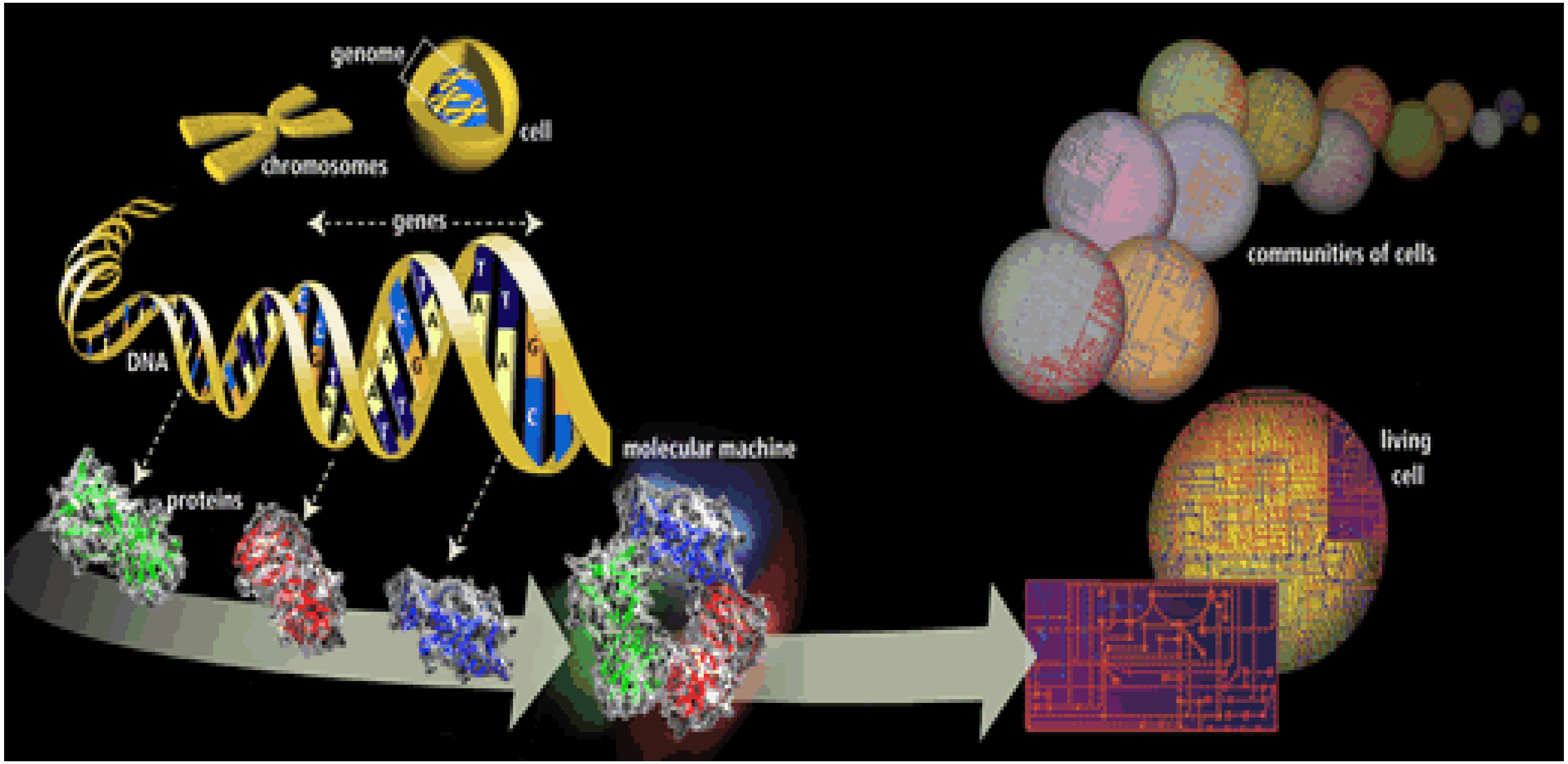
Bioinformatics (LGBIO2010)

Pierre Dupont and Michel Ghislain



Bioinformatics combines computer sciences, statistics, mathematics and engineering to analyse biological data and give interpretation

What kind of biology data are collected ?

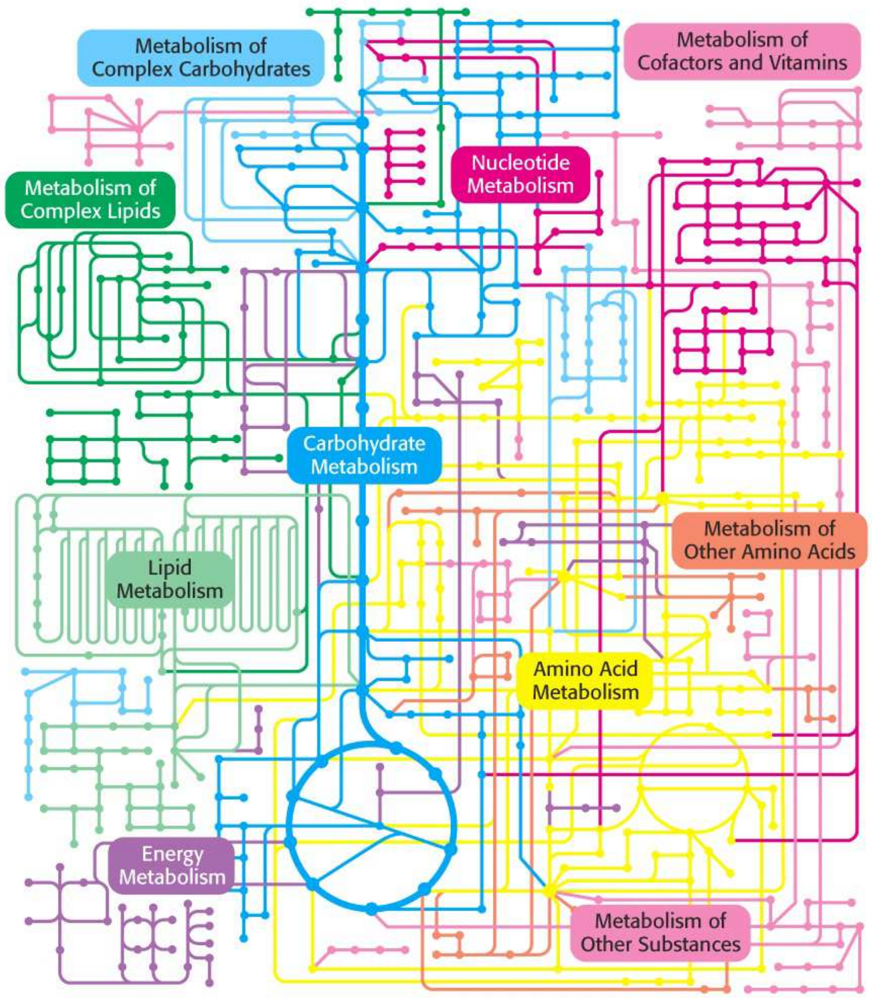
- Biology aims at the study of living organisms, including their structure, function, growth, evolution, distribution, …

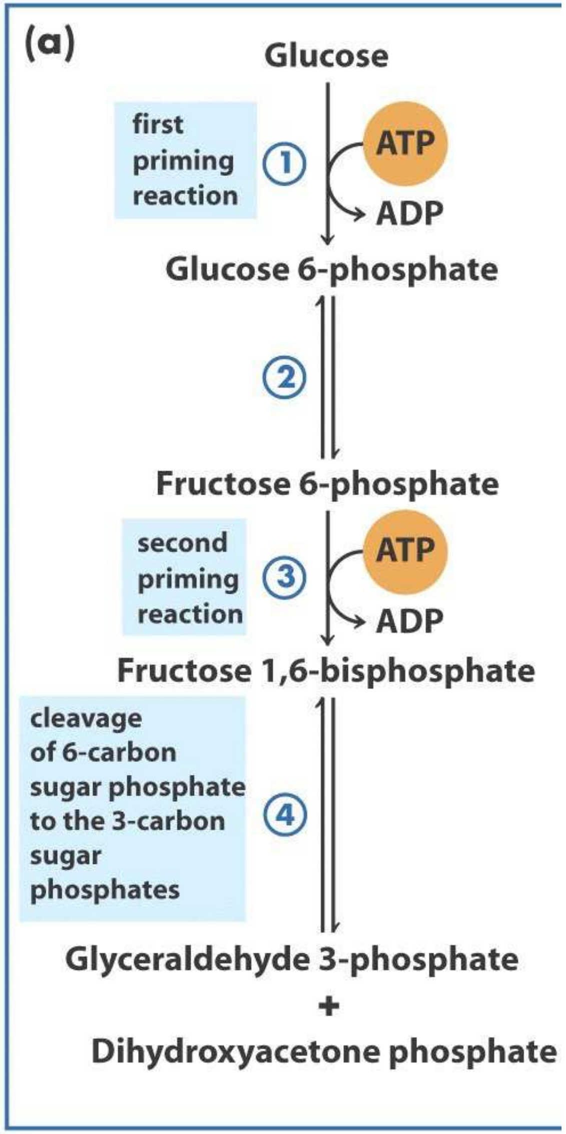
1. A living cell is a self contained, self-assembling, self-adjusting, self perpetuating system

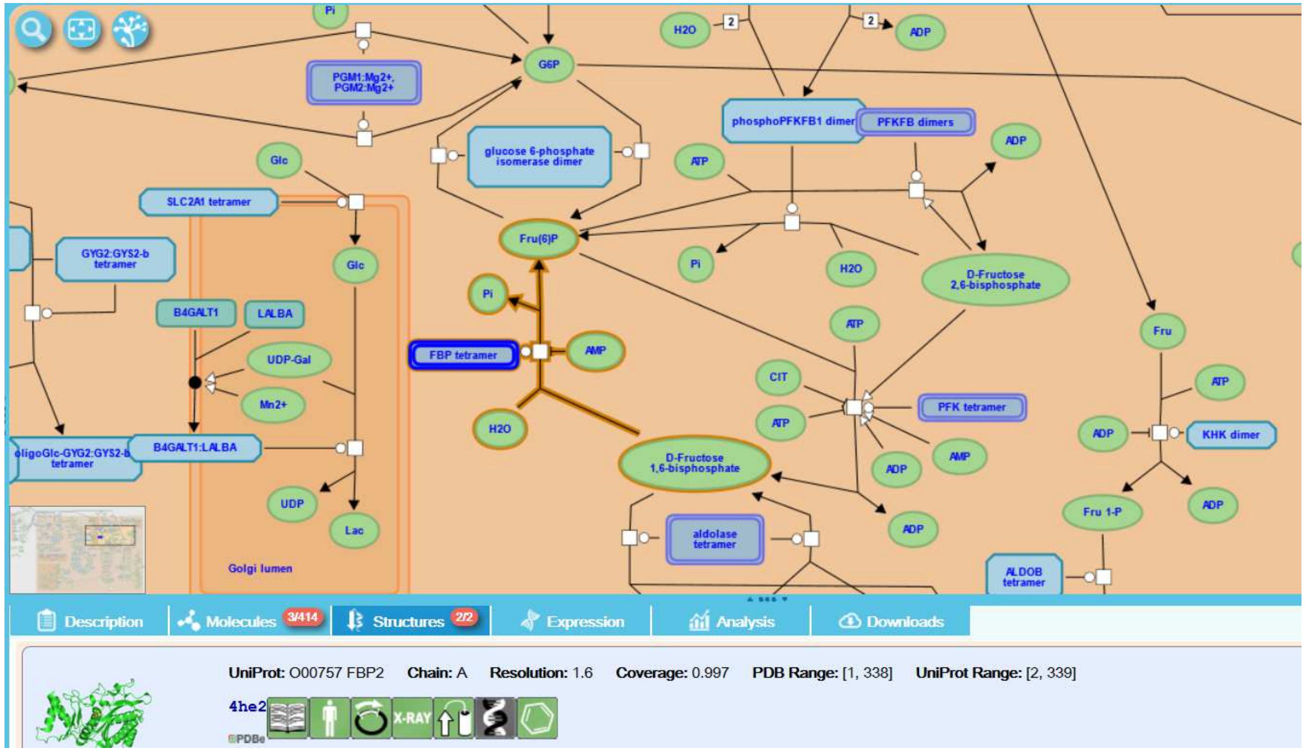
2. The cell extracts free energy and raw material from its environment

3. Control of gene expression allows cell differentiation and adaptation

1. Chemical transformations are organized into a network of reactions (metabolic pathways) with common regulation

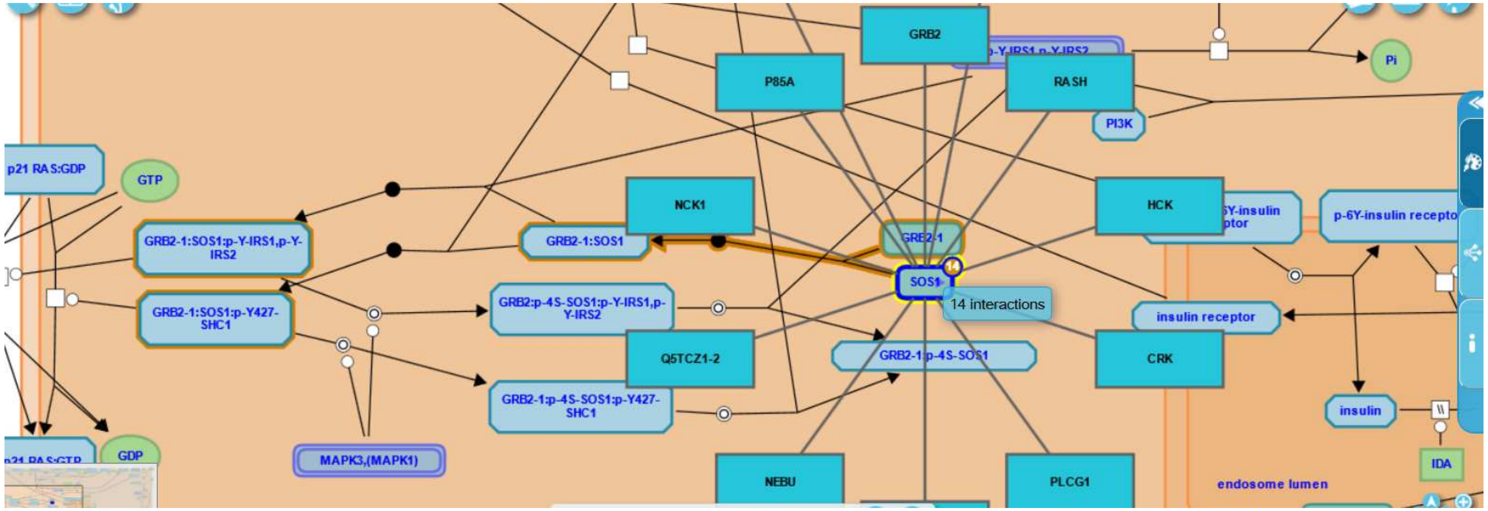

Metabolism of Complex Carbohydrates
Metabolism of Cofactors and Vitamins
Nucleotide Metabolism
Metabolism of Complex Lipids
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Other Substances


(a)
Glucose
first priming reaction
1
ATP
ADP
Glucose 6-phosphate
2
Fructose 6-phosphate
second priming reaction
3
ATP
ADP
Fructose 1,6-bisphosphate
cleavage of 6-carbon sugar phosphate to the 3-carbon sugar phosphates
4
Glyceraldehyde 3-phosphate
+
Dihydroxyacetone phosphate



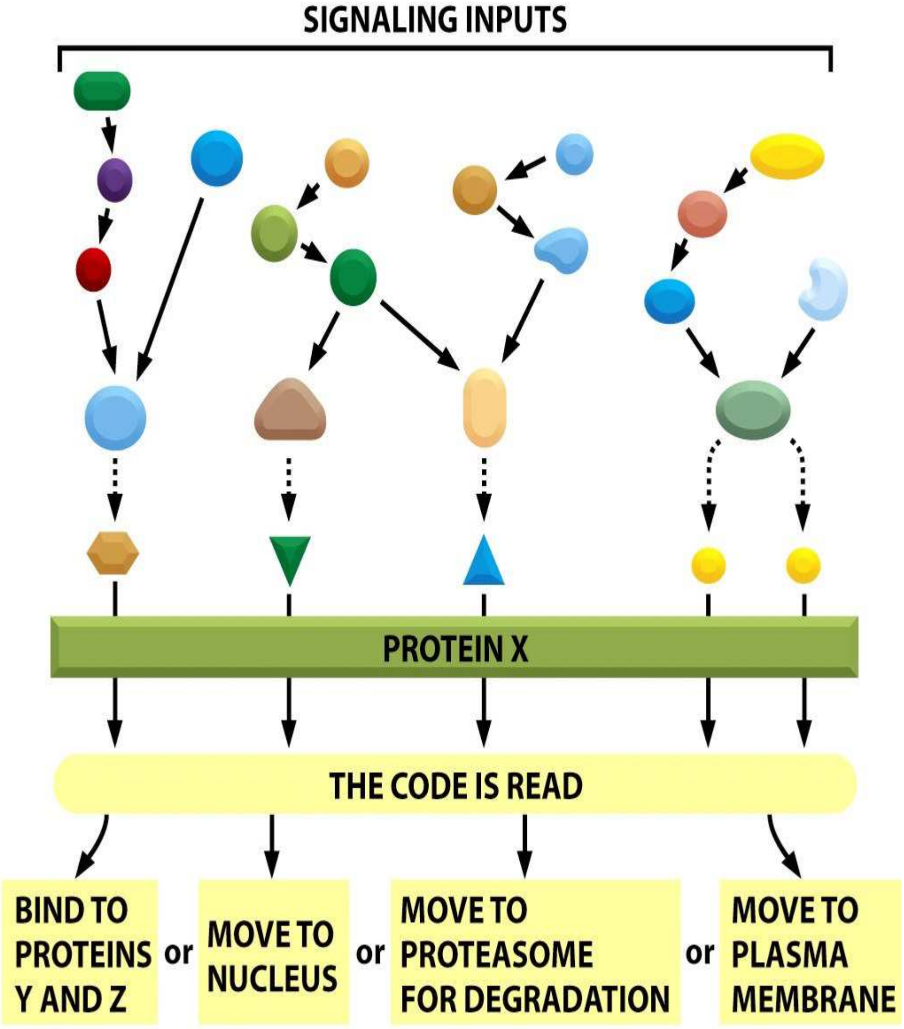
UniProt: O00757 FBP2 Chain: A Resolution: 1.6 Coverage: 0.997 PDB Range: [1, 338] UniProt Range: [2, 339]

A more comprehensive view of metabolism



Targets of the insulin-triggered transduction pathway and protein-protein interactions

2. Living organisms respond to environmental stimuli by activating signal transduction pathways



Proteins are subject to post-translational modifications affecting subcellular localisation, function and stability

3. Regulation of gene expression requires a network of interacting transcriptional factors that bind to target sequences

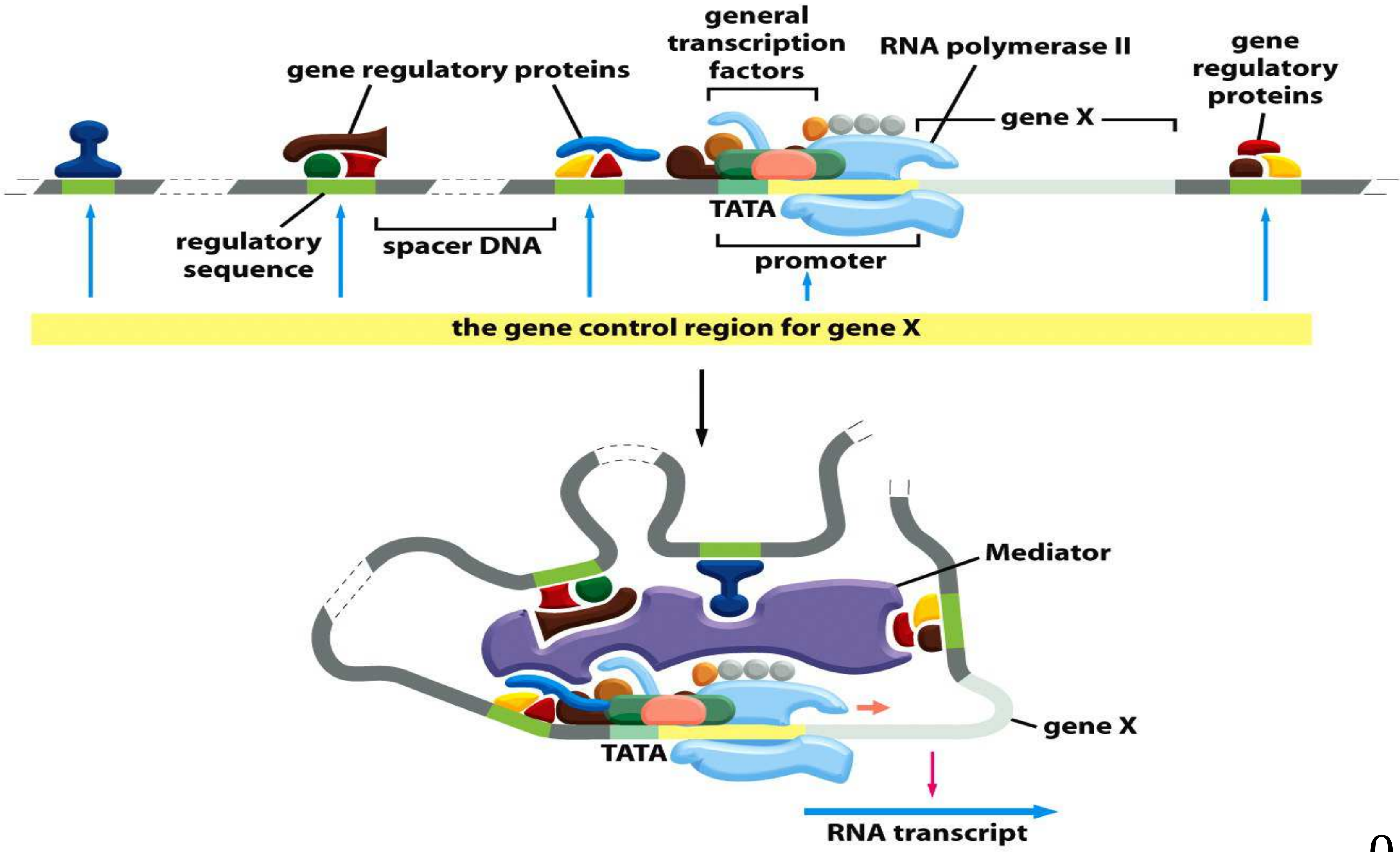


Figure 7-44 Molecular Biology of the Cell 5/e (© Garland Science 2008)

- Biological study implies the choice of:
  - an object (model organism, sequence, …)
  - an experimental approach (genetics, biochemistry, …)
  - a dedicated equipment (centrifuge) and several analysis methods and tools (electrophoresis, FACS,…)
- Molecular biology is the study of the interaction between DNA, RNA and proteins, and how these interactions are regulated

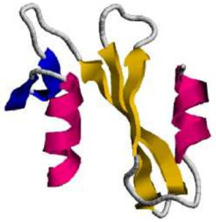
Molecular biology approaches used on yeast

Biochemistry

Activity : ATPase
H+-pumping

Purification:
A 100-kDa band

Sequence
Structure



Mechanism of
proton transport

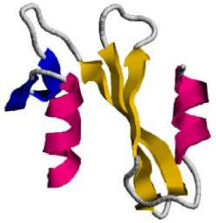
Genetics

Selection of a mutant
phenotype: Dio9R

Gene cloning
A 5 kb DNA fragment

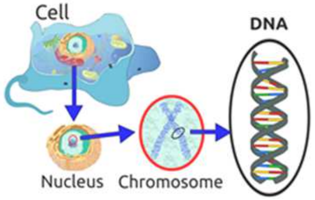
Nucleotide sequence
determination

Structure

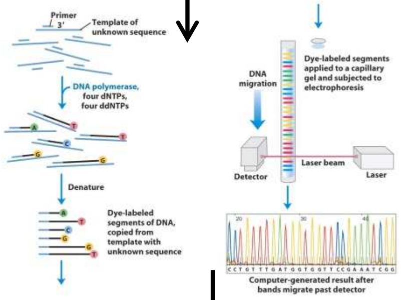


Mechanism :mutagenesis

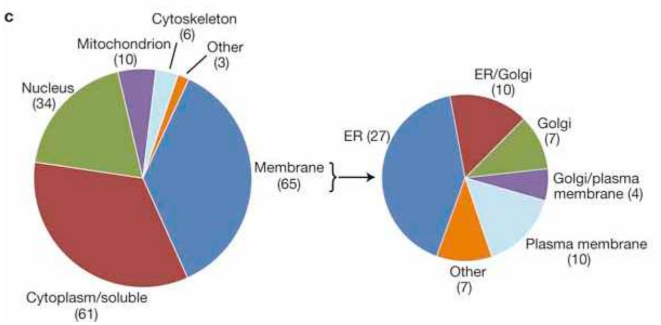
Omics



Genome-scale
investigation

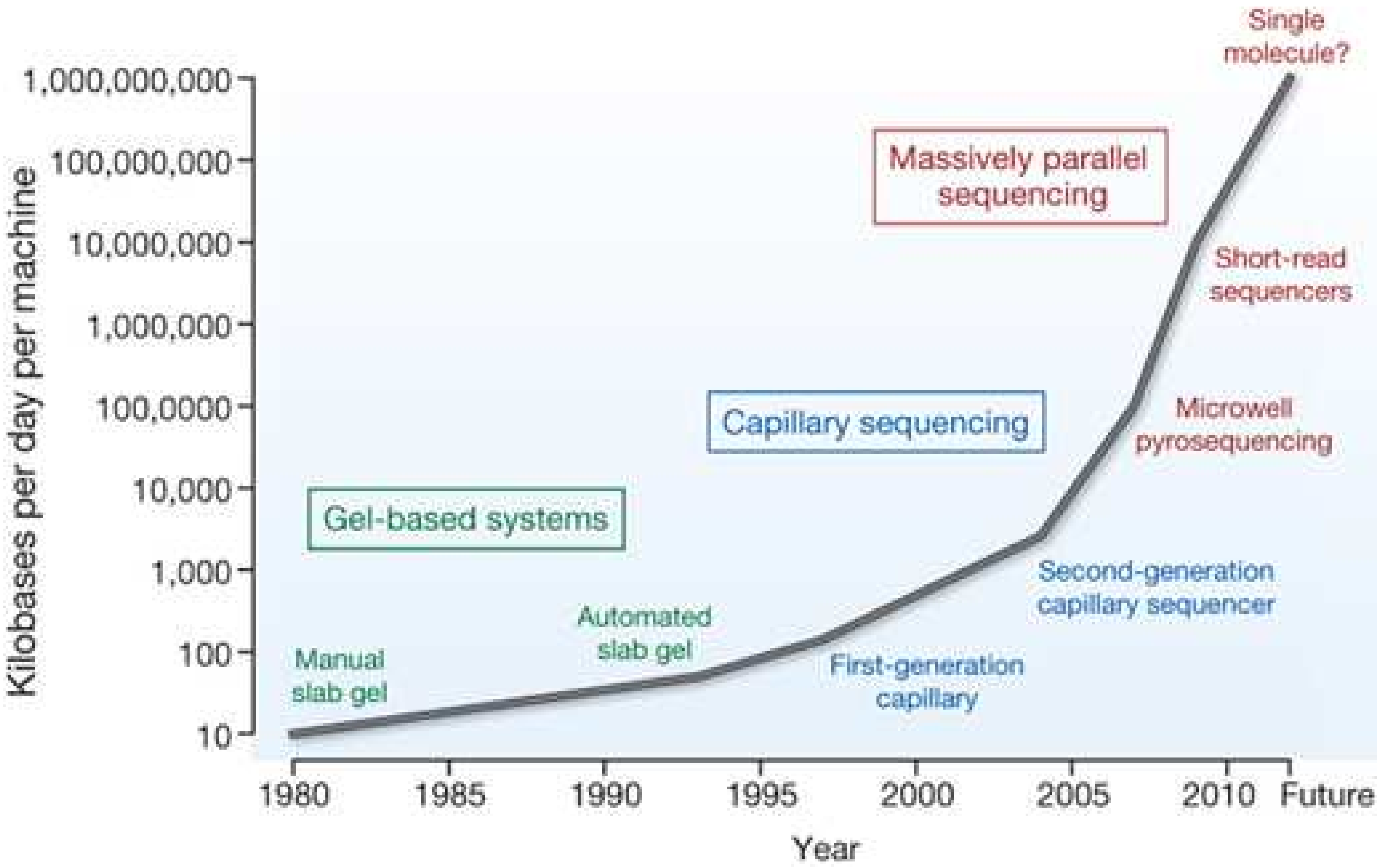


DNA/RNA
sequencing



Computational analysis

1. Improvements in DNA sequencing technologies and ever more powerful computers have created massive amounts of information (whole-genome assembly, variant detection, targeting resequencing)



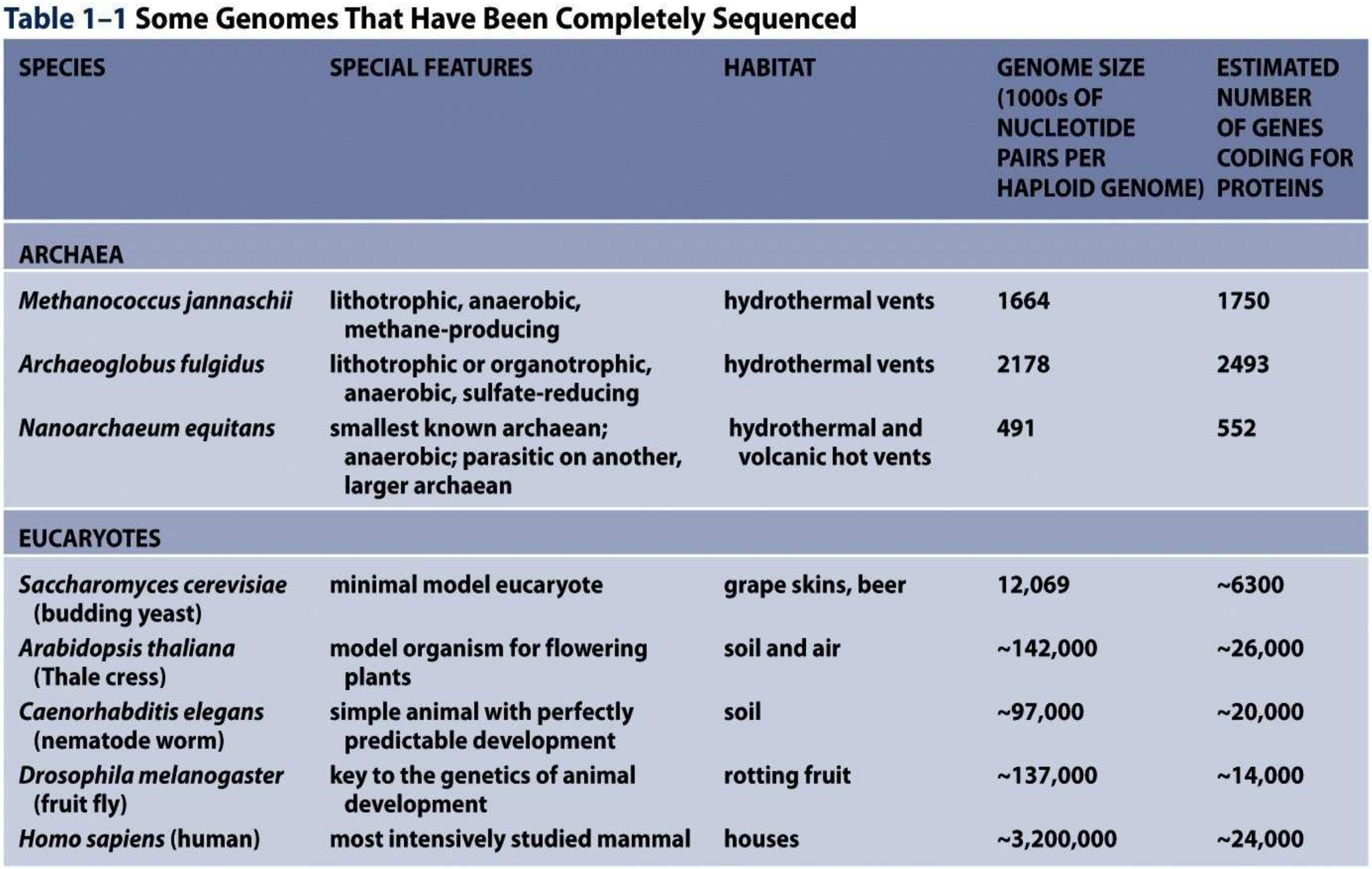


Table 1–1 Some Genomes That Have Been Completely Sequenced

SPECIES	SPECIAL FEATURES	HABITAT	GENOME SIZE (1000s OF NUCLEOTIDE PAIRS PER HAPLOID GENOME)	ESTIMATED NUMBER OF GENES CODING FOR PROTEINS
ARCHAEA				
Methanococcus jannaschii	lithotrophic, anaerobic, methane-producing	hydrothermal vents	1664	1750
Archaeoglobus fulgidus	lithotrophic or organotrophic, anaerobic, sulfate-reducing	hydrothermal vents	2178	2493
Nanoarchaeum equitans	smallest known archaean; anaerobic; parasitic on another, larger archaean	hydrothermal and volcanic hot vents	491	552
EUCARYOTES				
Saccharomyces cerevisiae (budding yeast)	minimal model eucaryote	grape skins, beer	12,069	~6300
Arabidopsis thaliana (Thale cress)	model organism for flowering plants	soil and air	~142,000	~26,000
Caenorhabditis elegans (nematode worm)	simple animal with perfectly predictable development	soil	~97,000	~20,000
Drosophila melanogaster (fruit fly)	key to the genetics of animal development	rotting fruit	~137,000	~14,000
Homo sapiens (human)	most intensively studied mammal	houses	~3,200,000	~24,000

1598 bacterial/85 archae/294 eukaryotic genomes (2010)

2. High througput analysis of gene expression

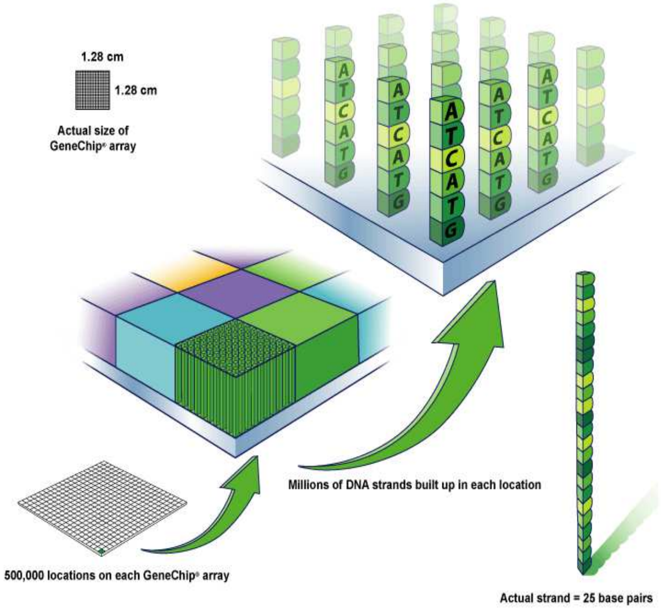
➢ Transcriptomics: microarrays

- An array works by exploiting the ability of a given mRNA molecule to hybridize to the DNA template.

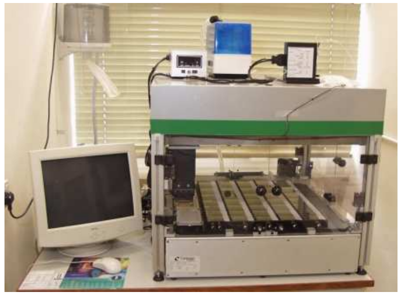
- Using an array containing many DNA samples in an experiment, the expression levels of hundreds or thousands genes within a cell by measuring the amount of mRNA bound to each site on the array.

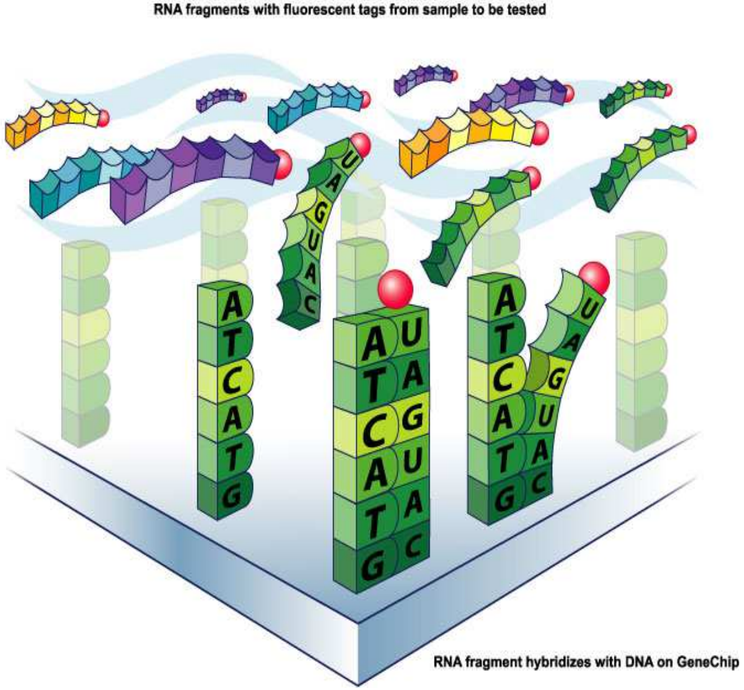
- With the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell.

DNA Arrays--Technical Foundations


1.28 cm
1.28 cm
Actual size of GeneChip® array
A T C A T G
Millions of DNA strands built up in each location
500,000 locations on each GeneChip® array
Actual strand = 25 base pairs

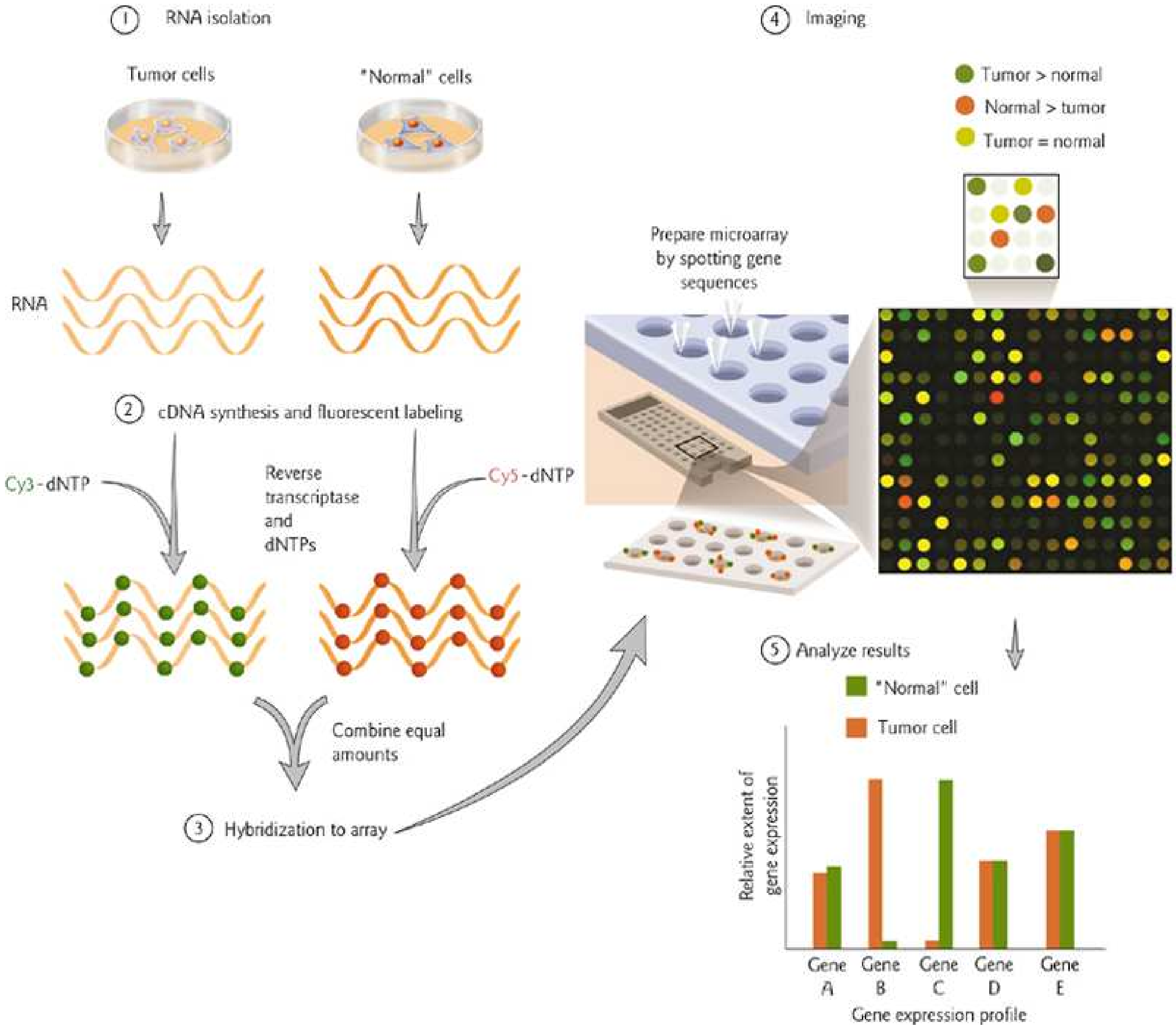
Millions of DNA strands build up on each location.




RNA fragments with fluorescent tags from sample to be tested
A T C A T G
U A G U A C
RNA fragment hybridizes with DNA on GeneChip

Tagged probes become hybridized to the DNA chip’s microarray.

Application: Wide-scale analysis of tumerogenesis

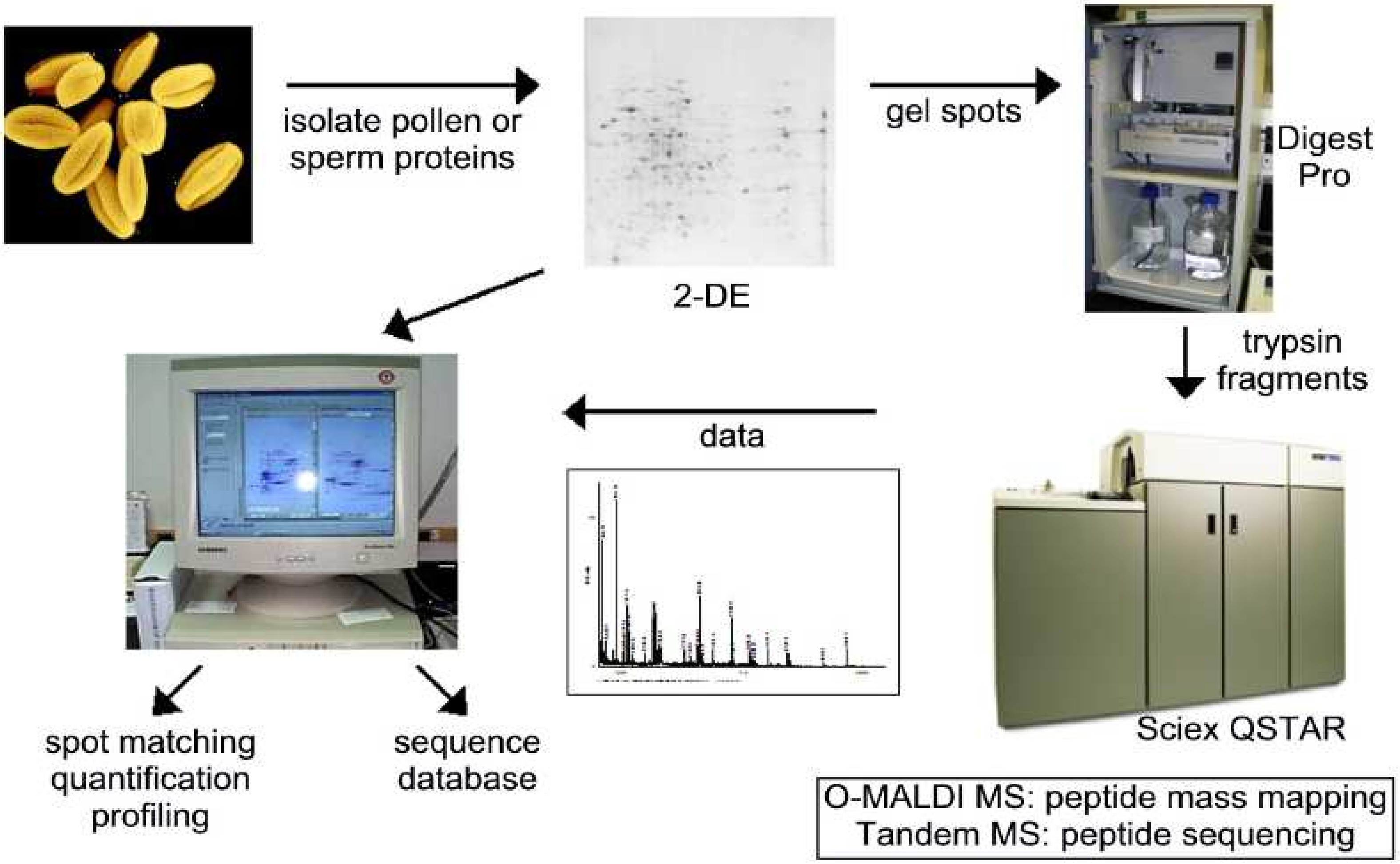


A gene chip made by Affymetrix. The well can contain probes for thousands of genes.

Imaging of a chip. The amount of fluorescence corresponds to the amount of a gene expressed.

Proteomics

Large-scale study of proteins, produced or modified by an organism according to environmental changes or tissue-specificity



isolate pollen or sperm proteins

2-DE

gel spots

Digest Pro

trypsin fragments

Sciex QSTAR

data

spot matching quantification profiling

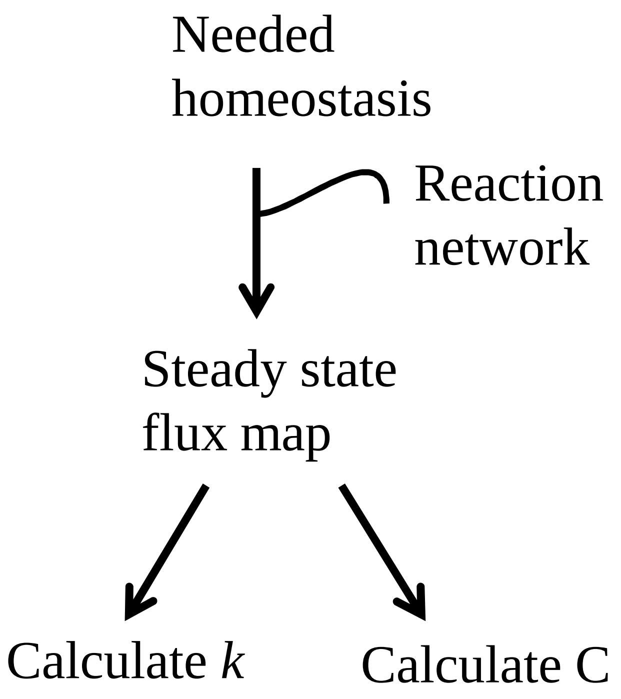
sequence database

O-MALDI MS: peptide mass mapping
Tandem MS: peptide sequencing

21st-Century biology : Systems integrative analysis

- Progress in DNA sequencing technologies and high-throughput experimental technologies have created massive amounts of information
- What do we do with genomics, transcriptomics, and proteomics ?
  - just developping list of cellular components and properties?
- Try a more integrative approach considering living organisms as systems
  - Approach combining bioinformatics, mathematical models, and computer simulation

Systems biology



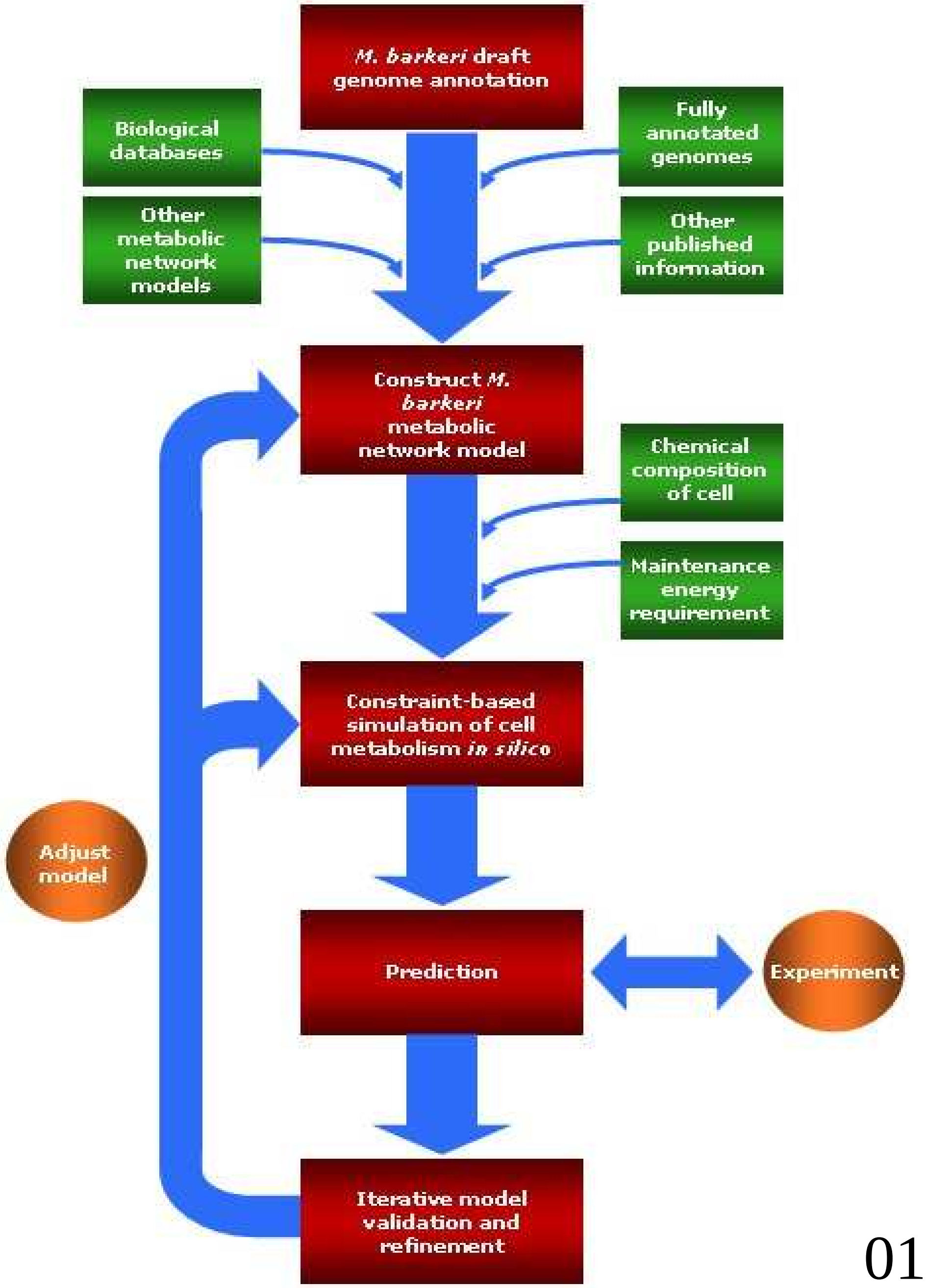
Needed homeostasis

Reaction network

Steady state flux map

Calculate *k*

Calculate C


M. barkeri draft genome annotation
Biological databases
Fully annotated genomes
Other metabolic network models
Other published information
Construct *M. barkeri* metabolic network model
Chemical composition of cell
Maintenance energy requirement
Constraint-based simulation of cell metabolism *in silico*
Adjust model
Prediction
Experiment
Iterative model validation and refinement

M. Barkeri metabolism modeling

Why bioinformatics ?

Bioinformatics is the use of computational methods to study biological data

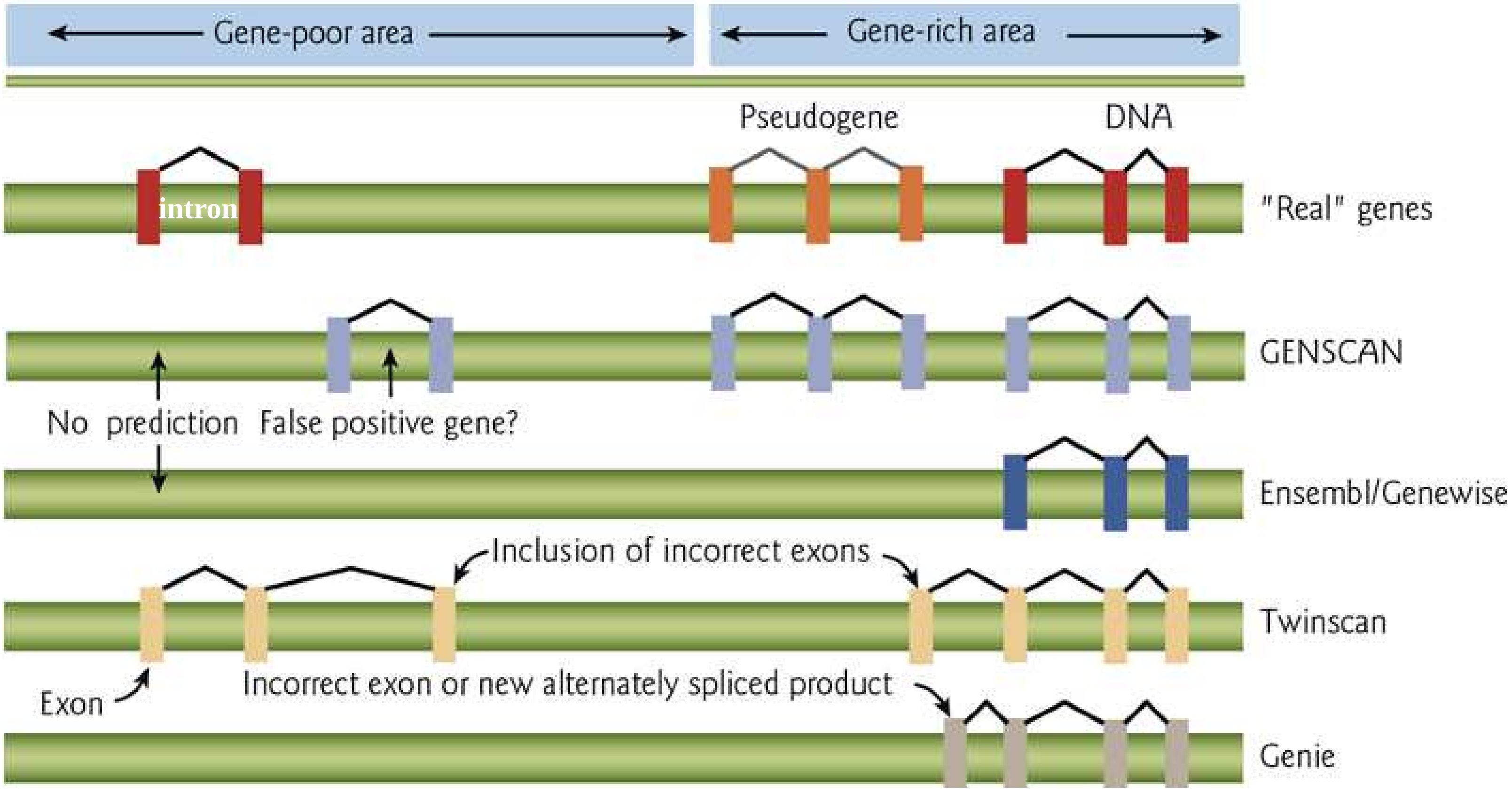
- Development of methods for the management and analysis of biological information arising from genomics and high-throughput experiments
- Development of computational methods for studying the structure, function, and evolution of genes, proteins, and whole genomes
- Data integration = Systems biology
  - Flux Balance Analysis
  - Organelle processes, expression modeling

1. Sequence analysis programs

Algorithms have been developed to recognize pattern matches and feature signatures from sequences:

- Identification of potential genes in new genome data
  - RNA splice sites
  - ORFs (signals, codon composition, …)
- Amino acid propensities in a protein
  - α-helix, TMS
- Conserved regions (motif, domain) of proteins and cis regulatory DNA elements in the promoter region
- Identification of evolutionary relationships = phylogenetic tree

Performance of various gene prediction programs



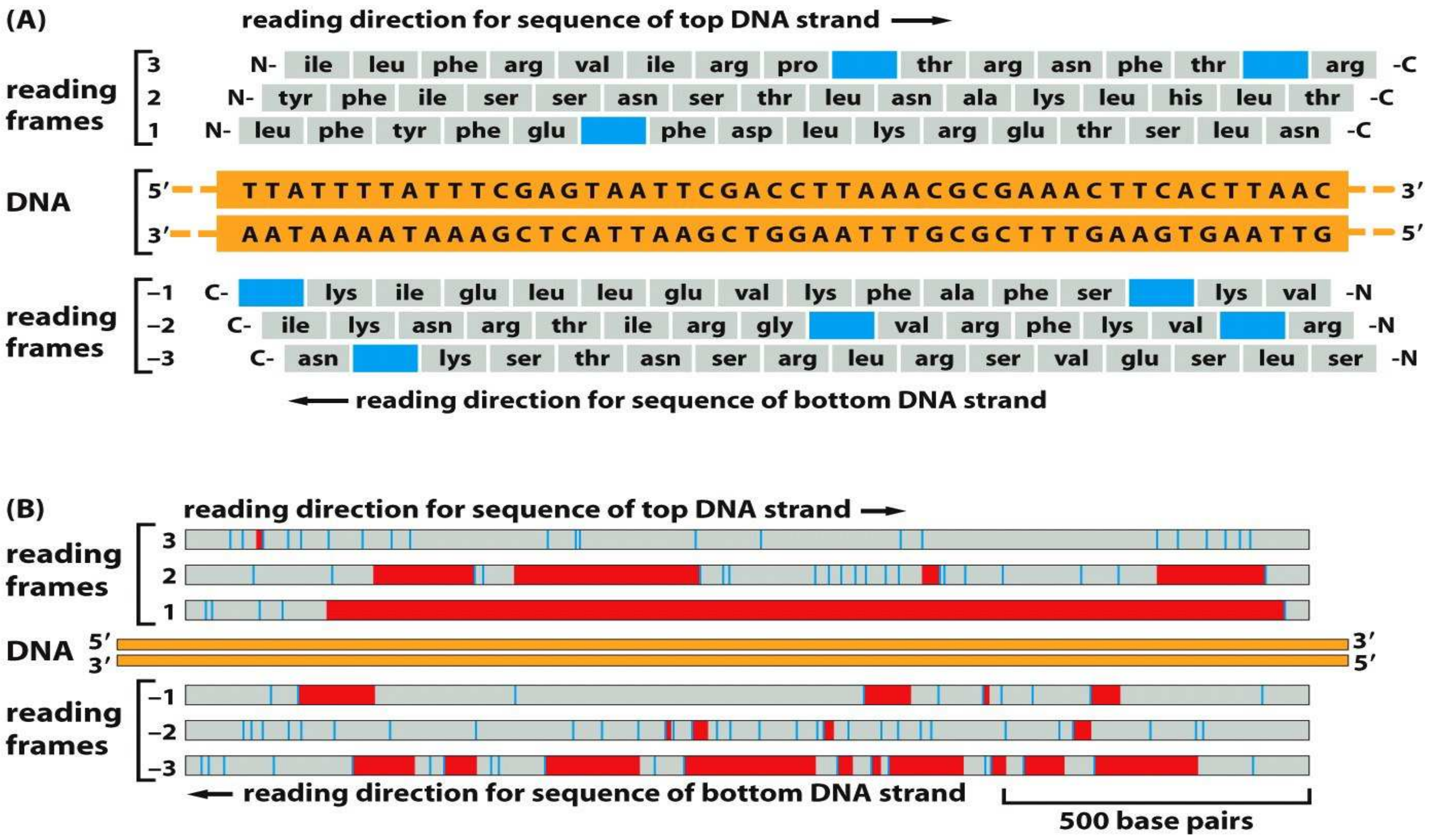
The comparison of different ORF prediction programs leads to the conclusion that ever more sophisticated methods are required

Analysis of the porcine circovirus genome for potential ORFs illustrates some of the problems and challenges faced by genome annotation and the underlying computer programs

1. Search genome sequence database for porcine circovirus 2
   - ➢ Refseq:nc_0051481 (accession number)
2. Use the *getorf* program to identify the encoded polypeptides
   - ➢ We need an algorithm for finding ORFs :
     - ✓ Given a DNA sequence **s**, and a positive integer *k*, for each reading frames decompose the sequence into triplets, and find all stretches of triplets starting with a start-codon and ending with a stop codon
     - ✓ Repeat also for the reverse complement of the sequence, **s'**
     - ✓ Output all ORFs longer than the prefixed threshold *k*
     - ✓ Once an ORF has been found, its translation is easy using the genetic code

ORF detection by using simple rules for prokaryote and lower eukaryote's genomes :

- Select the longest ORF delimited by stop codons
- Select the first AUG start codon downstream of the 5' stop codon


(A)
reading direction for sequence of top DNA strand
reading frames
3 N- ile leu phe arg val ile arg pro thr arg asn phe thr arg -C
2 N- tyr phe ile ser ser asn ser thr leu asn ala lys leu his leu thr -C
1 N- leu phe tyr phe glu phe asp leu lys arg glu thr ser leu asn -C
DNA
5' TTATTTTATTTCGAGTAATTCGACCTTAAACGCGAAACTTCACTTAAC 3'
3' AATAAAATAAAGCTCATTAAGCTGGAATTTGCGCTTTGAAGTGAATTG 5'
reading frames
−1 C- lys ile glu leu leu glu val lys phe ala phe ser lys val -N
−2 C- ile lys asn arg thr ile arg gly val arg phe lys val arg -N
−3 C- asn lys ser thr asn ser arg leu arg ser val glu ser leu ser -N
reading direction for sequence of bottom DNA strand
(B)
reading direction for sequence of top DNA strand
reading frames
3
2
1
DNA
5'
3'
3'
5'
reading frames
−1
−2
−3
reading direction for sequence of bottom DNA strand
500 base pairs

3. Gene annotation

   - ➢ Search for sequence similarity and the presence of conserved motives in protein or profile databases (experimental evidence)

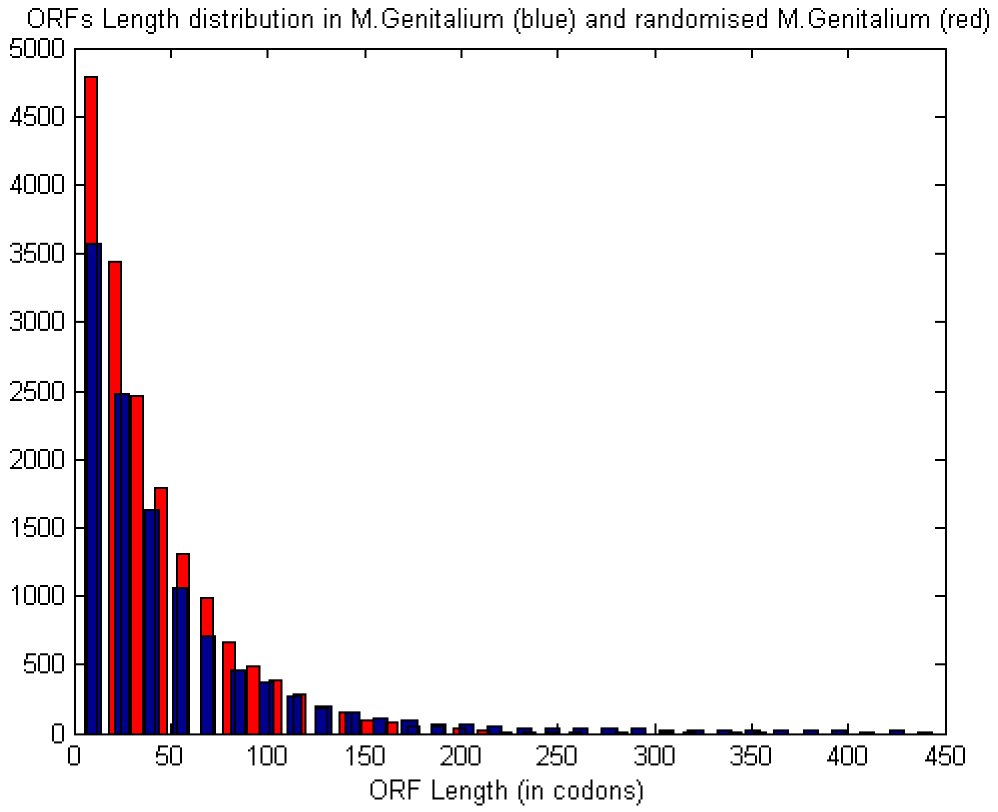
4. Statistics for evaluating the confidence that any short ORF with unknown function is real

   - ➢ Calculate the probability of seeing an ORF of length *L* in a random sequence
     - ✓ Use a sequence probability null model (hypothesis testing)
     - ✓ An ORF of a certain length is significant when it is highly unlikely under the null model
     - ✓ Make a choice as to how unlikely a given ORF has to be for us to accept it. The more stringent our conditions, the fewer candidates we will have

Computing a *p*-value for ORFs (N. Cristianini and MWHahn)

What is the probability of an ORF of *k* or more codons arising by chance?

What is the threshold value of *k* such that 95% of random ORFs are shorter than *k*?



- The *Mycoplasma genitalium* genomes contains 11 922 ORFs. After permutating the genome sequence we found 17 367 ORFs
- This list of ORFs lengths in the random sequence defines a null distribution
- If k >360 codons (the maximum ORF size in the randomized sequence), we find 326 ORFs longer than that.
  If k > 114 codons, we find 1520 ORFs
- The exact number is 470 genes

2. Micro array experiments allow us :

- to compare differences in expression for two different states
- To identify regulatory circuits and transcriptional factors networks

Informatics requirements:

- Algorithms for clustering groups of gene expression help point out possible regulatory networks
- Other algorithms perform statistical analysis to improve signal to noise contrast

Objectives of the course

- Understand how molecular biology problems are solved by computational methods
- Understand the structure of molecular biology data bases and the associated management tools
- Be able to explain how algorithms work
- Be able to select the most appropriate algorithm according to the question raised
- Be aware of the statistics used

Evaluation of acquired competences

- Theory aspects are assessed by examination (50% of the final mark)
- Proficiency in addressing a biological problem (sequence analysis) is assessed during the exam (30%)
- Semester exercise assignments (20%)

Contents of the course

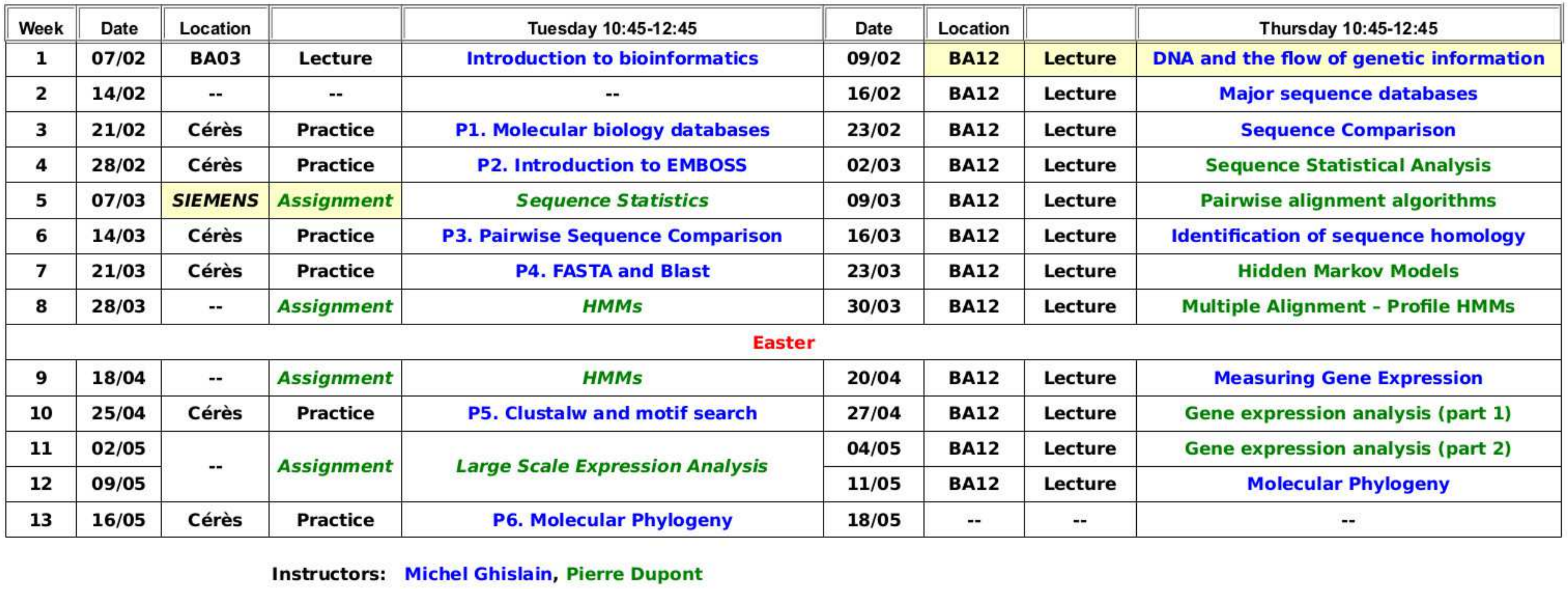
- Introduction
- Molecular biology fundamentals (optional)
- Major sequence and structure databases
- Sequence comparison
- Pairwise alignment algorithms
- Search of sequence homology

Contents of the course (contd)

- Sequence statistical analysis
- Hidden markov models (HMMs)
- Multiple alignment - Profile HMMs
- Gene expression measuring and analysis
- Molecular phylogeny

Teaching-training activities

- Lecture timetable

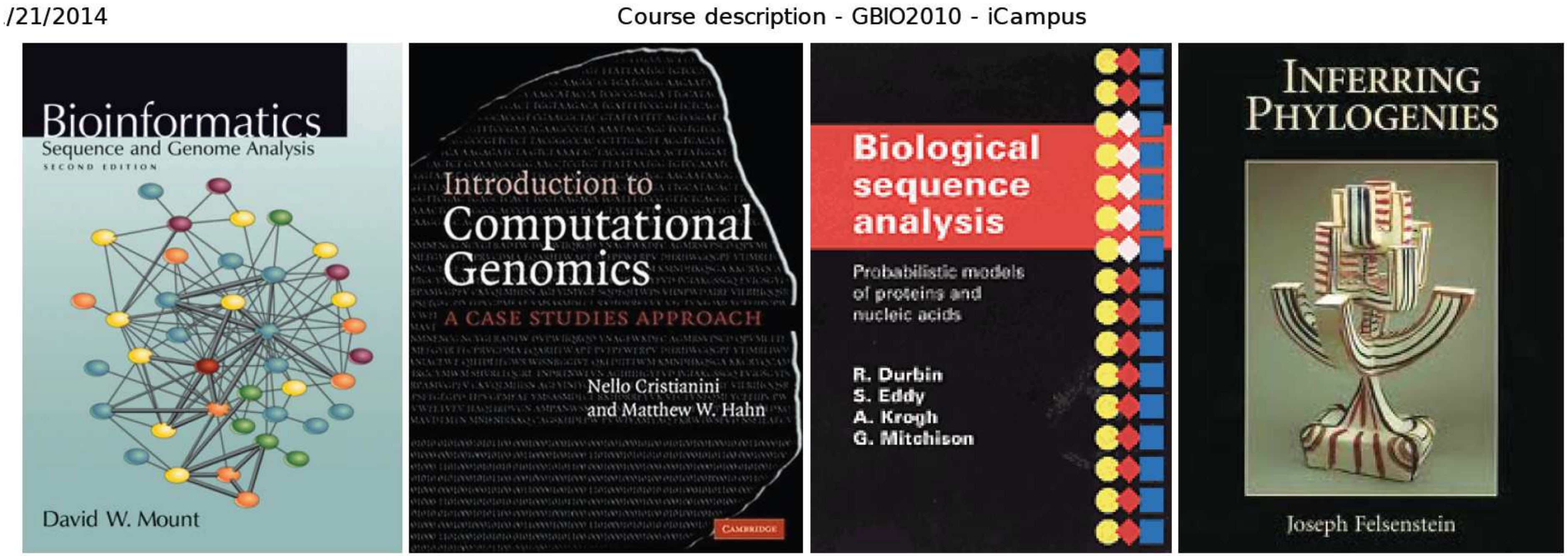


Week	Date	Location		Tuesday 10:45-12:45	Date	Location		Thursday 10:45-12:45
1	07/02	BA03	Lecture	Introduction to bioinformatics	09/02	BA12	Lecture	DNA and the flow of genetic information
2	14/02	--	--	--	16/02	BA12	Lecture	Major sequence databases
3	21/02	Cérès	Practice	P1. Molecular biology databases	23/02	BA12	Lecture	Sequence Comparison
4	28/02	Cérès	Practice	P2. Introduction to EMBOSS	02/03	BA12	Lecture	Sequence Statistical Analysis
5	07/03	*SIEMENS*	*Assignment*	*Sequence Statistics*	09/03	BA12	Lecture	Pairwise alignment algorithms
6	14/03	Cérès	Practice	P3. Pairwise Sequence Comparison	16/03	BA12	Lecture	Identification of sequence homology
7	21/03	Cérès	Practice	P4. FASTA and Blast	23/03	BA12	Lecture	Hidden Markov Models
8	28/03	--	*Assignment*	*HMMs*	30/03	BA12	Lecture	Multiple Alignment - Profile HMMs
Easter								
9	18/04	--	*Assignment*	*HMMs*	20/04	BA12	Lecture	Measuring Gene Expression
10	25/04	Cérès	Practice	P5. Clustalw and motif search	27/04	BA12	Lecture	Gene expression analysis (part 1)
11	02/05	--	*Assignment*	*Large Scale Expression Analysis*	04/05	BA12	Lecture	Gene expression analysis (part 2)
12	09/05				11/05	BA12	Lecture	Molecular Phylogeny
13	16/05	Cérès	Practice	P6. Molecular Phylogeny	18/05	--	--	--

Instructors: Michel Ghislain, Pierre Dupont

- Six training sessions with the EMBOSS software suite at the Cérès (AGRO) computational room (M. Ghislain)
- Three assignments for a pair of students on algorithmic and statistical problems (P. Dupont and V. Branders)

Suggested books:



/21/2014 Course description - GBIO2010 - iCampus