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

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

Akademgorodok,
Novosibirsk

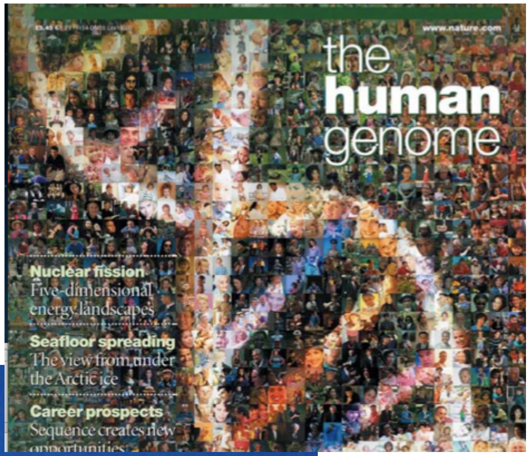
Novosibirsk State Univers

N
Map
Photos
US Dept of State Geographer

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

Baylor College of Medicine, Huston

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

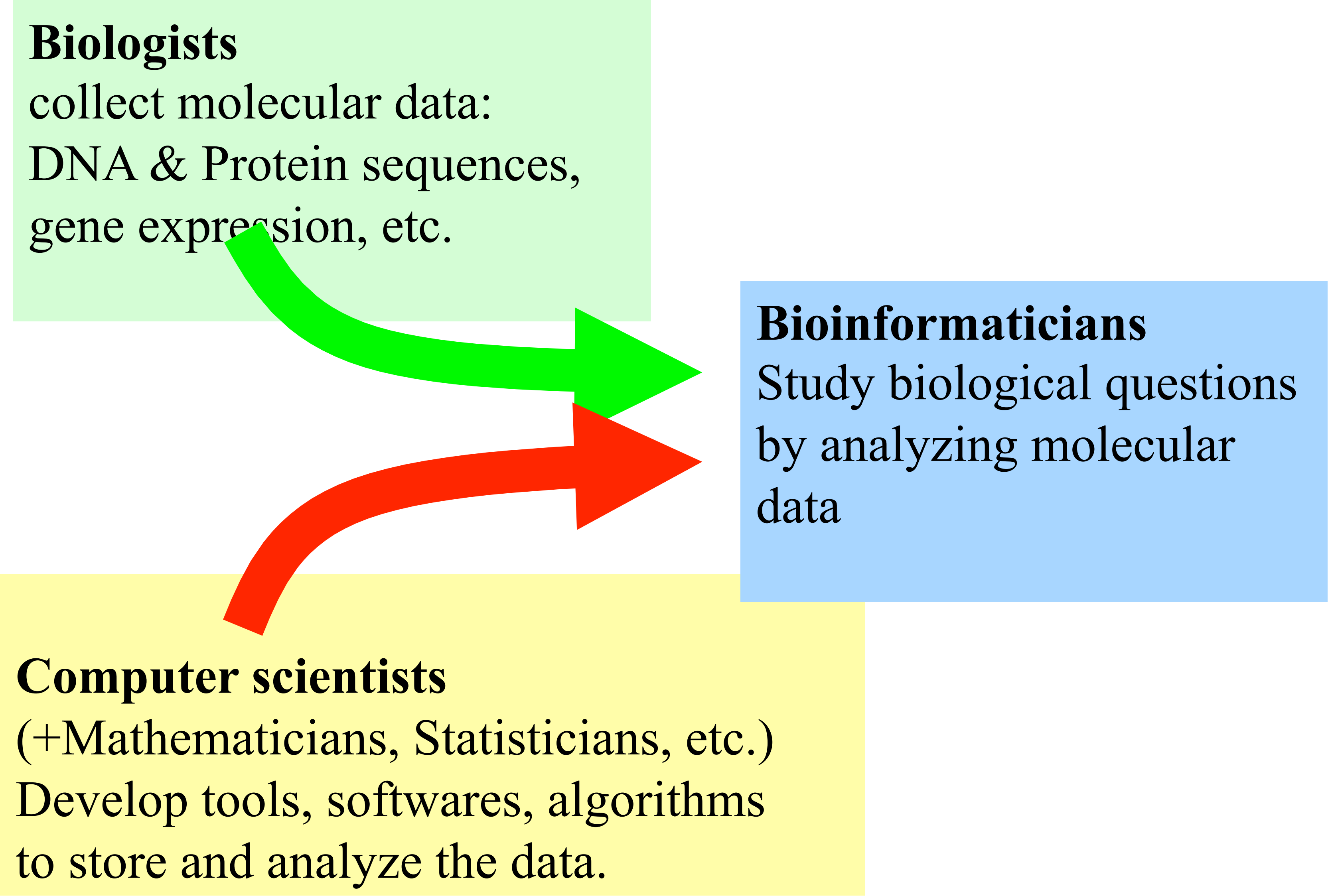
Joint Genome Institute, Berkeley National Lab. California

Genome annotation group

Royall Holloway, University of London

KAUST (Saudi Arabia)

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

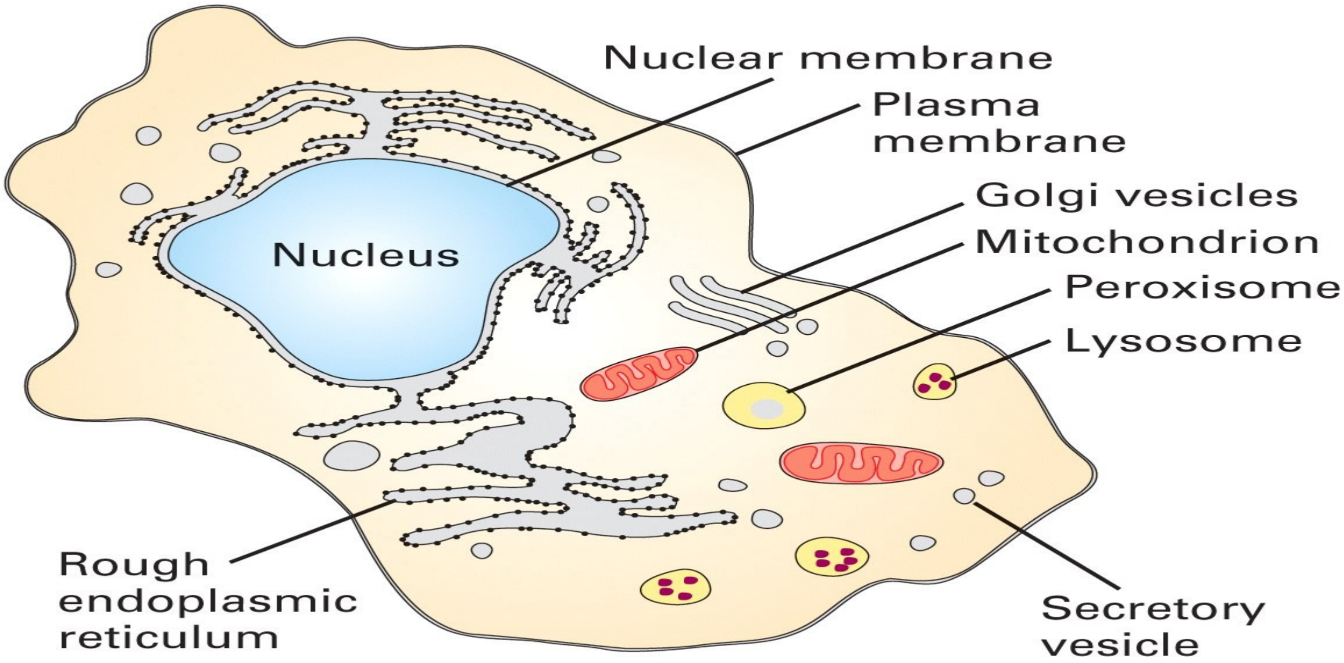


Biologists
collect molecular data:
DNA & Protein sequences,
gene expression, etc.

Bioinformaticians
Study biological questions
by analyzing molecular
data

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

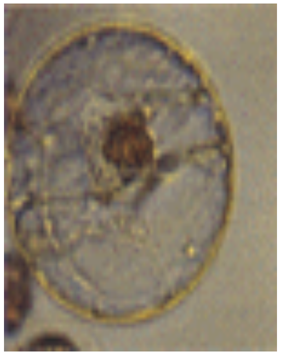
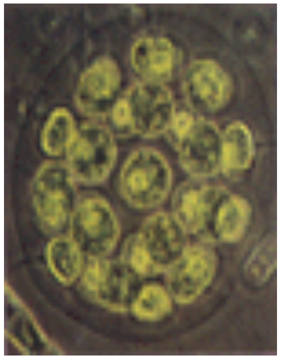
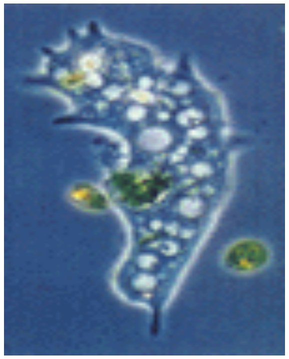
Life begins with the cell



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

Cell Information and Machinery

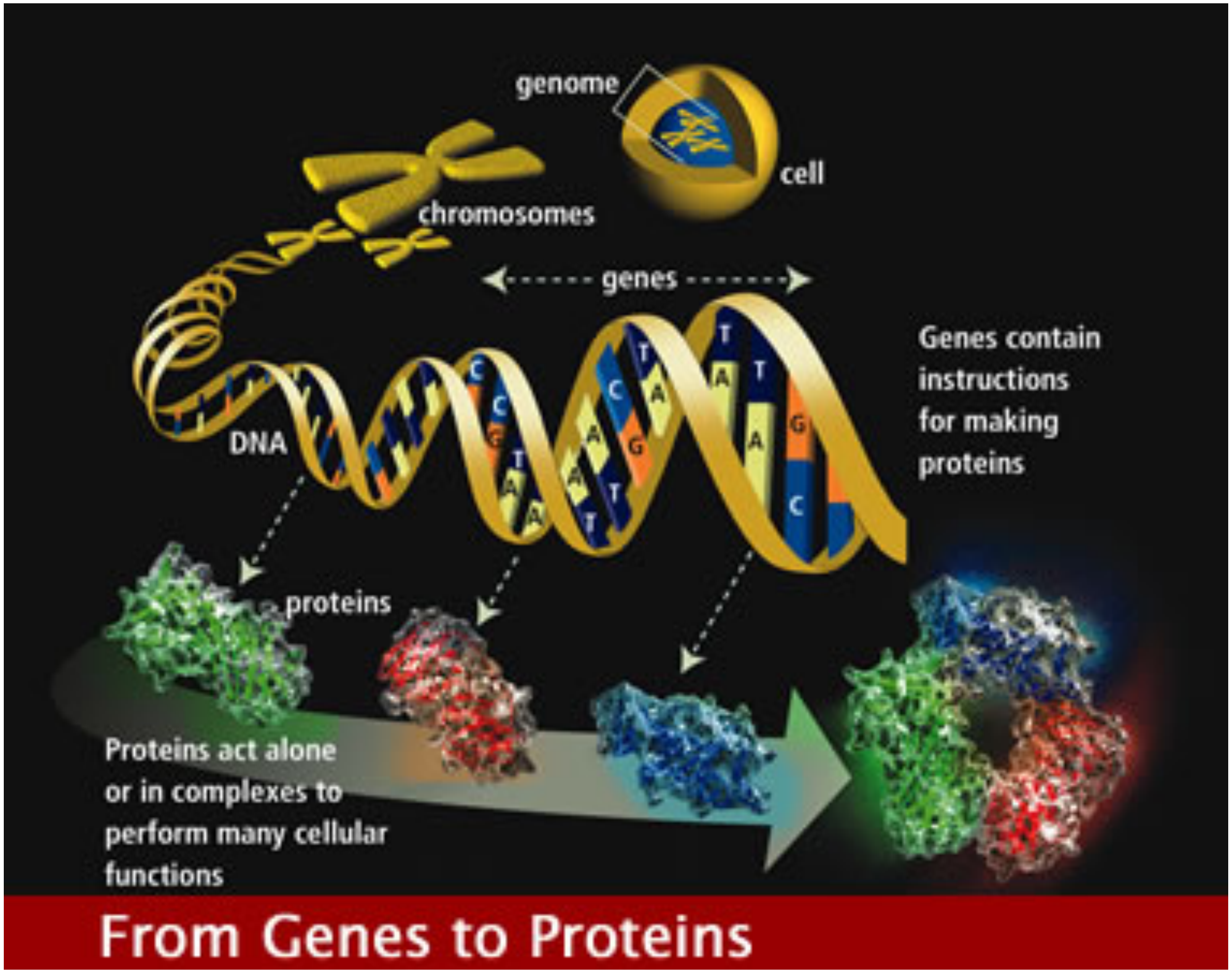
- A cell stores all information to replicate itself
  - Human genome is around 3 billion base pairs long
  - Almost every cell in human body contains same set of genes
  - But not all genes are used or expressed by those cells
- Machinery:
  - Collect and manufacture components
  - Carry out replication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

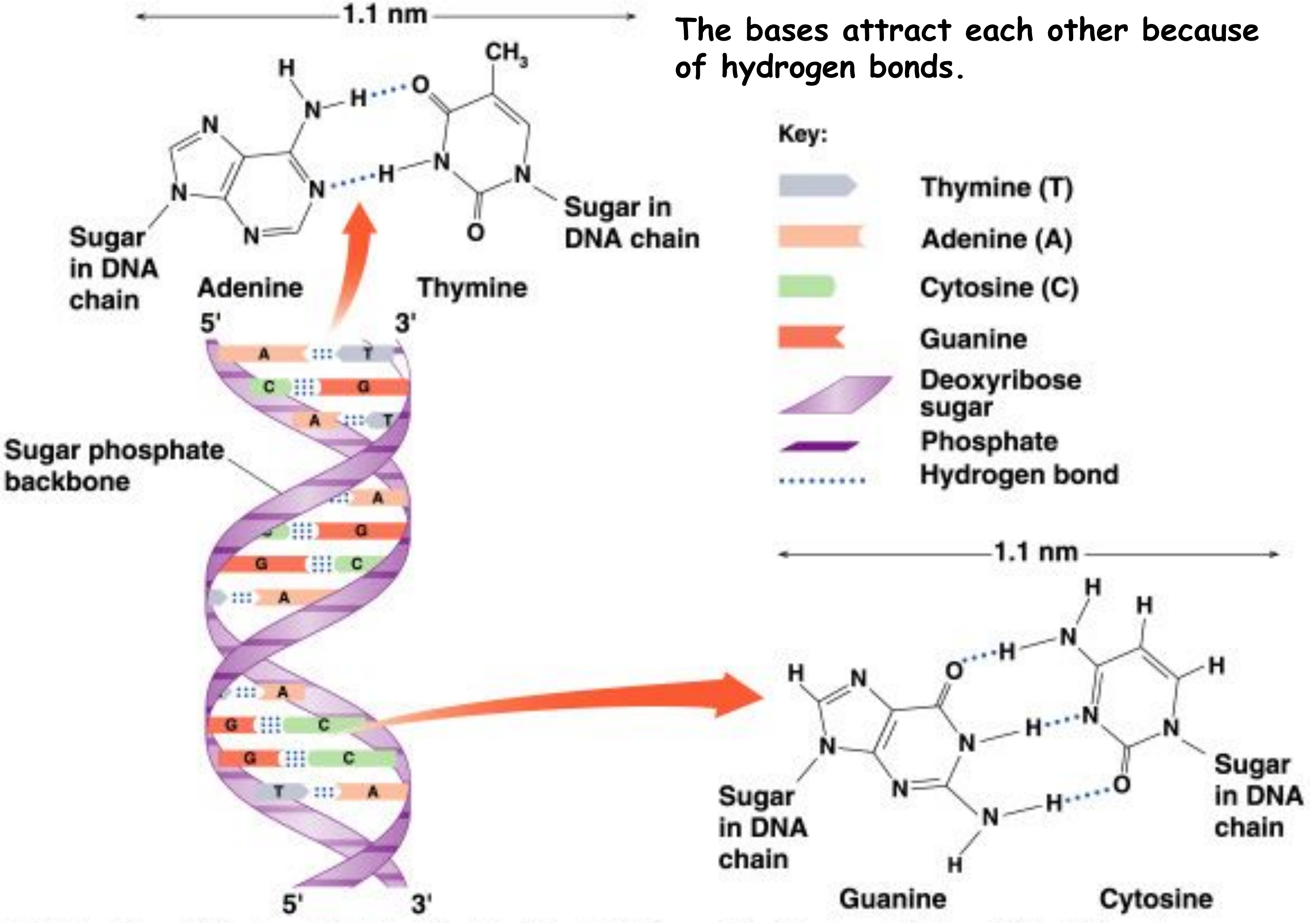

genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
From Genes to Proteins

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.**

Base Pairing in the DNA Double Helix

The bases attract each other because of hydrogen bonds.



Chemical structure DNA

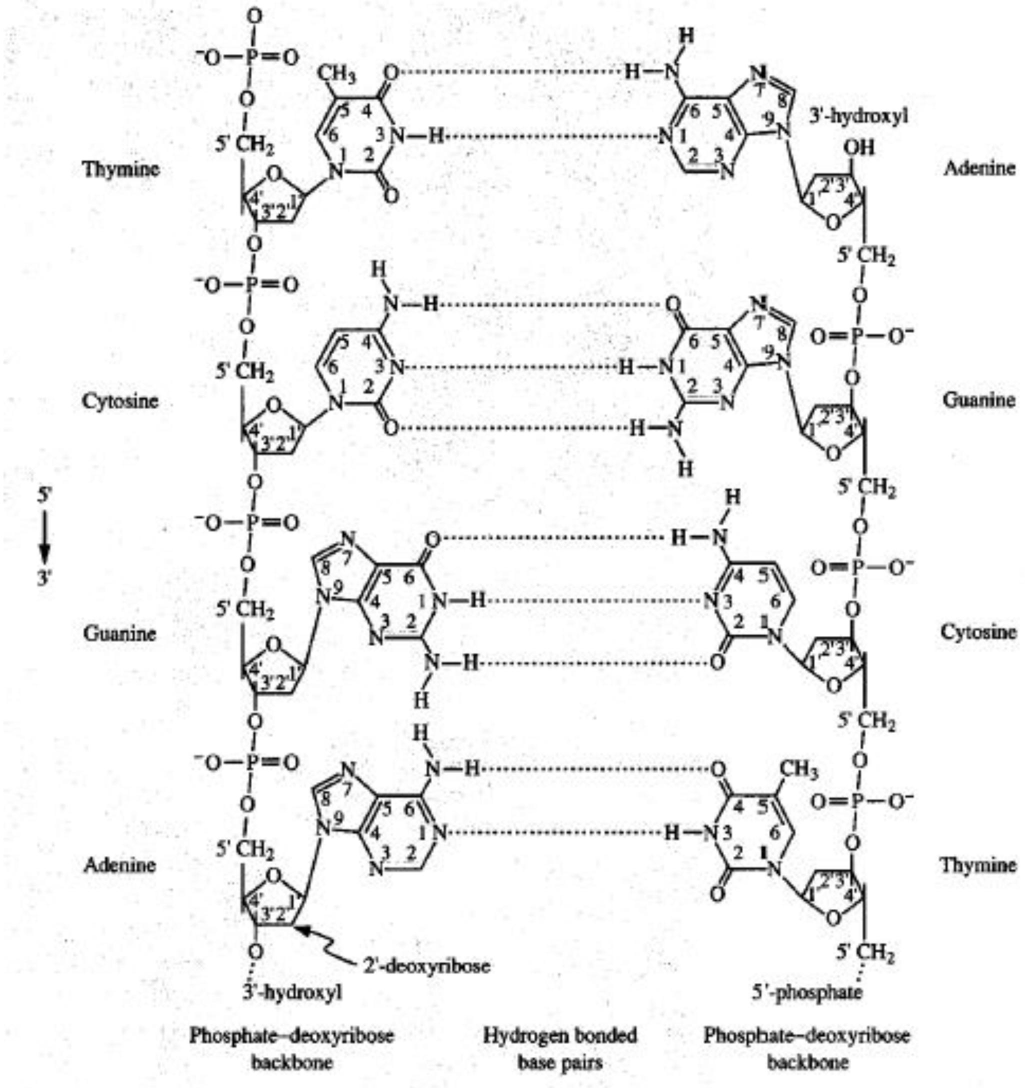
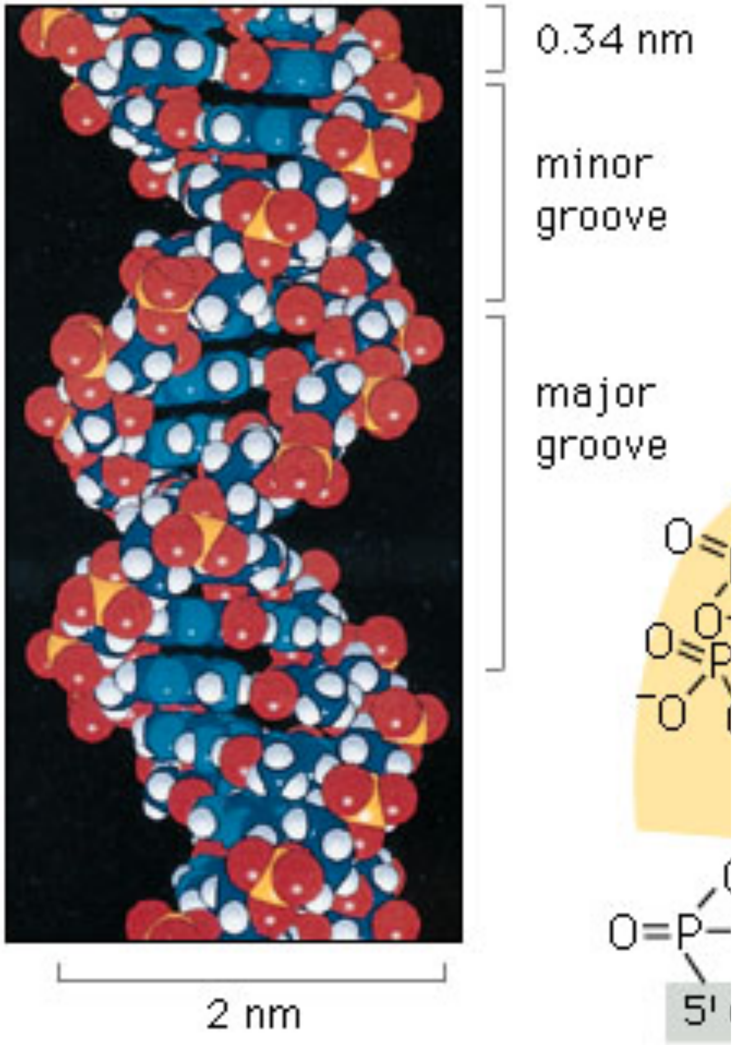
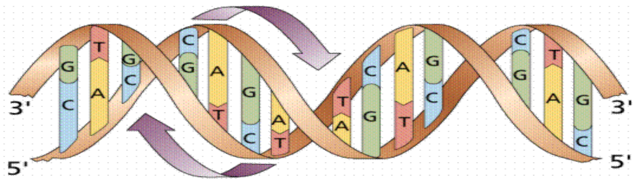


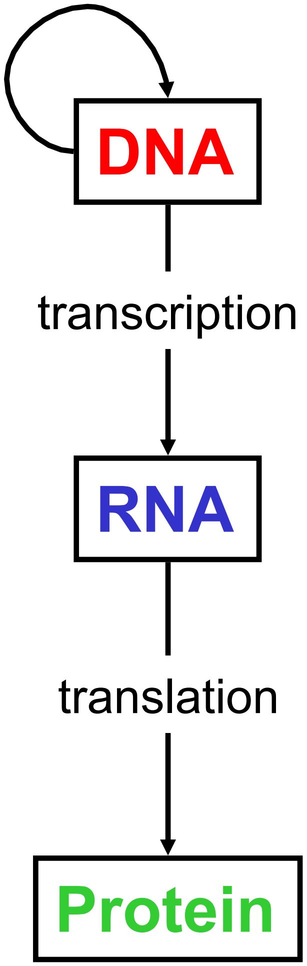
Fig. 1.2 Chemical structure and base pairing in double-stranded DNA.



The Central Dogma of Biology

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

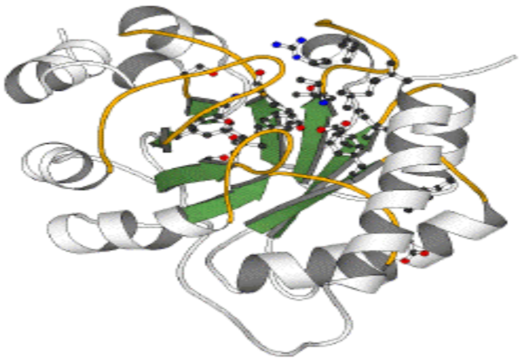




CCTGAGCCAACTATTGATGAA

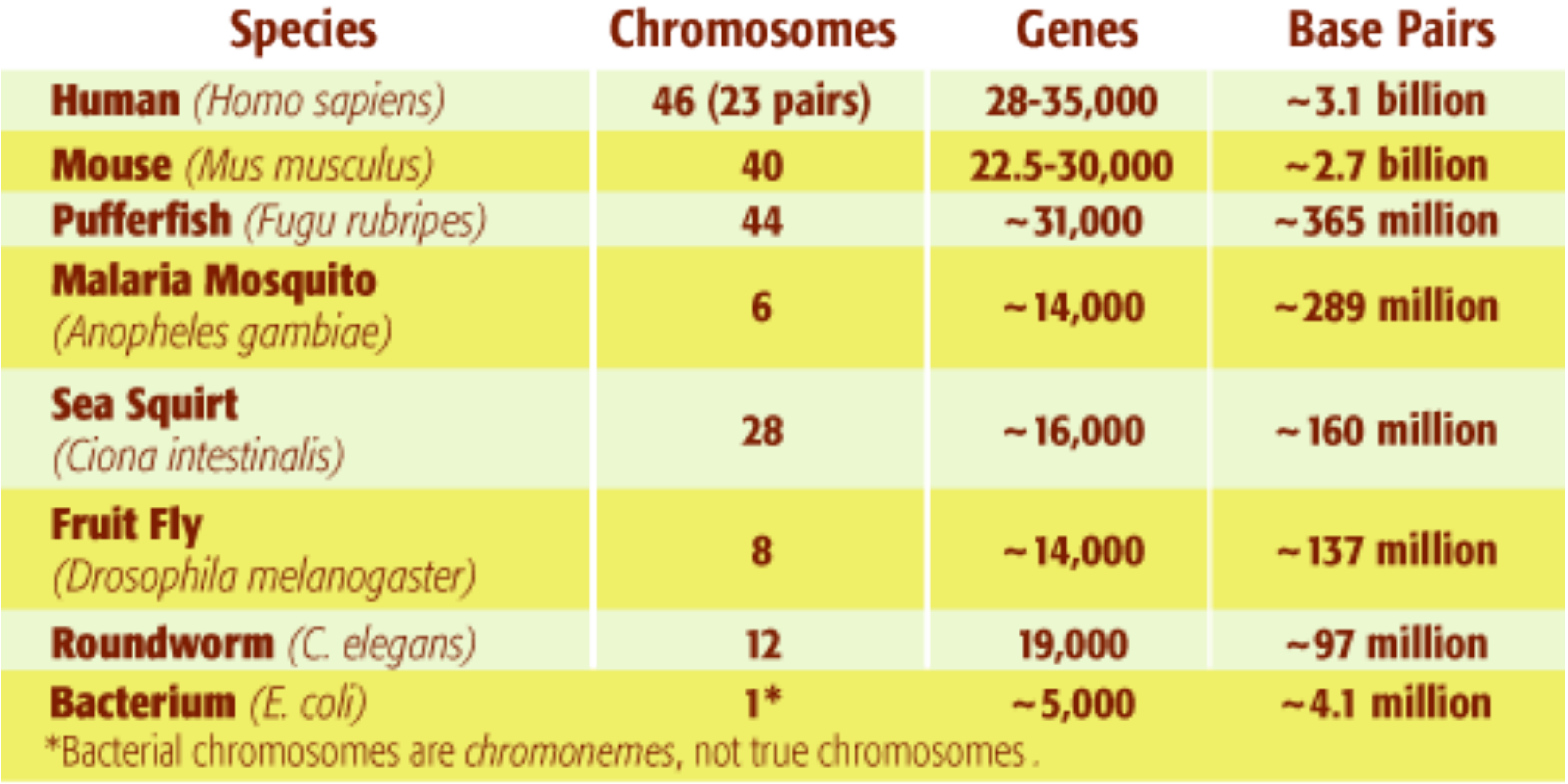
CCUGAGCCAACUAUUGAUGAA

PEPTIDE



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

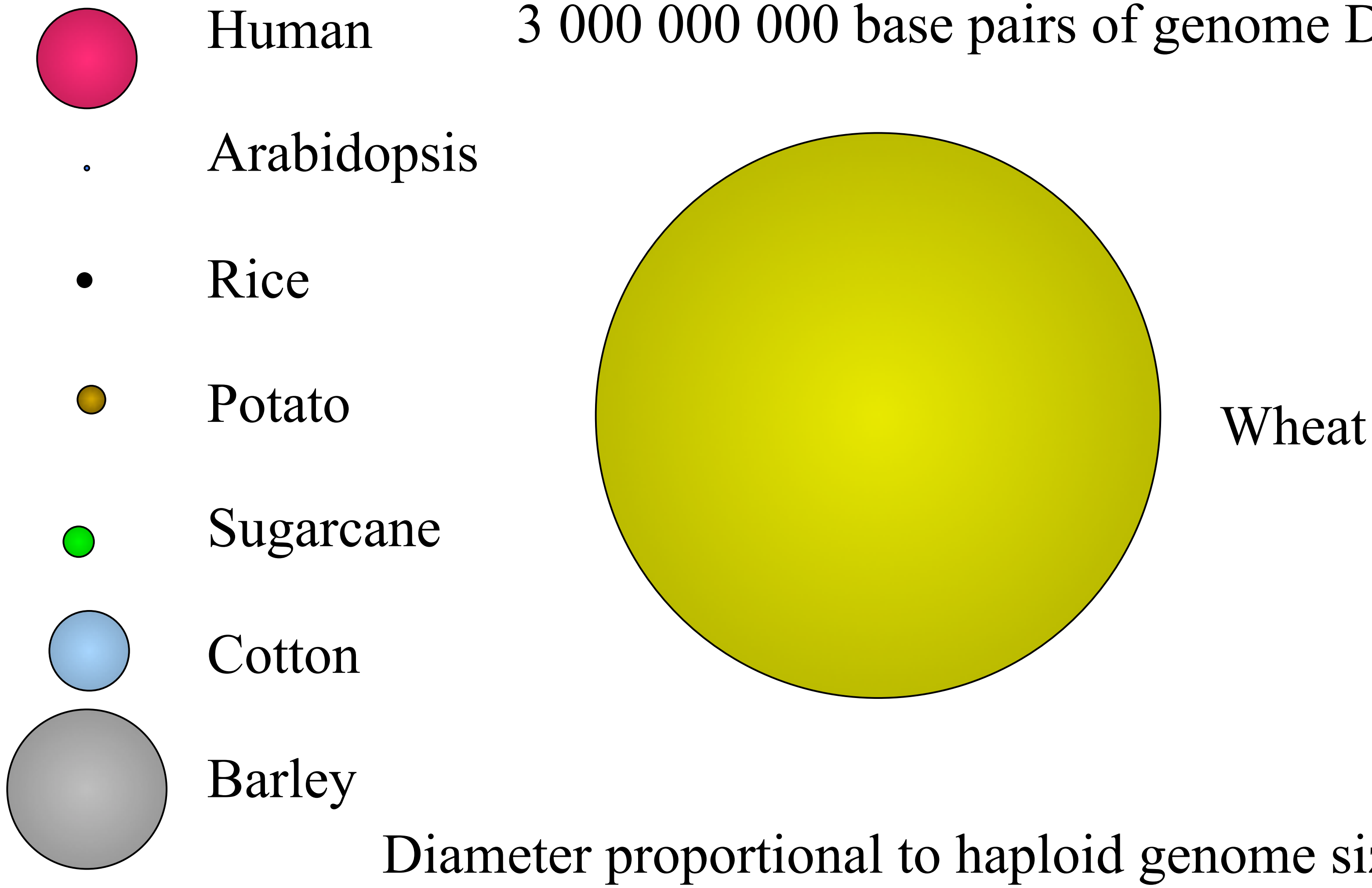
Genome sizes



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.

Genome size



Human

3 000 000 000 base pairs of genome DNA

Arabidopsis

Rice

Potato

Sugarcane

Cotton

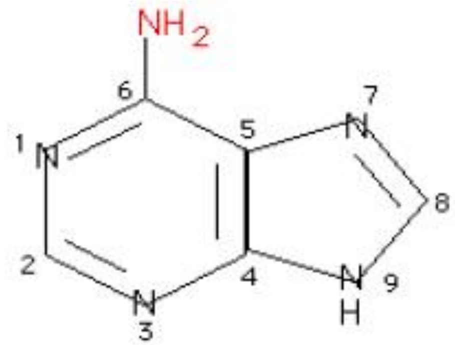
Barley

Wheat

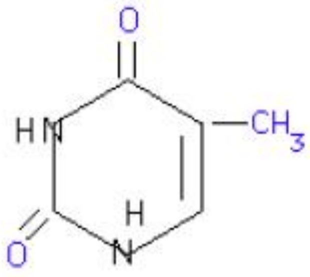
Diameter proportional to haploid genome size

Nitrogenous bases commonly found in RNA and DNA

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

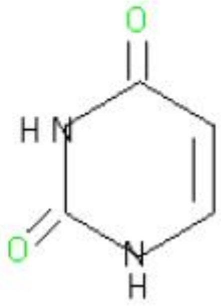


Adenine



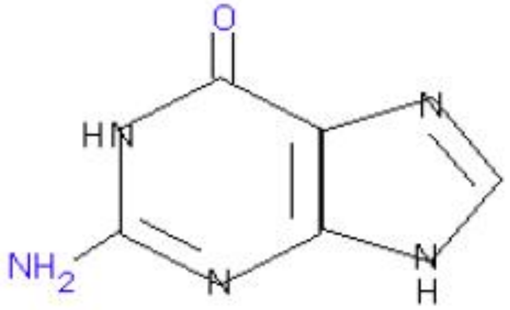
Thymine

T ----→ U

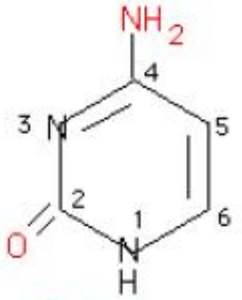


Uracil

DNA (AT GC)



Guanine



Cytosine

A-T (A-U) G=C

Complementary pairs

Hierarchical organization of RNA molecules

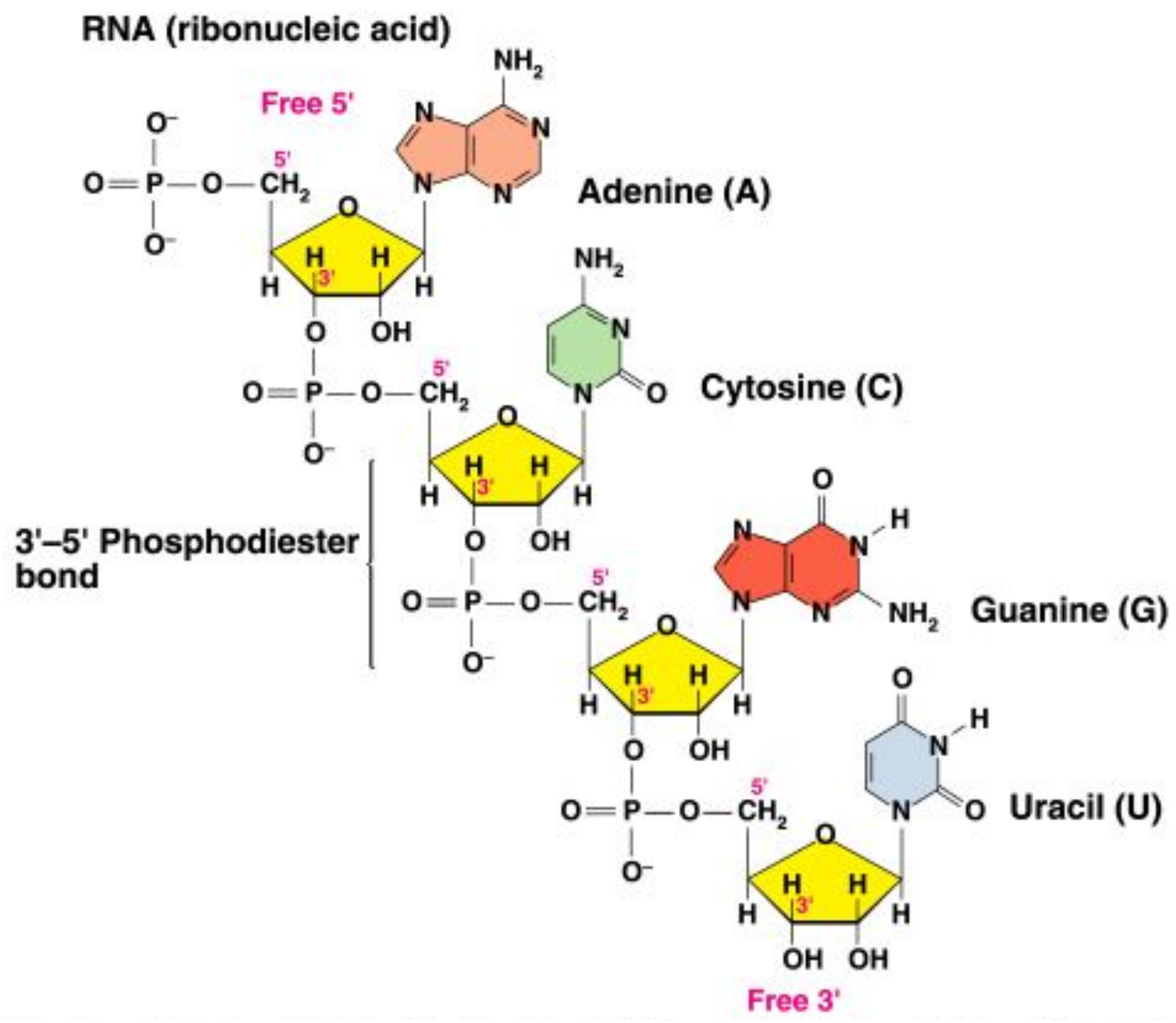
Primary structure:

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

•Commonly represented by a string S over the alphabet $\Sigma=\{A,C,G,U\}$

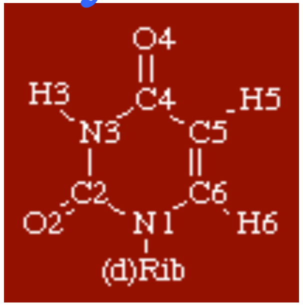
Example of RNA Primary Structure

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

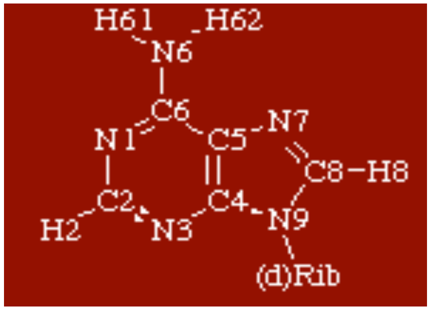


RNA synthesis and fold

- RNA immediately starts to fold when it is synthesized

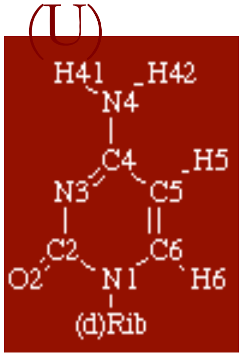


Uracyl (U)

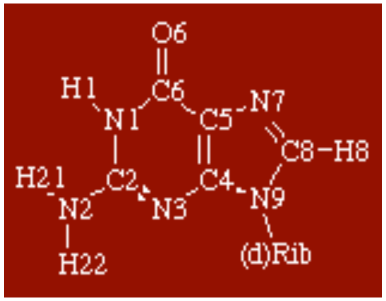


Adenine (A)

Wobble Base Pairing



Cytosine (C)

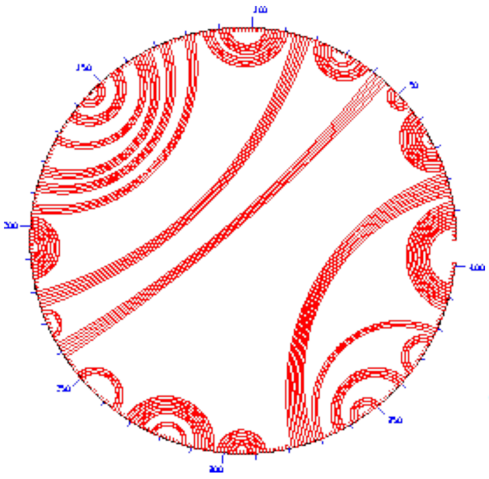


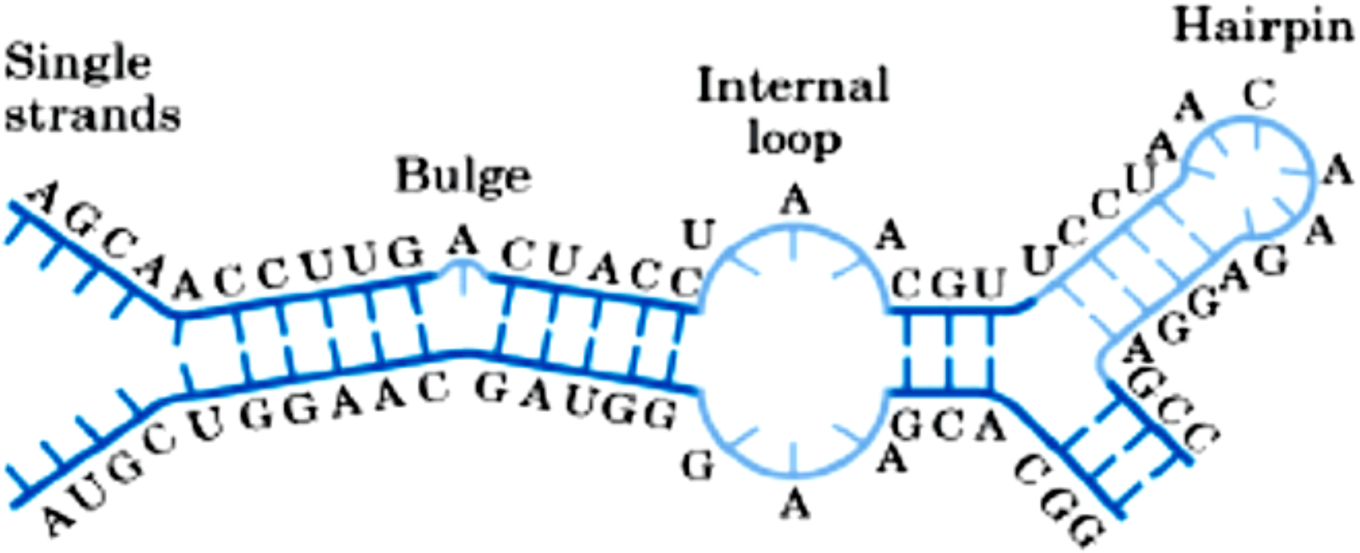
Guanine (G)

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.

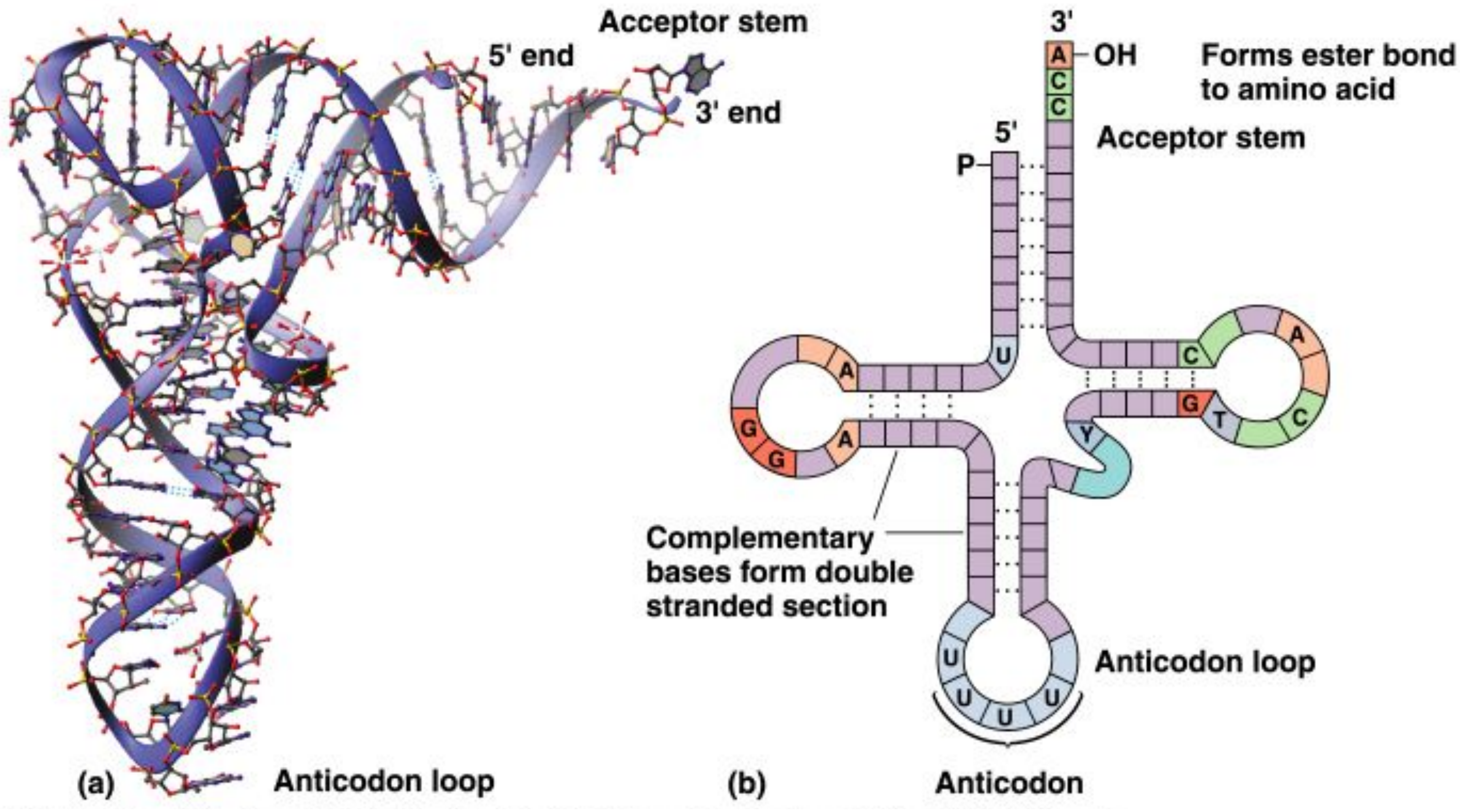




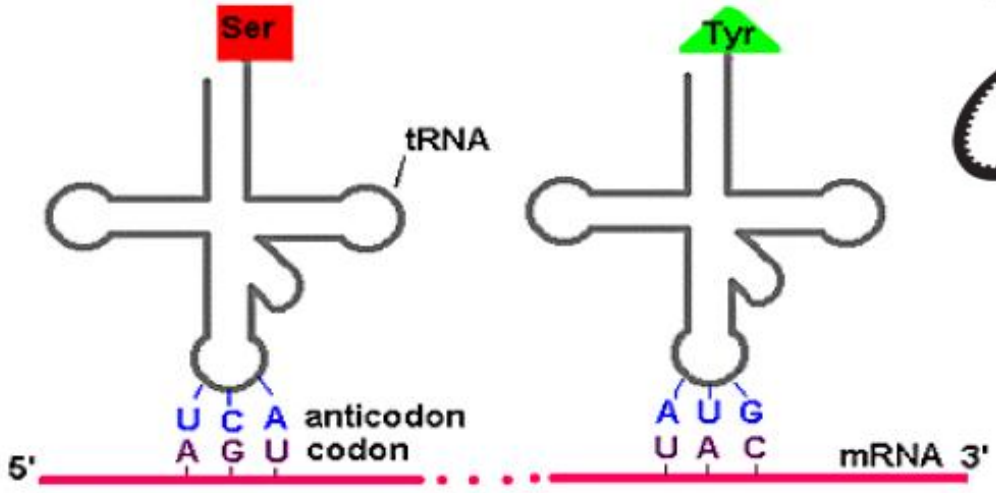
Transfer RNA

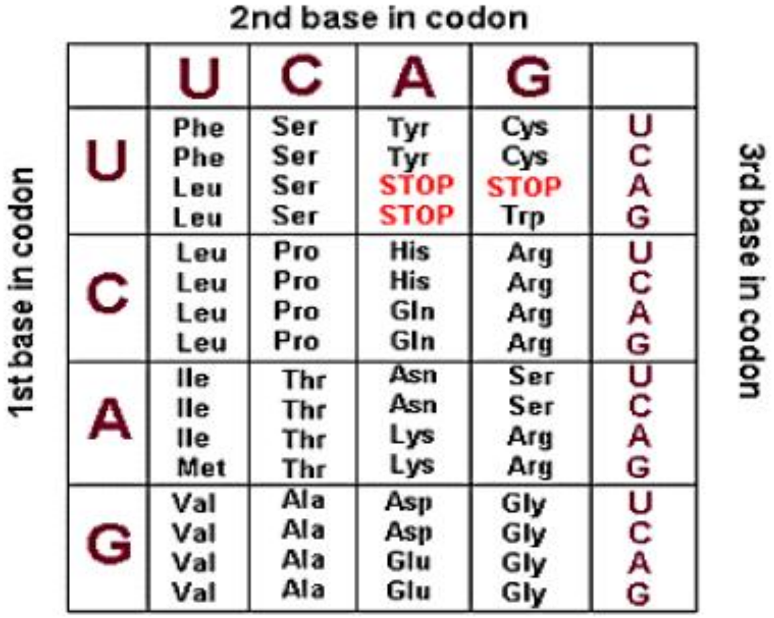
- tRNA has a tertiary structure that is L-shaped

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

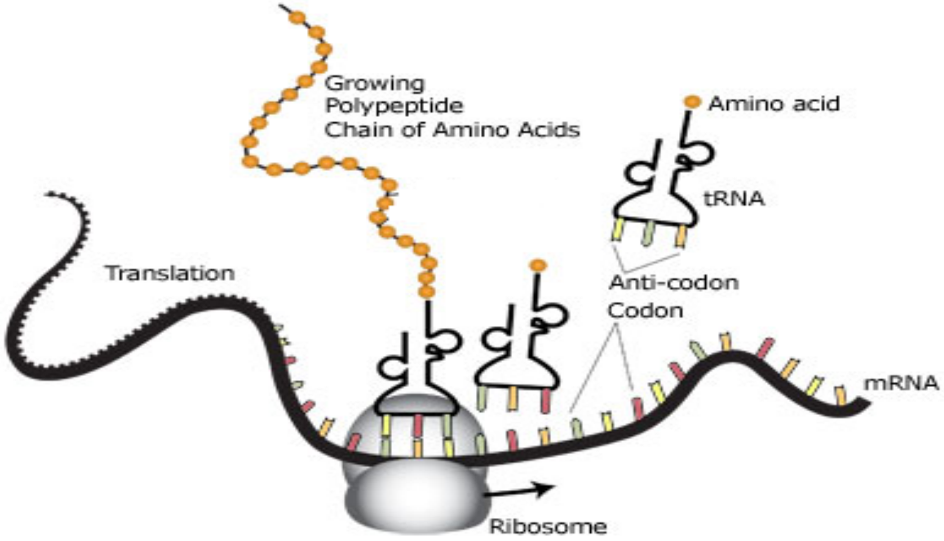


Genetic code

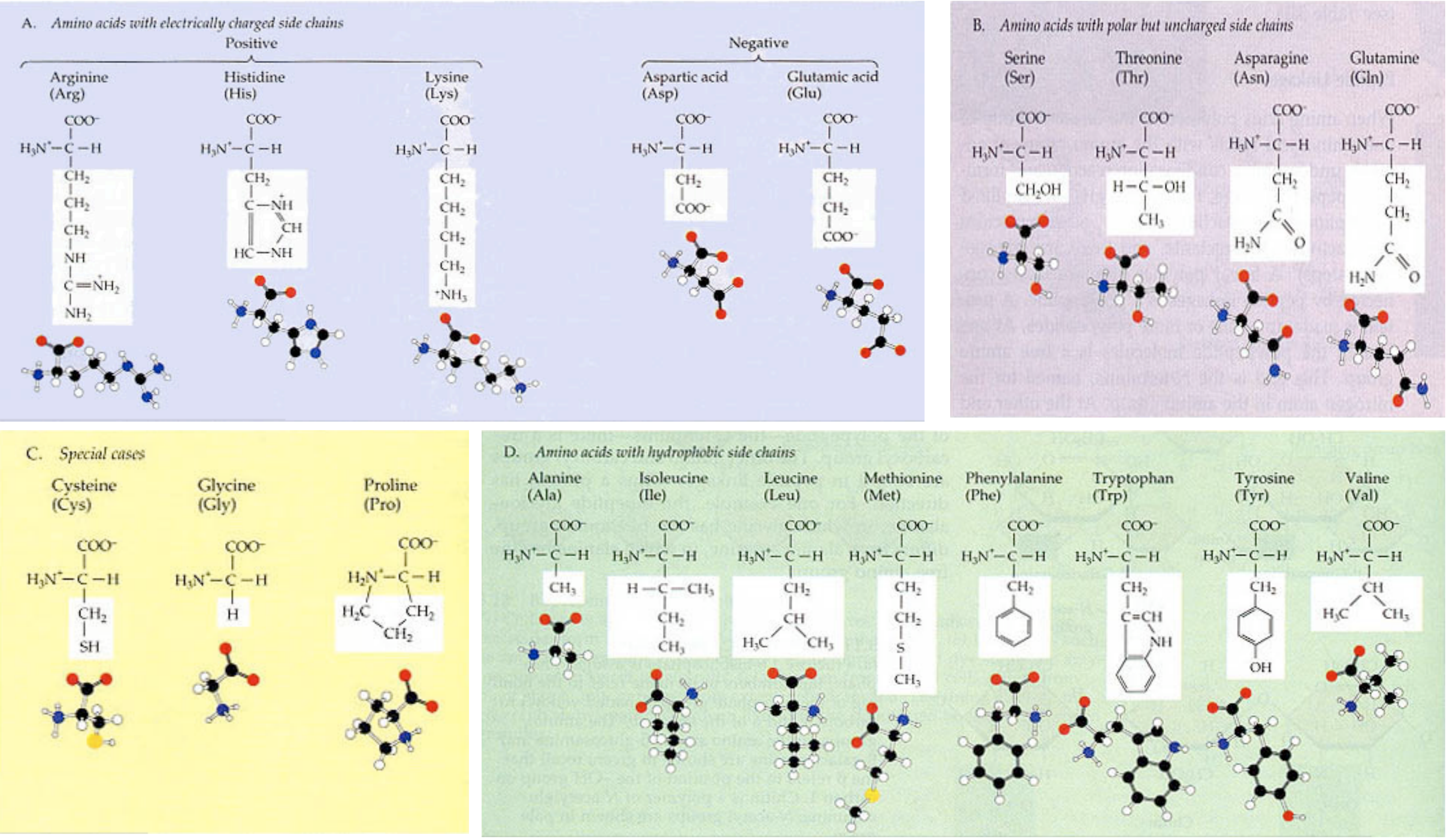



2nd base in codon

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



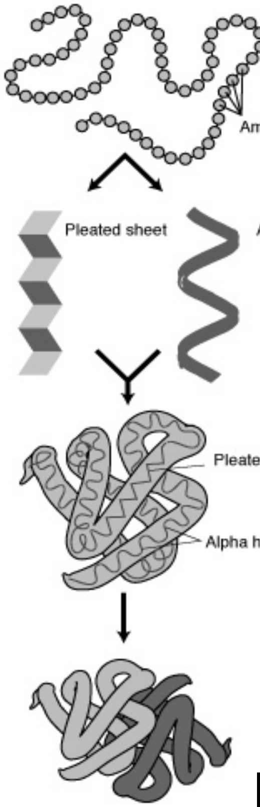
Amino acids - The protein building blocks



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

Amino Acids

Pleated sheet

Alpha helix

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

Pleated sheet

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

Alpha helix

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

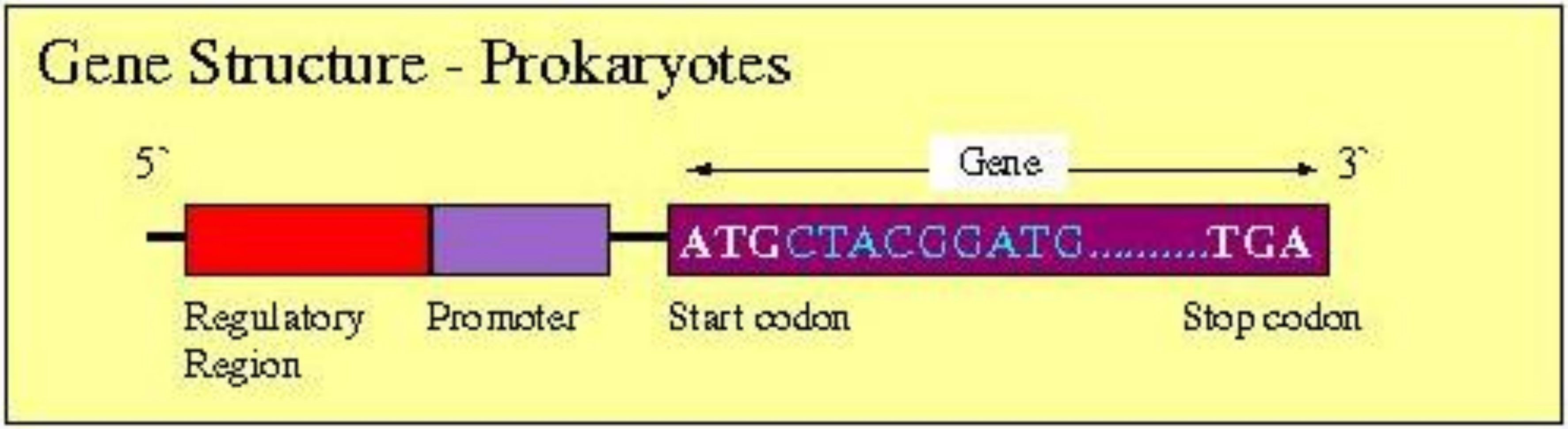
Molecular Bioinformatics involves the use of computational tools to discover new information in complex data sets (from the one-dimensional information of DNA through the two-dimensional information of RNA and the three-dimensional information of proteins, to the four-dimensional information of evolving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



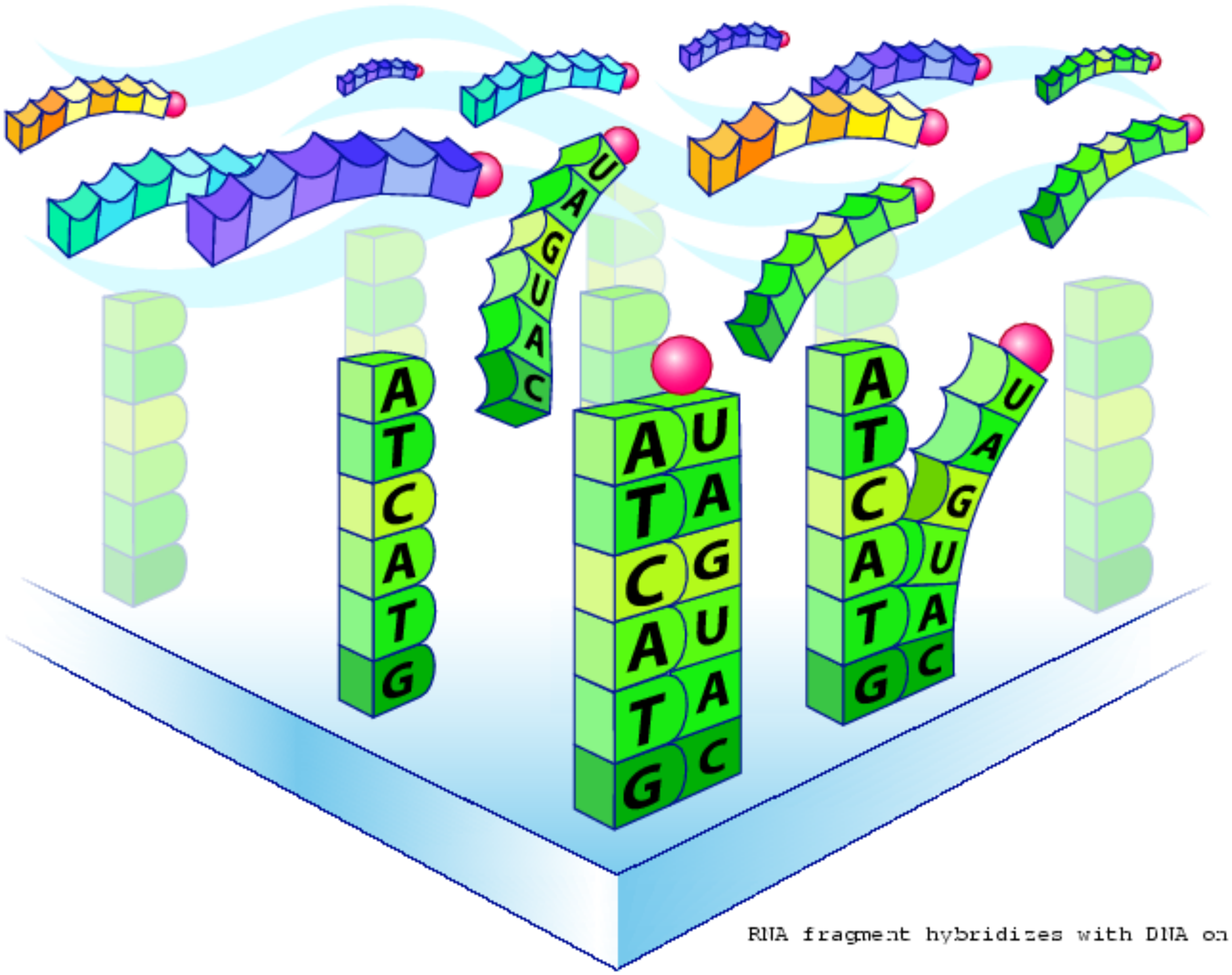
Gene Expression

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

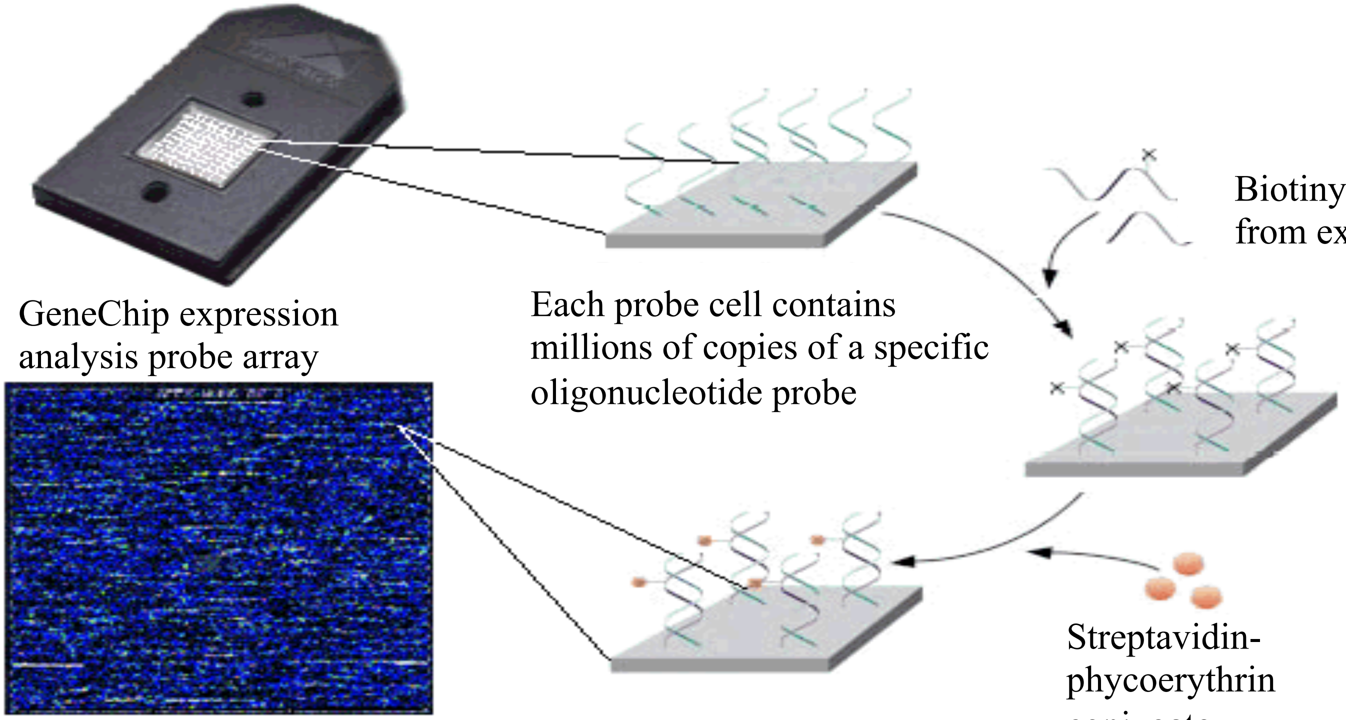


RNA fragment hybridizes with DNA on GeneChip

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

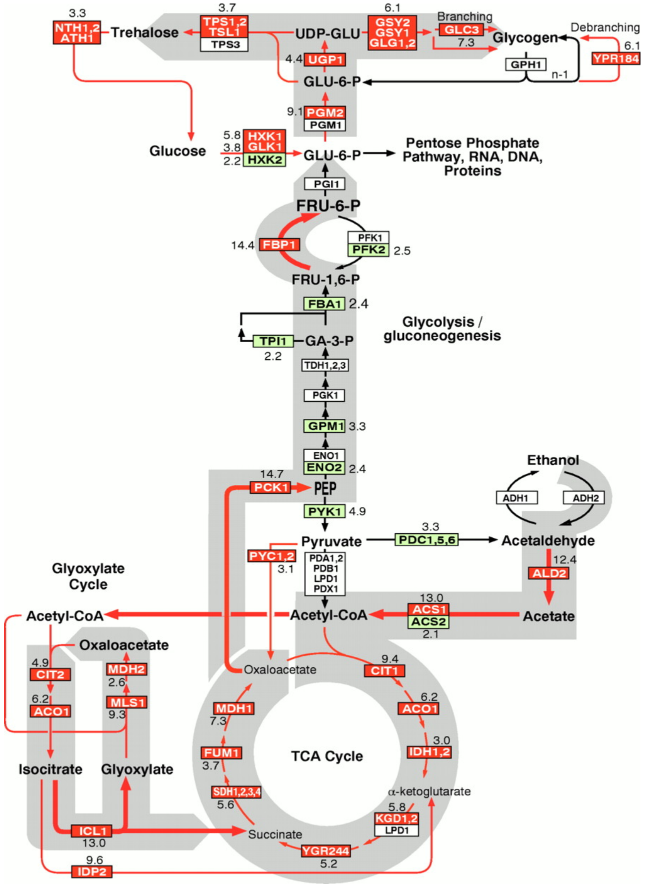
Streptavidin-phycoerythrin conjugate

Image of hybridized probe array

Information Derivable from Chip Data

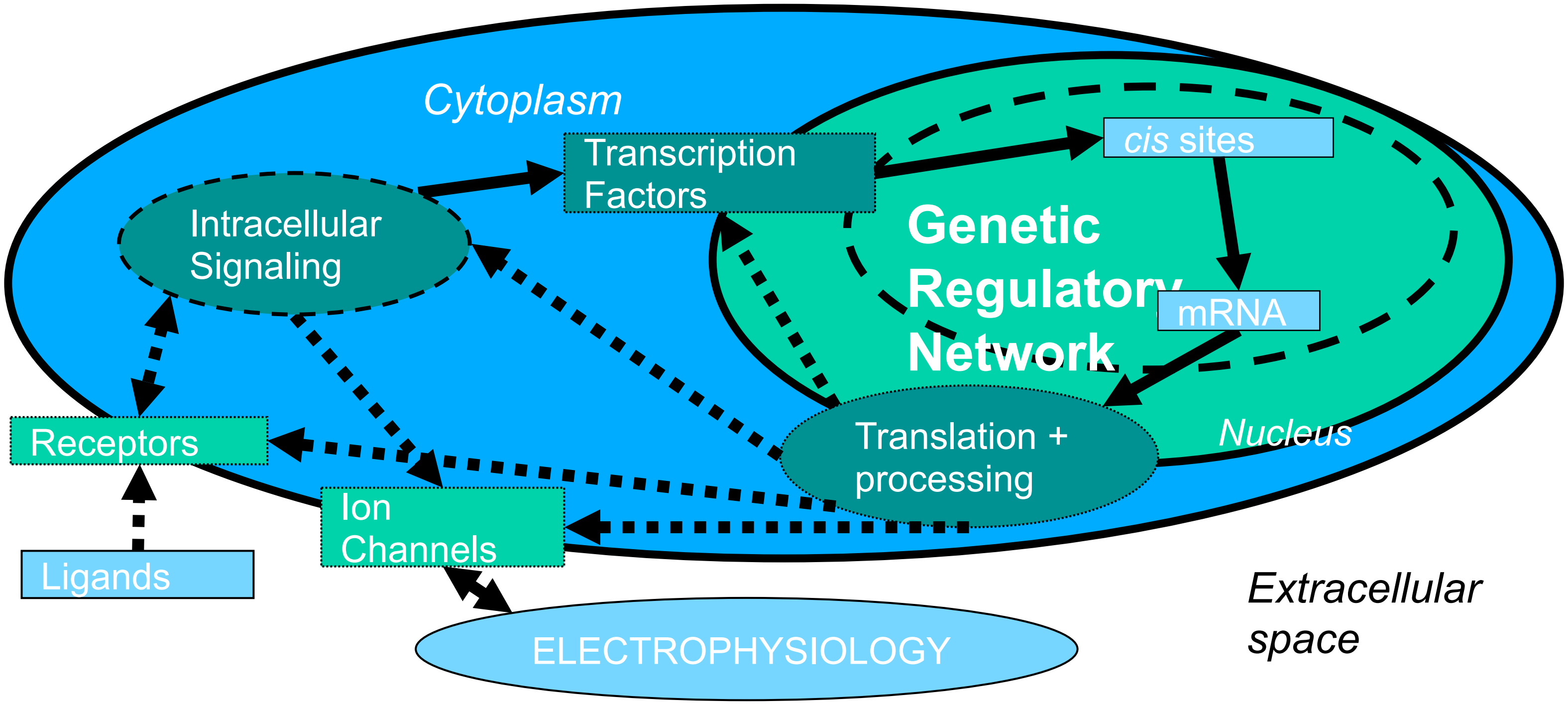
- Microarray data is becoming a key source of data for computational inference of biological networks
  - who interact with who
  - who regulate who
  - ….

How does this work?

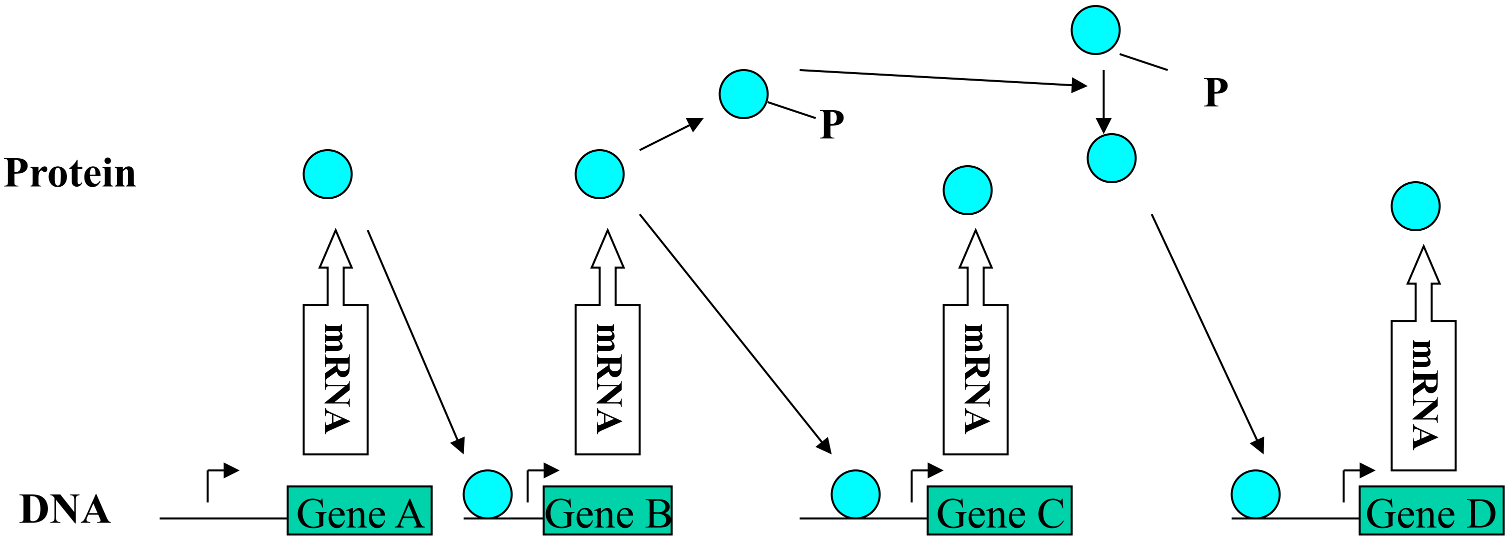


Genetic Regulatory Network

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



Microarray analysis model using gene expression profiles



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

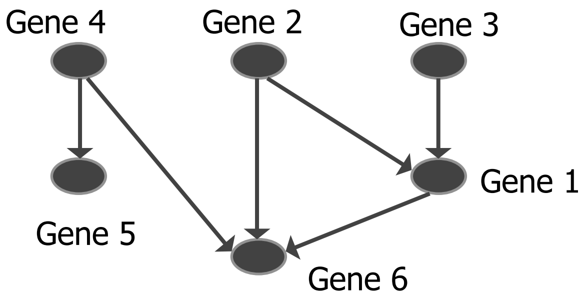
E X P matrix

	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

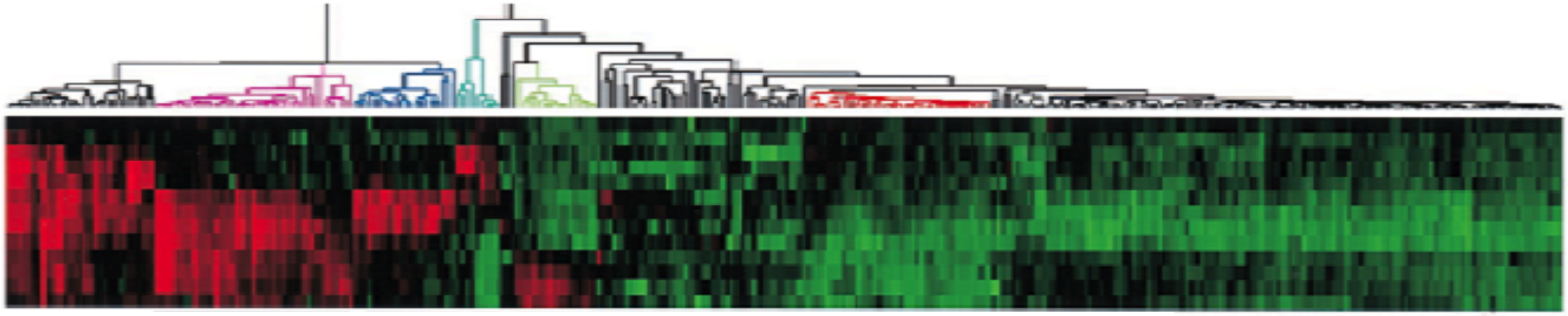
1. mRNA expression is only a partial picture
2. the number of sample is much smaller than the number of genes
3. high noise

Clustering

✓ Grouping genes with similar patterns of expression

Common role gene clustered together

Uncharacterized gene function guessed

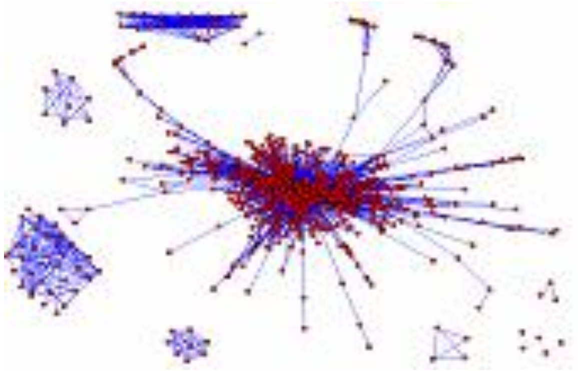


Similarity measure : standard correlation coefficient, ..

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

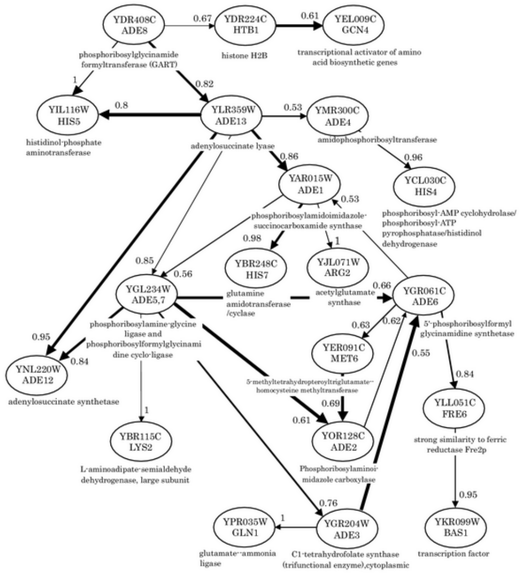
Can’t reveal the inner interaction structure !

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
0.96
YAR015W ADE1
phosphoribosylamidoimidazole-succinocarboxamide synthase
0.53
YCL030C HIS4
phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase/histidinol dehydrogenase
0.98
1
0.85
0.56
YBR248C HIS7
glutamine amidotransferase /cyclase
YJL071W ARG2
acetylglutamate synthase
0.66
YGL234W ADE5,7
phosphoribosylamine-glycine ligase and phosphoribosylformylglycinamidine cyclo-ligase
YGR061C ADE6
5'-phosphoribosylformyl glycinamidine synthetase
0.95
0.84
YNL220W ADE12
adenylosuccinate synthetase
0.63
0.62
YER091C MET6
5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase
0.55
0.84
YLL051C FRE6
strong similarity to ferric reductase Fre2p
1
YBR115C LYS2
L-aminoadipate-semialdehyde dehydrogenase, large subunit
0.69
0.61
YOR128C ADE2
Phosphoribosylaminoimidazole carboxylase
0.76
0.95
1
YPR035W GLN1
glutamate--ammonia ligase
YGR204W ADE3
C1-tetrahydrofolate synthase (trifunctional enzyme),cytoplasmic
YKR099W BAS1
transcription factor

Bayesian networks:

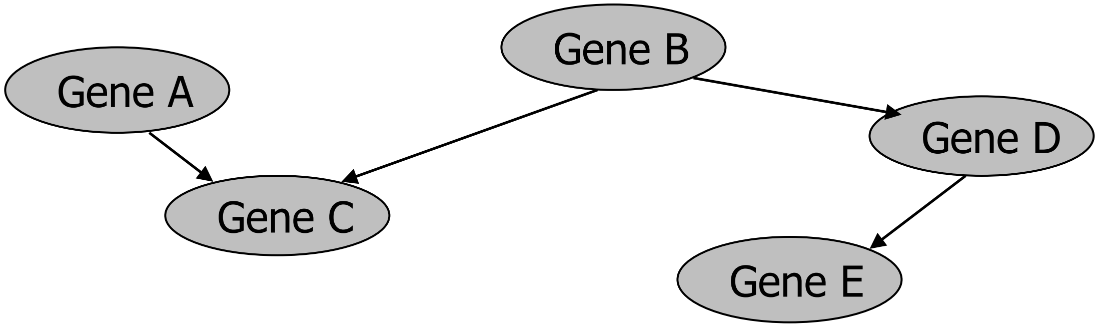
A directed, graphical representation of the probabilities of one observation given another. Nodes represent mRNA molecules; edges represent the probability of a particular expression value given the expression values of the parent nodes.

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

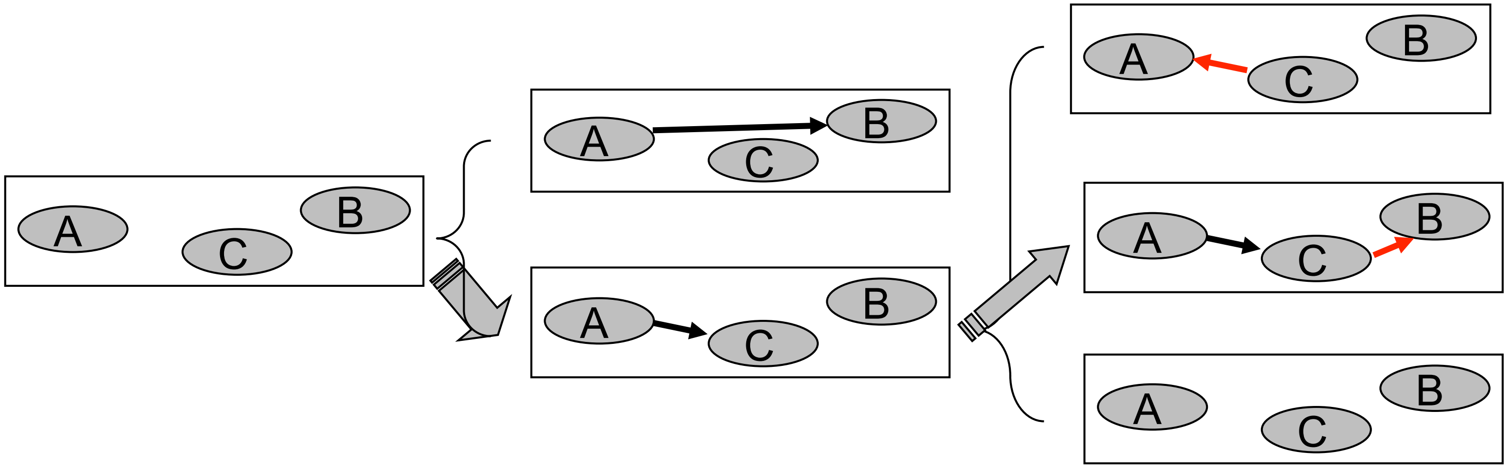


$$P(A, B, C, D, E) = \prod P(X_i \mid \text{Parent}(X_i))$$
$$= P(A)\,P(B)\,P(C|A,B)\,P(D|B)\,P(E|D)$$

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

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

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

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

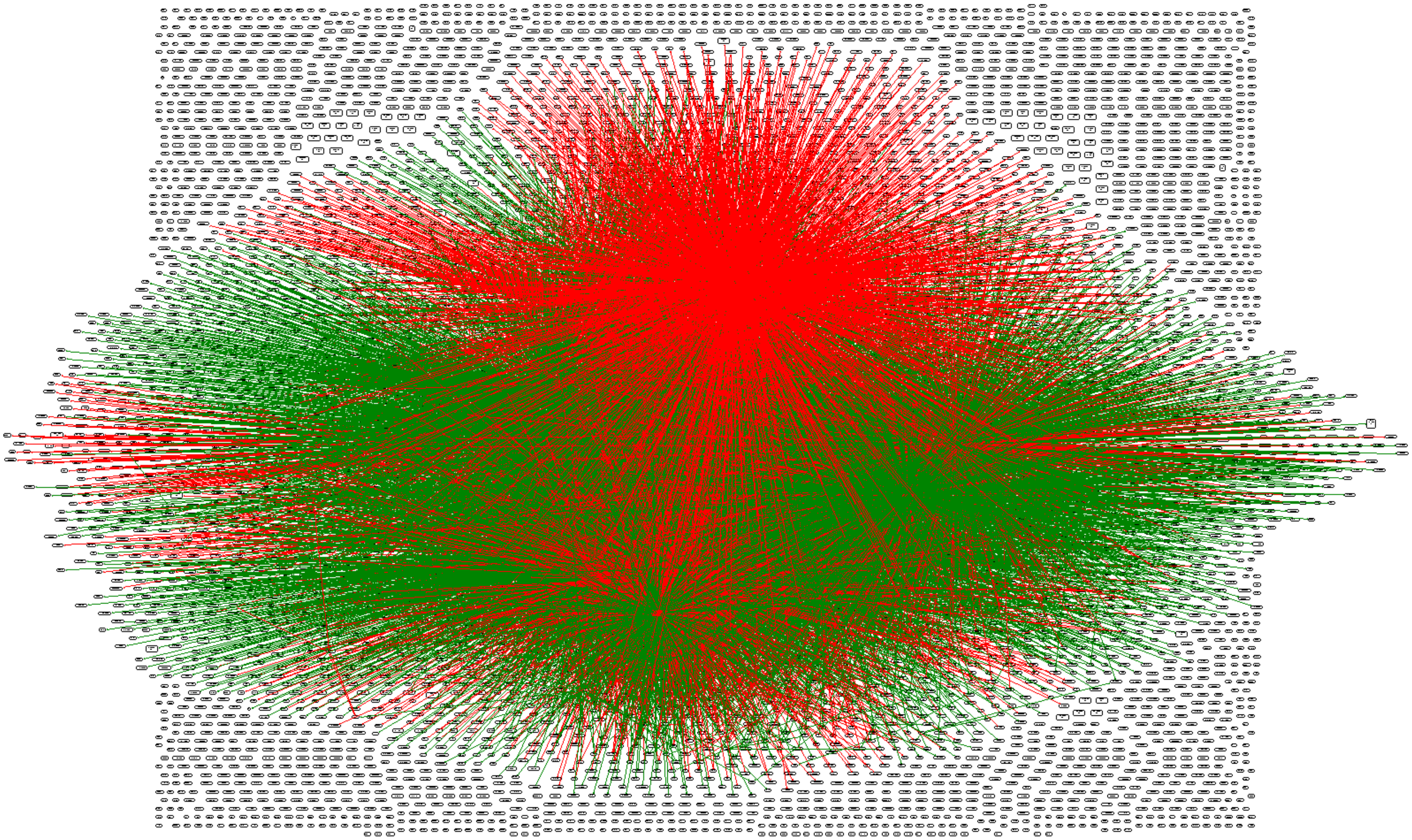
- **Differential Expression**

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.

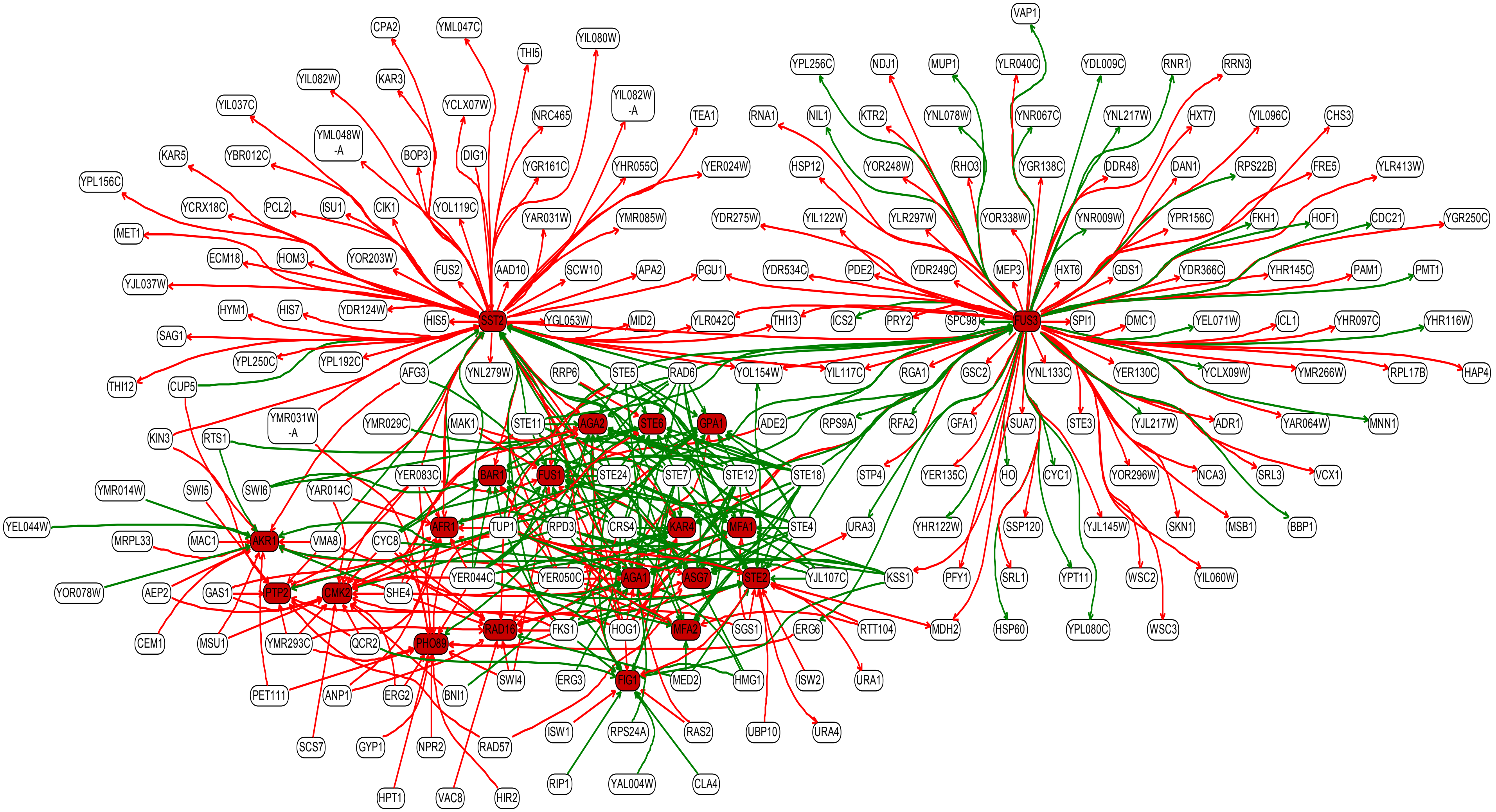
- **Bayesian Network**

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.

Mutation network for S. Cerevisiae



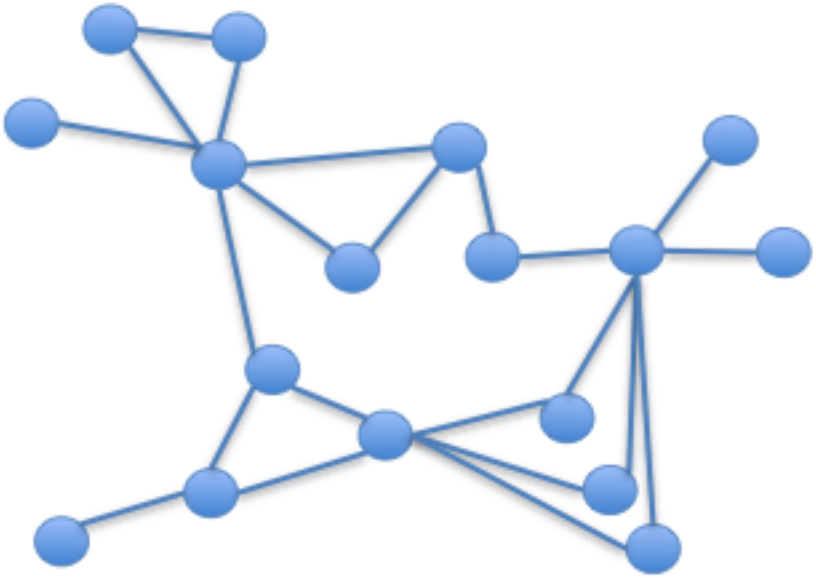
Mutation network filtered for the genes marked in red (mating)



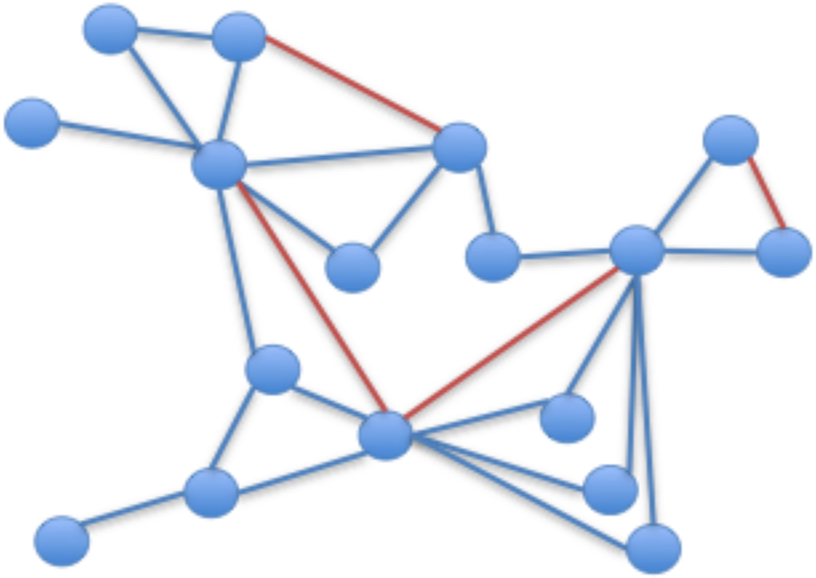
Thomas Schlitt, Johan Rung

Topological link prediction

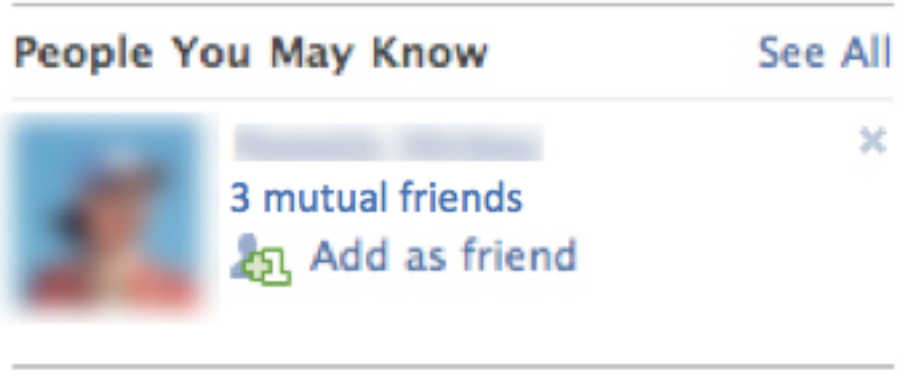
Observed network



Real/Future topology



A Local Community Approach to Link Prediction

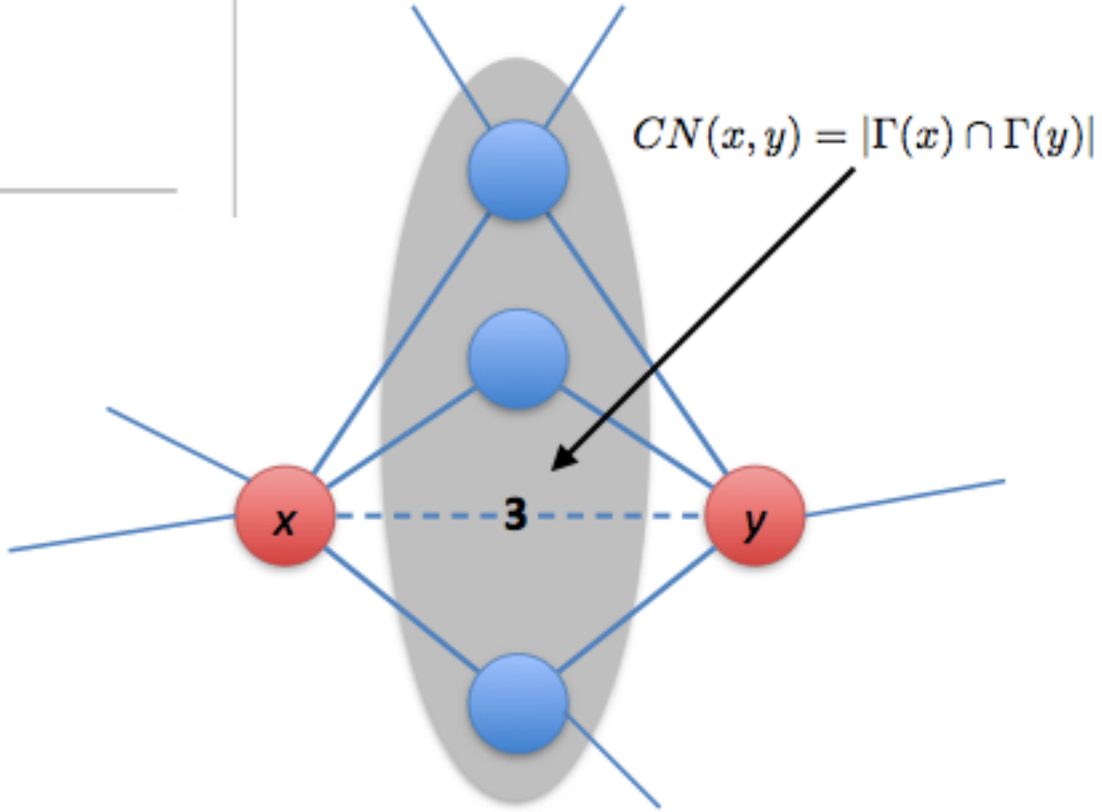


People You May Know See All

[redacted]

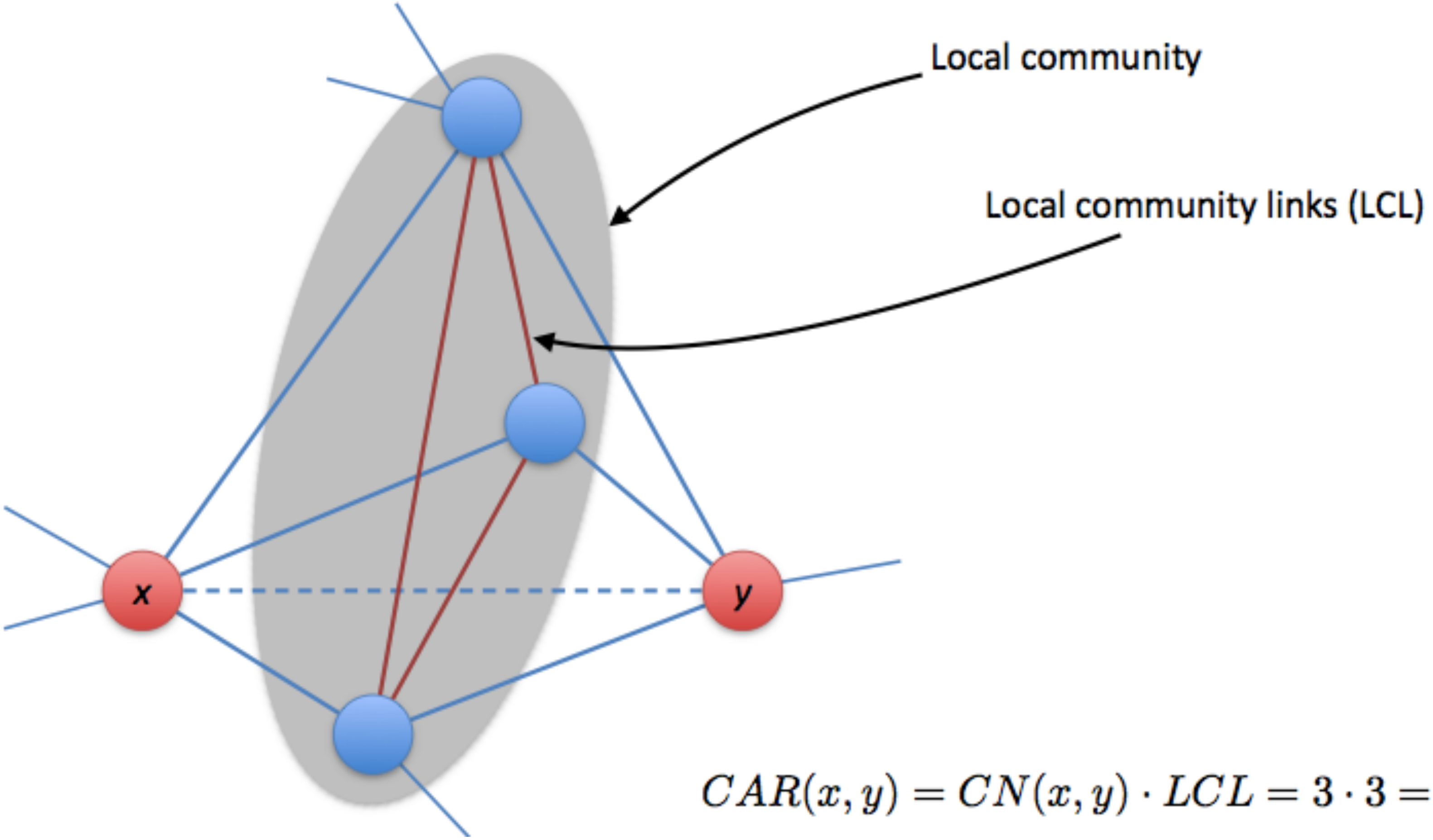
3 mutual friends

Add as friend



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

Shift from nodes to links: local community links and CAR

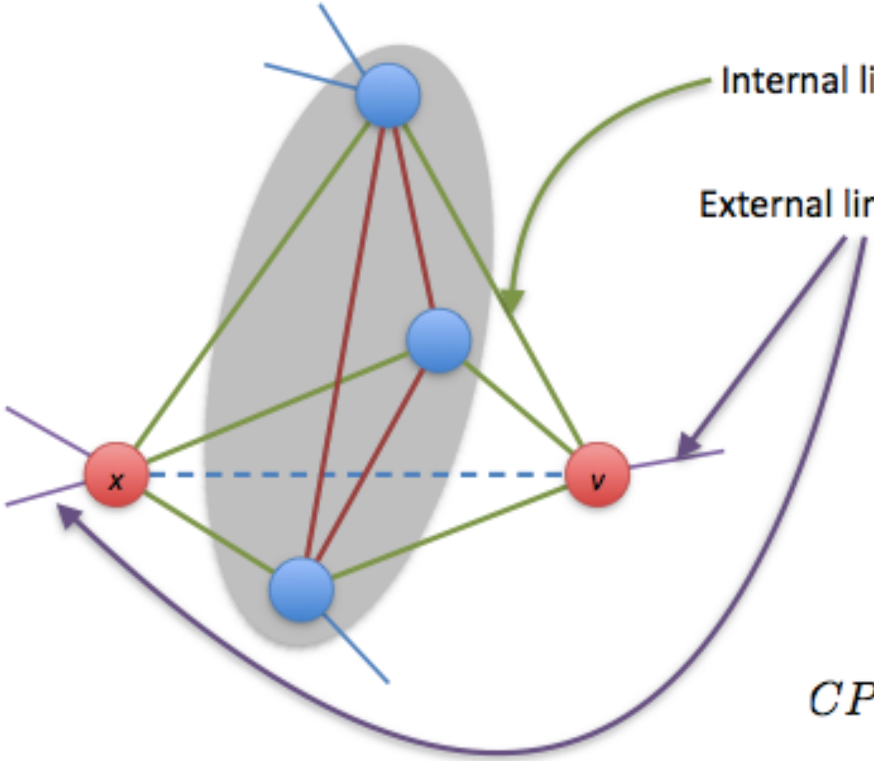


$$CAR(x,y) = CN(x,y) \cdot LCL = 3 \cdot 3 = 9$$

- 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, $i_x = i_y = CN(x,y)$

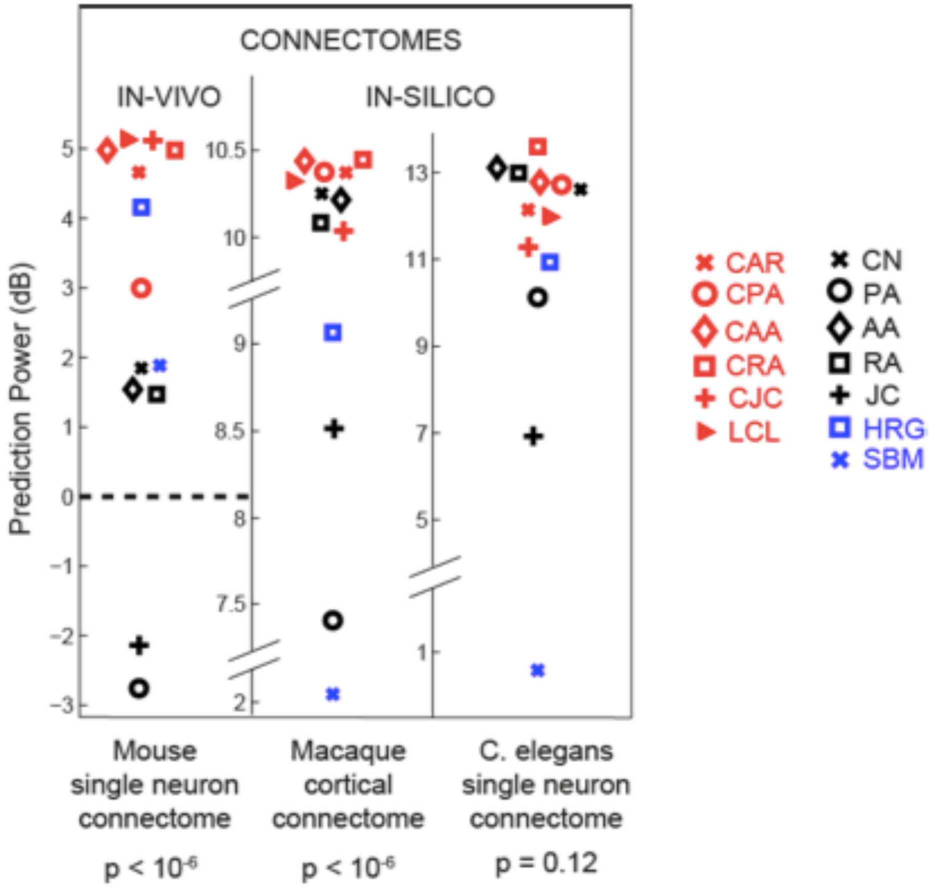
External links: e_x, e_y

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

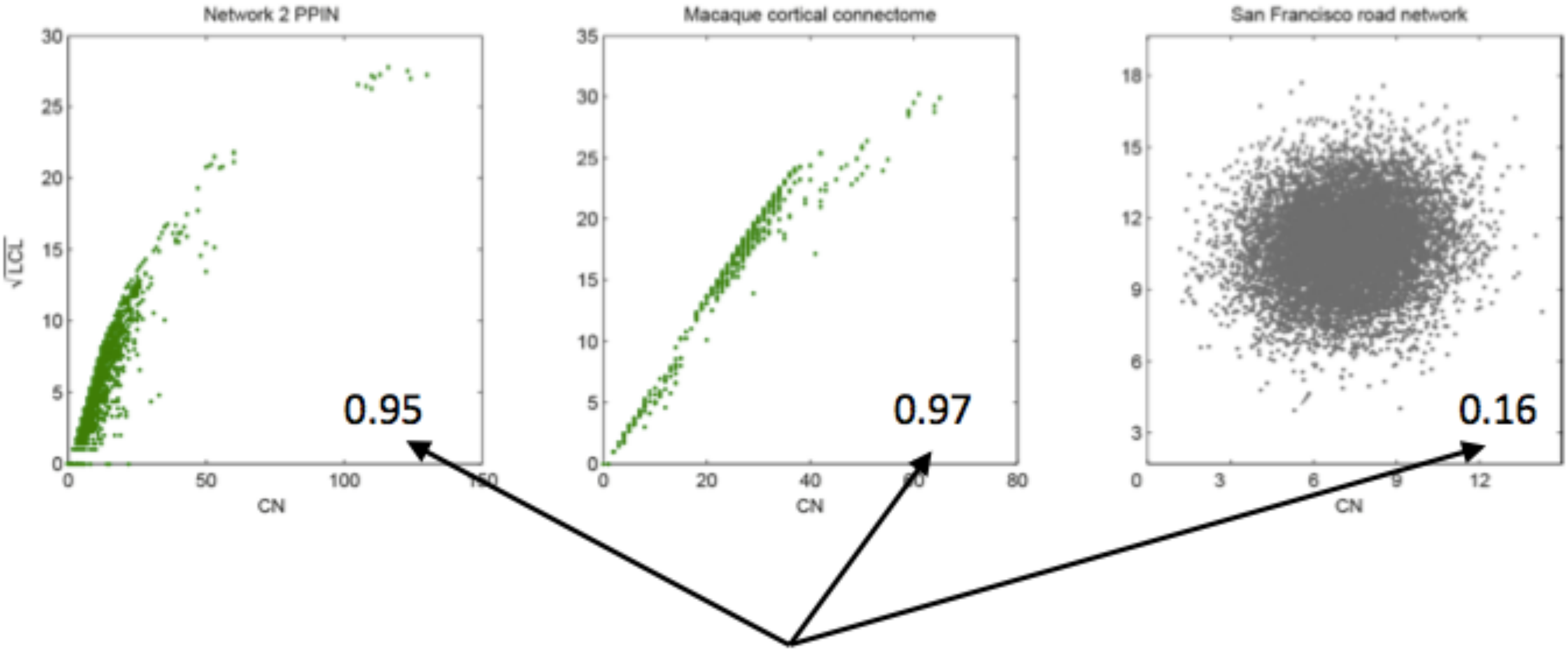
$$\downarrow$$

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

Testing CAR in brain connectomes

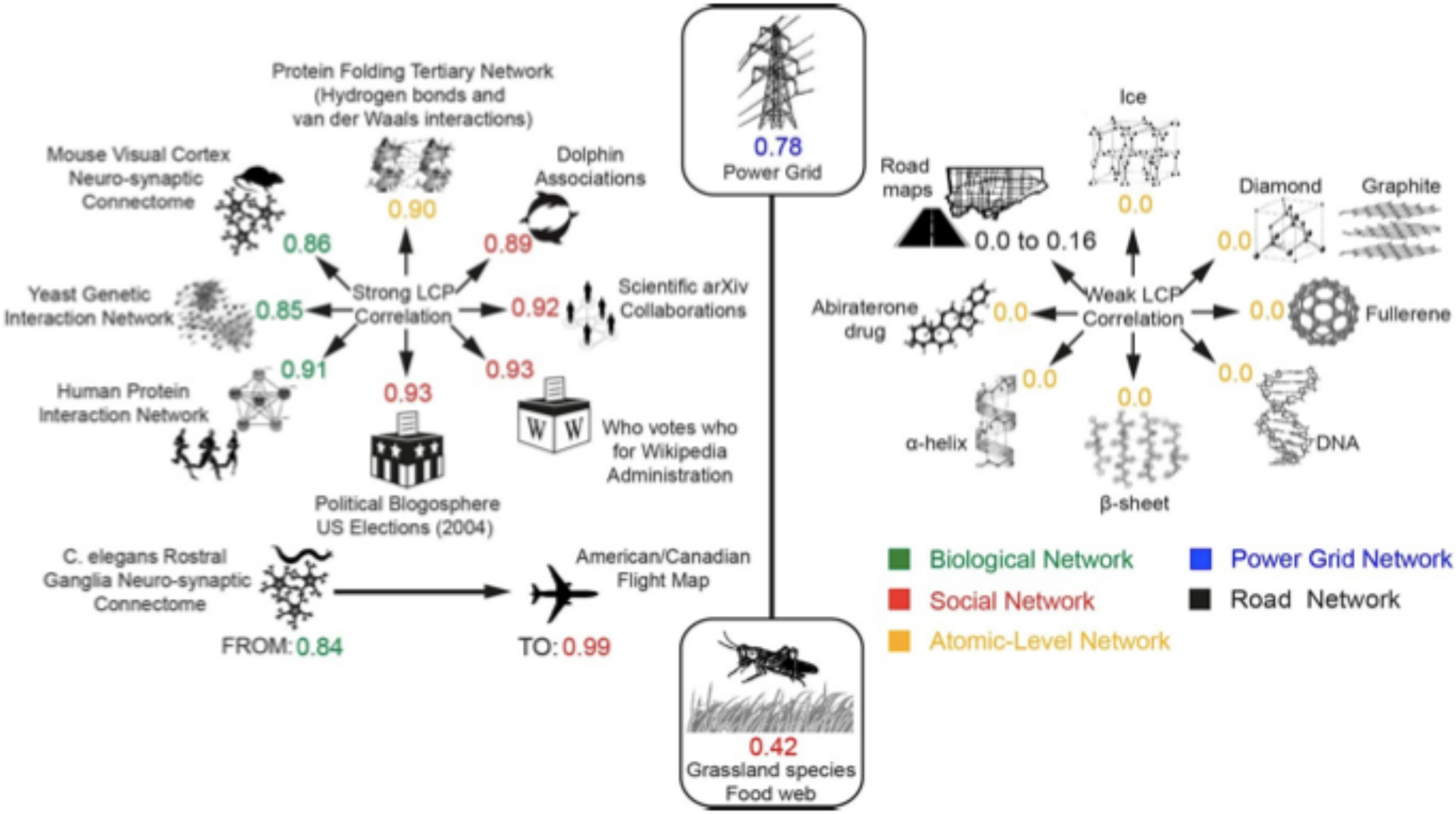


10% of links removed. Mean prediction precision considered relative to the mean random predictor performance



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

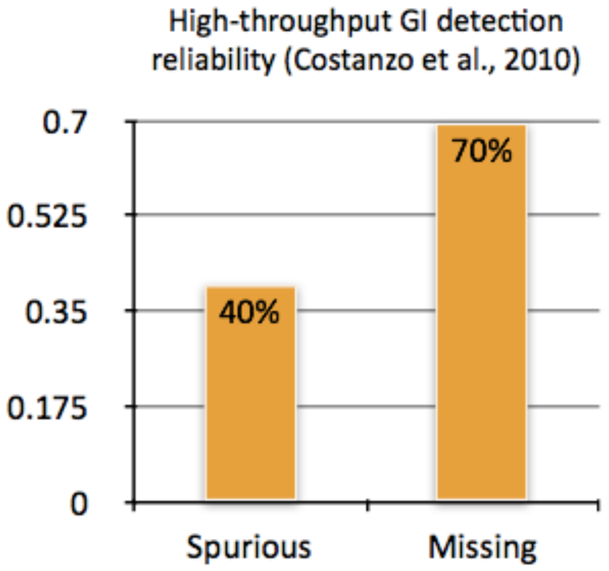
LCP and non-LCP networks

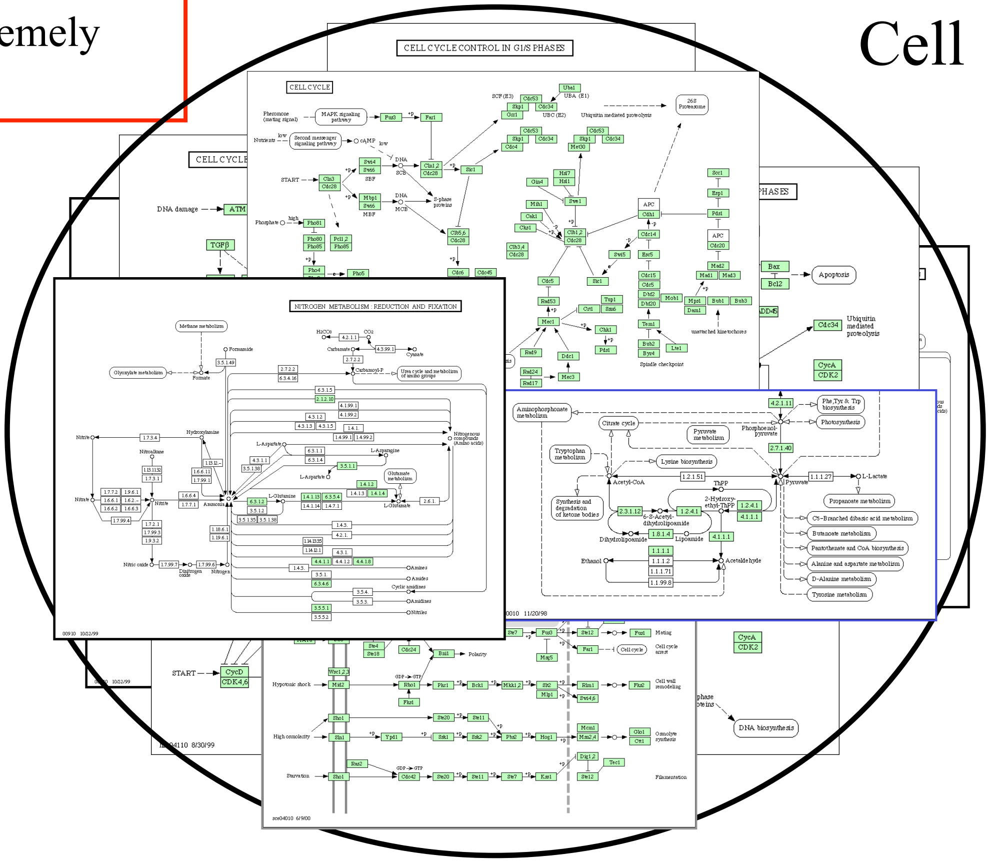


Information Derivable from Chip Data

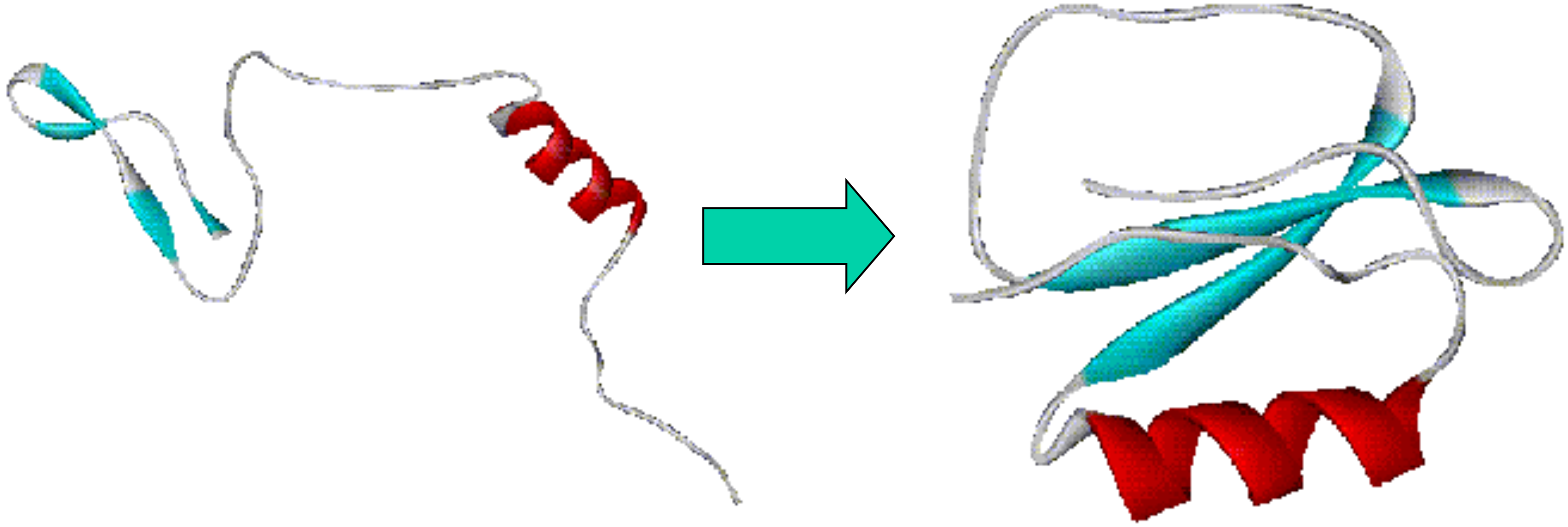
- The problem is the internal structure of a cell is very complex

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

Folding of chymotrypsin protein



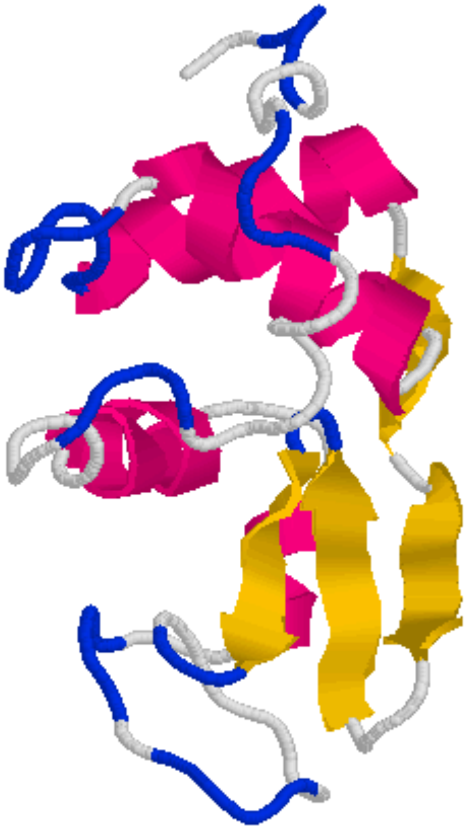
Protein Folding Problem

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

Can we predict structure (fold) from sequence?

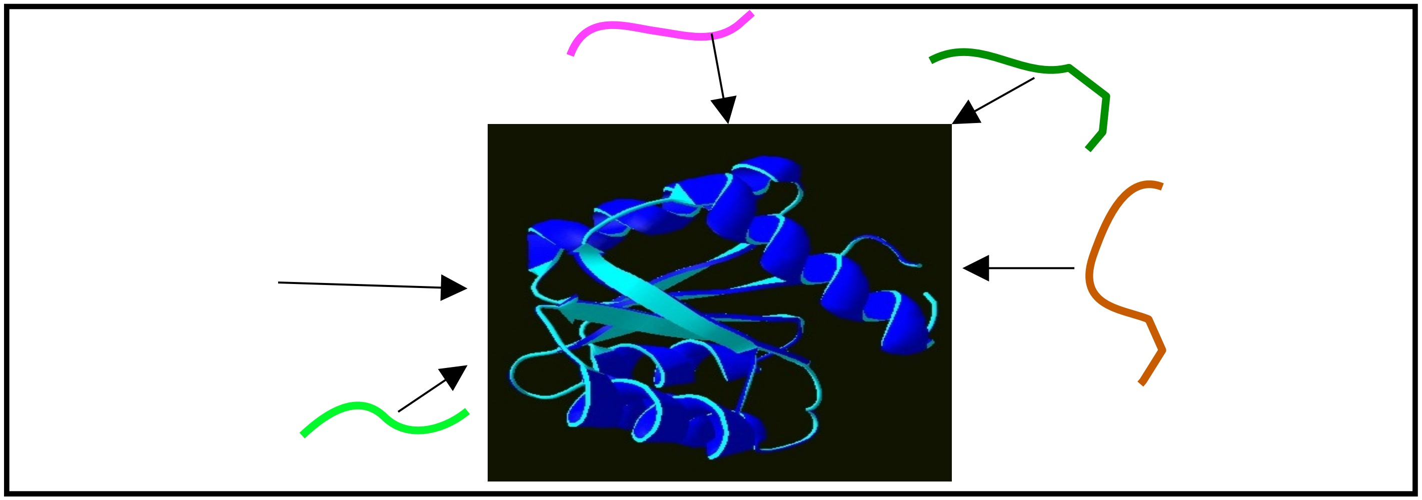
Lysozyme sequence:

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



Many proteins with dissimilar sequences fold into similar structures

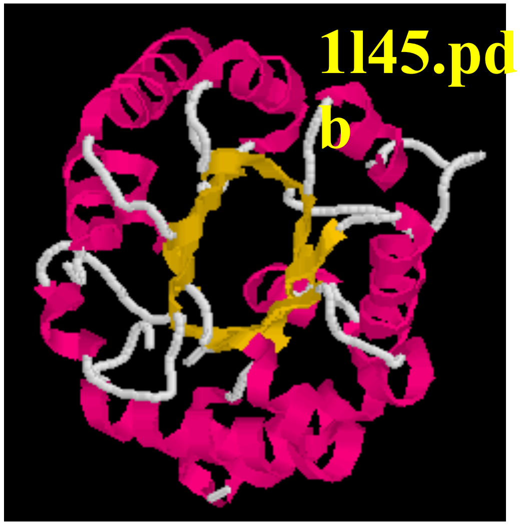
- Estimated number of folds: ~10000



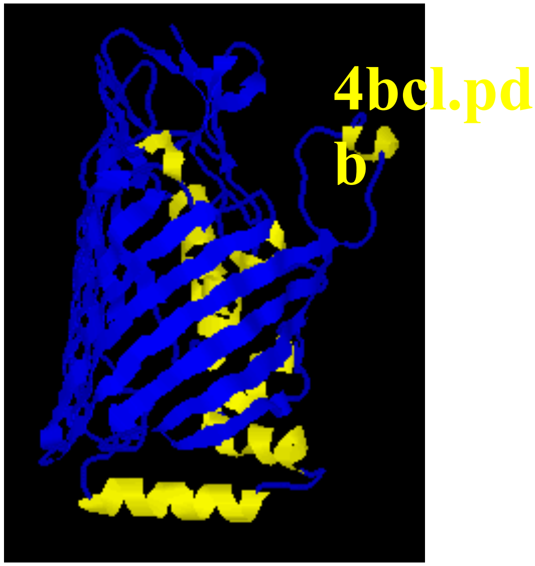
Protein Folds: sequential and spatial arrangement of secondary structures

Examples of different Folds

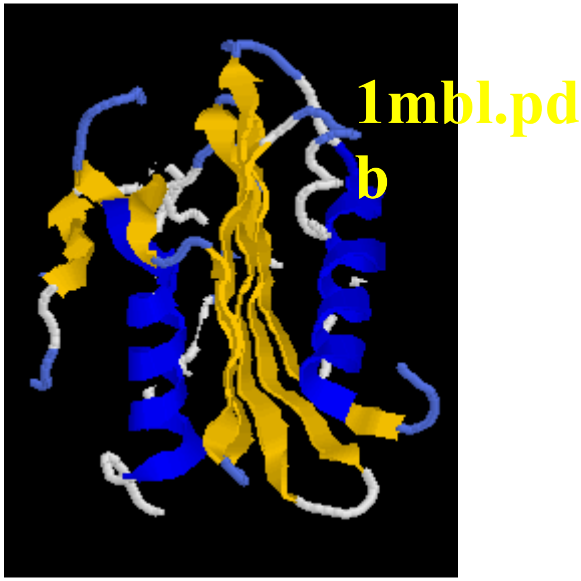
Refers to the spatial arrangement of its secondary structural elements (α-helices and β-strands)



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

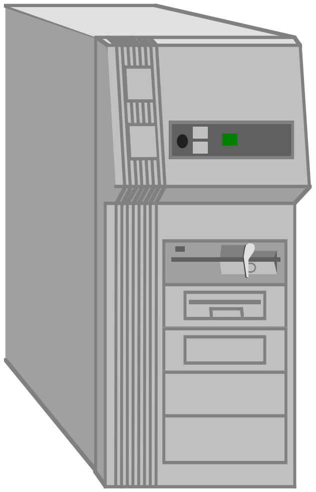
Problems:

- 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

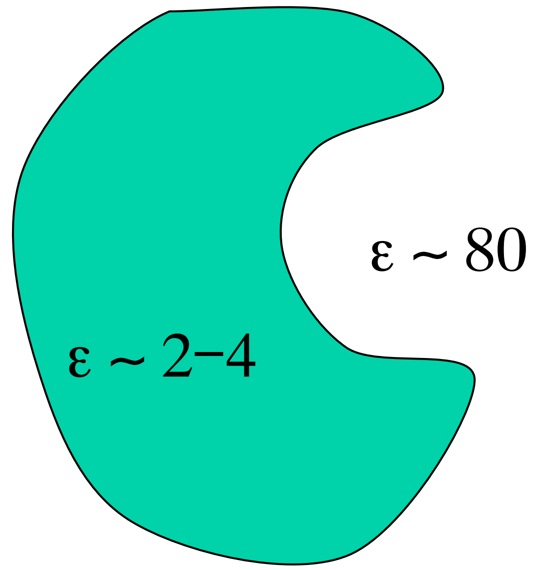
- Bond stretch
- Bond angle bending
- Torsion (bond rotation)
- Hydrogen bonding
- van der Waals interactions
- Electrostatic interactions
- Empirical solvation free energy



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

Electrostatic interactions: **Solvent dielectric model?**

- Problem: Inhomogeneous permittivity

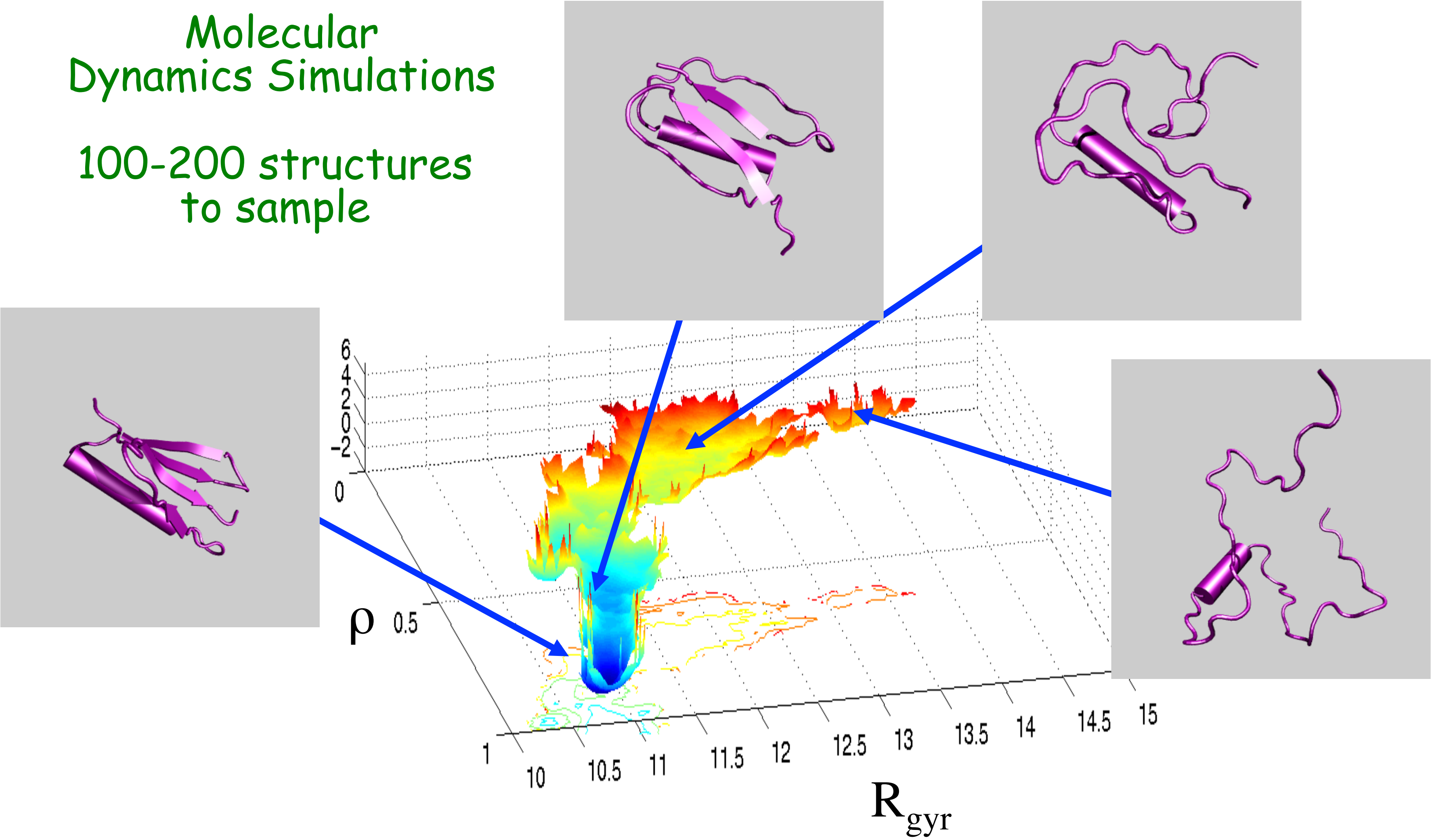


Depends on local structure and interactions with water

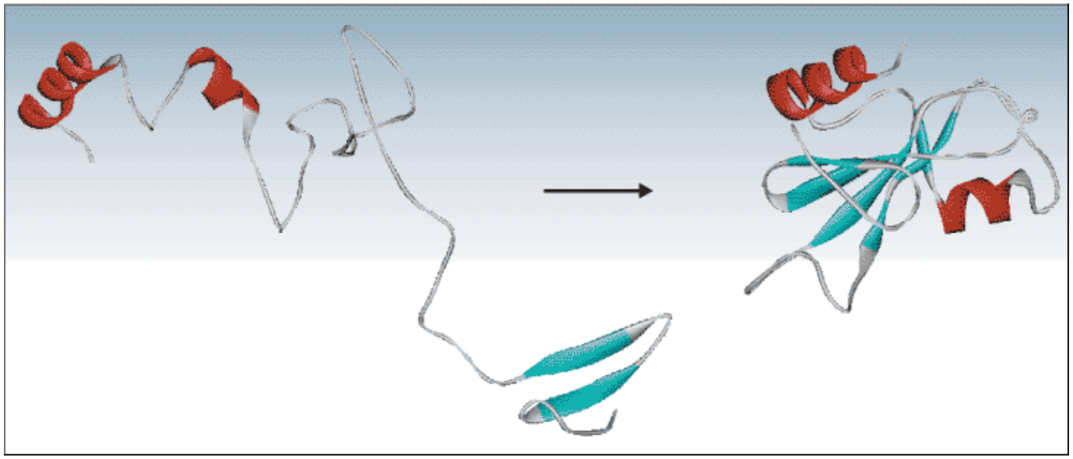
Folding Free Energy Landscape

Molecular
Dynamics Simulations

100-200 structures
to sample



Ab initio protein folding simulation



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})

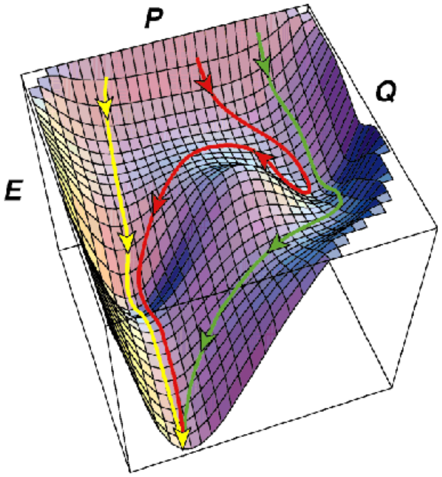
➔ **Blue Gene will need 3 years to simulate 100 μsec.**

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.



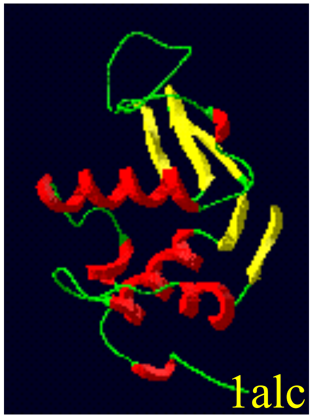
Comparative Modeling (homology modeling)

```
KQFTKCELSQNLYDIDGYGRIALPELICTMFH
TSGYDTQAIVENDESTEYGLFQISNALWCKSS
QSPQSRNICDITCDKFLDDDITDDIMCAKKIL
DIKGIDYWIAHKALCTEKLEQWLCEKE
```

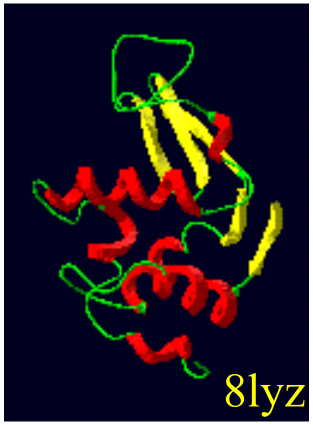
Homologous

Share Similar Sequence

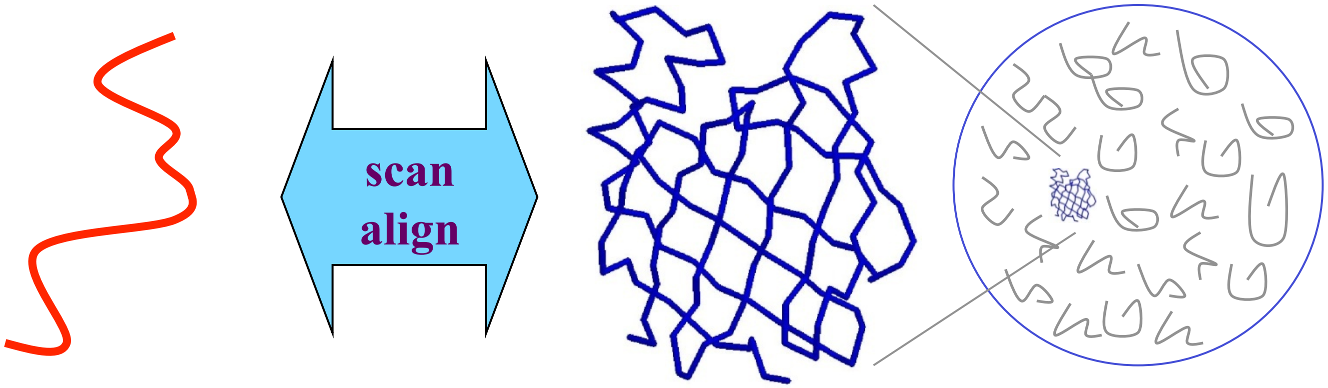
```
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKF
ESNFNTQATNRNTDGSTDYGILQINSRWWCNDGR
TPGSRNLCNIPCSALLSSDITASVNCAKKIVSDG
NGMNAWVAWRNRCKGTDVQAWIRGCRL
```



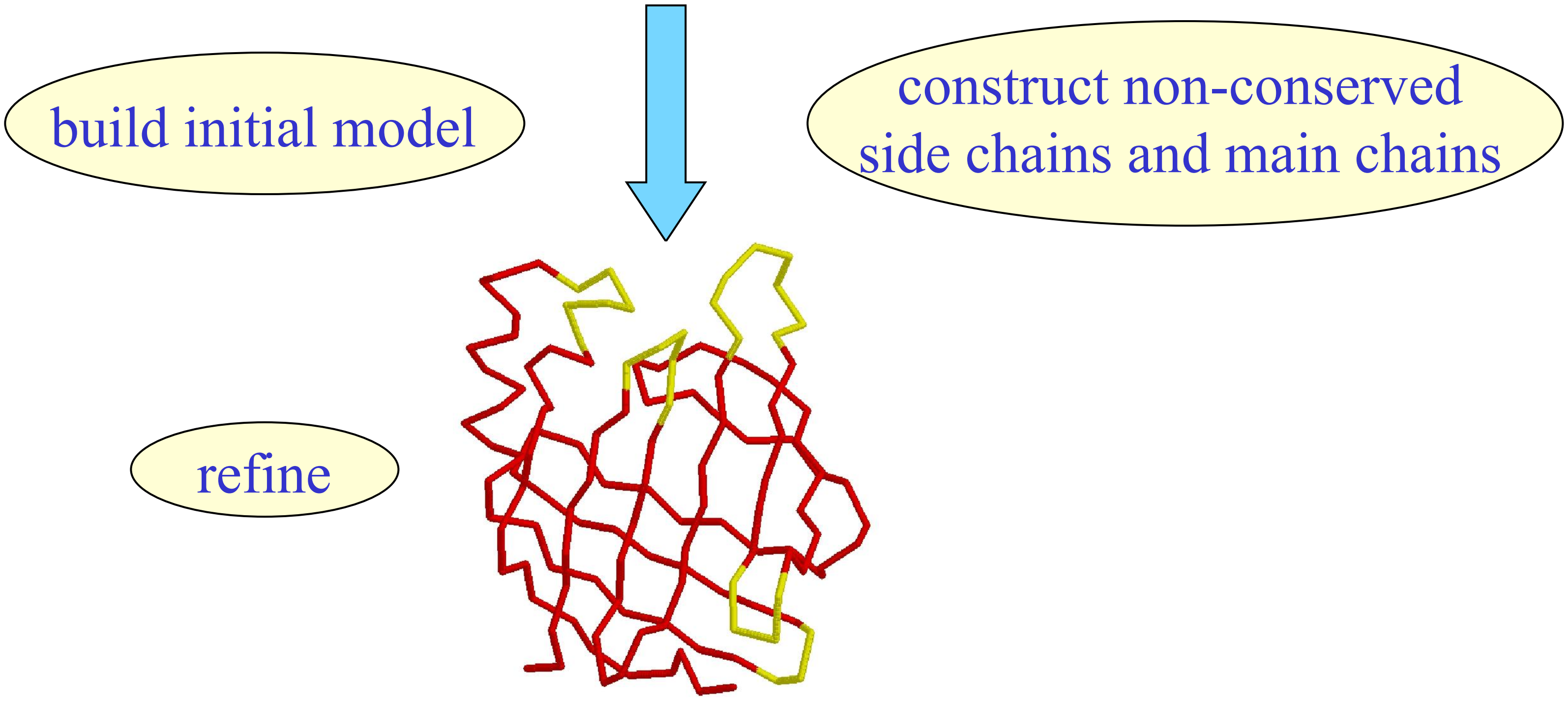
Use as template & model



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)

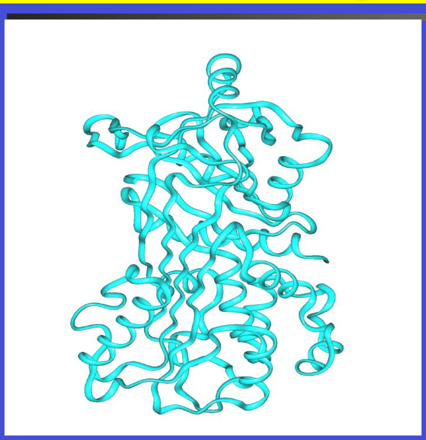
- *Use secondary structure prediction to guide the selection of a template, or to validate a template*

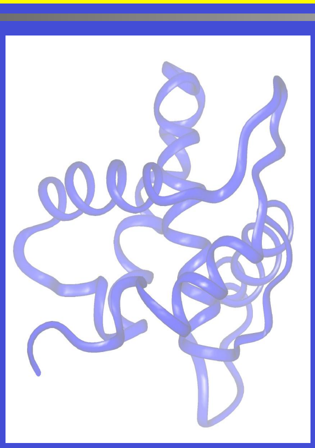
- *Use threading programs: sequence-structure alignments*

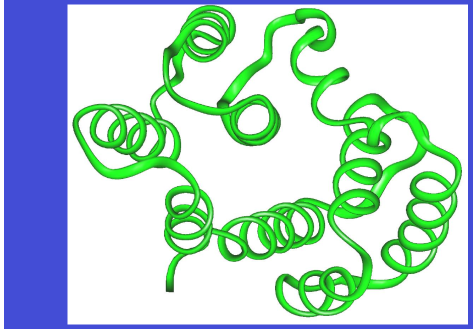
- *Use all of these methods! Meta-servers*

Fold Recognition: problem definition

A Library of Protein Folds (finite number)







Query sequence

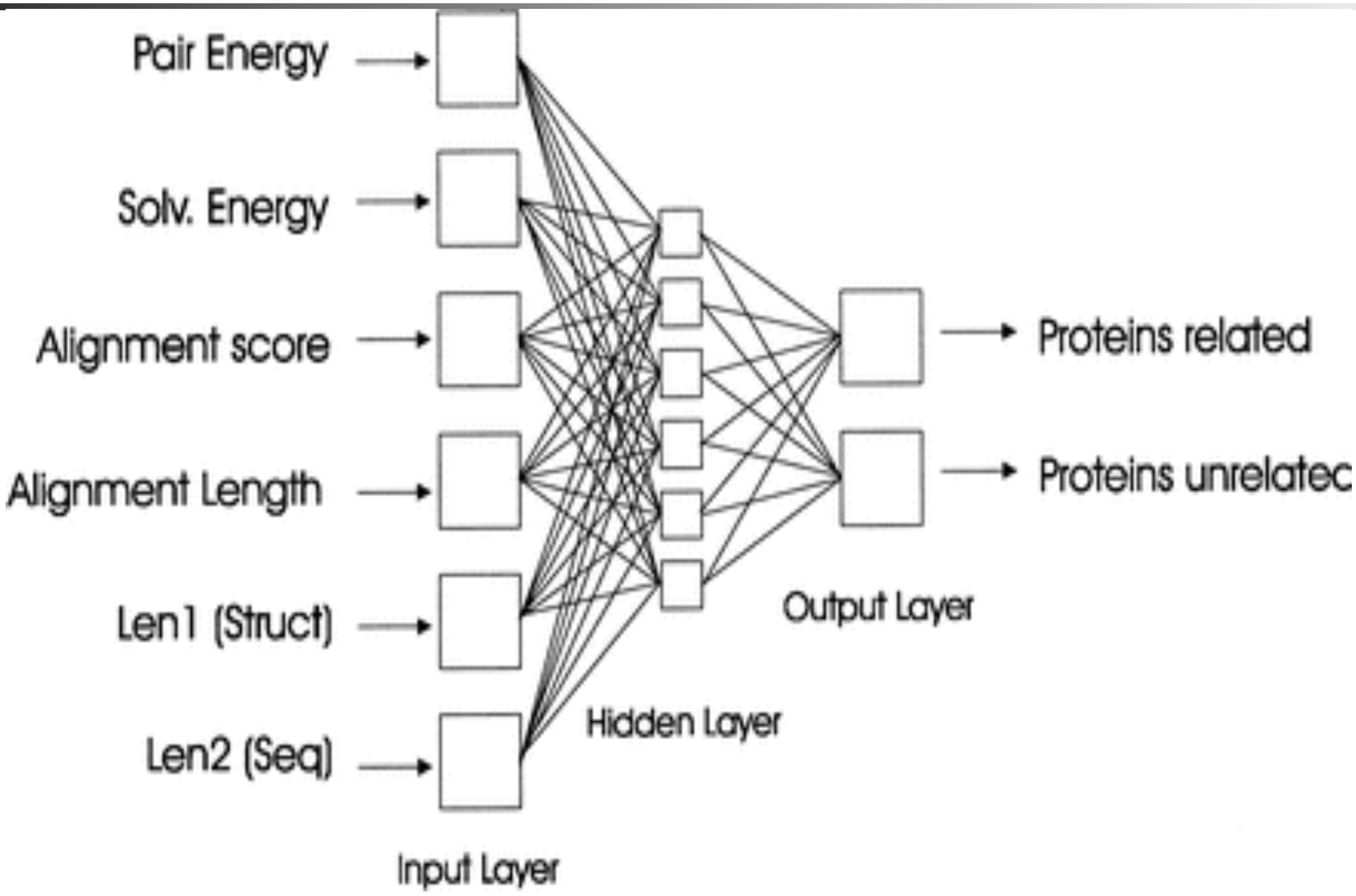
MTYGFRIPLNCERWGHKLSTVILKRP...

Goal: find to what folding template the sequence fits best

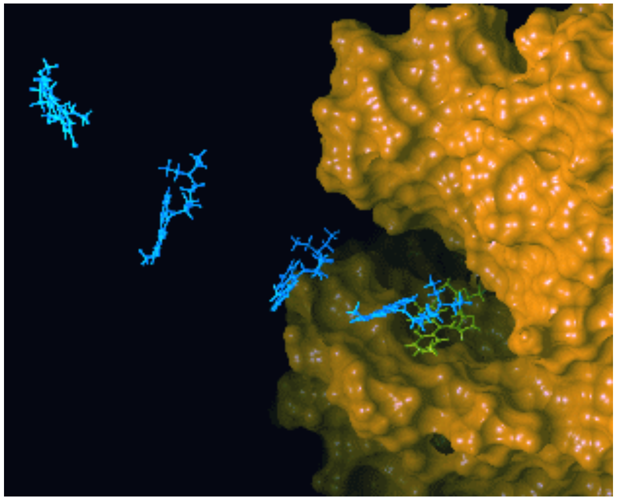
↓

Find ways to evaluate sequence-structure fit

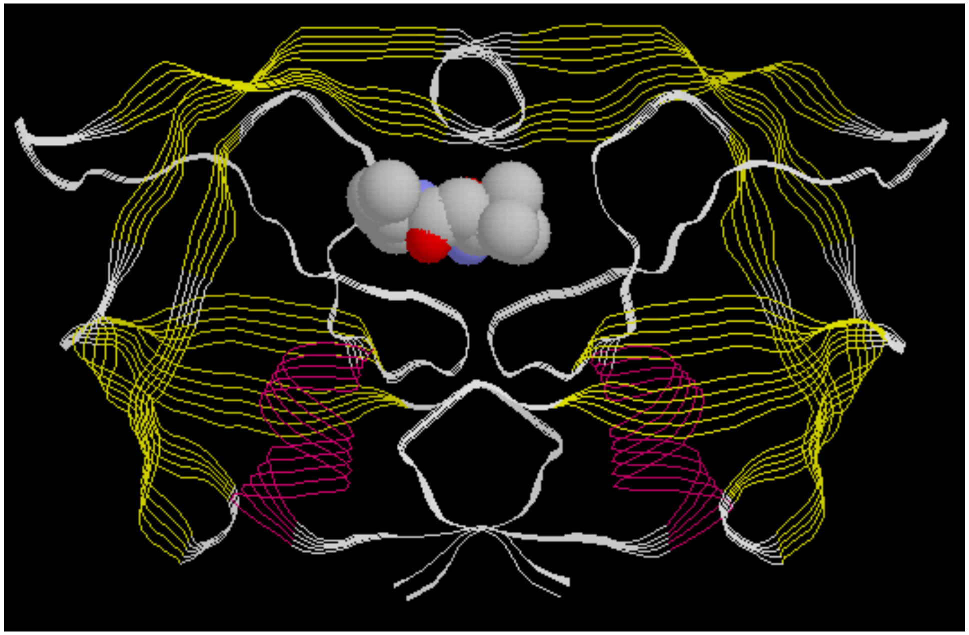
Essentials of GenTHREADER



Structure-Based Drug Design

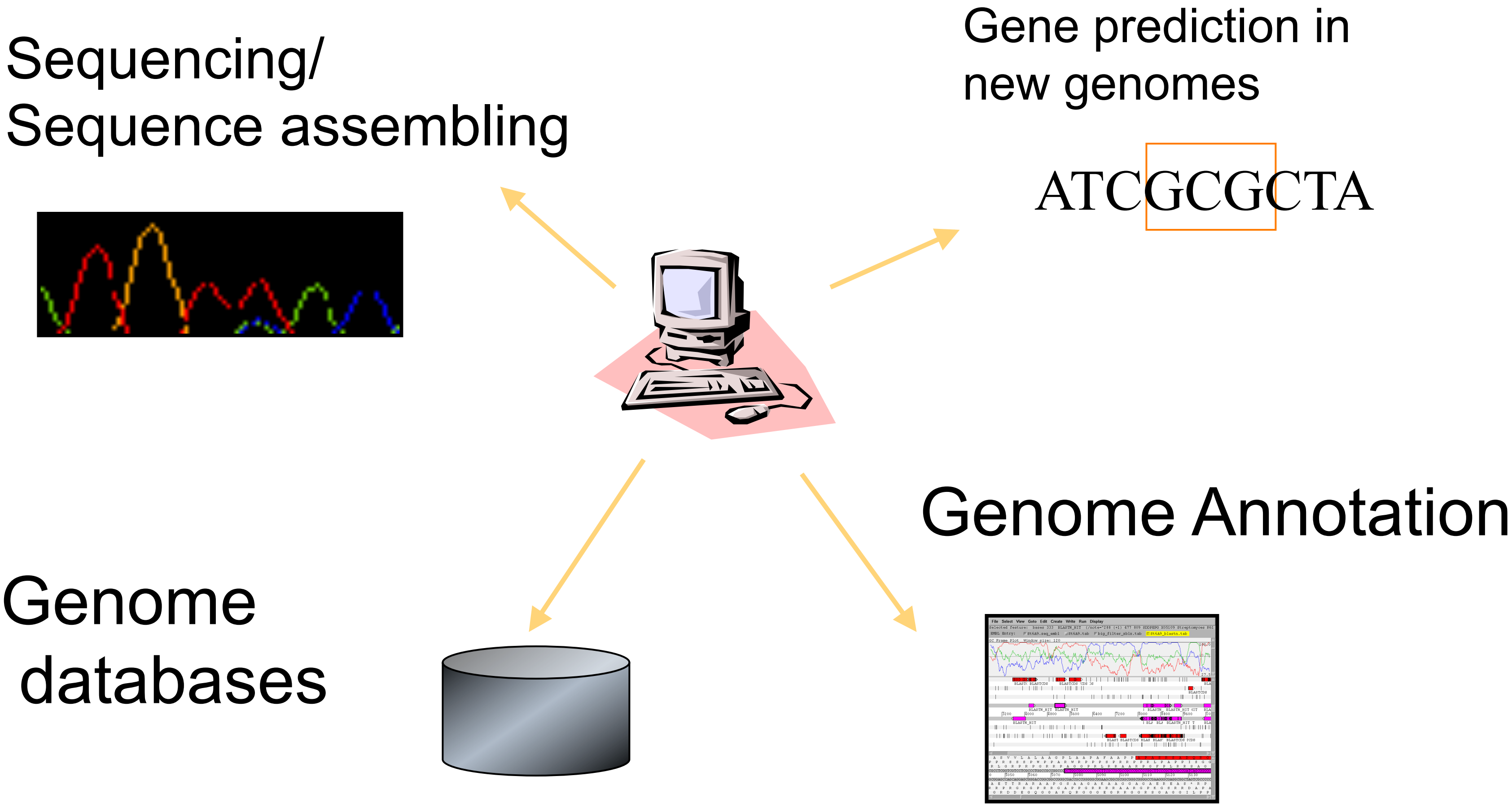


Structure-based rational drug design is still a major method for drug discovery.



HIV protease inhibitor

The role of Bioinformatics in support of genomics



Sequencing/
Sequence assembling

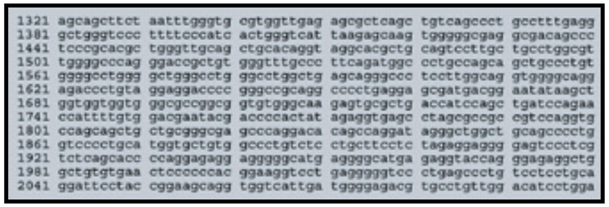
Gene prediction in
new genomes

ATCGCGCTA

Genome
databases

Genome Annotation

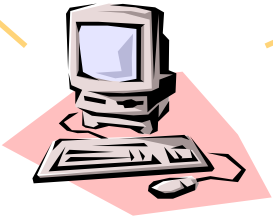
The role of bioinformatics supporting genetics

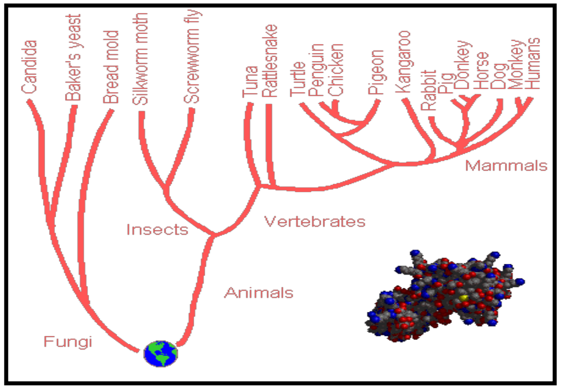


```
1321 agcagcttct aatttgggtg cgtggttgag agcgctcagc tgtcagccct gcctttgagg
1381 gctgggtccc ttttcccatc actgggtcat taagagcaag tgggggcgag gcgacagccc
1441 tcccgcacgc tgggttgcag ctgcacaggt aggcacgctg cagtccttgc tgcctggcgt
1501 tggggcccag ggaccgctgt gggtttgccc ttcagatggc cctgccagca gctgccctgt
1561 ggggcctggg 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

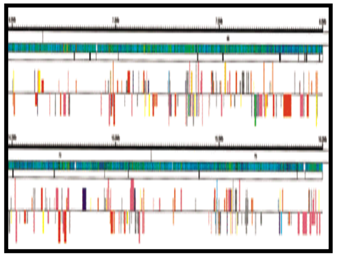




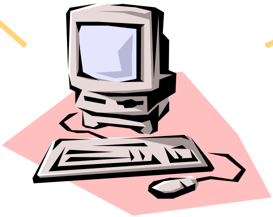
Phylogenetic trees

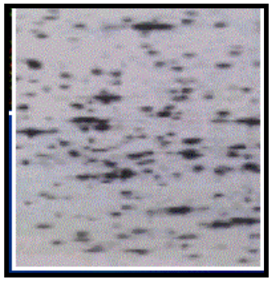
Structures

Bioinformatics in support of Post-Genomic Research

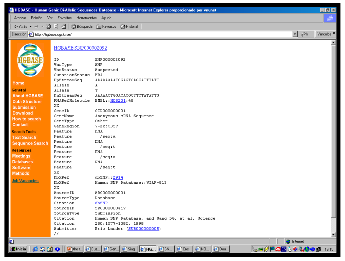


Genomes: Comparative Genomics (homology, evolution)





Proteomics (proteins in cells)

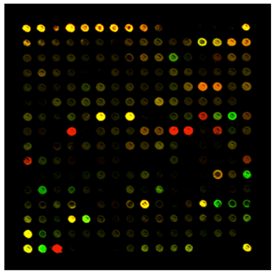

HGBASE - Human Genic Bi-Allelic Sequences Database - Microsoft Internet Explorer proporcionado por vnunet
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos Historial
Dirección http://hgbase.cgr.ki.se/
Ir Vínculos
HGBASE
Home
General
About HGBASE
Data Structure
Submission
Download
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

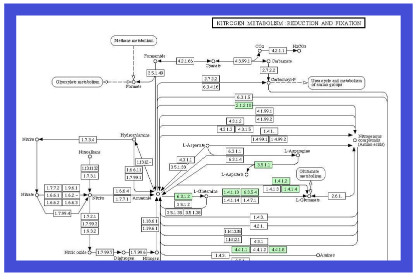
Individual Genome mutations/variations



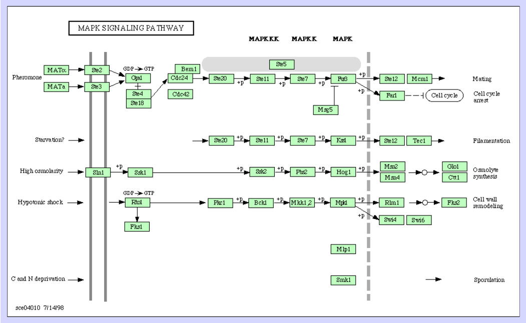
Functional Genomics (mRNAs)

DNA microarrays
Transcriptome Sequencing

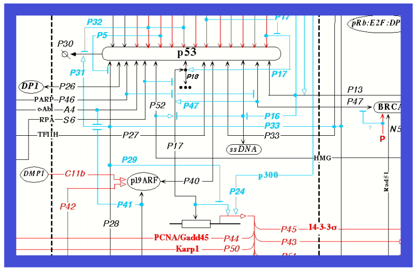
Bioinformatics in support of Systems Biology



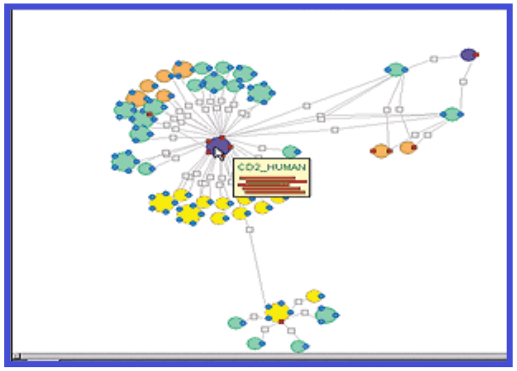
Metabolic Pathways



Signaling pathways



Genetic Networks



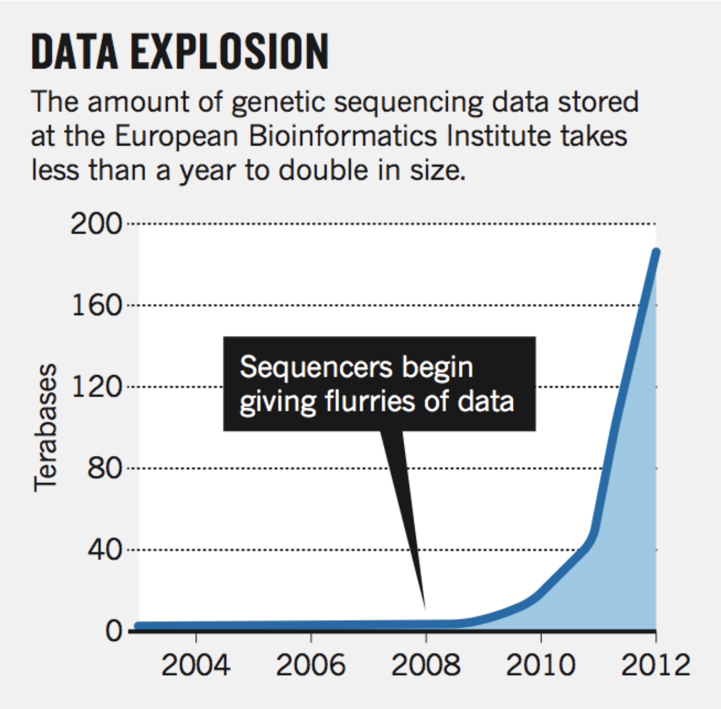
Interactions

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

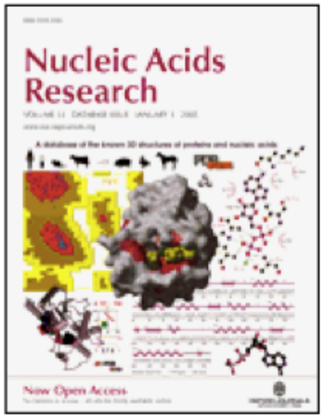
The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.

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/

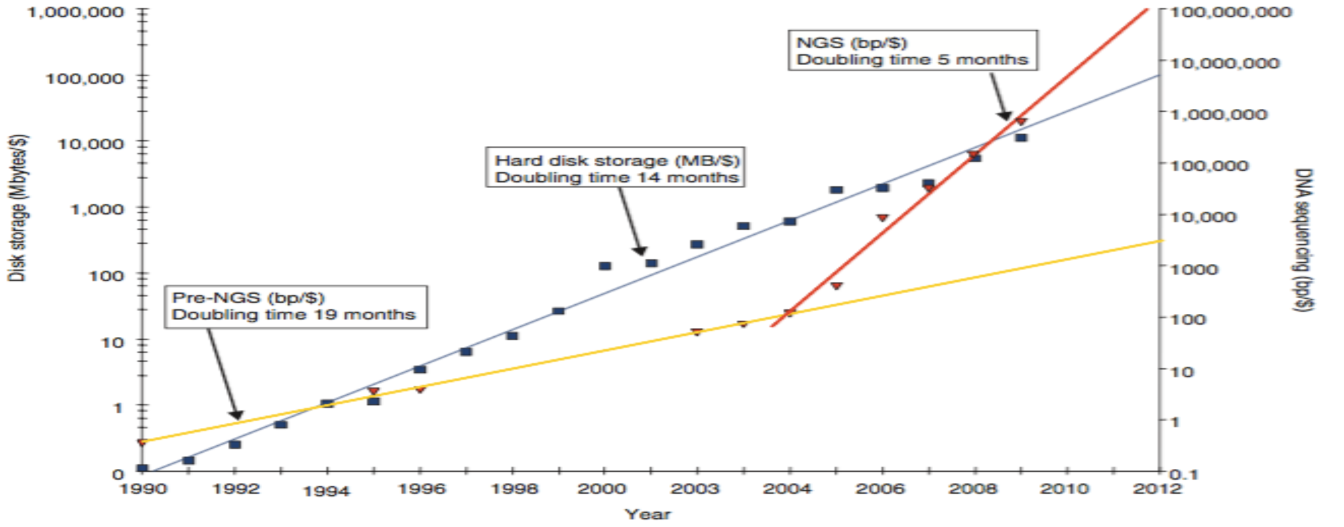
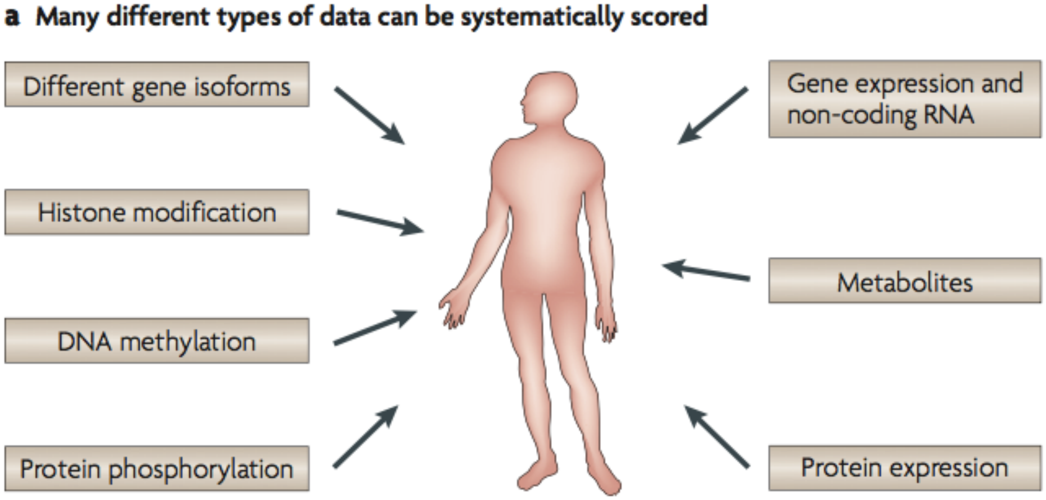


Figure 2. Historical trends in storage prices versus DNA sequencing costs. The blue squares describe the historic cost of disk prices in megabytes per US dollar. The long-term trend (blue line, which is a straight line here because the plot is logarithmic) shows exponential growth in storage per dollar with a doubling time of roughly 1.5 years. The cost of DNA sequencing, expressed in base pairs per dollar, is shown by the red triangles. It follows an exponential curve (yellow line) with a doubling time slightly slower than disk storage until 2004, when next generation sequencing (NGS) causes an inflection in the curve to a doubling time of less than 6 months (red line). These curves are not corrected for inflation or for the 'fully loaded' cost of sequencing and disk storage, which would include personnel costs, depreciation and overhead.

Over the coming years, the **National Cancer Institute will sequence a million genomes** to understand biological pathways and the genomic variation. Given that the whole genome of a tumor and a matching normal tissue sample consumes 1 TB of uncompressed data (this could be reduced by a factor of 10 if compressed); one million genomes will require 1 million TB, equivalent to 1000 petabyte (PB) or 1 Exabyte (EB)

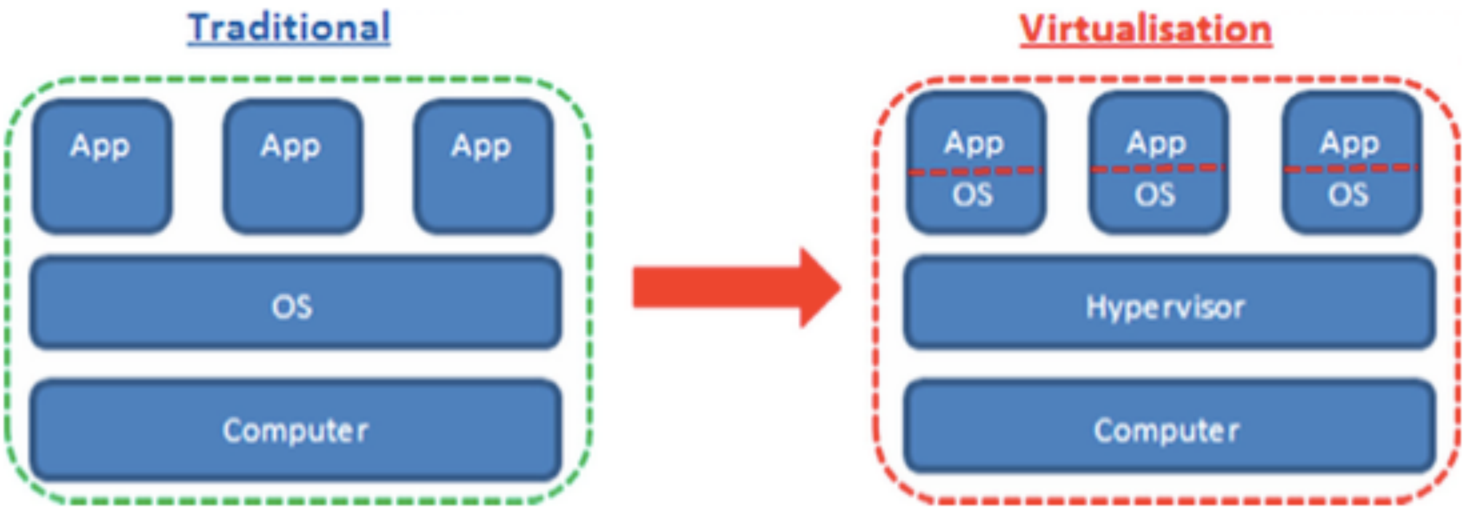
To Cloud computing

Biomedical research, driven by continued increases in data-generation capability, has become a data-intensive science.



For example, in the context of next-generation sequencing (NGS), a de novo assembly analysis step might require vastly more memory (RAM) in a single machine compared to a BLAST search step, which is much more limited by the clock speed of the CPU.

Fortunately, in recent years, cloud computing has emerged as a viable option to quickly and easily acquire computational resources required for an analysis.



The transition from traditional computing where applications interact with the hardware via one instance of the Operating System (OS), to virtualised environments where multiple OS images share the hardware resources (CPU, RAM, storage and networking), which are allocated and managed by virtualisation software known as a hypervisor or virtual machine monitor (VMM). Journal of Biomedical Informatics 46 (2013) 774–781

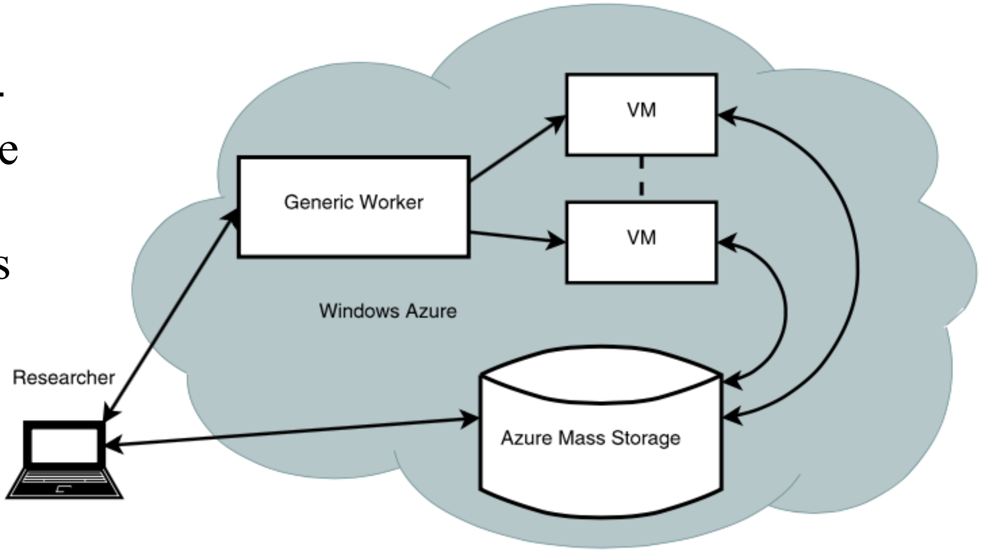
The earliest service provider to realize a practical cloud computing environment was **Amazon, with its Elastic Cloud Computing (EC2) service** introduced in 2005. It supports a variety of Linux and Windows virtual machines, a virtual storage system, and mechanisms for managing internet protocol (IP) addresses.

EC2 contains a variety of user selectable instance types that range in computing power and cost

An **EBS** volume is a storage device that can be attached to a running instance, similar to a USB thumb drive, and currently ranges in size from 1 GB to 1 TB.

S3 is an extremely reliable persistent storage system that also makes data readily available over the Internet.

Pay-per-use model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction



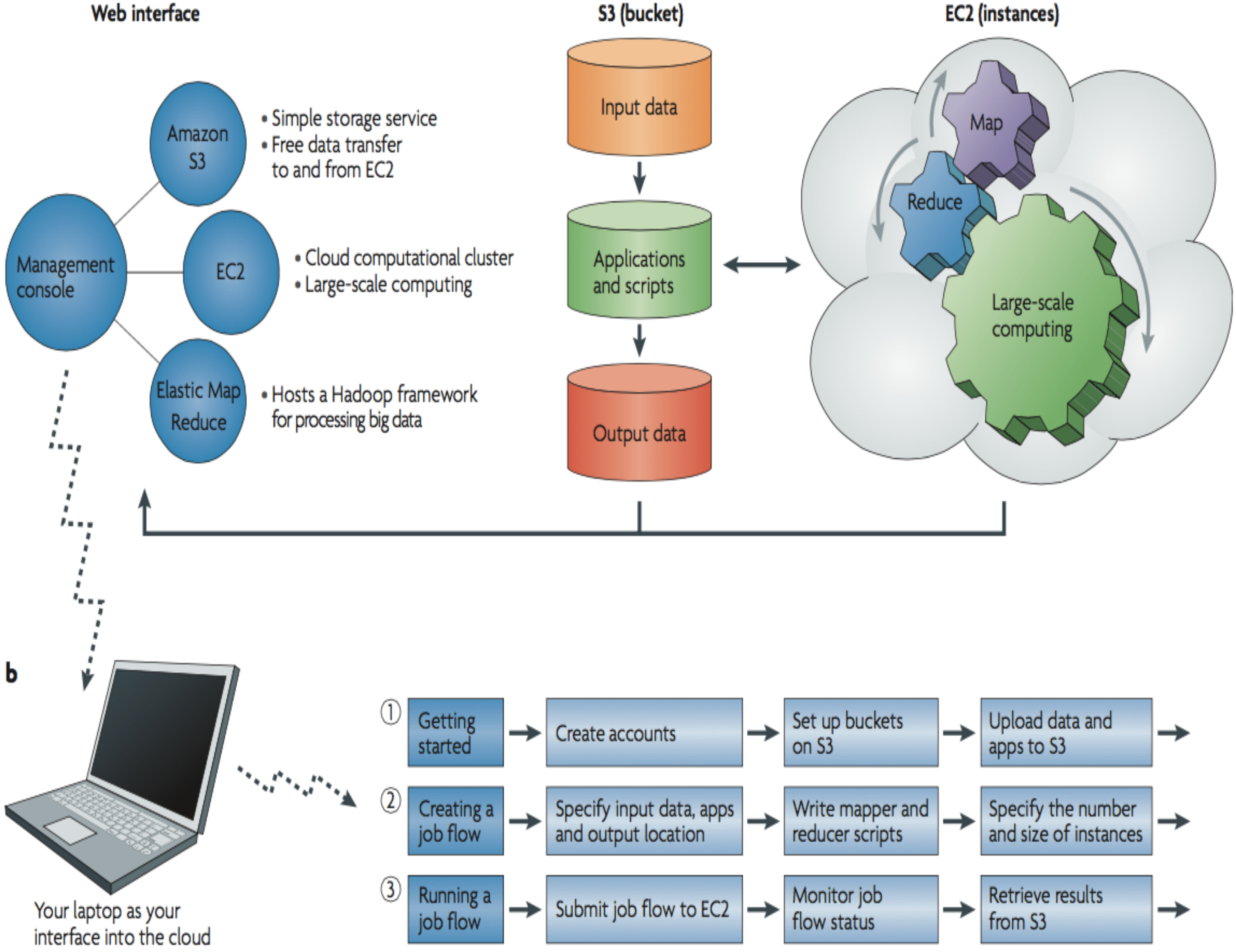
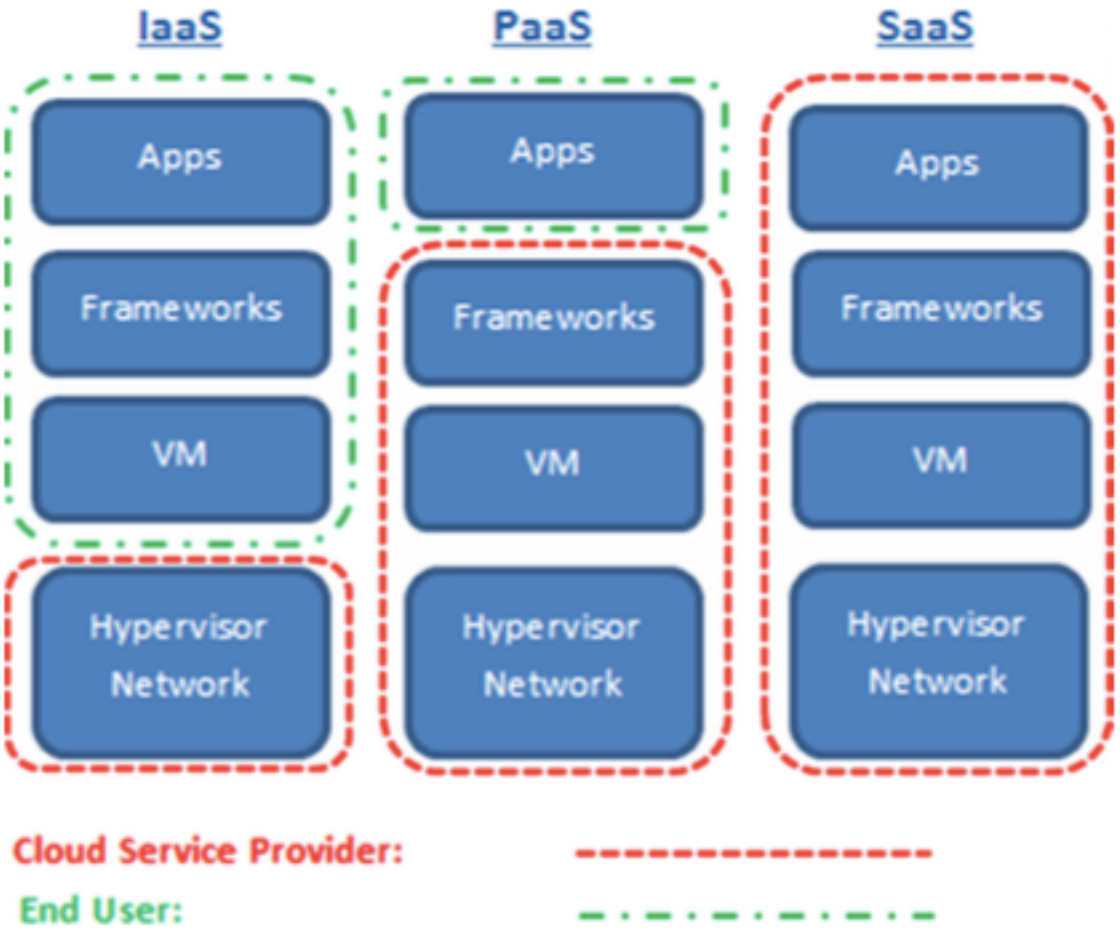


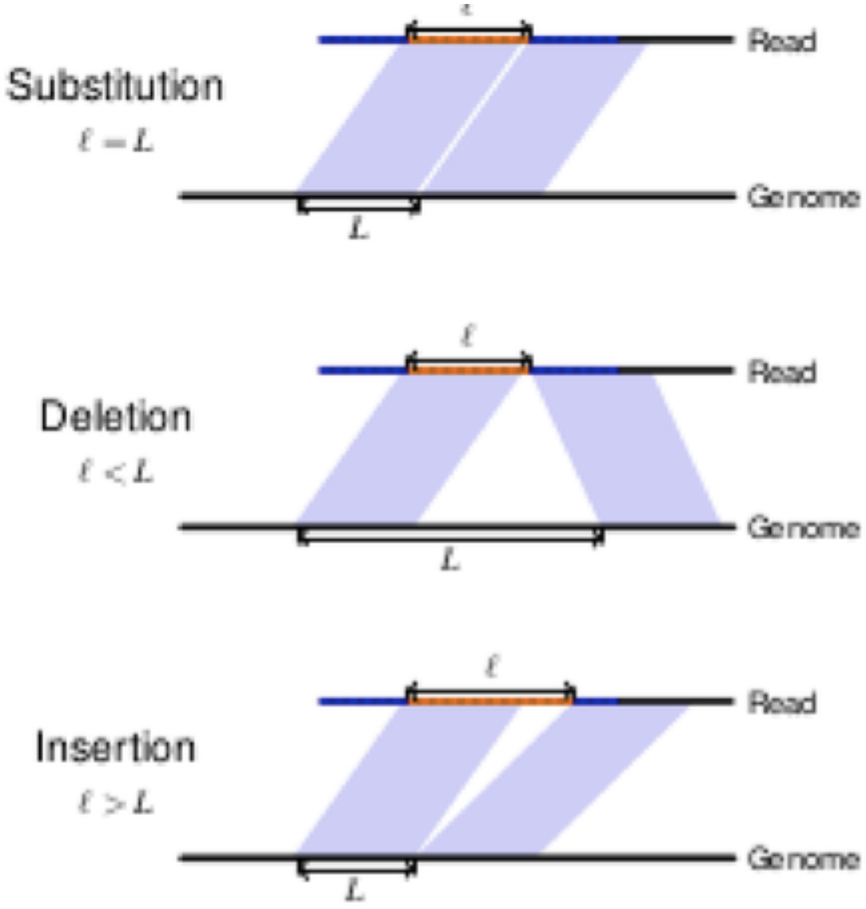
Figure 3 | **Amazon Web Services.** Amazon Web Services provides a simple and intuitive web-based interface into the



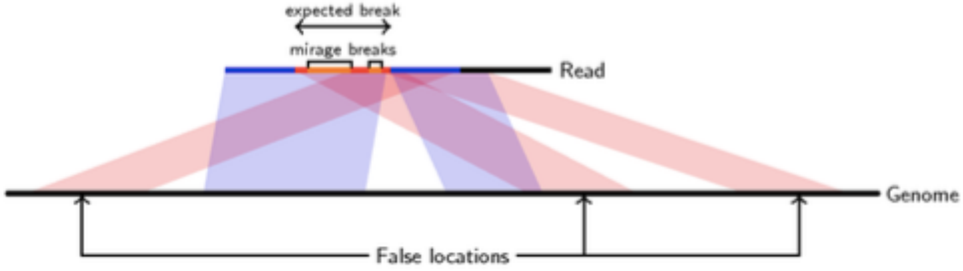
Cloud Computing Models are largely categorised as Infrastructure, Platform or Software as a Service (IaaS, PaaS, SaaS). Each model differs in the level of functionality provided to the user by the cloud provider.

Reads Mapping to reference genome

```
ATTTTATATTACATTAACAAGCTAATTTGCA
|||||||||||||||||||||||||||||||
8898989988848889888888889889888
ATTTTATATTACATTAACAAGCTAATTTGCA
ATTTTATATTACATTAACAAGCTAA......
ATTTTATATTACATTAACAAGCTNA......
ATTTTATATTACATTAACAANCTAA......
ATTTTATATTATATTAACAAGCTAA......
ATTTTATATTACATTNNCANNNNAA......
NTTTTATATTACATTAACNNGCTAA......
ATTTTATATTATATTAACAAGCNNN......
NTTTTATATTNCATTAACAAGCTNA......
ANNTTATATTATATTAACAAGCTAA......
ATTTTATATTATATTAACAANNTNA......
NTTTTATATTATATTAACAAGNTNN......
ATTTTATATTACATTAACAAGCTAAT.....
ATTTTATATTACATTAACNAGCTNNT.....
NNTTTATATTATATTAACAAGCTAAT.....
ATTTTATATTACNTTAACAAGCTNNT.....
ATTTTATATTANATTAACAANCTAAN.....
ATTTTATATTATATTAACAANCTAAT.....
ATTTTATATTACATTAACAAGCTAATT....
ATTTTATATTACATTAACAAGCTAATT....
ANNTTATATTACATTAACAAGCTAATT....
ATTTTATATTACATTAACAAGCNAATT....
NTTTTANATTACATTAACAAGCTAATT....
ATTTTATATTATATTAACAAGCTAATT....
ATTTTATATTATATTAACAAGCTAATT....
```



(c)



(d)

CloudBurst: highly sensitive read mapping with MapReduce

CloudBurst is a new parallel read-mapping algorithm optimized for mapping next-generation sequence data to the human genome and other reference genomes, for use in a variety of biological analyses including SNP discovery, genotyping and personal genomics.

CloudBurst uses the open-source Hadoop implementation of MapReduce to parallelize execution using multiple compute nodes.

MapReduce (Dean *et al*., 2008) is the software framework developed and used by GoogleTM to support parallel distributed execution of their data intensive applications. Google uses this framework internally to execute thousands of *MapReduce* applications per day, processing petabytes of data, all on commodity hardware.

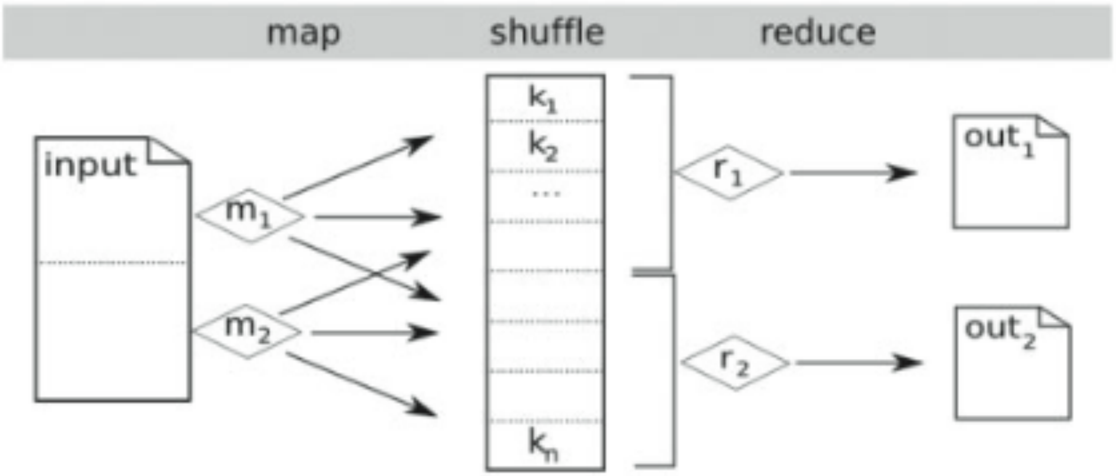


Fig. 1. Schematic overview of MapReduce. The input file(s) are automatically partitioned into chunks depending on their size and the desired number of mappers. Each mapper (shown here as m_1 and m_2) executes a user-defined function on a chunk of the input and emits key–value pairs. The shuffle phase creates a list of values associated with each key (shown here as k_1, k_2 and k_n). The reducers (shown here as r_1 and r_2) evaluate a user-defined function for their subset of the keys and associated list of values, to create the set of output files.

Unlike other parallel computing frameworks, which require application developers explicitly manage inter-process communication, computation in *MapReduce* is divided into two major phases called *map* and *reduce*, separated by an internal *shuffle* phase of the intermediate results (Fig. 1), and the framework automatically executes those functions in parallel over any number of processors.

MapReduce is designed for computations with extremely large datasets, far beyond what can be stored in RAM. Instead it uses files for storing and transferring Intermediate results, including the inter-machine communication between *map* and *reduce* functions.
This could become a severe bottleneck, so Google developed the robust distributed Google File System (GFS) (Ghemawat *et al*., 2003) to efficiently support *MapReduce.*
GFS is designed to provide very high-bandwidth for *MapReduce* by replicating and partitioning files across many physical disks. Files in the GFS are automatically partitioned into large chunks (64MB by default), which are replicated to several physical disks (three by default) attached to the compute nodes.
MapReduce is also ‘data aware’: it attempts to schedule computation at a compute node that has the required data instead of moving the data across the network.

Hadoop and the *Hadoop Distributed File System* (*HDFS*) are open source versions of *MapReduce* and the GFS implemented in Java and sponsored by AmazonTM, YahooTM, Google, IBMTM and other major vendors.

Like Google's proprietary *MapReduce* framework, applications developers need only write custom *map* and *reduce* functions, and the *Hadoop* framework automatically executes those functions in parallel. *Hadoop* and *HDFS* are used to manage production clusters with 10 000 + nodes and petabytes of data, including computation supporting every Yahoo search result. A Hadoop cluster of 910 commodity machines recently set a performance record by sorting 1 TB of data (10 billion 100 bytes records) in 209 s (http://www.hpl.hp.com/hosted/sortbenchmark/).

Amazon's Elastic Compute Cloud (EC2) (http://aws.amazon.com) contains tens of thousands of virtual machines, and supports *Hadoop* with minimal effort. In EC2, there are five different classes of virtual machines available providing different levels of CPU, RAM and disk resources with price ranging from $0.10 to $0.80 per hour per virtual machine.

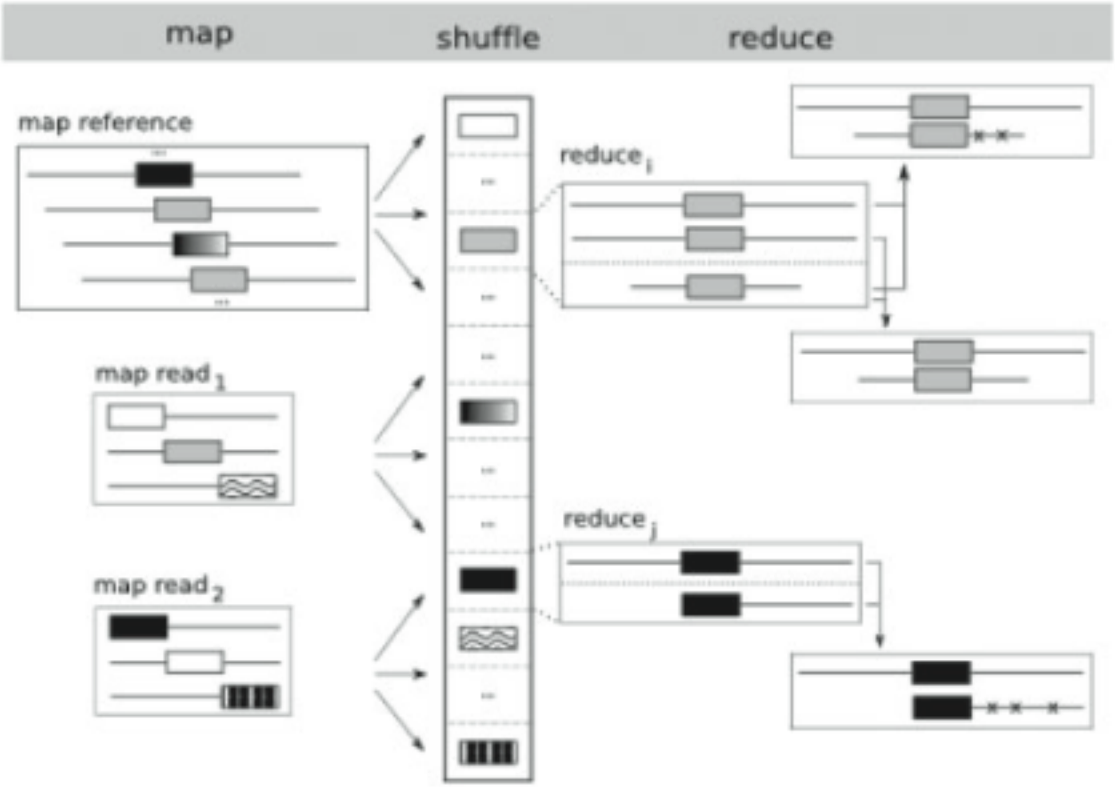
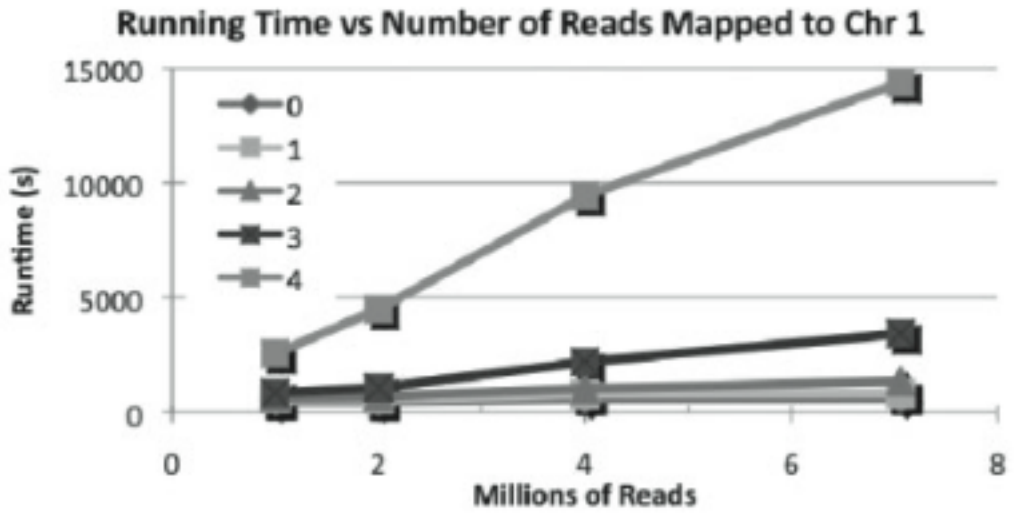


Fig. 2. Overview of the CloudBurst algorithm. The map phase emits k-mers as keys for every k-mer in the reference, and for all non-overlapping k-mers in the reads. The shuffle phase groups together the k-mers shared between the reads and the reference. The reduce phase extends the seeds into end-to-end alignments allowing for a fixed number of mismatches or indels. Here, two grey reference seeds are compared with a single read creating one alignment with two errors and one alignment with zero errors, while the black shared seed is extended to an alignment with three errors.



Results: CloudBurst's running time scales linearly with the number of reads mapped, and with near linear speedup as the number of processors increases. In a 24-processor core configuration, CloudBurst is up to 30 times faster than RMAP executing on a single core, while computing an identical set of alignments. Using a larger remote compute cloud with 96 cores, CloudBurst improved performance by >100-fold, reducing the running time from hours to mere minutes for typical jobs involving mapping of millions of short reads to the human genome.

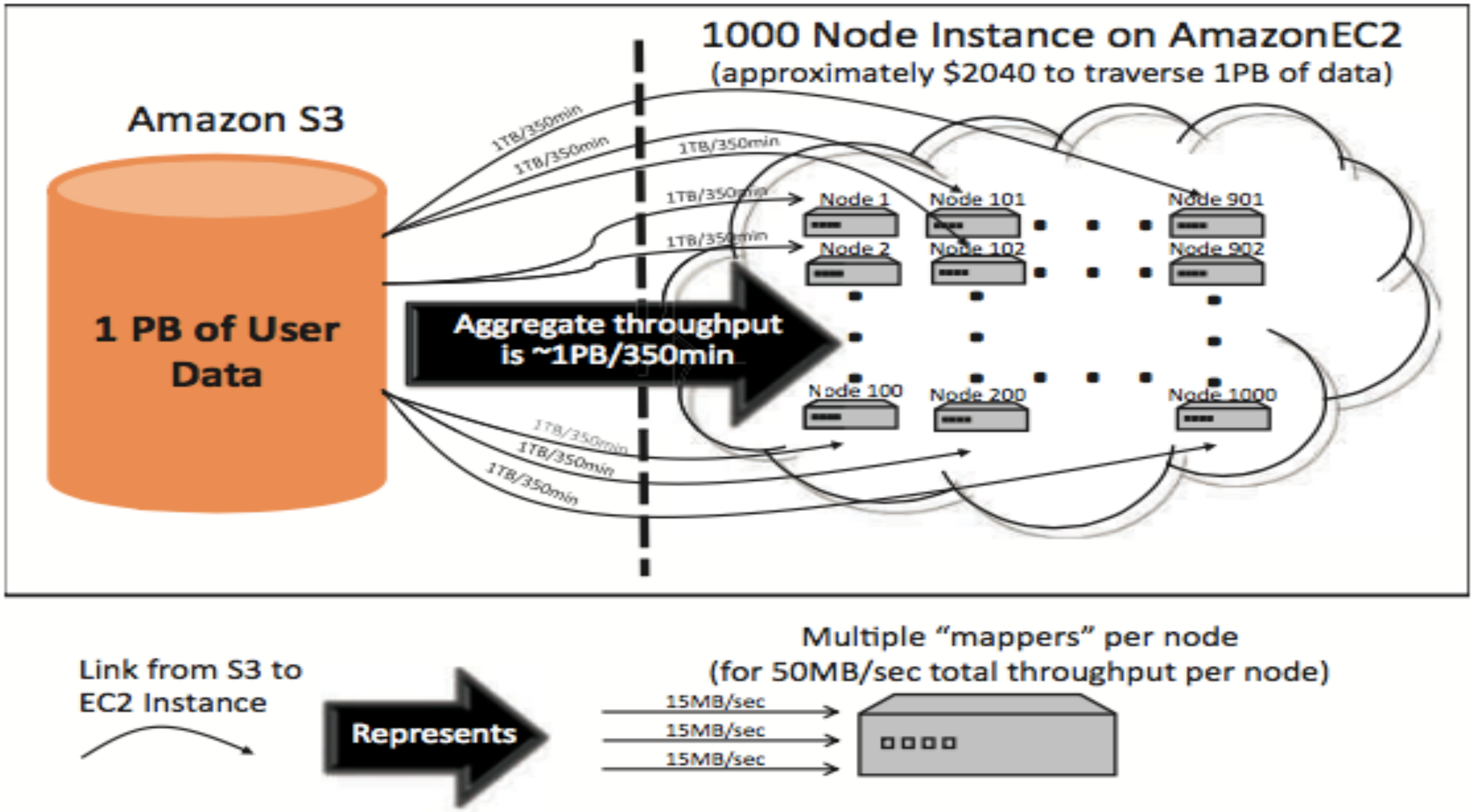

1000 Node Instance on AmazonEC2
(approximately $2040 to traverse 1PB of data)
Amazon S3
1 PB of User Data
Aggregate throughput is ~1PB/350min
1TB/350min
Node 1
Node 2
Node 100
Node 101
Node 102
Node 200
Node 901
Node 902
Node 1000
Link from S3 to EC2 Instance
Represents
Multiple "mappers" per node
(for 50MB/sec total throughput per node)
15MB/sec
15MB/sec
15MB/sec

Figure 1 | **Applying a MapReduce approach in the cloud to solve embarrassingly parallelizable problems.** To traverse a 1 petabyte (PB) data set, Trelles *et al.* mistakenly assume that the 1 PB data set needs to be traversed by every node. The ideal MapReduce application (depicted in the upper panel) instead distributes 1 terabyte (TB) to each of the 1,000 nodes for concurrent processing (the 'map' step in MapReduce). Furthermore, although Trelles *et al.* cite a paper that they claim indicates a 15 MB/s link between storage and nodes[6], the bandwidth quoted appears to be for a single input/output stream only. As shown in the lower panel, best practice is to launch multiple 'mappers' per node to saturate the available network bandwidth[7], which has been previously benchmarked at ~50 MB/s[8] (threefold higher than the 15 MB/s claimed) and consistent with the 90+ MB/s virtual machine (VM)-to-VM bandwidth reported[6]. Each node can process 1 TB at 50 MB/s at $0.34/h; therefore, the back-of-the-envelope calculations of Trelles *et al.* should be updated to state that 1,000 nodes could traverse 1 PB of data in ~350 minutes (not 750 days) at a cost of ~US$2,040 (not $6,000,000).

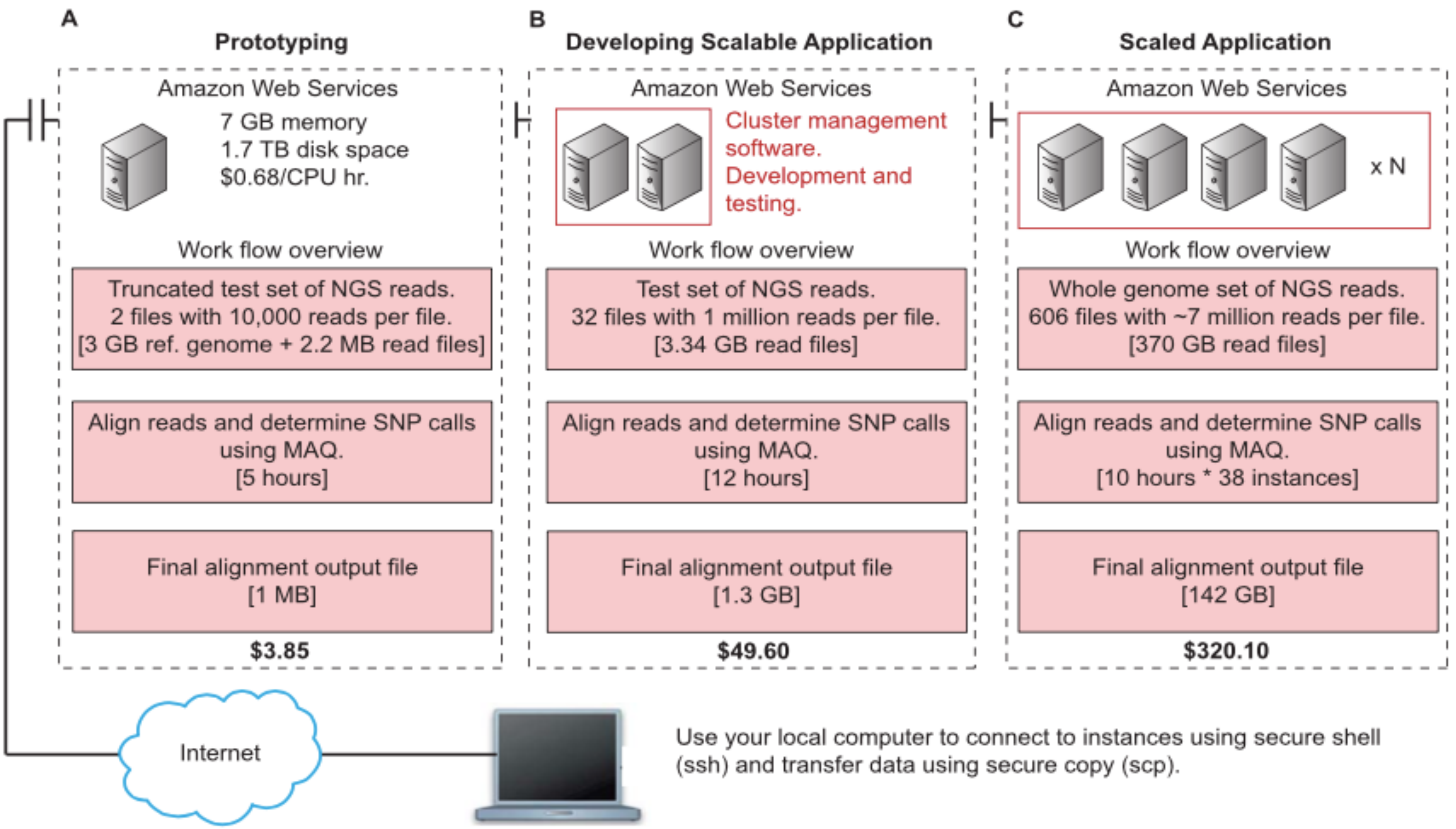

A
Prototyping
Amazon Web Services
7 GB memory
1.7 TB disk space
$0.68/CPU hr.
Work flow overview
Truncated test set of NGS reads.
2 files with 10,000 reads per file.
[3 GB ref. genome + 2.2 MB read files]
Align reads and determine SNP calls using MAQ.
[5 hours]
Final alignment output file
[1 MB]
$3.85
B
Developing Scalable Application
Amazon Web Services
Cluster management software.
Development and testing.
Work flow overview
Test set of NGS reads.
32 files with 1 million reads per file.
[3.34 GB read files]
Align reads and determine SNP calls using MAQ.
[12 hours]
Final alignment output file
[1.3 GB]
$49.60
C
Scaled Application
Amazon Web Services
x N
Work flow overview
Whole genome set of NGS reads.
606 files with ~7 million reads per file.
[370 GB read files]
Align reads and determine SNP calls using MAQ.
[10 hours * 38 instances]
Final alignment output file
[142 GB]
$320.10
Internet
Use your local computer to connect to instances using secure shell (ssh) and transfer data using secure copy (scp).

Figure 1. Step-wise framework for creating a scalable NGS computing application. Using your local computer, ssh into an instance running in AWS. The costs are representative of actual development time, data transfer into and out of the cloud, and the compute time using AWS (Table 1). The costs presented may vary, as AWS frequently updates their pricing structure. (A) An additional 3 hours were included for installing programs and testing the instance for the prototyping phase. (B) An additional 2 hours were included in developing the scalable application to learn how to use the cluster management software. (C) For the final scaled application, we used a 38-instance cluster.
doi:10.1371/journal.pcbi.1002147.g001

StarCluster was created to simplify the cluster creation, management, and job scheduling on AWS

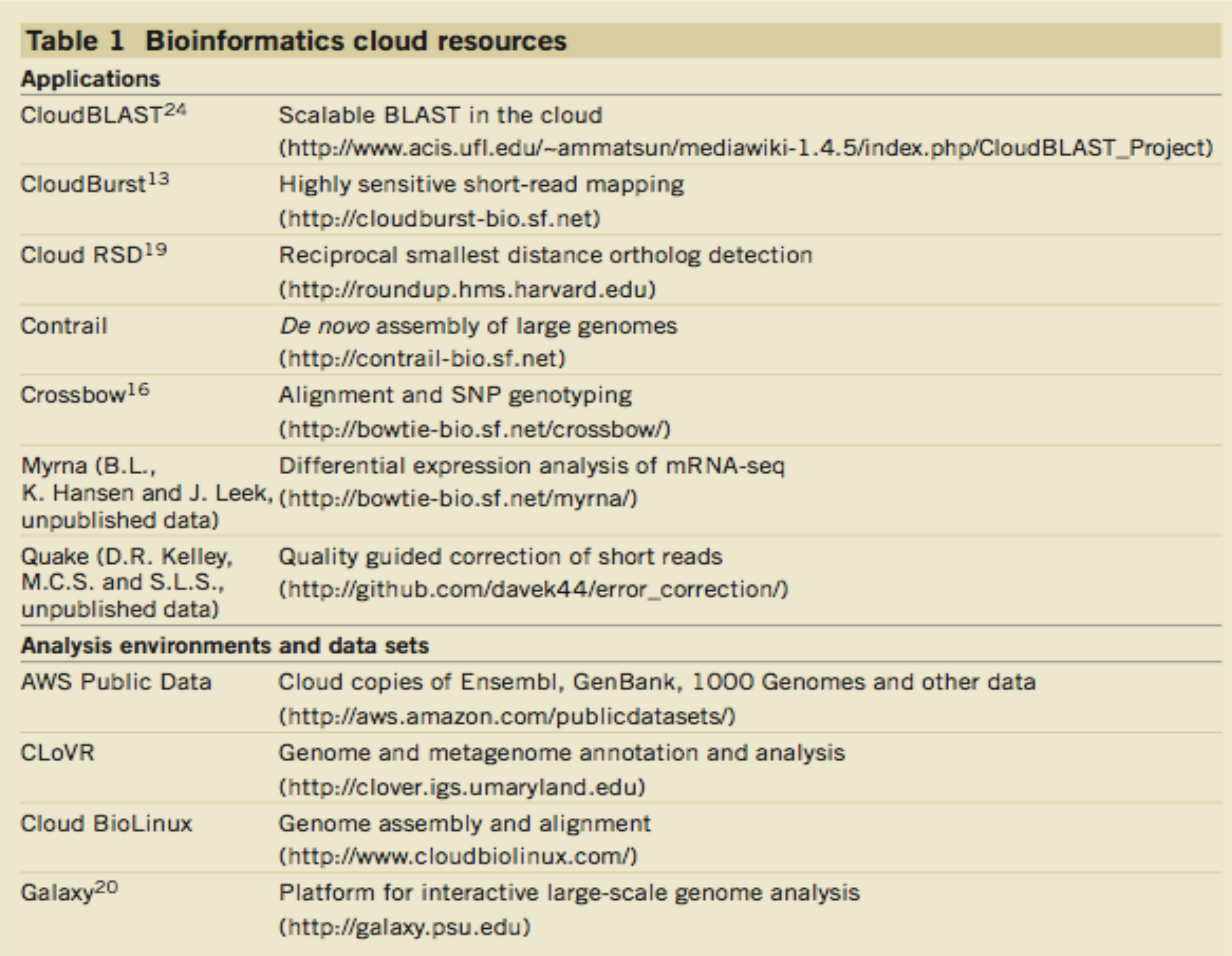


Table 1 Bioinformatics cloud resources

Applications	
CloudBLAST[24]	Scalable BLAST in the cloud (http://www.acis.ufl.edu/~ammatsun/mediawiki-1.4.5/index.php/CloudBLAST_Project)
CloudBurst[13]	Highly sensitive short-read mapping (http://cloudburst-bio.sf.net)
Cloud RSD[19]	Reciprocal smallest distance ortholog detection (http://roundup.hms.harvard.edu)
Contrail	*De novo* assembly of large genomes (http://contrail-bio.sf.net)
Crossbow[16]	Alignment and SNP genotyping (http://bowtie-bio.sf.net/crossbow/)
Myrna (B.L., K. Hansen and J. Leek, unpublished data)	Differential expression analysis of mRNA-seq (http://bowtie-bio.sf.net/myrna/)
Quake (D.R. Kelley, M.C.S. and S.L.S., unpublished data)	Quality guided correction of short reads (http://github.com/davek44/error_correction/)
Analysis environments and data sets	
AWS Public Data	Cloud copies of Ensembl, GenBank, 1000 Genomes and other data (http://aws.amazon.com/publicdatasets/)
CLoVR	Genome and metagenome annotation and analysis (http://clover.igs.umaryland.edu)
Cloud BioLinux	Genome assembly and alignment (http://www.cloudbiolinux.com/)
Galaxy[20]	Platform for interactive large-scale genome analysis (http://galaxy.psu.edu)

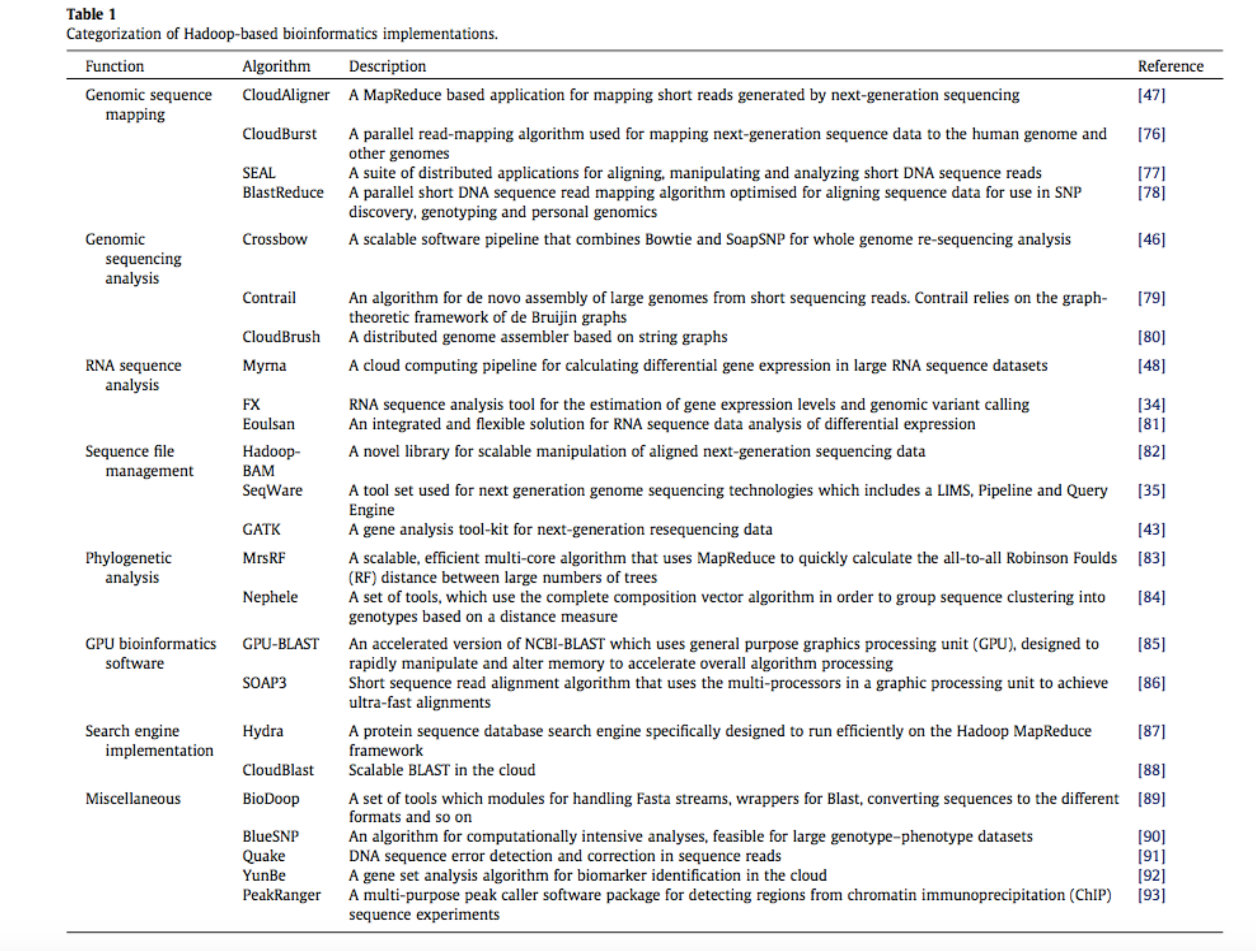
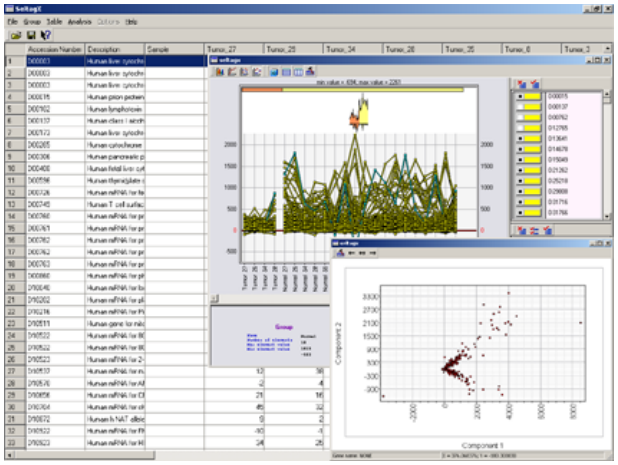


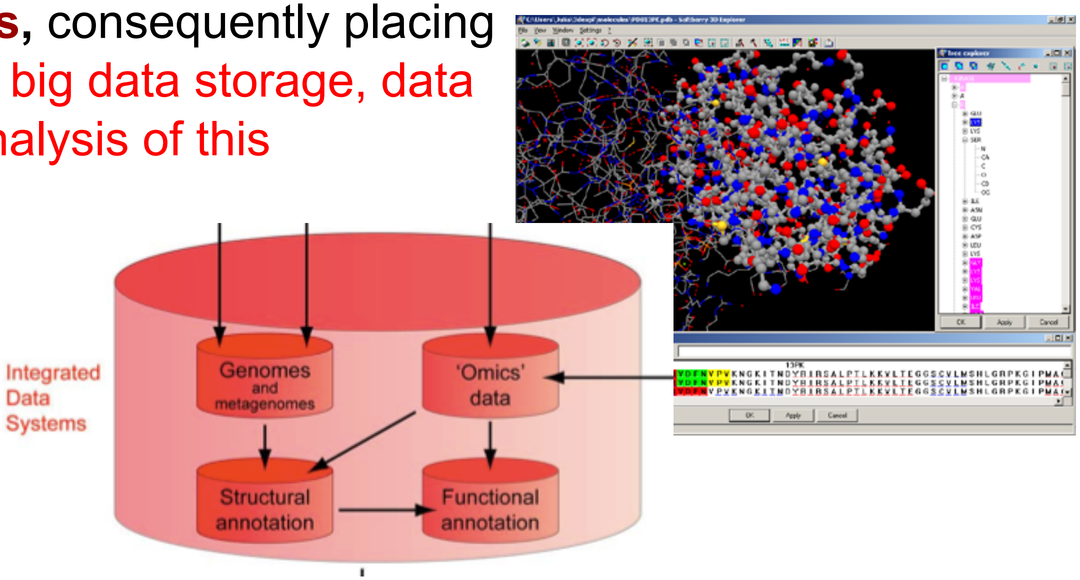
Table 1
Categorization of Hadoop-based bioinformatics implementations.

Function	Algorithm	Description	Reference
Genomic sequence mapping	CloudAligner	A MapReduce based application for mapping short reads generated by next-generation sequencing	[47]
	CloudBurst	A parallel read-mapping algorithm used for mapping next-generation sequence data to the human genome and other genomes	[76]
	SEAL	A suite of distributed applications for aligning, manipulating and analyzing short DNA sequence reads	[77]
	BlastReduce	A parallel short DNA sequence read mapping algorithm optimised for aligning sequence data for use in SNP discovery, genotyping and personal genomics	[78]
Genomic sequencing analysis	Crossbow	A scalable software pipeline that combines Bowtie and SoapSNP for whole genome re-sequencing analysis	[46]
	Contrail	An algorithm for de novo assembly of large genomes from short sequencing reads. Contrail relies on the graph-theoretic framework of de Bruijin graphs	[79]
	CloudBrush	A distributed genome assembler based on string graphs	[80]
RNA sequence analysis	Myrna	A cloud computing pipeline for calculating differential gene expression in large RNA sequence datasets	[48]
	FX	RNA sequence analysis tool for the estimation of gene expression levels and genomic variant calling	[34]
	Eoulsan	An integrated and flexible solution for RNA sequence data analysis of differential expression	[81]
Sequence file management	Hadoop-BAM	A novel library for scalable manipulation of aligned next-generation sequencing data	[82]
	SeqWare	A tool set used for next generation genome sequencing technologies which includes a LIMS, Pipeline and Query Engine	[35]
	GATK	A gene analysis tool-kit for next-generation resequencing data	[43]
Phylogenetic analysis	MrsRF	A scalable, efficient multi-core algorithm that uses MapReduce to quickly calculate the all-to-all Robinson Foulds (RF) distance between large numbers of trees	[83]
	Nephele	A set of tools, which use the complete composition vector algorithm in order to group sequence clustering into genotypes based on a distance measure	[84]
GPU bioinformatics software	GPU-BLAST	An accelerated version of NCBI-BLAST which uses general purpose graphics processing unit (GPU), designed to rapidly manipulate and alter memory to accelerate overall algorithm processing	[85]
	SOAP3	Short sequence read alignment algorithm that uses the multi-processors in a graphic processing unit to achieve ultra-fast alignments	[86]
Search engine implementation	Hydra	A protein sequence database search engine specifically designed to run efficiently on the Hadoop MapReduce framework	[87]
	CloudBlast	Scalable BLAST in the cloud	[88]
Miscellaneous	BioDoop	A set of tools which modules for handling Fasta streams, wrappers for Blast, converting sequences to the different formats and so on	[89]
	BlueSNP	An algorithm for computationally intensive analyses, feasible for large genotype–phenotype datasets	[90]
	Quake	DNA sequence error detection and correction in sequence reads	[91]
	YunBe	A gene set analysis algorithm for biomarker identification in the cloud	[92]
	PeakRanger	A multi-purpose peak caller software package for detecting regions from chromatin immunoprecipitation (ChIP) sequence experiments	[93]

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.