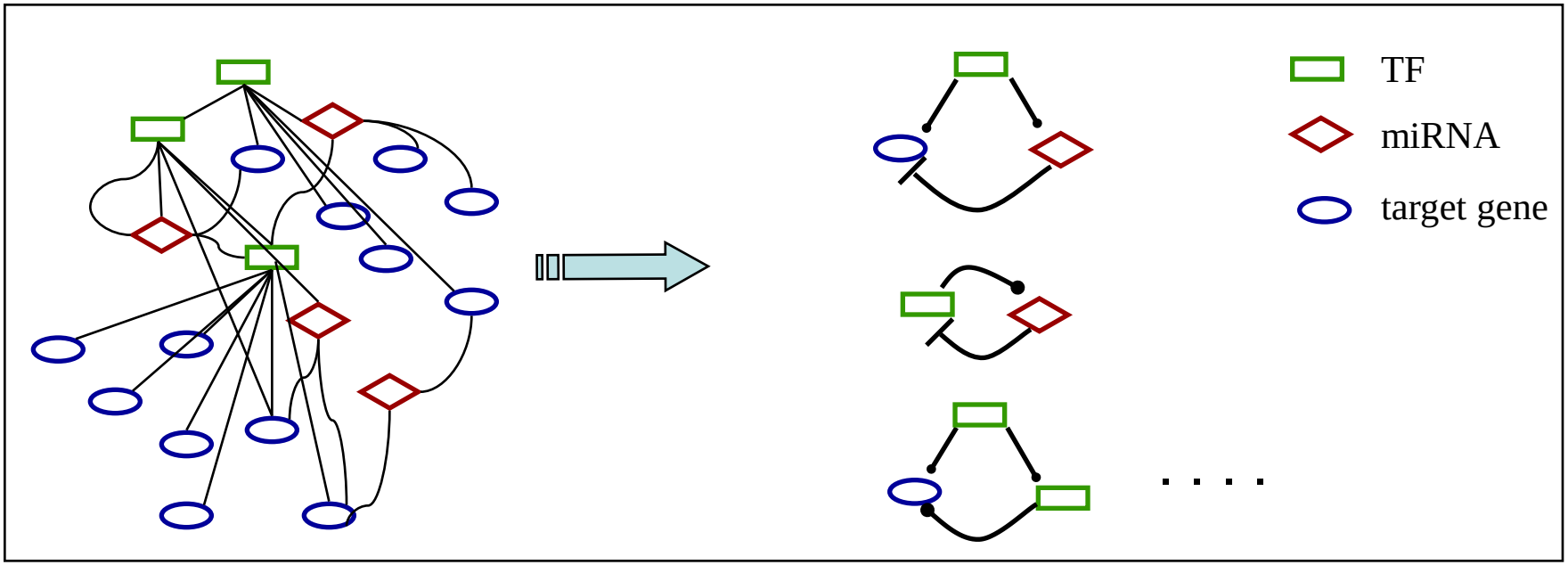
**miRNAs** also act in a combinatorial and one-to-many way, and, moreover, are transcribed from same POL-II promotes of TFs.
**--> Post-Transcriptional Network**



# Regulatory Networks 2

**Key 2 -->** Biological functions are performed by groups of genes which act in an interdependent and synergic way. A complex network can be divided into simpler, distinct regulatory patterns called **network motifs**, typically composed by 3 or 4 interacting components which are able to perform elementary signal processing functions.

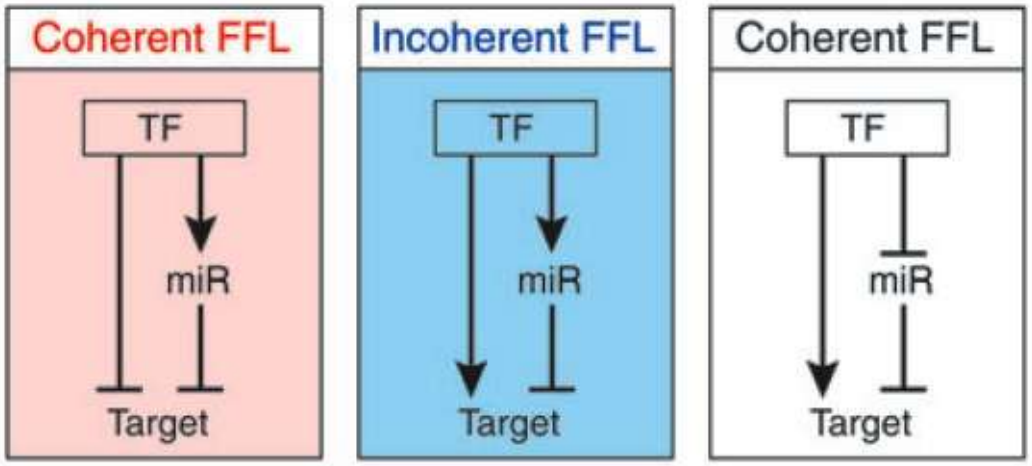


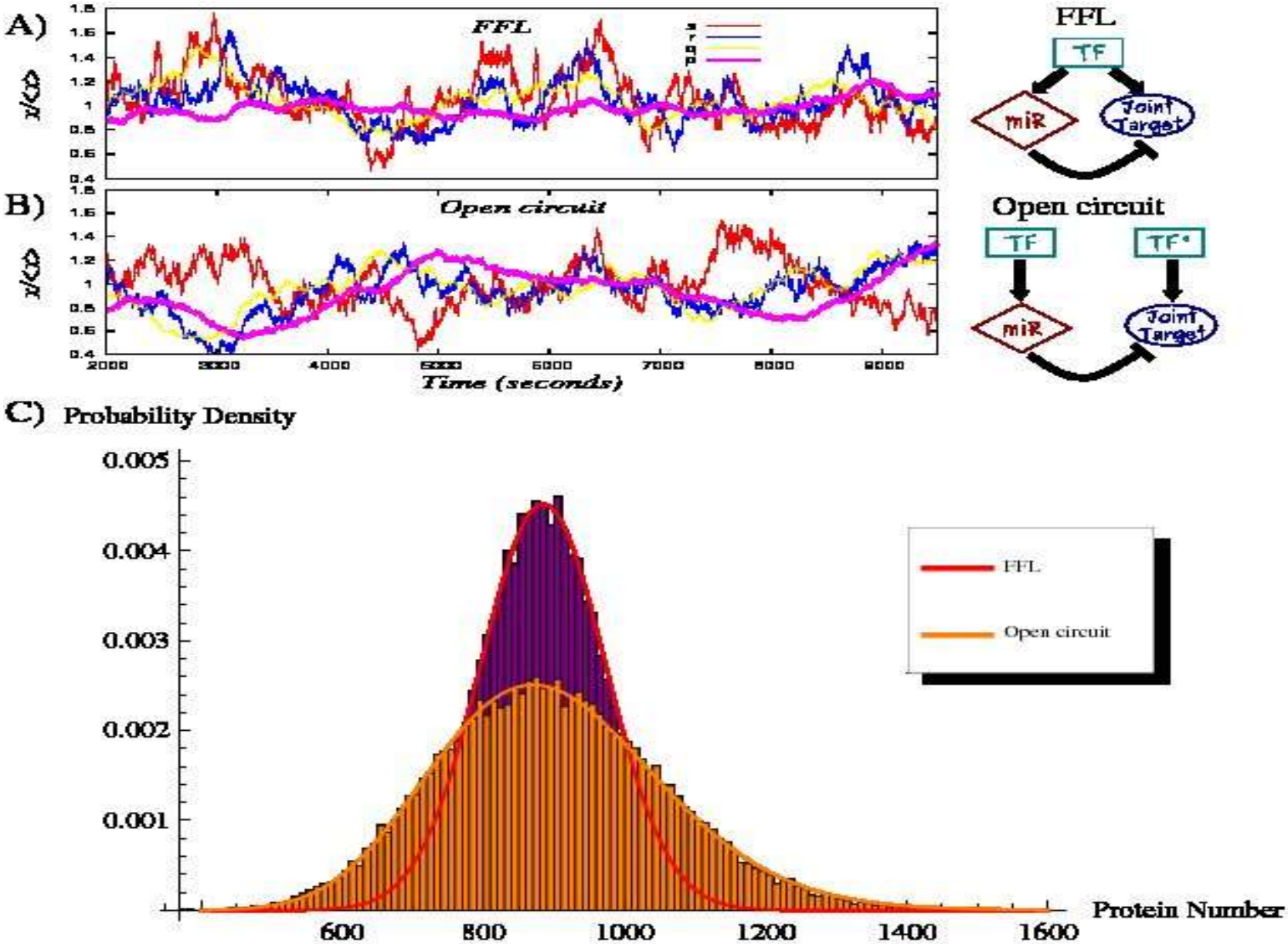
# Network Motifs I

Network motifs can be studied using standard tools of theoretical physics:
- Ordinary differential equations
- Stochastic equations
- Montecarlo (Gillespie) simulations.

- Goal: understand the functional role of the motif and why it was selected by evolution

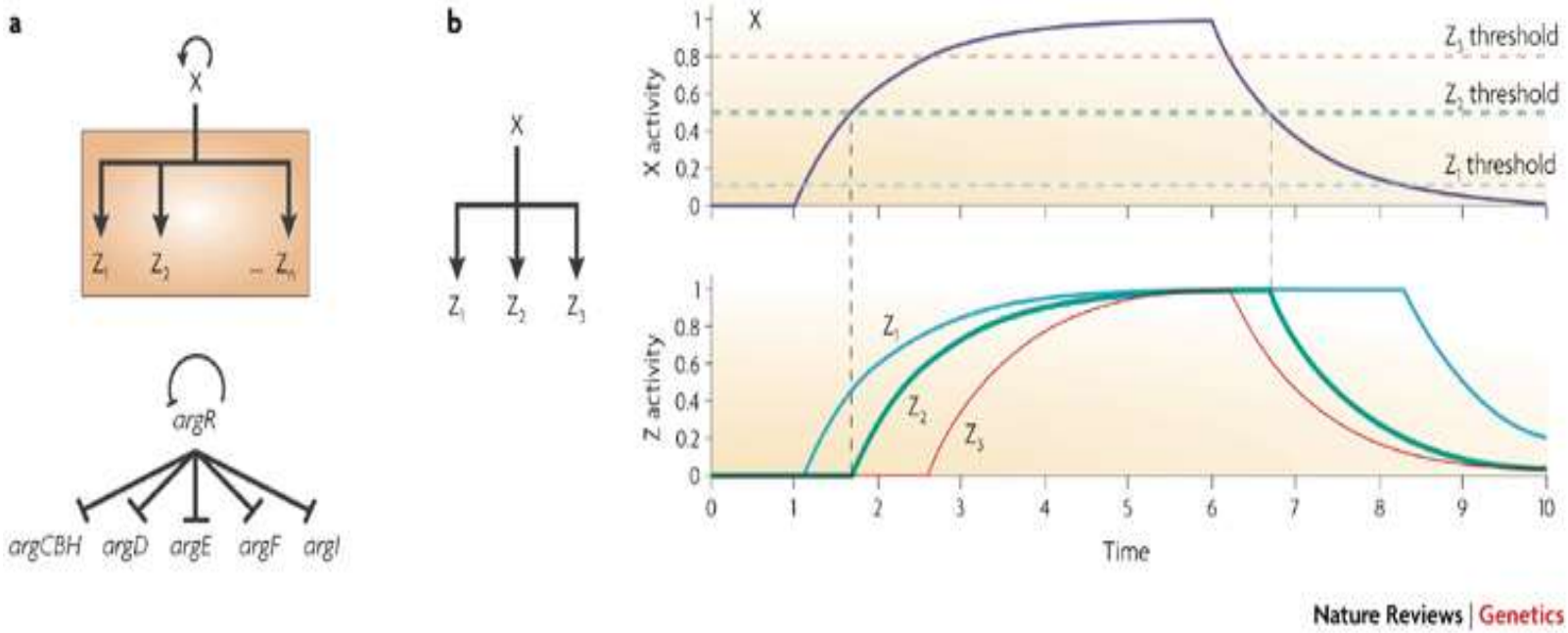
- Example 1: incoherent feedforward loops can reduce the noise in the amount of produced proteins.




A)
FFL
x/<x>
B)
Open circuit
x/<x>
Time (seconds)
FFL
TF
miR
Joint Target
Open circuit
TF
TF*
miR
Joint Target
C) Probability Density
0.005
0.004
0.003
0.002
0.001
600
800
1000
1200
1400
1600
Protein Number
FFL
Open circuit

# Network motifs II

Example 2: SIM (Single Input Module) (a) experimental realization: arginine biosynthesis b) Circuit behaviour: different genes are activated at different times as a function of their different activation threshold as the concentration of X (master regulator) changes in time R.Milo et al. Science 298 (2002) 824



# Chemotaxis

Chemotaxis is the process which allows eukaryotic cells to identify and follow spatial gradients of extracellular guidance cues (chemoattractors)

Chemotaxis can be understood as a phase separation process (like the Ising model phase transition).

The process which drives chemotaxis is a complex combination of protein interactions in the so called **signalling network**.

The architecture of this network is very similar to that of **multilayer perceptrons** and, as for MLP, the signalling network is able to organize non trivial strategies

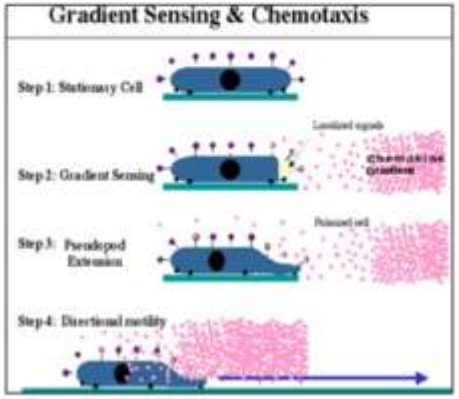
# Eukaryotic chemotaxis (1)

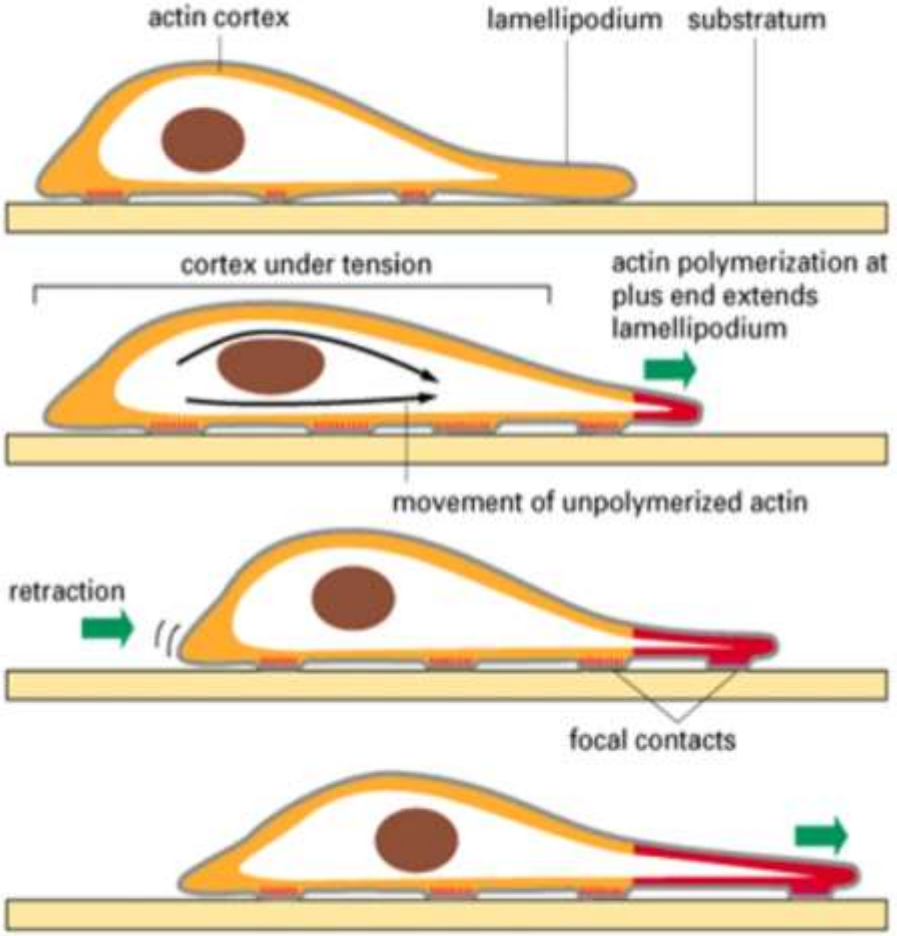
**Crowling movement on a surface in 3 steps:**

Extension of protusions from the leading edge;

Anchorage of protrusions to the substrate (trans-membrane proteins, contractile apparatus,...);

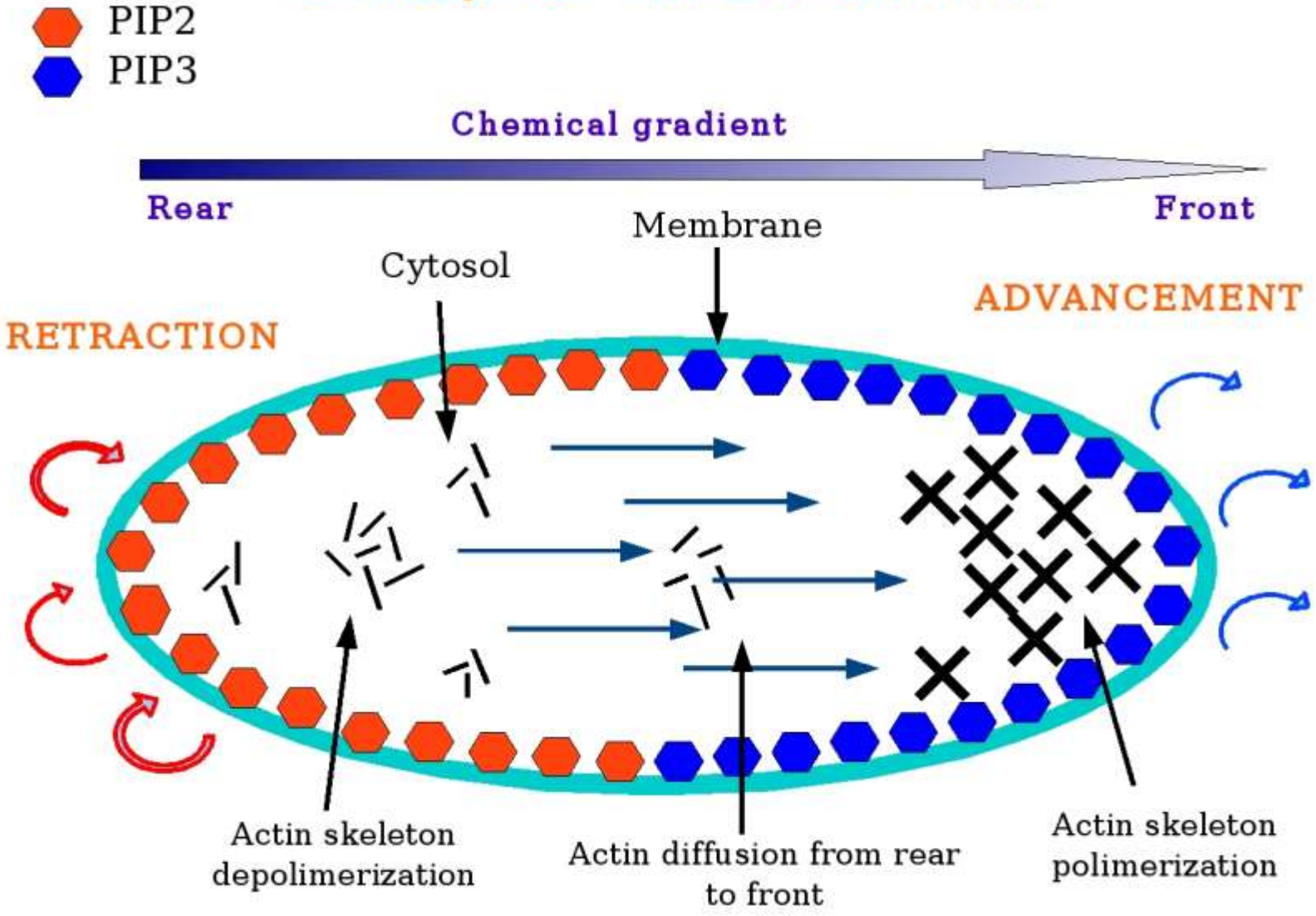
Rear edge retraction.



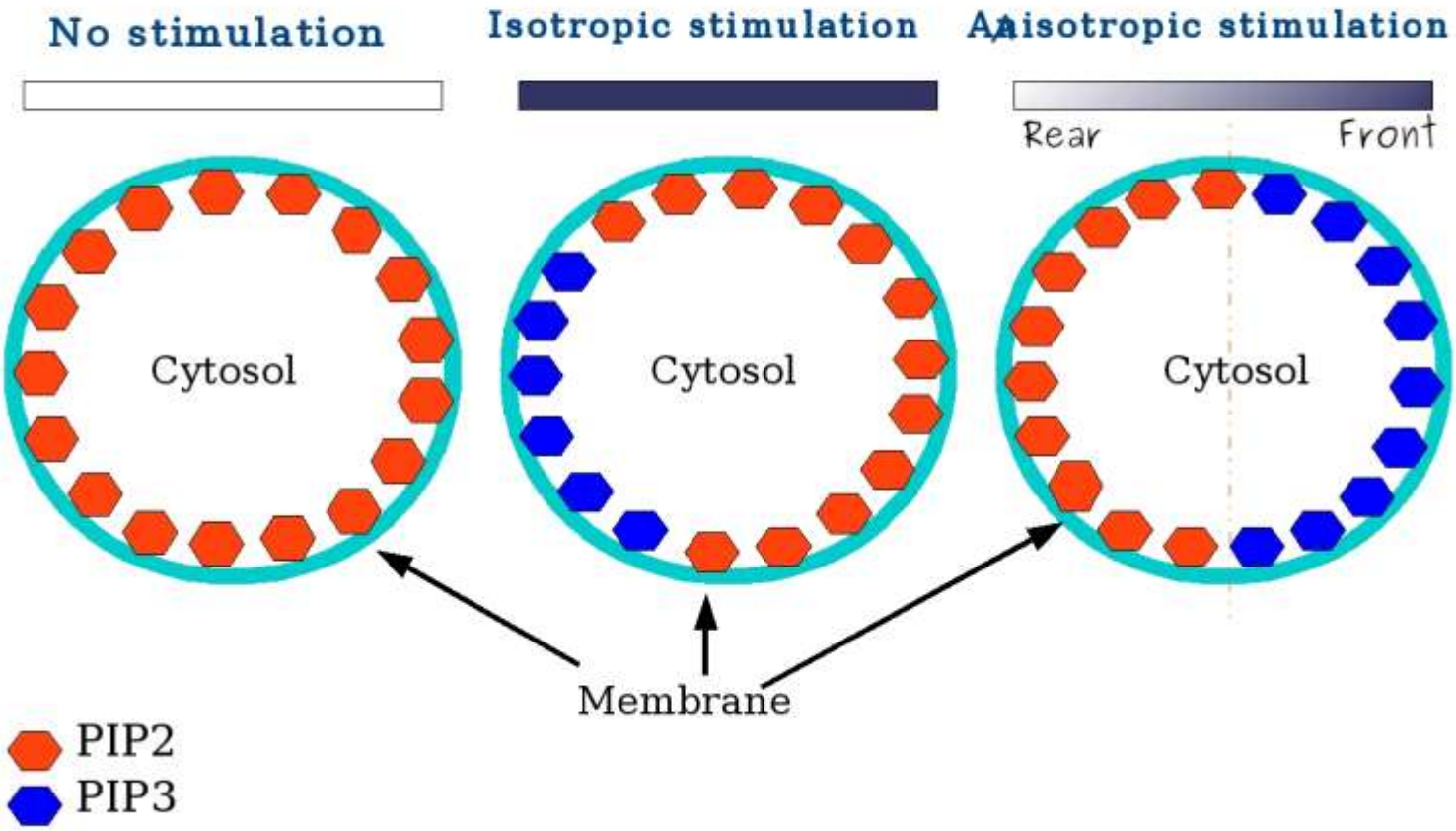




http://klemkelab.ucsd.edu/research/cell.htm

# Eukaryotic chemotaxis (2)

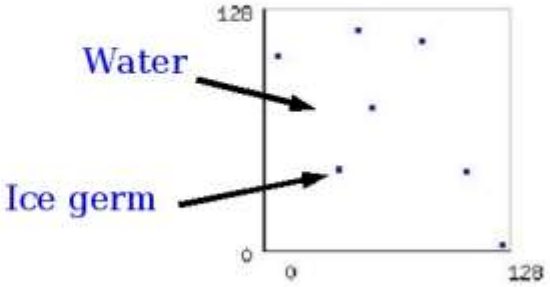


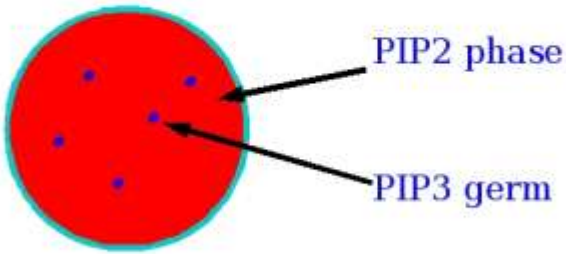
# Eukaryotic chemotaxis (4)



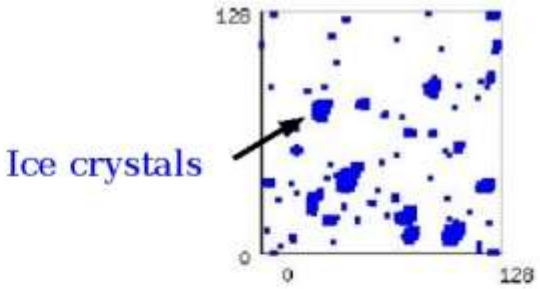
# Physical analogy: changes of state

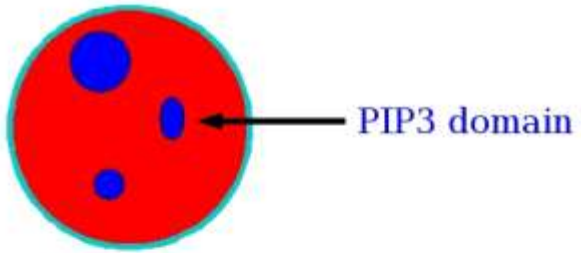
**Nucleation/ metastability**



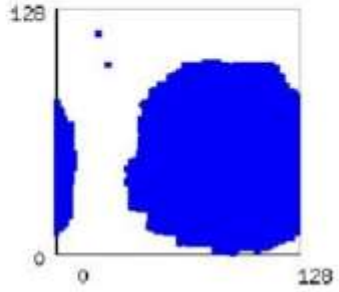




**Growth**

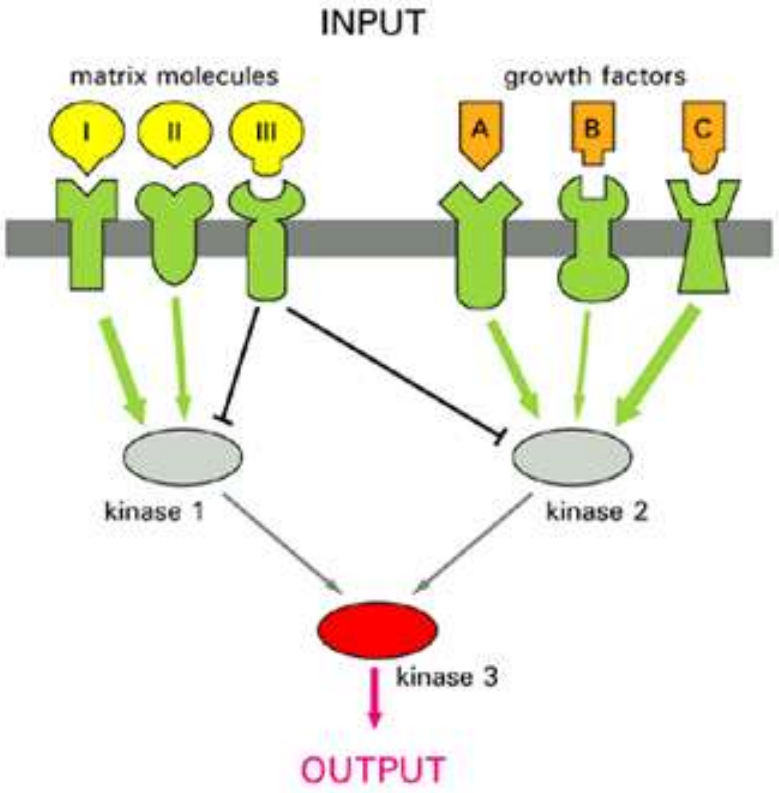


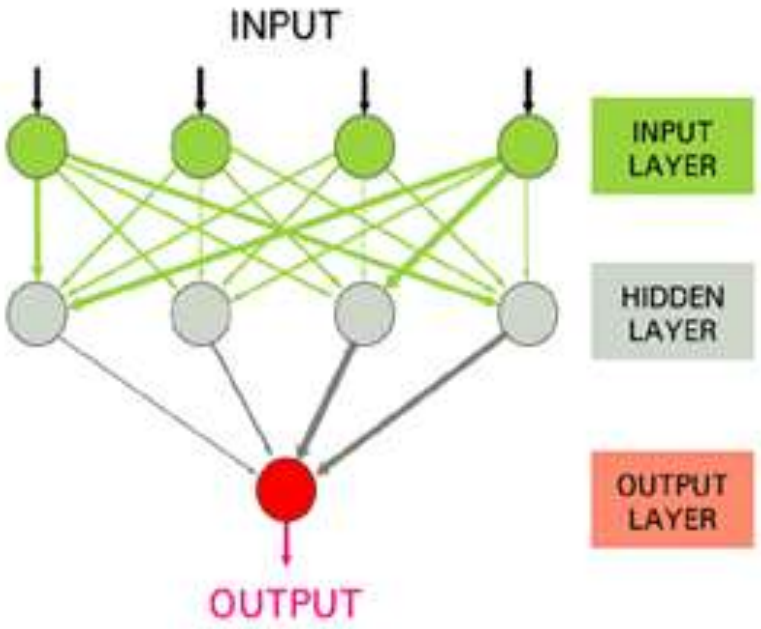




**Coalescence**




# Signalling Network versus MLP



# Conclusions

Quantitative biology offers a lot of interesting challenges for physicists, both from the experimental point of view:

- nanotechnologies
- microfluidics

and from the theoretical point of view:

- modeling
- inference techniques
- simulations