Quantitative Biology

La Thuile 03/03/2011

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

Plan of the talk

1. Introduction: DNA, genes and proteins

2. The last ten years: The “genomic revolution”

3. New tools and ideas:
   Computational Biology and Systems biology

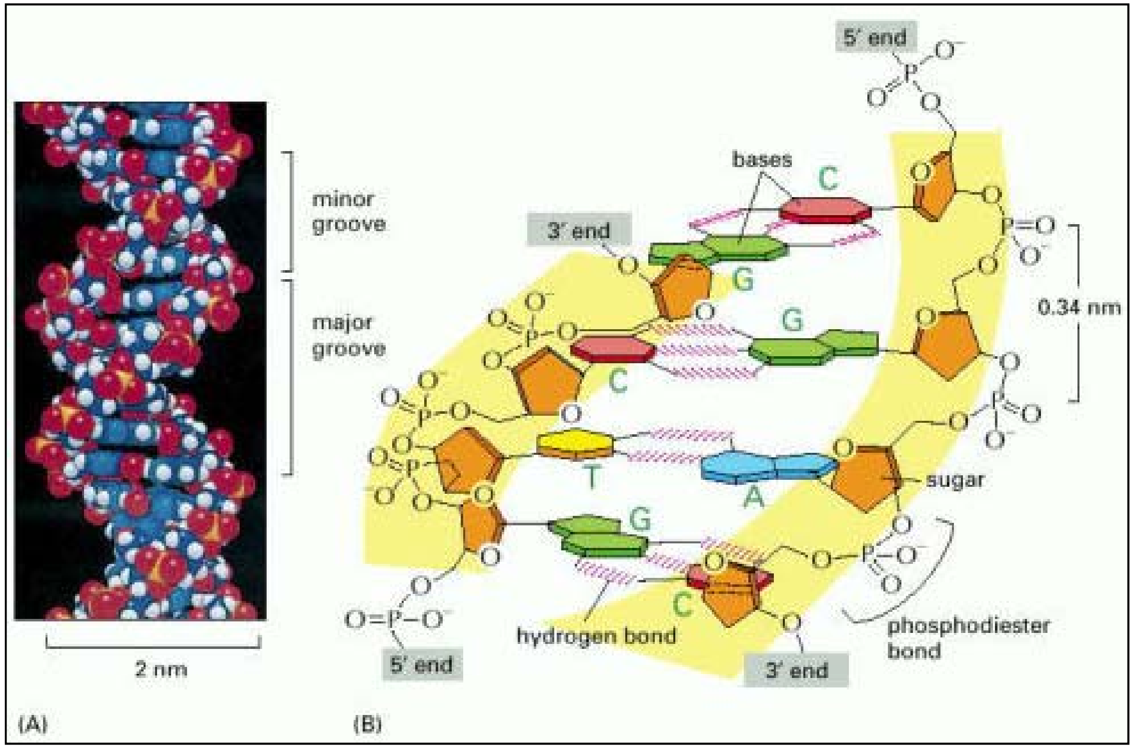
4. Example 1: Evolutionary models

5. Example 2: Gene Regulation

6. Example 3: Chemotaxis

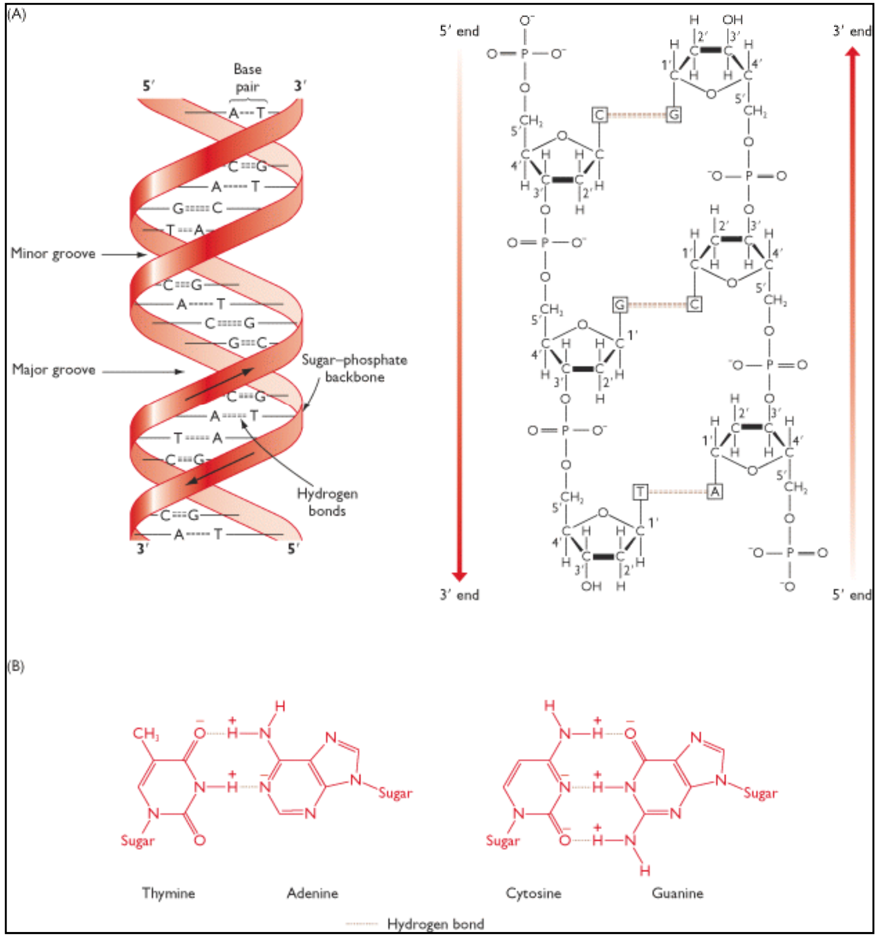
DNA

- 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.

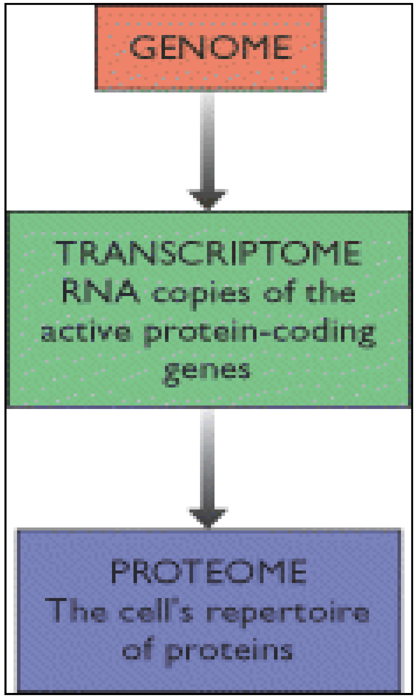


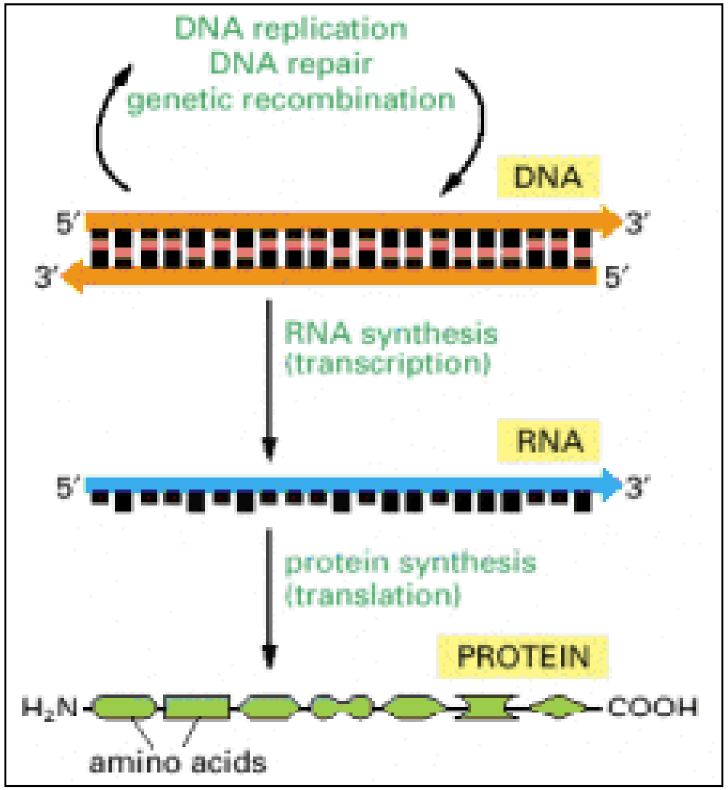
Proteins

Most of the functions in the cell are performed by **proteins** which are composed by 20 different types of elemantary constituents: the aminoacids

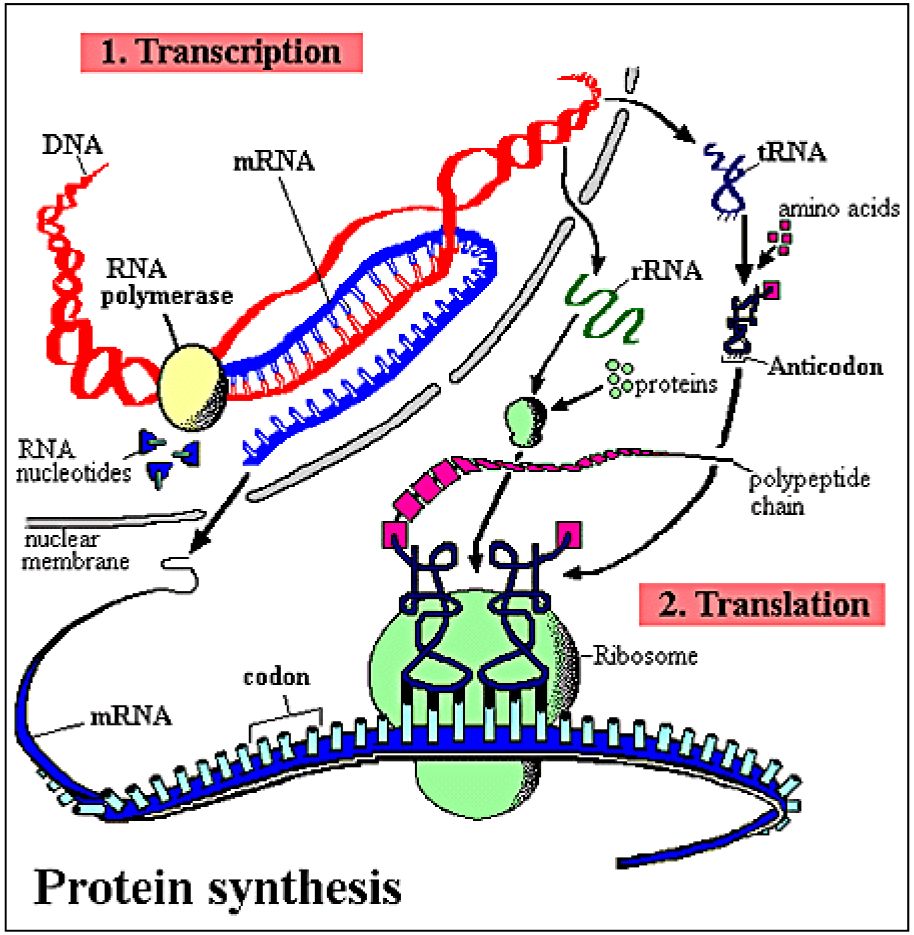
Information flow in the cell

“Central Dogma” of molecular biology


GENOME
TRANSCRIPTOME
RNA copies of the active protein-coding genes
PROTEOME
The cell's repertoire of proteins

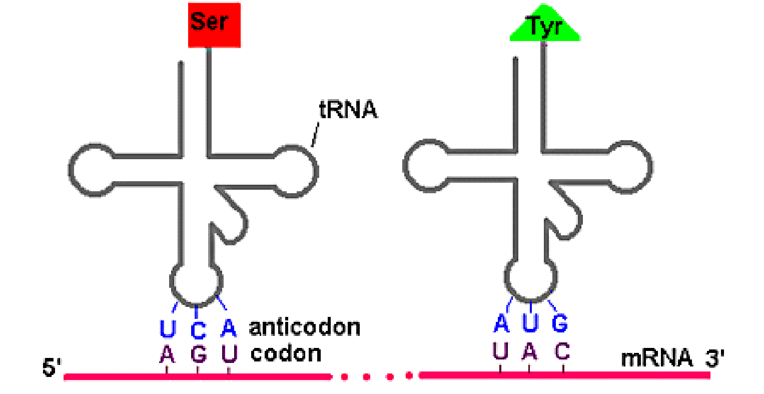


Protein synthesis



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

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

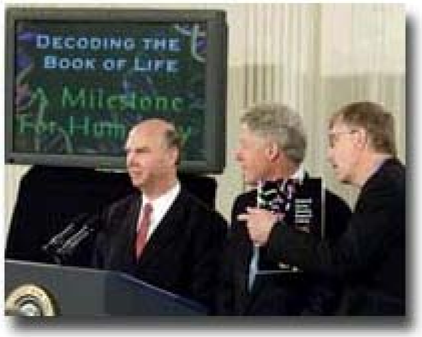
And on the theoretical side:

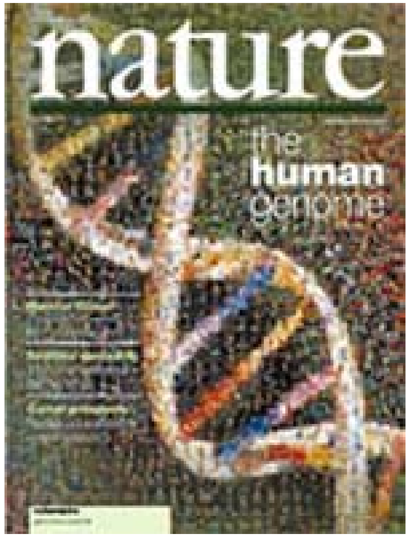
- new inference methods
- modeling of complex systems
- network theory
- alignment tools

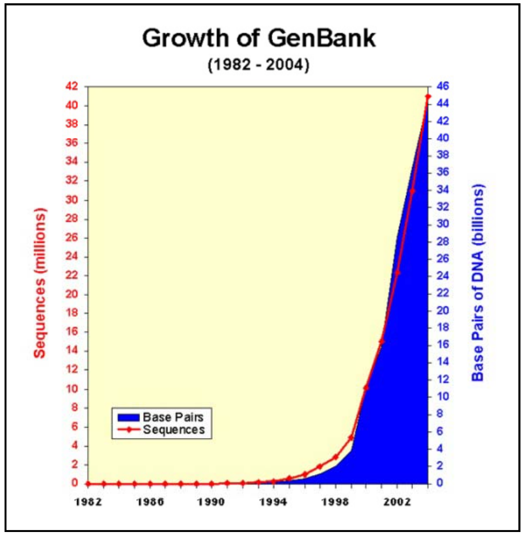
Genomic Revolution: *sequences*

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

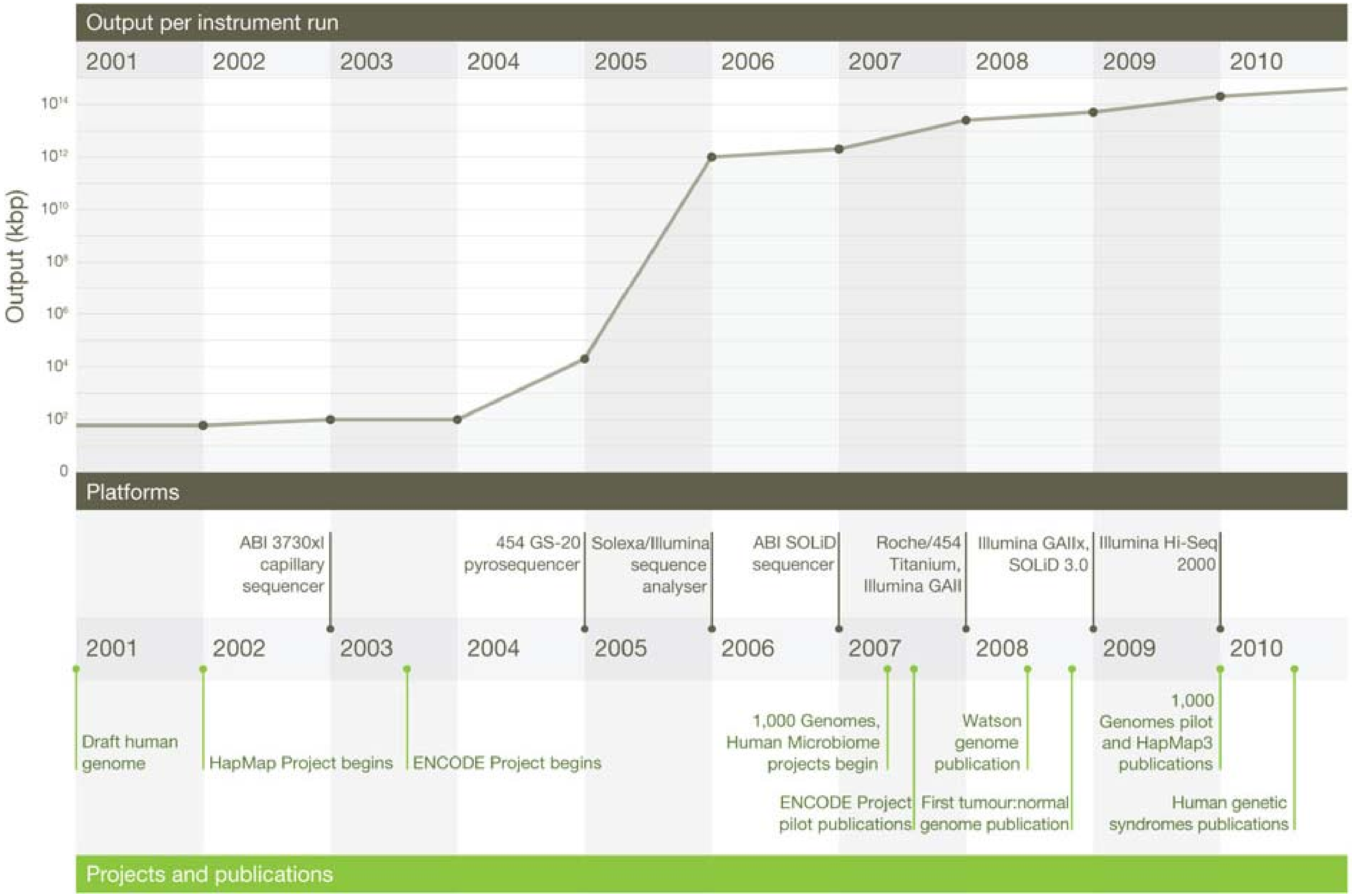
> *homo_sapiens*
ACTTTTTTACCCTCGTGTGTTGC
AGACTTTTTGCCACTTTTAAAAC
GCTGACAATTCGACCCTTTCCAA
GTGCAAAAAGTGCCAAGATTTA
CGATAAAATTCCCCCGAGAGAC
GTGTGCA………





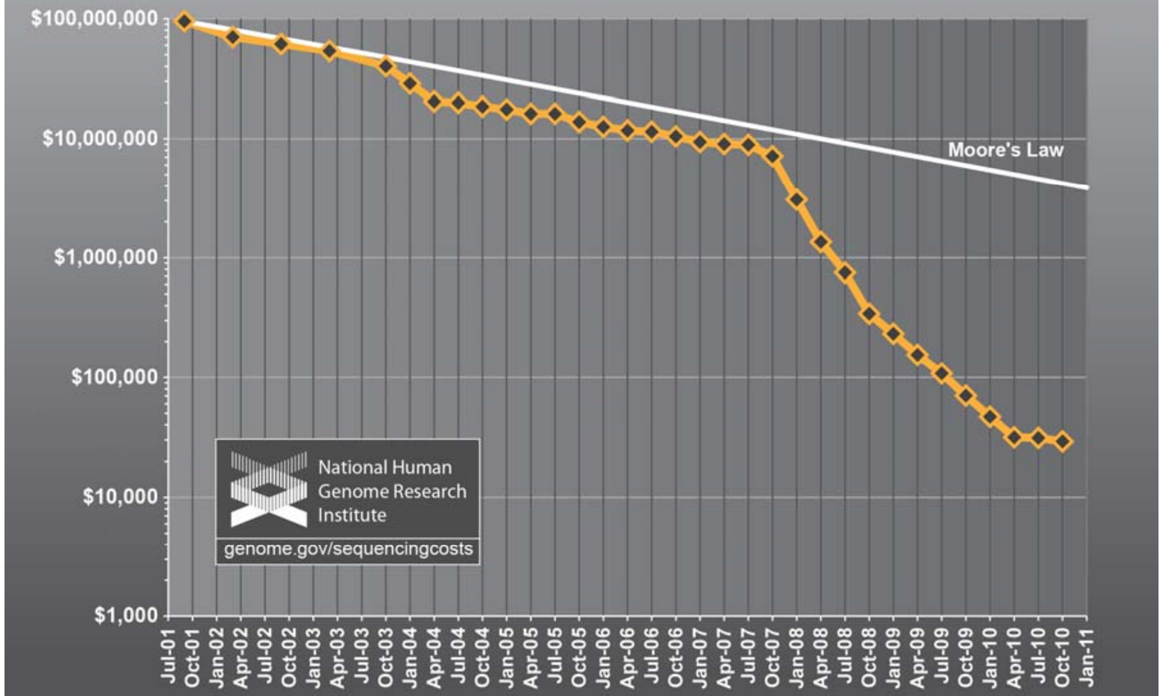


Changes in instrument capacity over the past decade



Timing of the major sequencing projects

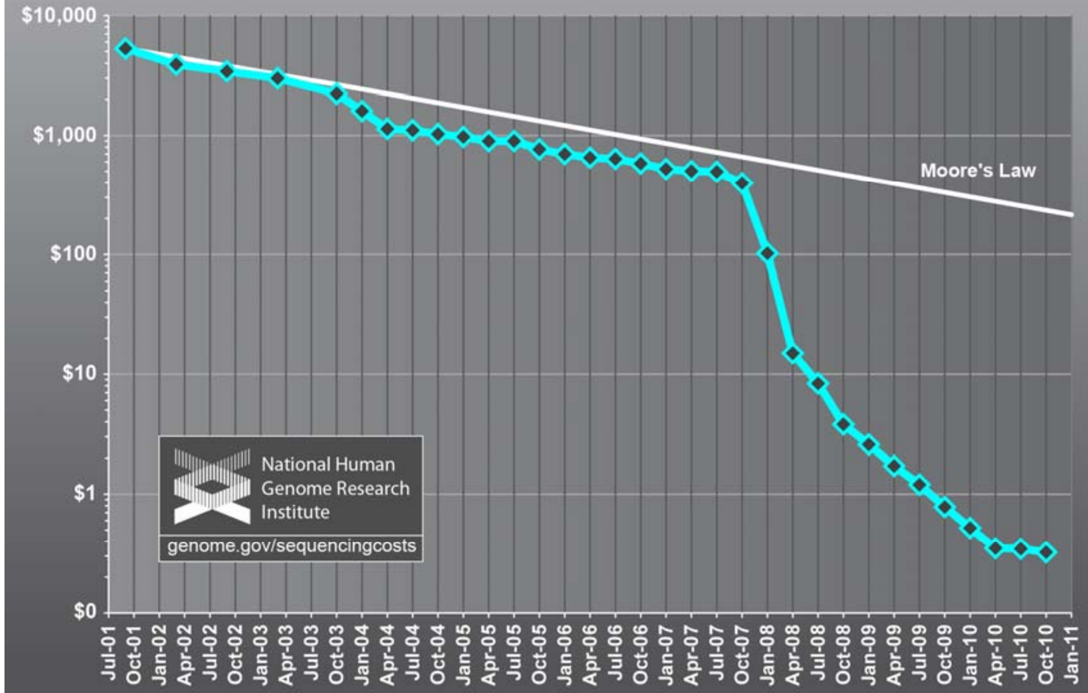
Cost per Genome



National Human
Genome Research
Institute

genome.gov/sequencingcosts

Cost per Megabase of DNA Sequence

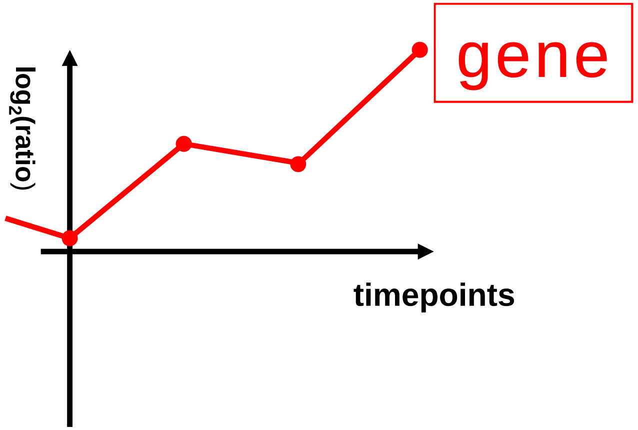


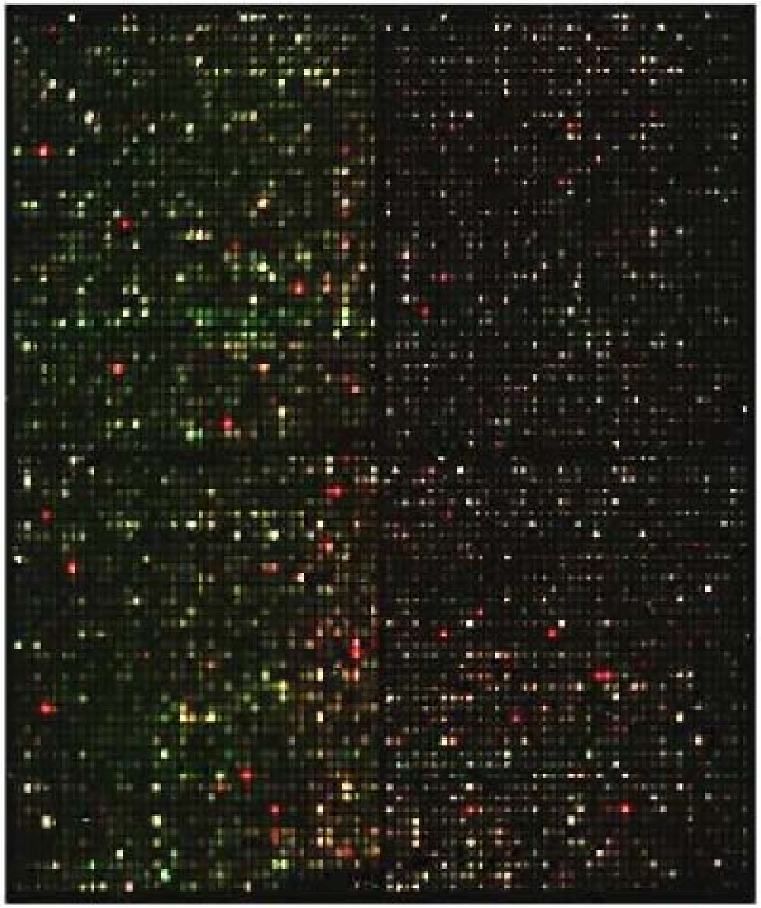
National Human Genome Research Institute

genome.gov/sequencingcosts

Transcriptomics: *microarray*

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

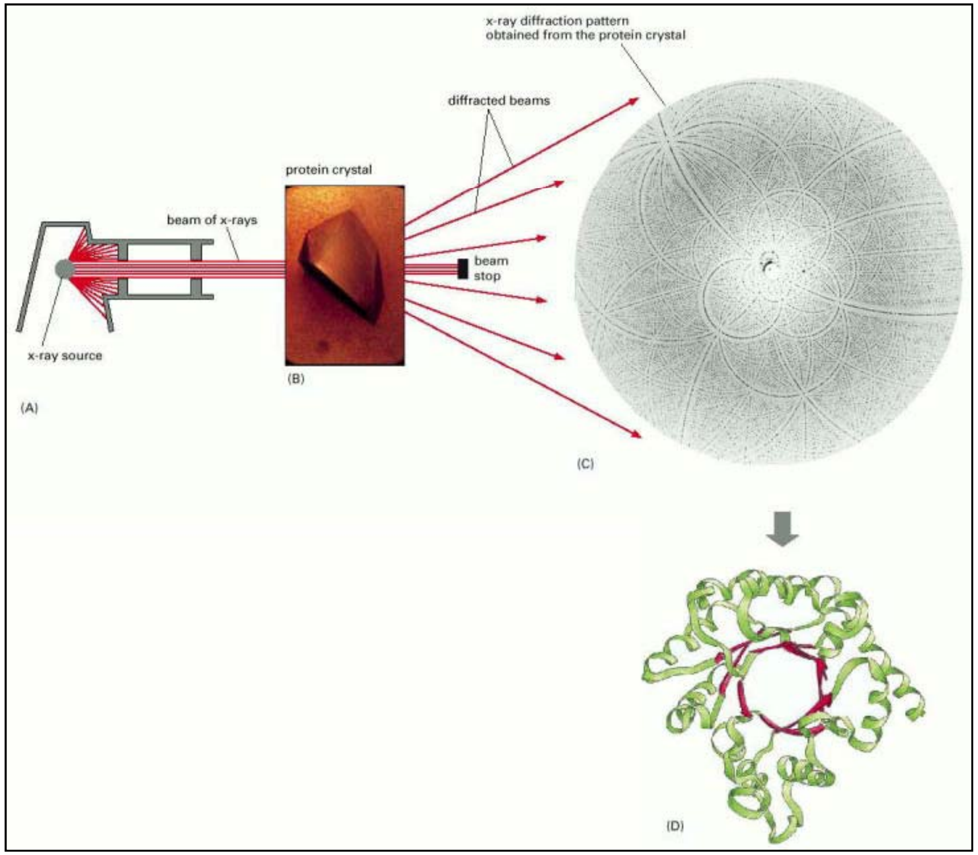




Proteomics:

- 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?

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

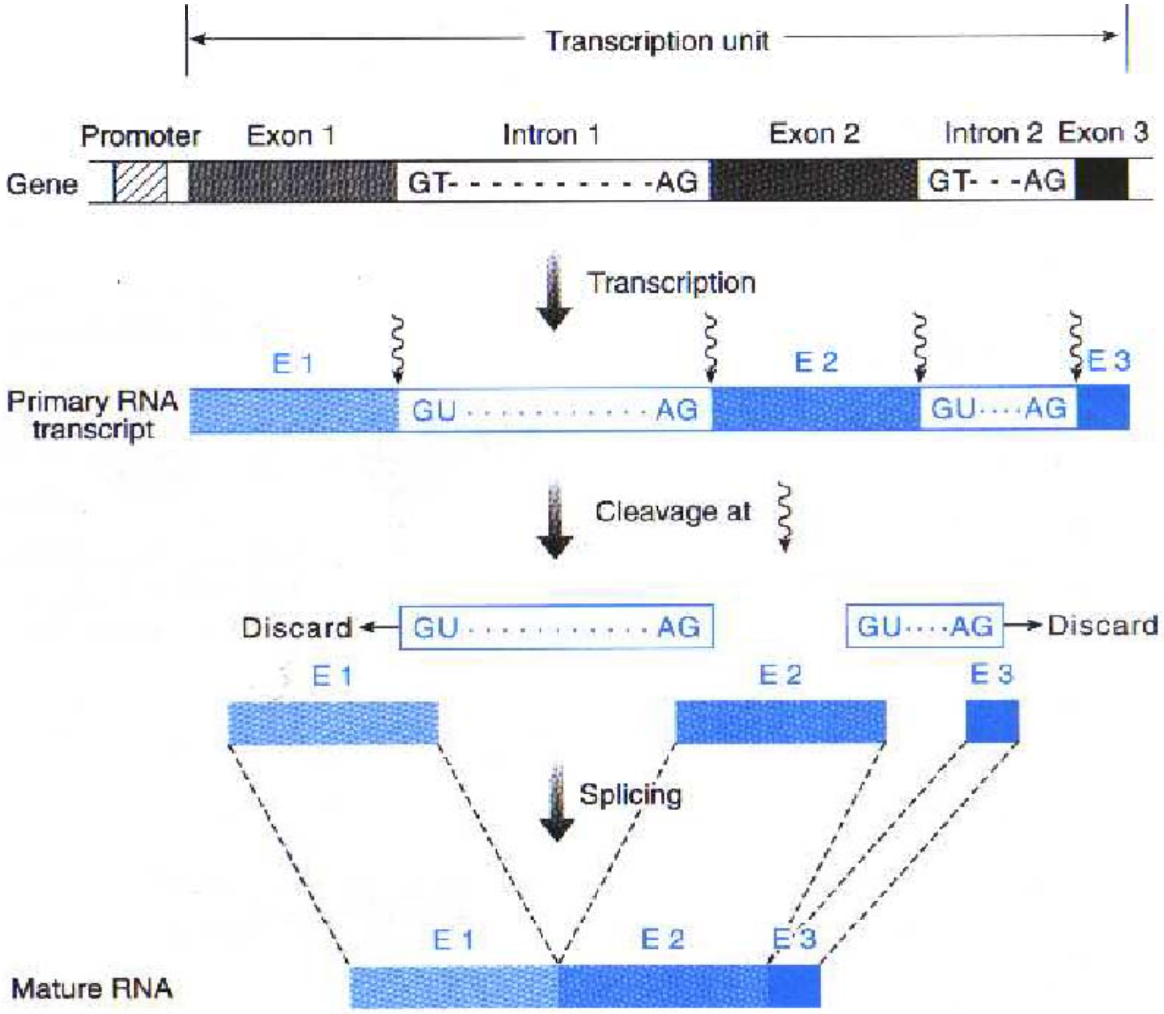
Genome Organization

- 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.

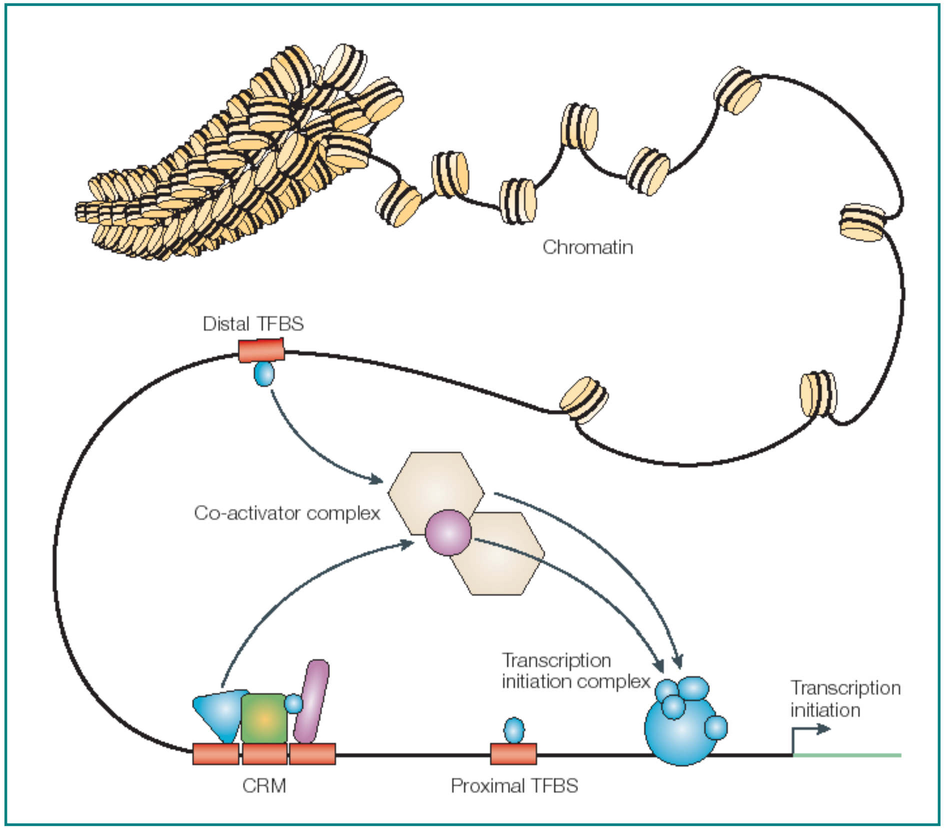
Gene structure

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)

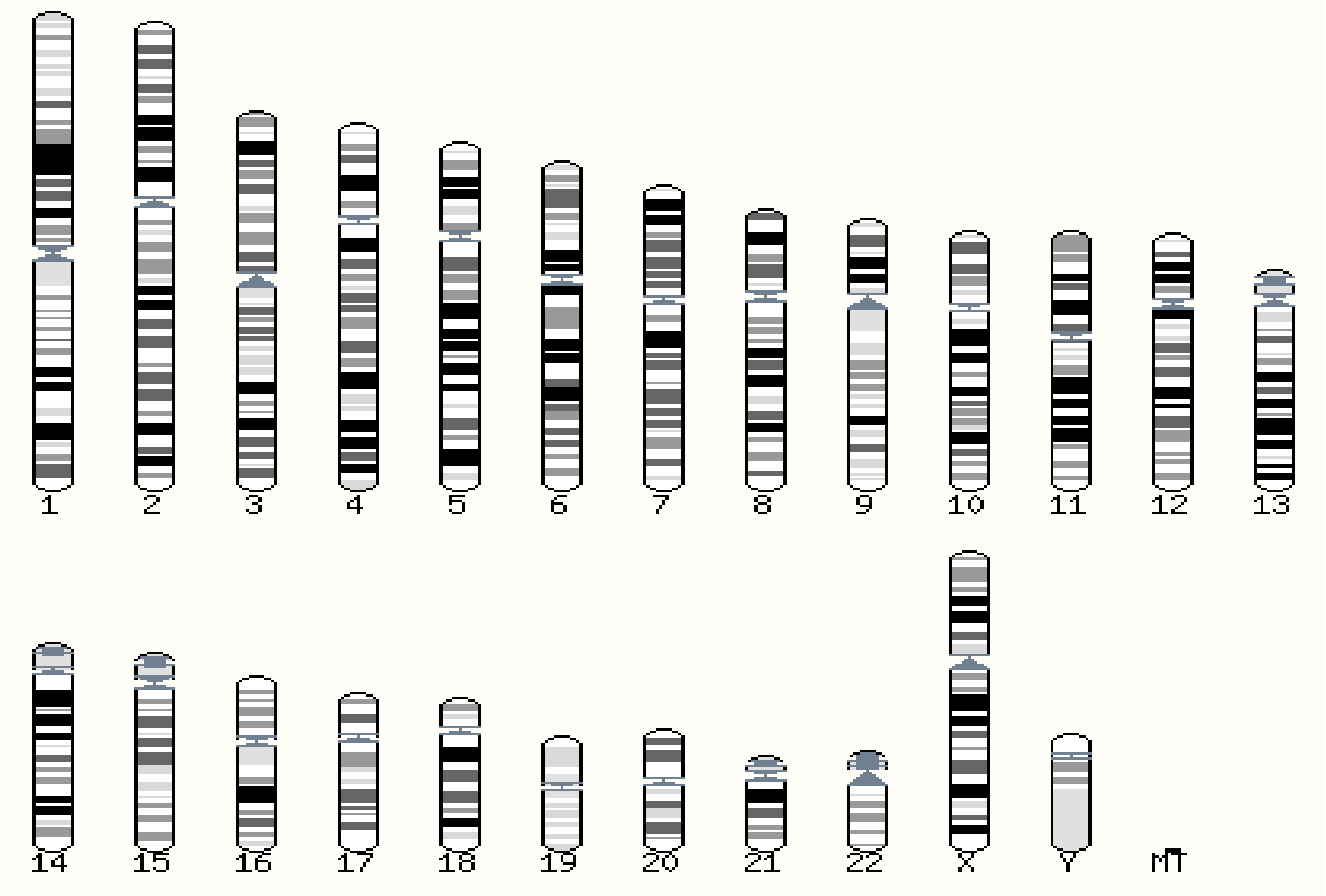

Transcription unit
Promoter
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Gene
GT- - - - - - - - - - -AG
GT- - -AG
Transcription
E 1
E 2
E 3
Primary RNA transcript
GU · · · · · · · · · · · AG
GU · · · · AG
Cleavage at
Discard
GU · · · · · · · · · · · AG
GU · · · · AG
Discard
E 1
E 2
E 3
Splicing
E 1
E 2
E 3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

Gene Regulation

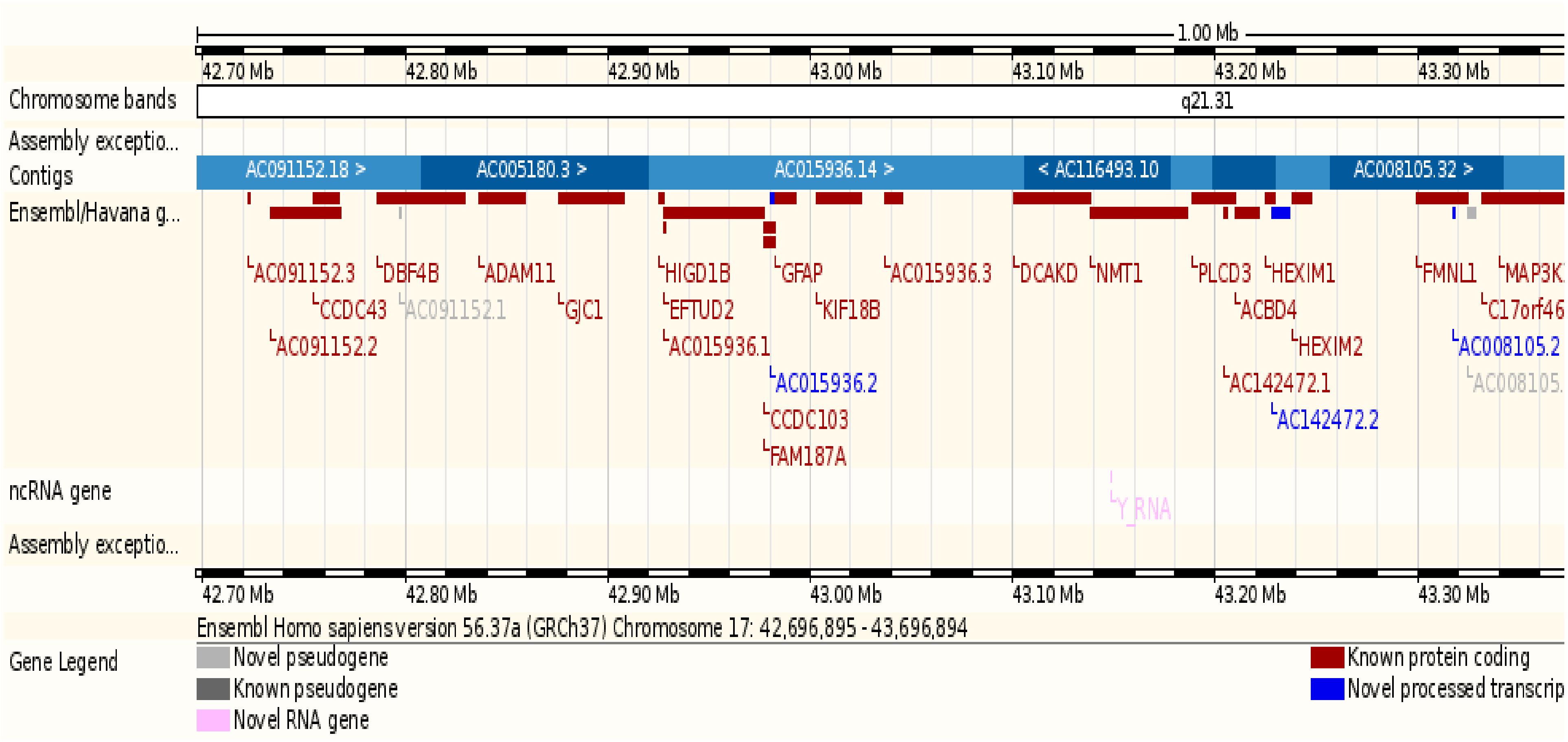
- Gene expression is tightly regulated. All cells in the body carry the full set of genes, but only expres about 20% of them at any particular time



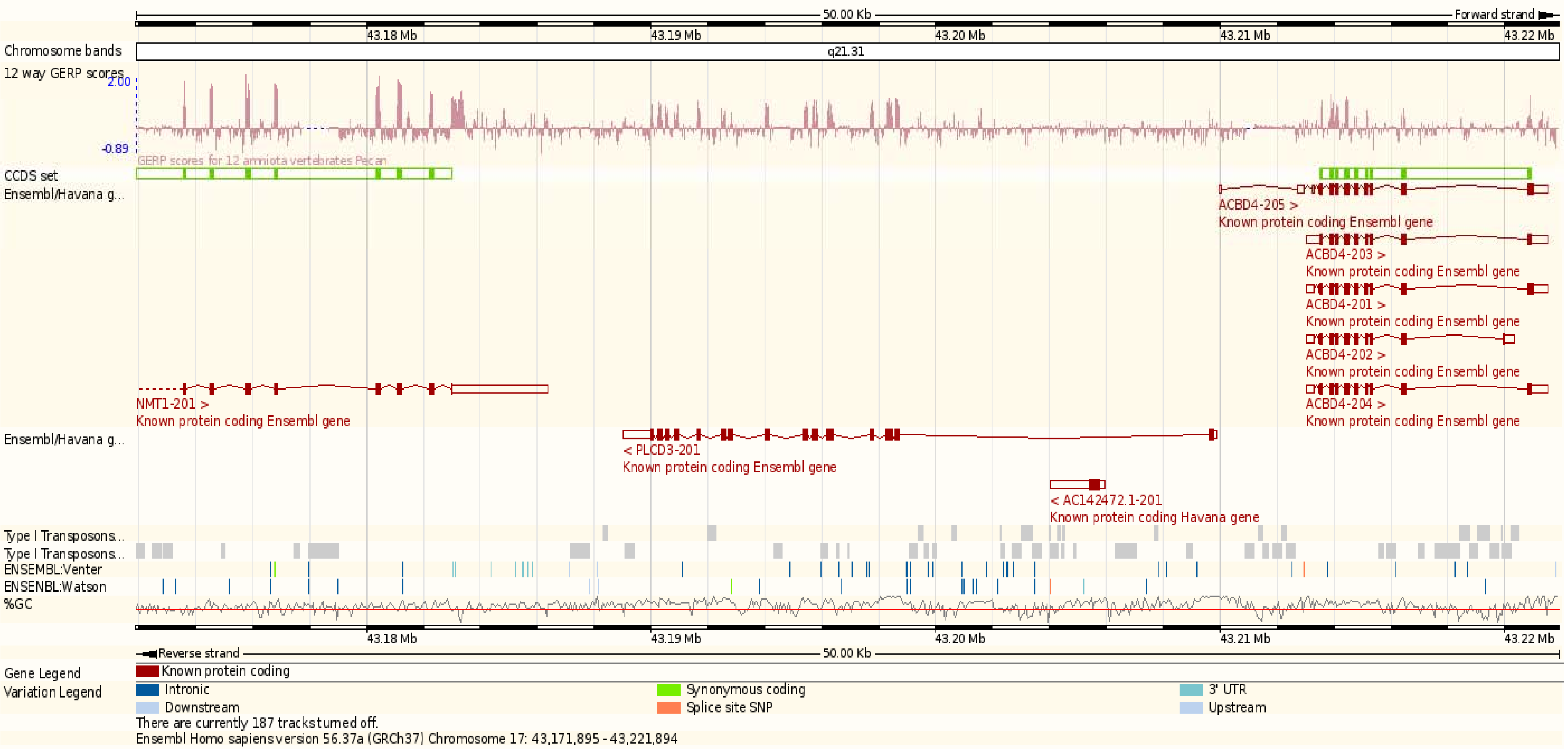
Human Genome



Ensembl Genome Browser



Zoom !



There are currently 187 tracks turned off.

Ensembl Homo sapiens version 56.37a (GRCh37) Chromosome 17: 43,171,895 - 43,221,894

New questions, a few tentative answers

- Most of the human DNA is not coding and regulates protein expression.
- The “central dogma” is wrong:

  from a single gene the cell can create a huge amount of different proteins

  (alternative splicing)
- The information flow can be reverted: from RNA to DNA (Retrotransposons)

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.

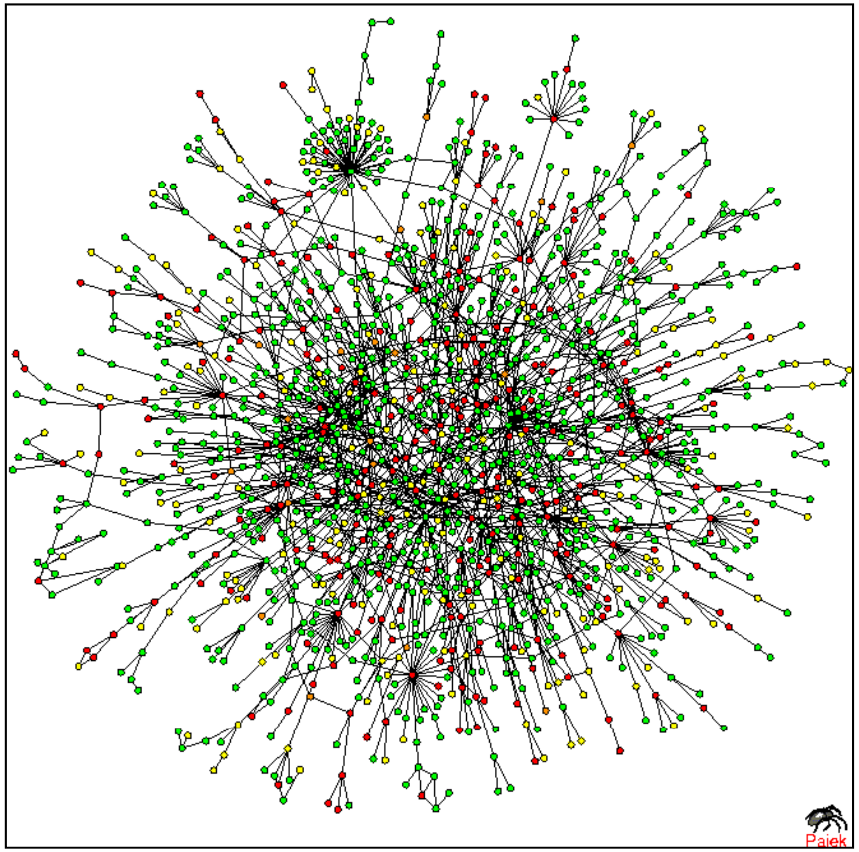
Ontologies: biological (and medical) information must be organized in a quantitative and standardized way

Modern Genomics: *networks*

- genes and proteins of a given organism are organized in networks .

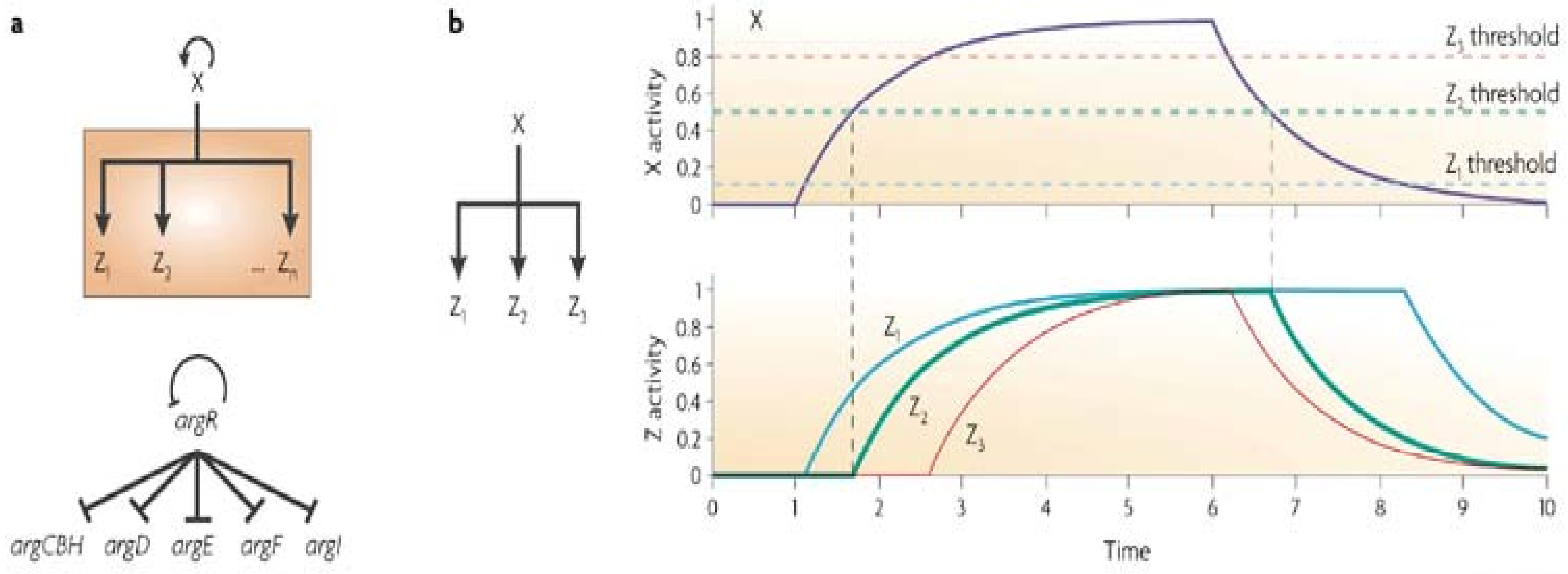
- Cells react to external stimuli in a “global” way.

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



Nature Reviews | Genetics

Modern Genomics:
Gene Ontology

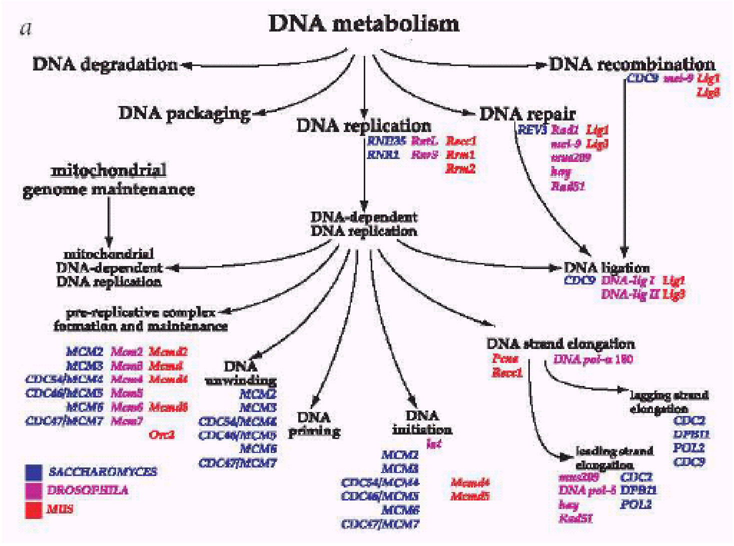
- **Gene Ontology** is an example of standardization of biological data.

- 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

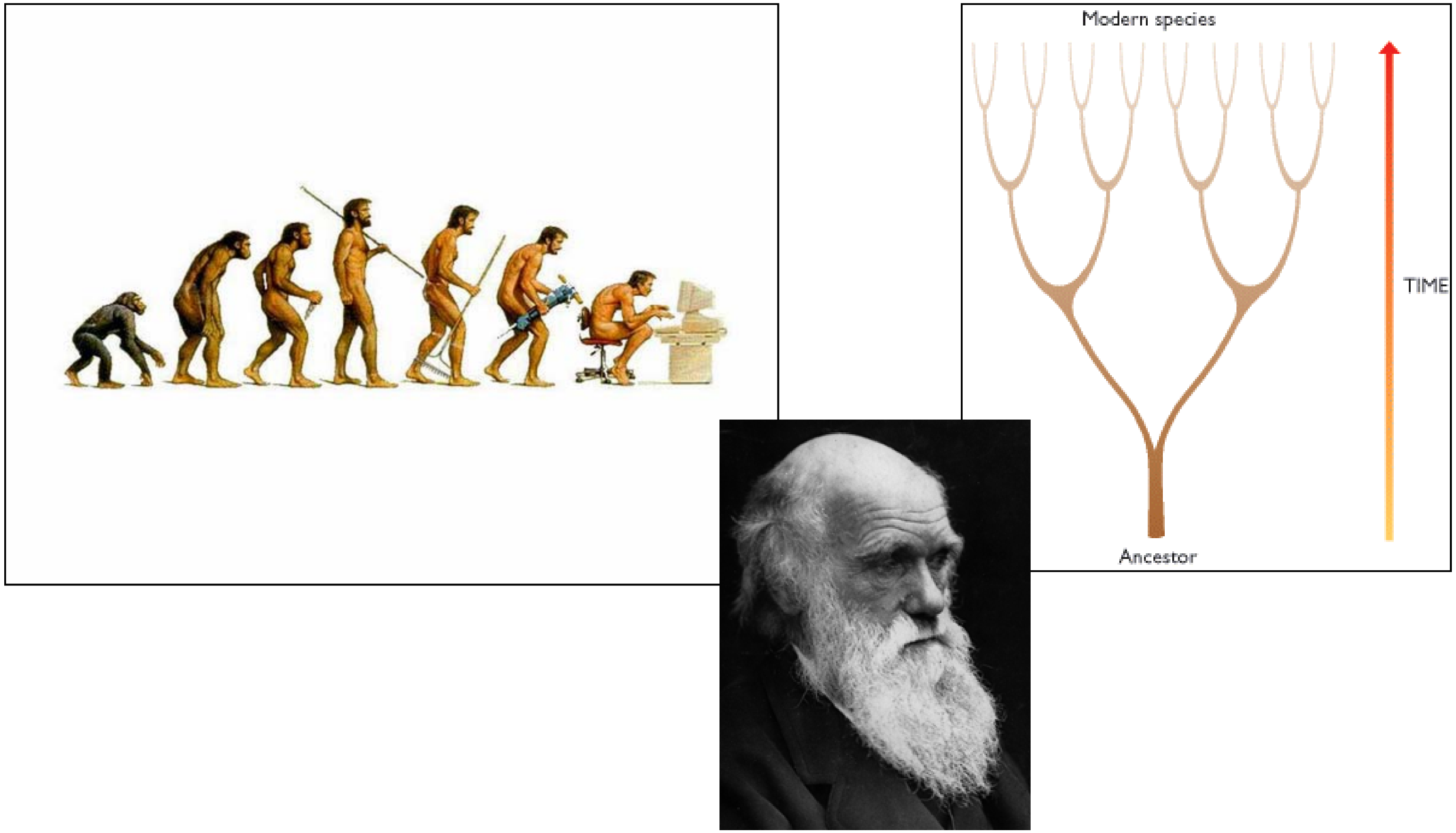
Three examples of applications

§ Evolutionay models

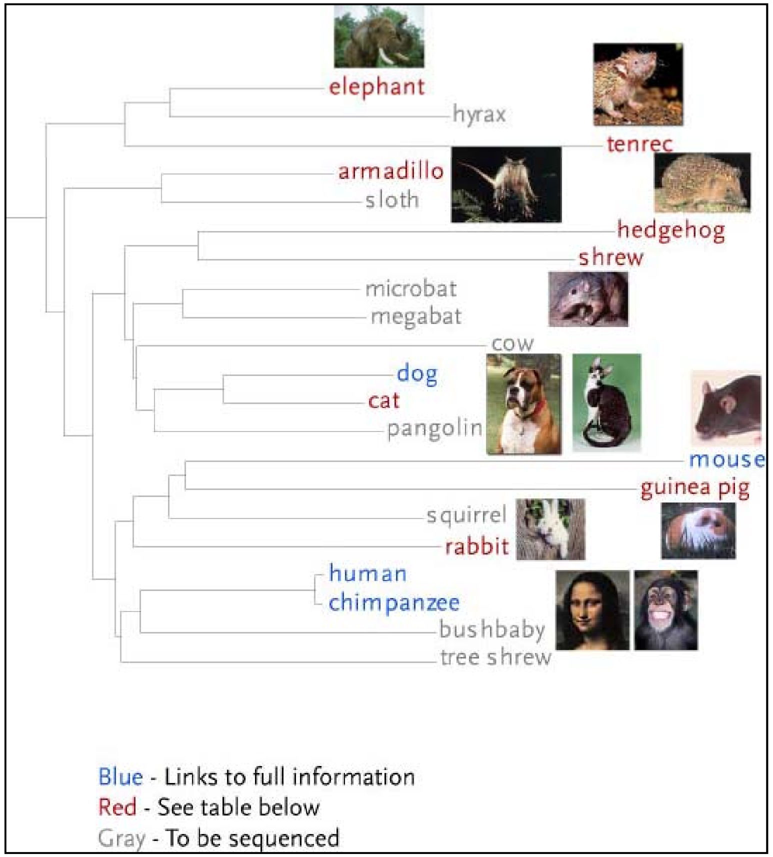
§ Gene Regulation

§ Chemotaxis

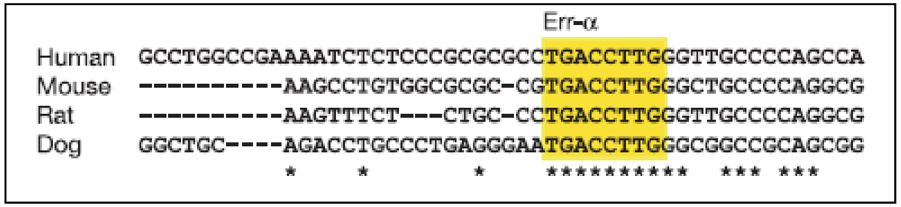
Evolutionary models



Taxonomic versus Genomic trees

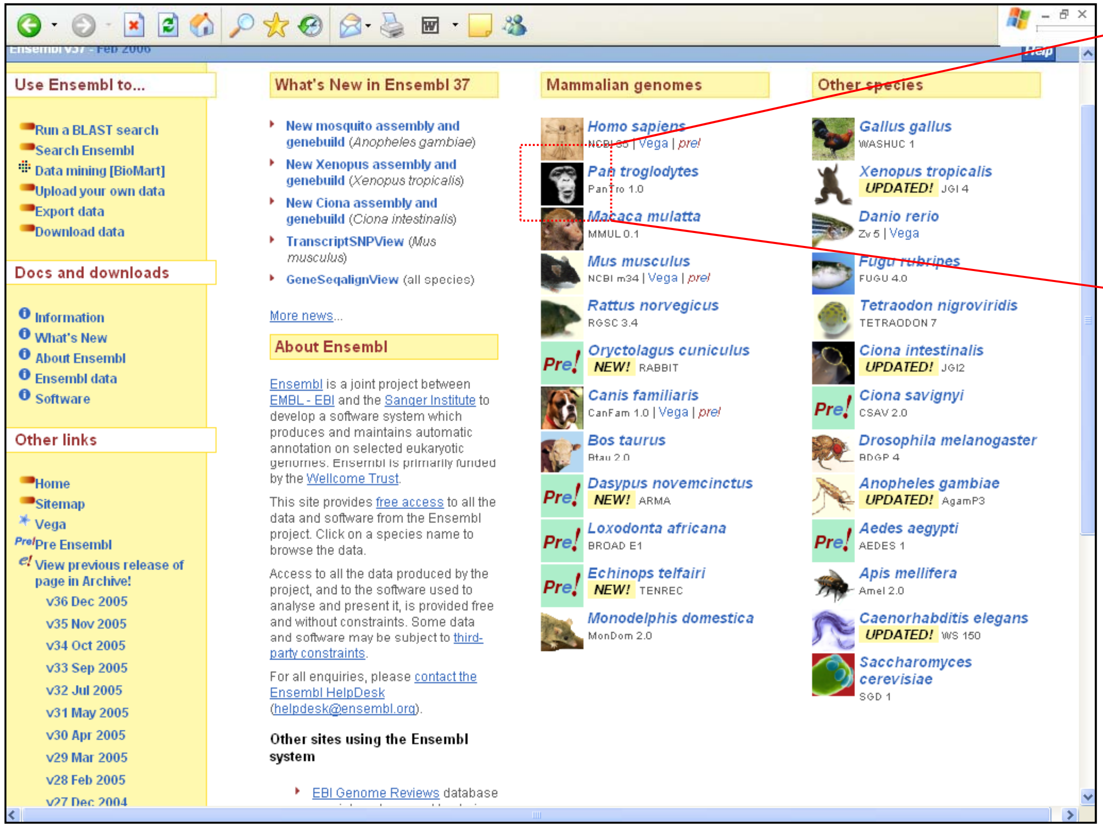


Genomic trees may be obtained using alignment algorithms. They are impressively similar to the taxonomic trees. This is a highly non trivial test of Evolution theory.



```
                                     Err-α
Human  GCCTGGCCGAAAATCTCTCCCGCGCGCCTGACCTTGGGTTGCCCCAGCCA
Mouse  ----------AAGCCTGTGGCGCGC-CGTGACCTTGGGCTGCCCCAGGCG
Rat    ----------AAGTTTCT---CTGC-CCTGACCTTGGGTTGCCCCAGGCG
Dog    GGCTGC----AGACCTGCCCTGAGGGAATGACCTTGGGCGGCCGCAGCGG
                 *    *       *    **********  *** ***
```

Human and Chimps



Ensembl v37 - Feb 2006

Help

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
  - v36 Dec 2005
  - v35 Nov 2005
  - v34 Oct 2005
  - v33 Sep 2005
  - v32 Jul 2005
  - v31 May 2005
  - v30 Apr 2005
  - v29 Mar 2005
  - v28 Feb 2005
  - v27 Dec 2004

What's New in Ensembl 37

- New mosquito assembly and genebuild (*Anopheles gambiae*)
- New Xenopus assembly and genebuild (*Xenopus tropicalis*)
- New Ciona assembly and genebuild (*Ciona intestinalis*)
- TranscriptSNPView (*Mus musculus*)
- GeneSeqalignView (all species)

More news...

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints.

For all enquiries, please contact the Ensembl HelpDesk (helpdesk@ensembl.org).

Other sites using the Ensembl system

- EBI Genome Reviews database

Mammalian genomes

- *Homo sapiens* NCBI 35 | Vega | *pre!*
- *Pan troglodytes* PanTro 1.0
- *Macaca mulatta* MMUL 0.1
- *Mus musculus* NCBI m34 | Vega | *pre!*
- *Rattus norvegicus* RGSC 3.4
- *Oryctolagus cuniculus* *NEW!* RABBIT
- *Canis familiaris* CanFam 1.0 | Vega | *pre!*
- *Bos taurus* Btau 2.0
- *Dasypus novemcinctus* *NEW!* ARMA
- *Loxodonta africana* BROAD E1
- *Echinops telfairi* *NEW!* TENREC
- *Monodelphis domestica* MonDom 2.0

Other species

- *Gallus gallus* WASHUC 1
- *Xenopus tropicalis* *UPDATED!* JGI 4
- *Danio rerio* Zv 5 | Vega
- *Fugu rubripes* FUGU 4.0
- *Tetraodon nigroviridis* TETRAODON 7
- *Ciona intestinalis* *UPDATED!* JGI2
- *Ciona savignyi* CSAV 2.0
- *Drosophila melanogaster* BDGP 4
- *Anopheles gambiae* *UPDATED!* AgamP3
- *Aedes aegypti* AEDES 1
- *Apis mellifera* Amel 2.0
- *Caenorhabditis elegans* *UPDATED!* WS 150
- *Saccharomyces cerevisiae* SGD 1

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

Evolution and gene regulation

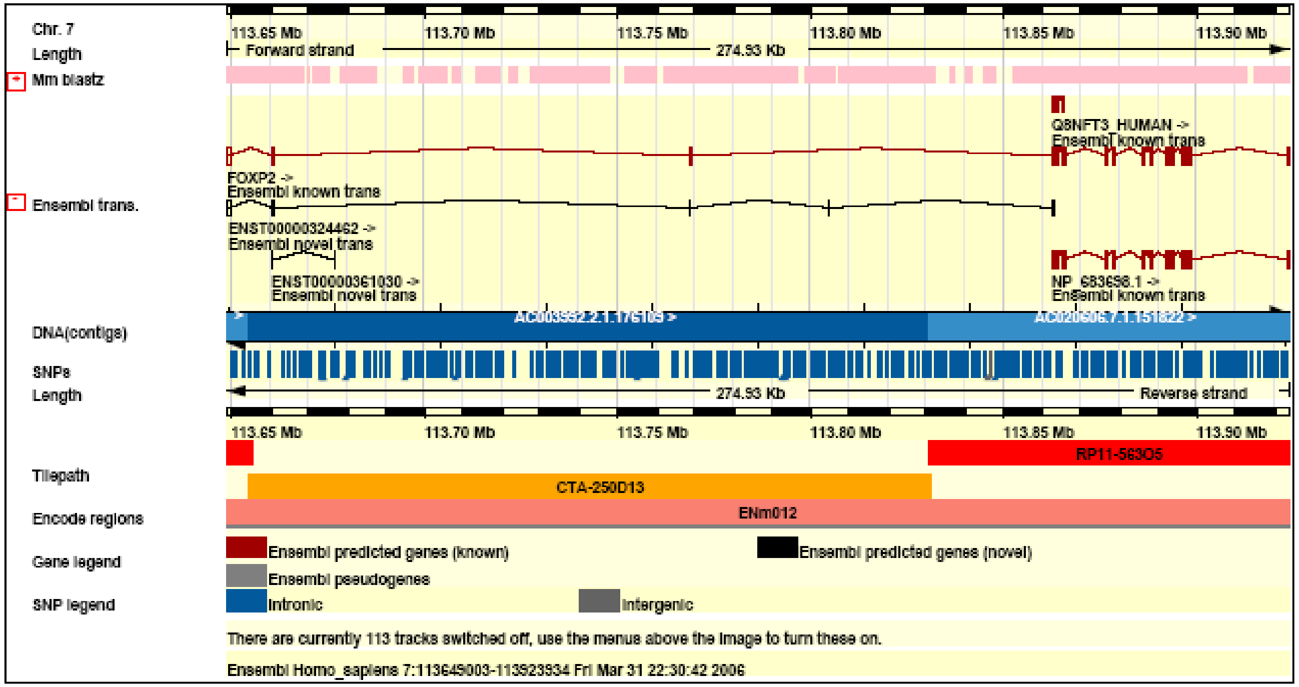
- Goal: use evolutionary conservation to identify functionally important regions of the genome. Different regions show different levels of conservation

"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.



There are currently 113 tracks switched off, use the menus above the image to turn these on.

Ensembl Homo_sapiens 7:113649003-113923934 Fri Mar 31 22:30:42 2006

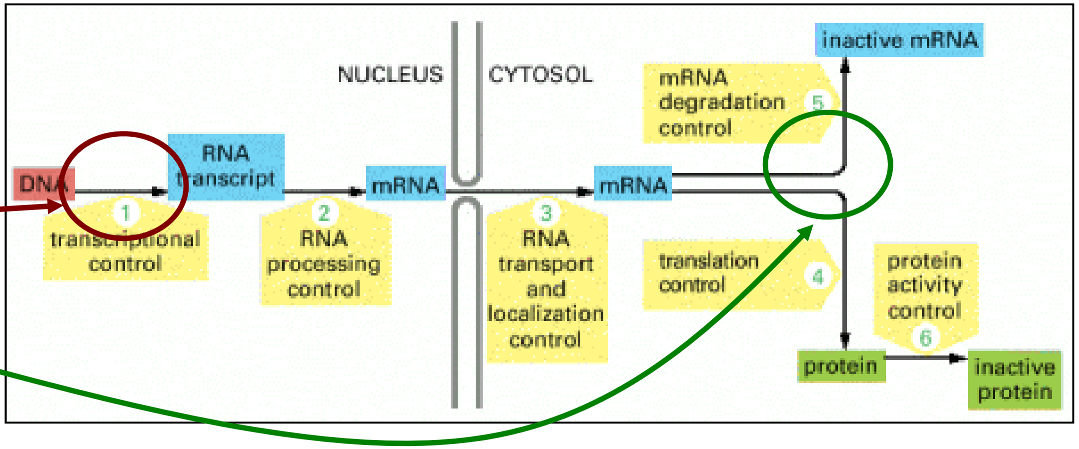
Gene Regulation

Gene expression in eukaryotes is carefully controlled.

Among the various regulatory steps the most important ones are:

□ transcriptional control, by **Transcription Factors.**

□ post-transcriptional control, by **microRNAs.**

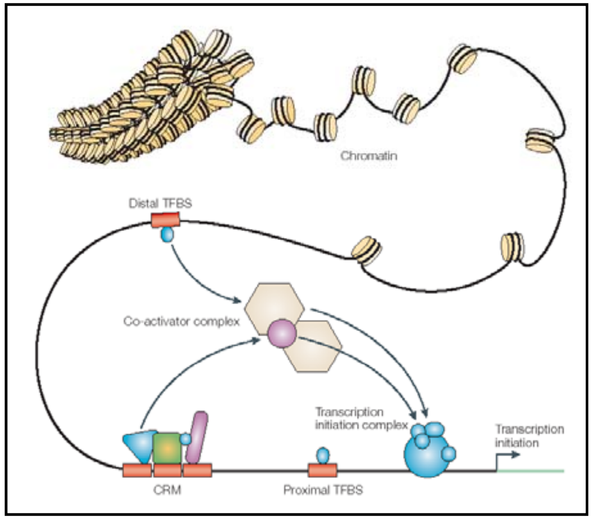


Alberts, *Molecular Biology of the Cell*

Transcription Factors and miRNAs

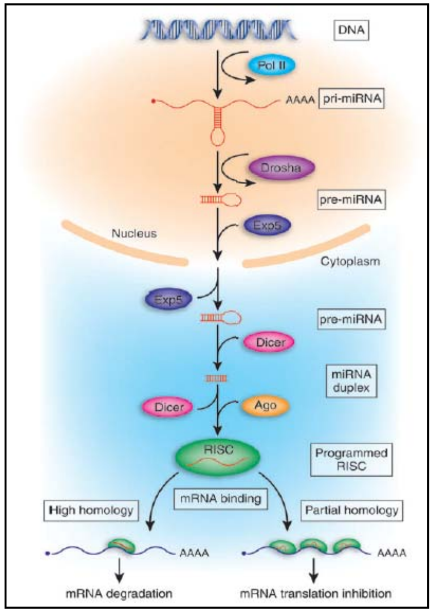
- **Regulation of gene expression** mainly mediated by:

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

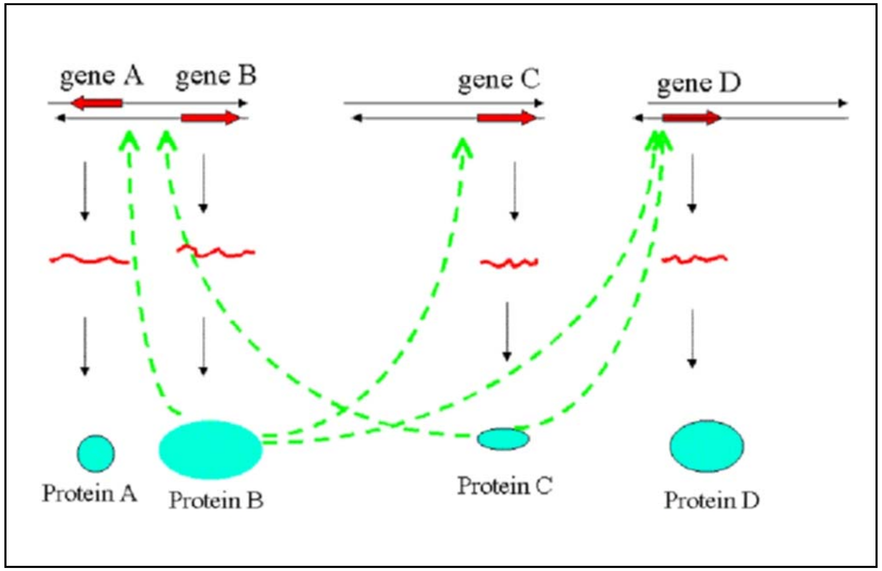
MicroRNAs (miRNAs) are a family of small RNAs (typically **21 - 25** nucleotide long) that **negatively regulate gene expression at the posttranscriptional level**, (usually) thanks to the "seed" region in 3'-UTR regions.

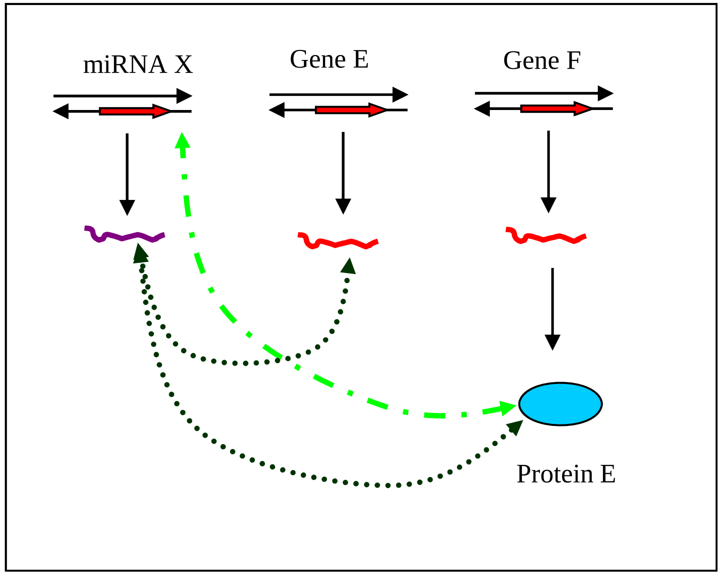


Regulatory Networks 1

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

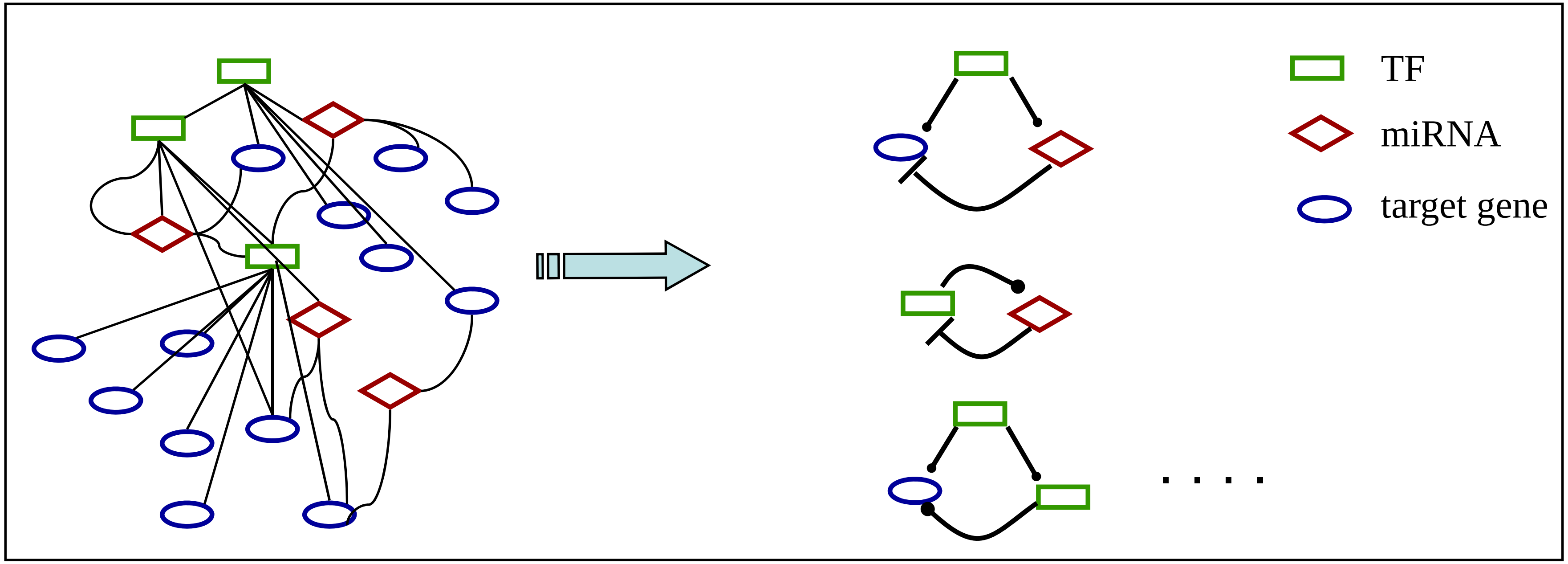
miRNAs also act in a combinatorial and one-to-many way, and, moreover, are transcribed from same POL-II promotes of TFs.
--> Post-Transcriptional Network





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.



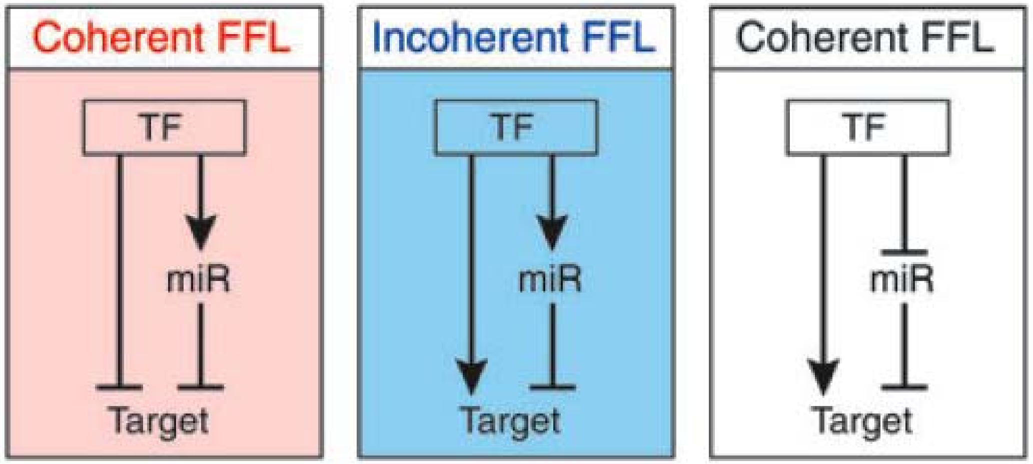
Network Motifs

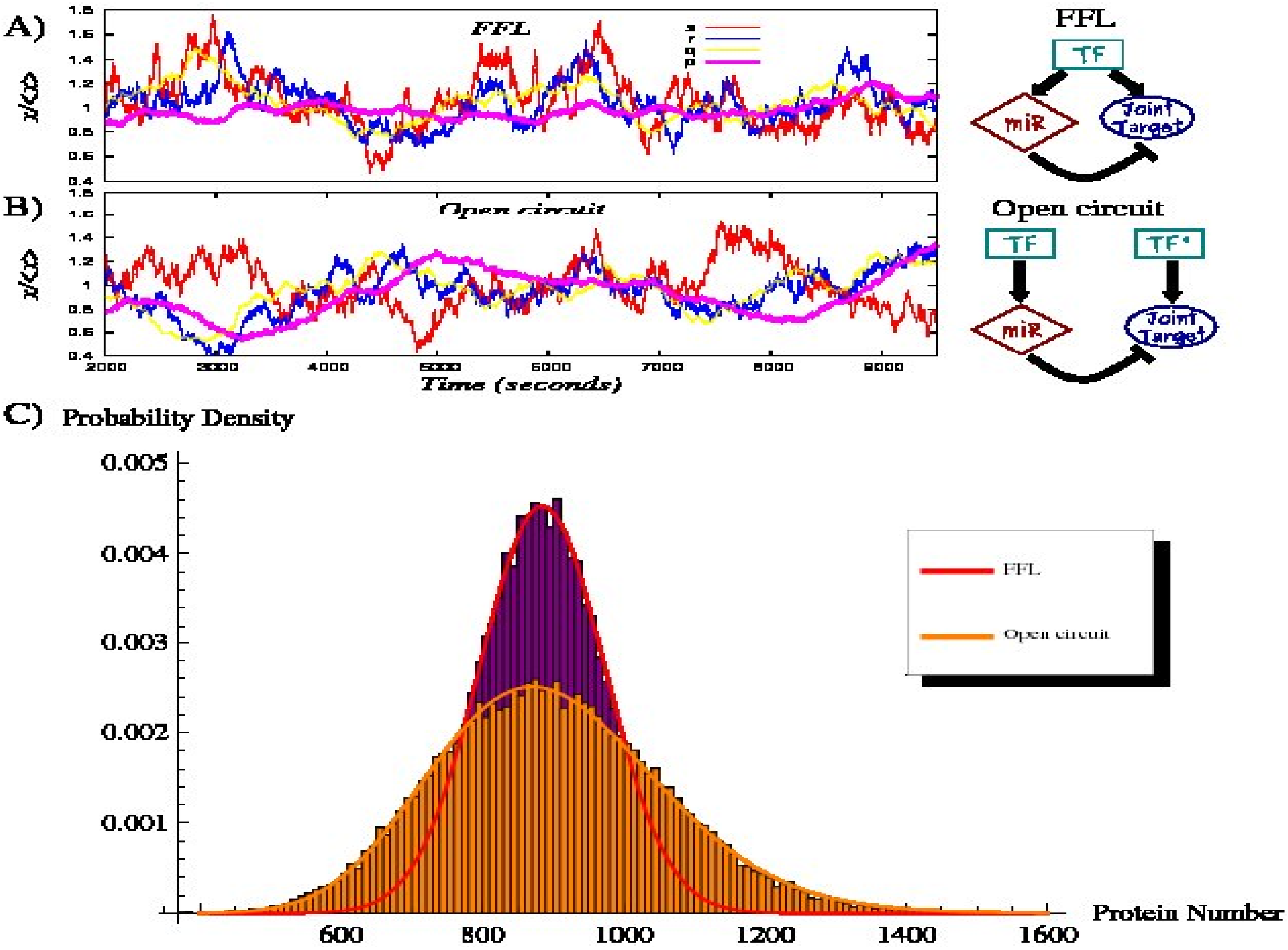
Network motifs can be studied using standard tools of theoretical physics:

- **Ordinary differential equations**
- **Stochastic equations**
- **Montecarlo (Gillespie) simulations.**

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

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




A)
FFL
s
r
g
p
x/<x>
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
FFL
TF
miR
Joint Target
B)
Open circuit
x/<x>
2000
3000
4000
5000
6000
7000
8000
9000
Time (seconds)
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

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

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

Backup Slides

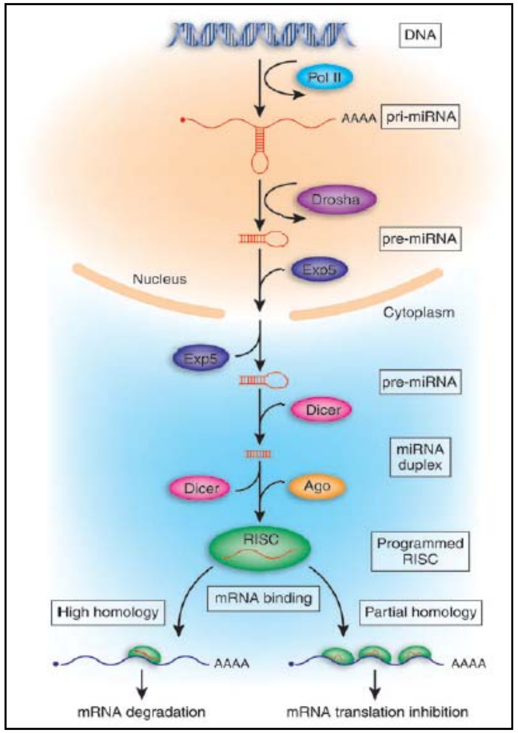
MicroRNA biogenesis

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.