

02-716

Cross species analysis of genomics data

Ziv Bar-Joseph
zivbj@cs.cmu.edu
GHC 8006

<http://www.cs.cmu.edu/~zivbj/class11/crossSP.html>

Topics

- Introduction
- Sequence: Whole genome alignments, motifs and microRNAs
- Expression: Gene expression in similar studies across species and databases
- Interactions: Protein-protein, protein-DNA and genetic interaction networks across species

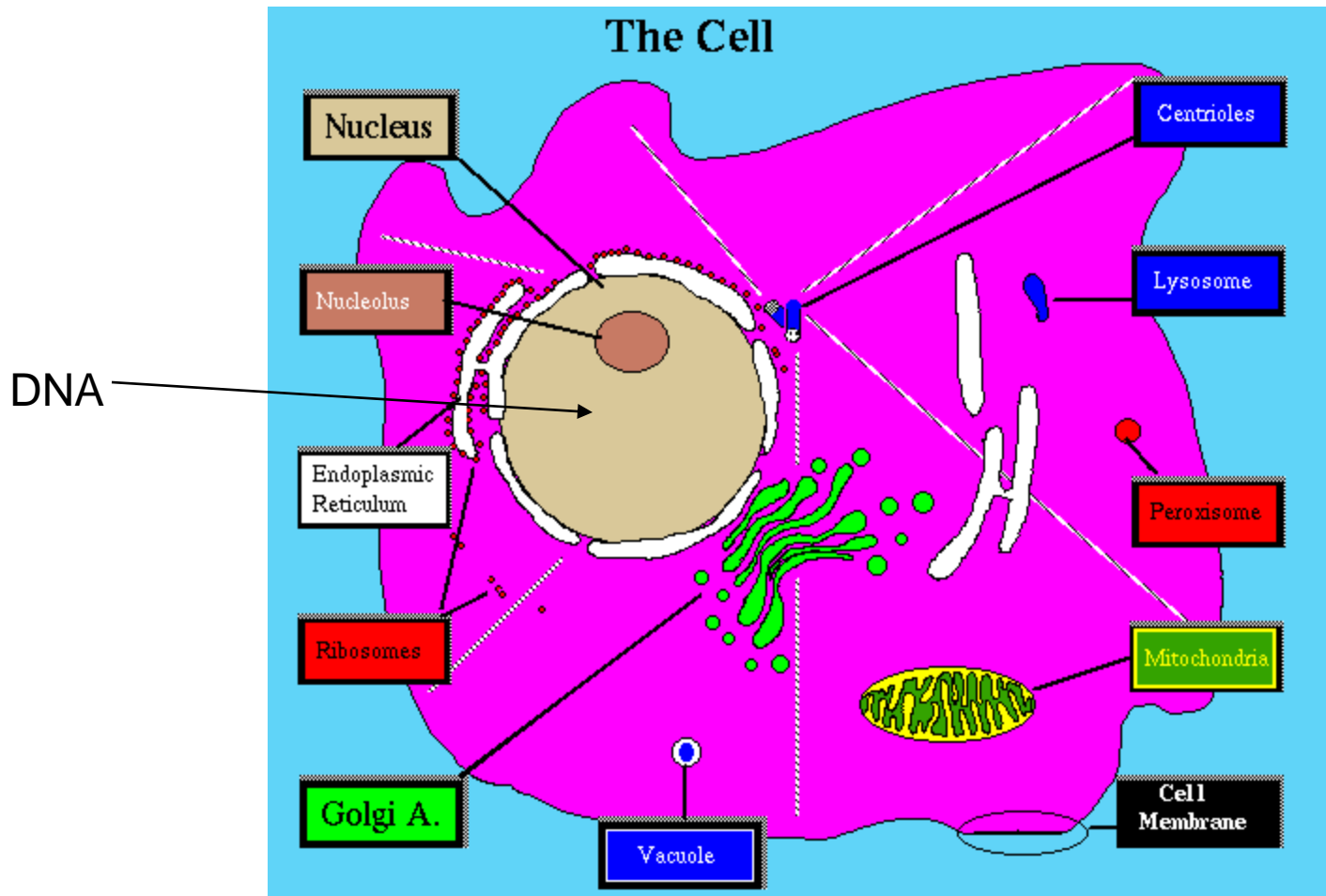
Grades

- Class presentations: 50%
- Projects: 35%
- Class participation and reading: 15%

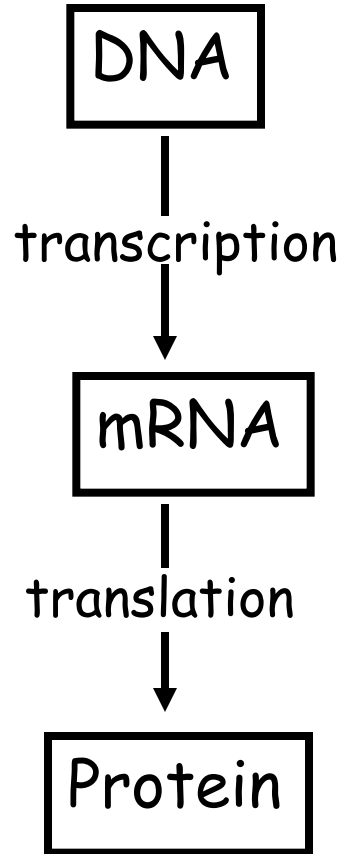
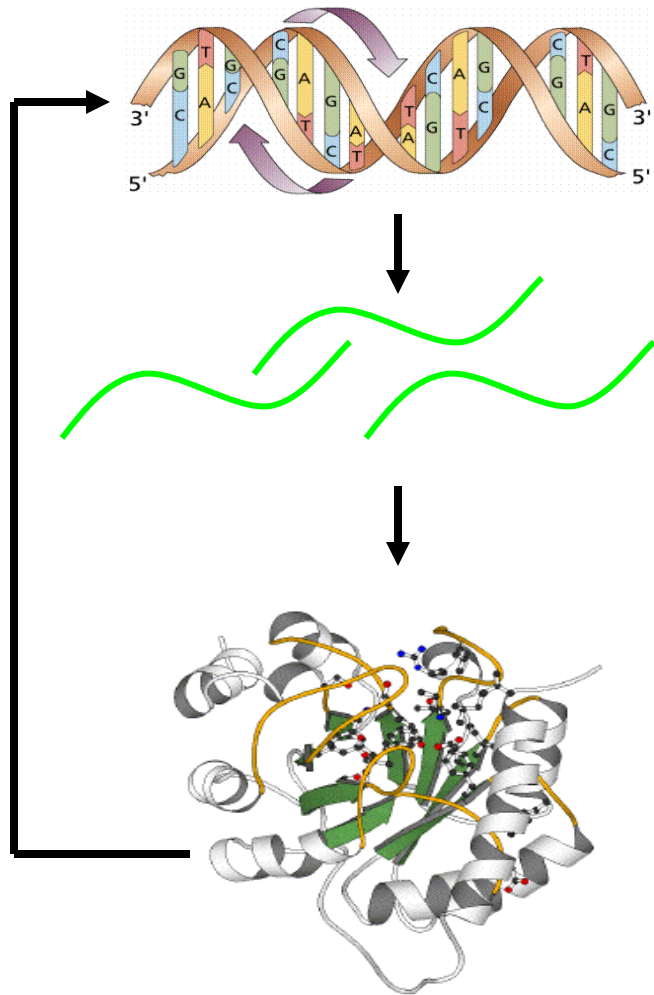
Introduction to Molecular Biology (with special emphasis on integrating data from multiple species)

- Genomes
- Genes
- Regulation
- microRNAs
- mRNAs
- Proteins
- Biological systems

The Eukaryotic Cell



Central dogma



CCTGAGCCAAC TATTGATGAA

CCUGAGCCAACU AUUGAUGAA

PEPTIDE

Genes and Genomes

Genome

- A genome is an organism's complete set of DNA (including its genes).
- In humans less than 3% of the genome actually encodes for genes. In other organisms that percentage is much higher.
- A part of the rest of the genome serves as a control regions (though that's also a small part).
- The goal of the rest of the genome is not completely known (it includes non coding RNAs, control regions, enhancer regions, etc.)

Comparison of Different Organisms

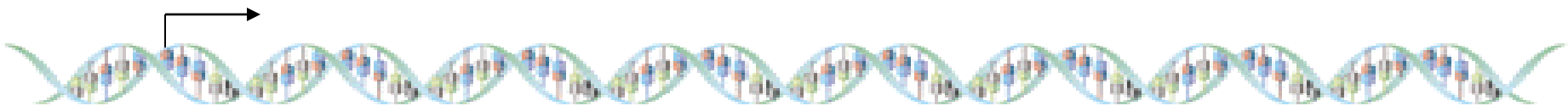
	Genome size	Num. of genes
E. coli	$.05 \times 10^8$	4,200
Yeast	$.15 \times 10^8$	6,000
Worm	1×10^8	18,400
Fly	1.8×10^8	13,600
Human	30×10^8	25,000
Plant	1.3×10^8	25,000

What is a gene?

Promoter

Protein coding sequence

Terminator

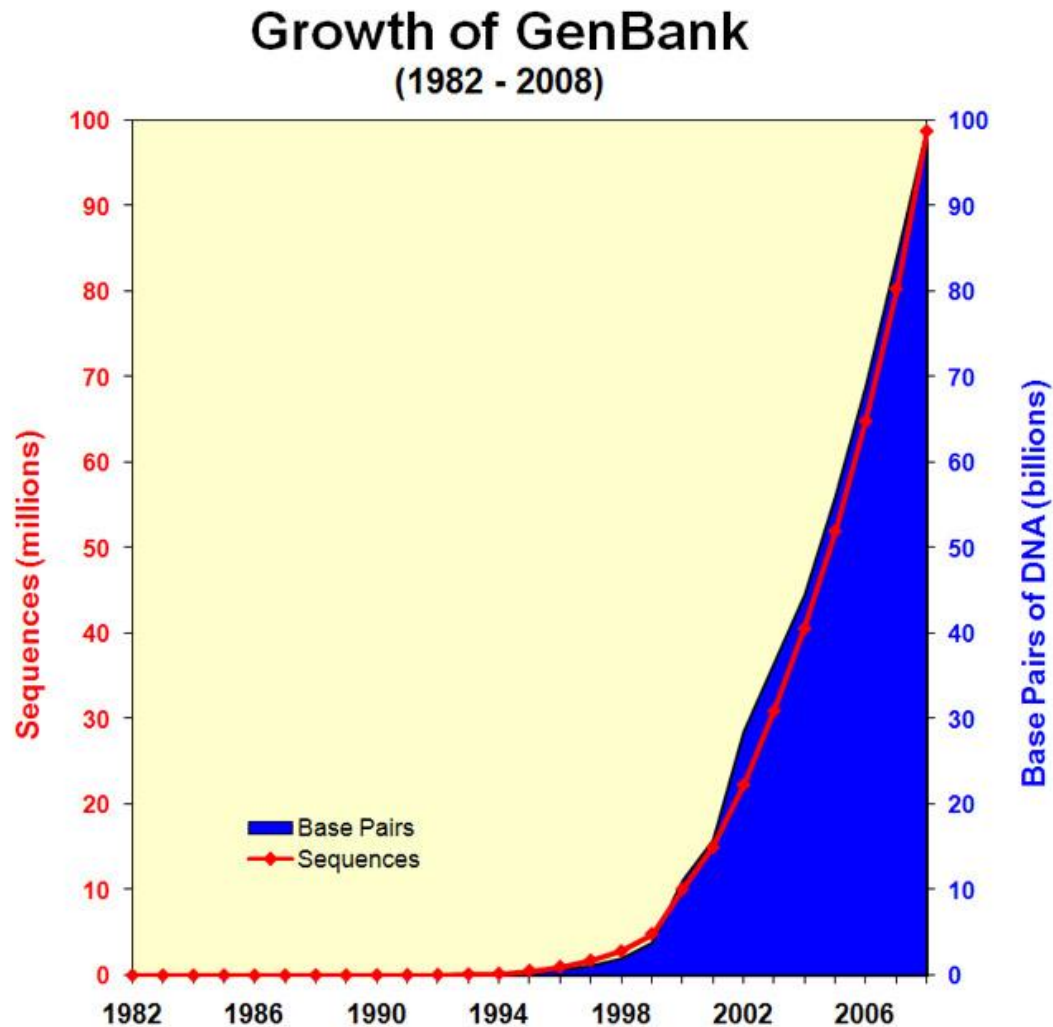


Genomic DNA

Genes Encode for Proteins

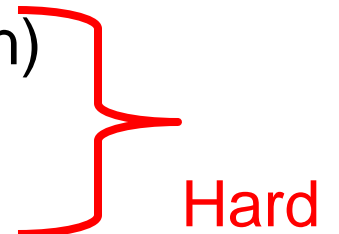
		Second Letter					
		U	C	A	G		
1st letter	U	UUU Phe UUC UUA Leu UUG	UCU UCC Ser UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G	3rd letter
	C	CUU CUC Leu CUA CUG	CCU CCC Pro CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG	U C A G	
	A	AUU AUC Ile AUA AUG Met	ACU ACC Thr ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G	
	G	GUU GUC Val GUA GUG	GCU GCC Ala GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG	U C A G	

Number of Genes in Public Databases




Assigning function to genes / proteins

- One of the main goals of molecular (and computational) biology.
- There are 25000 human genes and the vast majority of their functions is still unknown
- Several ways to determine function
 - Direct experiments (knockout, overexpression)
 - Interacting partners
 - 3D structures
 - Sequence homology



Hard



Easier

Function from sequence homology

- We have a query gene: **ACTGGTGTACCGAT**
- Given a database with genes with a known function, our goal is to find another gene with similar sequence (possibly in another organism)
- When we find such gene we predict the function of the query gene to be similar to the resulting database gene
- Problems
 - How do we determine similarity?

Sequence analysis techniques

- A major area of research within computational biology.
- Initially, based on deterministic (dynamic programming) or heuristic (Blast) alignment methods
- More recently, based on probabilistic inference methods (HMMs).

Identifying Genes in Sequence Data

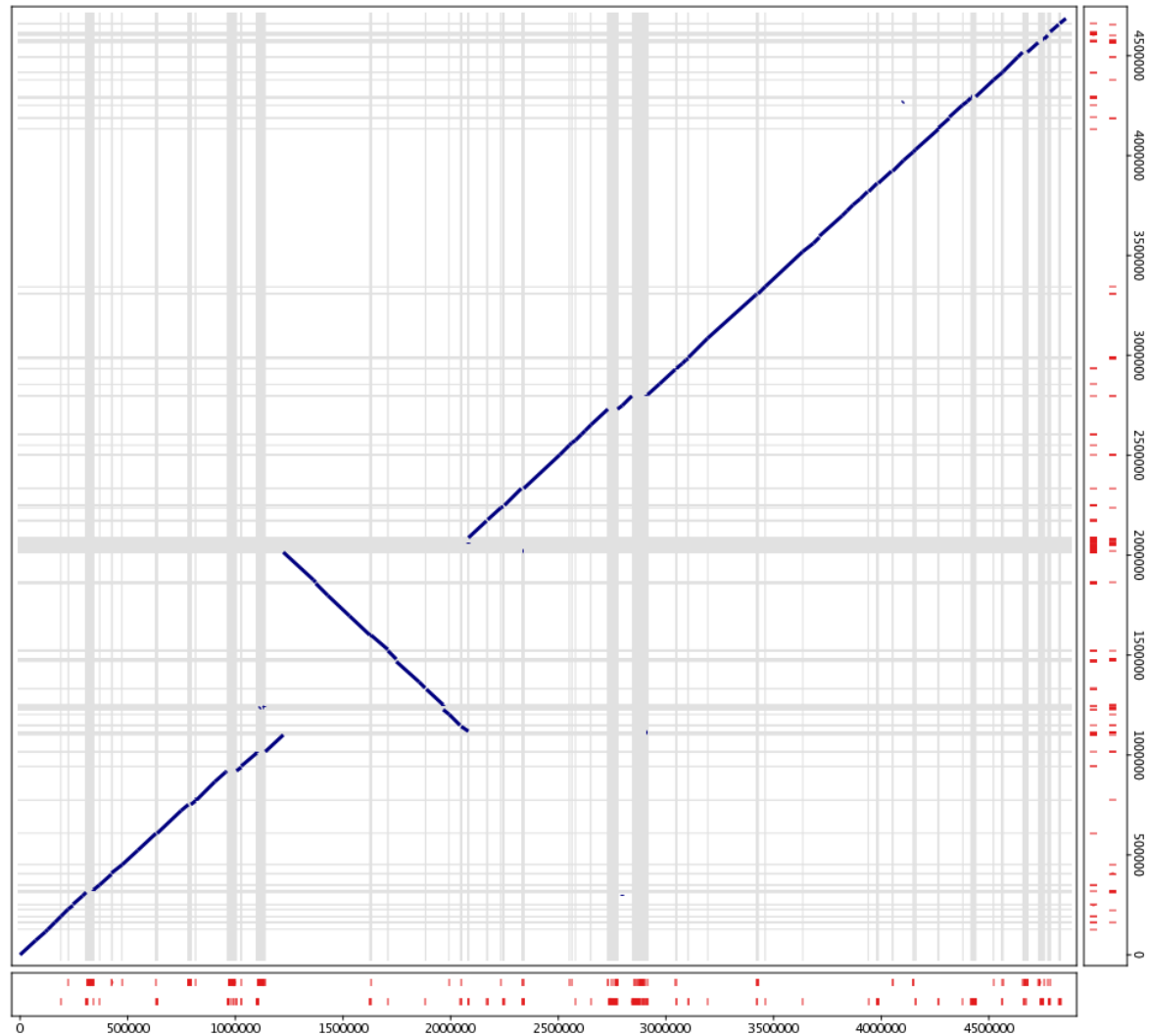
- Predicting the start and end of genes as well as the introns and exons in each gene is one of the basic problems in computational biology.

In many cases cross species analysis is used to improve gene finding and alternative splicing analysis

the start codon, end with one of the end codons, and do not contain any other end codon in between.

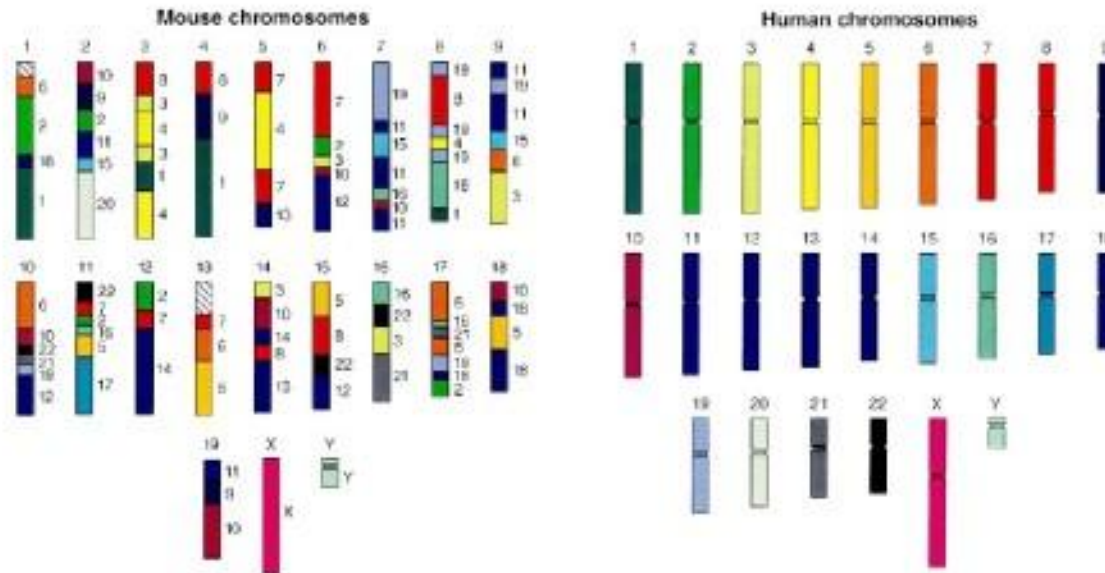
- Splice site prediction has received a lot of attention in the literature.

Whole Genome Alignment



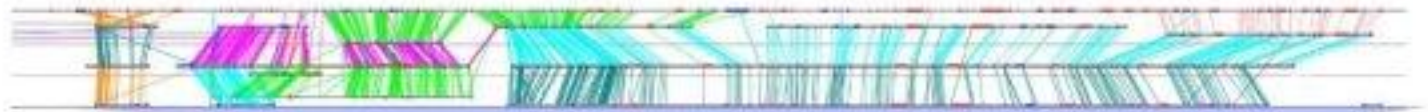
Comparative genomics

Mouse and Human Genetic Similarities

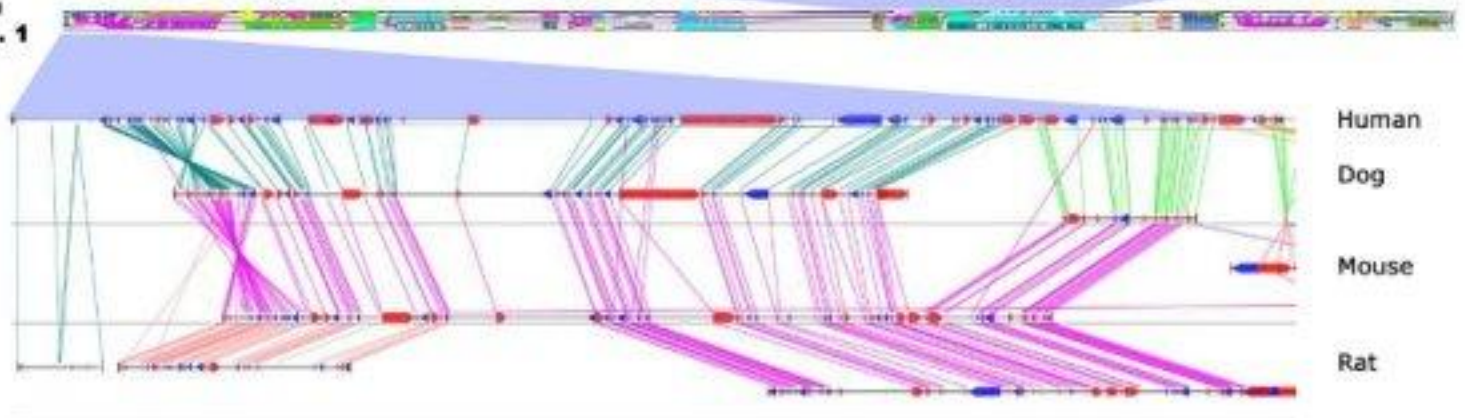


VCA 94-07582

Courtesy Lisa Stubbs
Oak Ridge National Laboratory



**human
chrom. 1**



Human

Dog

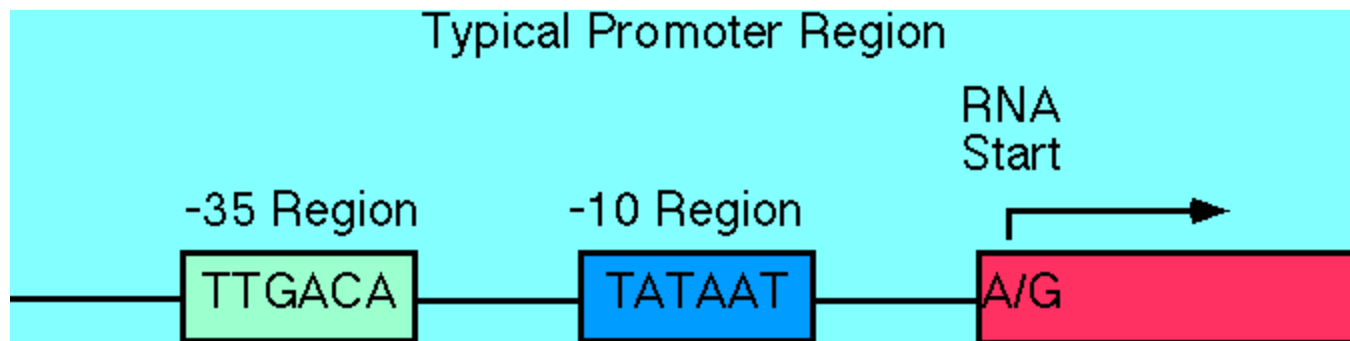
Mouse

Rat

Regulatory Regions

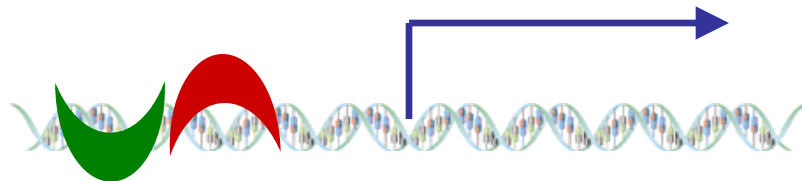
Promoter

The promoter is the place where RNA polymerase binds to start transcription. This is what determines which strand is the coding strand.

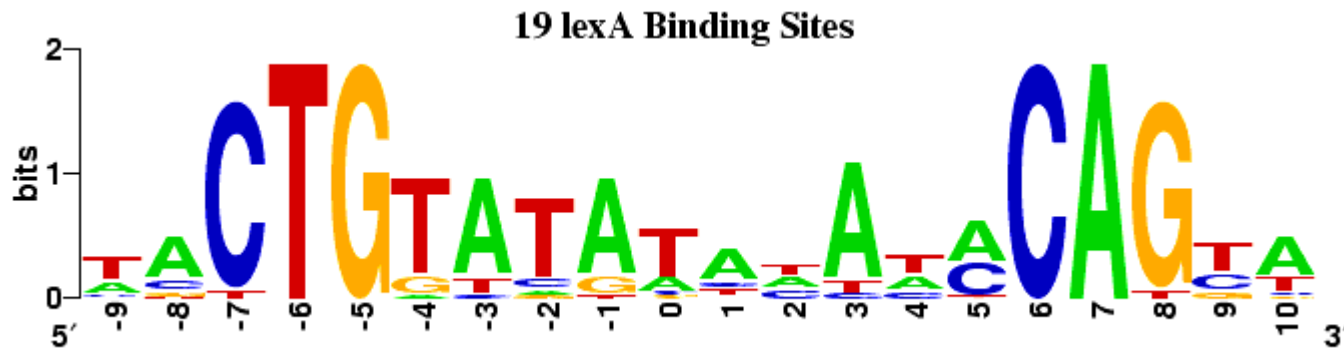


DNA Binding Motifs

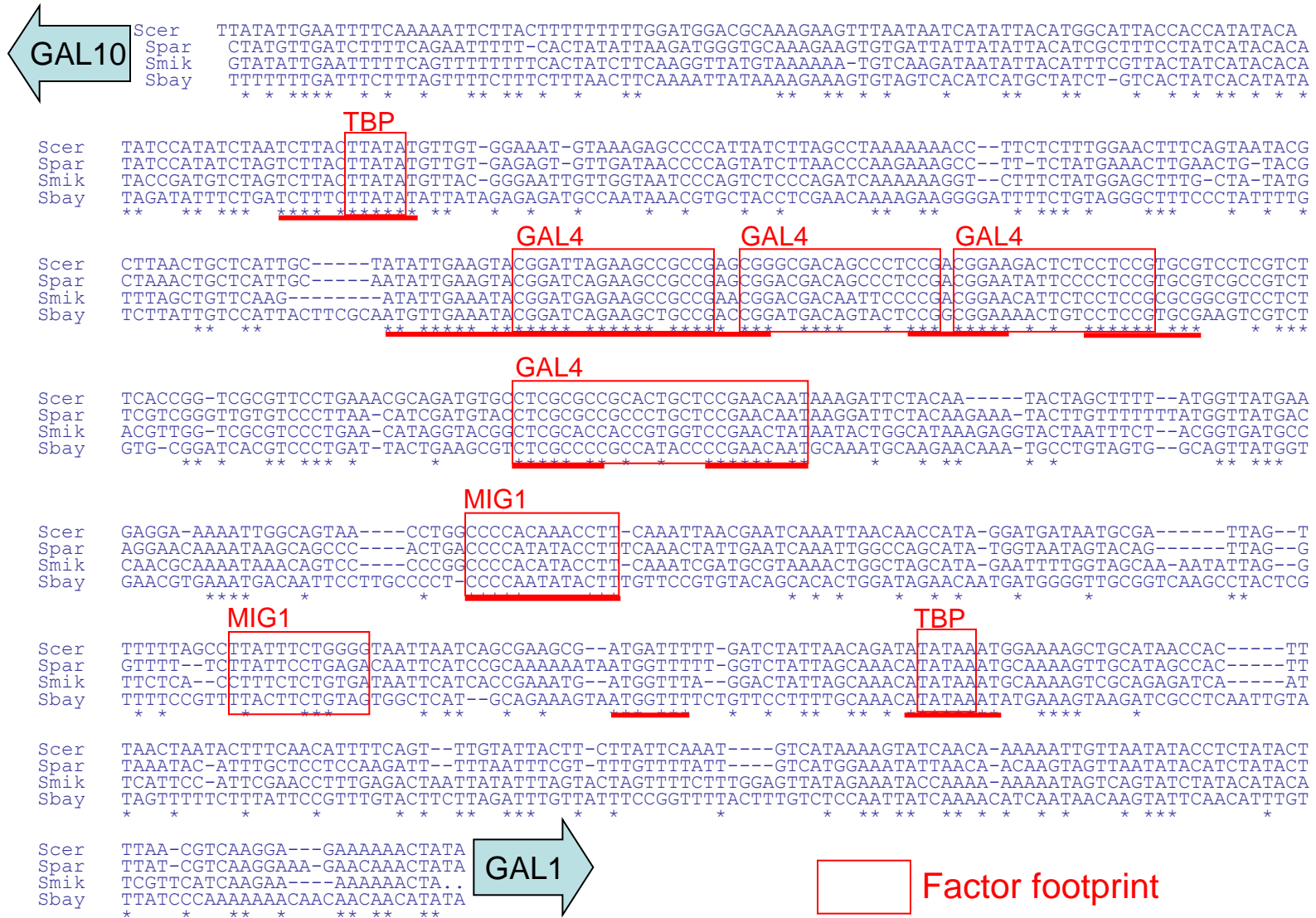
- In order to recruit the transcriptional machinery, a transcription factor (TF) needs to bind the DNA in front of the gene.
- TFs bind in to short segments which are known as DNA binding motifs.
- Usually consists 6 – 8 letters, and in many cases these letters generate palindromes.



Example of Motifs



Conservation of Motifs



RNA: mRNA, microRNAs

RNA

Four major types (one recently discovered regulatory RNA).

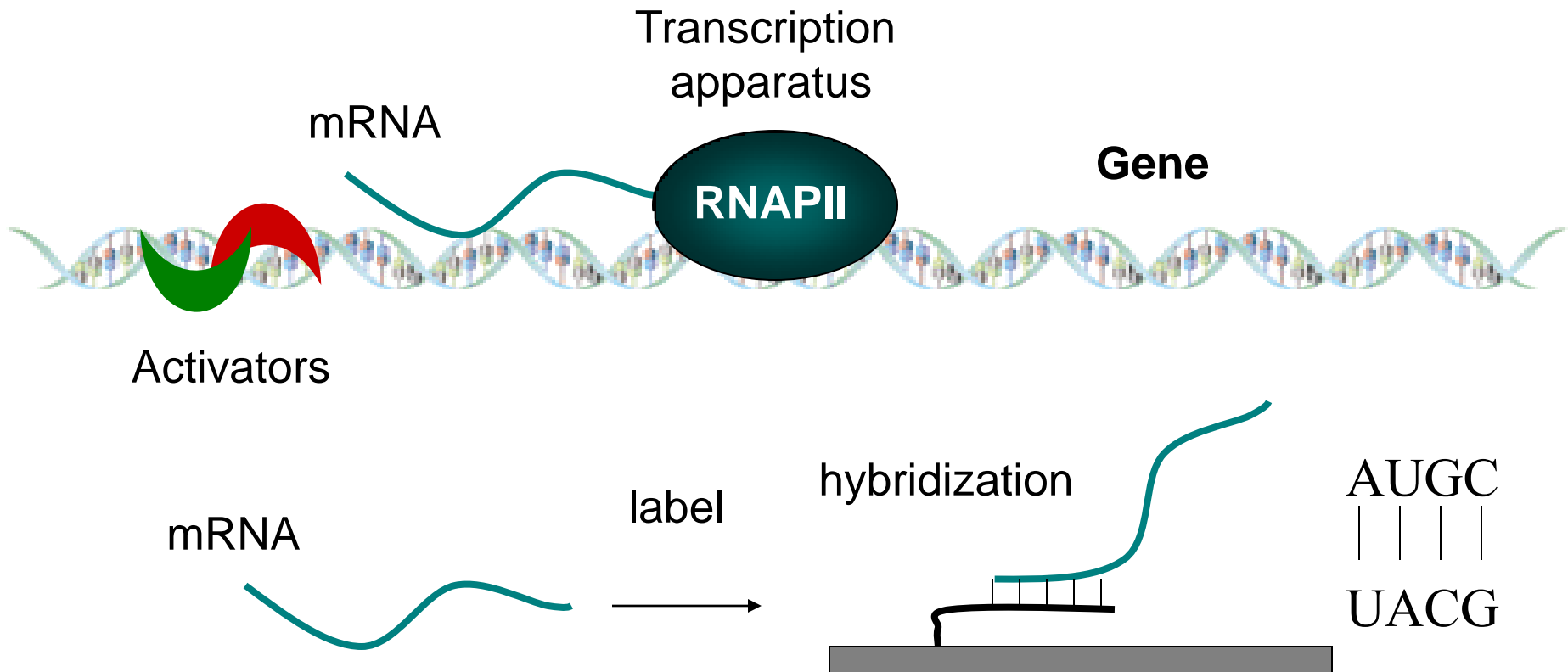
- mRNA – messenger RNA
- tRNA – Transfer RNA
- rRNA – ribosomal RNA
- microRNA – small non coding RNAs

Messenger RNA

- Basically, an intermediate product
- Transcribed from the genome and translated into protein
- Number of copies correlates well with number of proteins for the gene.
- Unlike DNA, the amount of messenger RNA (as well as the number of proteins) differs between different cell types and under different conditions.

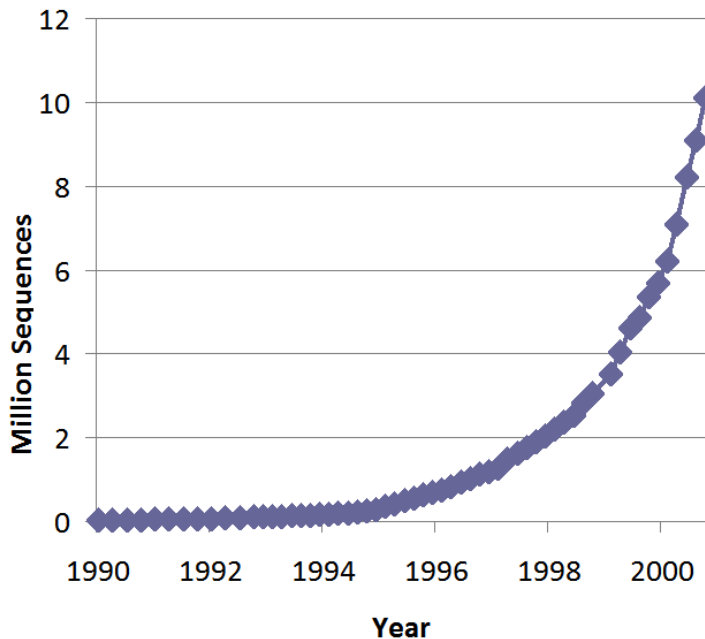
Complementary base-pairing

- mRNA is transcribed from the DNA
- mRNA (like DNA, but unlike proteins) binds to its complement

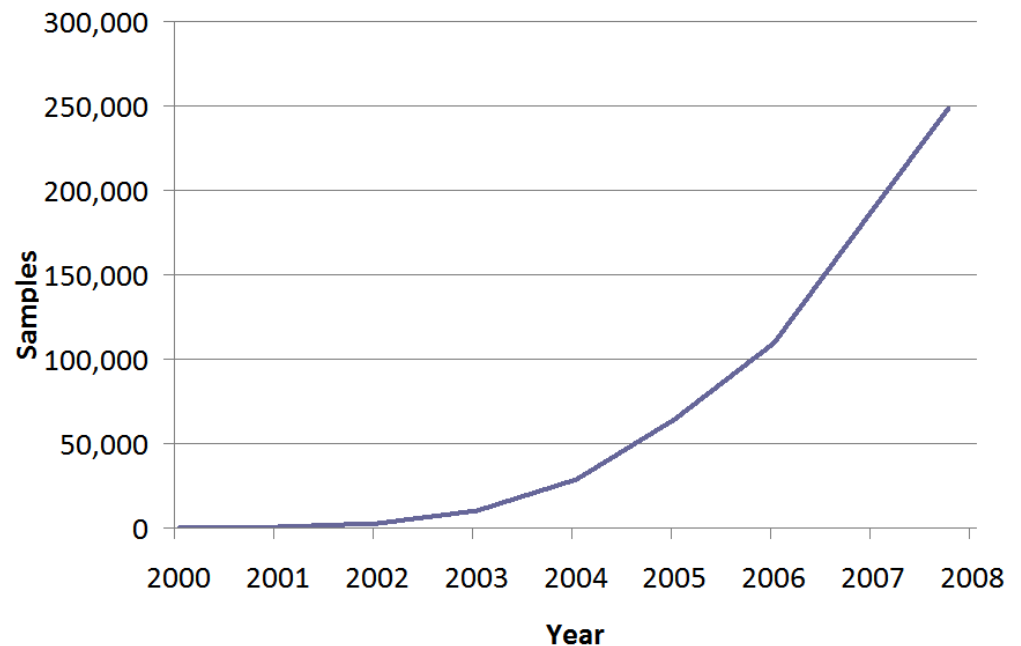


Expression databases are also exploding

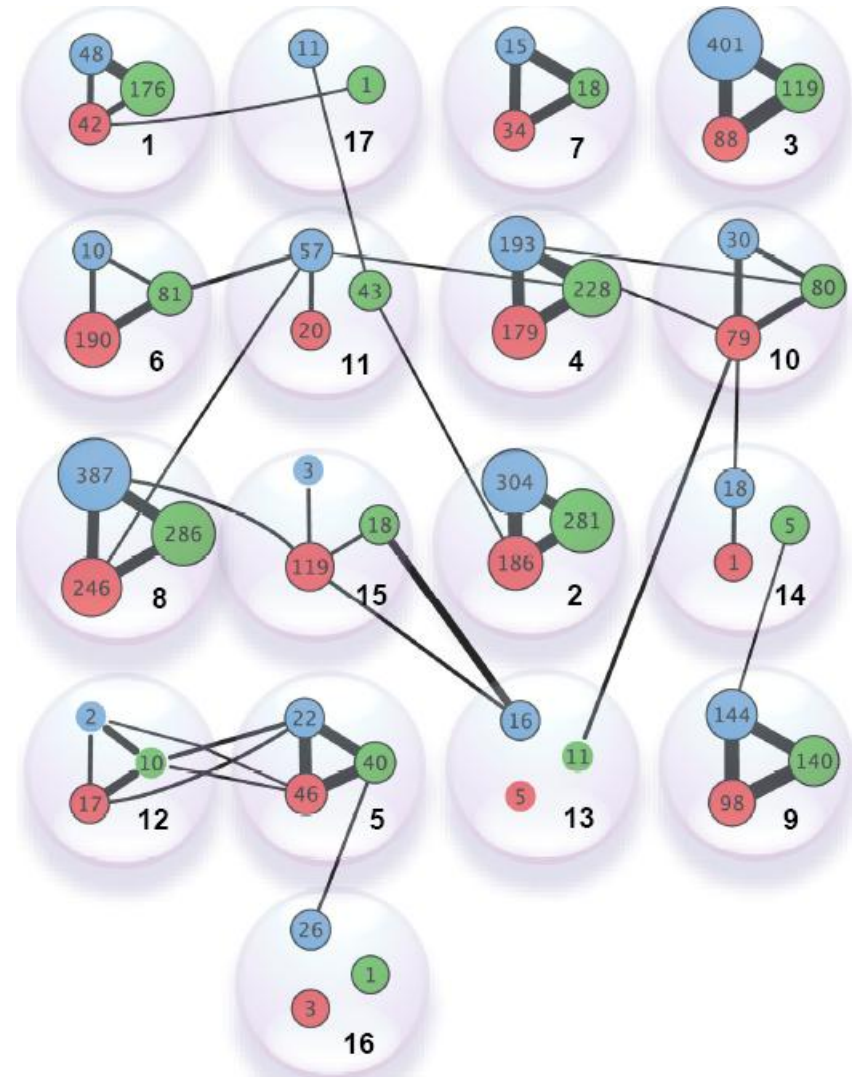
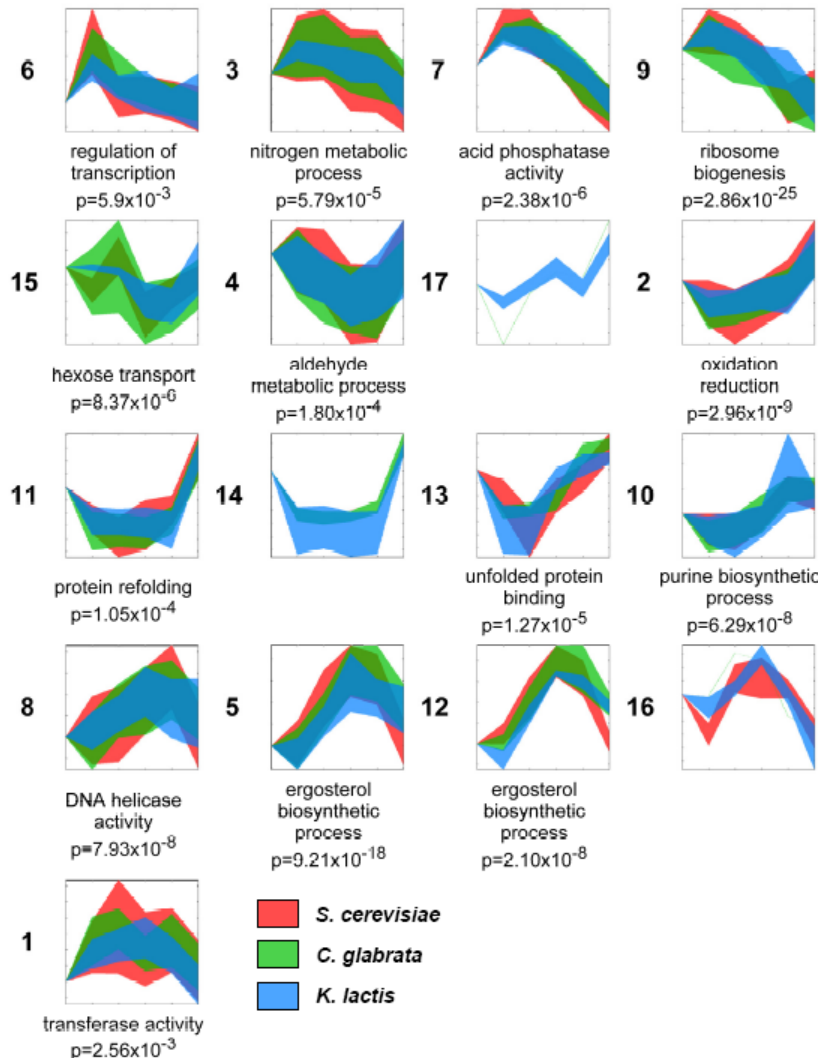
Growth of GenBank in the 90's



Growth of GEO in 2000's

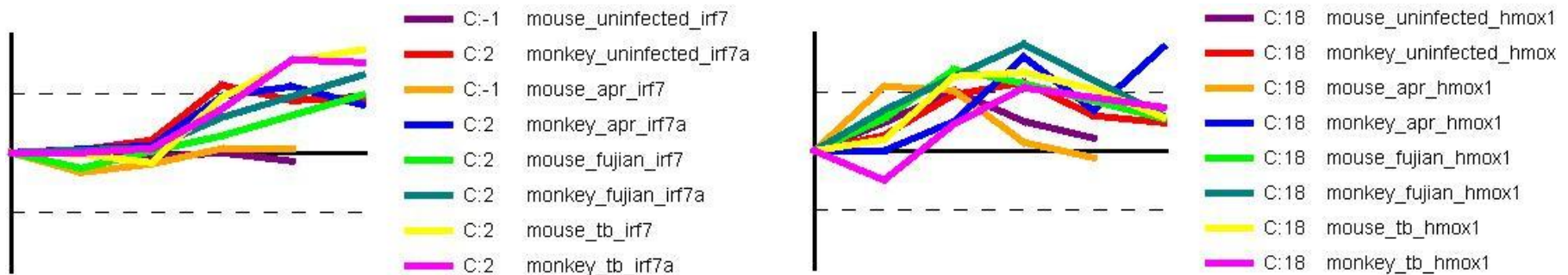


Comparison of expression changes in yeast response to drugs

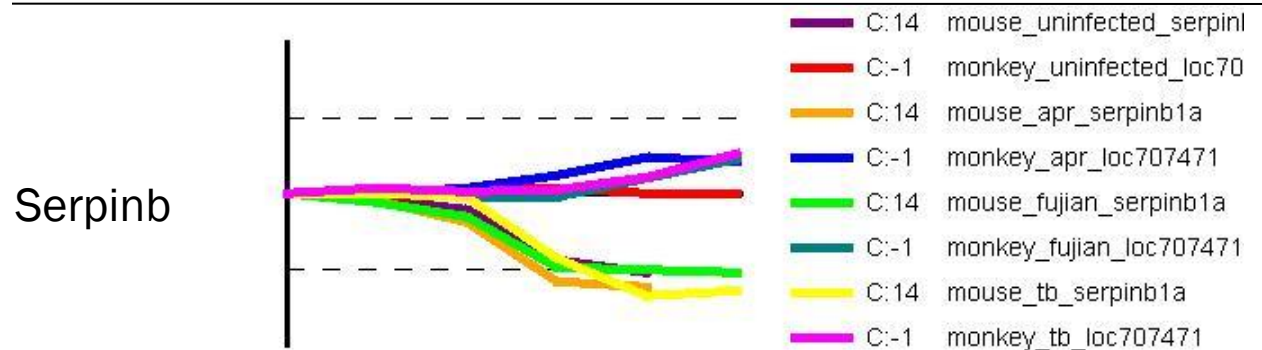
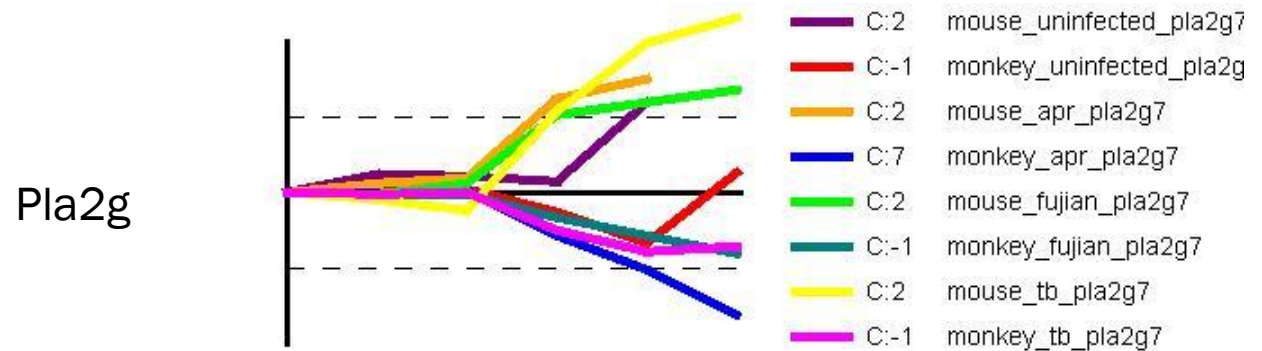
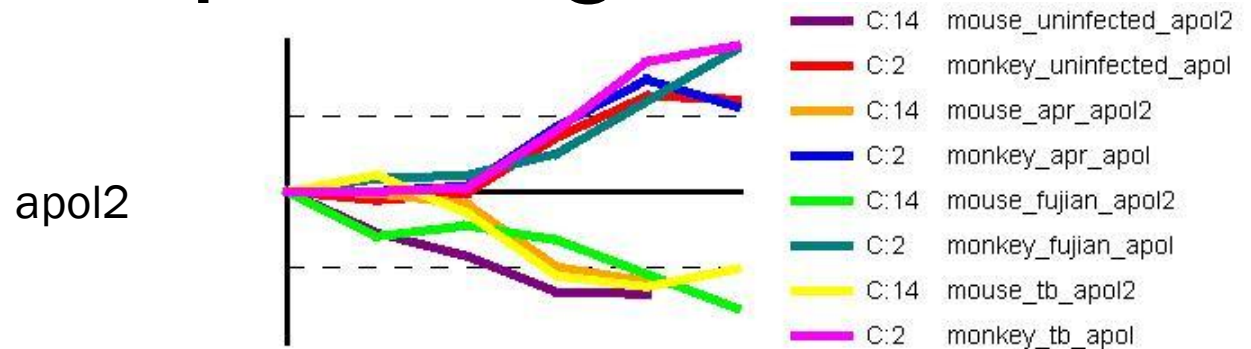


Comparison of immune response in mice and monkeys

- Looked for common and unique response patterns.
- Initial goal: core genes; genes that respond in a similar way to all infections in both species.

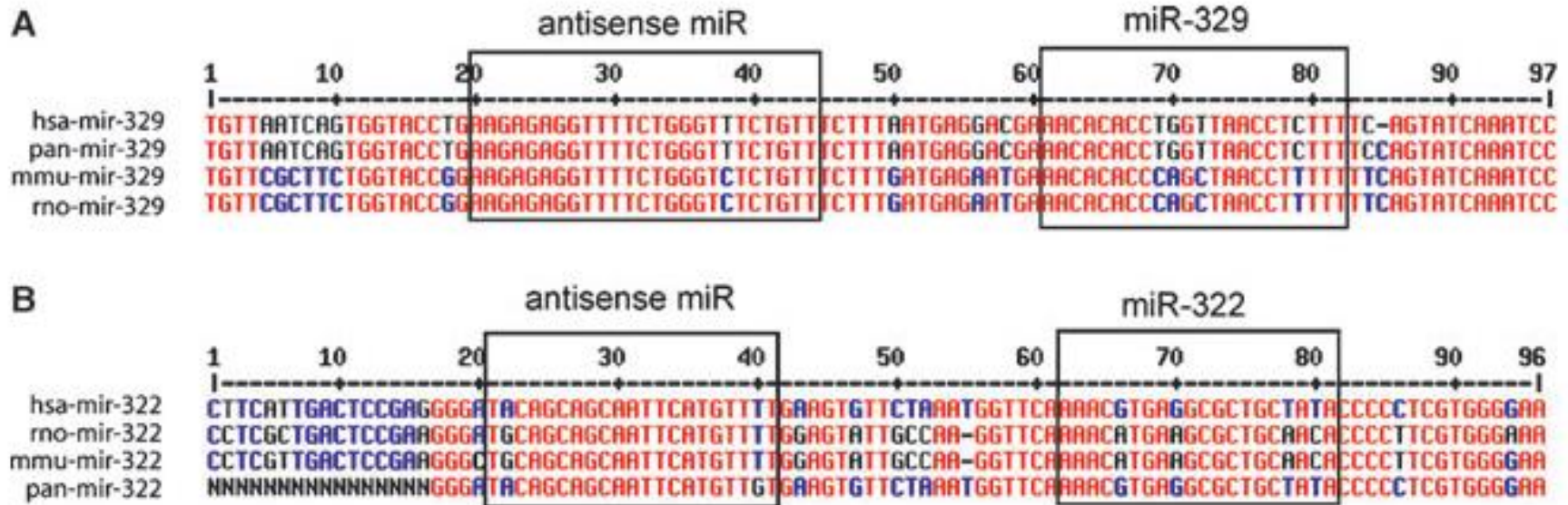


Species – specific genes



➤ Gene are more differentiated among animals than among pathogens

60 new potential miRNAs (15 for human and 45 for mouse)



- Mature miRNA were either perfectly conserved or differed by only 1 nucleotide between human and mouse.

Interactions: Protein-protein,
protein-DNA, genetic etc.

Protein Interaction

In order to fulfill their function, proteins interact with other proteins in a number of ways including:

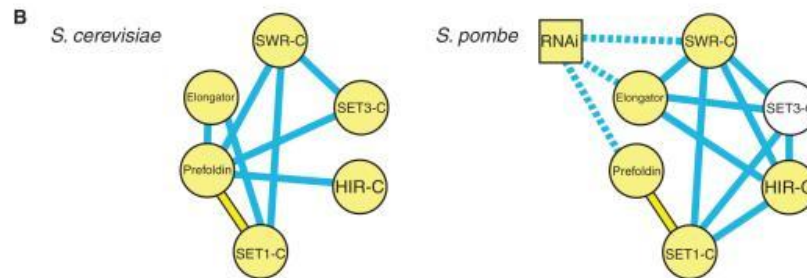
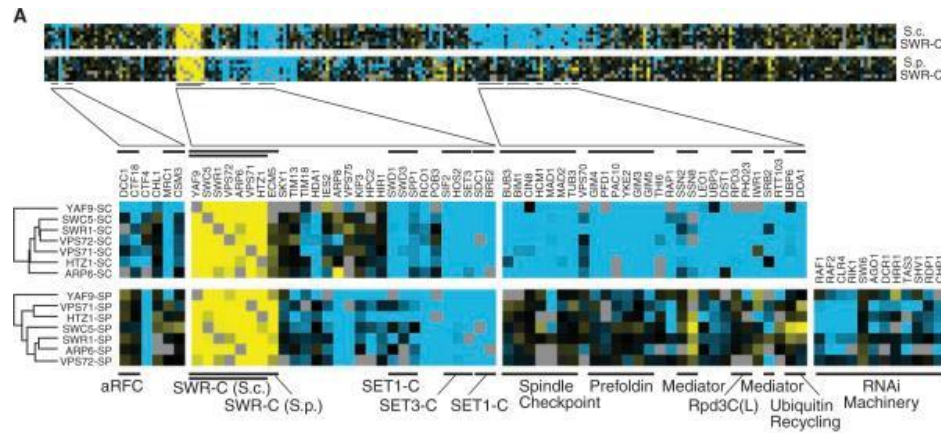
- Regulation
- Pathways, for example $A \rightarrow B \rightarrow C$
- Post translational modifications
- Forming protein complexes

Other types of interactions

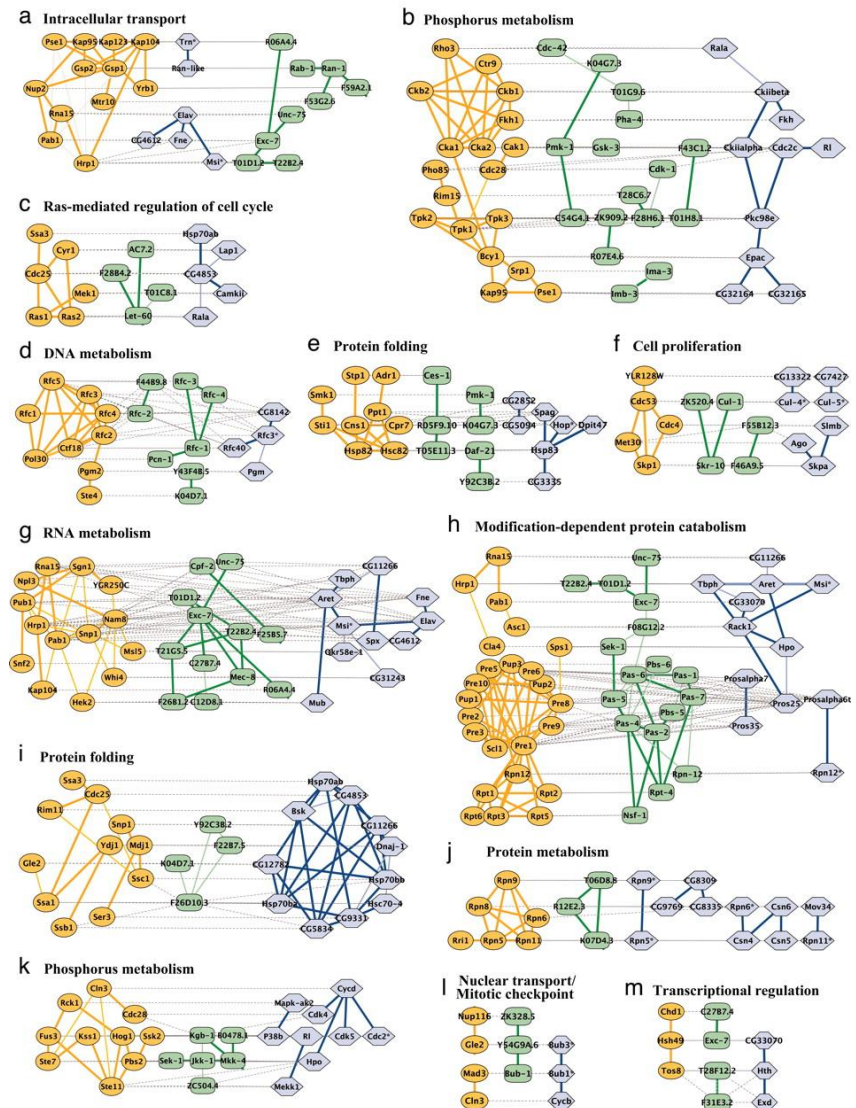
In addition to protein-DNA and protein-protein interactions there are many other physical interactions that are studied using cross species analysis. These include:

- Genetic interactions
- microRNA-mRNA interactions
- Co-expression
- etc.

Genetic interactions



Sharan et al PNAS 2005

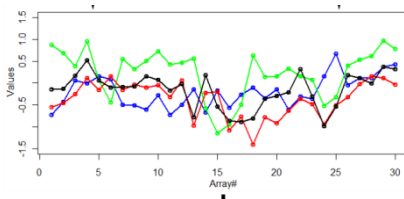


Not all interactions are well conserved ...

Sequence Highly conserved (90%
human / mouse, 50+%
budding and fission yeast)



Expression



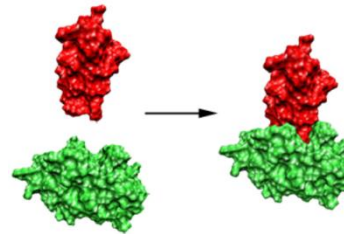
Weakly conserved. 15%
between the same
tissues in human and
mouse

Protein- DNA interactions



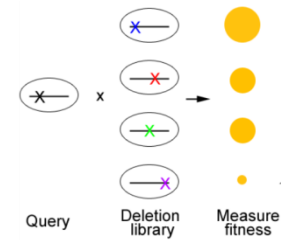
Weakly conserved (20%
between human and
mouse, 20% between 3
yeasts with 85%
sequence identify)

Protein- protein interactions



Weakly conserved.
~3% conservation
between the two
yeasts.

Genetic interactions (positive & negative)



Not conserved.
~2%
conservation
between the two
yeasts.

... but those within strong modules usually are

Budding and fission yeast DNA replication initiation modules

