Machine Learning in Bioinformatics

Gunnar Rätsch

Friedrich Miescher Laboratory, Tübingen

August 20, 2007

Machine Learning Summer School 2007, Tübingen, Germany

Help with slides: Alexander Zien, Cheng Soon Ong and Jean-Philippe Vert

Overview

1. Introduction to Bioinformatics
   - Basic Biology and Central Dogma
   - Typical Data Types
   - Common Analysis Tasks
2. Sequence Analysis (with SVMs)
   - String Kernels
   - Large Scale Data Structures
   - Heterogeneous Data
3. Structured Output Learning
   - Hidden Markov Models & Dynamic Programming
   - Discriminative Approaches (CRFs & HMSVMs)
   - Large Scale Approaches
4. Some Applications
   - Spliced Alignments (PALMA)
   - Gene Finding (mGene & CRAIG)
   - Analysis of Resequencing Arrays (SNPs and Polymorphic regions)

Overview



1. Introduction to Bioinformatics
   - Basic Biology and Central Dogma
   - Typical Data Types
   - Common Analysis Tasks
2. Sequence Analysis (with SVMs)
   - String Kernels
   - Large Scale Data Structures
   - Heterogeneous Data
3. Structured Output Learning
   - Hidden Markov Models & Dynamic Programming
   - Discriminative Approaches (CRFs & HMSVMs)
   - Large Scale Approaches
4. Some Applications
   - Spliced Alignments (PALMA)
   - Gene Finding (mGene & CRAIG)
   - Analysis of Resequencing Arrays (SNPs and Polymorphic regions)

Part I: Introduction to Bioinformatics



Part I: Introduction to Bioinformatics

- Basic Biology and the Central Dogma
- Typical Data Types
- Common Analysis Tasks

NB: Large parts are from tutorials at MLSS 2004 by Alexander Zien and at MLSS 2006 by Jean-Philippe Vert.

What is Bioinformatics?

One Opinion (Christopher Lee, UCLA):

Bioinformatics is the study of the inherent structure of biological information and biological systems. It brings together the avalanche of systematic biological data (e.g. genomes) with the analytic theory and practical tools of computer science and mathematics.

Bioinformatics is the application of computer technology to the management and analysis of molecular biological data.

- bioinformatics = molecular bioinformatics
- approx. synonymous: computational (molecular) biology
- borderline: medical (clinical, phenotypic) data
- NOT: biologically inspired computer science (e.g., neural networks, genetic algorithms)

More definitions can be found here: `http://tinyurl.com/2xgsy7`

What is Bioinformatics?

One Opinion (Christopher Lee, UCLA):

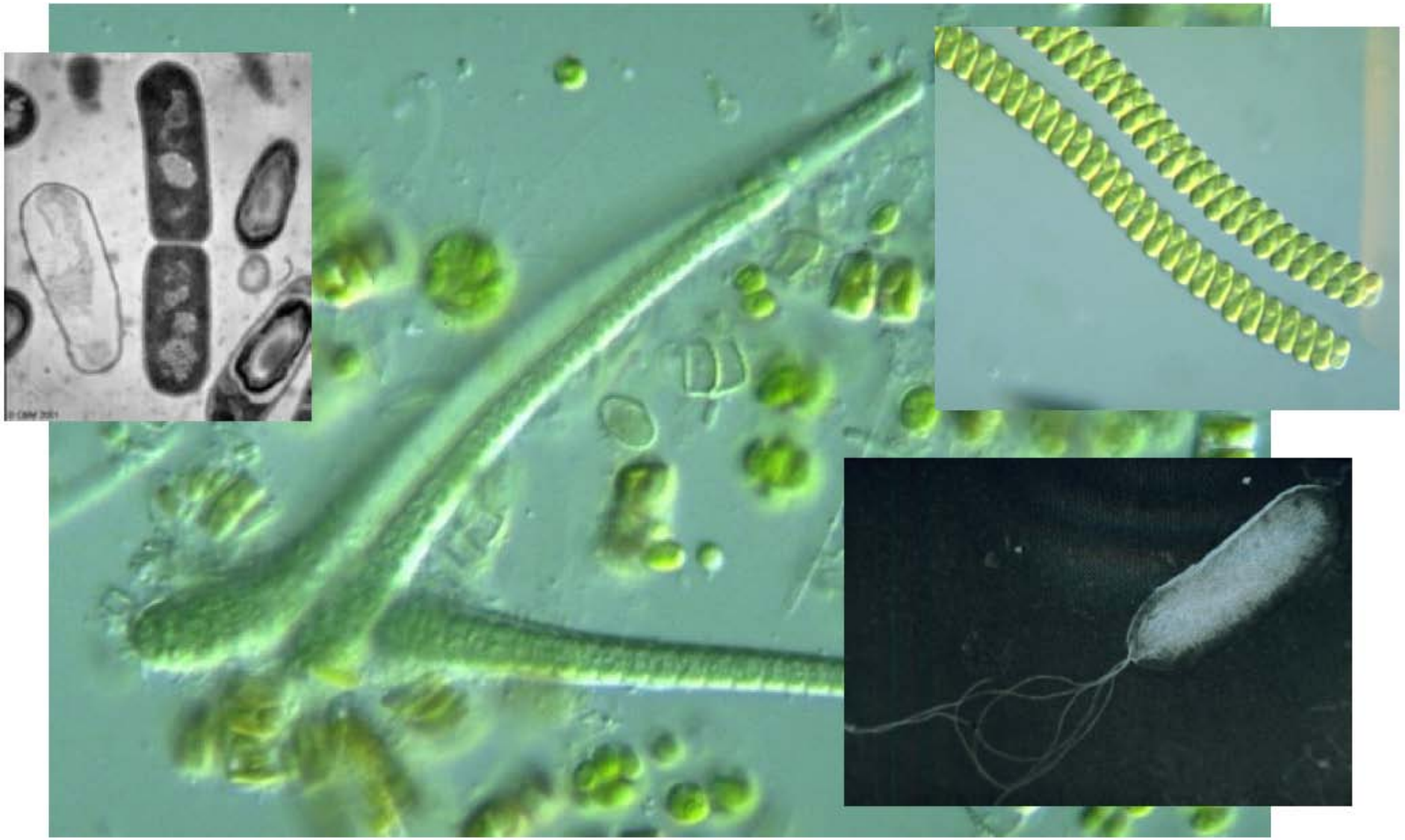
Bioinformatics is the study of the inherent structure of biological information and biological systems. It brings together the avalanche of systematic biological data (e.g. genomes) with the analytic theory and practical tools of computer science and mathematics.

Bioinformatics is the application of computer technology to the management and analysis of molecular biological data.

- bioinformatics = molecular bioinformatics
- approx. synonymous: computational (molecular) biology
- borderline: medical (clinical, phenotypic) data
- NOT: biologically inspired computer science (e.g., neural networks, genetic algorithms)

More definitions can be found here: `http://tinyurl.com/2xgsy7`

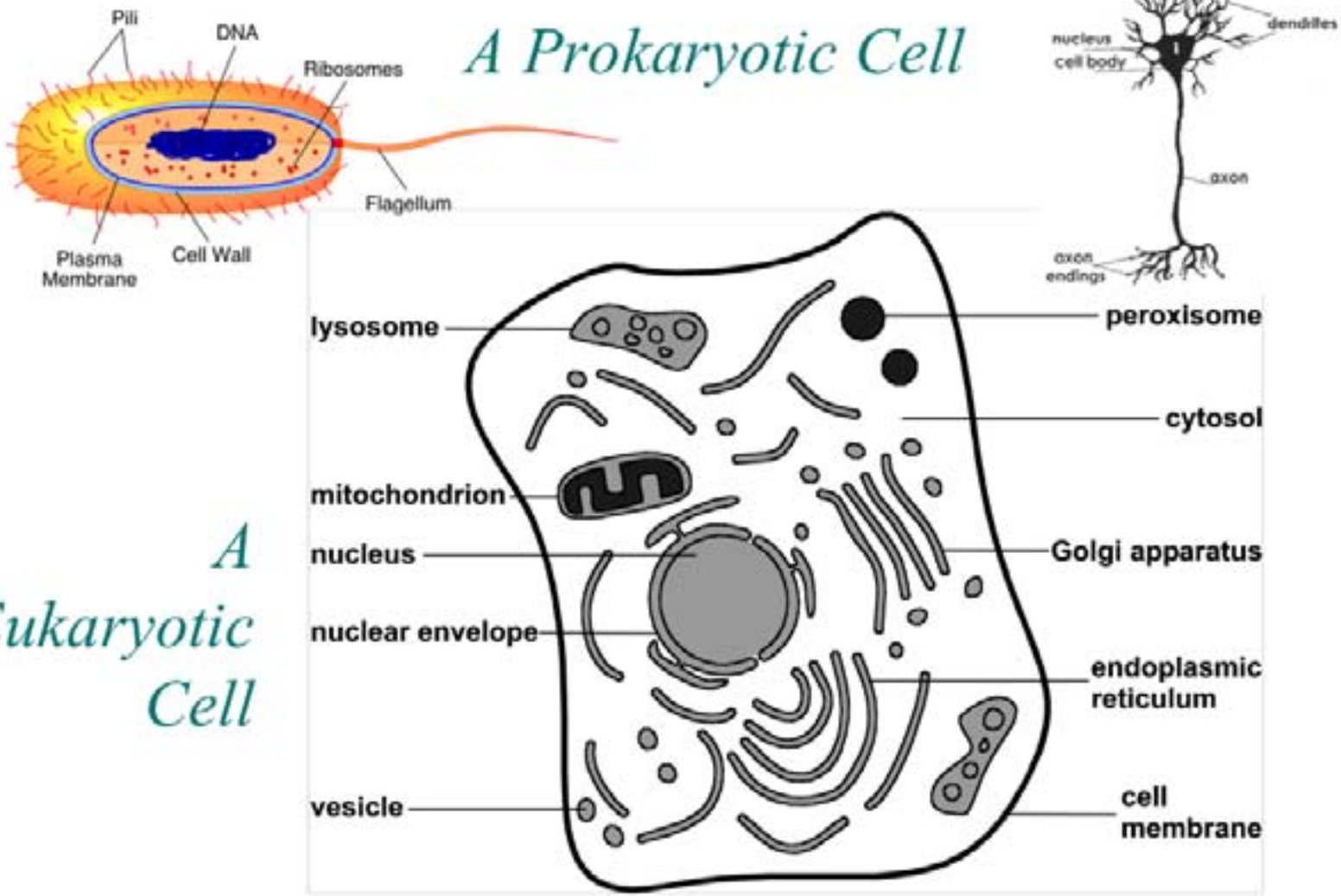
The Subject of Bioinformatics



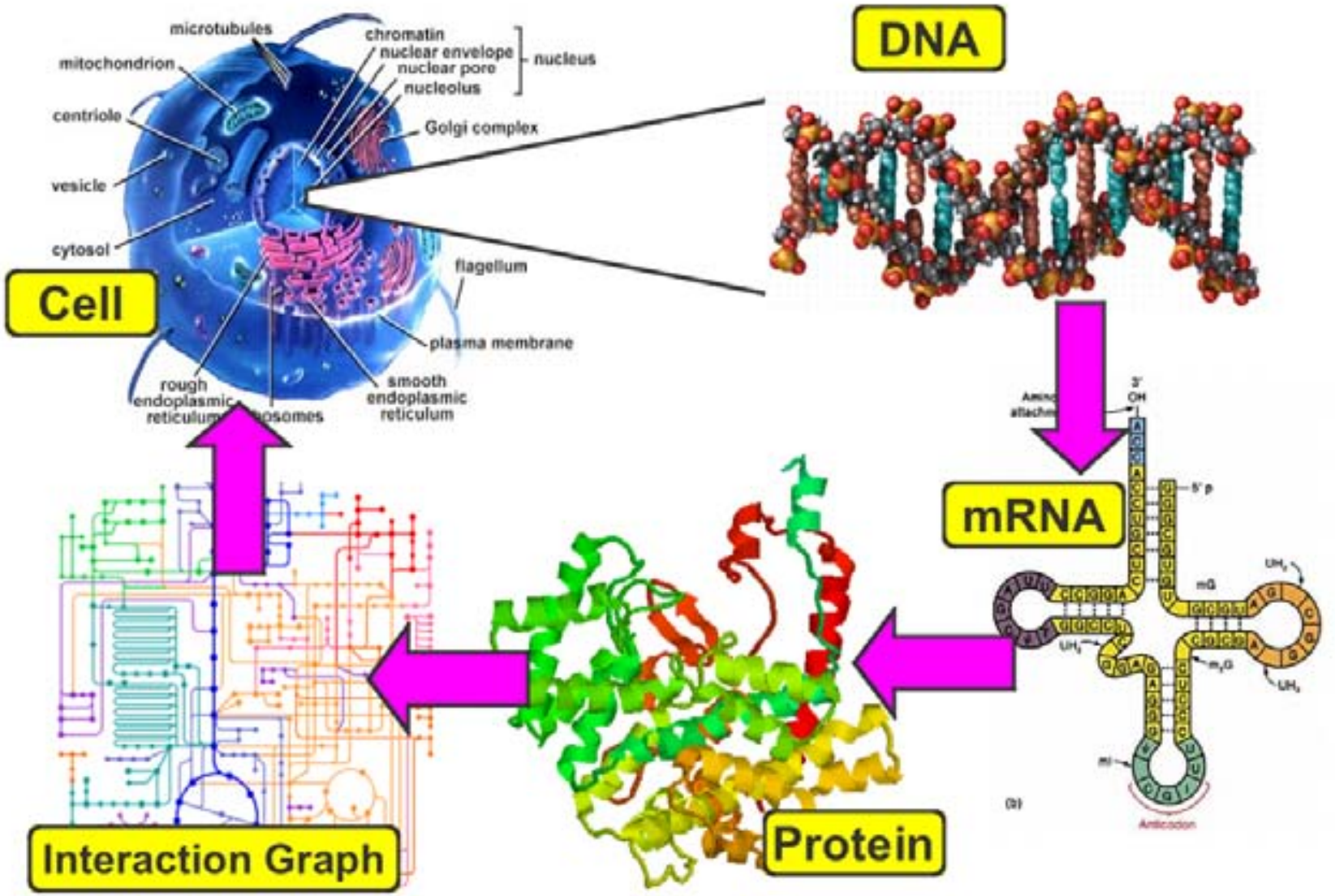
The Three Kingdoms of Life



Basic Unit of Life: the Cell



Biology of the Cell



Short History of Genomics



1866 Laws of heredity (Mendel)
1869 Discovery of the Nuclein (Miescher)
1909 Fruit-fly genetics (Morgan)
1944 DNA supports heredity (Avery)
1953 Structure of DNA (Crick and Watson)
1966 Genetic code (Nirenberg)
1977 Method for DNA sequencing (Sanger)
1982 Creation of Genbank
1990 Human genome project launched
2003 Human genome project completed

Discovery of the Nuclein (Friedrich Miescher, 1869)







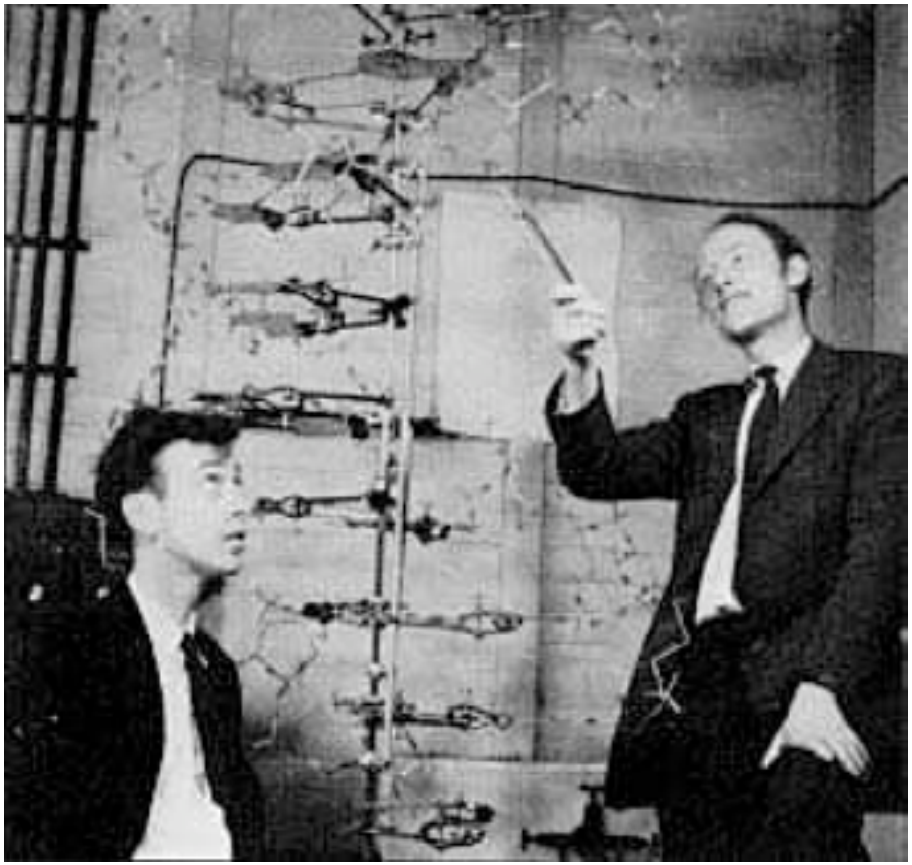


Discovery of Nuclein:

- isolated from Neckar salmon
- "multi-basic acid" (≥ 4)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost consider nuclein" [Miescher, 1874]

Structure of DNA



Watson and Crick [1953]

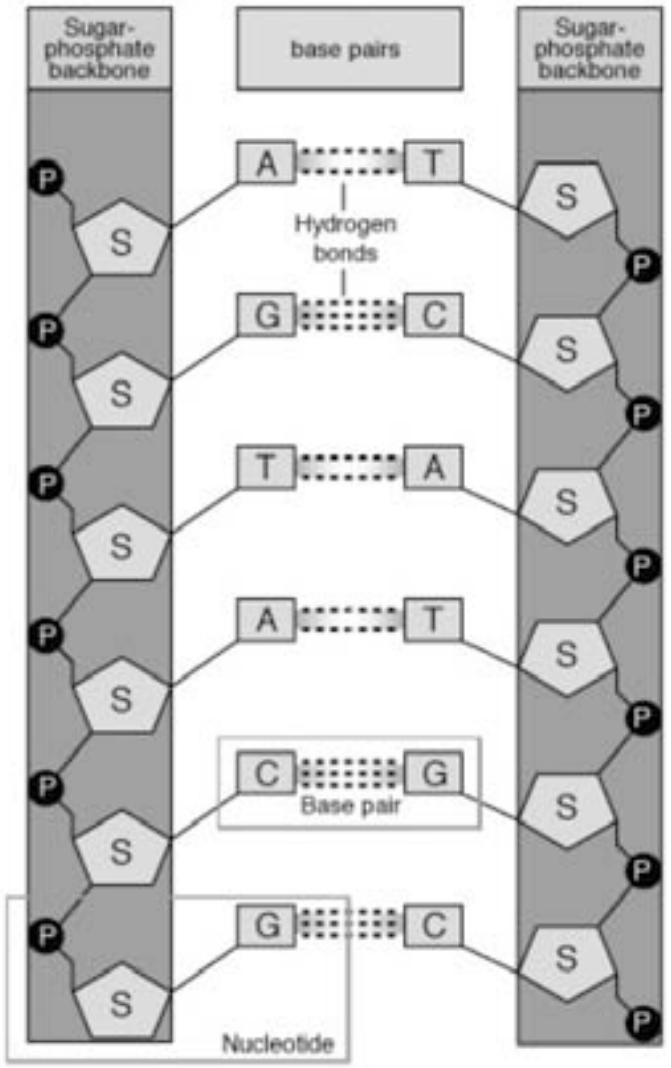
"We wish to suggest a structure for the salt of desoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest"

DNA Double Helix

- linear chain molecule of nucleotides A,C,G,T
- very long: 10^6 (bacteria) to 10^9 (plants, mammals)
- complementarity of double helix:

  ...ACGTTC →
  ← TGCAAG...

- passive, static information storage "genome"

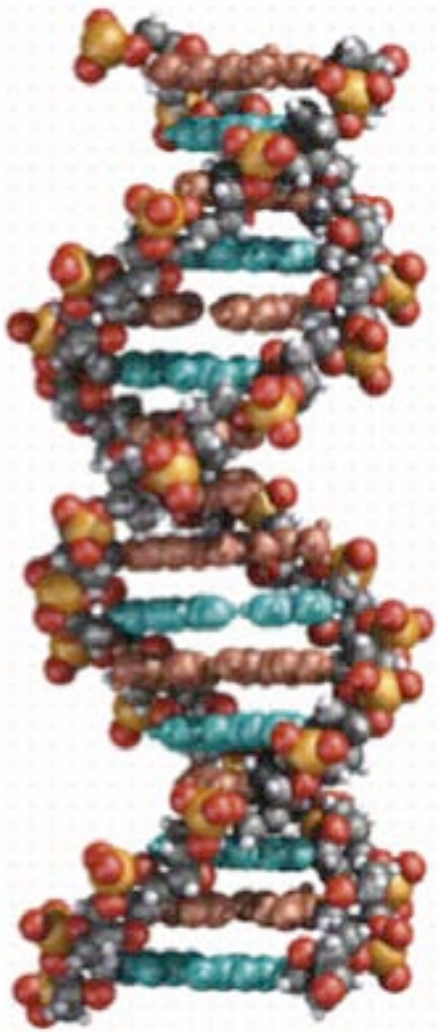


DNA Double Helix

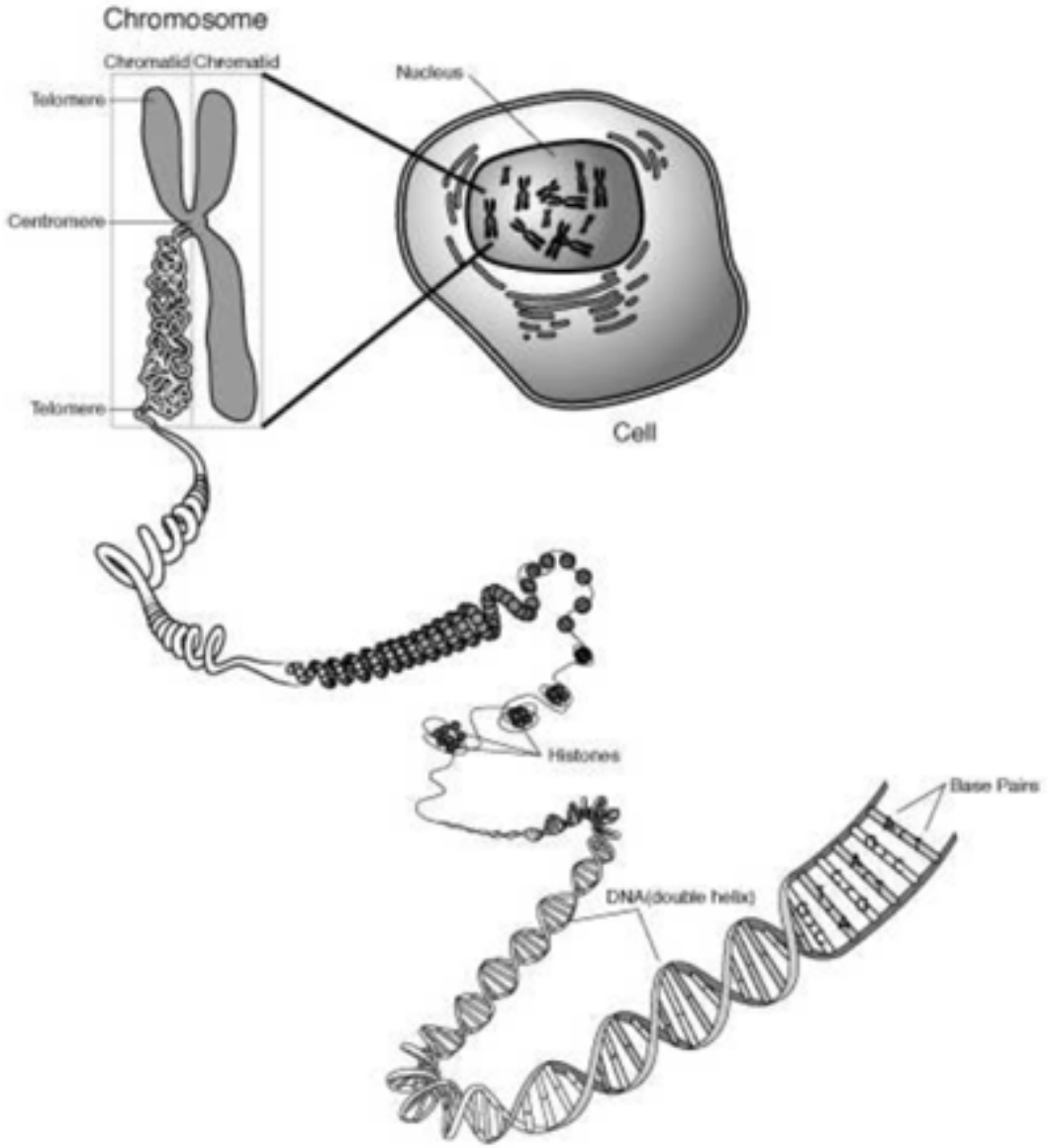
- linear chain molecule of nucleotides A,C,G,T
- very long: 10^6 (bacteria) to 10^9 (plants, mammals)
- complementarity of double helix:

  ...ACGTTC →
  ← TGCAAG...

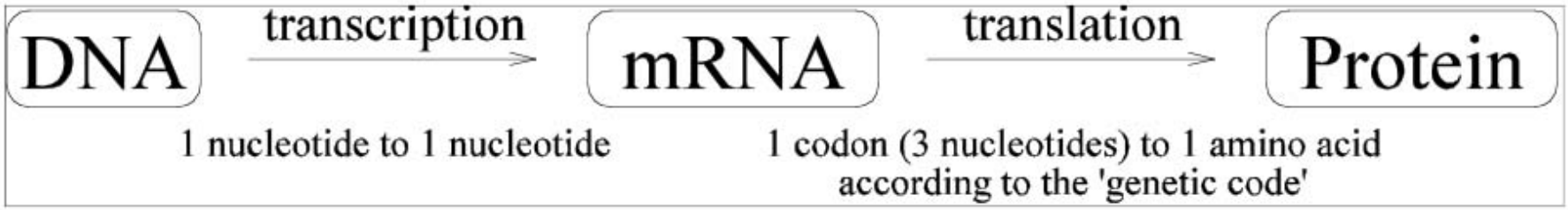
- passive, static information storage "genome"

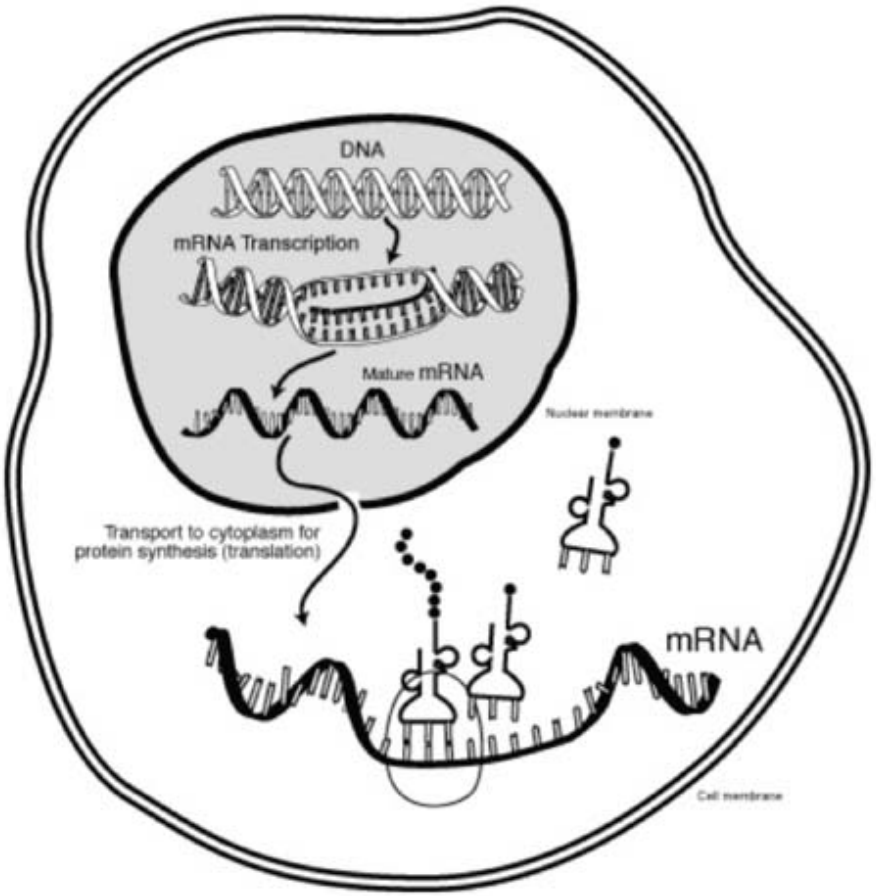


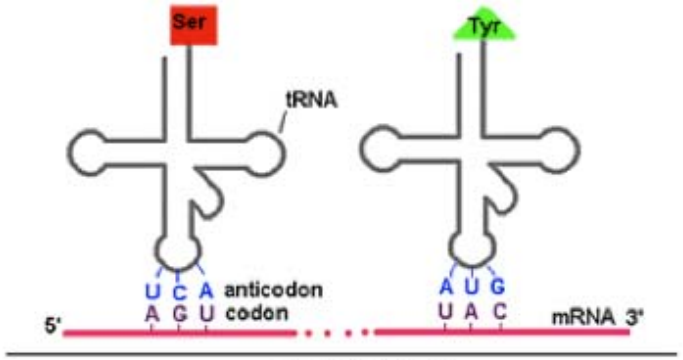
Chromosomes and DNA



Central Dogma





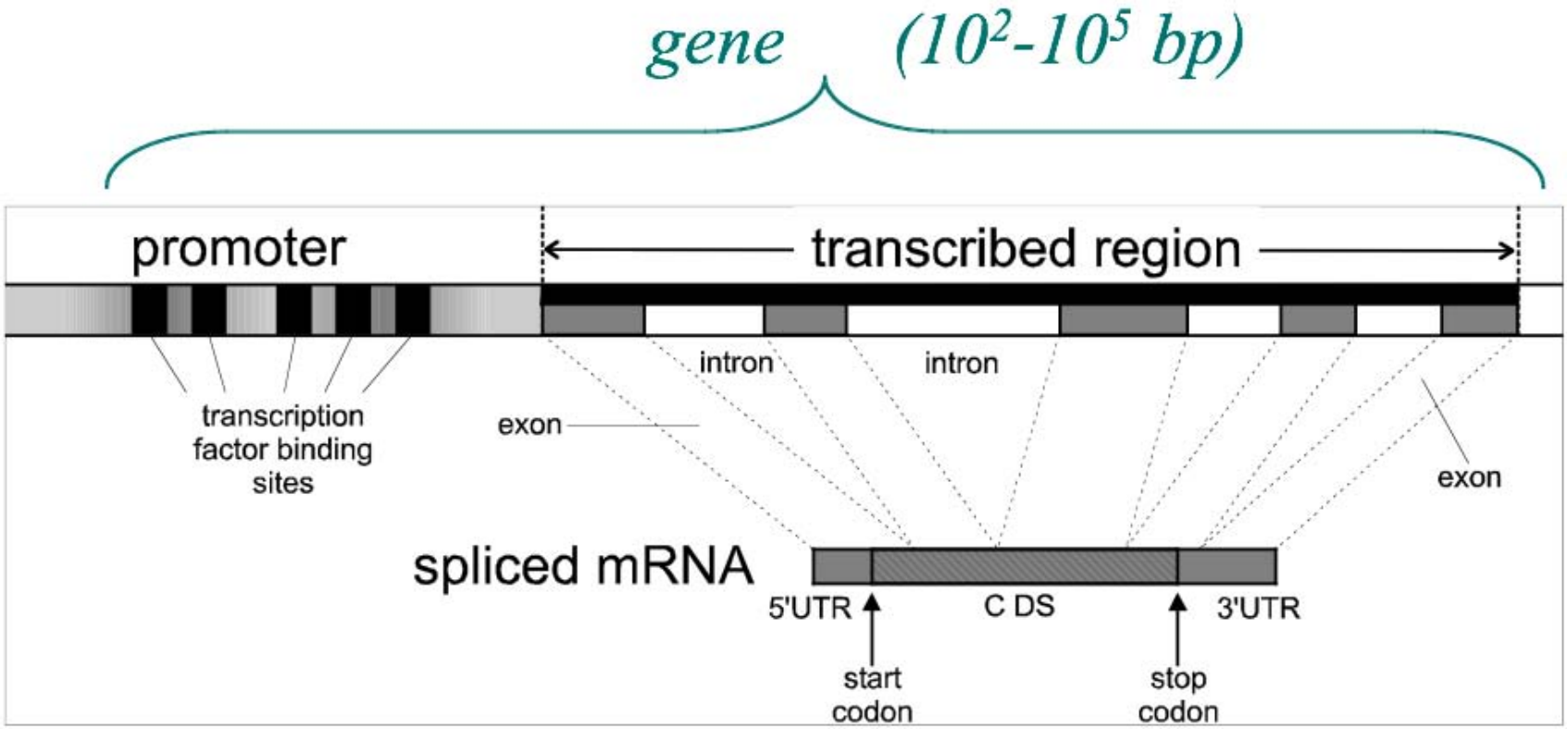


2nd base in codon

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

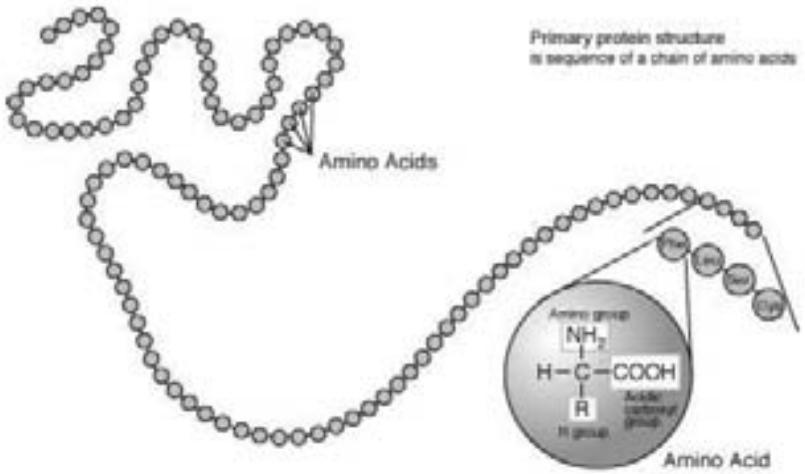
Genetic code

The Genome Encodes Genes



For illustrations of transcription, splicing and translation see for instance:
`http://vcell.ndsu.nodak.edu/animations`

Protein Sequence

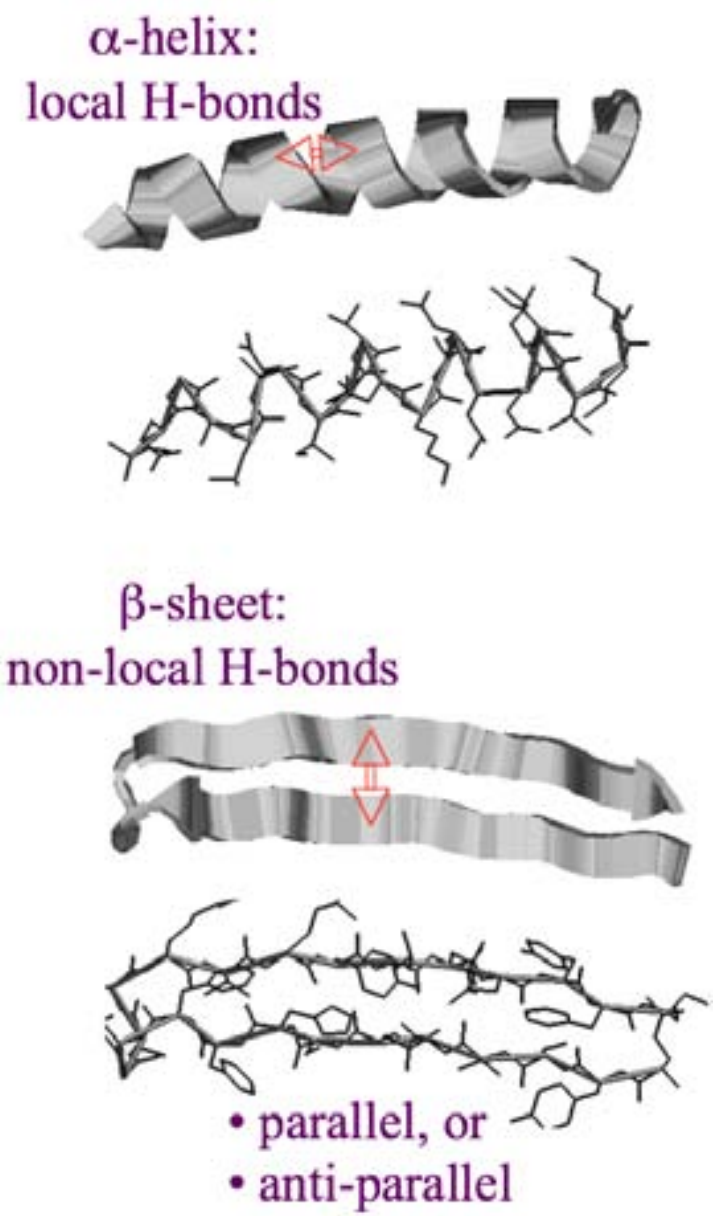


20 Amino different acids:

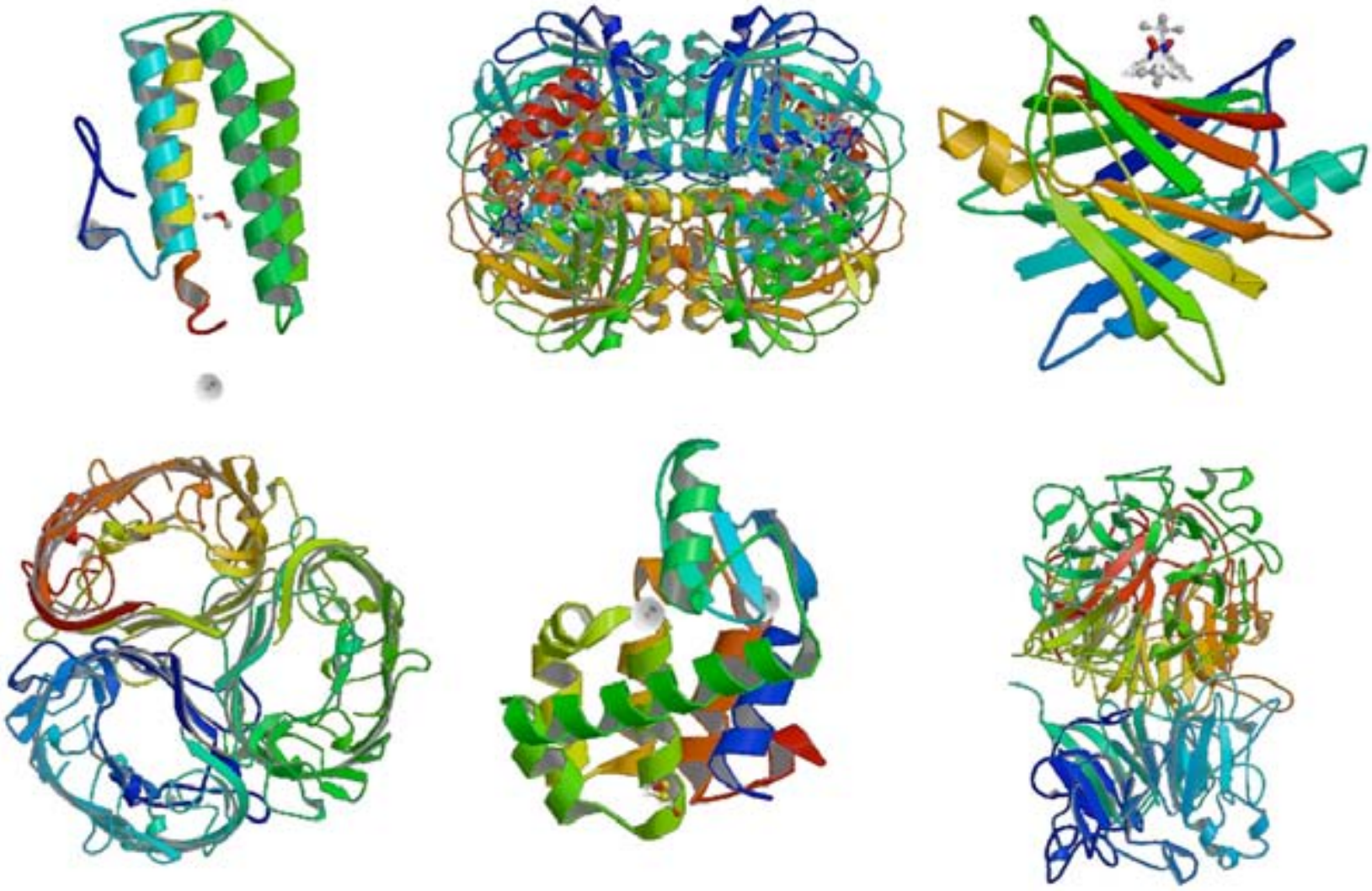
A : Alanine	V : Valine	L : Leucine
F : Phenylalanine	P : Proline	M : Méthionine
E : Acide glutamique	K : Lysine	R : Arginine
T : Threonine	C : Cysteine	N : Asparagine
H : Histidine	V : Thyrosine	W : Tryptophane
I : Isoleucine	S : Sérine	Q : Glutamine
D : Acide aspartique	G : Glycine	

Levels of Protein Structure

- Primary Structure: The sequence of amino acids
- Secondary Structure: Structural Motifs
  - α-helix
  - β-strand
  - other (coil, loop)
- Tertiary Structure: 3D-Structure, packing of secondary structure (in domains)
- Quaternary Structure: Assembly of >1 proteins into single complex



Proteins Have Complex Shapes



Functions of Proteins

- Information processing
  - Extracellular sensing (receptors)
  - Signal transduction
  - Regulation of gene expression
  - Cell cycle regulation
  - Cell differentiation regulation
- Metabolism
  - Energy
  - Protein/DNA/RNA synthesis, ...
  - Biogenesis of cell. components
- Cell structure
  - Cytoskeleton
  - Cellular transport

- Conflicting aspects of function
  - Catalyzed reaction
  - Purpose for cell
- Ambiguity of function
  - Proteins with several functions
  - Function may depend on context

$\Rightarrow$ Functional classification is

- To high degree arbitrary
- Not necessarily a tree

Functions of Proteins

- Information processing
  - Extracellular sensing (receptors)
  - Signal transduction
  - Regulation of gene expression
  - Cell cycle regulation
  - Cell differentiation regulation
- Metabolism
  - Energy
  - Protein/DNA/RNA synthesis, ...
  - Biogenesis of cell. components
- Cell structure
  - Cytoskeleton
  - Cellular transport

- Conflicting aspects of function
  - Catalyzed reaction
  - Purpose for cell
- Ambiguity of function
  - Proteins with several functions
  - Function may depend on context

$\Rightarrow$ Functional classification is

- To high degree arbitrary
- Not necessarily a tree

Functions of Proteins

- Information processing
  - Extracellular sensing (receptors)
  - Signal transduction
  - Regulation of gene expression
  - Cell cycle regulation
  - Cell differentiation regulation
- Metabolism
  - Energy
  - Protein/DNA/RNA synthesis, ...
  - Biogenesis of cell. components
- Cell structure
  - Cytoskeleton
  - Cellular transport

- Conflicting aspects of function
  - Catalyzed reaction
  - Purpose for cell
- Ambiguity of function
  - Proteins with several functions
  - Function may depend on context

$\Rightarrow$ Functional classification is

- To high degree arbitrary
- Not necessarily a tree

Functions of Proteins

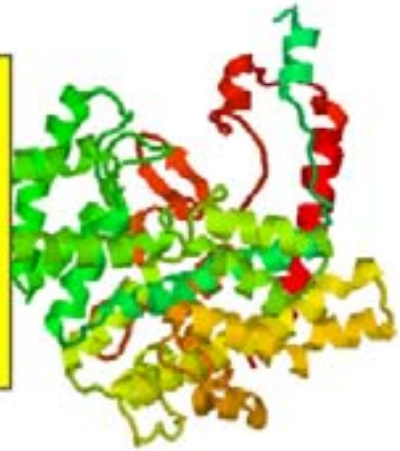
- Information processing
  - Extracellular sensing (receptors)
  - Signal transduction
  - Regulation of gene expression
  - Cell cycle regulation
  - Cell differentiation regulation
- Metabolism
  - Energy
  - Protein/DNA/RNA synthesis, ...
  - Biogenesis of cell. components
- Cell structure
  - Cytoskeleton
  - Cellular transport

- Conflicting aspects of function
  - Catalyzed reaction
  - Purpose for cell
- Ambiguity of function
  - Proteins with several functions
  - Function may depend on context

$\Rightarrow$ Functional classification is
- To high degree arbitrary
- Not necessarily a tree

Most Popular Types of Measurement Data

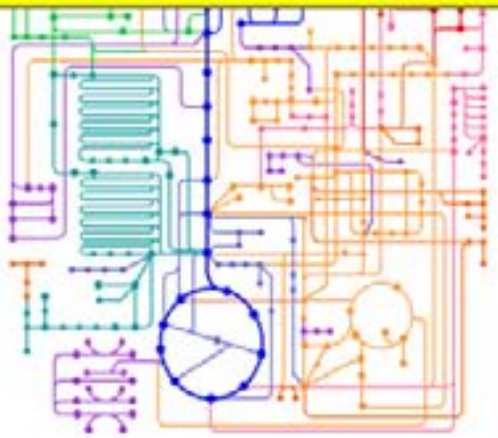
- **Molecule Structures**
  graphs of 3D-coordinates of atoms



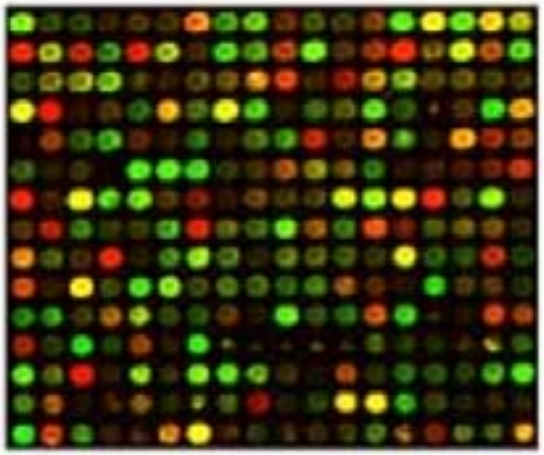
- **Sequences (Genomes, Genes, Proteins, etc.)**
  texts over molecular alphabets

...TGAACTACGA...

- **Protein Networks**
  graphs with annotated nodes and edges



- **Population Data**
  SNPs, Haplotypes, etc



- **Expression Data**
  real-valued matrices

DNA Sequences

Subsequences of $\sim$ 1000 Nucleotides

- RNA converted to cDNA (complementary DNA)
- Prior amplification required

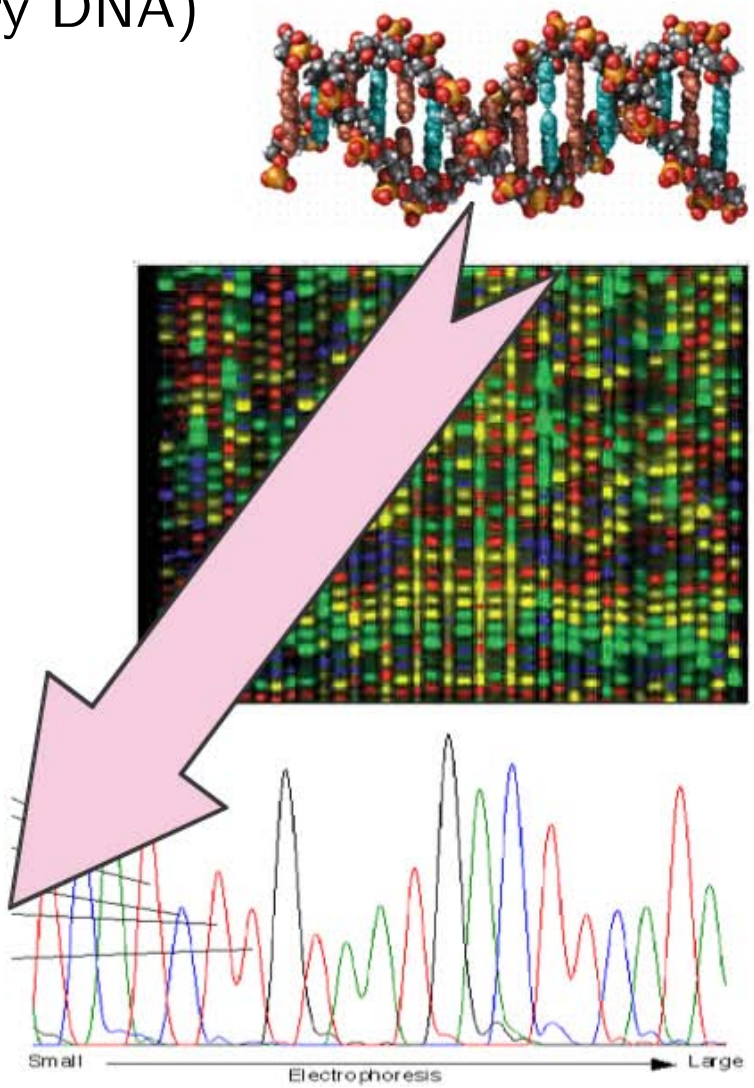
Assembly of genomes

- Repeats hard to resolve
- Strategies:
  - Human Genome Project
  - Celera shotgun sequencing

Cost

- 3 billion dollar for the human genome
- Next generation sequencing: $\leq$\$1million (Illumina, 454 Roche, ABI, ...)





Population Data

- Not just one genome!
- New technology allows to determine variation between populations or individuals
  - Genotyping arrays
  - Next generation sequencing
- $\approx$ 2010: \$1000 genome
- Typically recorded as deviation from a reference
  - Single Nucleotide Polymorphisms (SNPs)
  - Insertions or deletions (Indel)
- Heterozygosity complicates analyses
- Identification of Haplotypes
  - Human Hapmap project

About 1 SNP per 1kb sequence, often clustered:

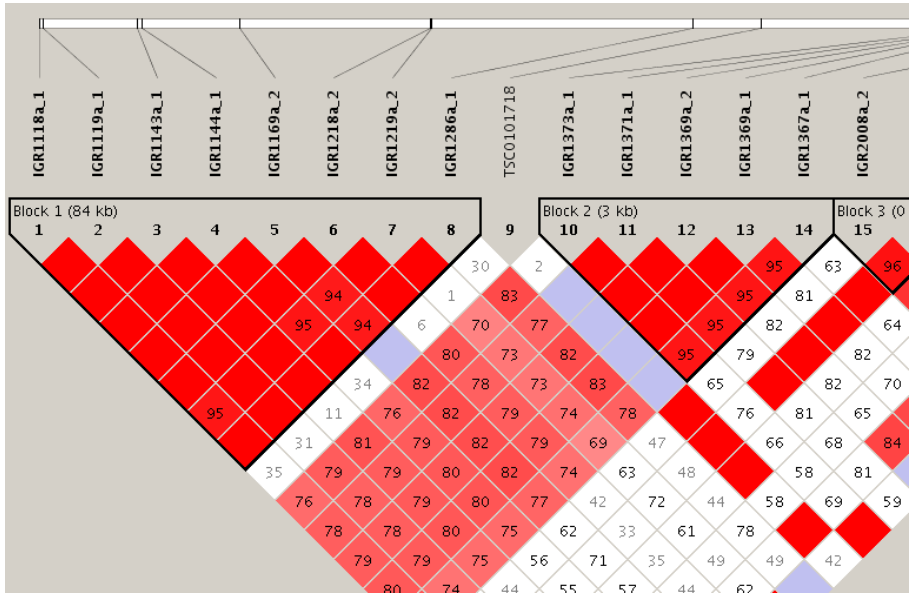
```
GT 1 ...ACTGTGCATGAGTCCTGTAAATGCTCCCTTACGTGAGCGCG...
GT 2 ...ACTGTGCATGAGTCCTGTAAATGCTCCCTAACGGGAGCGCG...
GT 3 ...ACTGTGCATGAGTCCTGTAAATGCTCCCTAACGGGAGCGCG...
GT 4 ...ACTGTGCACGAGTACTGTAAATGCTCCCTAACGGGAGCGCG...
GT 5 ...ACTGTGCACGAGTACTGTAAATGCTCCCTTACGTGAGCGCG...
GT 6 ...ACTGTGCACGAGTACTGTAAATGCTCCCTTACGTGAGCGCG...
```

Population Data

- Not just one genome!
- New technology allows to determine variation between populations or individuals
  - Genotyping arrays
  - Next generation sequencing
- $\approx$ 2010: \$1000 genome
- Typically recorded as deviation from a reference
  - Single Nucleotide Polymorphisms (SNPs)
  - Insertions or deletions (Indel)
- Heterozygosity complicates analyses
- Identification of Haplotypes
  - Human Hapmap project

About 1 SNP per 1kb sequence, often clustered:

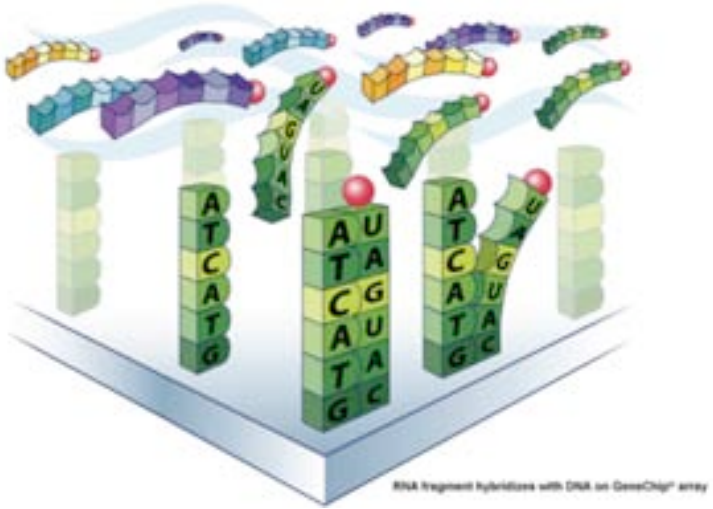




Gene Expression

Microarrays (also: DNA-Chips)

- two-channel spotted cDNA
- one-channel spotted cDNA
- on-chip synthesis of oligo-nucleotides (Affymetrix)

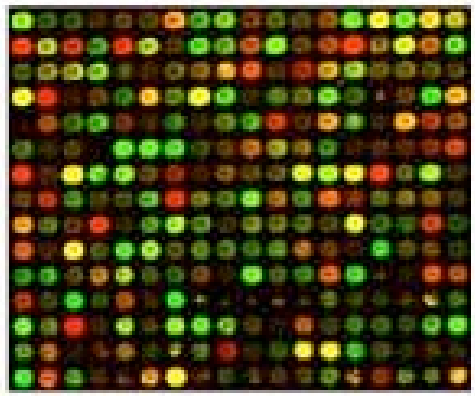


Many Non-Trivial Low-Level Problems

- image analysis – noise models
- normalization – replicates

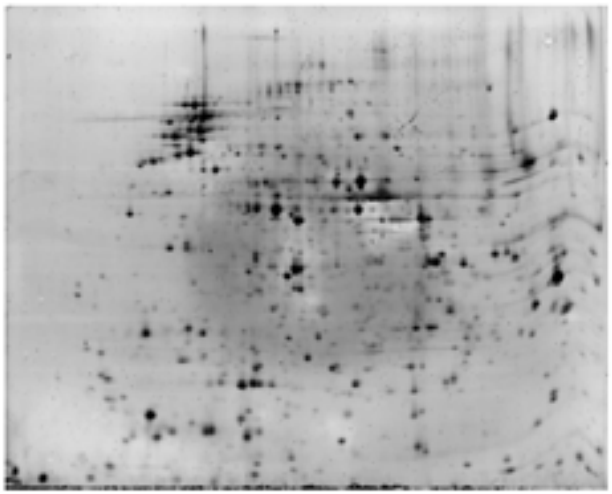
Many Other Technologies

- qPCR (quantitative polymerase chain reaction)
- MPSS (massively parallel signature sequencing)
- SAGE (serial analysis of gene expression)
- etc.

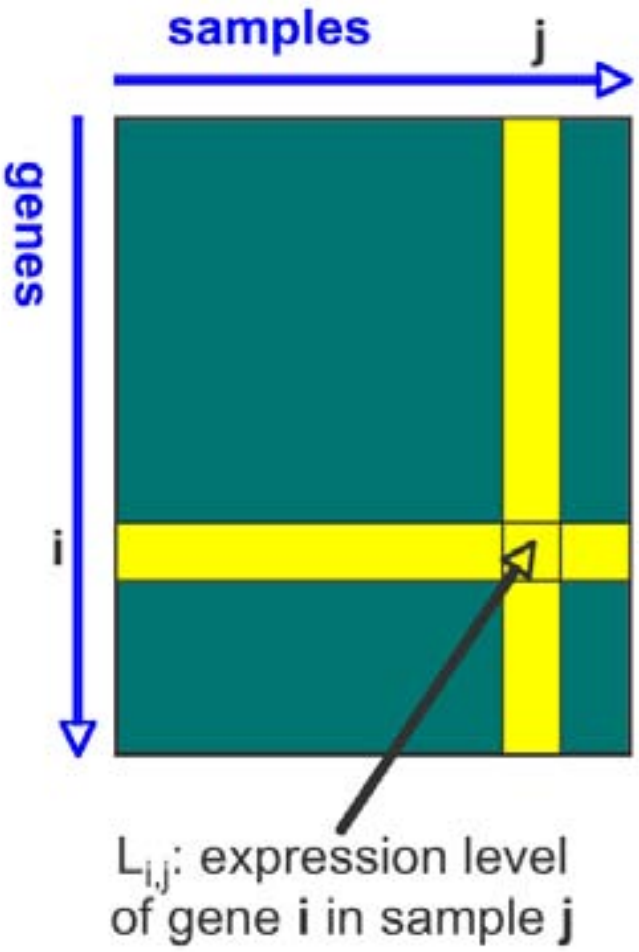


Expression (of Proteins, Metabolites)

- Protein expression
  - 2D Gel Electrophoresis
  - Mass Spectrometry
  - Immunoprecipitation
  - Protein-Chips (Microarrays)
- Metabolite abundances
- Similar in gene expression
  - Low level problems
  - But higher error rates

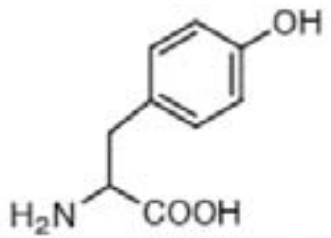


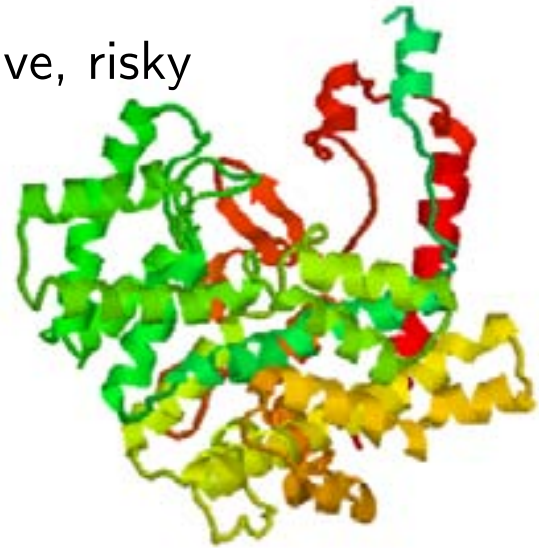
- Common Data Structure:
  The (Gene) Expression Matrix



3D-Structures (of Proteins)

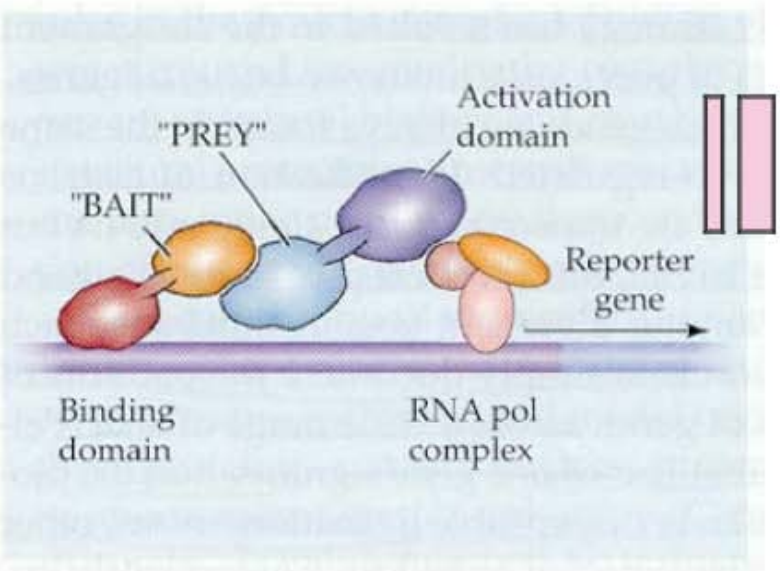
- Nuclear Magnetic Resonance (NMR)
  - get distance of heavy atoms that carry magnetic spin
  - works for molecules in solution
  - not very precise for macromolecules
- X-Ray Crystallography
  - determine electron density from crystal
  - not necessarily identical to native structure
  - proteins do not like to crystallize: slow, expensive, risky
- Further interesting structures
  - Metabolites (small molecules)
  - DNA
  - RNA
  - Complexes

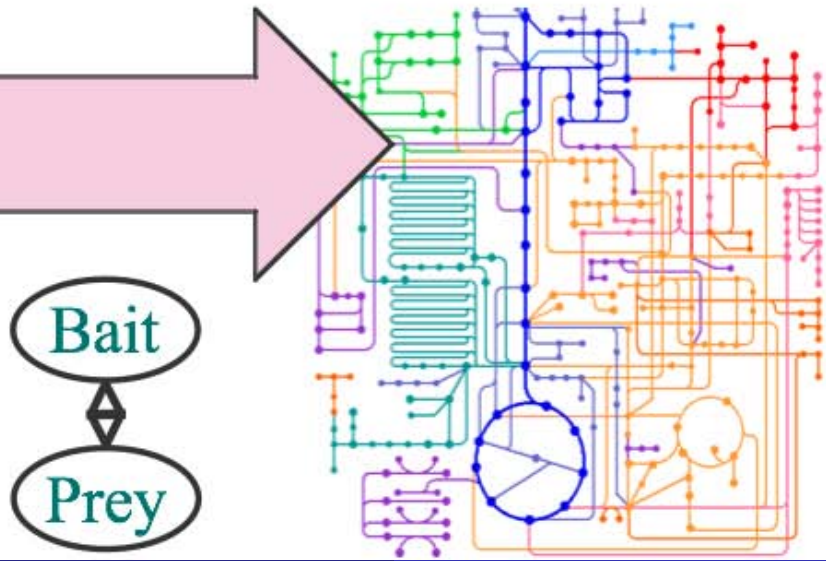




Interactions

- Protein-Ligand (small molecule)
  - High-throughput screening
- Protein-Protein
  - Yeast-2-hybrid
  - Tandem-affinity purification
  - High-throughput mass spectrometric protein complex identification
- Protein-DNA/RNA
  - Transcriptional regulation
  - RNA processing regulation





Further Available Data Types ...



- Protein-DNA Interaction
- Subcellular Localization
- Protein Sequence (MS: mass spectrometry)
- Protein Modifications (MS)
- High-throughput Screening (HTS)
- many more exist and ...
- new ones are constantly being invented

Why Analyze These Data?



. . . to replace other measurements (save costs)

- e.g., predict protein structure from sequence

. . . to understand biology

- e.g., reconstruct rules underlying biological mechanisms

. . . to reconstruct the past

- e.g., infer evolution of species

. . . to predict the future

- e.g., medical diagnosis

. . . to improve organisms

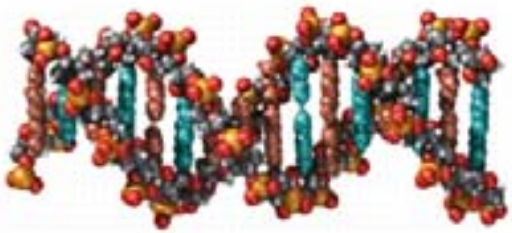
- e.g., make better beer!

Lots of Different, Specialized Problems



- Find Structure in DNA
- Predict Properties of Proteins
- Relate Molecular to Macroscopic Data
- Low level problems:
  - Image analysis
  - Measurement normalization
  - Sequence processing
- Many others . . .

Find Structure in DNA



find structure in DNA

Gene Finding

reconstruct evolutionary history

find genes and resolve gene structure

Genome Comparison

Transriptional Regulation

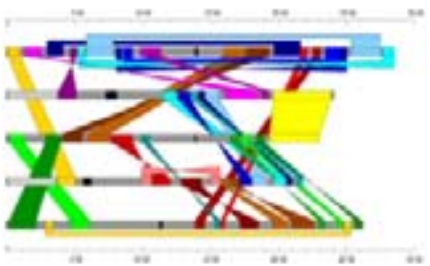
find rearrangements and duplications

reconstruct phylogeny of species

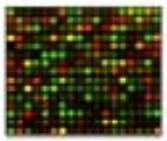
identify promoter elements

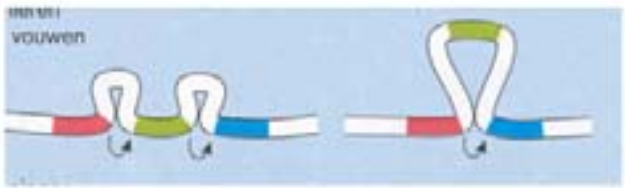
identify splice sites

identify coding sequence

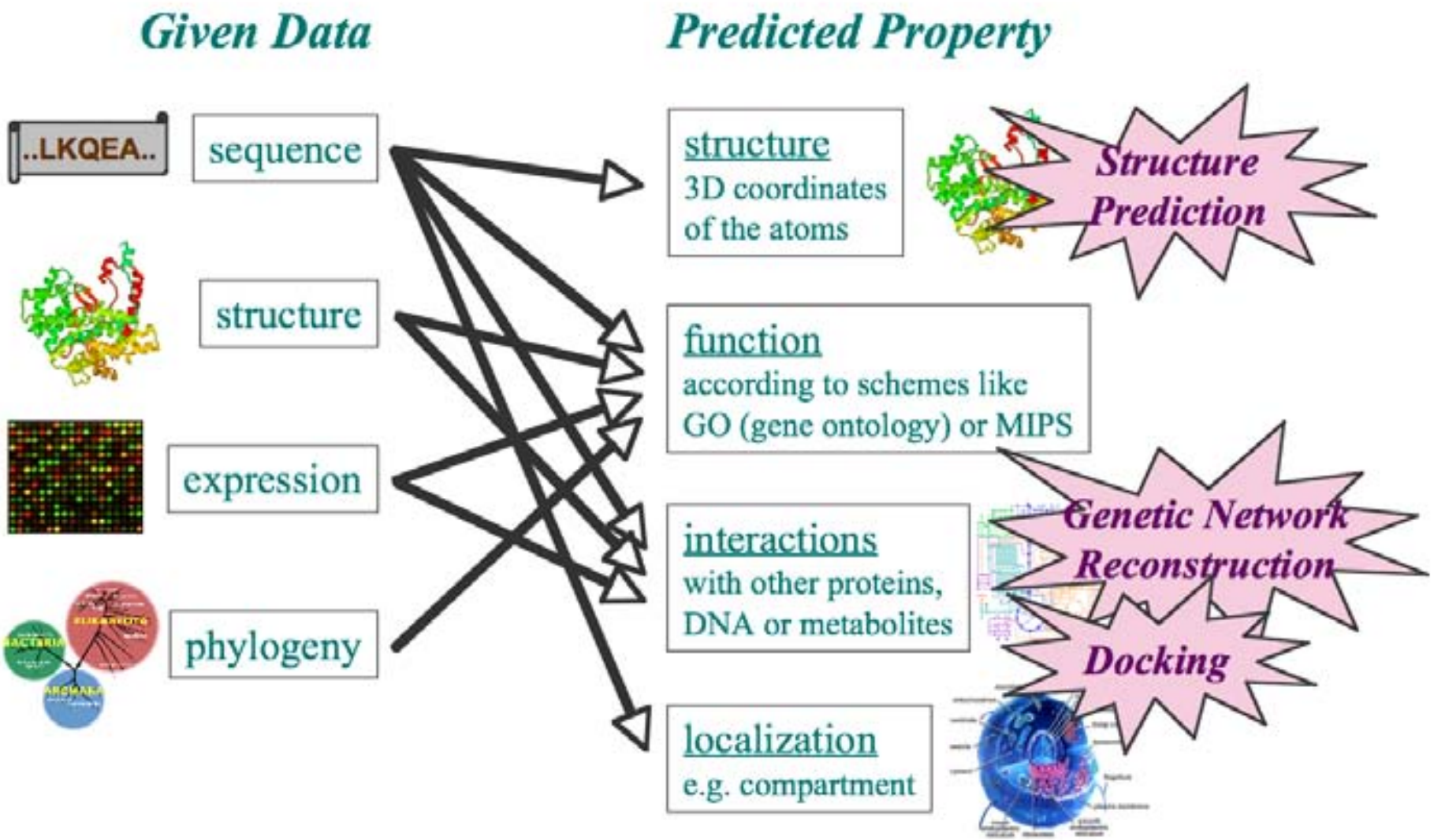






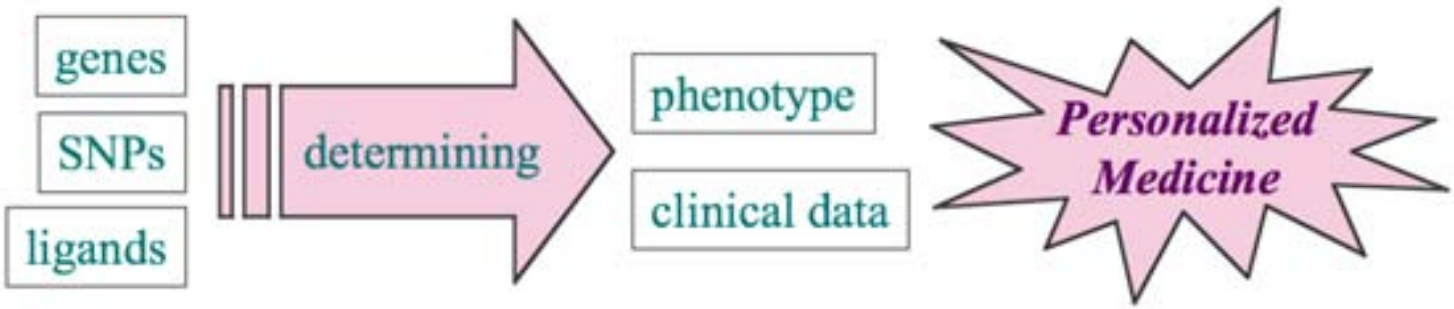


Predict Protein Properties



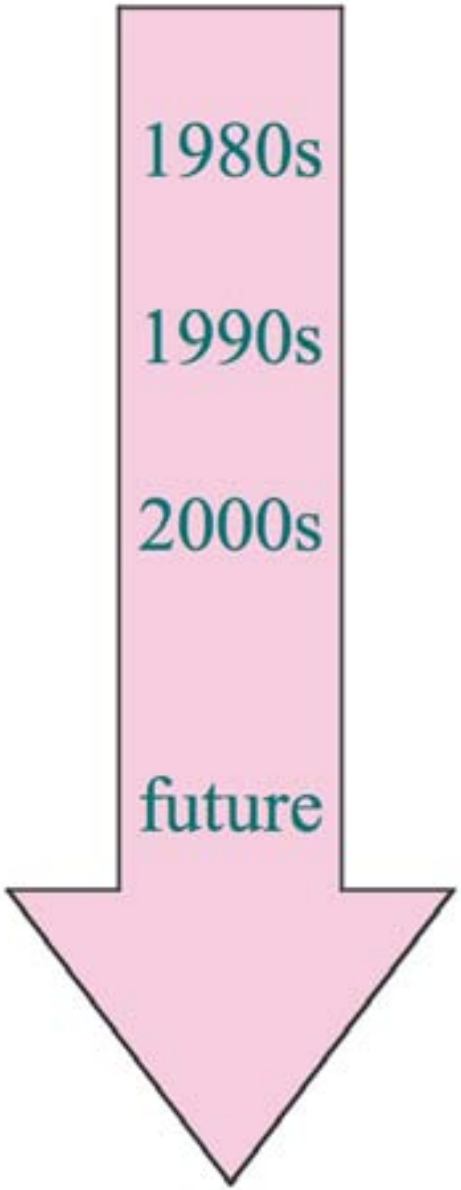
Relate Molecular to Macroscopic Data

- Identify Molecular Causes of Macroscopic States
  - classification or regression with feature selection



- Identify or Refine Macroscopic States
  - clustering of gene expression data
  - blocking of haplotypes (?)
- Simulate Dynamics of Entire Cells and Organs
  - Systems Biology

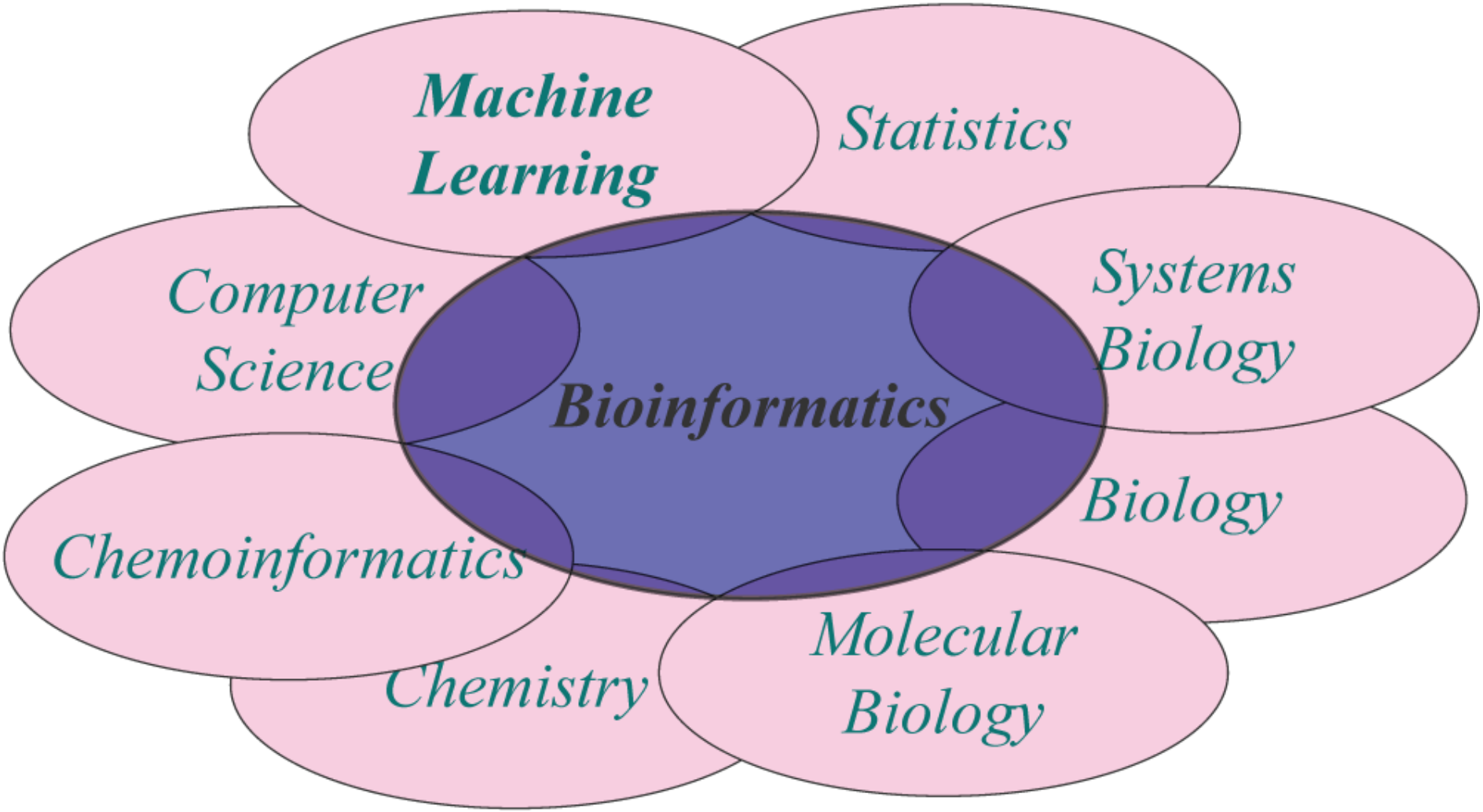
History of Bioinformatics



- sequence similarity (DNA, protein)
- evolution / phylogeny
- protein structures
- interactions of molecules
- dynamics (protein folding, docking)
- model promoters (regulation of expression)
- "personalized medicine" (understanding polymorphisms)
- "systems biology" (simulation of cells)
- interactions of cells
- developmental biology
- simulating organs and their interactions
- understanding the brain
- designing new organisms

$\Rightarrow$ enough to do for the next decades!

Bioinformatics



Challenges for Machine Learning



Why solve these problems with machine learning?

- Too complex to be solved from first principles
- More or less labeled data available

In which ways are they challenging for machine learning?

- Large amounts of data (e.g., many genomes)
- Noisy measurements
- Unlabeled data
- Structured data (e.g., vectorial, strings, graphs)
- Structured output (e.g., trees, sequences, graphs)
- Combination of different data types

Resources



Conferences Intelligent Systems in Molecular Biology (ISMB), Research in Computational Molecular Biology, European Conference on Bioinformatics (ECCB), Pacific Symposium on Biocomputing (PSB) (for more see `http://www.iscb.org/events/event_board.php`)

Journals PLoS Computational Biology, Bioinformatics, Journal of Computational Biology, BMC Bioinformatics (for more see e.g. `http://www.iscb.org/journals.shtml`)

Books/Review Papers Genes IX [Lewin, 2007], Molecular Biology of the Cell [Alberts et al., 2002], Biological Sequence Analysis [Durbin et al., 1998], also check out the growing Education Collection at PLoS CompBio

Data sources NCBI databases (including Genbank), Genome browsers (e.g. UCSC, Ensembl), Swissprot, Protein Data Bank (PDB), etc.; see for instance NAR Database issue