


Bioinformatics

- “half a year in the lab can easily save you an afternoon in front of the computer....”

(unknown)

Bioinformatics - N 2004

- Home page: www.birc.dk/studies
- Book: Krane & Raymer “Fundamental Concepts of Bioinformatics”
- Teachers: Leif Schauser, Thomas Mailund and others
- Instructors: Philippe Lamy

Defining Bioinformatics.....

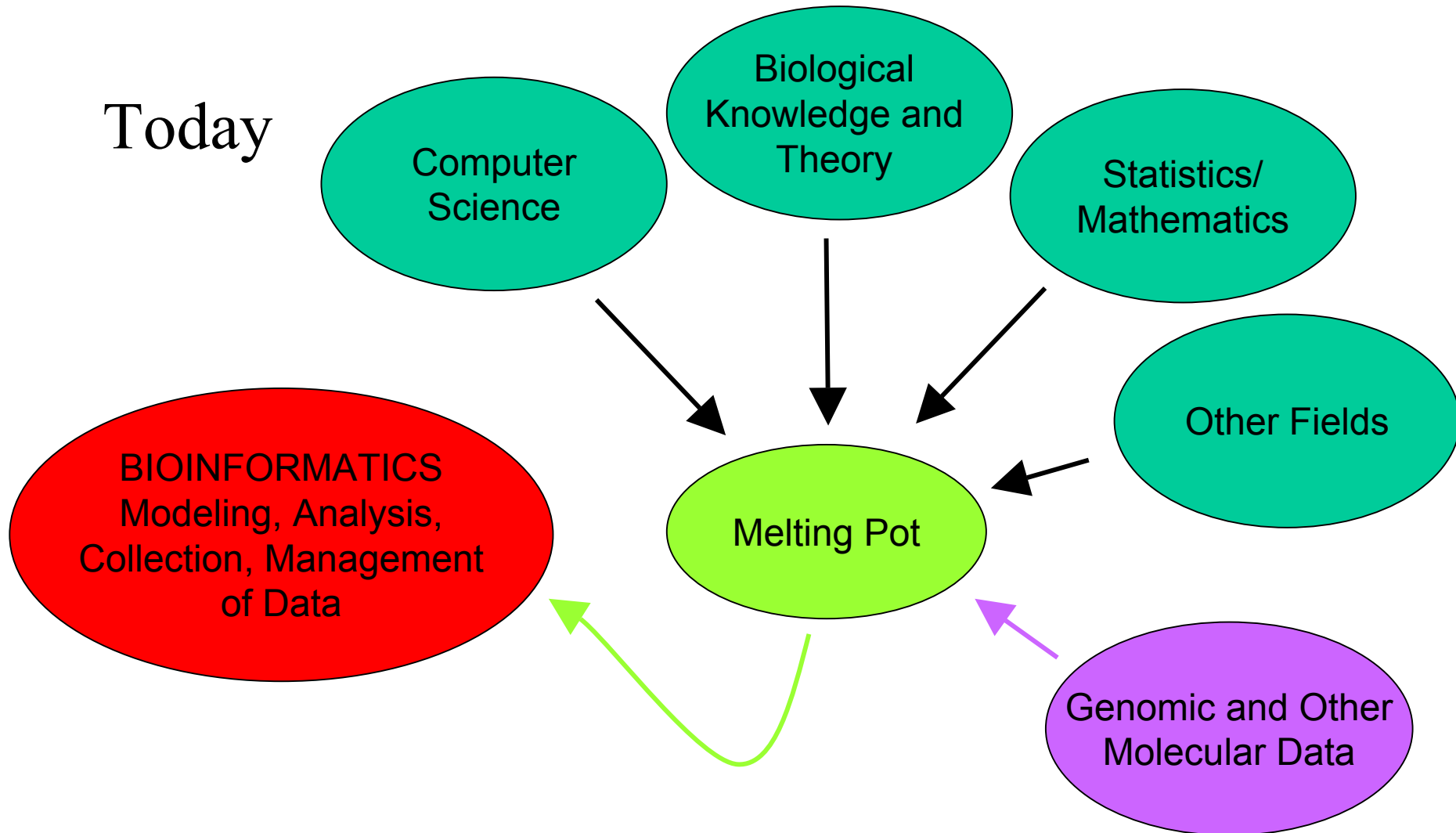
The term Bioinformatics was introduced in 1988 by Hwa Lim, combining 'bio' and 'informatics' into one word

Early days

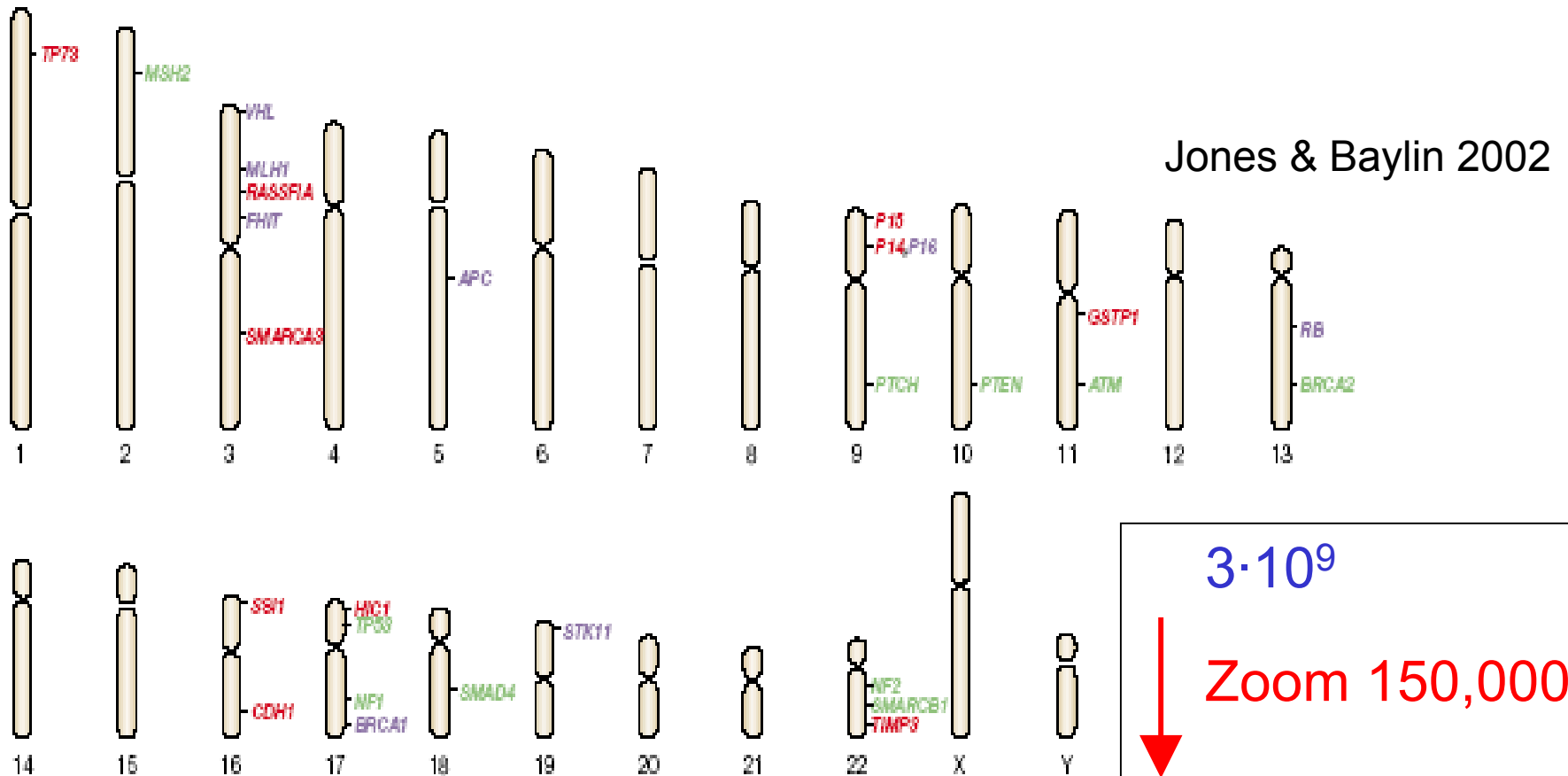


Data compilation, organization
Tracking, storing, retrieving ...
Data dissemination
Analysis (string algorithms)

Defining Bioinformatics.....



Our Own Genome



$3 \cdot 10^9$
↓
Zoom 150,000
 $2 \cdot 10^4$

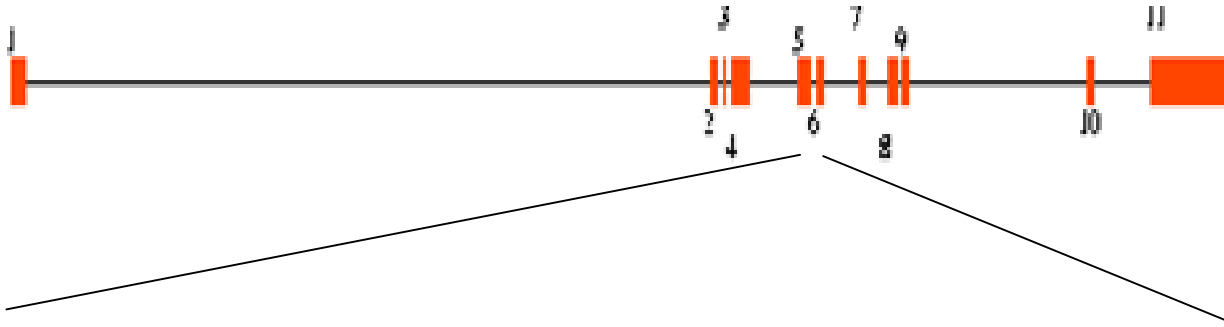
TP53



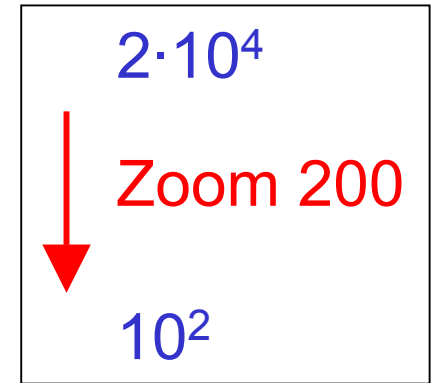
One of 30.000 genes

Our Own Genome

EXON 6 consists of 113 base pairs



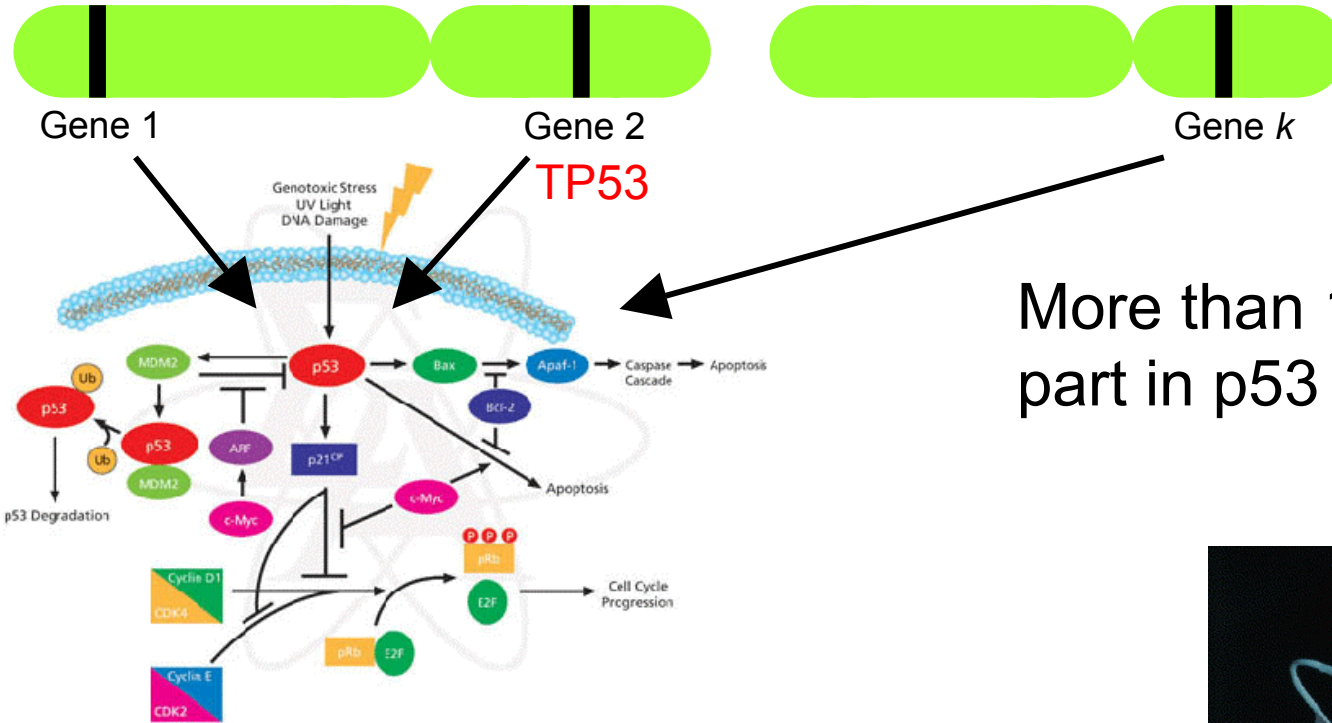
GTCTGGCCCCTCCTCAGCATCTTATCCGAGTGGAAGG
AAATTTGCGTGTGGAGTATTTGGATGACAGAAACACTT
TTCGACATAGTGTGGTGGTGCCCTATGAGCCGCCTGAG



...that are translated, joint with other exons, into phosphoprotein p53. Part of the protein reads

.....LWKLLPENVLSPLPSQAMDDL.....

Stepping Back



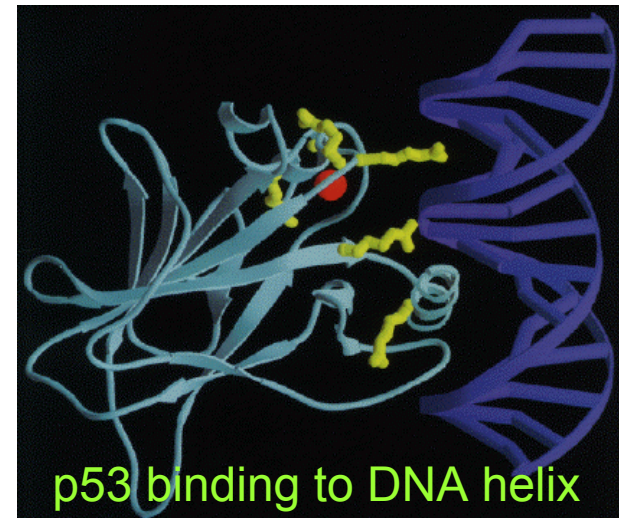
More than 100 genes take part in p53 pathways

The p53 signaling pathway

Networks: Pathways, interactions

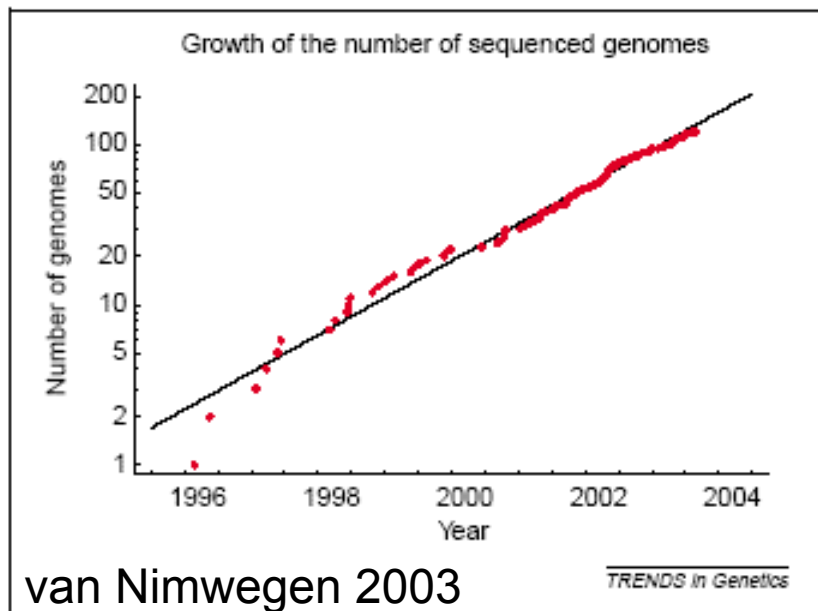
Qualitative Features: size, directions

Quantitative Features: expression levels

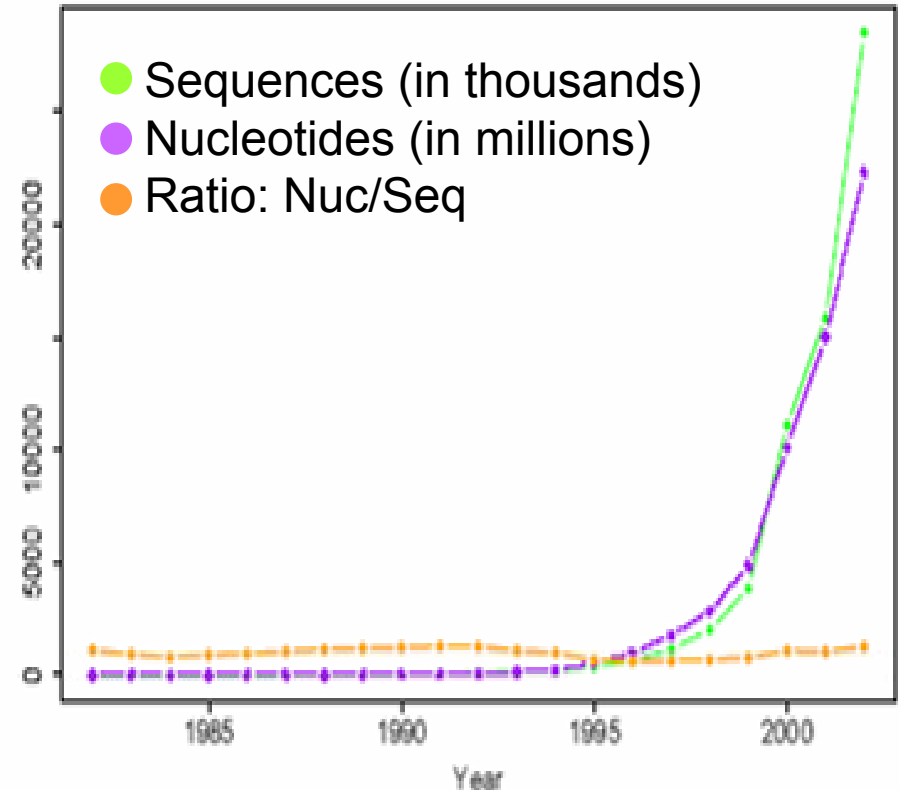


The Amount of Data

Over the years there has been an explosion in the amount of data

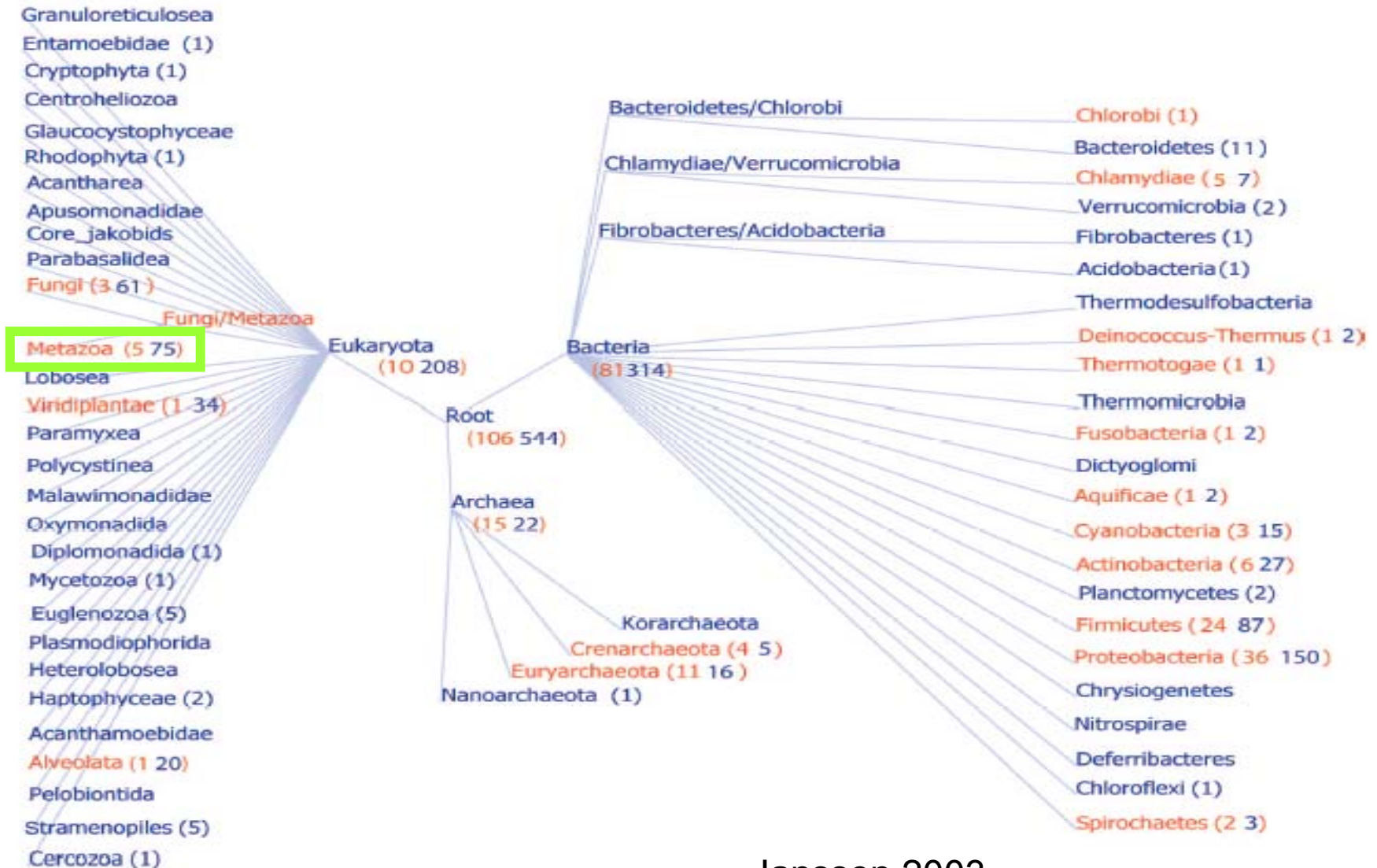


GenBank webpage 2004



Improved technology
More resources
'Hot' research areas

The Amount of Data



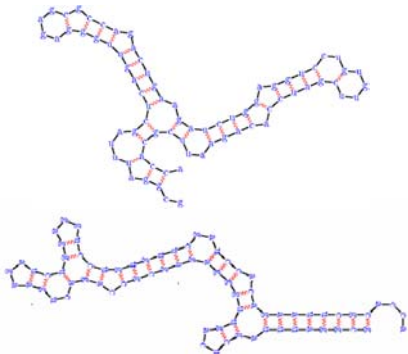
Leaving the cell

Species Level

DNA sequence

Man: ATTCGT
 Mouse: ATCCT
 Zebra: TTCAATA

Other structures



Population Level

DNA sequence

1st: AAAGTACG
 2nd: AACGTACG
 3rd: AAAGTATG

Gene expression

1st: High
 2nd: High
 3rd: Low

Individual Level

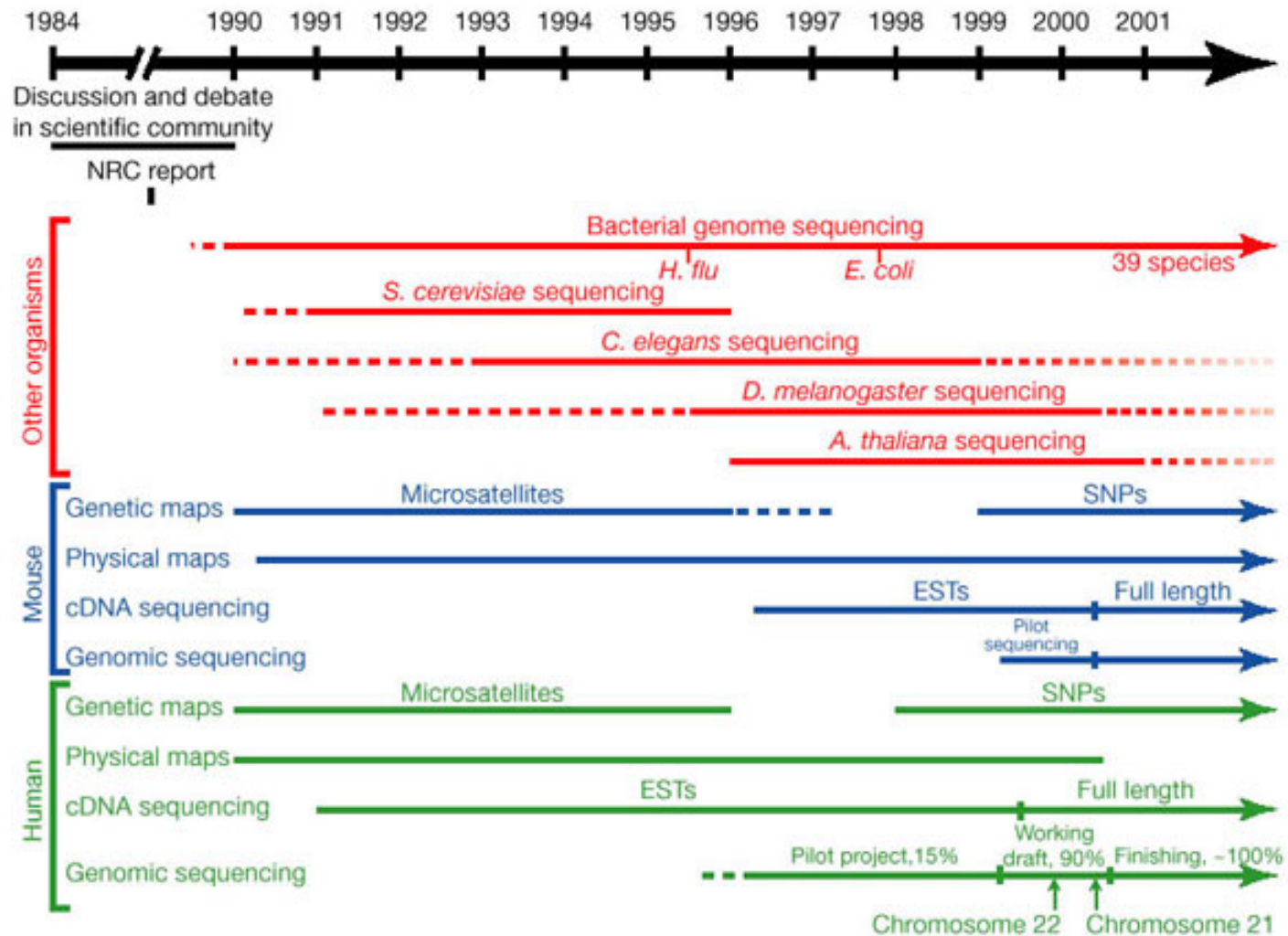
Gene expression

Tumor: High
 Blood: Low
 Liver: Medium

Gene activation

Tumor: Active
 Blood: Inactive

Variation within a structure at a given level



Genome projects

| Organism | Year | Size (Mbp) | # of Genes |
|---------------------------------|-------------|-------------------|-------------------|
| <i>Saccharomyces cerevisiae</i> | 1996 | 12 | 5.200 |
| <i>Caenorhabditis elegans</i> | 1998 | 100 | 19.000 |
| <i>Drosophila melanogaster</i> | 2000 | 115 | 13.000 |
| <i>Arabidopsis thaliana</i> | 2000 | 115 | 25.000 |
| <i>Human</i> | 2001 | 2700 | 39.000 |

Genomes

- First genomes were selected in order to reflect biological diversity.
- Database contains 20×10^9 bp
- Doubling time: 15 month
 - CPU doubling time 18 month
- Effective tools for sequence analysis needed

Bioinformatics Master

| | | |
|------------------------------|---|---|
| Masters Thesis | | |
| Algorithms in Bioinformatics | Complex systems | Protein structure |
| Algorithms and Datastructure | Molecular Population Genetics and Evolution | Biostatistics |
| Basics in Programming | Mathematics basic Molecular biology basics | Intro: Bioinformatics Genome analysis |

Topics

- Substitution matrices
- Pairwise alignment
- Multiple alignment
- Phylogenetic analysis
- Database searching
- RNA / protein secondary structure prediction
- Regulatory networks

Objectives

- Overview: understanding of topics and techniques
 - Motivation / principles
 - Mathematical and statistical models
 - Algorithms
- User-focus
 - When and how to use applications

Non-objectives

- To learn how to write programs
- To construct mathematical and statistical models
- To improve algorithms

Alignment: a central problem

- Alignments are basis of many analysis
 - Phylogeny reconstruction
 - Database searches
 - Predicting RNA / protein secondary structure
 - Genome analysis mm.

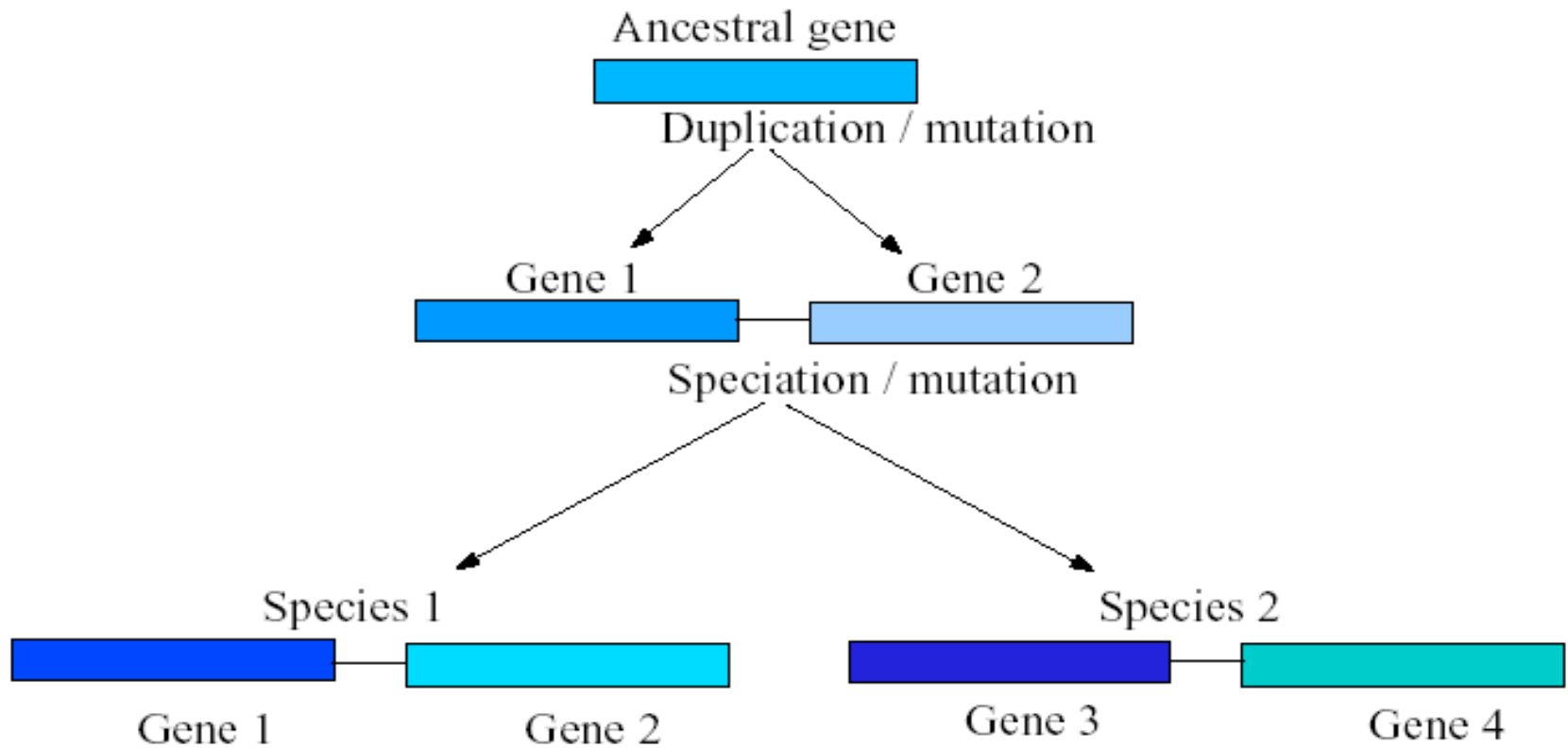
Why alignment?

- Discuss with your neighbour:
 - Which principle does an alignment represent

Why alignment?

- Biological sequences are related
 - Common ancestors
 - Duplication, mutation, speciation, variation
 - Principle of evolution

Why alignment?



What is an Alignment?

- HIGHLY RELATED:

```
HBA_HUMAN      GSAQVKGHGKKVADALTNVAHAVDDMPNALSALSDDLHAKL
                G+ +VK+HGKKV  A++++AH+D++  +++++LS+LH KL
HBB_HUMAN      GNPVKKAHGKKVLGAFSDGLAHL DNLKGT FATLSELHCDKL
```

- RELATED:

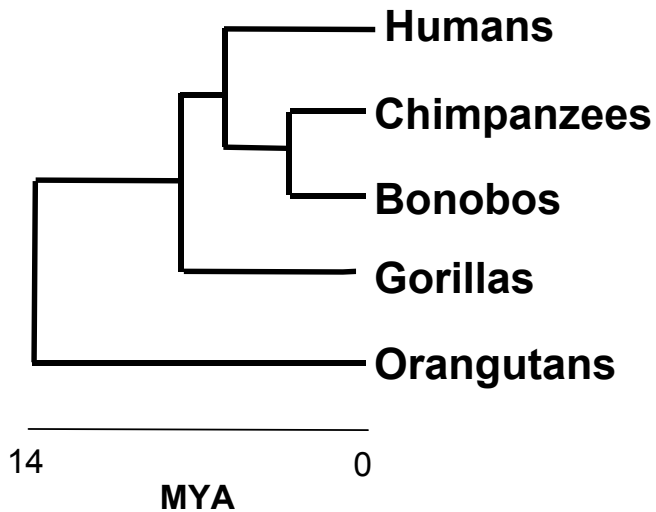
```
HBA_HUMAN      GSAQVKGHGKKVADALTNVAHV---D--DMPNALSALSDDLHAKL
                ++ ++++H+ KV   + +A   ++                +L  L+++H+ K
LGB2_LUPLU     NNPELQAHAGKVFKLVYEAAIQ LQVTGVVVTDATLKNLGSVHVS KG
```

- SPURIOUS ALIGNMENT:

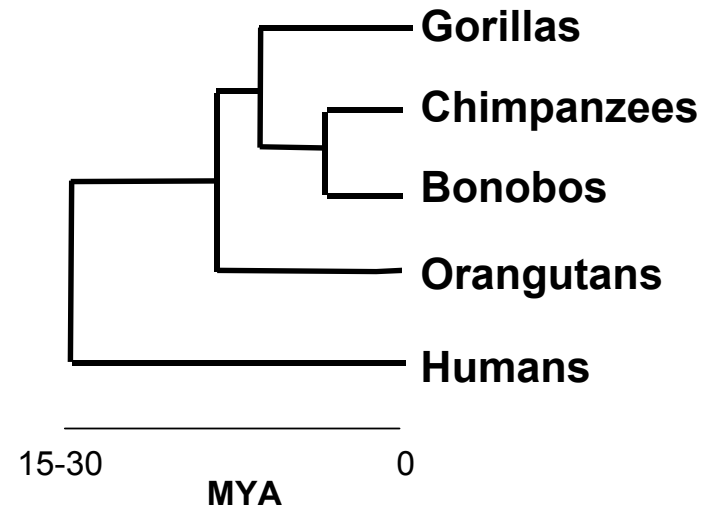
```
HBA_HUMAN      GSAQVKGHGKKVADALTNVAHAVDDMPNALSALSD----LHAKL
                GS+ + G +   +D L  ++ H+ D+  A +AL D   ++AH+
F11G11.2       GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEF PQFKAHQE
```

- How to filter out the last one & pick up the second?

Which species are the closest living relatives of modern humans?

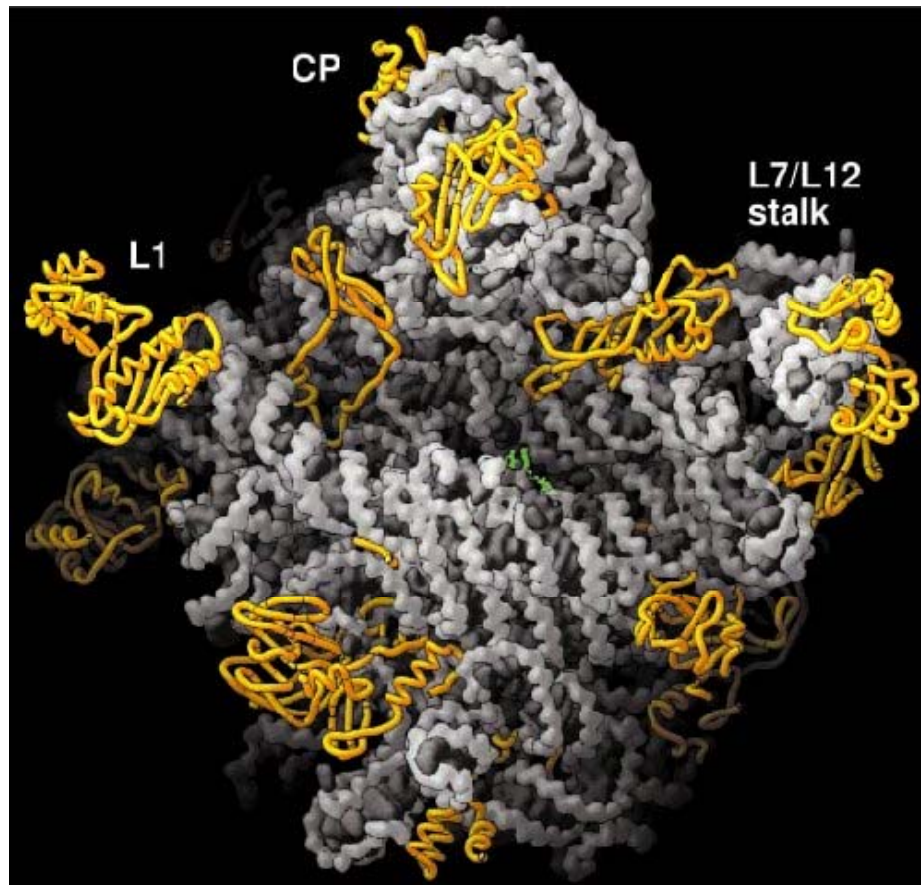


Mitochondrial DNA, most nuclear DNA-encoded genes, and DNA/DNA hybridization all show that bonobos and chimpanzees are related more closely to humans than either are to gorillas.

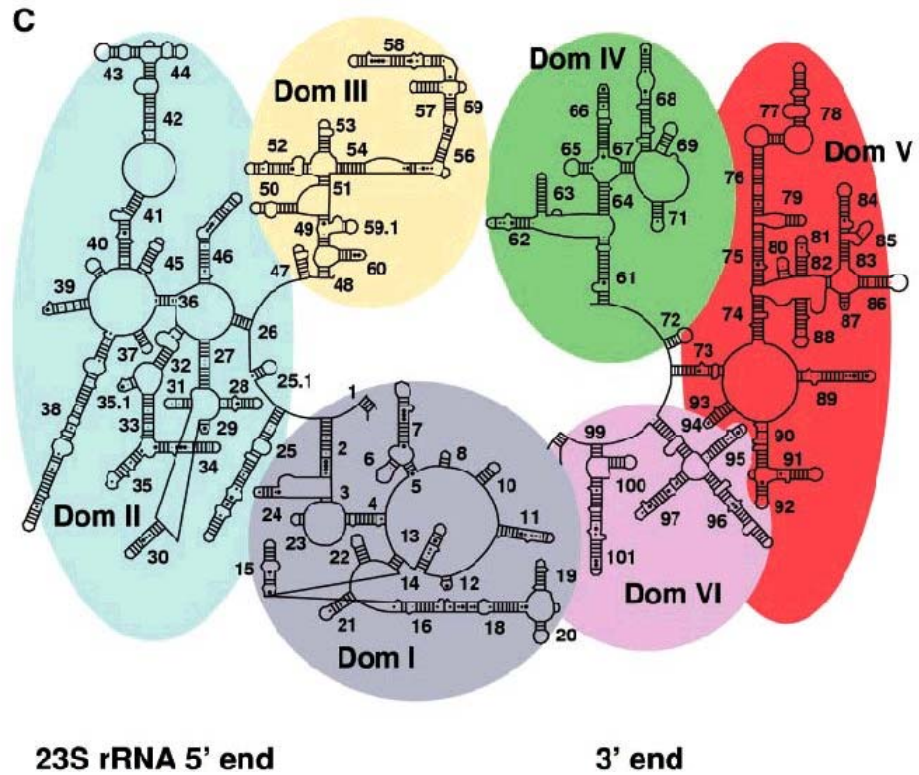


The pre-molecular view was that the great apes (chimpanzees, gorillas and orangutans) formed a clade separate from humans, and that humans diverged from the apes at least 15-30 MYA.

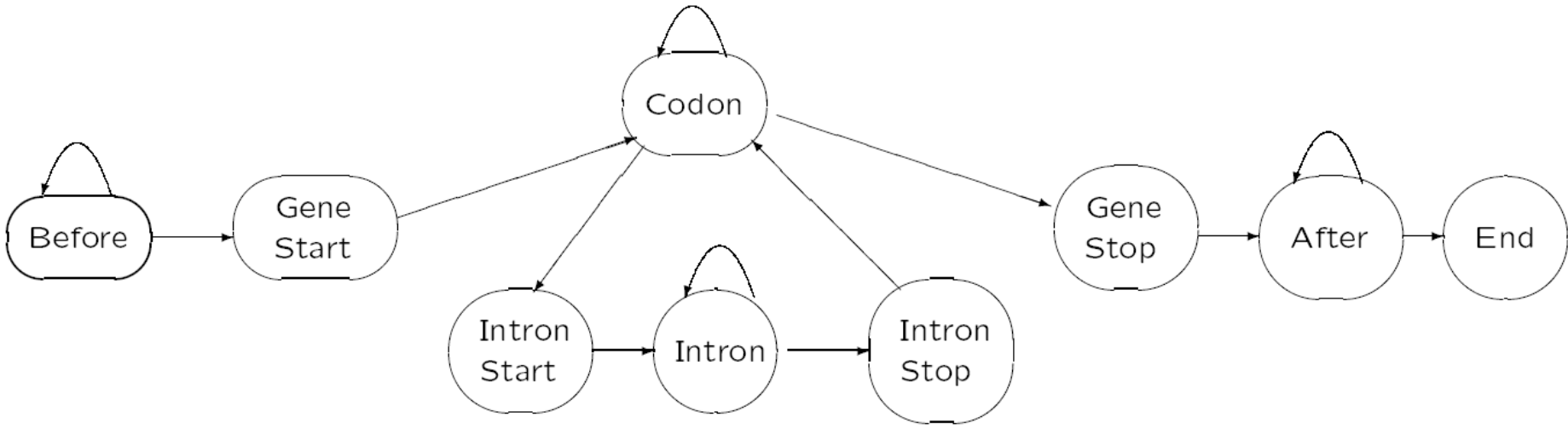
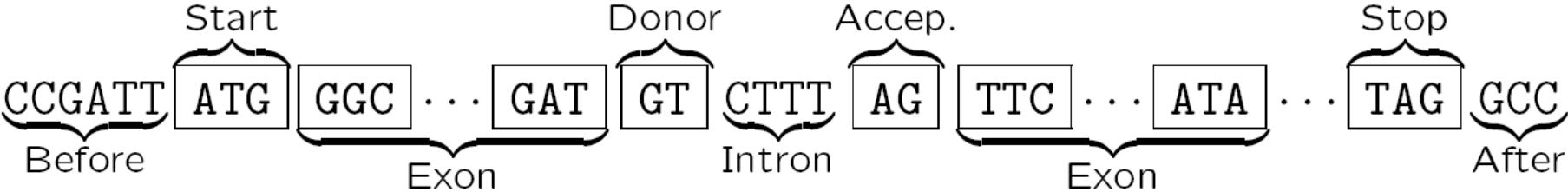
Ribosome structure

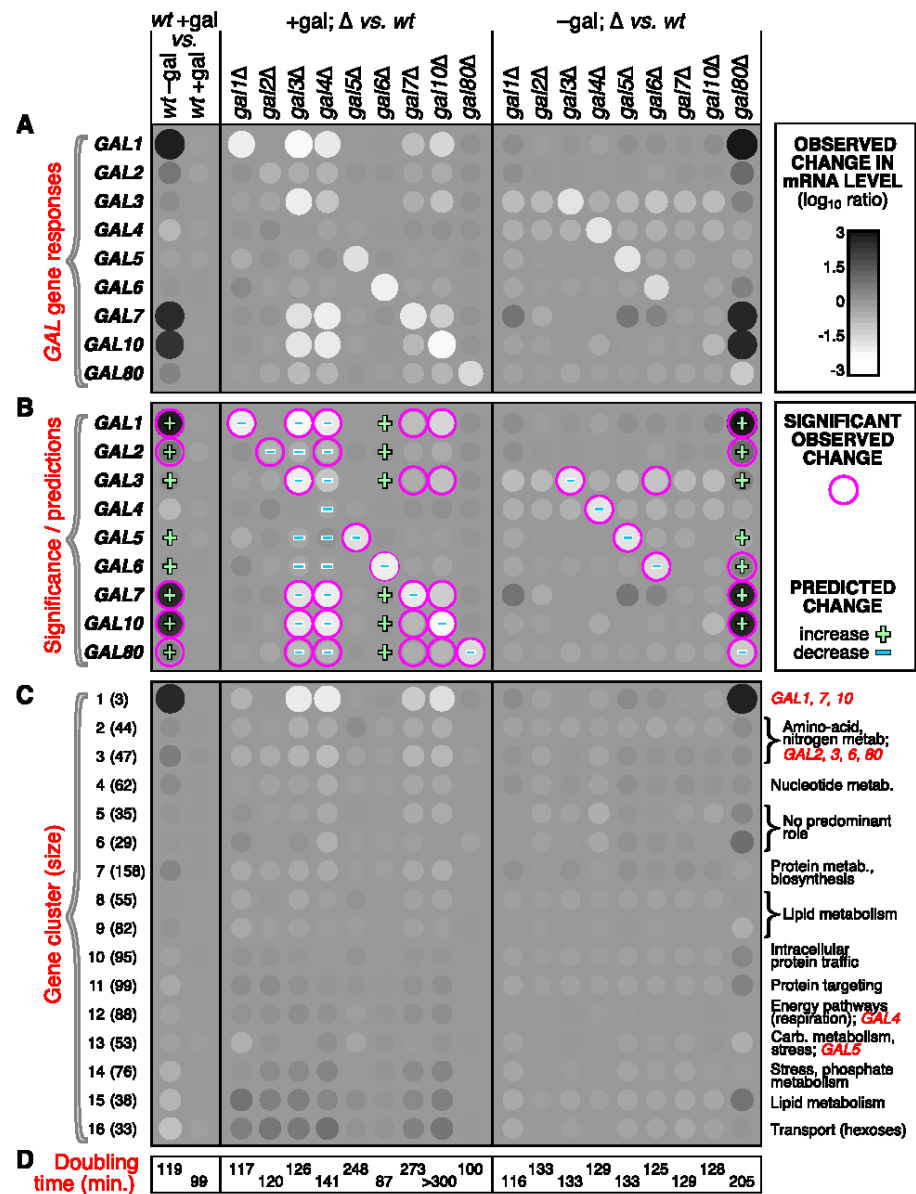


Rimosome rRNA

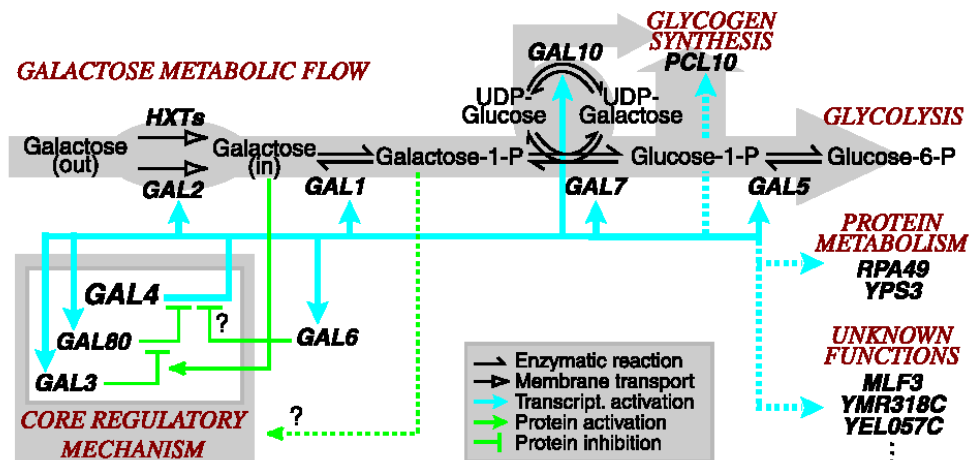


Graphical representation of an eukaryotic HMM gene finder





GALACTOSE METABOLIC FLOW



Conclusions

- Bioinformatic methods are motivated by the explosion of sequence data
- This course gives a broad introduction to a number of analysis tools
- Most of these tools rely on the principle of evolution

Schedule

- Lectures take place Mondays 11-13, Auditorium G1, Department of Mathematical Sciences and on Wednesday 11-12, Auditorium D1.
- Computer / theoretical exercises take place on Wednesdays 13-16 (HOLD 1), and Monday 12-15 (HOLD 2) at G31.