# 1. Bioinformatics and Computational tools for high-throughput analysis of biological data

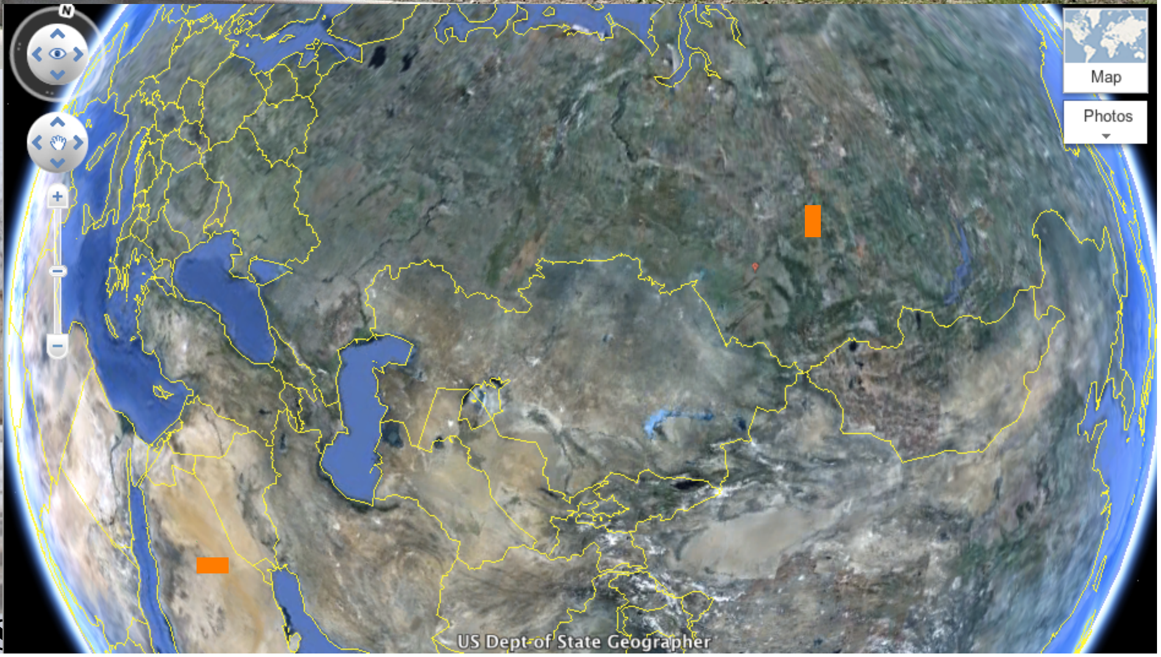
1. Bioinformatics and Big problems in Biology
2. Next Generation Sequencing, Genome assembling and bacterial gene identification
3. HMM eukaryotic gene finding, fast sequence reads alignment, big data analysis

Victor Solovyev
Computer, Electrical and Mathematical Sciences and Engineering Division
KAUST, Saudi Arabia

*The lecture 1 uses personal as well as publicly available WEB and publications materials*

Akademgorodok,
Novosibirsk

**Novosibirsk State Univers**

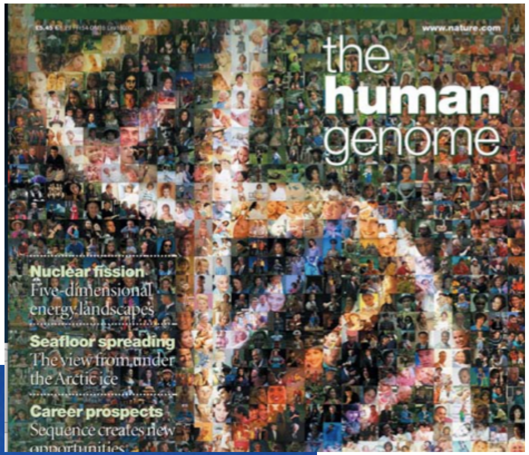


# Supercomputer Computations Research Institute (SCRI), the Florida State University

# Baylor College of Medicine, Huston

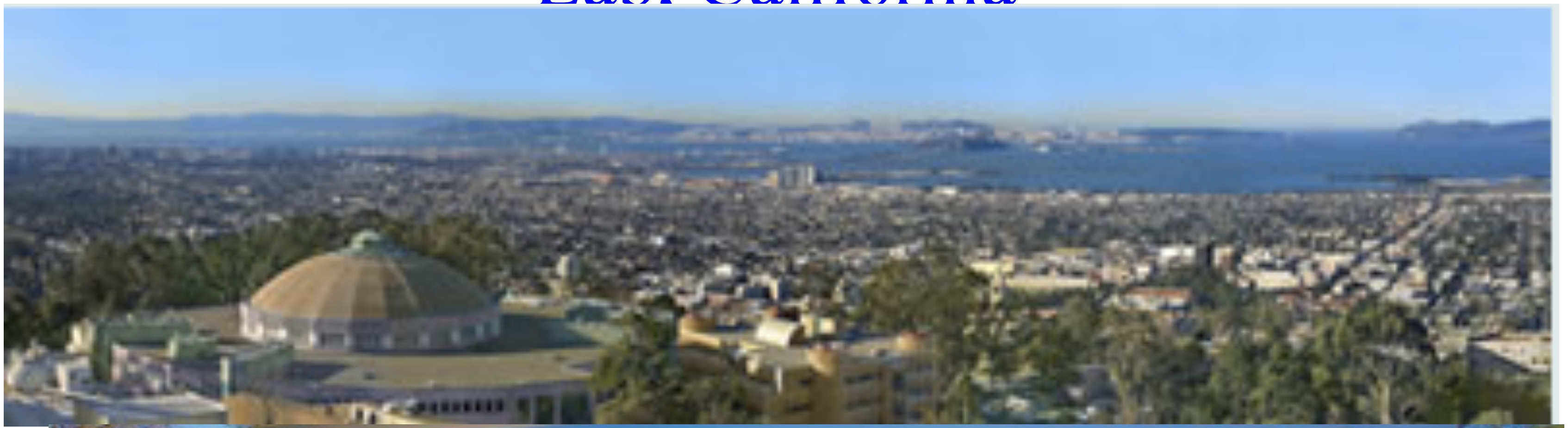
# Amgen Inc., Los Angeles

# The Sanger Centre, Cambridge, UK


www.nature.com
the
human
genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities

Computational Genomic group
Human genome Sequencing era

# Joint Genome Institute, Berkeley National Lab. California

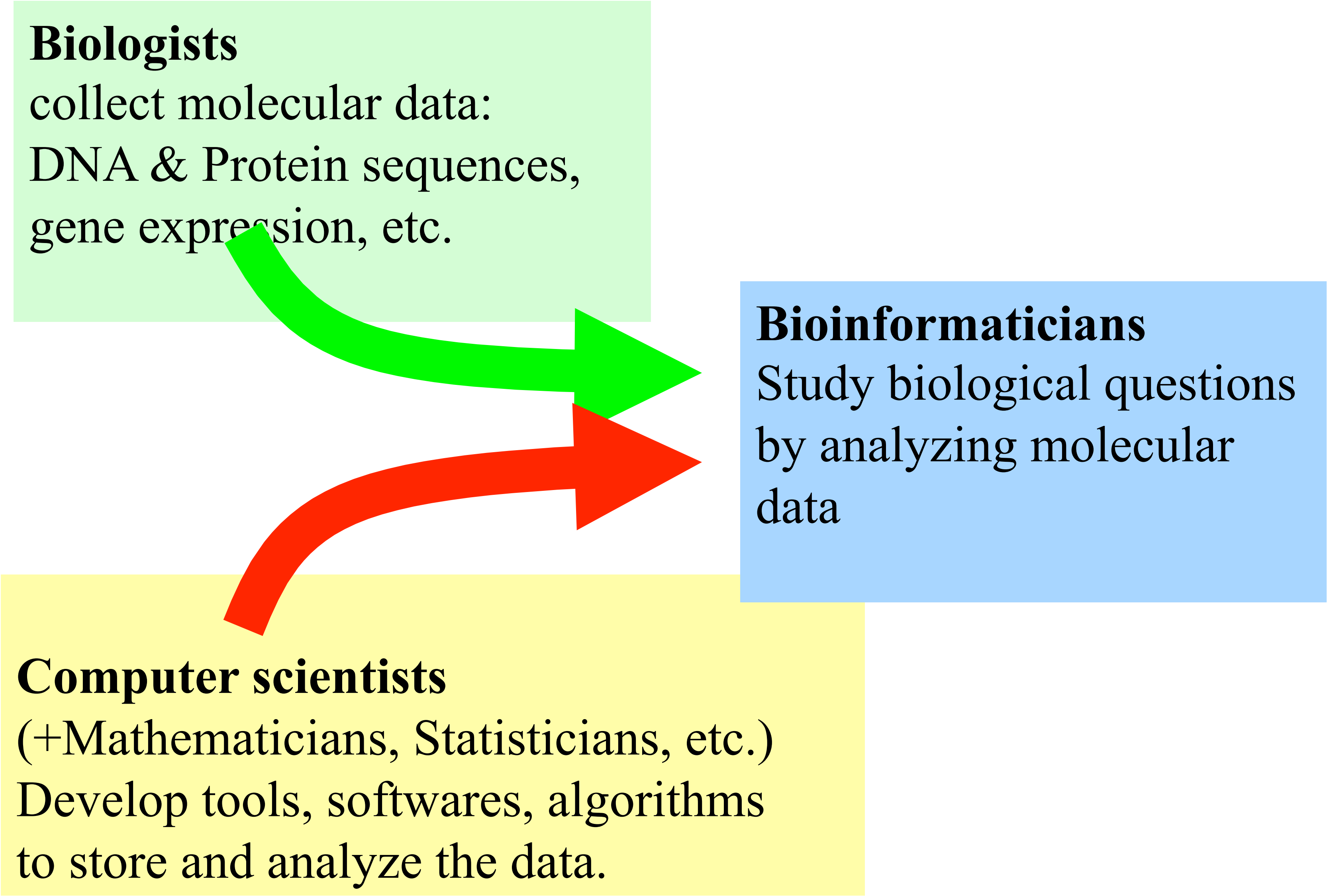


Genome annotation group

# Royall Holloway, University of London

KAUST (Saudi Arabia)

# Bioinformatics - The application of computer science and mathematics to solve biological problems

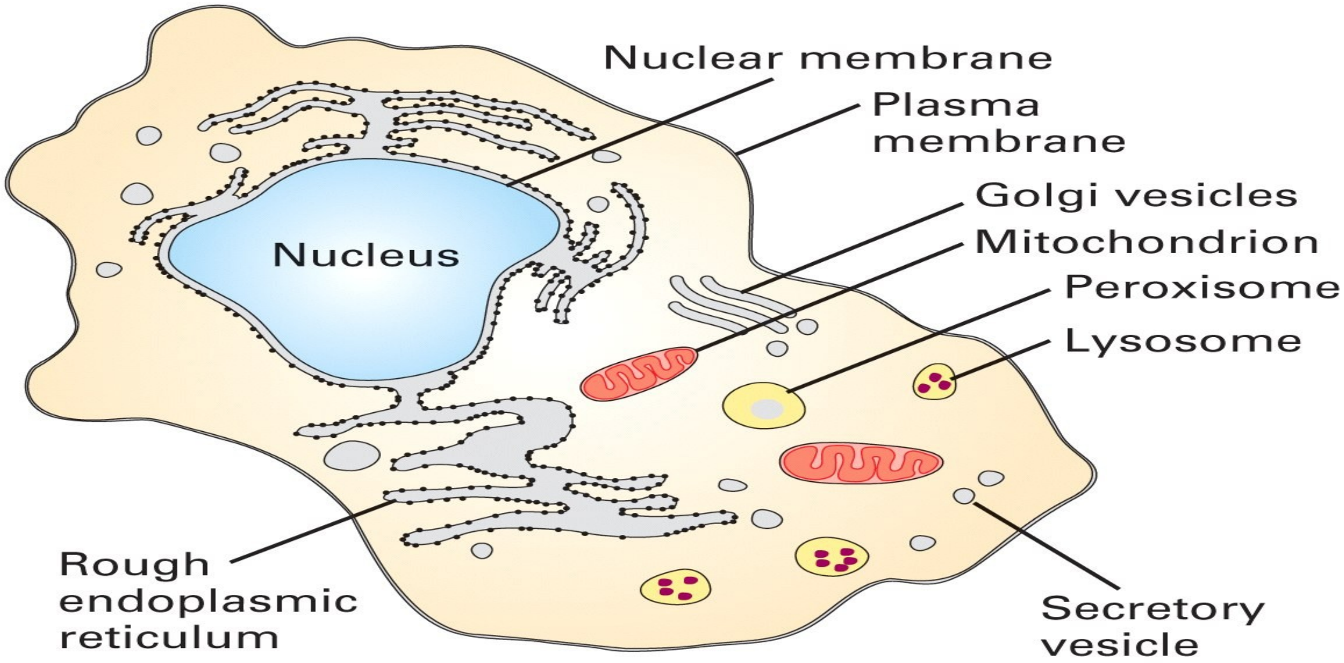


**Biologists**
collect molecular data:
DNA & Protein sequences,
gene expression, etc.

**Bioinformaticians**
Study biological questions
by analyzing molecular
data

**Computer scientists**
(+Mathematicians, Statisticians, etc.)
Develop tools, softwares, algorithms
to store and analyze the data.

# Life begins with the cell



- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All cells have some common features

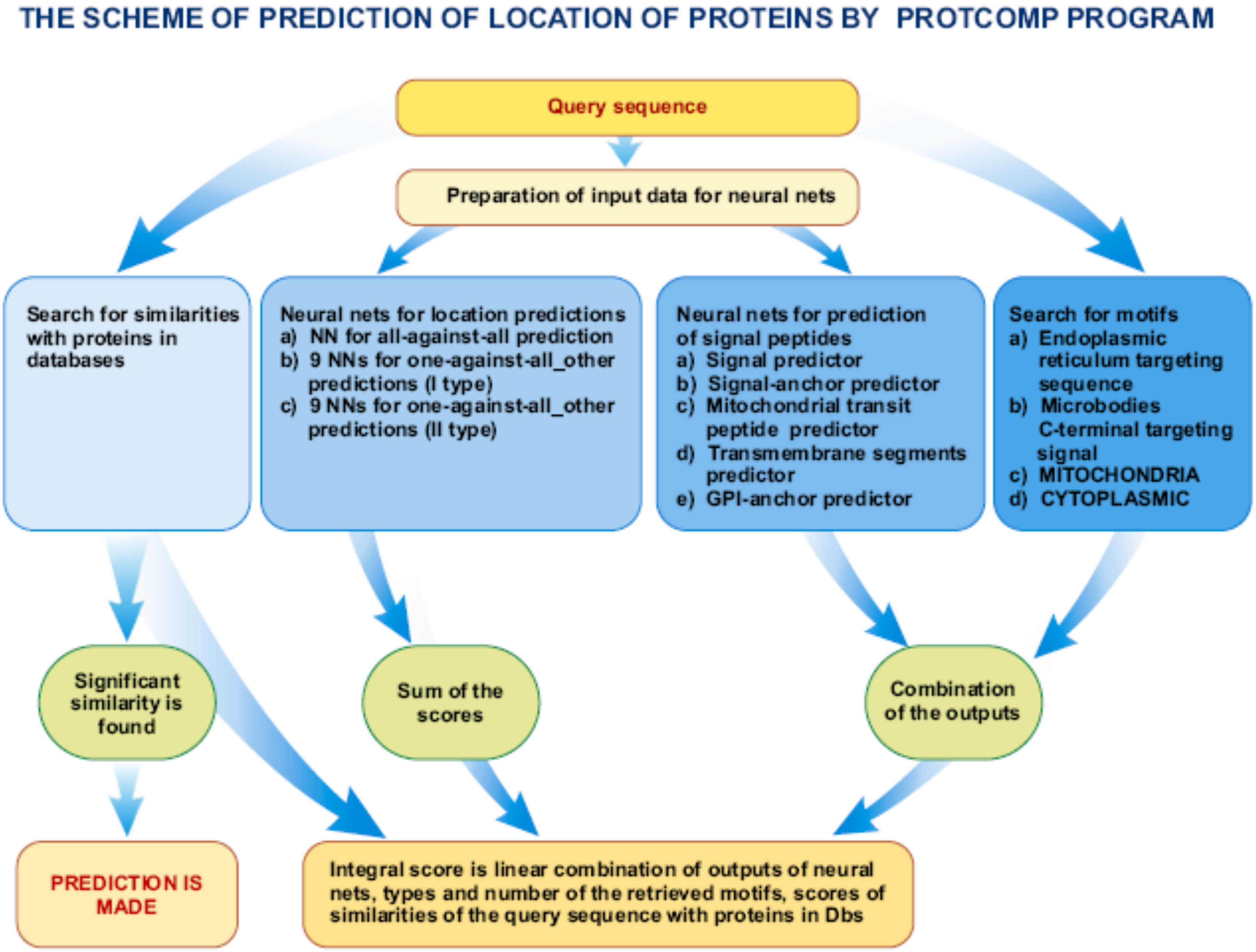

THE SCHEME OF PREDICTION OF LOCATION OF PROTEINS BY PROTCOMP PROGRAM
Query sequence
Preparation of input data for neural nets
Search for similarities with proteins in databases
Neural nets for location predictions
a) NN for all-against-all prediction
b) 9 NNs for one-against-all_other predictions (I type)
c) 9 NNs for one-against-all_other predictions (II type)
Neural nets for prediction of signal peptides
a) Signal predictor
b) Signal-anchor predictor
c) Mitochondrial transit peptide predictor
d) Transmembrane segments predictor
e) GPI-anchor predictor
Search for motifs
a) Endoplasmic reticulum targeting sequence
b) Microbodies C-terminal targeting signal
c) MITOCHONDRIA
d) CYTOPLASMIC
Significant similarity is found
Sum of the scores
Combination of the outputs
PREDICTION IS MADE
Integral score is linear combination of outputs of neural nets, types and number of the retrieved motifs, scores of similarities of the query sequence with proteins in Dbs

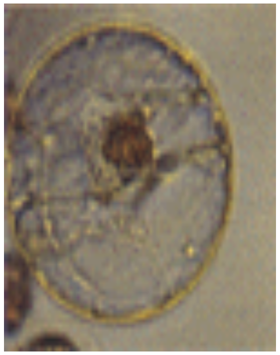
**Figure 2**. Logical scheme of protein subcellular localization prediction by ProtComp.

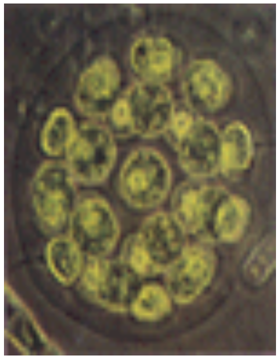
```
ProtComp  Identifying sub-cellular location (Plants)

Seq name: Q9LVV5 Location:Chloroplast  DE  Thylakoid lumenal 19.6 kDa
protein, chloroplast precursor. 179
Significant similarity in Potential Location DB -  Location:Chloroplast
Database sequence: AC=Q9LVV5 Location:Chloroplast  DE  Thylakoid
lumenal 19.6 kDa protein, chloroplast
Score=9050, Sequence length=179, Alignment length=179
Predicted by Neural Nets - Chloroplast with score    2.7
******** Chloroplast Transit peptide 1-31 is found
******** Transmembrane segments are found: .+52:75-.
Integral Prediction of protein location: Membrane bound Chloroplast
with score    3.7
Location weights:     LocDB / PotLocDB / Neural Nets / Integral
 Nuclear                0.0 /      0.0 /        0.73 /     0.73
 Plasma membrane        0.0 /      0.0 /        0.87 /     0.87
 Extracellular          0.0 /      0.0 /        0.80 /     0.80
 Cytoplasmic            0.0 /      0.0 /        0.71 /     0.71
 Mitochondrial          0.0 /      0.0 /        0.60 /     0.60
 Chloroplast            0.0 /   9050.0 /        2.65 /     3.66
 Endoplasm. retic.      0.0 /      0.0 /        0.71 /     0.71
 Peroxisomal            0.0 /      0.0 /        0.60 /     0.60
```

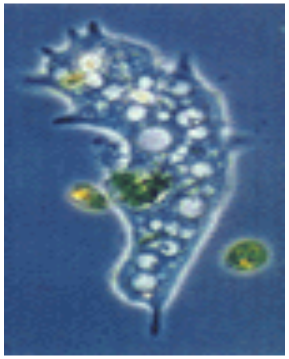
| Compartment | Percent predicted correctly |
|---|---|
| | ver. 5 |
| Nucleus | 88 |
| Plasma Membrane | 87 |
| Extracellular | 83 |
| Cytoplasm | 63 |
| Mitochondria | 82 |
| Endoplasmic Retic | 83 |
| Peroxisome | 97 |
| Lysosome | 91 |
| Golgi | 77 |

# Cell Information and Machinery

- A cell stores all information to replicate itself
  - Human genome is around 3 billion base pairs long
  - Almost every cell in human body contains same set of genes
  - But not all genes are used or expressed by those cells
- Machinery:
  - Collect and manufacture components
  - Carry out replication
  - Kick-start its new offspring



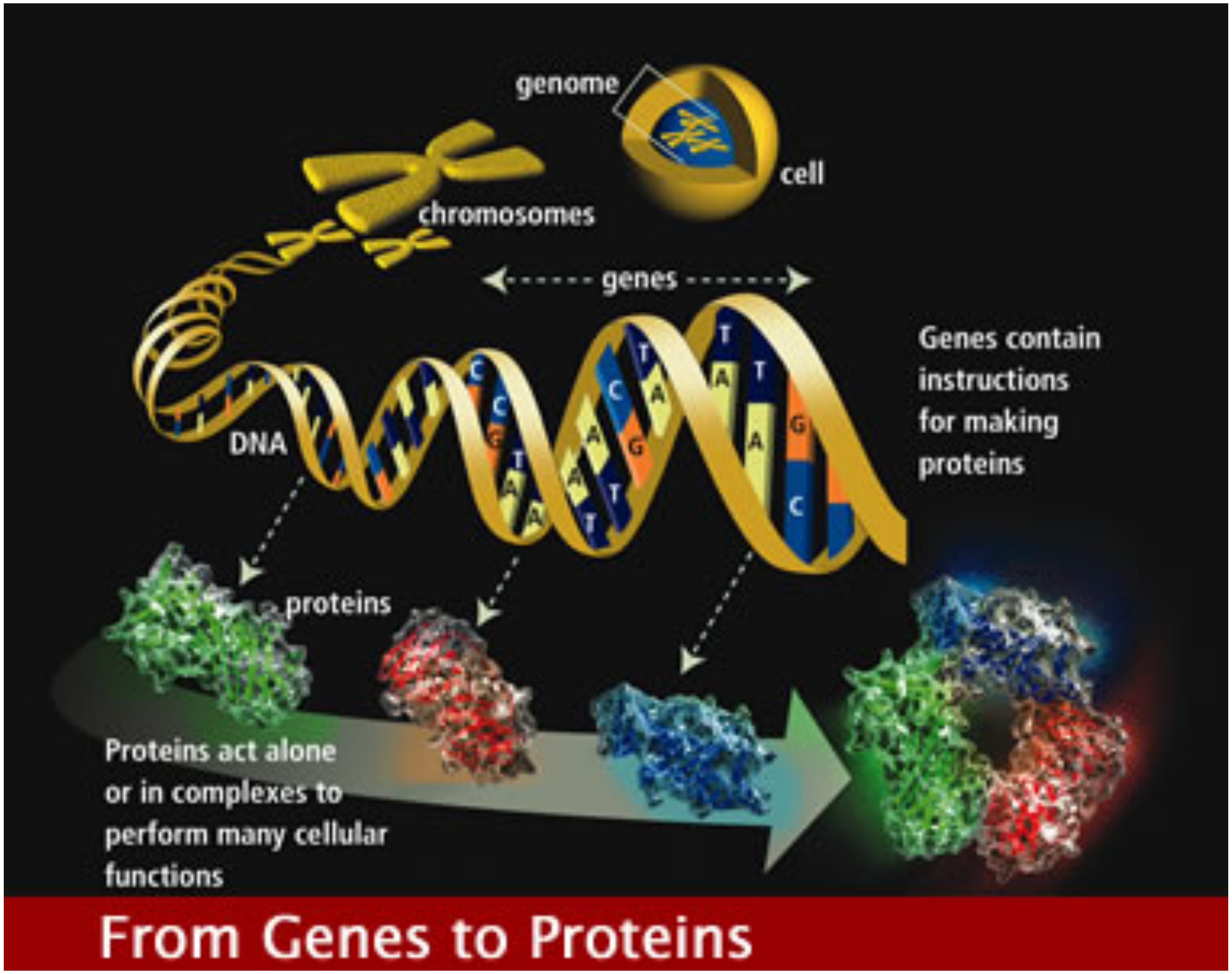




# All life depends on 3 critical molecules

- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body’s major components (e.g. hair, skin, etc.)

# Chromosomes and **genes**



Genes contain instructions for making proteins

Proteins act alone or in complexes to perform many cellular functions

From Genes to Proteins

DNA in the human genome is arranged into 24 distinct **chromosomes**

Each chromosome contains many **genes**, the basic physical and functional units of heredity. **Genes are specific sequences of bases that encode instructions on how to make proteins.**

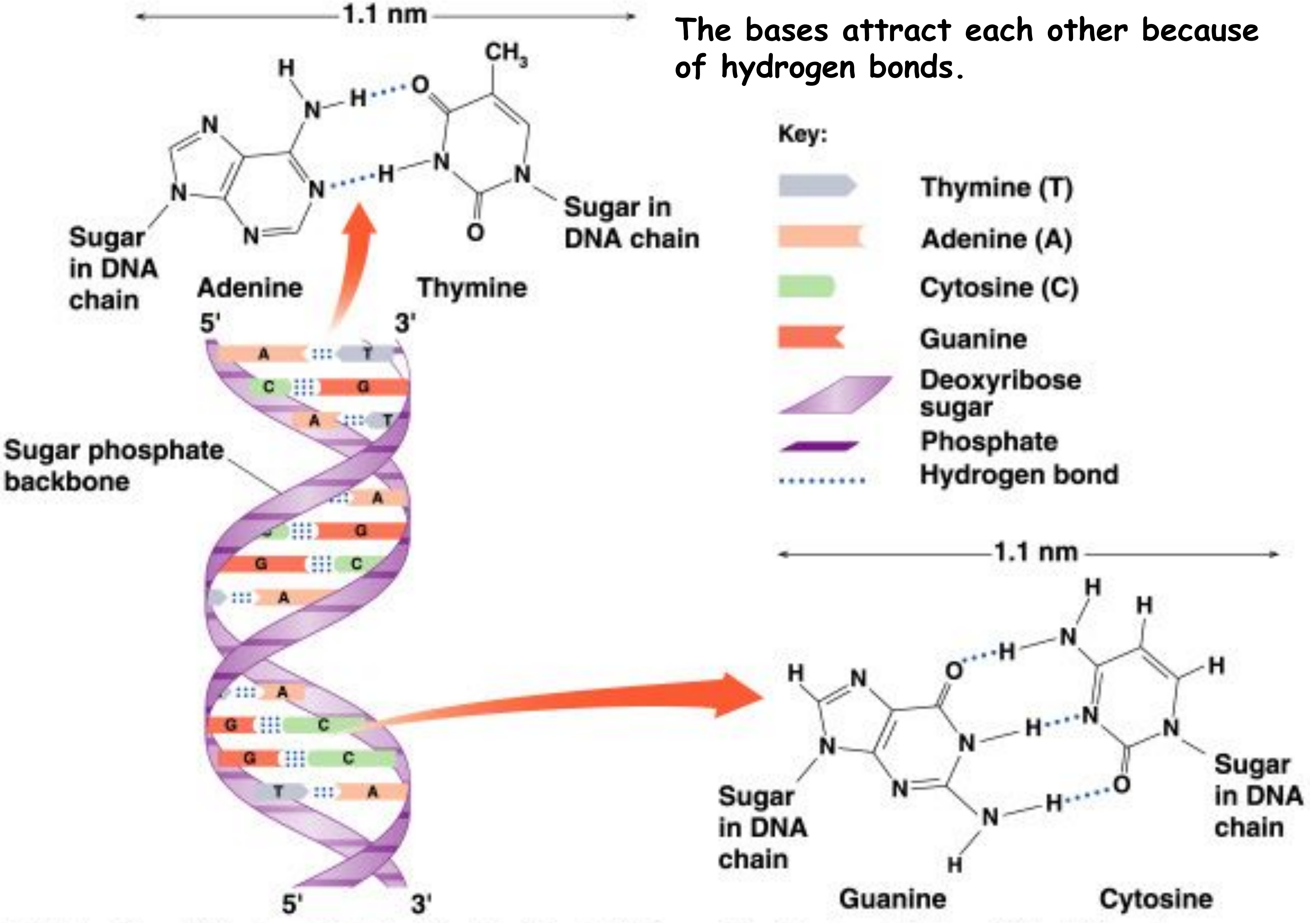
# DNA by the Numbers

- Each cell has about 2 m of DNA.
- The average human has 75 trillion cells.
- The average human has enough DNA to go from the earth to the sun more than 400 times.
- DNA has a diameter of only 0.000000002 m.

The earth is 150 billion m or 93 million miles from the sun.

# Base Pairing in the DNA Double Helix

The bases attract each other because of hydrogen bonds.



# Chemical structure DNA

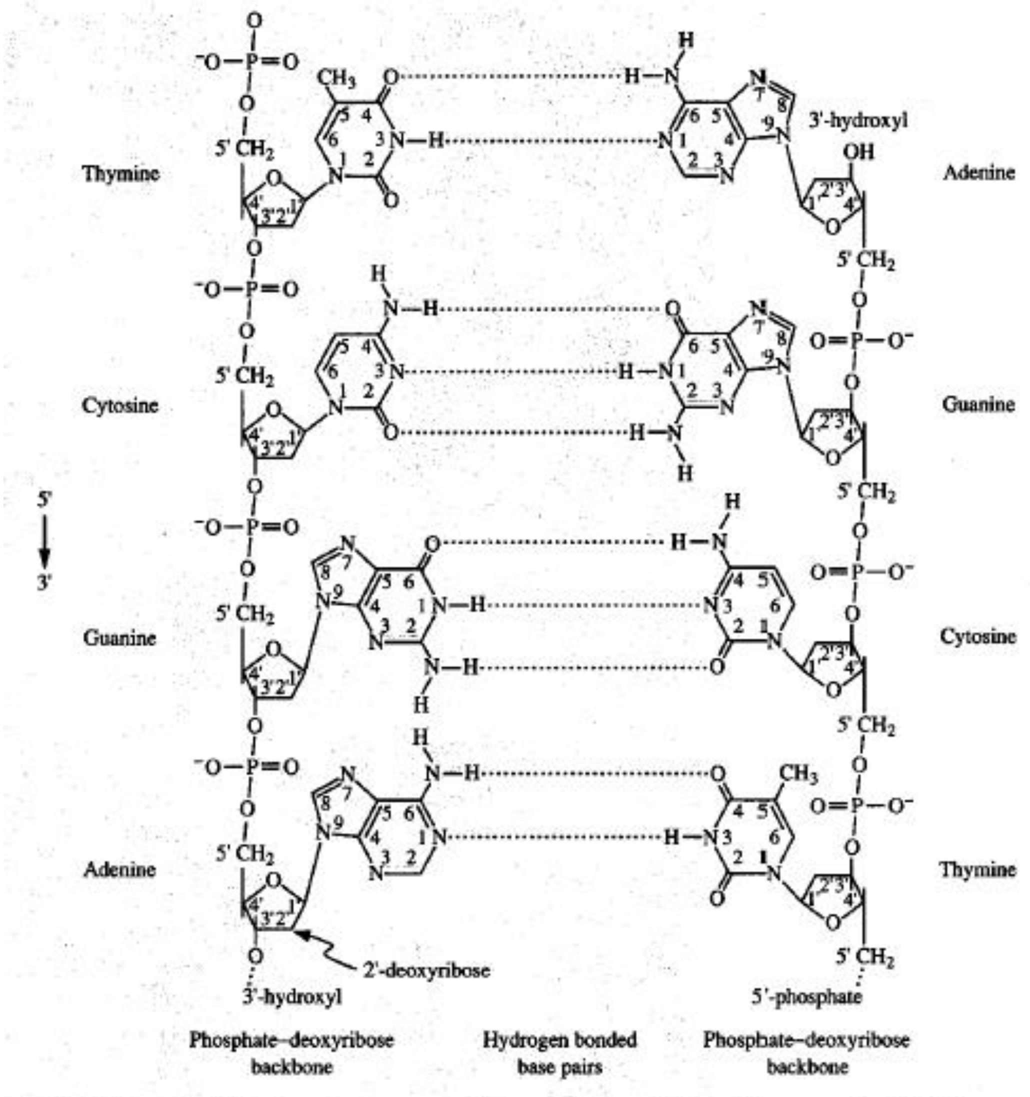
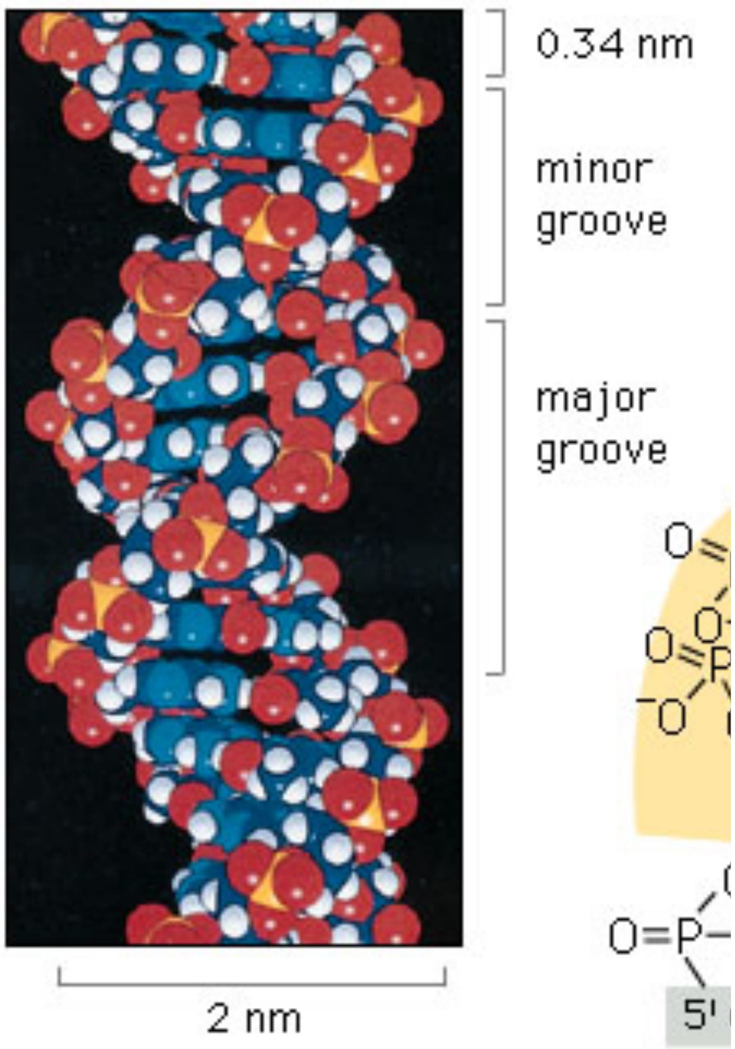
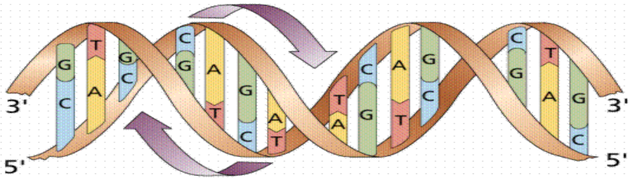


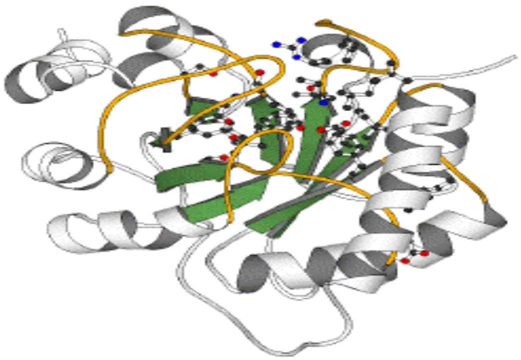
**Fig. 1.2** Chemical structure and base pairing in double-stranded DNA.

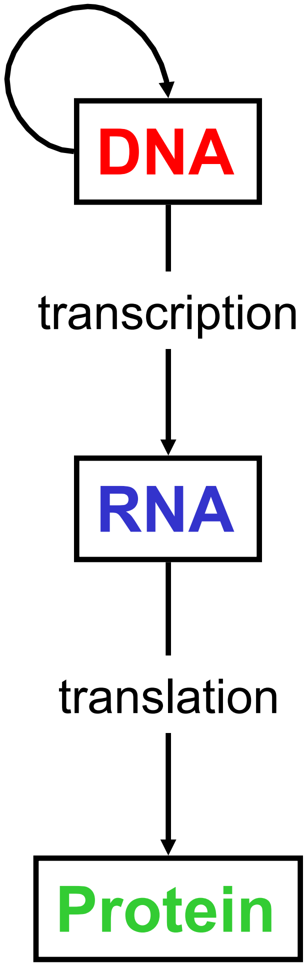


# The Central Dogma of Biology

**Genetic information in genes flows into proteins: DNA → RNA → protein**







CCTGAGCCAACTATTGATGAA

↓

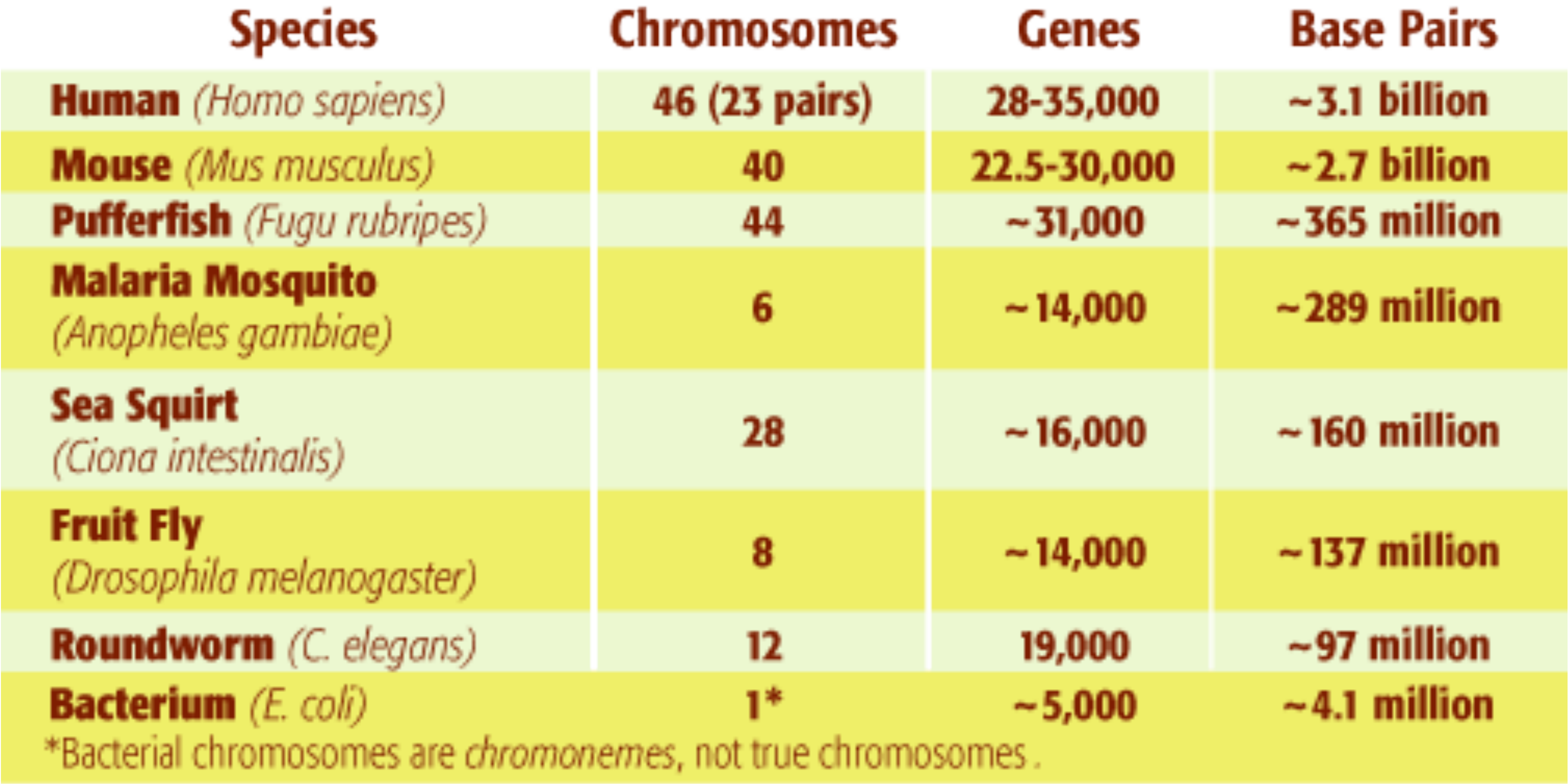
CCUGAGCCAACUAUUGAUGAA

↓

PEPTIDE

It was first stated by Francis Crick in 1958 and re-stated in a Nature paper published in 1970

# Genome sizes

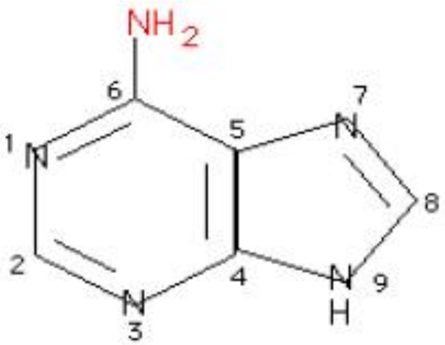


| Species | Chromosomes | Genes | Base Pairs |
|---|---|---|---|
| **Human** *(Homo sapiens)* | 46 (23 pairs) | 28-35,000 | ~3.1 billion |
| **Mouse** *(Mus musculus)* | 40 | 22.5-30,000 | ~2.7 billion |
| **Pufferfish** *(Fugu rubripes)* | 44 | ~31,000 | ~365 million |
| **Malaria Mosquito** *(Anopheles gambiae)* | 6 | ~14,000 | ~289 million |
| **Sea Squirt** *(Ciona intestinalis)* | 28 | ~16,000 | ~160 million |
| **Fruit Fly** *(Drosophila melanogaster)* | 8 | ~14,000 | ~137 million |
| **Roundworm** *(C. elegans)* | 12 | 19,000 | ~97 million |
| **Bacterium** *(E. coli)* | 1* | ~5,000 | ~4.1 million |

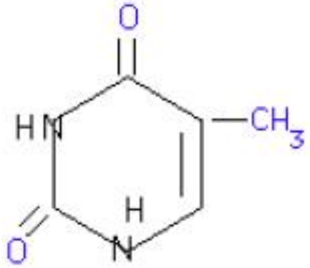
*Bacterial chromosomes are *chromonemes*, not true chromosomes .

# Nitrogenous bases commonly found in RNA and DNA

PURINES PYRIMIDINES ***RNA (AU GC)***

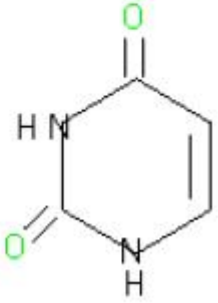


Adenine



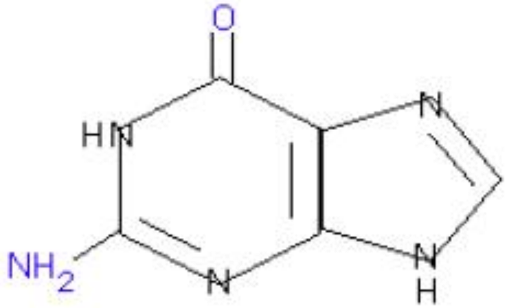
Thymine

**T ----→ U**

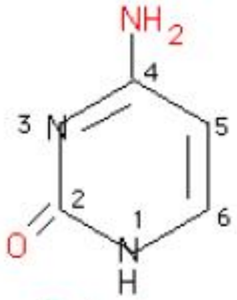


Uracil

***DNA (AT GC)***



Guanine



Cytosine

*A-T (A-U) G=C*

***Complementary pairs***

# Hierarchical organization of RNA molecules

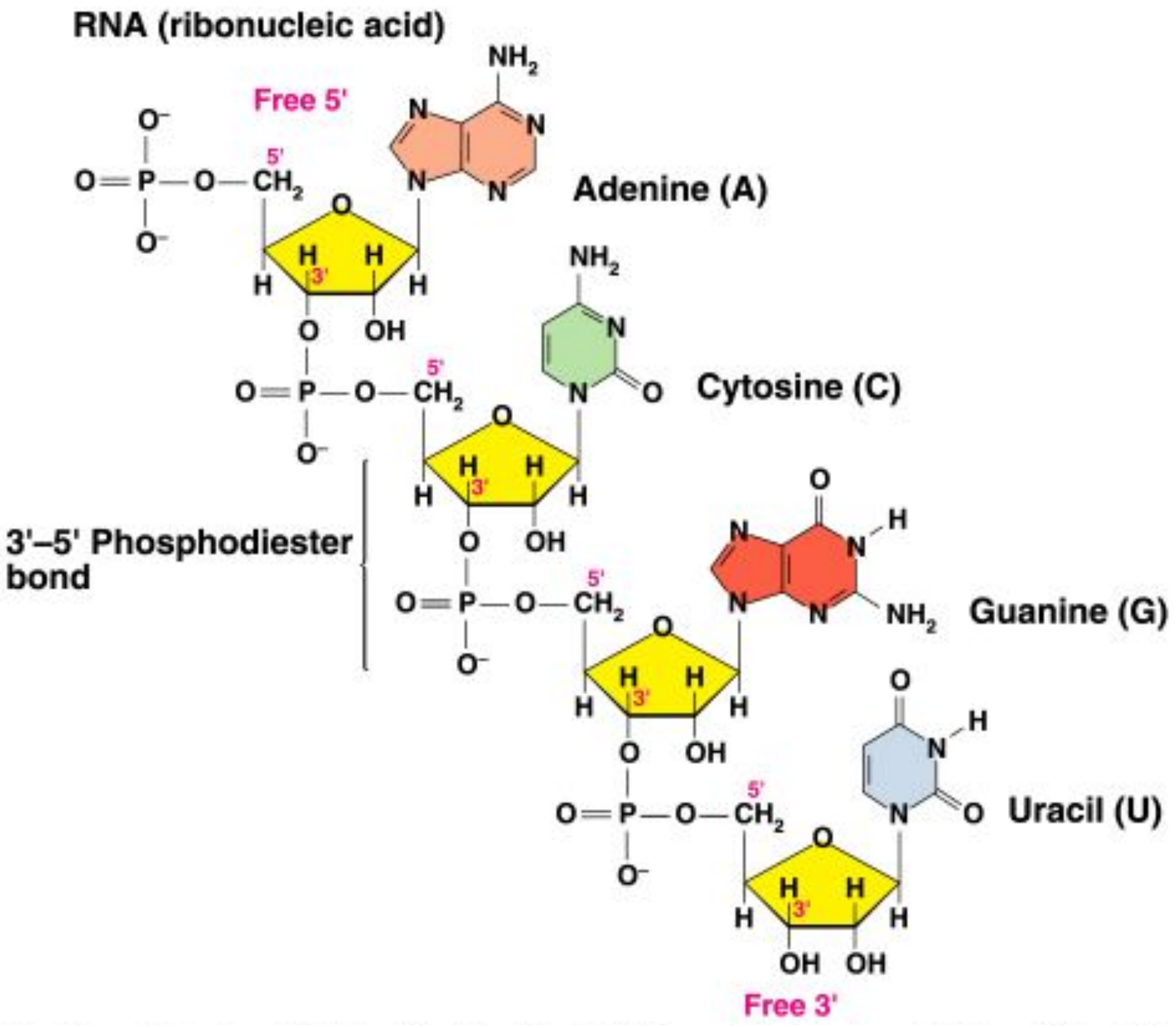
***Primary structure:***

- 5’ to 3’ list of covalently linked nucleotides, named by the attached base

- Commonly represented by a string S over the alphabet Σ={A,C,G,U}

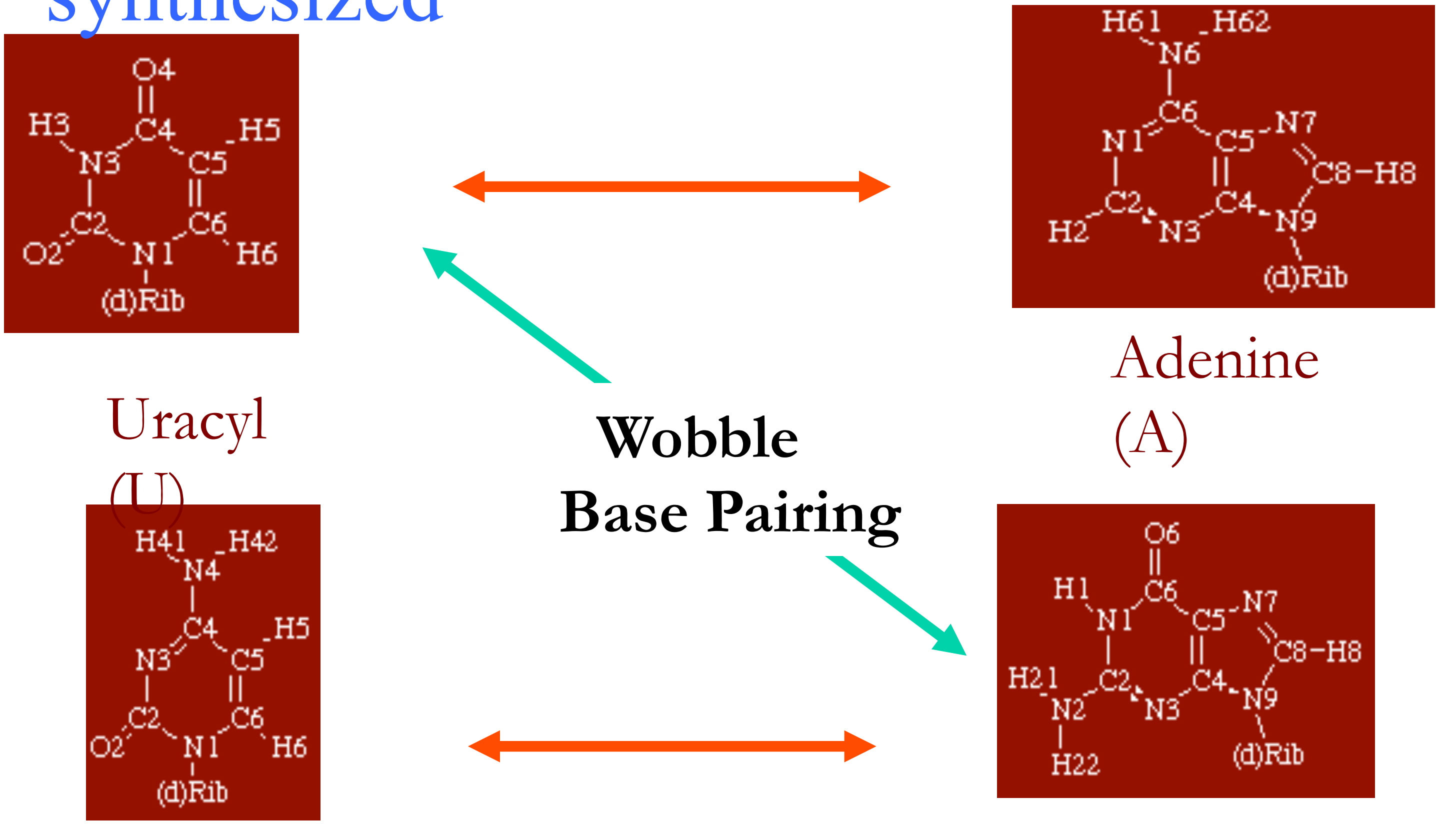
# Example of RNA Primary Structure

- In RNA, A, C, G, and U are linked by 3’-5’ ester bonds between ribose and phosphate



# RNA synthesis and fold

- RNA immediately starts to fold when it is synthesized



Uracyl (U)

Adenine (A)

**Wobble Base Pairing**

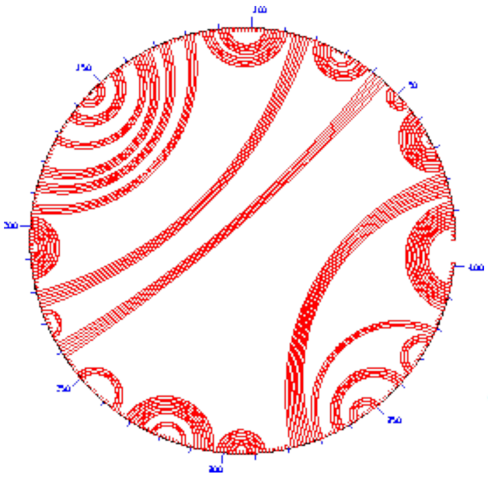
Cytosine (C)

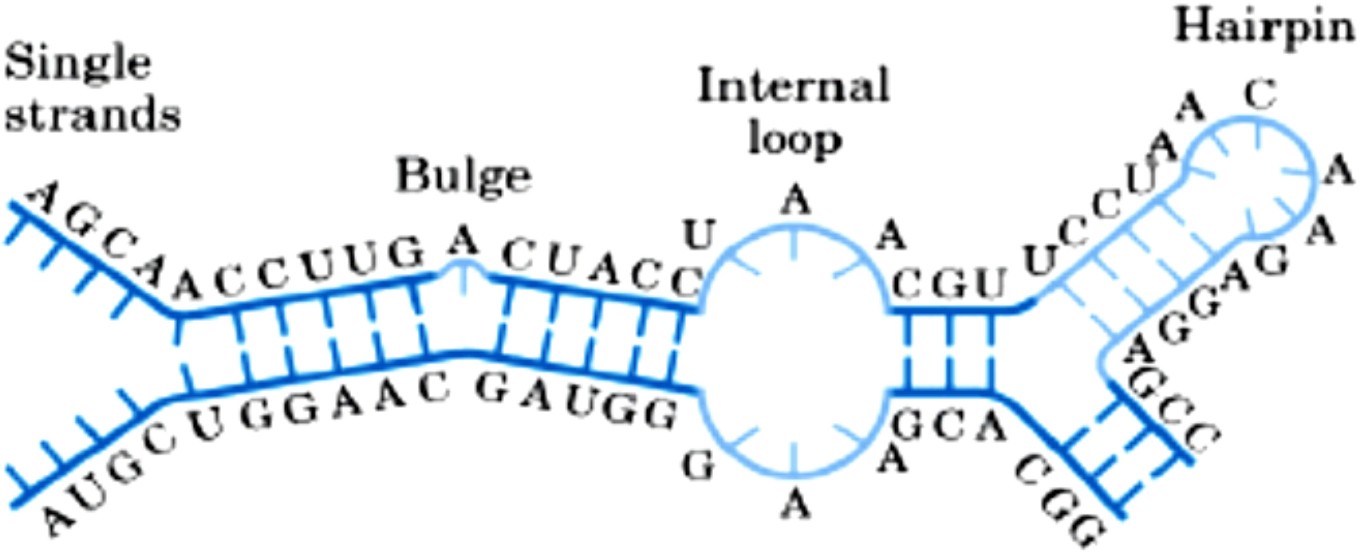
Guanine (G)

# RNA secondary structures

Single stranded bases within a stem are called a bulge of bulge loop if the single stranded bases are on only one side of the stem.

If single stranded bases interrupt both sides of a stem, they are called an internal (interior) loop.

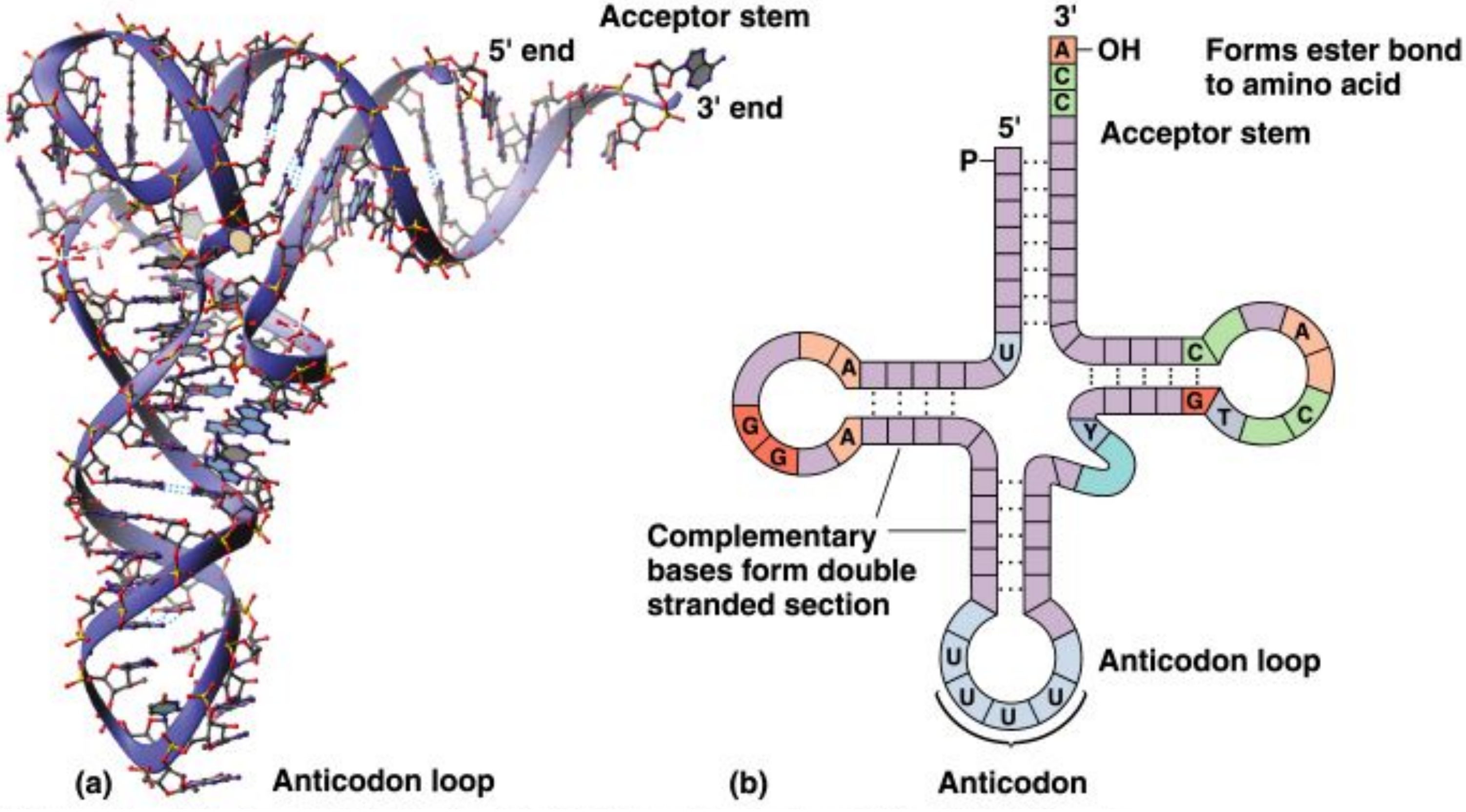




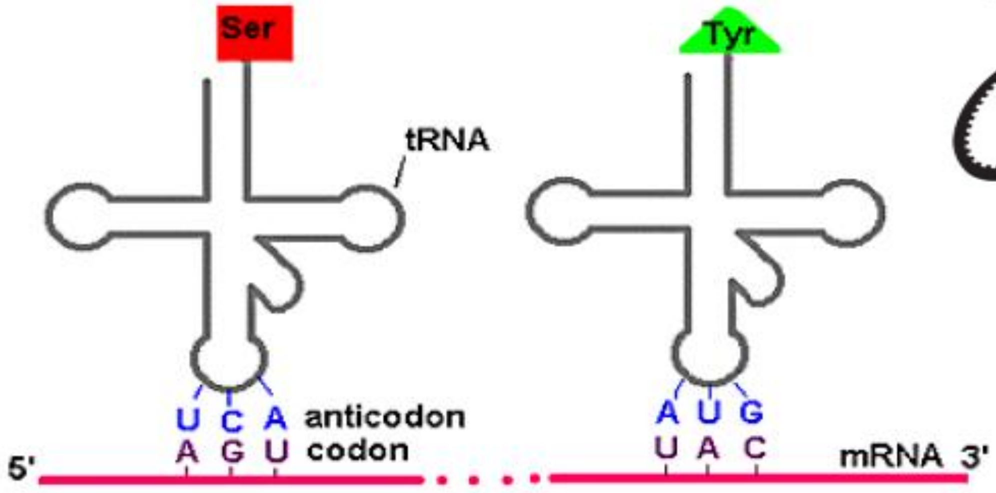
# Transfer RNA

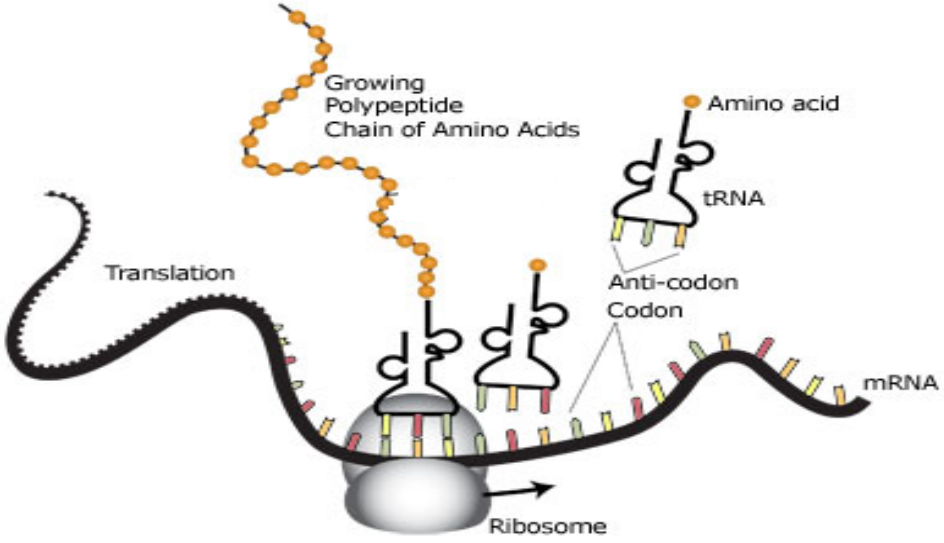
- tRNA has a tertiary structure that is L-shaped

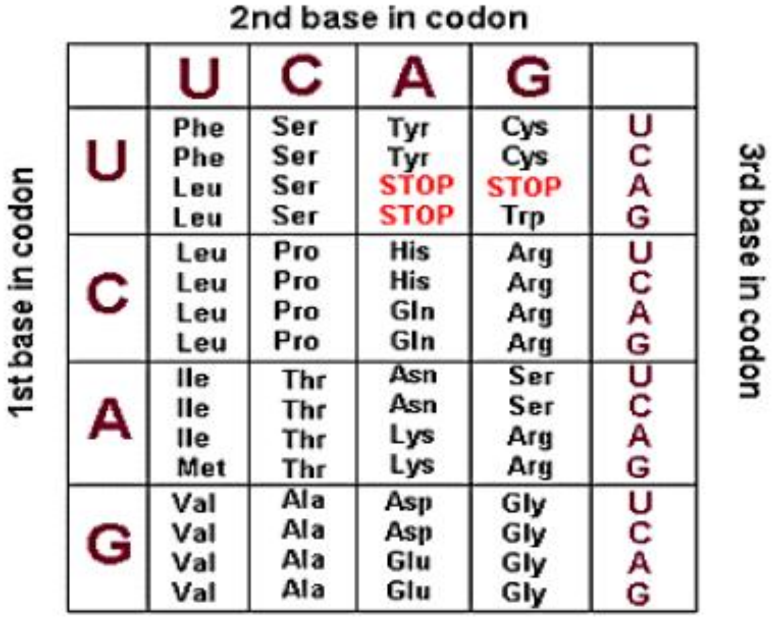
  - one end attaches to the amino acid and the other binds to the mRNA by a 3-base complimentary sequence



# Genetic code



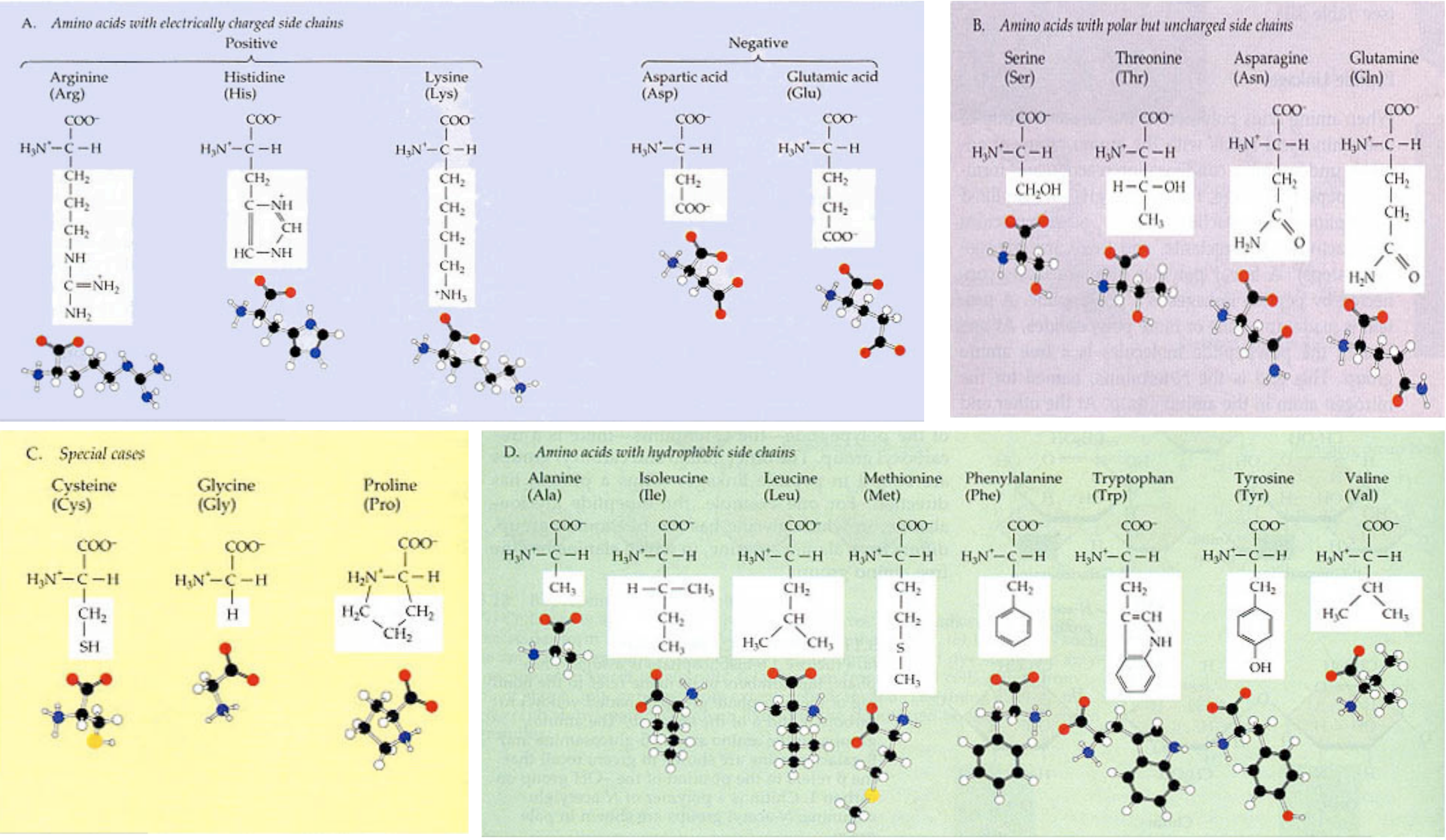




2nd base in codon

| 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 |

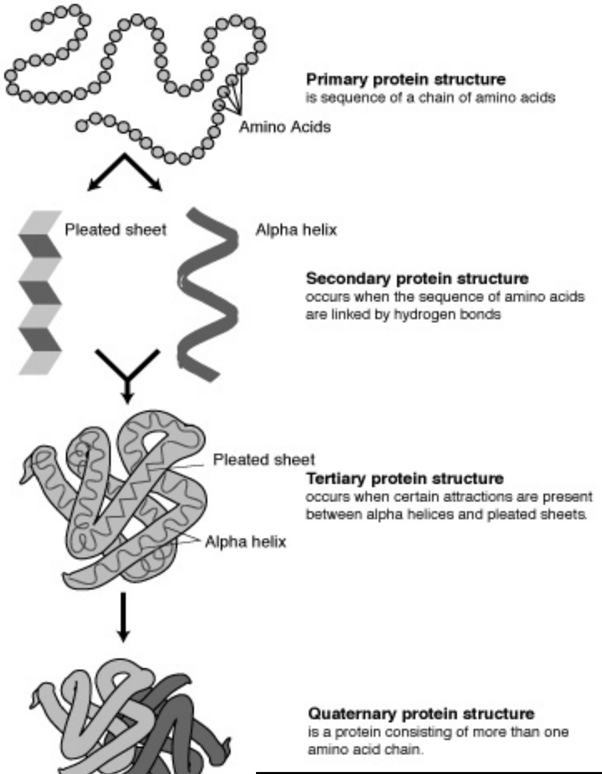
# **Amino acids** - The protein building blocks



C G P

# Protein Folding

- The structure that a protein adopts is vital to its chemistry
- Its structure determines which of its amino acids are exposed to carry out the protein's function
- Its structure also determines what substrates it can react with



**Primary protein structure**
is sequence of a chain of amino acids

**Secondary protein structure**
occurs when the sequence of amino acids are linked by hydrogen bonds

**Tertiary protein structure**
occurs when certain attractions are present between alpha helices and pleated sheets.

**Quaternary protein structure**
is a protein consisting of more than one amino acid chain.

# How do we commonly represent DNA sequences?

- ***Both strands depicted*** *with bases only*
- **5′ ATCTTTGGCTCAGTCTAGTGCACCCAGTT 3′**
- **3′ TAGAAACCGAGTCAGATCACGAGGGTCAA 5′**

- ***The coding strand, 5' to 3'.*** *The coding strand is the strand whose sequence is the same as the corresponding mRNA sequence*

```
DNA   ATCTTTGGCTCAGTCTAGTGCACCCAGTT
mRNA  AUCUUUGGCUCAGUCUAGUGCACCCAGUU
```

- Protein: **F G S V**

# *Molecular Bioinformatics*

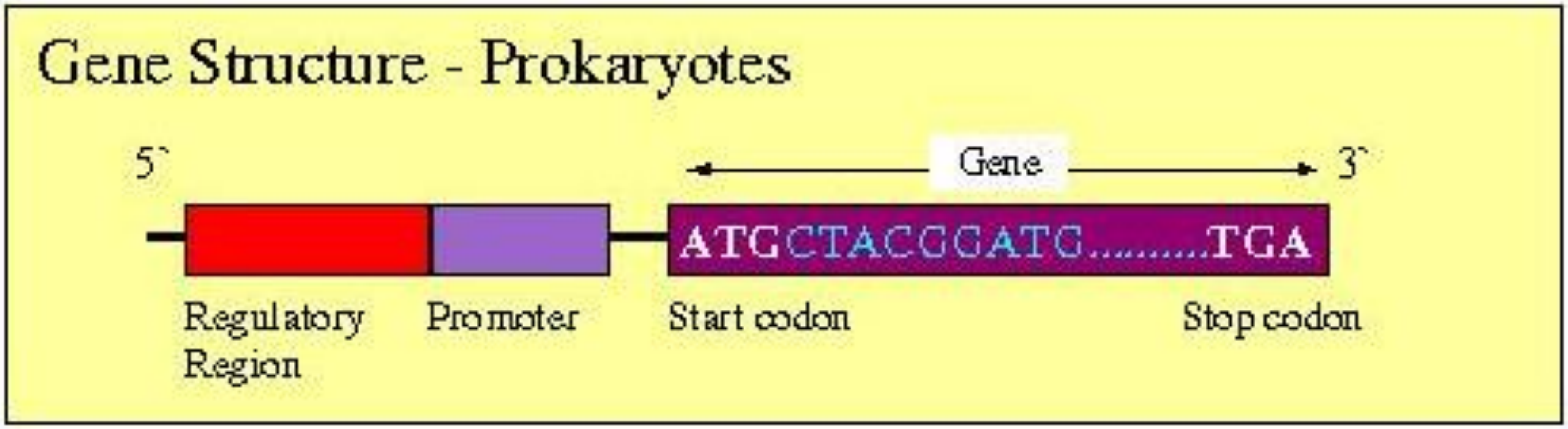
**Molecular Bioinformatics** involves the use of computational tools to discover new information in complex data sets (from the one-dimensional information of DNA through the two-dimensional information of RNA and the three-dimensional information of proteins, to the four-dimensional information of evolving living systems).

# Examples of some important Problems from the Biological side

- Protein folding
- Find Homologies (Similarities)
- Finding genes in new genomes
- Phylogenetic Trees
- Analysis of Gene Expression data
- Prediction of special (regulatory) sites in DNA
- Determine Pathways/gene interaction networks
- Databases/Data mining
- Stochastic Modelling / Simulation of biosystems

# Find genes in DNA sequence

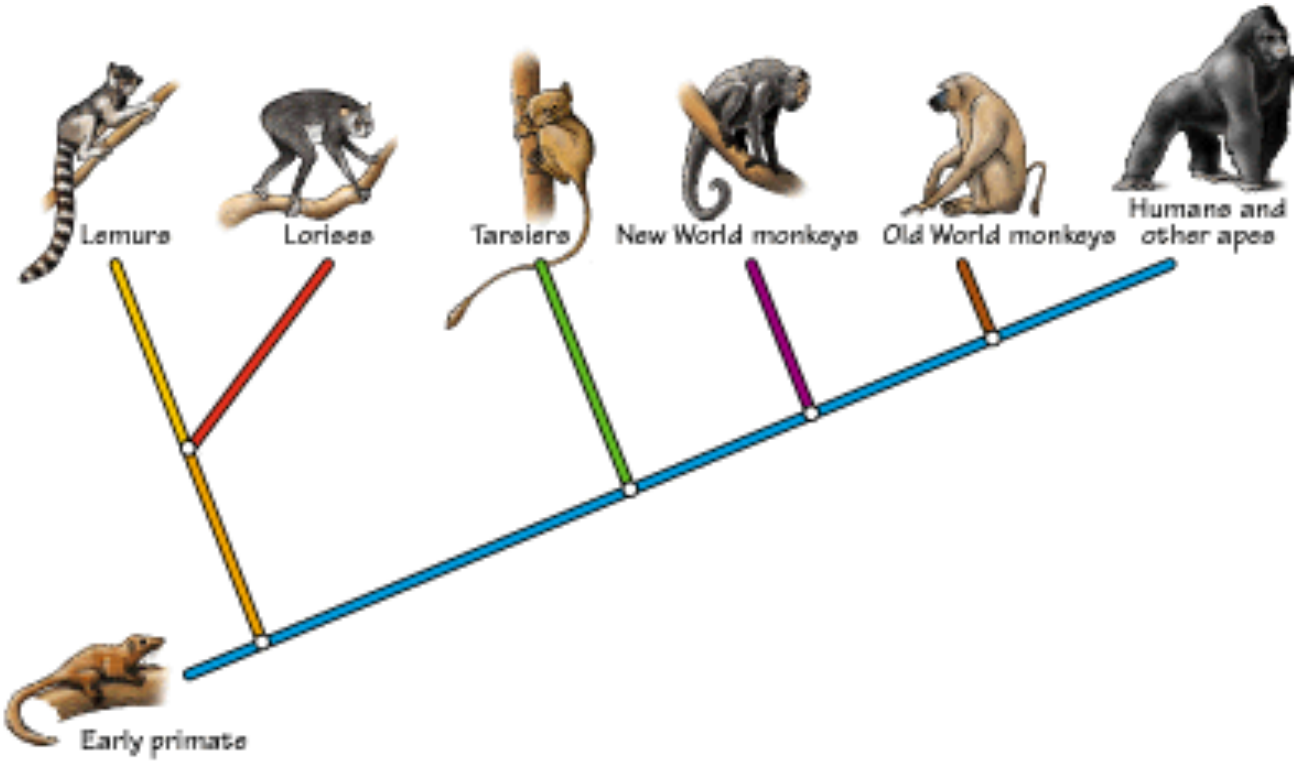
```
GAATTCTAATCTCCCTCTCAACCCTACAGTCACCCATTTGGTATATTAAAGATGTGTTGTCTACTGTCTAGTATCCCTCA
AGTAGTGTCAGGAATTAGTCATTTAAATAGTCTGCAAGCCAGGAGTGGTGGCTCATGTCTGTAATTCCAGCACTGGAGAG
GTAGAAGTGGGAGGACTGCTTGAGCTCAAGAGTTTGATATTATCCTGGACAACATAGCAAGACCTCGTCTCTACTTAAAA
AAAAAAAAATTAGCCAGGCATGTGATGTACACCTGTAGTCCCAGCTACTCAGGAGGCCGAAATGGGAGGATCCCTTGAGC
TCAGGAGGTCAAGGCTGCAGTGAGACATGATCTTGCCACTGCACTCCAGCCTGGACAGCAGAGTGAAACCTTGCCTCACG
AAACAGAATACAAAAACAAACAAACAAAAAACTGCTCCGCAATGCGCTTCCTTGATGCTCTACCACATAGGTCTGGGTAC
TTTGTACACATTATCTCATTGCTGTTCGTAATTGTTAGATTAATTTTGTAATATTGATATTATTCCTAGAAAGCTGAGGC
CTCAAGATGATAACTTTTATTTTCTGGACTTGTAATAGCTTTCTCTTGTATTCACCATGTTGTAACTTTCTTAGAGTAGT
AACAATATAAAGTTATTGTGAGTTTTTGCAAACACATGCAAACACAACGACCCATATAGACATTGATGTGAAATTGTCTAT
TGTCAATTTATGGGAAAACAAGTATGTACTTTTTCTACTAAGCCATTGAAACAGGAATAACAGAACAAGATTGAAAGAAT
ACATTTTCCGAAATTACTTGAGTATTATACAAAGACAAGCACGTGGACCTGGGAGGAGGGTTATTGTCCATGACTGGTGT
GTGGAGACAAATGCAGGTTTATAATAGATGGGATGGCATCTAGCGCAATGACTTTGCCATCACTTTTAGAGAGCTCTTGG
GGACCCCAGTACACAAGAGGGGACGCAGGGTATATGTAGACATCTCATTCTTTTTCTTAGTGTGAGAATAAGAATAGCCA
TGACCTGAGTTTATAGACAATGAGCCCTTTTCTCTCTCCCACTCAGCAGCTATGAGATGGCTTGCCCTGCCTCTCTACTA
GGCTGACTCACTCCAAGGCCCAGCAATGGGCAGGGCTCTGTCAGGGCTTTGATAGCACTATCTGCAGAGCCAGGGCCGAG
AAGGGGTGGACTCCAGAGACTCTCCCTCCCATTCCCGAGCAGGGTTTGCTTATTTATGCATTTAAATGATATATTTATTT
TAAAAGAAATAACAGGAGACTGCCCAGCCCTGGCTGTGACATGGAAACTATGTAGAATATTTTGGGTTCCATTTTTTTTT
CCTTCTTTCAGTTAGAGGAAAAGGGGCTCACTGCACATACACTAGACAGAAAGTCAGGAGCTTTGAATCCAAGCCTGATC
```



# Phylogenetic Trees

How did our genome evolve?  How close are we related to other  species?

## Primate evolution



# Morphological vs. Molecular

- Classical phylogenetic analysis: **morphological** features
  - number of legs, lengths of legs, etc.

- Modern biological methods allow to use **molecular** features
  - Gene sequences
  - Protein sequences

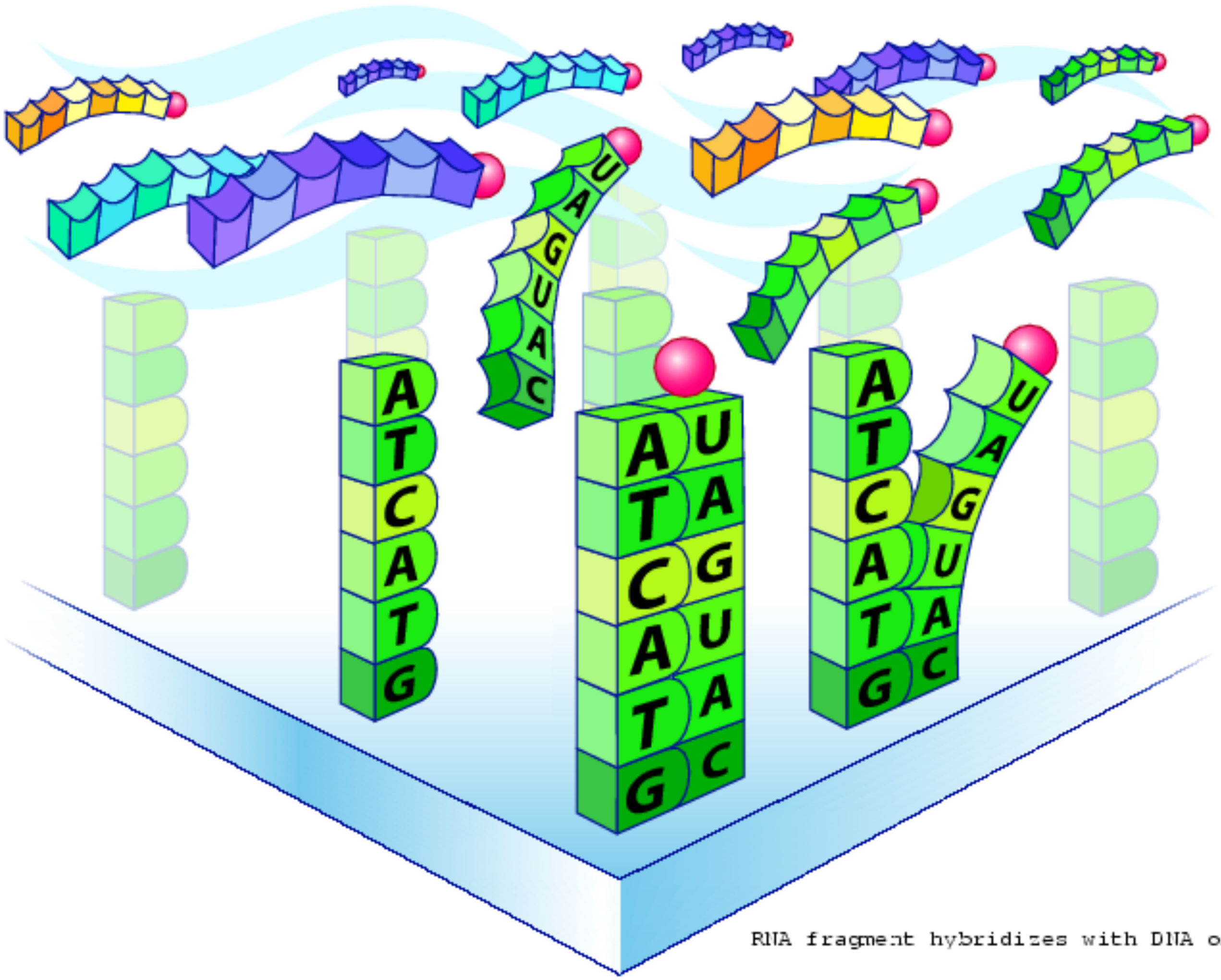
# Gene Expression

**How do genes in one cell work together over time?**

**What is the difference of gene activity between a young and old cell or between healthy and sick cell?**

**What set of genes is activated in cancer cells?**

RNA fragments with fluorescent tags from sample to be tested

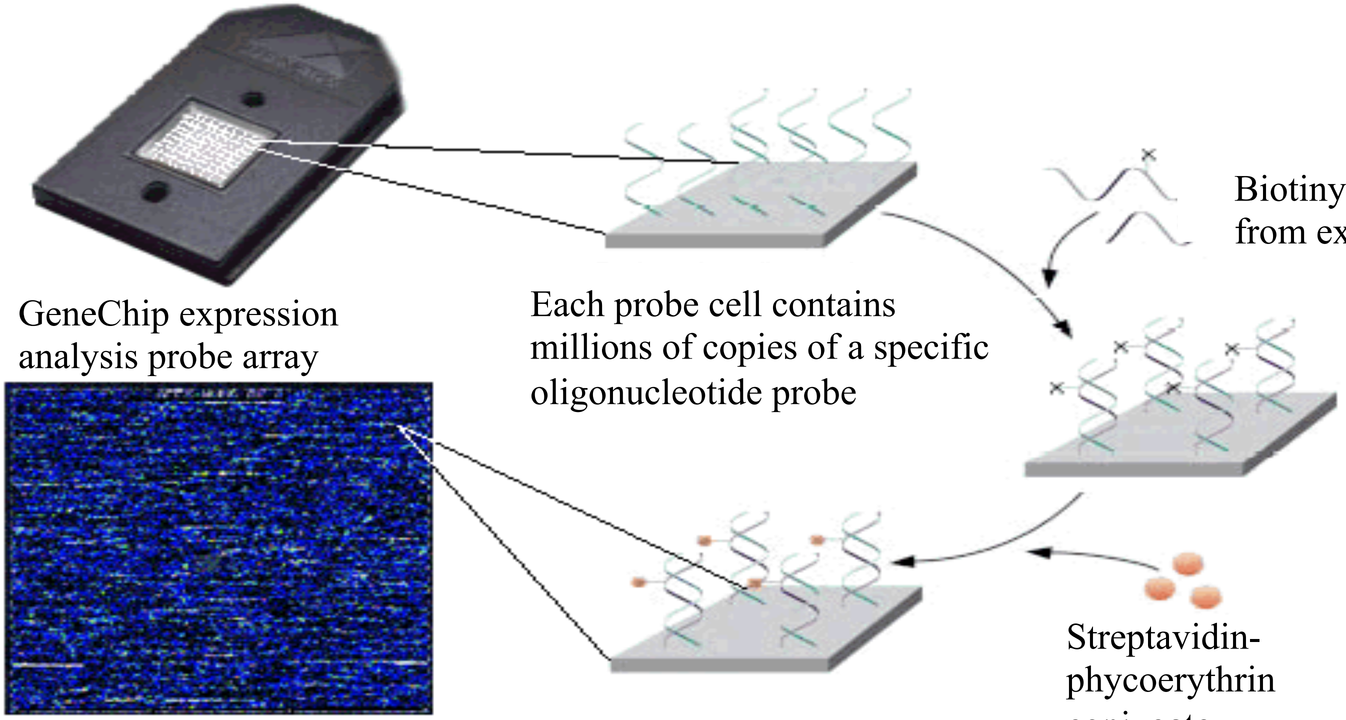


RNA fragment hybridizes with DNA on GeneChip

# GeneChip Expression Analysis

AFFYMETRIX®

## GeneChip® Expression Analysis Process



GeneChip expression analysis probe array

Each probe cell contains millions of copies of a specific oligonucleotide probe

Biotinylated RNA from experiment

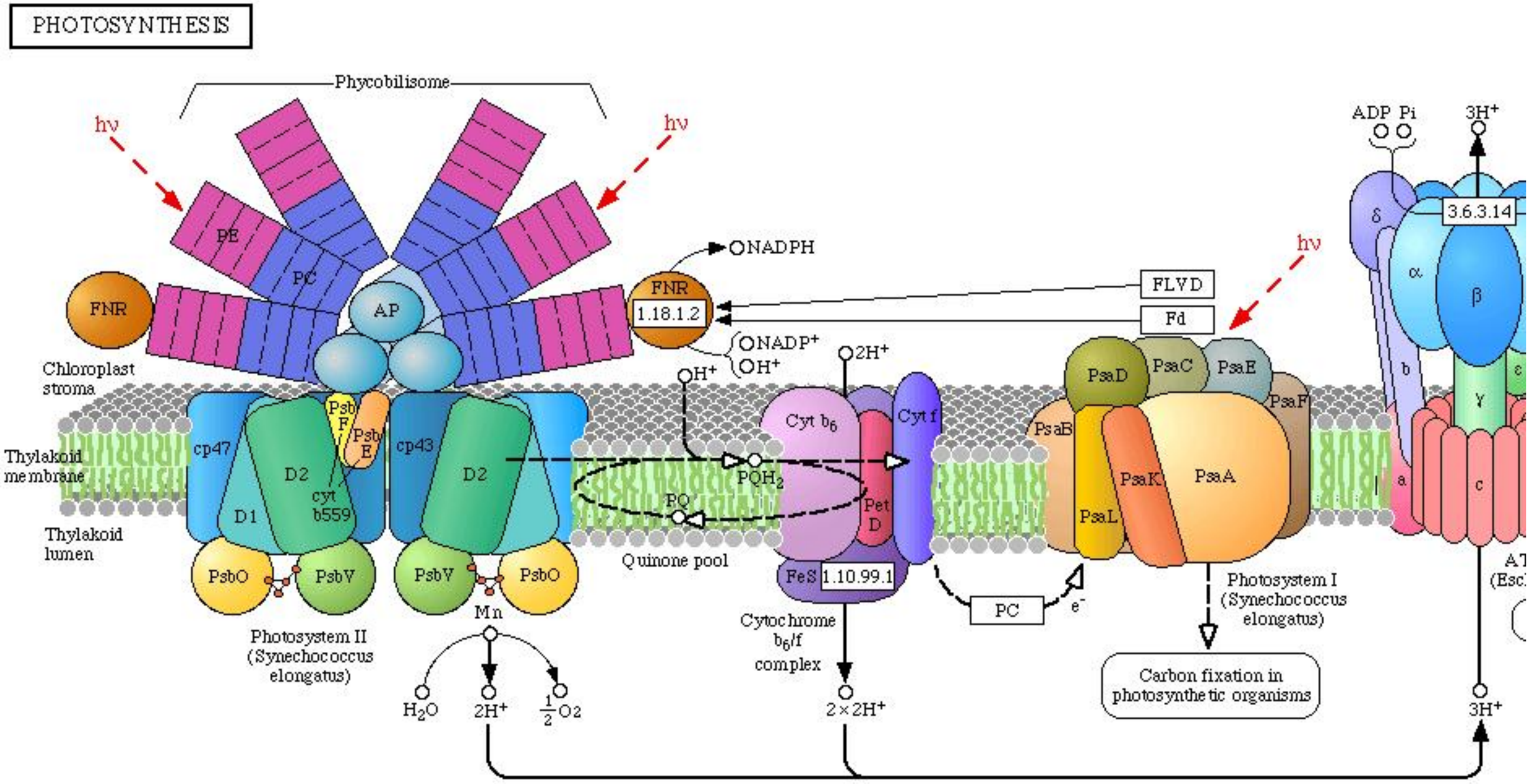
Streptavidin-phycoerythrin conjugate

Image of hybridized probe array

# Determine Pathways

Which genes work together? Which genes are active at which times in which situations in which cells?

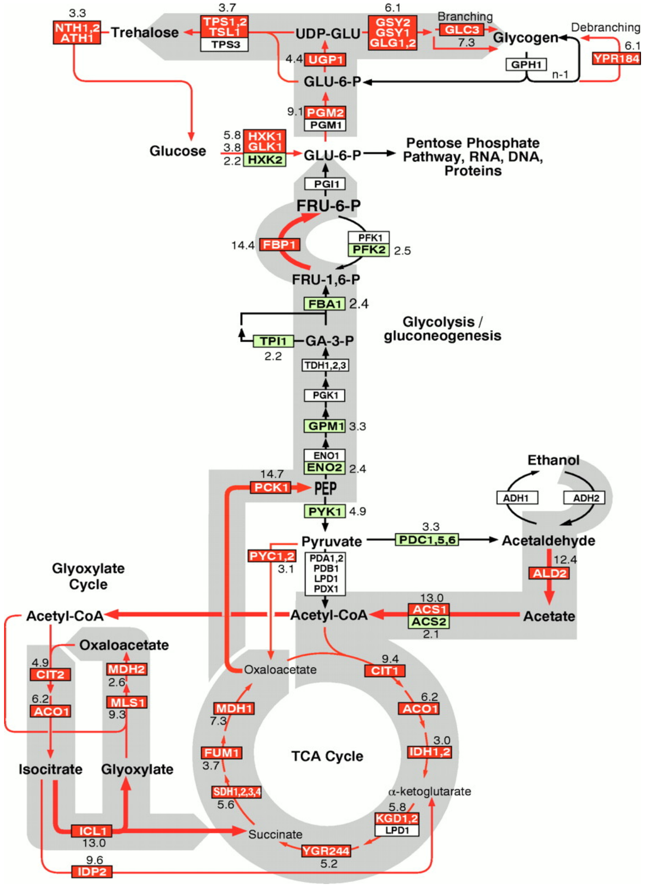
How are the functions of different proteins interconnected?



# Information Derivable from Chip Data

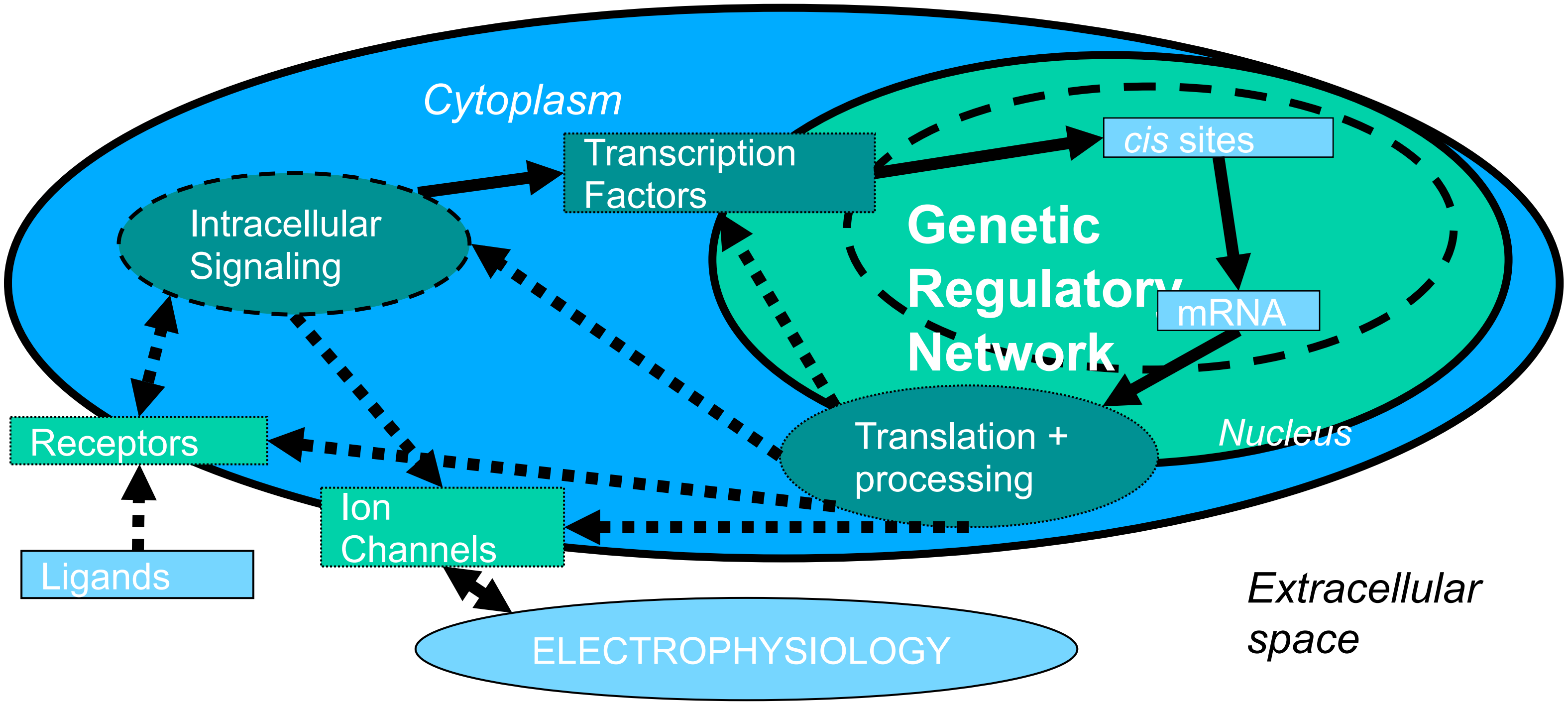
- Microarray data is becoming a key source of data for computational inference of biological networks
  - who interact with who
  - who regulate who
  - ….

How does this work?

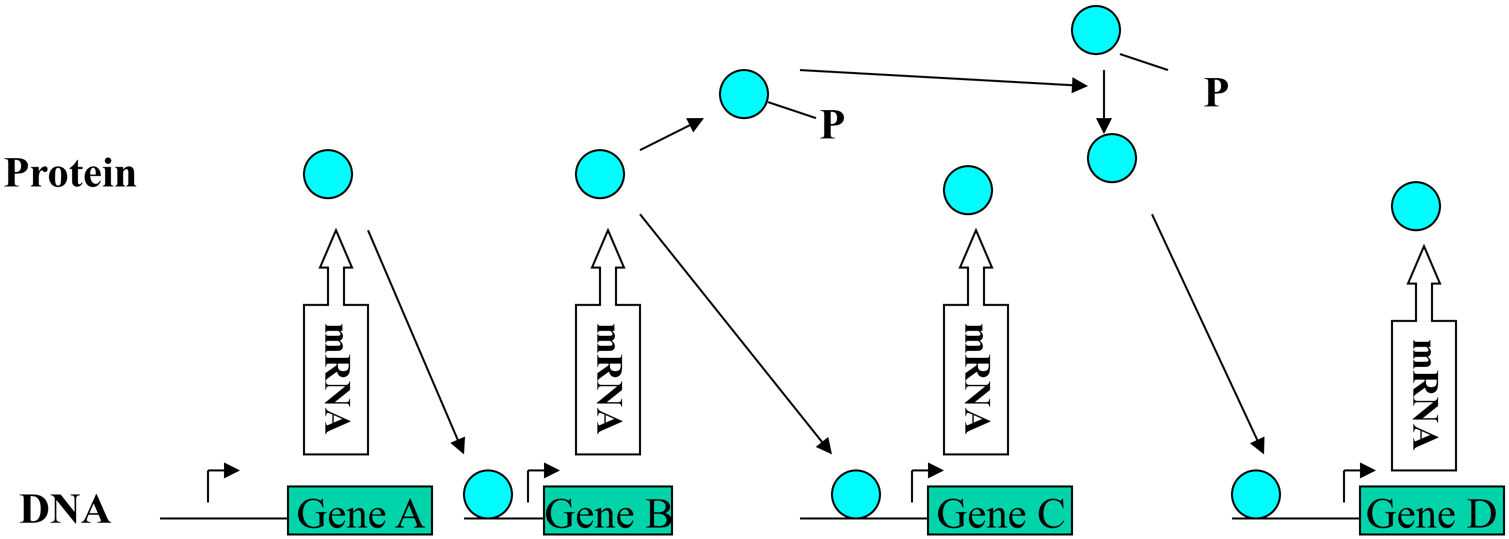


# Genetic Regulatory Network

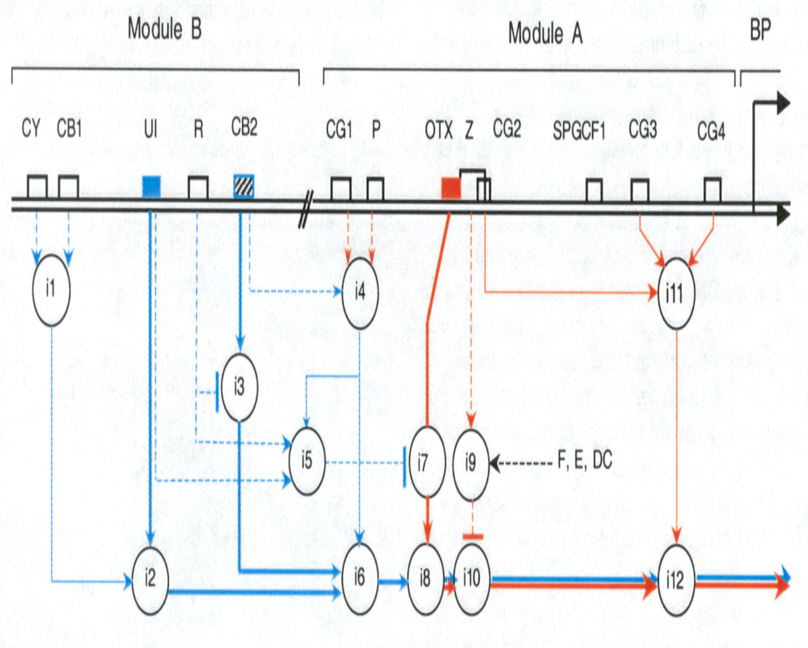
the set of mutually activating and repressing genes and gene products and their interactions

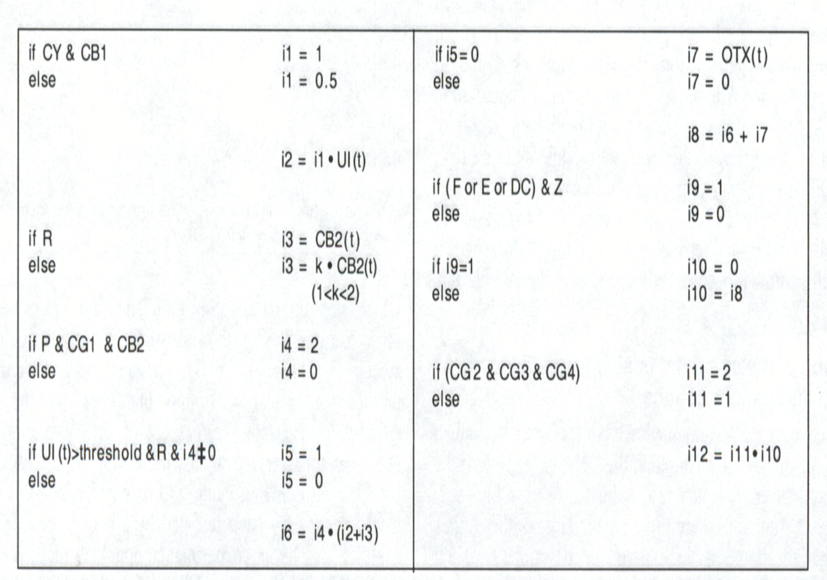


# Microarray analysis model using gene expression profiles



# Gene Regulatory Systems





| | | | |
|---|---|---|---|
| if CY & CB1 | i1 = 1 | if i5=0 | i7 = OTX(t) |
| else | i1 = 0.5 | else | i7 = 0 |
| | | | i8 = i6 + i7 |
| | i2 = i1 • UI(t) | | |
| | | if ( F or E or DC) & Z | i9 = 1 |
| | | else | i9 = 0 |
| if R | i3 = CB2(t) | | |
| else | i3 = k • CB2(t) | if i9=1 | i10 = 0 |
| | (1<k<2) | else | i10 = i8 |
| if P & CG1 & CB2 | i4 = 2 | | |
| else | i4 = 0 | if (CG2 & CG3 & CG4) | i11 = 2 |
| | | else | i11 =1 |
| if UI (t)>threshold &R & i4‡0 | i5 = 1 | | i12 = i11•i10 |
| else | i5 = 0 | | |
| | i6 = i4 • (i2+i3) | | |

"Programs built into the DNA of every animal."

Eric H. Davidson

# mRNA Expression Data Format

## From cDNA microarray

| | Intensity (treated) | Intensity (wild type) | Ratio |
|---|---|---|---|
| Gene A | 0.22 | 0.24 | 0.917 |
| Gene B | 0.67 | 1.21 | 0.598 |
| Gene C | 1.13 | 0.43 | 2.630 |
| Gene D | 2.45 | 2.44 | 1.01 |

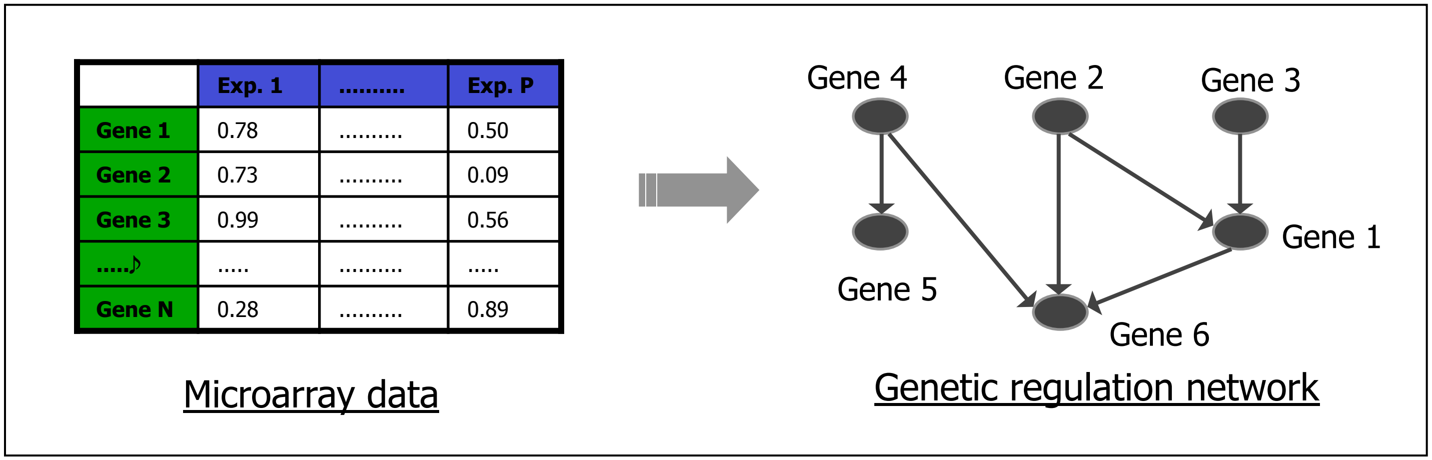
0 < ratio < Inf.

-Inf. < $\log_2$(ratio) < + Inf.
where
$\log_2$(ratio) > 0: increase
$\log_2$(ratio) < 0: decrease

## E X P matrix

| | Exp. 1 | ...... | Exp. P |
|---|---|---|---|
| Gene 1 | 0.78 | ...... | 0.50 |
| Gene 2 | 0.73 | ...... | 0.09 |
| Gene 3 | 0.99 | ...... | 0.56 |
| Gene 4 | 0.60 | ...... | 0.41 |
| Gene 5 | 0.44 | ...... | 0.86 |
| Gene 6 | 0.07 | ...... | 0.05 |
| Gene 7 | 0.28 | ...... | 0.89 |
| Gene 8 | 0.91 | ...... | 0.00 |
| ..... | ..... | ...... | ..... |
| Gene N | 0.28 | ...... | 0.89 |

# Problem Definition



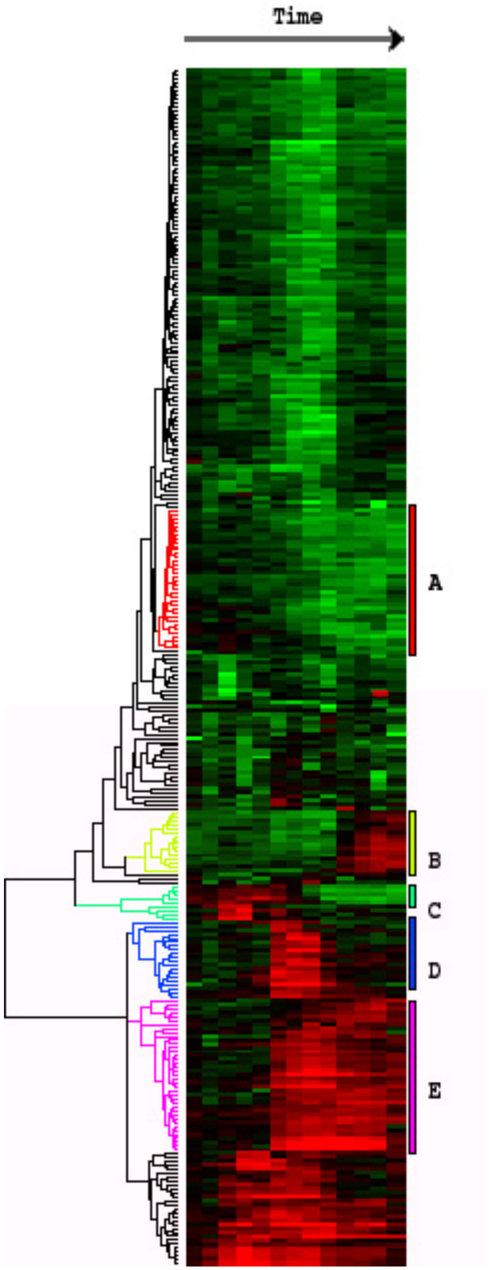
| | Exp. 1 | .......... | Exp. P |
|---|---|---|---|
| Gene 1 | 0.78 | .......... | 0.50 |
| Gene 2 | 0.73 | .......... | 0.09 |
| Gene 3 | 0.99 | .......... | 0.56 |
| .....♪ | ..... | .......... | ..... |
| Gene N | 0.28 | .......... | 0.89 |

Microarray data

Genetic regulation network

## Difficulty in Reconstructing Genetic Regulatory Network

1. mRNA expression is only a partial picture
2. the number of sample is much smaller than the number of genes
3. high noise



# Clustering

Eisen *et al.* (1998):

**FIG. 1.** Cluster display of data from time course of serum stimulation of primary human fibroblasts.

**Expemeriments:**
Foreskin fibroblasts were grown in culture and were deprived of serum for 48 hr. Serum was added back and samples taken at time 0, 15 min, 30 min, 1hr, 2 hr, 3 hr, 4 hr, 8 hr, 12 hr, 16 hr, 20 hr, 24 hr.

**Clustering:**
Correlation Coefficient + Centroid Hierarchical Clustering

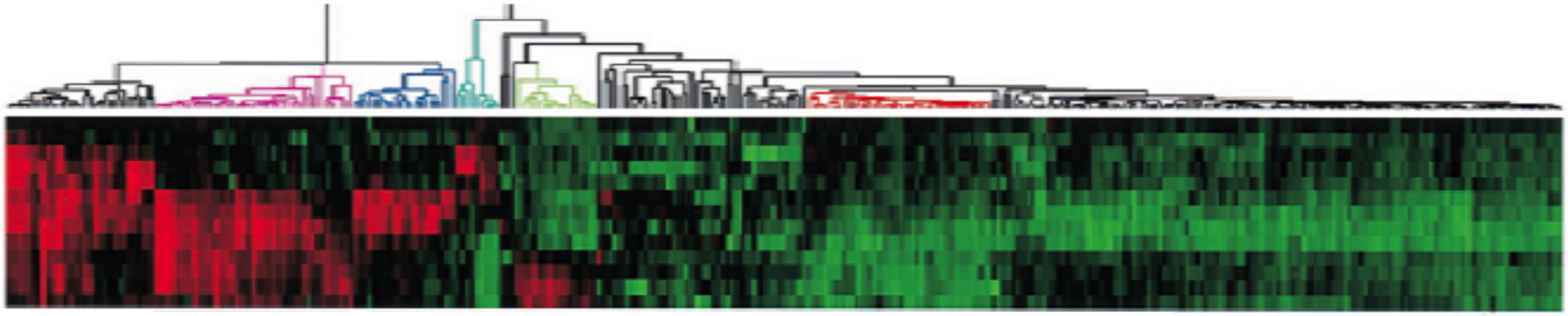
**Clusters:**
(A) cholesterol biosynthesis,
(B) the cell cycle,
(C) the immediate-early response,
(D) signaling and angiogenesis,
(E) wound healing and tissue remodeling.

# Clustering

✓ Grouping genes with similar patterns of expression

Common role gene clustered together

Uncharacterized gene function guessed



Similarity measure : standard correlation coefficient, ..

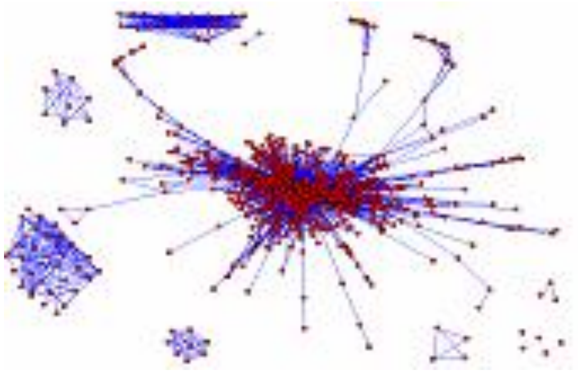
Method : Hierarchical clustering, K-means, SOM ..

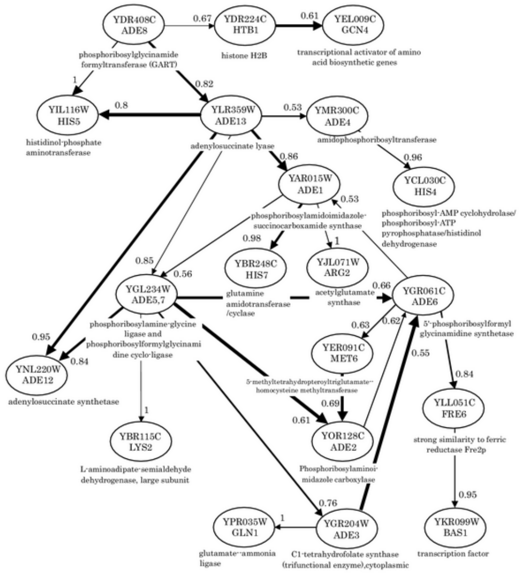
Can’t reveal the inner interaction structure !

# Molecular Networks Constructed from High-throughput assays

## Correlation or co-expression network:

A graphical representation that averages over observed expression data. Nodes are mRNA molecules, edges represent correlations between expression levels of connected nodes.




YDR408C
ADE8
phosphoribosylglycinamide formyltransferase (GART)
0.67
YDR224C
HTB1
histone H2B
0.61
YEL009C
GCN4
transcriptional activator of amino acid biosynthetic genes
1
0.82
YIL116W
HIS5
histidinol-phosphate aminotransferase
0.8
YLR359W
ADE13
adenylosuccinate lyase
0.53
YMR300C
ADE4
amidophosphoribosyltransferase
0.86
YAR015W
ADE1
0.53
phosphoribosylamidoimidazole-succinocarboxamide synthase
0.96
YCL030C
HIS4
phosphoribosyl-AMP cyclohydrolase/ phosphoribosyl-ATP pyrophosphatase/histidinol dehydrogenase
0.98
YBR248C
HIS7
glutamine amidotransferase /cyclase
1
YJL071W
ARG2
acetylglutamate synthase
0.85
0.56
YGL234W
ADE5,7
phosphoribosylamine-glycine ligase and phosphoribosylformylglycinamidine cyclo-ligase
0.66
YGR061C
ADE6
5'-phosphoribosylformyl glycinamidine synthetase
0.63
0.62
YER091C
MET6
5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase
0.55
0.95
0.84
YNL220W
ADE12
adenylosuccinate synthetase
0.84
YLL051C
FRE6
strong similarity to ferric reductase Fre2p
0.69
0.61
YOR128C
ADE2
Phosphoribosylaminoimidazole carboxylase
1
YBR115C
LYS2
L-aminoadipate-semialdehyde dehydrogenase, large subunit
0.76
0.95
1
YPR035W
GLN1
glutamate--ammonia ligase
YGR204W
ADE3
C1-tetrahydrofolate synthase (trifunctional enzyme),cytoplasmic
YKR099W
BAS1
transcription factor

## Bayesian networks:

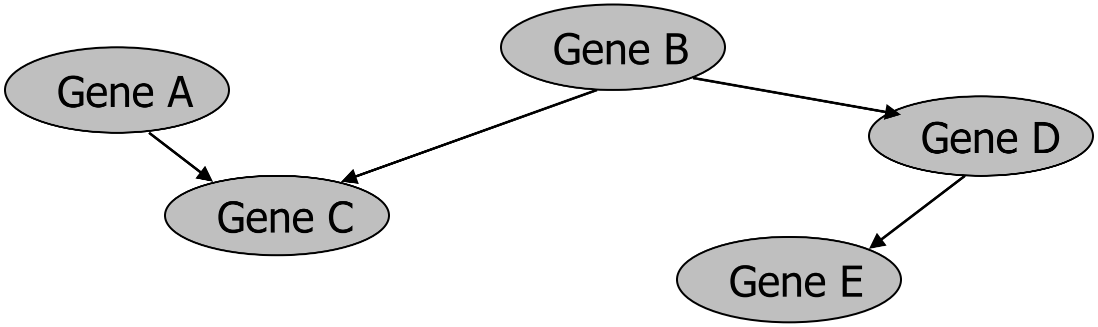
A directed, graphical representation of the probabilities of one observation given another. Nodes represent mRNA molecules; edges represent the probability of a particular expression value given the expression values of the parent nodes.

# Bayesian Network

Probabilistic framework for inference of interactions in the presence of noise

✓ *G*: a directed-acyclic graph structure

✓ Θ: a set of parameters for conditional distribution of each variable



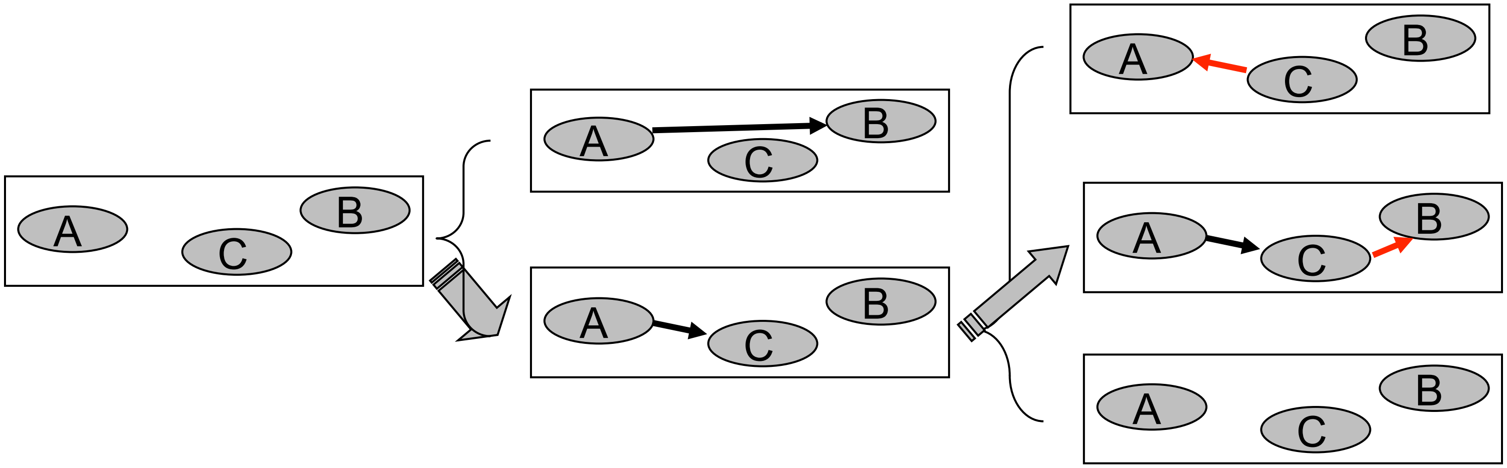
$$P(A, B, C, D, E) = \prod P(X_i \mid \text{Parent}(X_i))$$
$$= P(A)\,P(B)\,P(C|A,B)\,P(D|B)\,P(E|D)$$

# Bayesian Network - Structure Learning

The two key components of a structure learning algorithm are
a) searching for/generating "good" structures and
b) scoring these structures

✓ Heuristic Search Approaches

greedy-hill climbing, simulated annealing etc



# Bayesian Network – Structure Learning

Get the score for each network with respect to the training data

$$\mathbf{S(G{:}D) = \log \mathit{p}(D, S^h) = \log \overset{\text{prior}}{\overset{\downarrow}{\mathit{p}}}(S^h) + \log \overset{\text{likelihood}}{\overset{\downarrow}{\mathit{p}}}(D|S^h)}$$

$$\mathbf{\text{Likelihood } \log \mathit{p}(D|S^h) = \sum \log \mathit{p}(x_i \mid pa(x_i), S^h)}$$

Model with the highest log likelihood is a model that is the best predictor of the data D

# Summary

Bayesian network is suitable for genetic network reconstruction

- ✓ Can deal with stochastic nature
- ✓ Ideal for sparse domain (Useful for locally interacting components)
- ✓ Can handle noisy data
- ✓ Missing data
- ✓Inference reasoning

More research needed

- ✓ Incorporation of more biological information
- ✓ To model feedback process

=> Dynamic Bayesian networks

# References on networks building

- **Differential Expression**

1. Inferring Gene Regulator Networks from Time-Ordered Gene Expression Data Using Differential Equation
   by Michiel de Hoon et al. 2002.
2. Stability of Genetic Regulatory Network with Time Delay
   by Luonan chen et al. 2002.
3. Modeling Gene Expression with Differential Equations
   by Ting Chen et al. 1999.

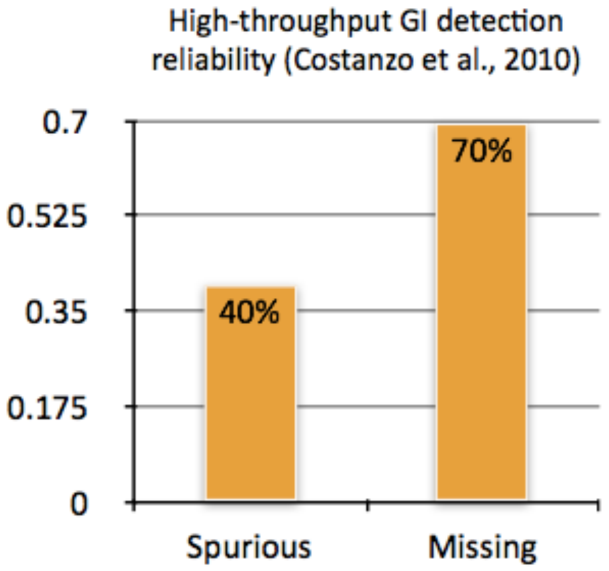
- **Bayesian Network**

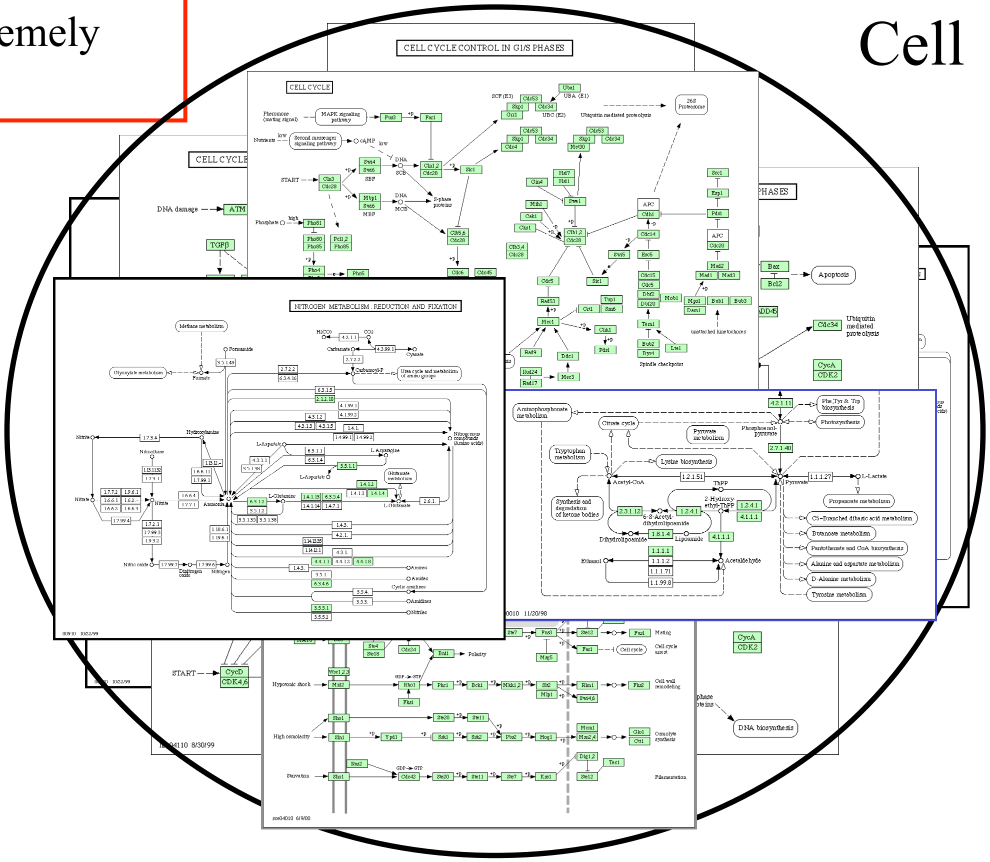
1. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection
   by Yoshinori et al. 2003.
2. Combining Location and Expression data for Principled Discovery of Genetic Regulatory Network Models
   by Hartemink et al. 2002.
3. Inferrring Subnetworks from Perturbed Expression Profiles
   by Pe’er et al. 2001.
4. Using Bayesian Networks to Analyze Expression Data
   by Friedman et al. 2000.

# Information Derivable from Chip Data

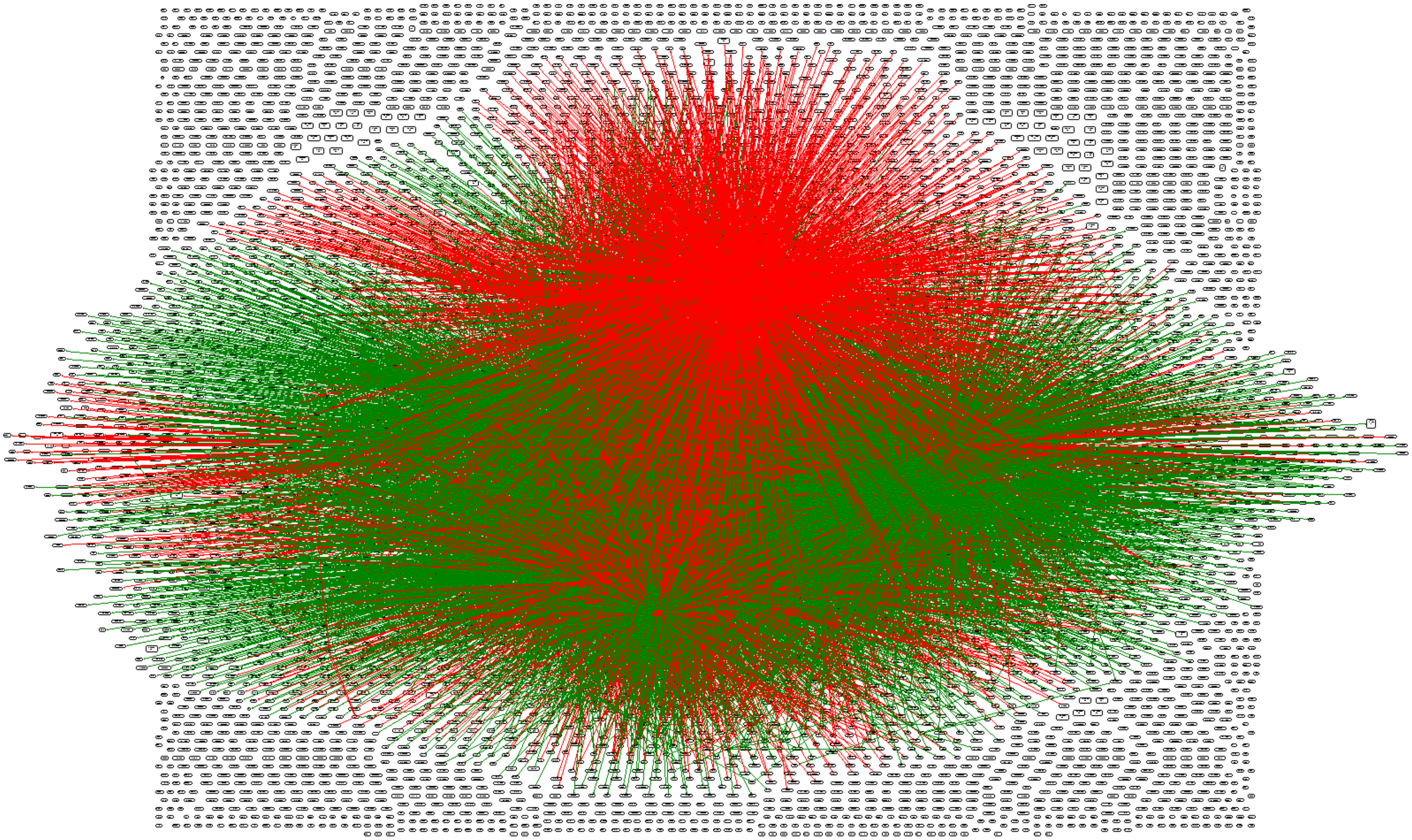
- The problem is the internal structure of a cell is very complex

Deciphering internal structure of a cell networks through computational prediction is extremely challenging and exciting problem!

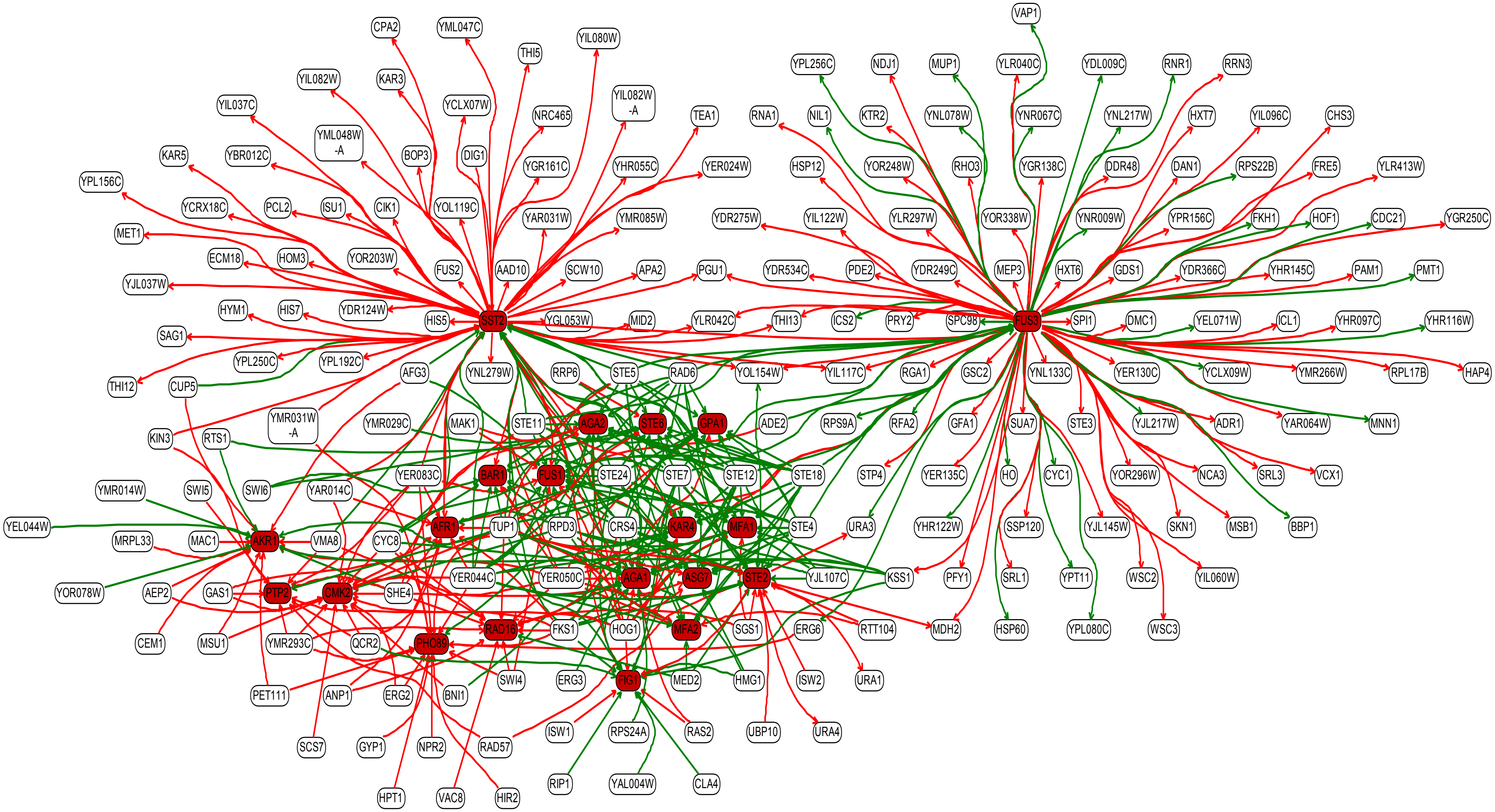



Cell
CELL CYCLE CONTROL IN G1/S PHASES
CELL CYCLE
NITROGEN METABOLISM REDUCTION AND FIXATION

# Mutation network for S. Cerevisiae

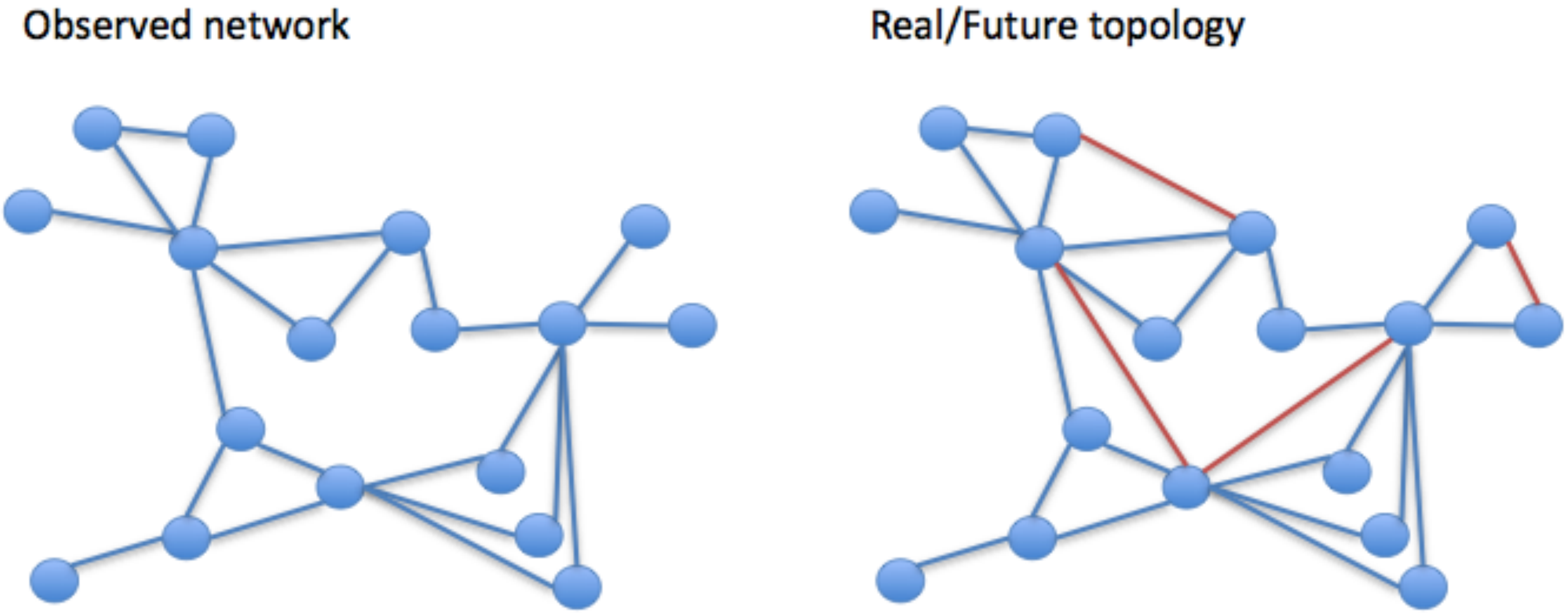


# Mutation network filtered for the genes marked in red (mating)



Thomas Schlitt, Johan Rung

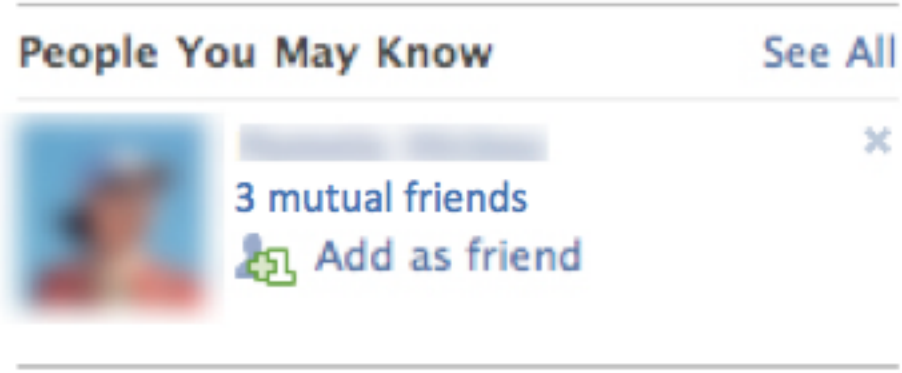
# Topological link prediction



Observed network

Real/Future topology

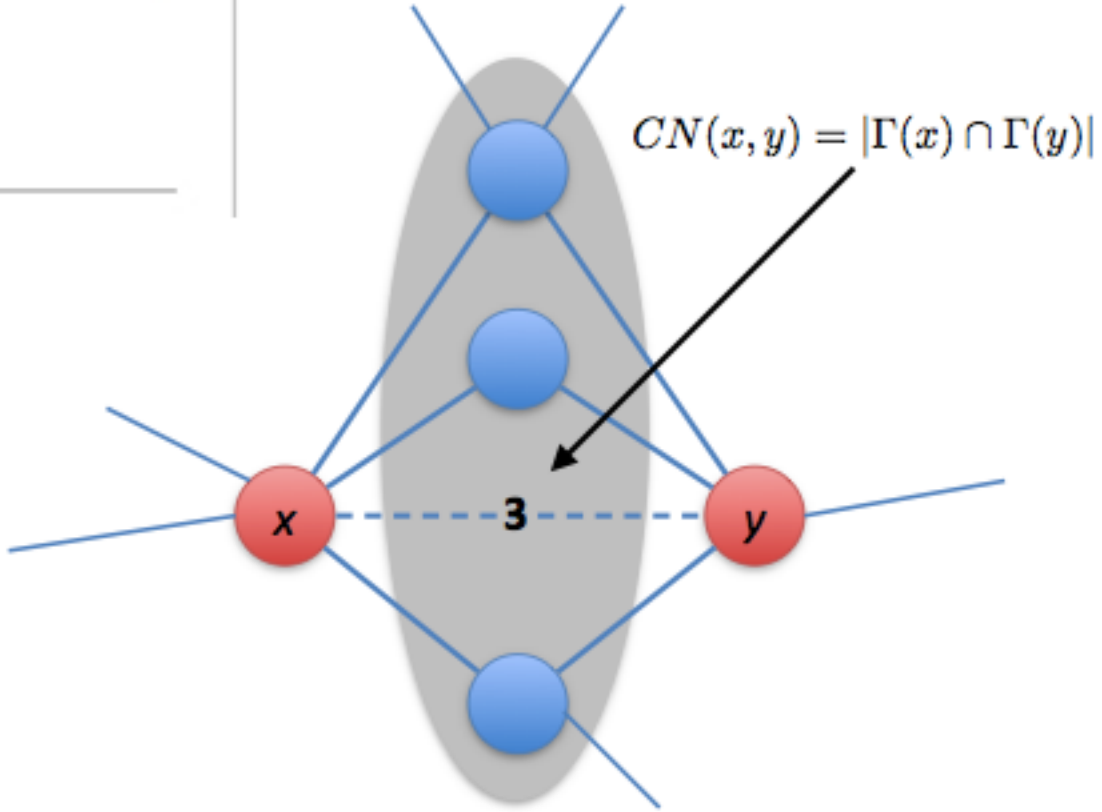
# A Local Community Approach to Link Prediction



People You May Know See All

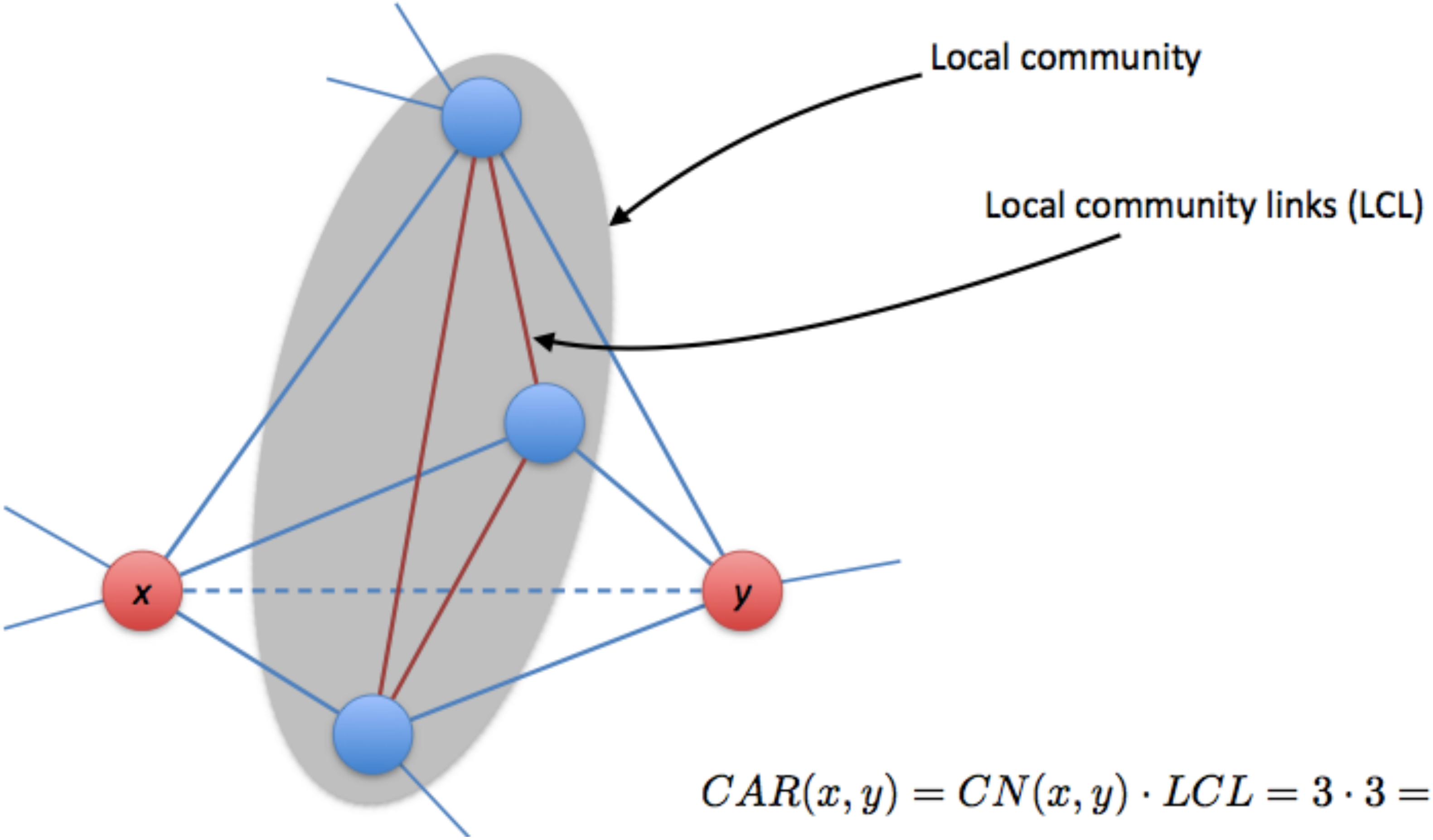
3 mutual friends

Add as friend



$CN(x,y) = |\Gamma(x) \cap \Gamma(y)|$

# Shift from nodes to links: local community links and CAR

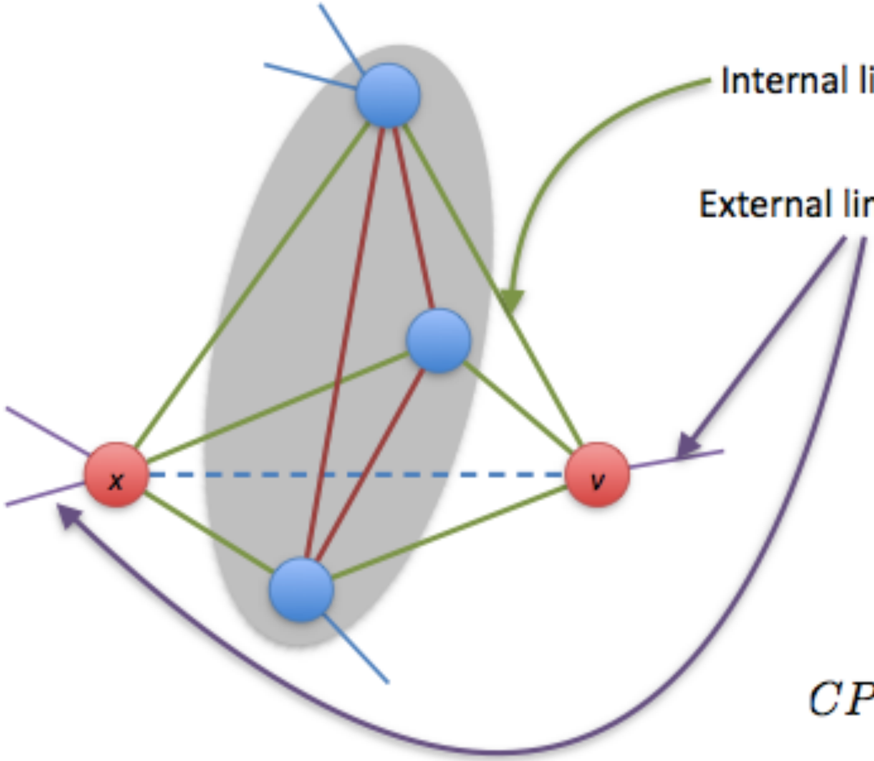


$$CAR(x, y) = CN(x, y) \cdot LCL = 3 \cdot 3 = 9$$

- Cannistraci, C.V., Alanis-Lobato, G. & Ravasi, T. (2013) From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. Scientific Reports 3, 1613. http://dx.doi. org/10.1038/srep01613. ©The Author 2013. Published by Nature Publishing Group.

# CAR variants of classical link predictors

$$JC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} = \frac{CN(x,y)}{|\Gamma(x) \cup \Gamma(y)|} \longrightarrow CJC(x,y) = \frac{CAR(x,y)}{|\Gamma(x) \cup \Gamma(y)|}$$



Internal links, $i_x = i_y = CN(x,y)$

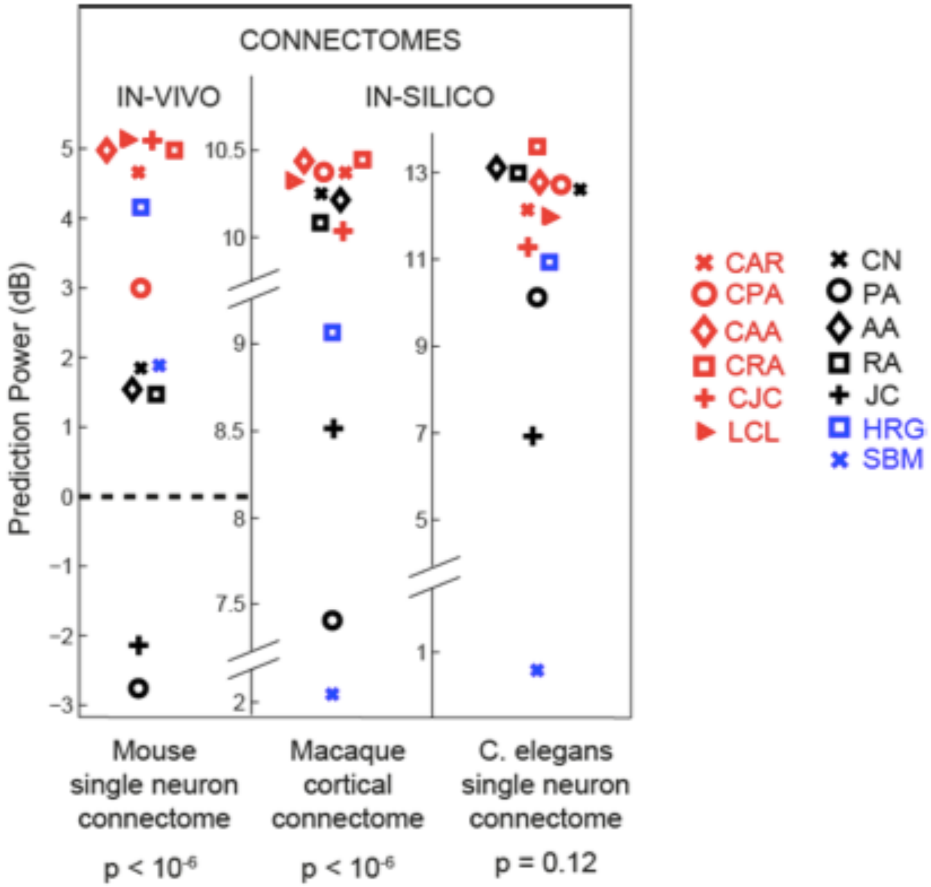
External links: $e_x, e_y$

$$\begin{aligned} PA(x,y) &= |\Gamma(x)| \cdot |\Gamma(y)| \\ &= (i_x + e_x)(i_y + e_y) \end{aligned}$$

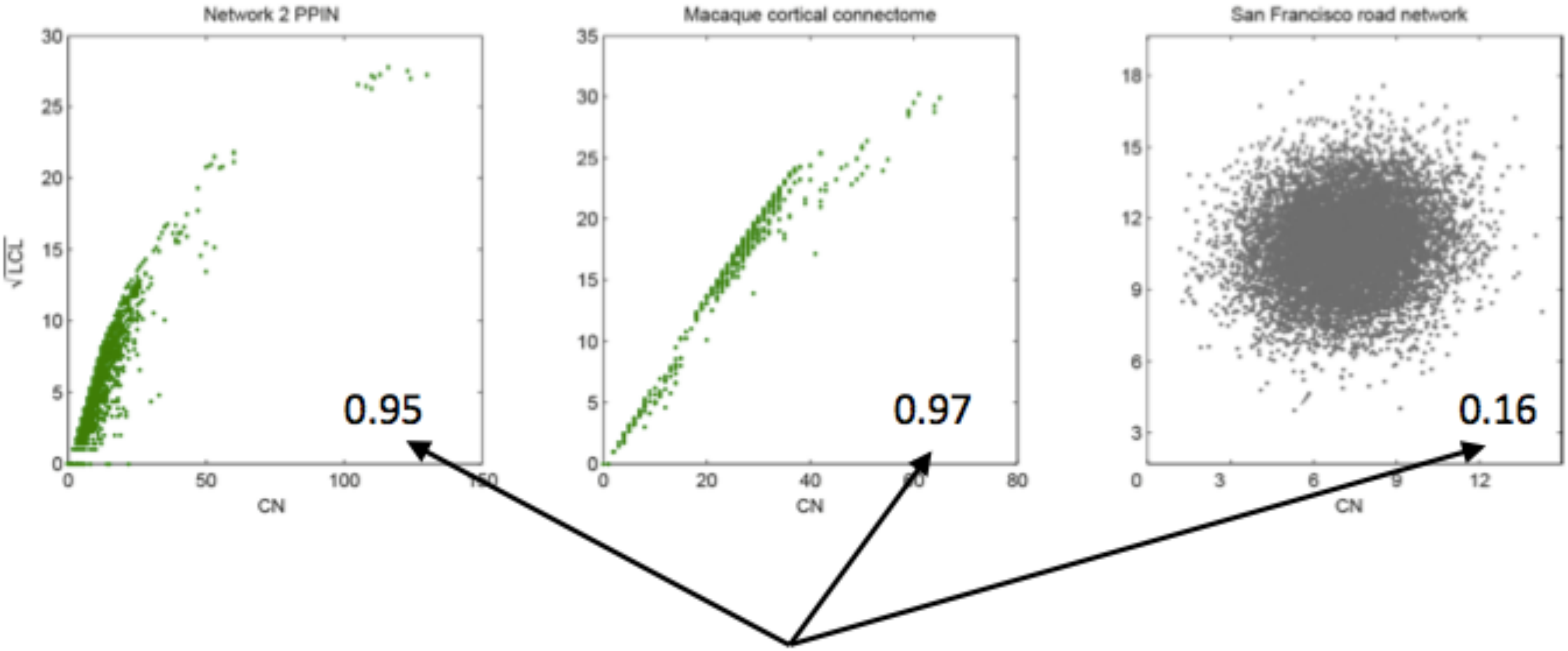
$$\big\downarrow$$

$$CPA(x,y) = (CAR(x,y) + e_x)(CAR(x,y) + e_y)$$

# Testing CAR in brain connectomes

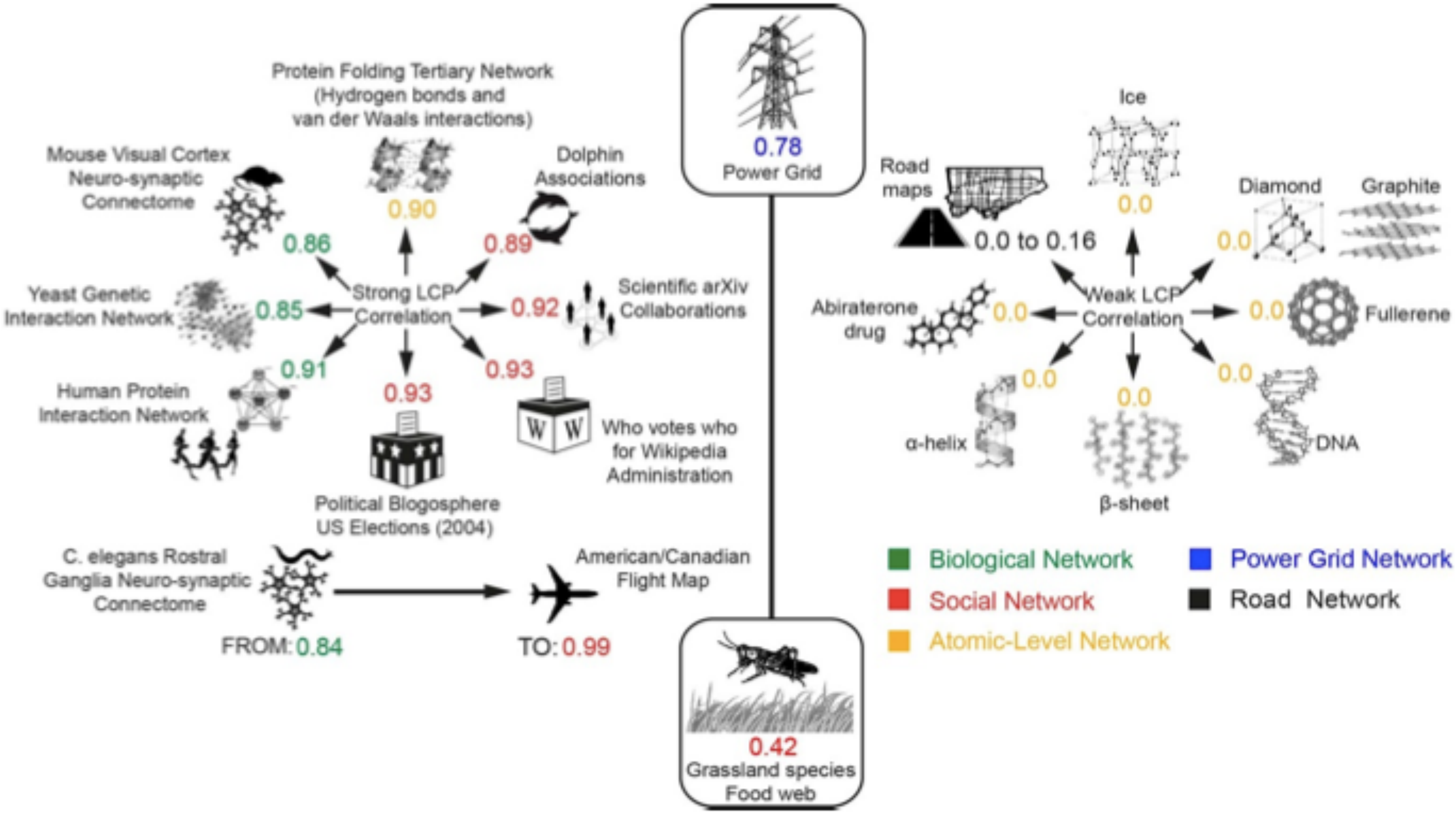


10% of links removed. Mean prediction precision considered relative to the mean random predictor performance



$$\text{LCP-corr}(G) = \text{Pearson}(CN, \sqrt{LCL})$$

# LCP and non-LCP networks



# Autoimmune Disease Network

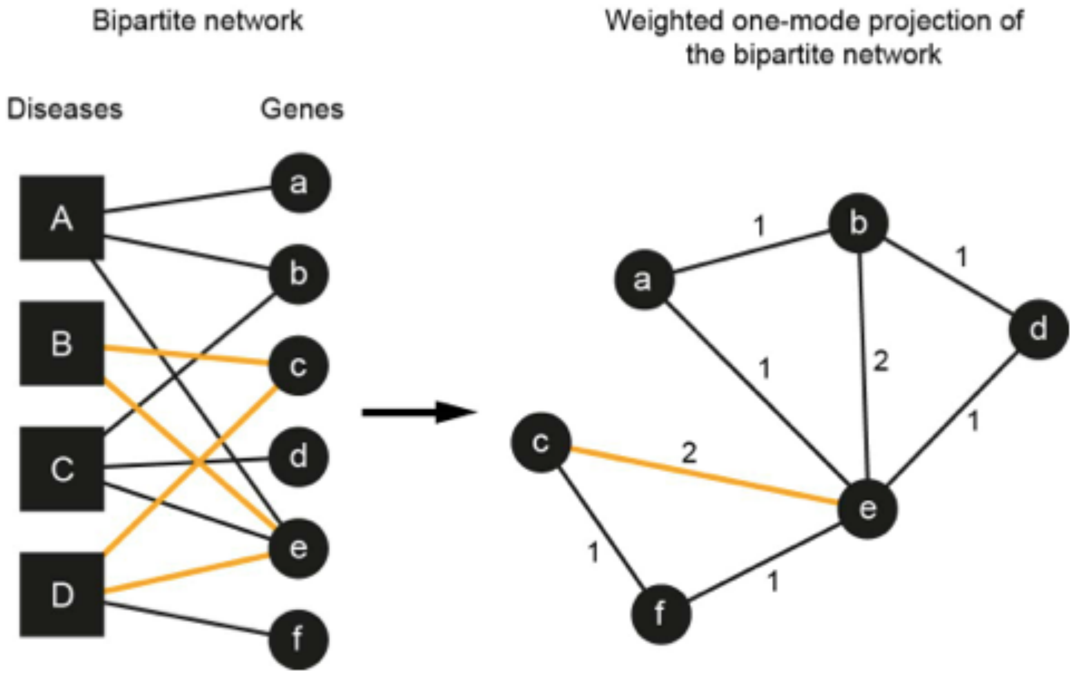
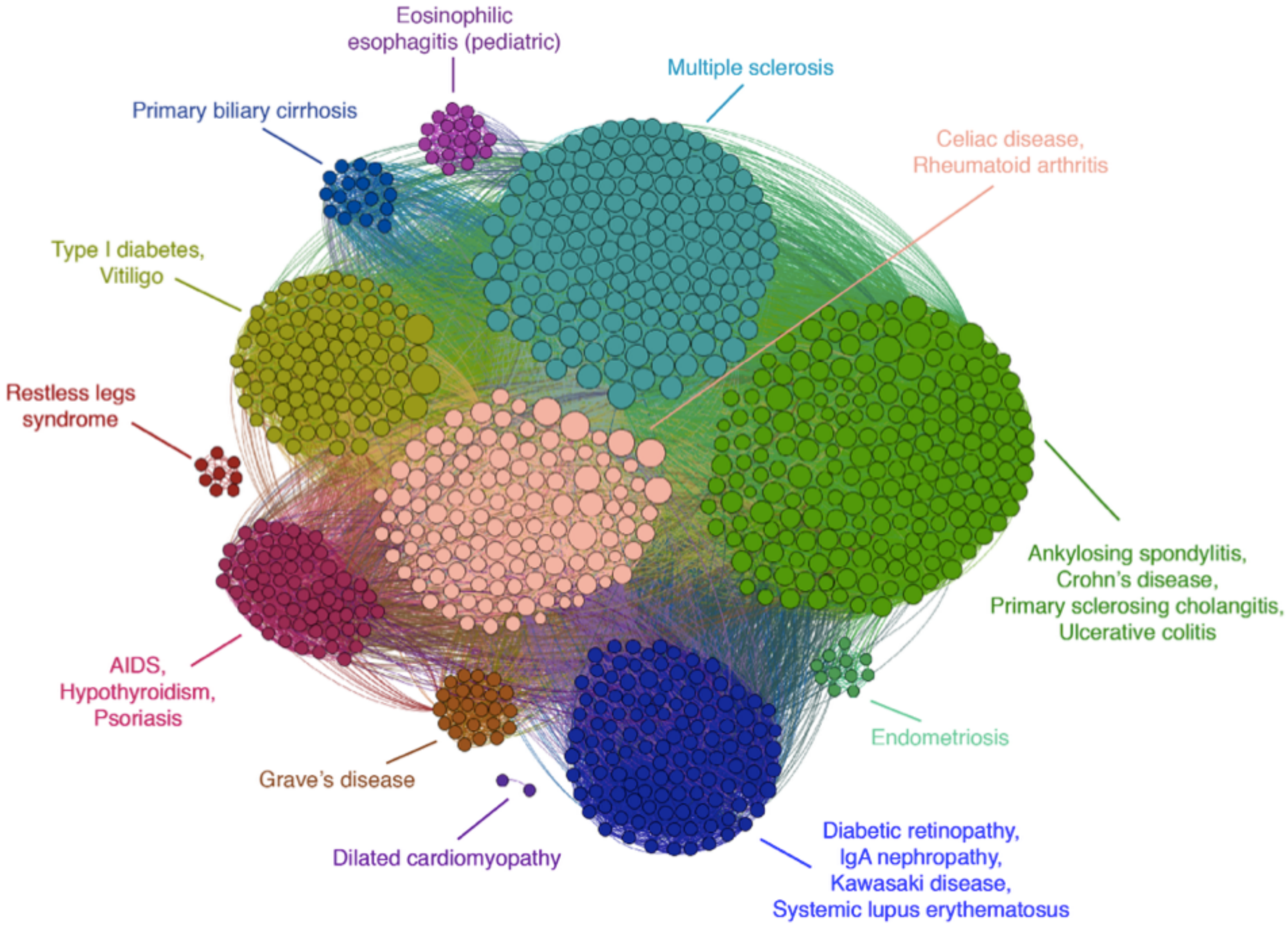
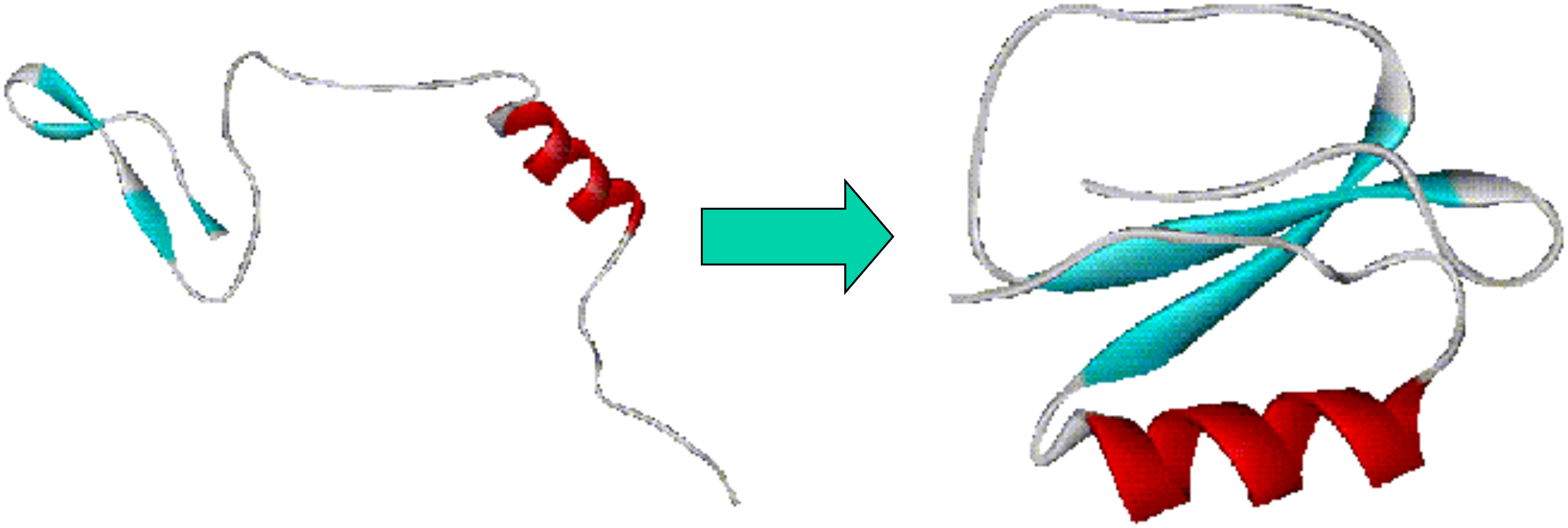


Figure 6.1: One-mode projection of the bipartite network of genes and diseases. In the highlighted example, gene c is associated with diseases B and D whereas gene e is associated with diseases A, B, C and D. Since they have two diseases in common (B and D), they are linked with a weight of 2 in the projection of the bipartite network to the gene space.

- Alanis-Lobato, G., Cannistraci, C.V. & Ravasi, T. (2014) Exploring the Genetics Underlying Autoimmune Diseases with Network Analysis and Link Prediction. In Proceedings of the MECBME 2014, 167-170. http://dx.doi.org/10.1109/MECBME.2014.6783232. ©IEEE 2014. All rights reserved.


Eosinophilic esophagitis (pediatric)
Multiple sclerosis
Primary biliary cirrhosis
Celiac disease, Rheumatoid arthritis
Type I diabetes, Vitiligo
Restless legs syndrome
Ankylosing spondylitis, Crohn's disease, Primary sclerosing cholangitis, Ulcerative colitis
AIDS, Hypothyroidism, Psoriasis
Endometriosis
Grave's disease
Dilated cardiomyopathy
Diabetic retinopathy, IgA nephropathy, Kawasaki disease, Systemic lupus erythematosus

# Folding of chymotrypsin protein



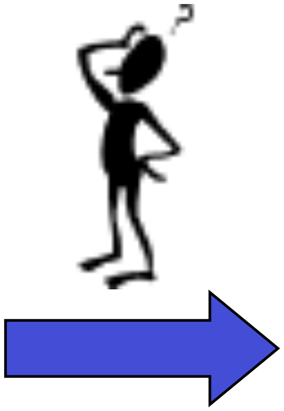
# Protein Folding Problem

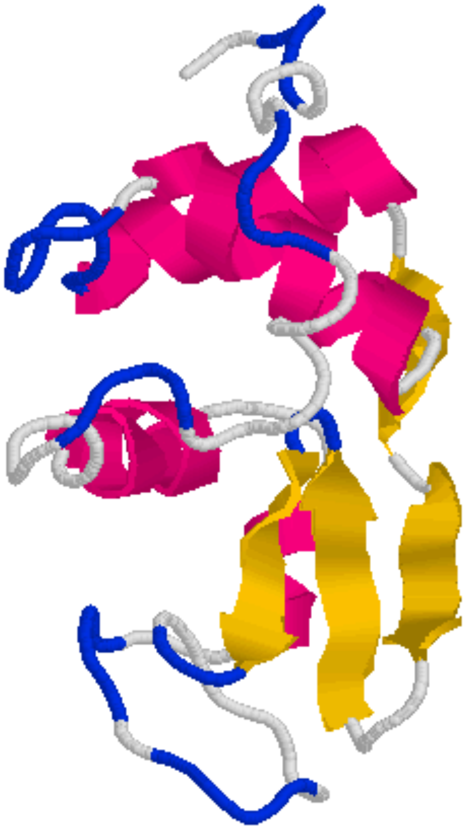
A protein folds into a unique 3D structure under the physiological condition.

Can we predict structure (fold) from sequence?

**Lysozyme sequence:**

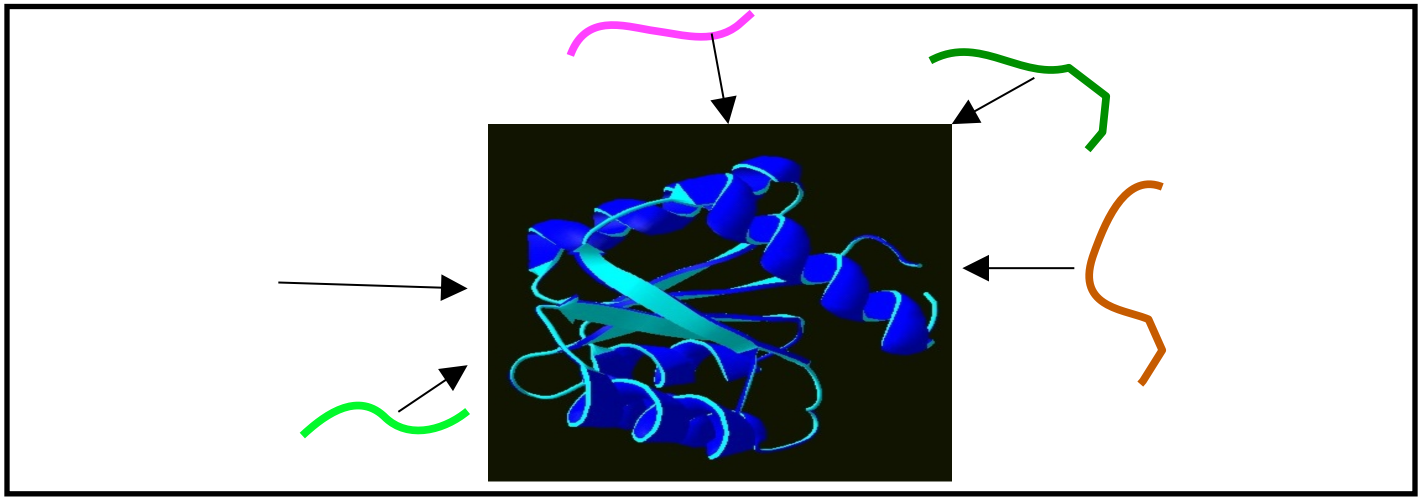
```
KVFGRCELAA AMKRHGLDNY
RGYSLGNWVC AAKFESNFNT
QATNRNTDGS TDYGILQINS
RWWCNDGRTP GSRNLCNIPC
SALLSSDITA SVNCAKKIVS
DGNGMNAWVA WRNRCKGTDV
QAWIRGCRL
```





# Many proteins with dissimilar sequences fold into similar structures

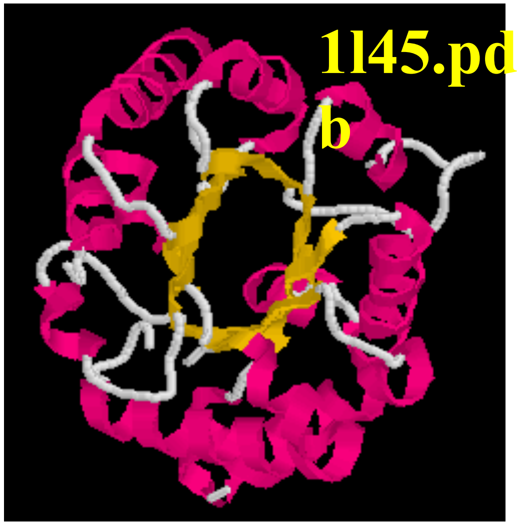
- Estimated number of folds: ~10000



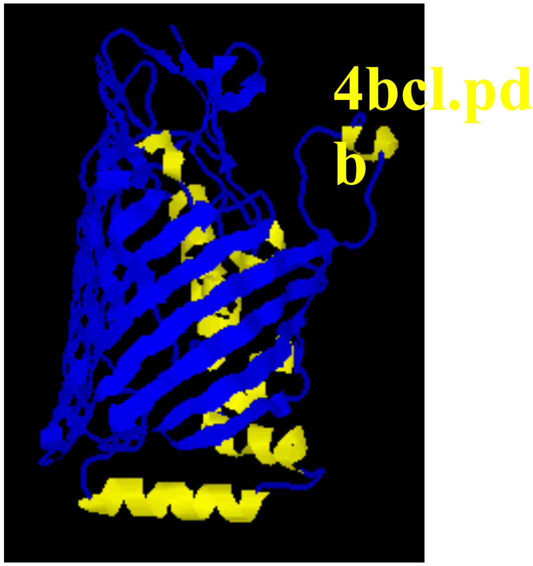
Protein Folds: sequential and spatial arrangement of secondary structures

# Examples of different Folds

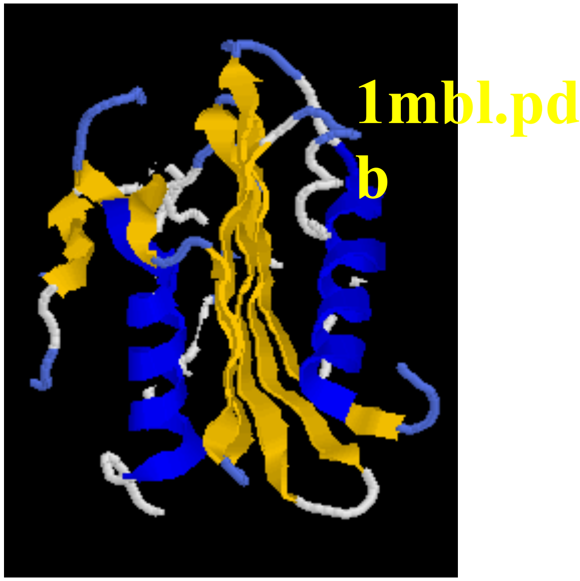
Refers to the spatial arrangement of its secondary structural elements (α-helices and β-strands)



α/β-barrel



β-barrel



α/β-sandwich

# Predicting Protein Structure: Alternative Methods

•***Ab initio* prediction**
(no similarity with any sequence of known structure)
Given only the sequence, predict the 3D structure from “first principles”, based on energetic or statistical principles.

•**Sequence-structure threading = Fold recognition**
(sequences with <= 30% sequence identity to sequences of known structure)
Given the sequence, and a set of folds observed in PDB, see if any of the sequences could adopt one of the known folds.

•**Homology Modelling**
Given a sequence with homology (> 30%) to a known structure in PDB, use known structure as template to create a 3D model from the sequence.

# Approaches to Ab-initio Prediction

**Molecular Mechanics**

- folded form is the minimal energy conformation of the protein

**Molecular Dynamics**

- Simulates the forces that governs the protein within water

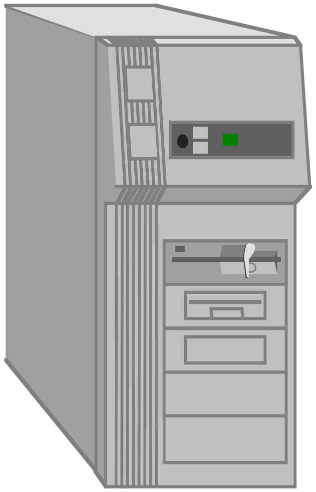
Problems:

- Thousands of atoms
- Huge number of time steps to reach folded protein
- There is no correct energy function
- Optimization in multi-minima space (most methods can reach only local minimum)

➔Intractable problem

# Forces Involved in Molecular Interactions

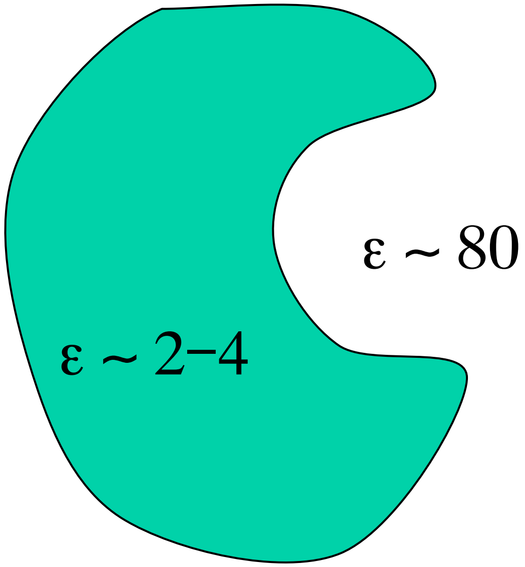
- Bond stretch
- Bond angle bending
- Torsion (bond rotation)
- Hydrogen bonding
- van der Waals interactions
- Electrostatic interactions
- Empirical solvation free energy



$$
\begin{aligned}
V = {} & \Sigma_{\text{bond}}\ 1/2K_b\ (r-r_{eq})^2 + \\
& \text{Sangle}\ ½\ K_\theta\ (\theta-\theta_{eq})^2 + \\
& \Sigma_{\text{torsions}}\ 1/2\ V_n\ [\ 1 + \cos(n\phi-\gamma')\ ] + \\
& \Sigma_{\text{H bonds}}\ [\ V_0\ (1-e^{-a(r-r0)}\ )^2 - V_0\ ] + \\
& \Sigma_{\text{non bonded}}\ [\ A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6} + q_i q_j\ /\varepsilon_r\ r_{ij}] + \\
& \Sigma_{\text{atoms i}}\ \ \Delta\sigma_i\ A_i
\end{aligned}
$$

# Electrostatic interactions: Solvent dielectric model?

- Problem: Inhomogeneous permittivity

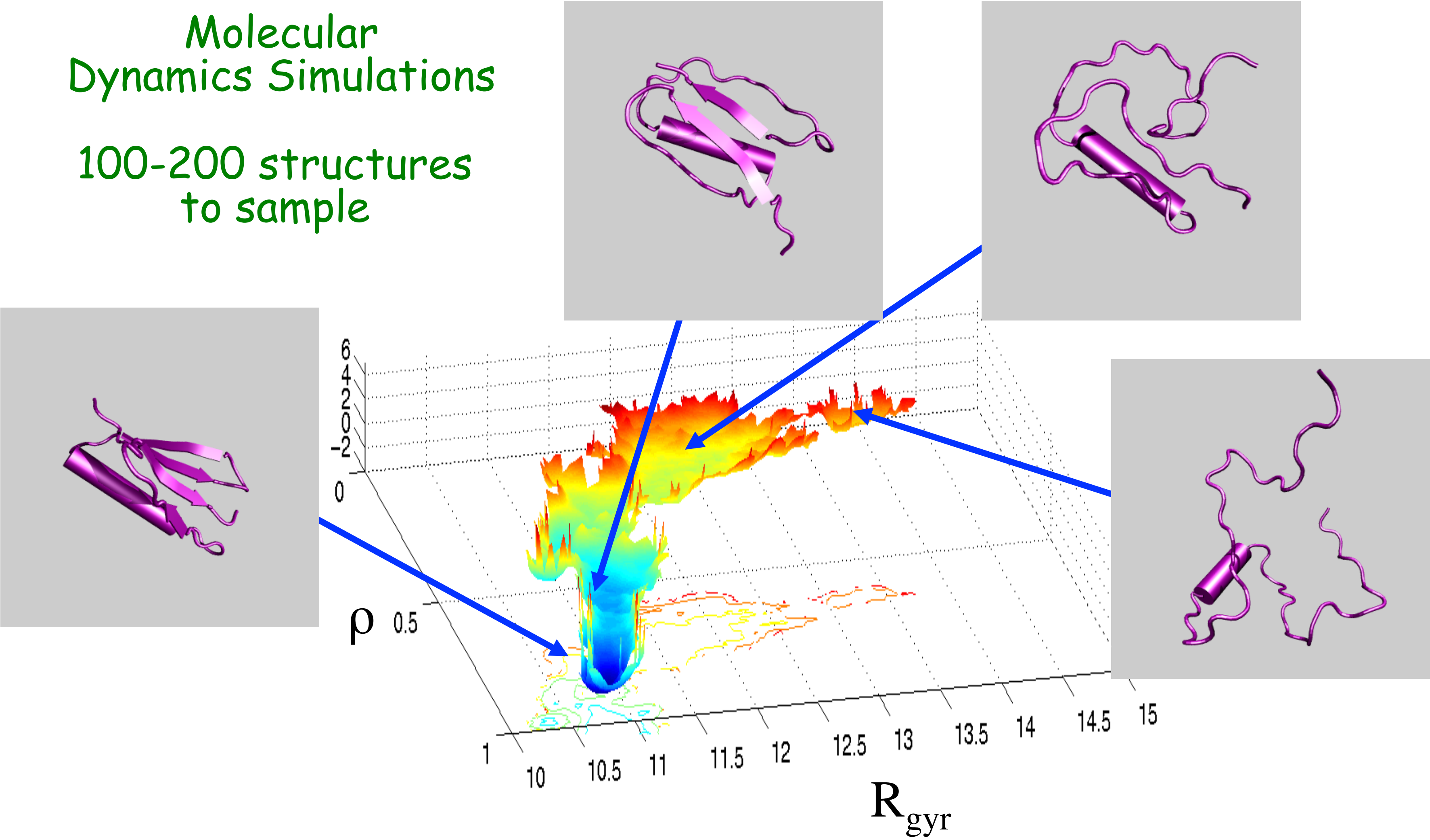


Depends on local structure and interactions with water

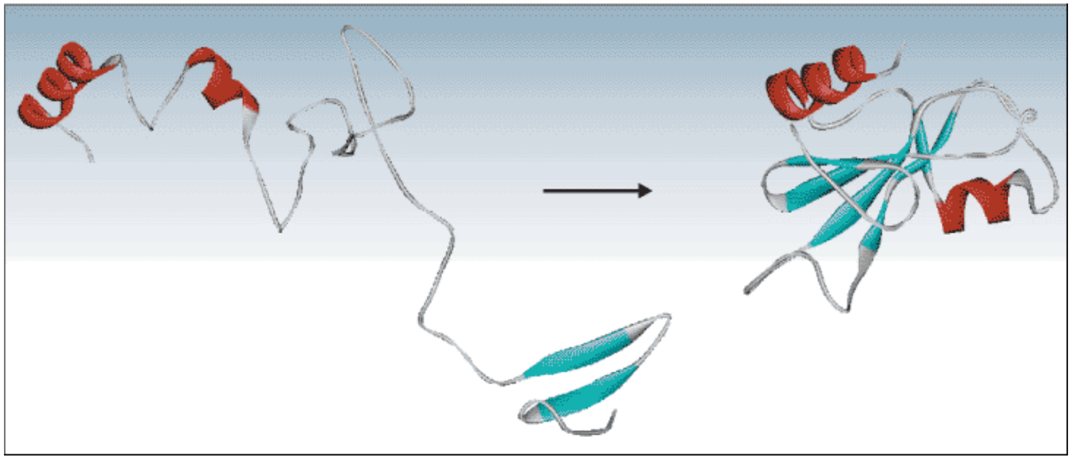
# Folding Free Energy Landscape

Molecular
Dynamics Simulations

100-200 structures
to sample



# *Ab initio* protein folding simulation



BlueGene

| | |
|---|---|
| Physical time for simulation | $10^{-4}$ seconds |
| Typical time-step size | $10^{-15}$ seconds |
| Number of MD time steps | $10^{11}$ |
| Atoms in a typical protein and water simulation | 32’000 |
| Approximate number of interactions in force calculation | $10^{9}$ |
| Machine instructions per force calculation | 1000 |
| Total number of machine instructions | $10^{23}$ |
| BlueGene capacity (floating point operations per second) | ($10^{15}$) |

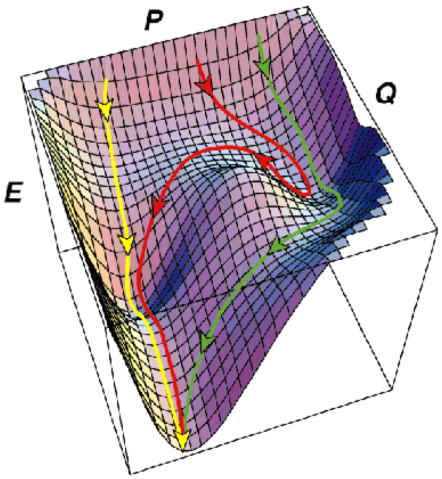
➔ **Blue Gene will need 3 years to simulate 100 μsec.**

# Why Do We Need Homology Modelling?

- *Ab Initio* protein folding (“random” sampling):
  - 100 aa, 10 conf./residue gives approximately $10^{100}$ different overall conformations!

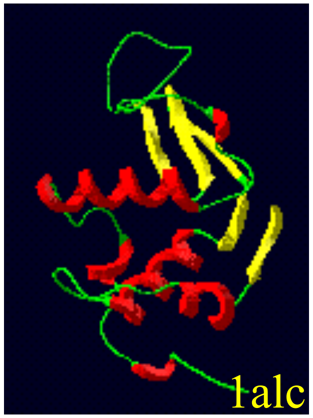
- Random sampling is *NOT feasible*, even if conformations can be sampled at picosecond ($10^{-12}$ sec) rates.
  - Levinthal’s paradox if a protein were to attain its correctly folded configuration by sequentially sampling all the possible conformations, it would require a time longer age of the universe to arrive at its correct native conformation

- Do fold recognition or homology modelling instead.



# Comparative Modeling (homology modeling)

```
KQFTKCELSQNLYDIDGYGRIALPELICTMFH
TSGYDTQAIVENDESTEYGLFQISNALWCKSS
QSPQSRNICDITCDKFLDDDITDDIMCAKKIL
DIKGIDYWIAHKALCTEKLEQWLCEKE
```

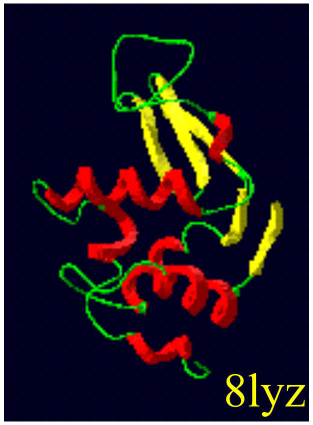


Homologous

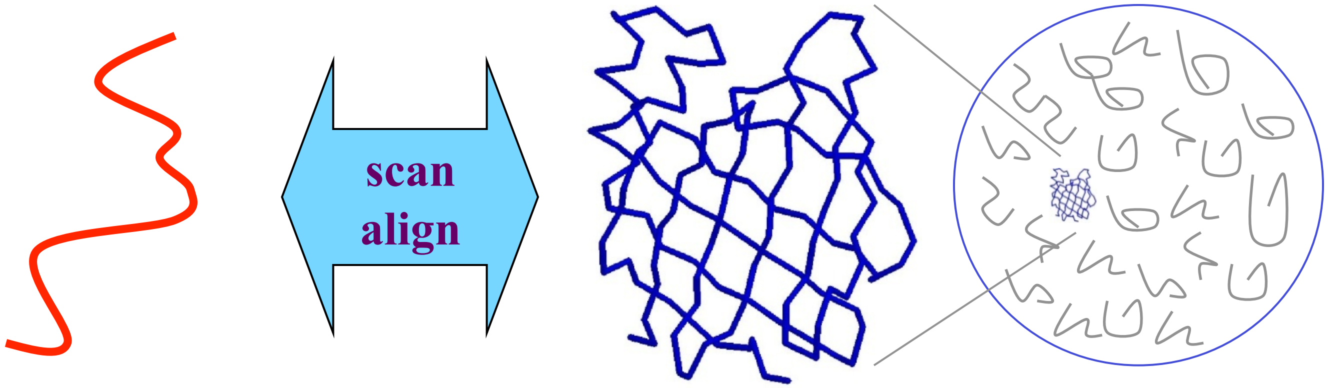
**Share**
**Similar**
**Sequence**

**Use as template**
**& model**

```
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKF
ESNFNTQATNRNTDGSTDYGILQINSRWWCNDGR
TPGSRNLCNIPCSALLSSDITASVNCAKKIVSDG
NGMNAWVAWRNRCKGTDVQAWIRGCRL
```



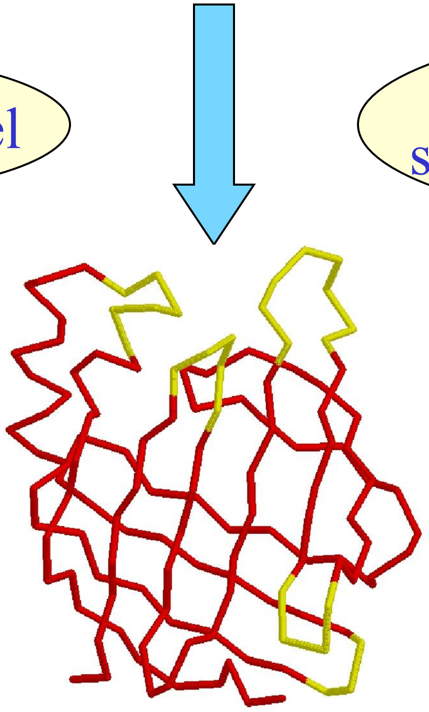
# Comparative modelling of protein structure



```
... KDHPFGFAVPTKNPDGTMNLMNWECAIP ...
    KDPPAGIGAPQDN----QNIMLWNAVIP
    ** * *   *  *     * * *   **
```

build initial model

construct non-conserved side chains and main chains



refine

# Fold Recognition

*Homology modeling refers to the easy case when the template structure can be identified using BLAST alone.*

What to do when BLAST fails to identify a template?

- *Use more sophisticated sequence methods*
  - Profile-based BLAST: PSIBLAST
  - Hidden Markov Models (HMM)

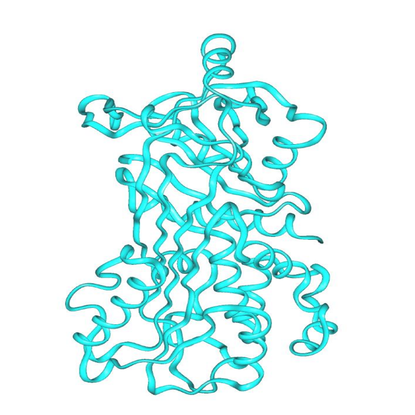
- *Use secondary structure prediction to guide the selection of a template, or to validate a template*

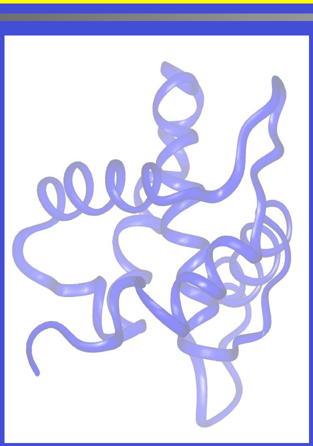
- *Use threading programs: sequence-structure alignments*

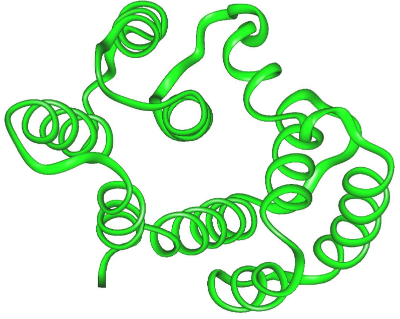
- *Use all of these methods! Meta-servers*

# Fold Recognition: problem definition

**A Library of Protein Folds (finite number)**







**Query sequence**

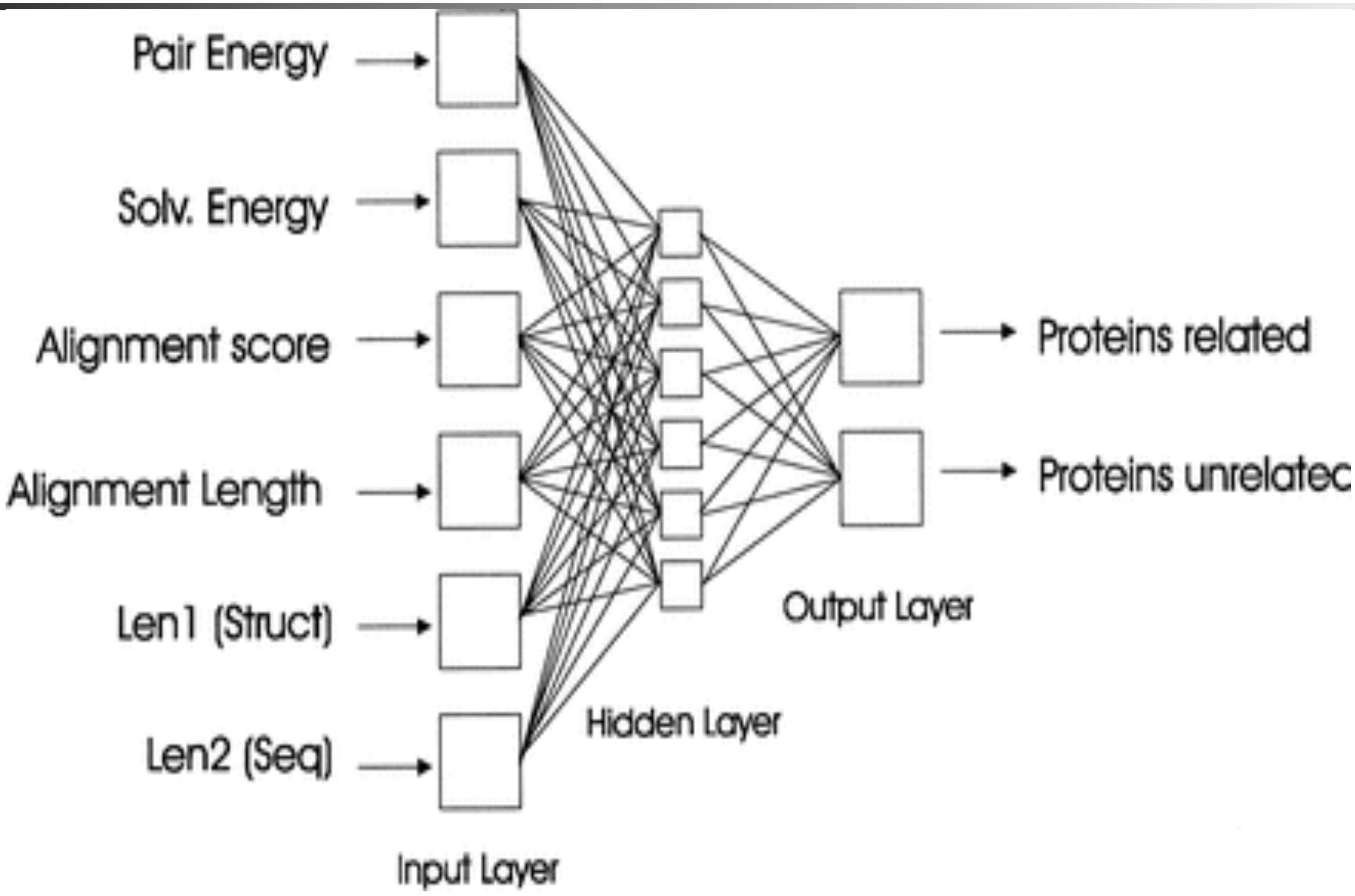
MTYGFRIPLNCERWGHKLSTVILKRP...

**Goal: find to what folding template the sequence fits best**

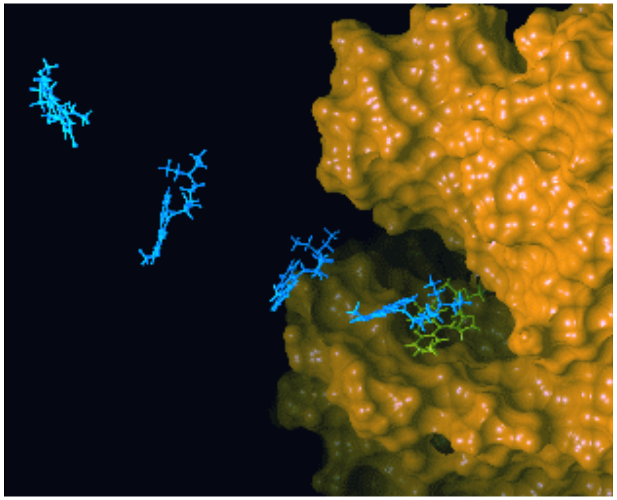
↓

**Find ways to evaluate sequence-structure fit**

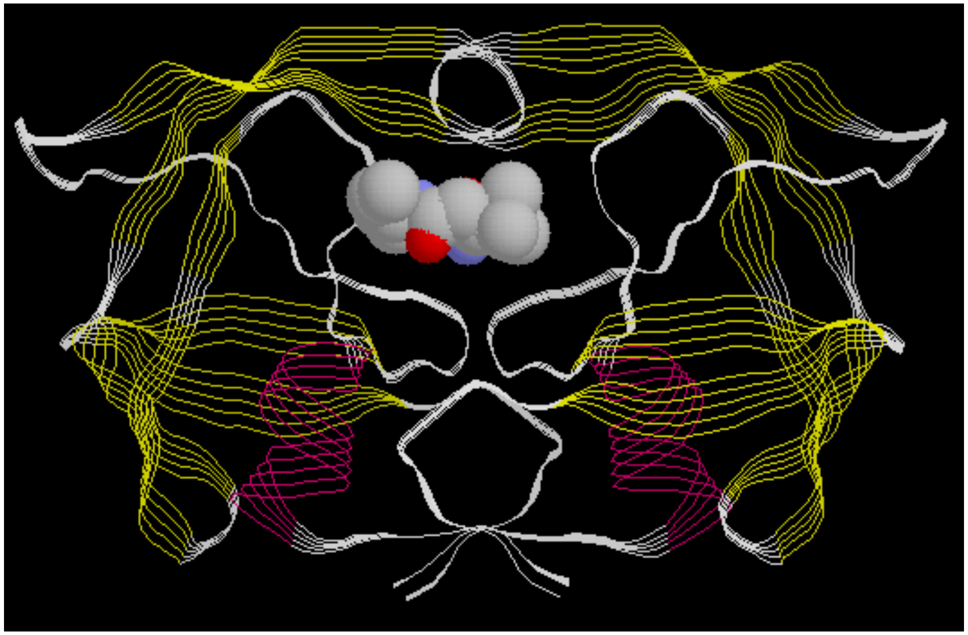
# Essentials of GenTHREADER



# Structure-Based Drug Design

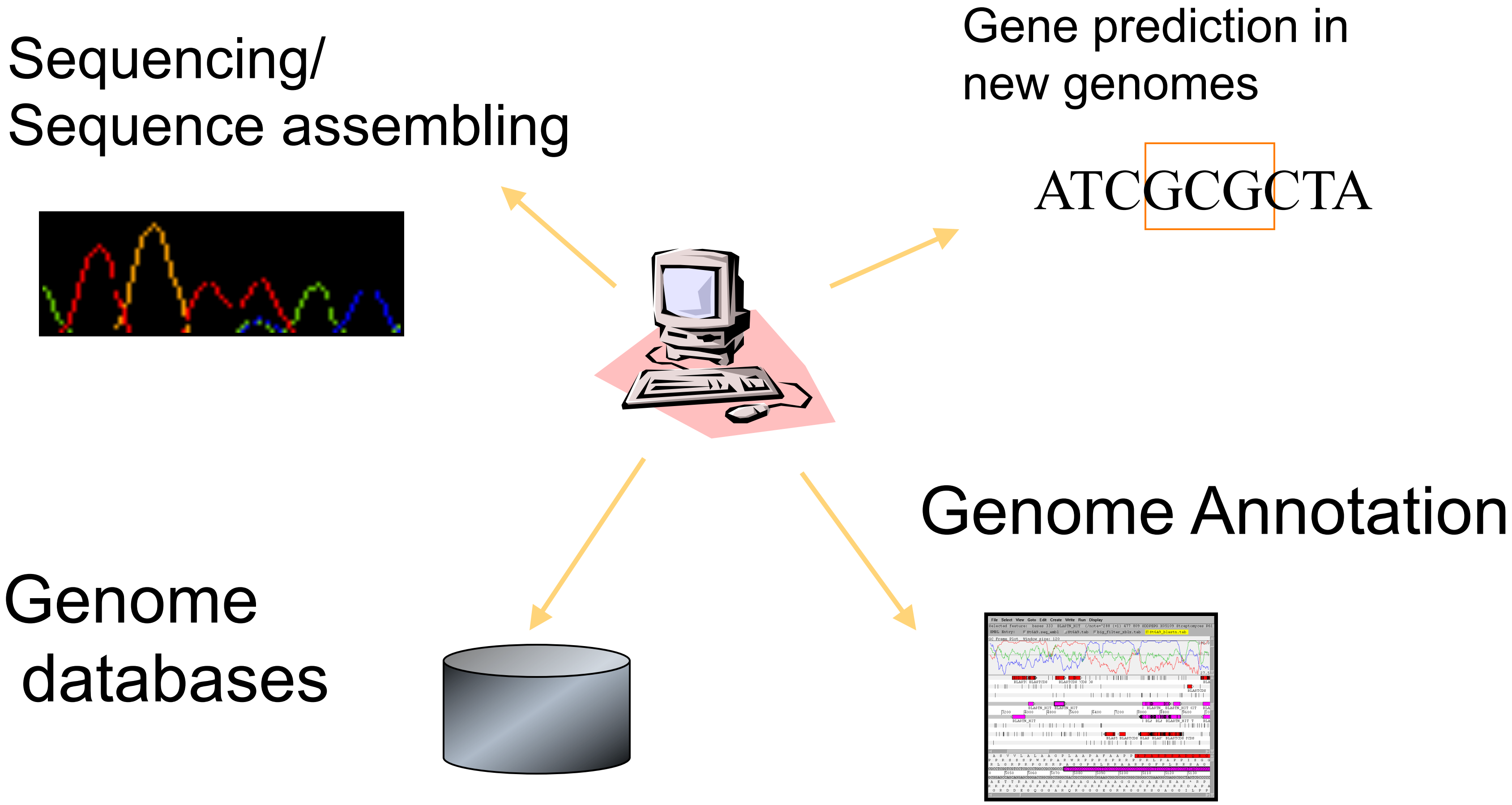


**Structure-based rational drug design is still a major method for drug discovery.**



HIV protease inhibitor

# The role of Bioinformatics in support of genomics



Sequencing/
Sequence assembling

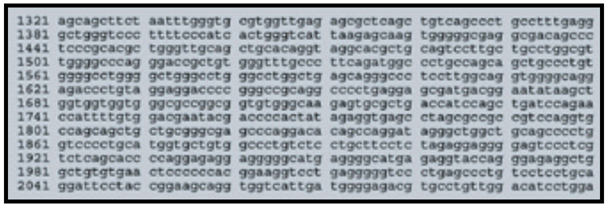
Gene prediction in
new genomes

ATCGCGCTA

Genome Annotation

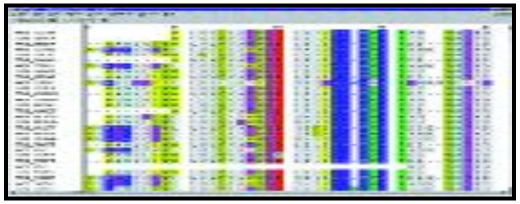
Genome
databases

# The role of bioinformatics supporting genetics

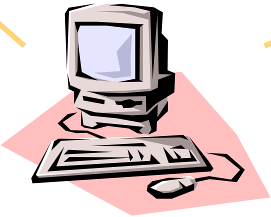


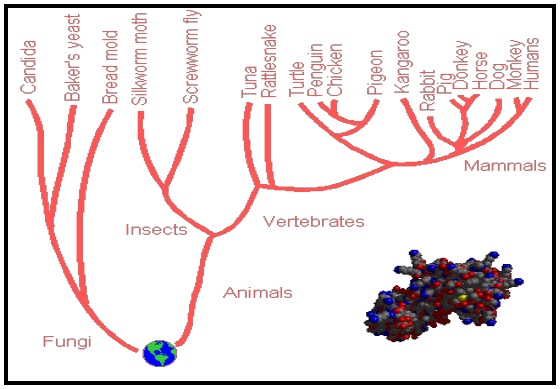
```
1321 agcagcttct aatttgggtg cgtggttgag agcgctcagc tgtcagccct gcctttgagg
1381 gctgggtccc ttttcccatc actgggtcat taagagcaag tgggggcgag gcgacagccc
1441 tcccgcacgc tgggttgcag ctgcacaggt aggcacgctg cagtccttgc tgcctggcgt
1501 tggggcccag ggaccgctgt gggtttgccc ttcagatggc cctgccagca gctgccctgt
1561 ggggcctggg gctgggcctg ggcctggctg agcagggccc tccttggcag gtggggcagg
1621 agaccctgta ggaggacccc gggccgcagg cccctgagga gcgatgacgg aatataagct
1681 ggtggtggtg ggcgccggcg gtgtgggcaa gagtgcgctg accatccagc tgatccagaa
1741 ccattttgtg gacgaatacg accccactat agaggtgagc ctagcgccgc cgtccaggtg
1801 ccagcagctg ctgcgggcga gcccaggaca cagccaggat agggctggct gcagcccctg
1861 gtcccctgca tggtgctgtg gccctgtctc ctgcttcctc tagaggaggg gagtccctcg
1921 tctcagcacc ccaggagagg agggggcatg aggggcatga gaggtaccag ggagaggctg
1981 gctgtgtgaa ctccccccac ggaaggtcct gaggggtcc ctgagccctg tcctcctgca
2041 ggattcctac cggaagcagg tggtcattga tggggagacg tgcctgttgg acatcctgga
```

Identification of sequence functions and functional signals



Alignments

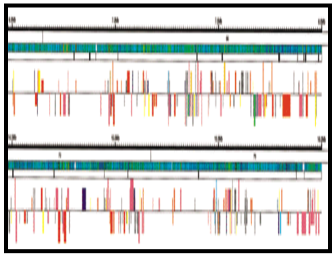




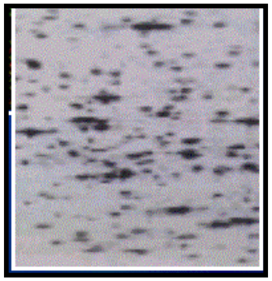
Structures

Phylogenetic trees

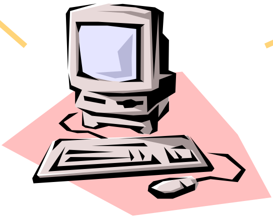
# Bioinformatics in support of Post-Genomic Research

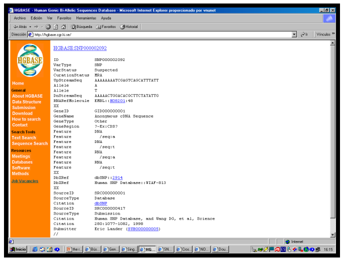


Genomes: Comparative Genomics (homology, evolution)



Proteomics (proteins in cells)





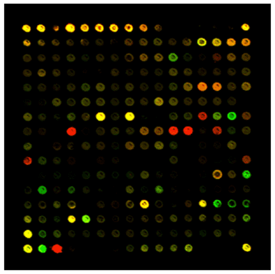
HGBASE - Human Genic Bi-Allelic Sequences Database - Microsoft Internet Explorer proporcionado por vnunet
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos Historial
Dirección http://hgbase.cgr.ki.se/
Ir Vínculos
HGBASE
Home
General
About HGBASE
Data Structure
Submission
Download
How to search
Contact
Search Tools
Text Search
Sequence Search
Resources
Meetings
Databases
Software
Methods
Job Vacancies
Internet
Inicio
16:15

HGBASE:SNP000002092

```
ID              SNP000002092
VarType         SNP
VarStatus       Suspected
CurationStatus  MRA
UpStreamSeq     AAAAAAAATCGAGTCAGCATTTATT
Allele          A
Allele          T
DnStreamSeq     AAAAACTGGACACGCTTCTATATTG
RNARefMolecule  EMBL::H08201:48
XX
GeneID          GID000000001
GeneName        Anonymous cDNA Sequence
GeneType        Other
GeneRegion      ?-Ex:CDS?
Feature         DNA
Feature           /seq:a
Feature         DNA
Feature           /seq:t
Feature         RNA
Feature           /seq:a
Feature         RNA
Feature           /seq:t
XX
DbXRef          dbSNP::2914
DbXRef          Human SNP Database::WIAF-813
XX
SourceID        SRC000000001
SourceType      Database
Citation        dbSNP
SourceID        SRC000000417
SourceType      Submission
Citation        Human SNP Database, and Wang DG, et al, Science
Citation        280:1077-1082, 1998
Submitter       Eric Lander (SUB000000005)
//
```

SNPs

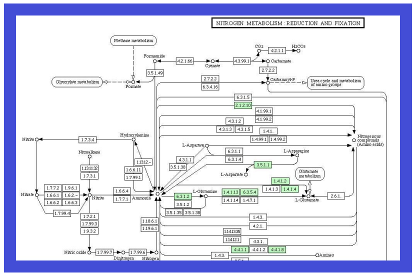
Individual Genome mutations/variations



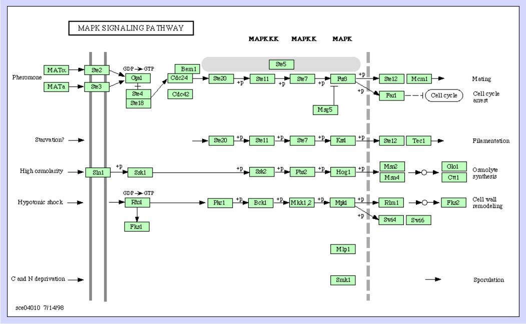
Functional Genomics (mRNAs)

DNA microarrays
Transcriptome Sequencing

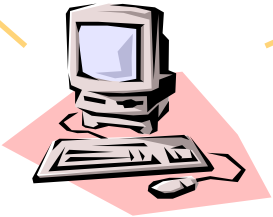
# Bioinformatics in support of Systems Biology

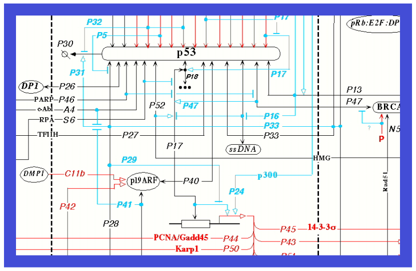


Metabolic Pathways

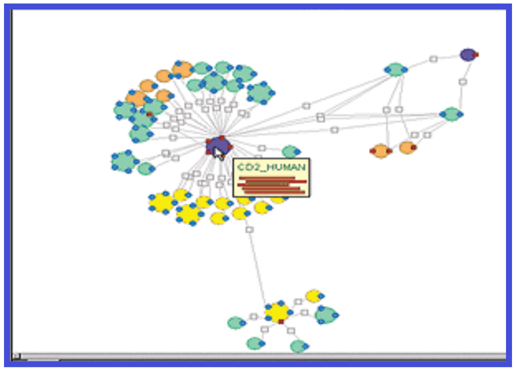


Signaling pathways



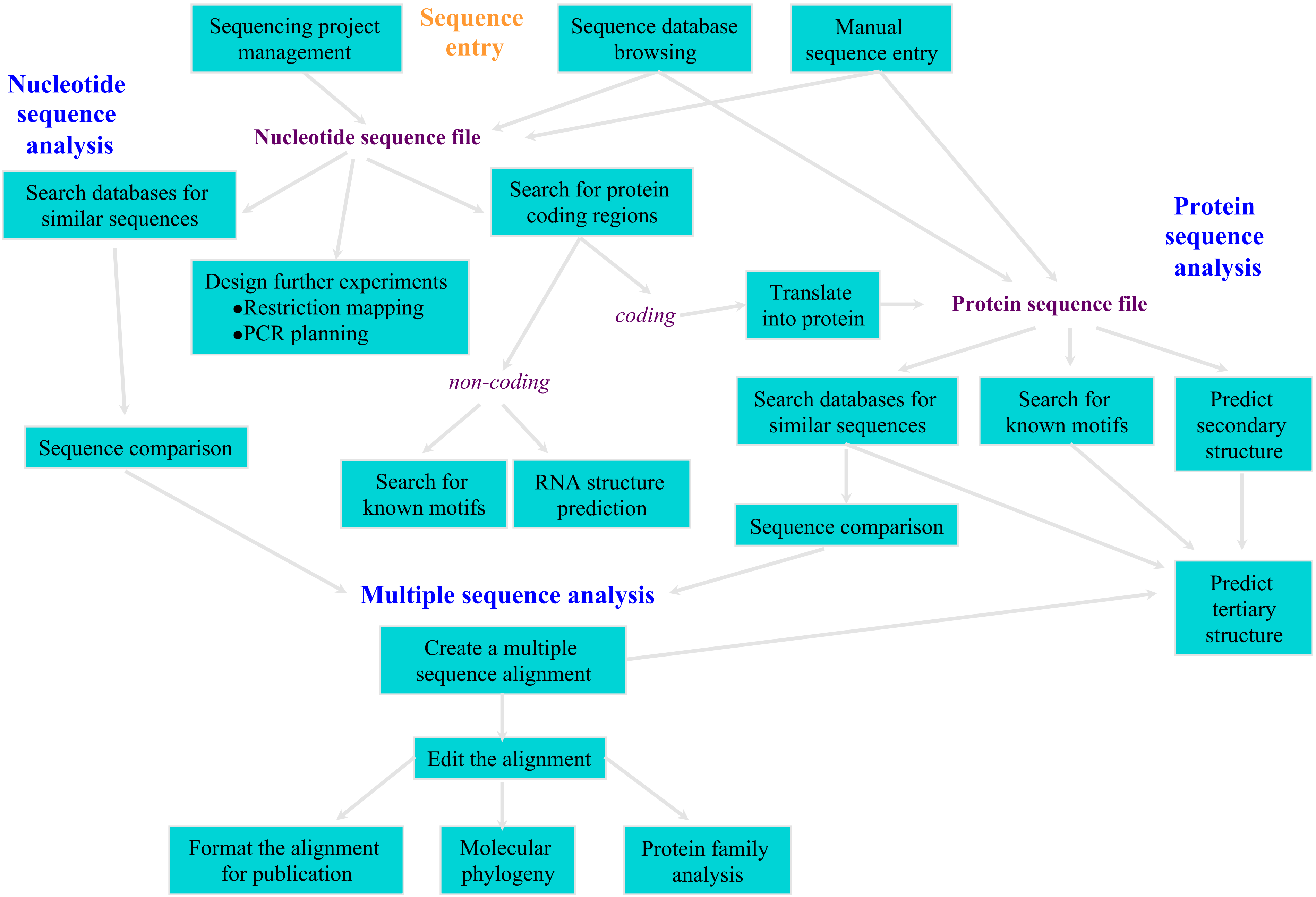


Genetic Networks



Interactions

# Sequence analysis: overview



Sequence entry

Sequencing project management

Sequence database browsing

Manual sequence entry

Nucleotide sequence analysis

Nucleotide sequence file

Search databases for similar sequences

Design further experiments
- Restriction mapping
- PCR planning

Search for protein coding regions

*coding*

Translate into protein

*non-coding*

Search for known motifs

RNA structure prediction

Sequence comparison

Protein sequence analysis

Protein sequence file

Search databases for similar sequences

Search for known motifs

Predict secondary structure

Sequence comparison

Predict tertiary structure

Multiple sequence analysis

Create a multiple sequence alignment

Edit the alignment

Format the alignment for publication

Molecular phylogeny

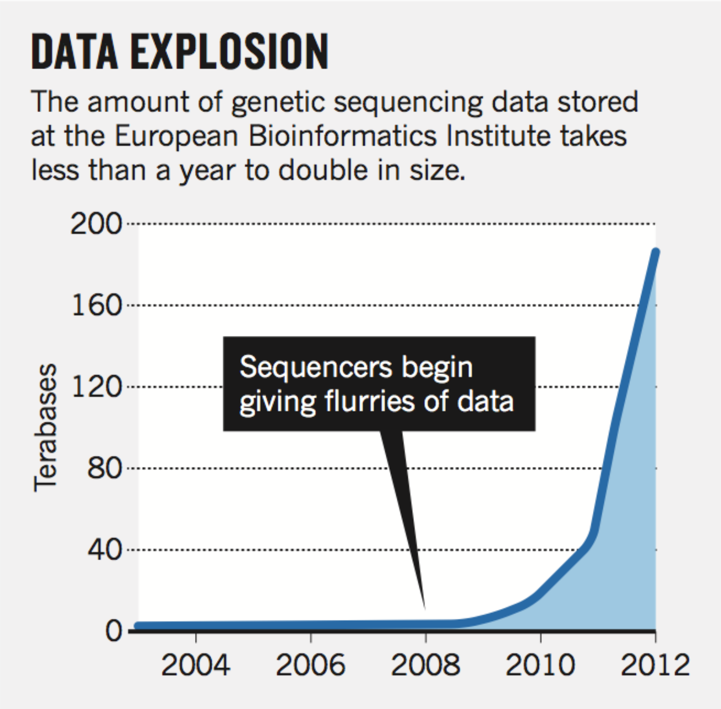
Protein family analysis

# Why is Computing and Mathematics necessary to solve bio-medical problems?

The big change: New technology allows biologists to perform experiments much more efficiently (using complex machines).

- This provides a growing amount of information/data from experiments.
- The data has to be analyzed in a hopefully efficient way.

The European Bioinformatics Institute (EBI) in Hinxton, UK,
currently stores **20 petabytes** (1 petabyte is $10^{15}$ bytes) of data
and back-ups about genes, proteins and small molecules.



**DATA EXPLOSION**

The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.

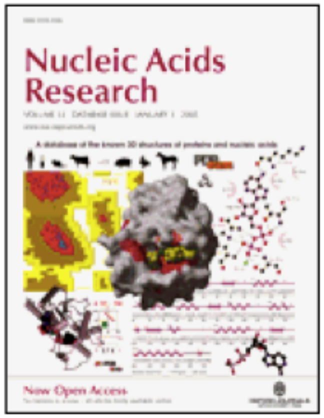
Sequencers begin giving flurries of data

# Tools

**1996**: first annual compilation of databases and tools lists **57 databases and tools**

**2000**: **230 databases and tools** listed in compilation

**2006: 856 databases and tools**



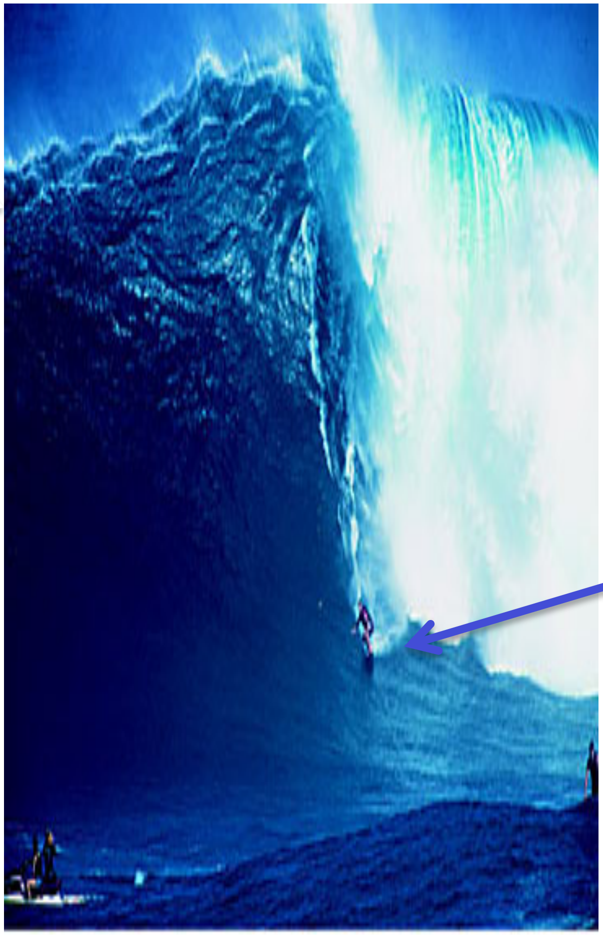
**2010: 1230 databases and tools**

**The annual database issue of Nucleic Acids Research (NAR) has grown exponentially**

The online 2011 NAR Database Collection lists **1330** molecular biology databases
http://www.oxfordjournals.org/nar/database/a/

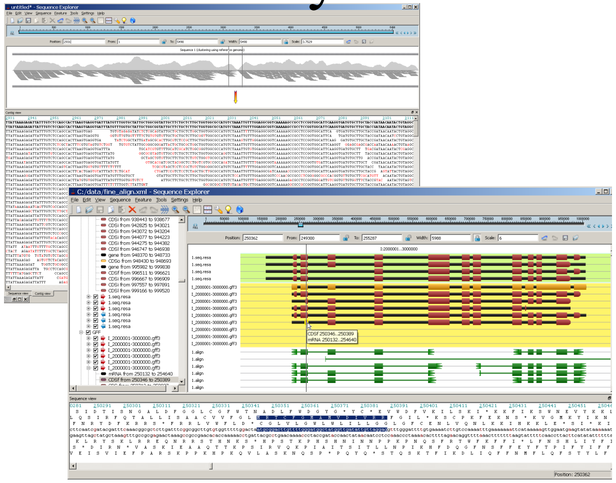
# Data exceeds analysis

Data



Bioinformatician

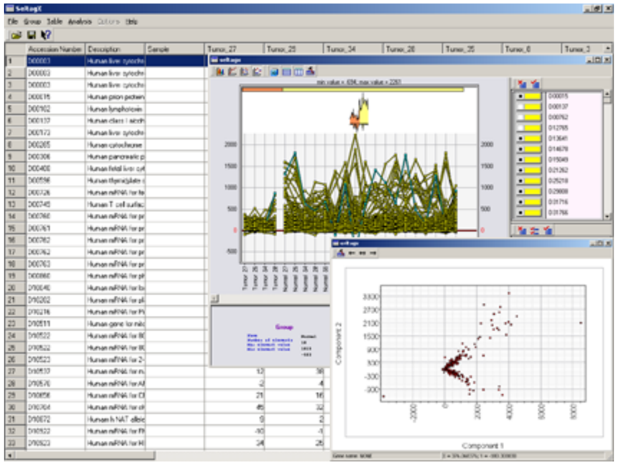
days-months



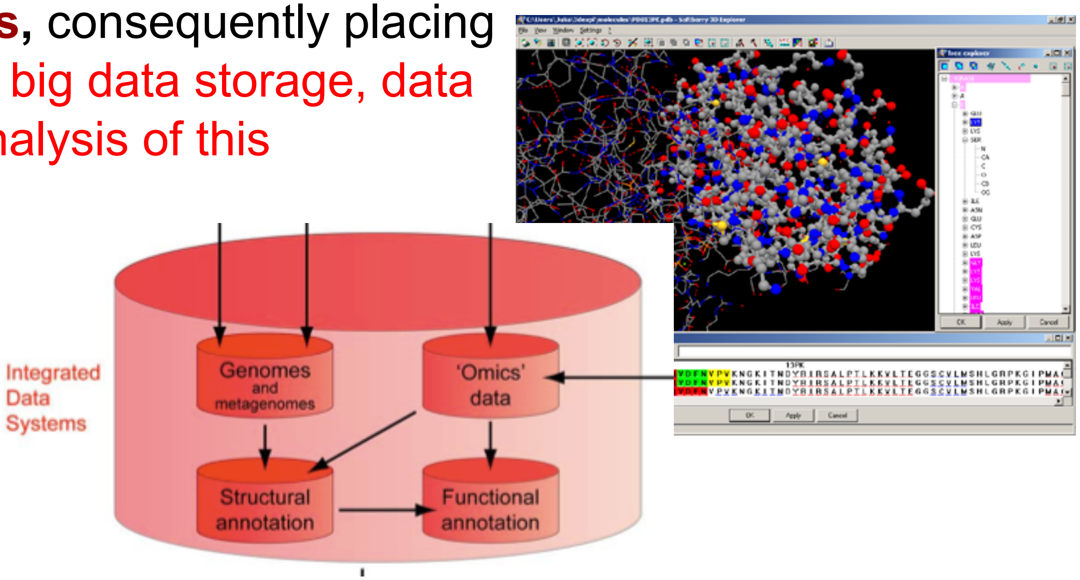
1-2 days

A tsunami of NGS data:

High-throughput experimental technique created vast amounts of biological data



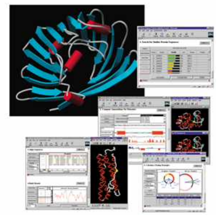
**Digging out the “treasure” from massive biological data represents the primary challenge in bioinformatics,** consequently placing unprecedented demands on big data storage, data manipulation and efficient analysis of this information.



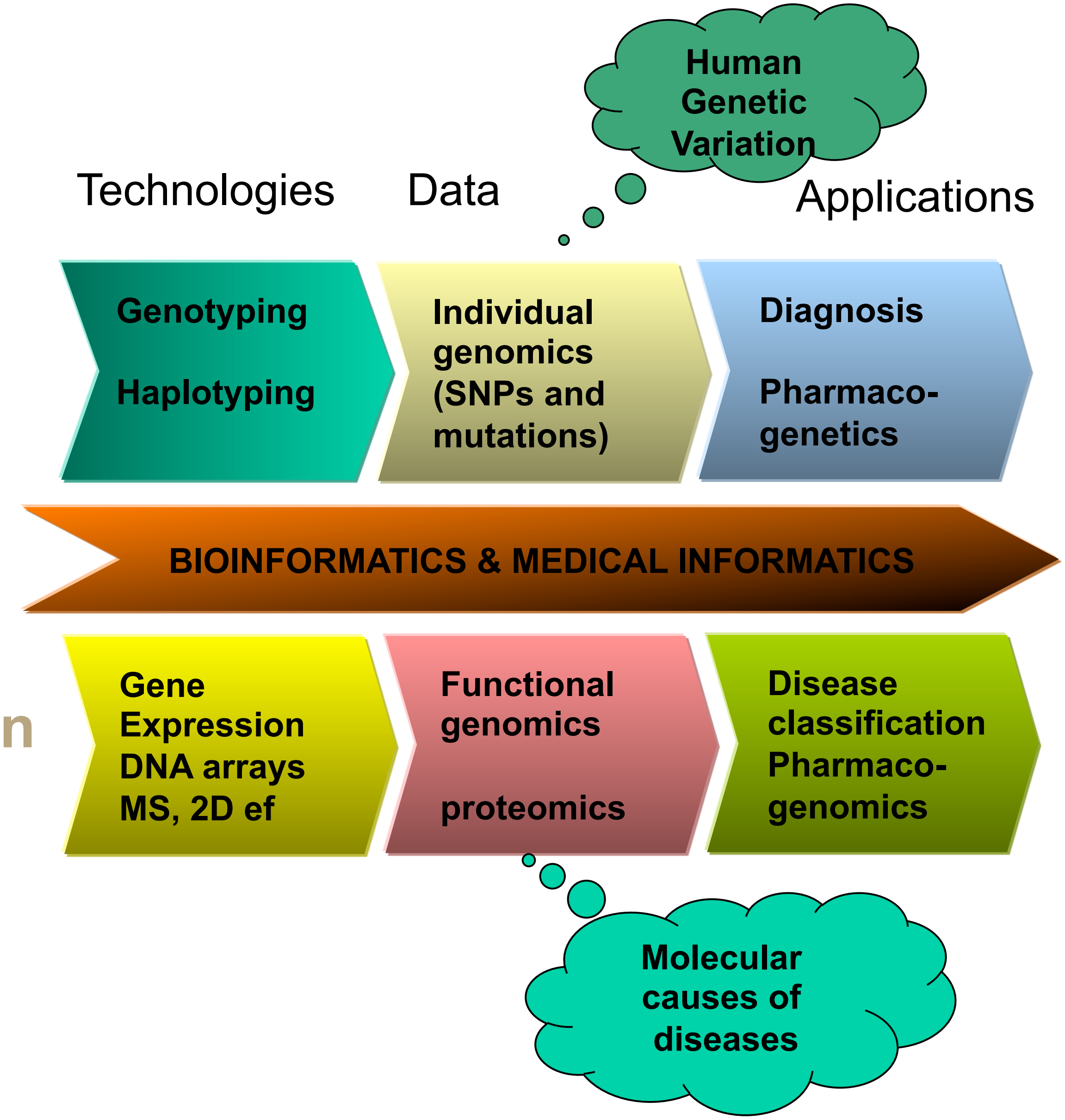
Biologists are increasingly finding that the management of complex data sets is becoming a bottleneck for scientific advances. Therefore, **bioinformatics** is rapidly become a key technology in all fields of biology.

# Bioinformatics and Medicne

Genome



Information



Technologies

Data

Human Genetic Variation

Applications

Genotyping

Haplotyping

Individual genomics (SNPs and mutations)

Diagnosis

Pharmaco-genetics

BIOINFORMATICS & MEDICAL INFORMATICS

Gene Expression DNA arrays MS, 2D ef

Functional genomics

proteomics

Disease classification Pharmaco-genomics

Molecular causes of diseases

Individualised healthcare

Molecular medicine