Bioinformatics: Basic Concepts and Recent Trends

Ujjwal Maulik
Dept. of CSE
Jadavpur University

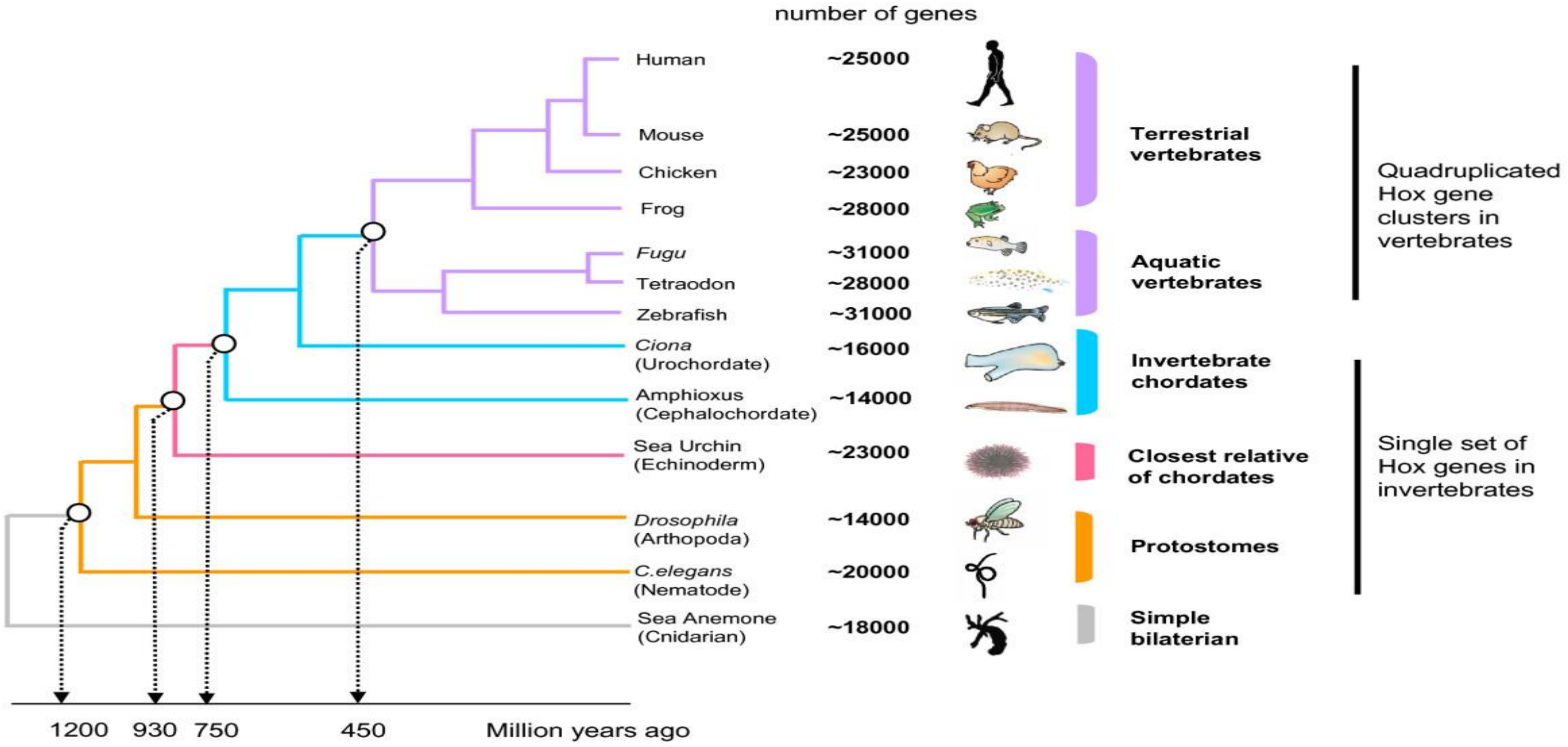
Outline of the Presentation

- Basics of molecular biology
  - Central dogma of molecular biology
- What is bioinformatics and computational biology
- Biological data and important tasks
- Challenges
- Some computational biology methods
- Future trends
- Summary

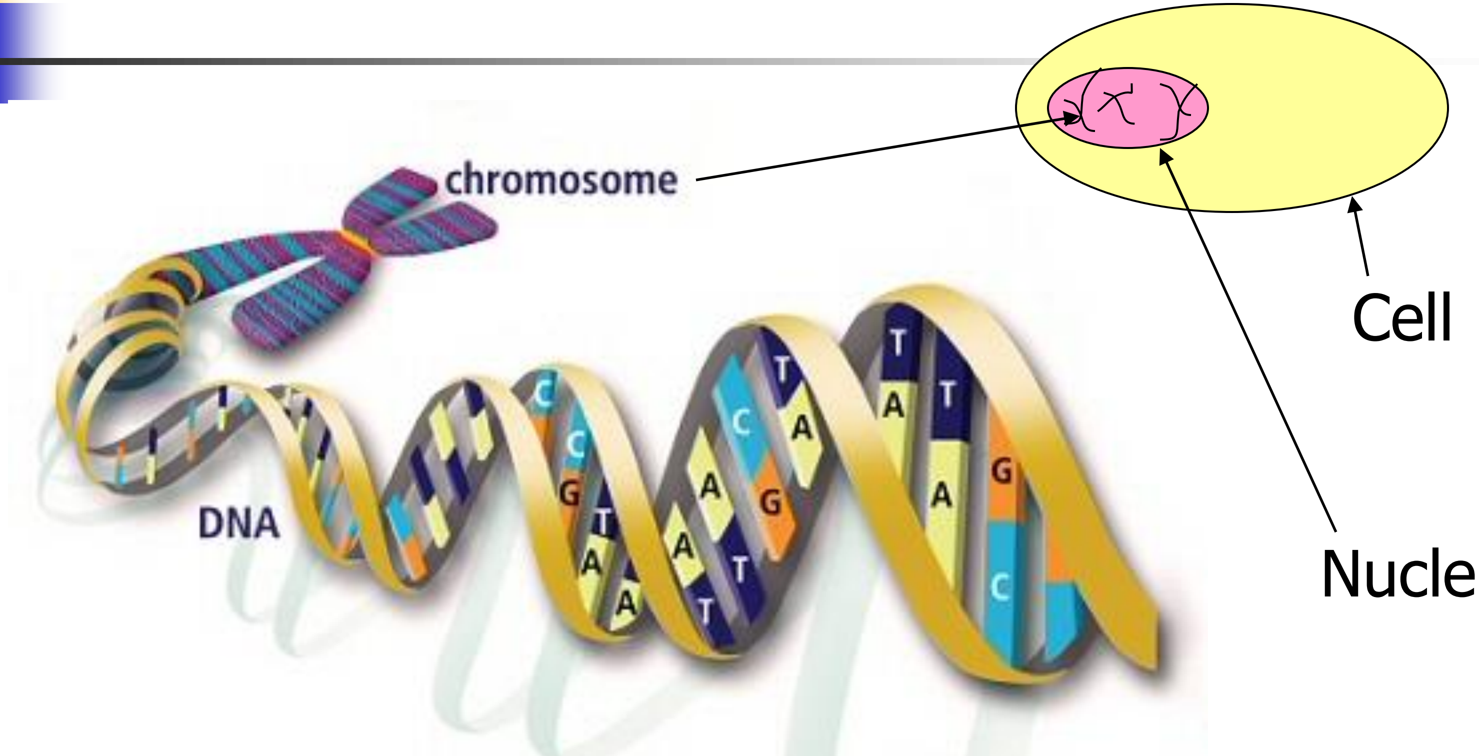
Molecular Biology
-Some basic concepts

- Cells → Tissues → Organs → Organism
- Main actors in the chemistry of life
  - Nucleic Acids
  - Proteins
- Molecular biology research is basically devoted to the understanding of structures and functions of proteins and nucleic acids.

Phylogeny of organisms


number of genes
Human ~25000
Mouse ~25000
Chicken ~23000
Frog ~28000
Fugu ~31000
Tetraodon ~28000
Zebrafish ~31000
Ciona (Urochordate) ~16000
Amphioxus (Cephalochordate) ~14000
Sea Urchin (Echinoderm) ~23000
Drosophila (Arthopoda) ~14000
C.elegans (Nematode) ~20000
Sea Anemone (Cnidarian) ~18000
Terrestrial vertebrates
Aquatic vertebrates
Invertebrate chordates
Closest relative of chordates
Protostomes
Simple bilaterian
Quadruplicated Hox gene clusters in vertebrates
Single set of Hox genes in invertebrates
1200
930
750
450
Million years ago

Deoxy-ribonucleic acid (DNA)



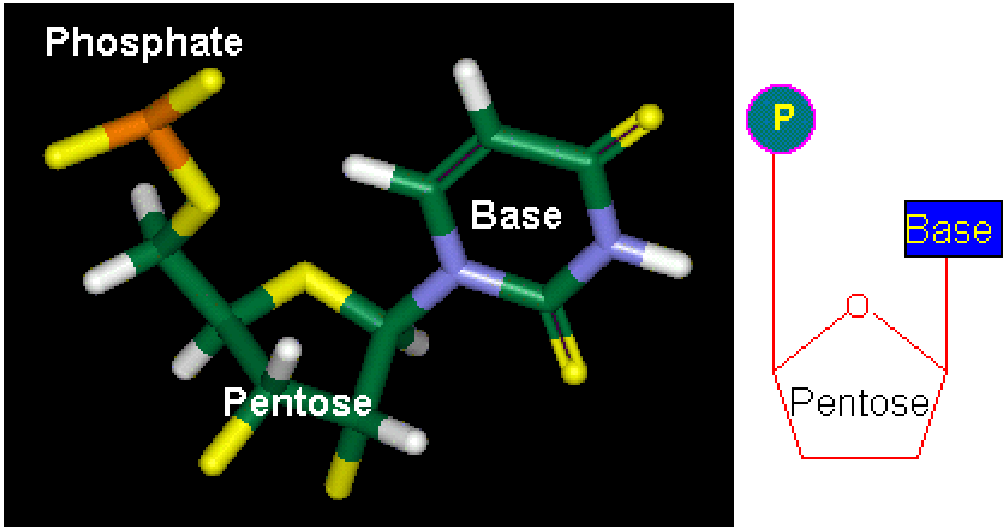
DNA made up of 4 bases – A, T, C and G

A pairs with T, C pairs with G

The entire genetic information is stored in the DNA strand

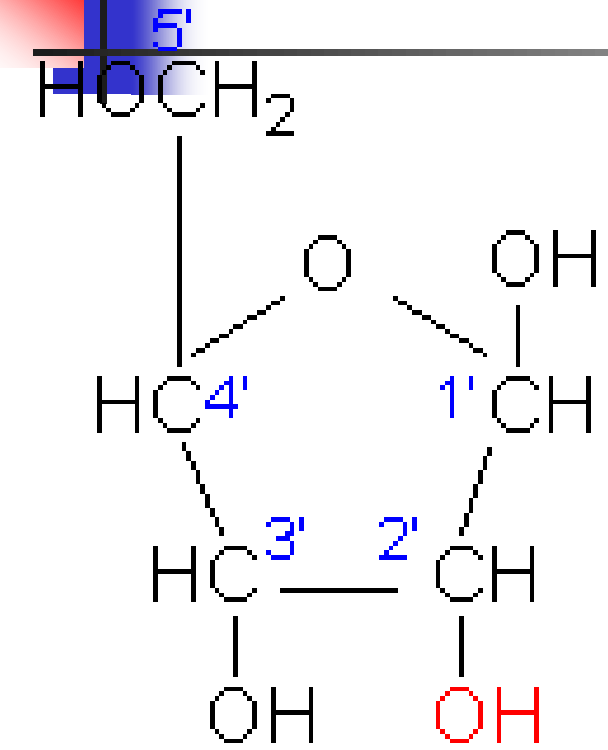
Ack: http://www.childrensmuseum.org/teachers/unitsofstudy/biotechnology

Nucleotide

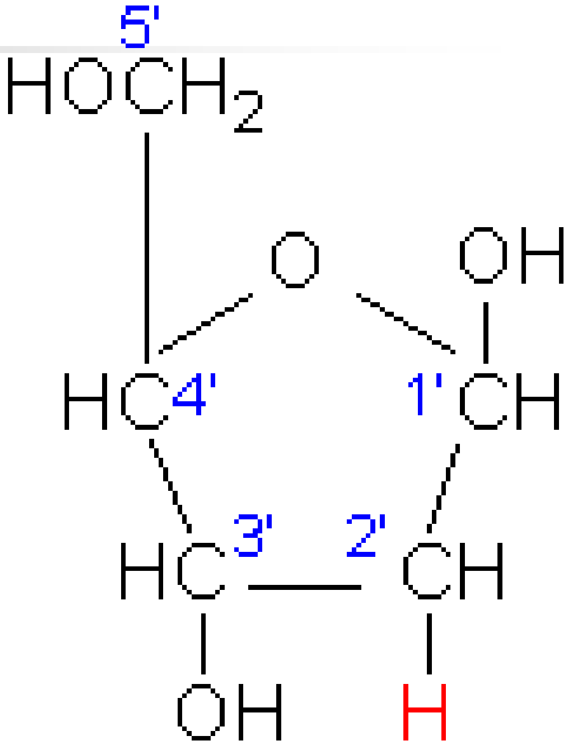


The general structure of nucleotides. Left: computer model. Right: a simplified representation.

If the phospate is removed, then we get **nucleoside**.



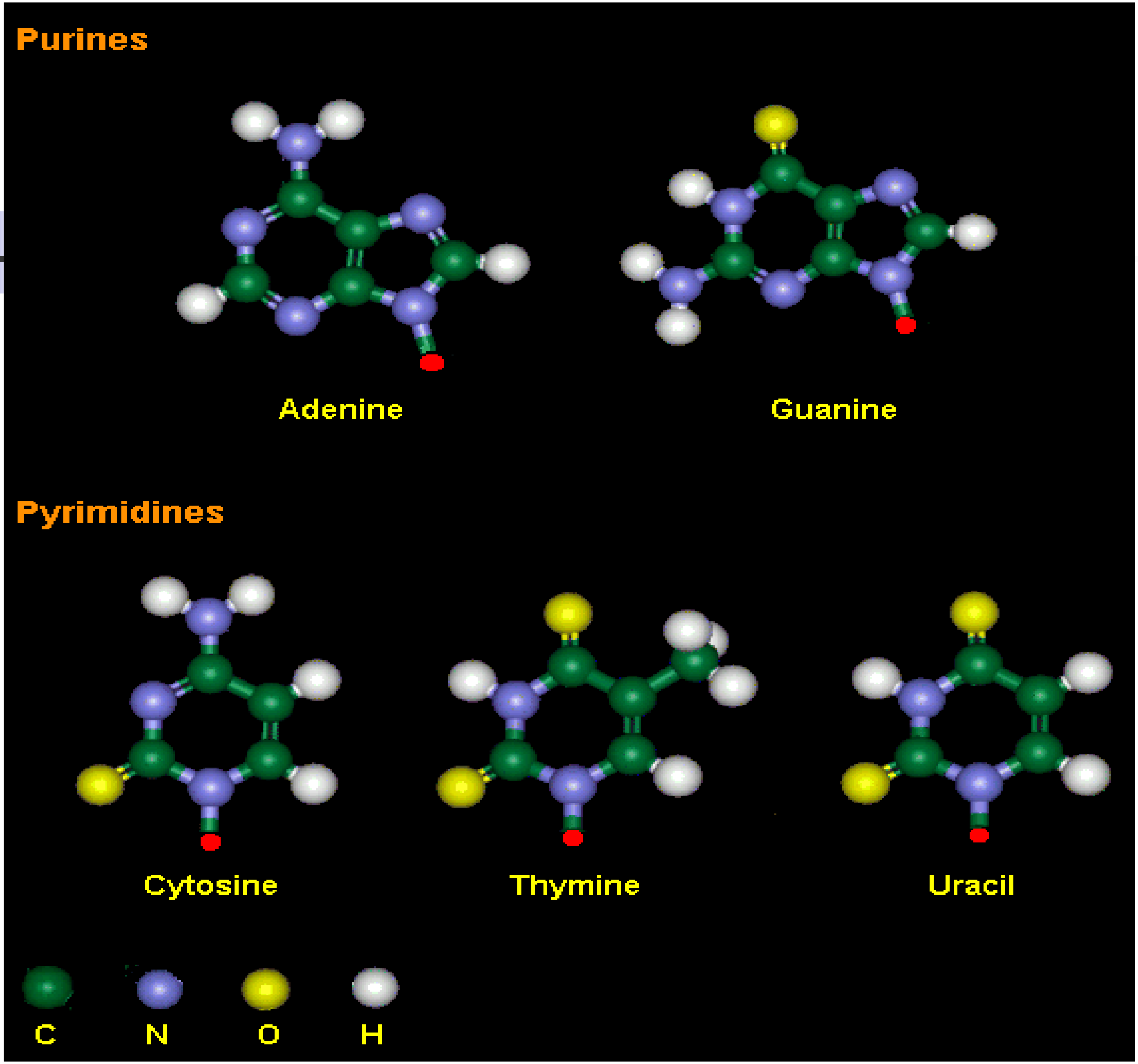
Ribose
(in RNA)



2'-Deoxyribose
(in DNA)

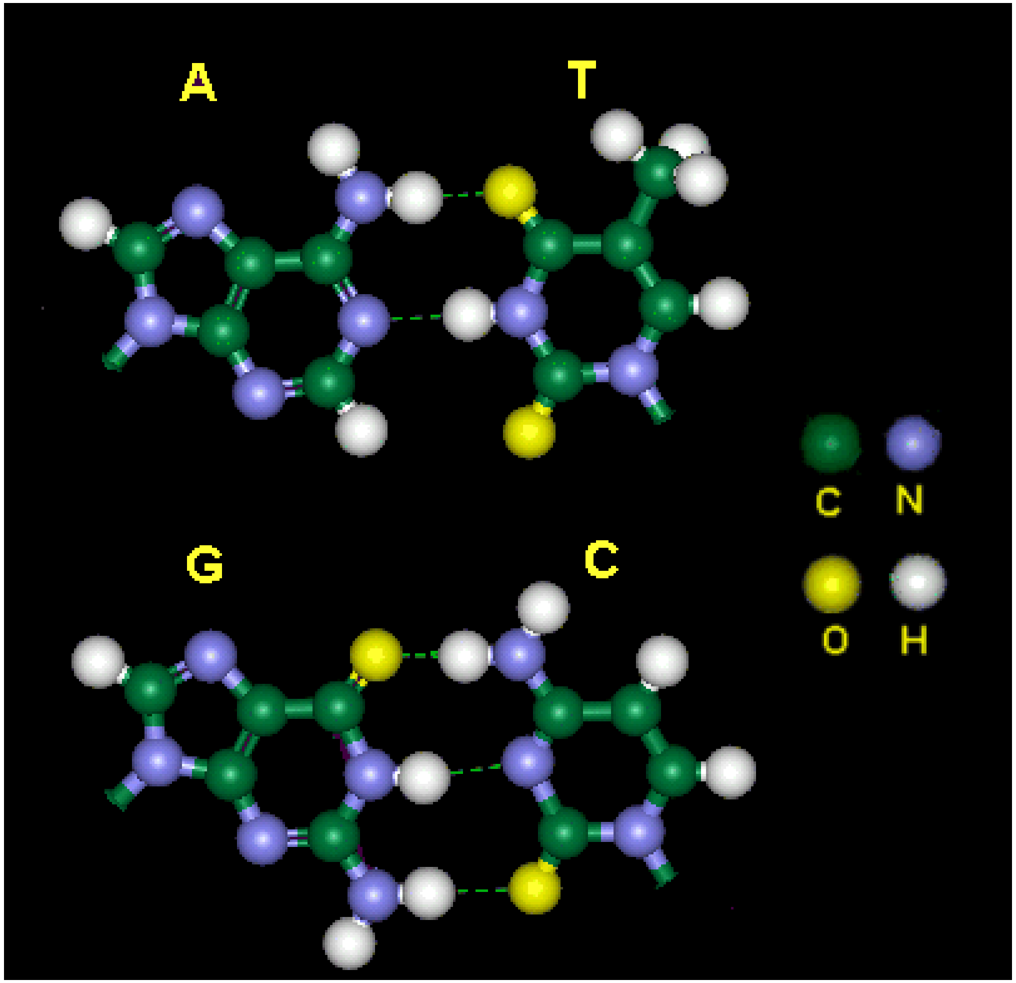
Bases

- Five different bases, each is denoted by a single letter
  - **Adenine (A), Cytosine (C), Guanine (G), Thymine (T), and Uracil (U).**
  - A, C, G and T exist in DNA;
  - A, C, G and U exist in RNA
- A and G contain a pair of fused rings
  - classified as **purines.**
- C, T, and U contain only one ring,
  - classified as **pyrimidines**.

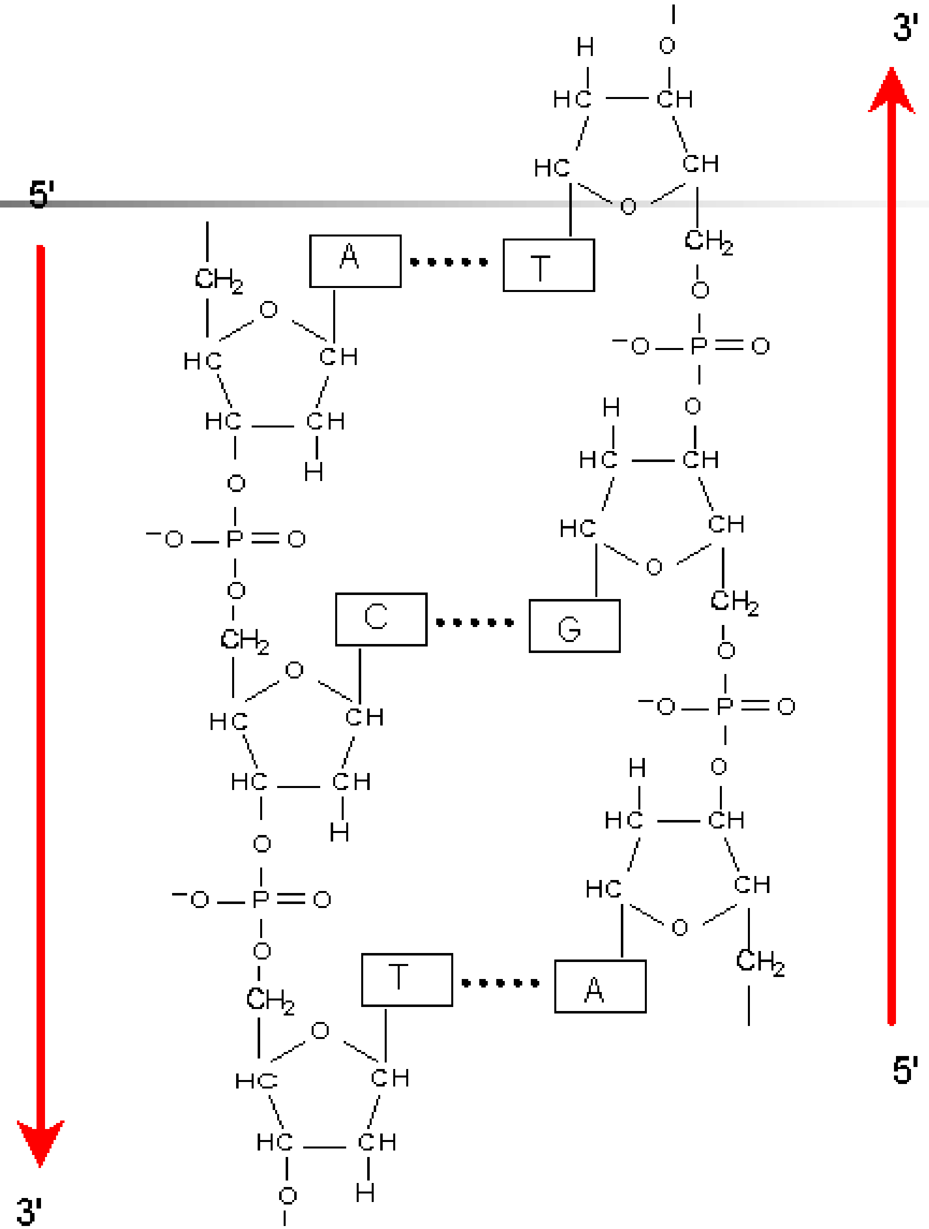

Purines
Adenine
Guanine
Pyrimidines
Cytosine
Thymine
Uracil
C
N
O
H

Pairing of bases – Base pairs

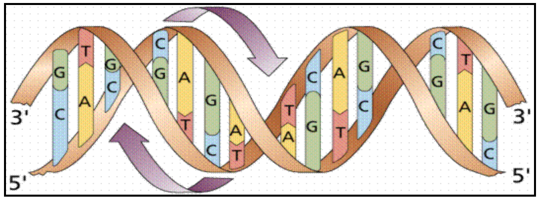
- IN DNA
  - A pairs with T
    - 2 H bonds
  - C pairs with G
    - 3 H bonds
- IN RNA
  - A pairs with U
    - 2 H bonds
  - C pairs with G
    - 3 H bonds
- Other base pairs [e.g., (G:T) and (C:T)] may also form H-bonds
  - strengths are not as much as (C:G) and (A:T) found in natural DNA molecules.



DNA Strand



Central Dogma of Molecular Biology



DNA

transcription

mRNA

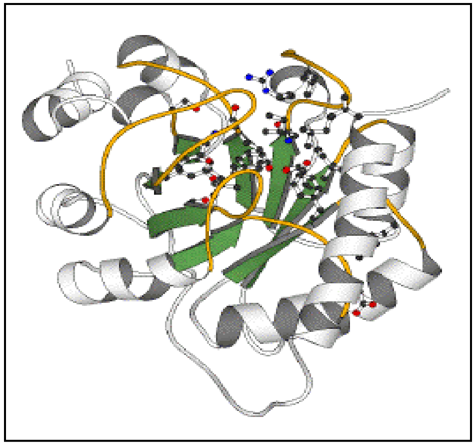
translation

Protein

CCTGAGCCAACTATTGATGAA

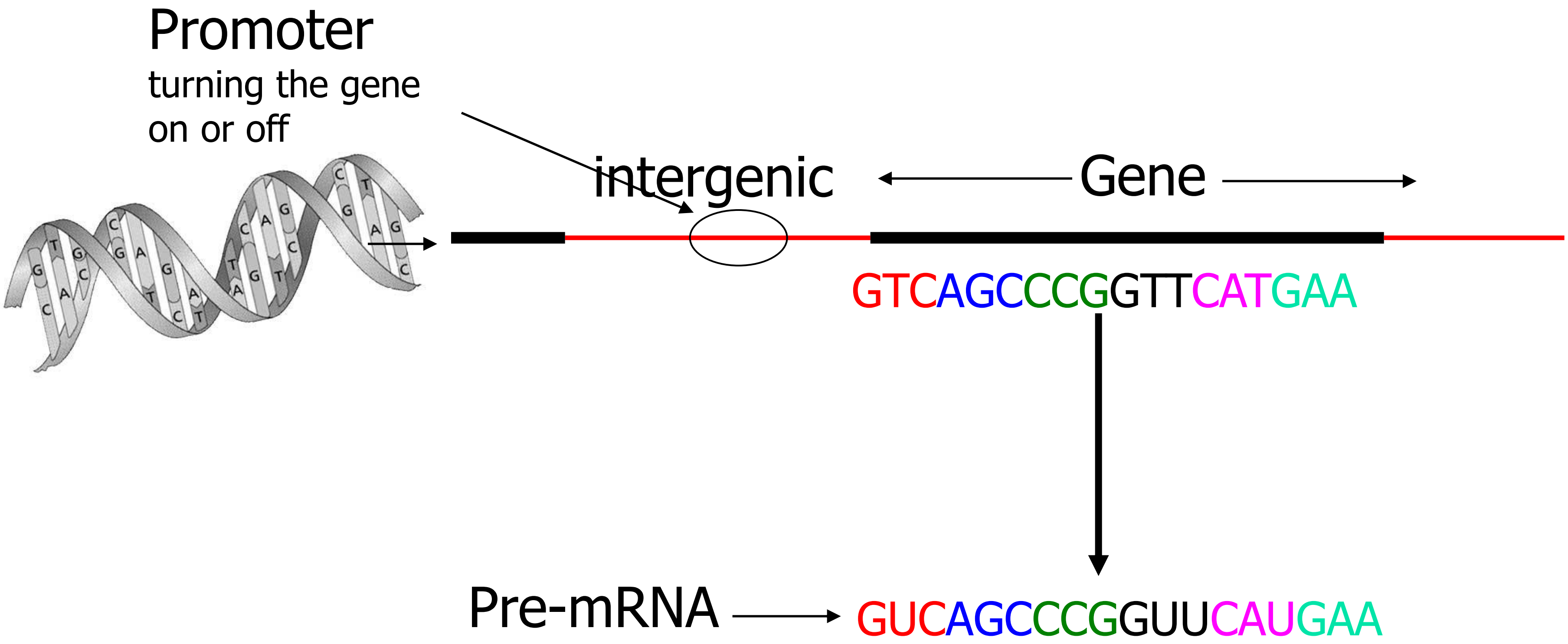
CCUGAGCCAACUAUUGAUGAA

PEPTIDE



Transcription

- Process by which DNA forms RNA



Promoter

turning the gene on or off

intergenic

Gene

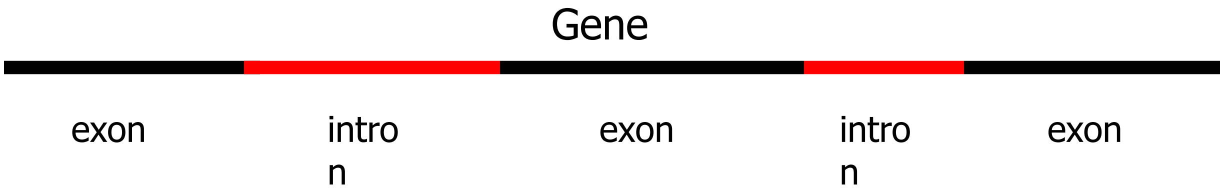
GTCAGCCCGGTTCATGAA

Pre-mRNA

GUCAGCCCGGUUCAUGAA

Transcription contd…

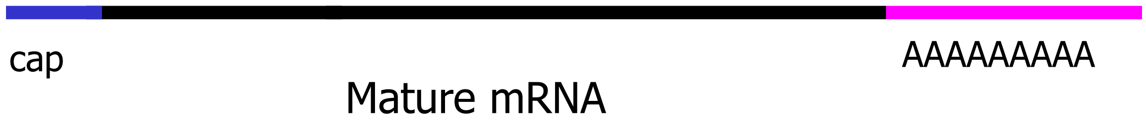
- Pre-mRNA to mature RNA
    - Capping using modified guanine



Gene

exon intron exon intron exon

-
    - Removal of introns
    - Splicing of exons
    - Addition of a polyadenine tail (polyA)



cap AAAAAAAAA

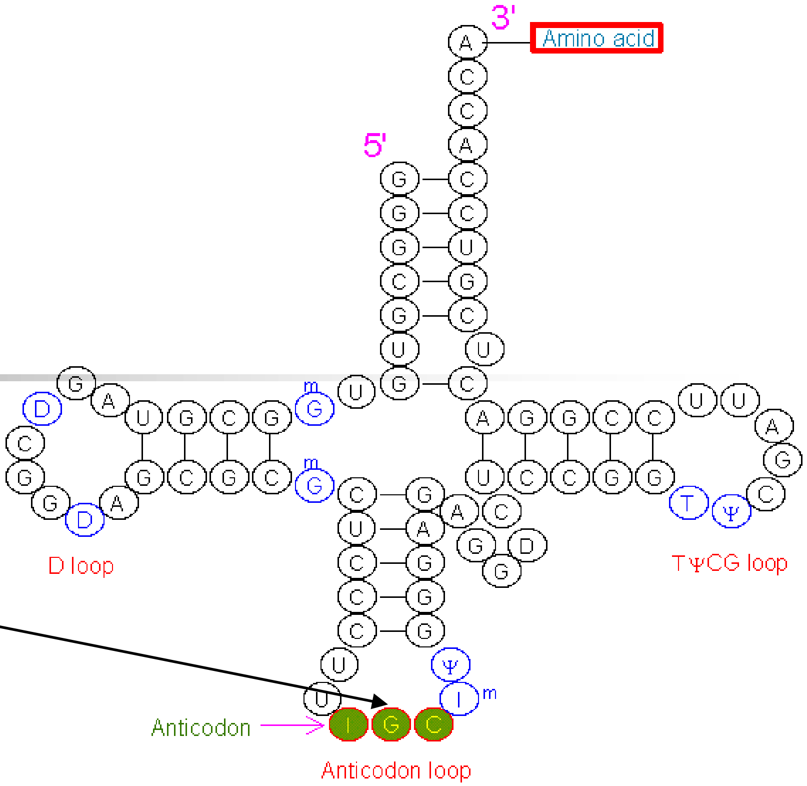
Mature mRNA

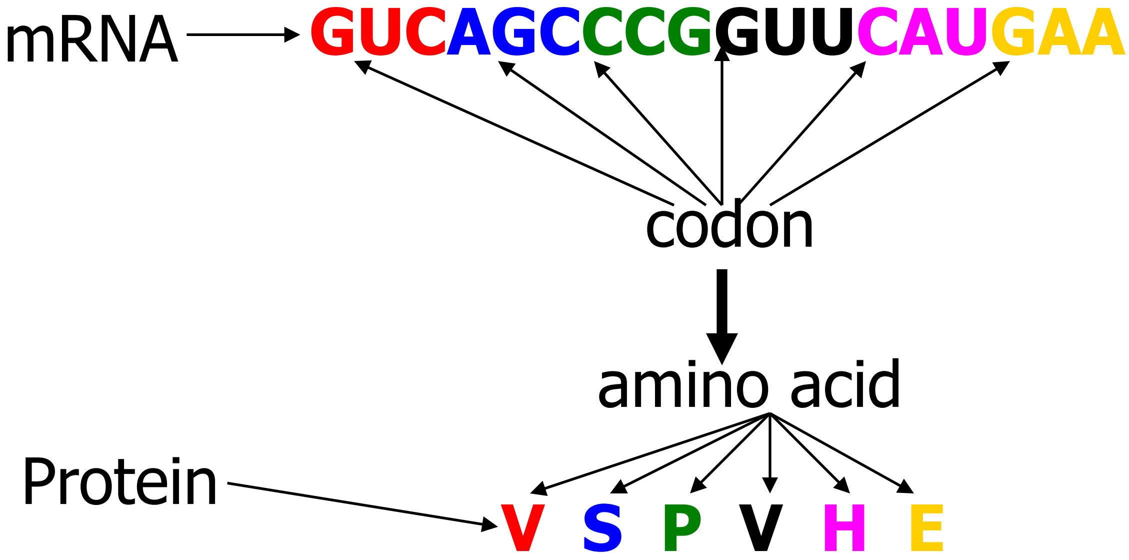
Transcription

Transcription Video

Translation

tRNA carries the anticodon and the Corresponding amino acid





mRNA GUCAGCCCGGUUCAUGAA

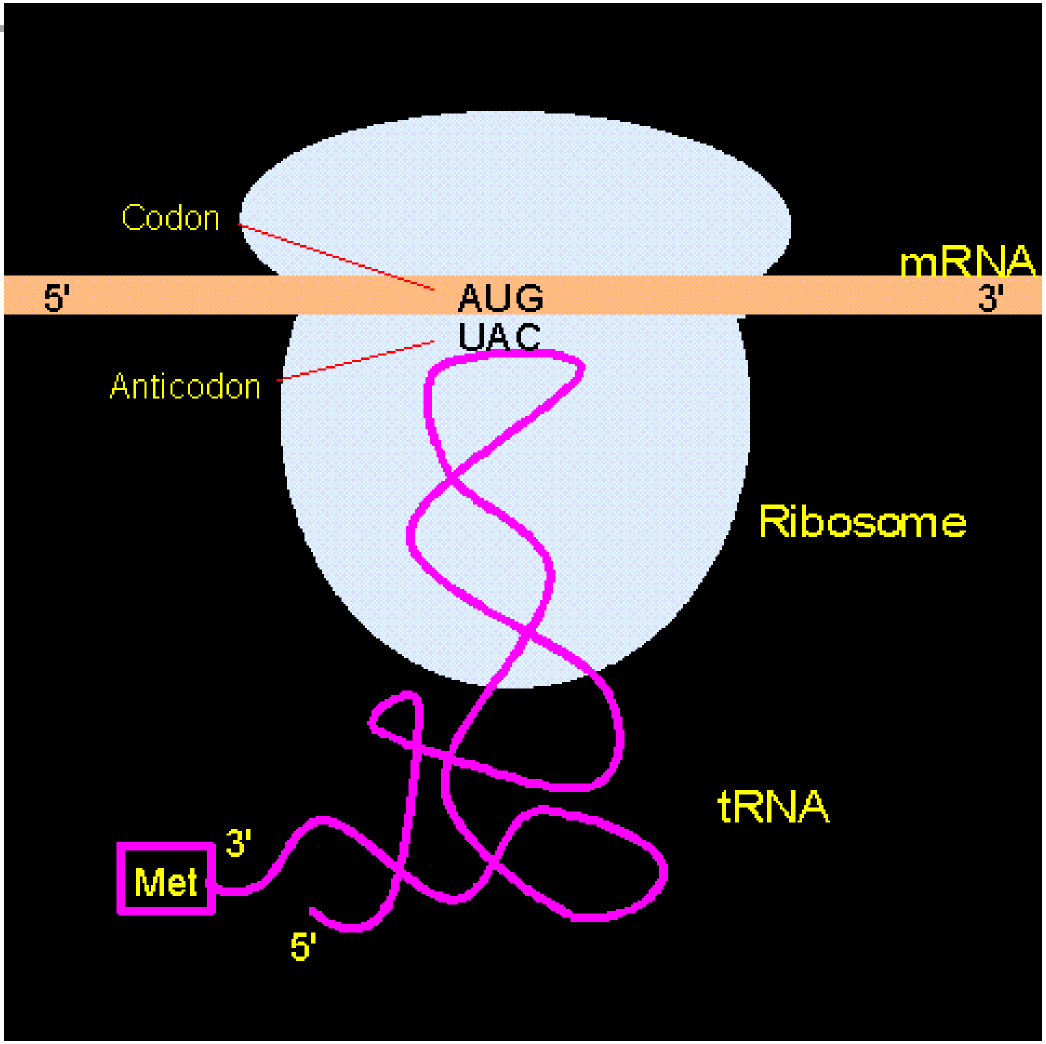
codon

amino acid

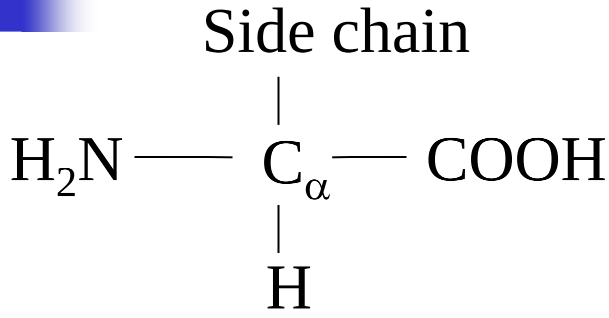
Protein V S P V H E

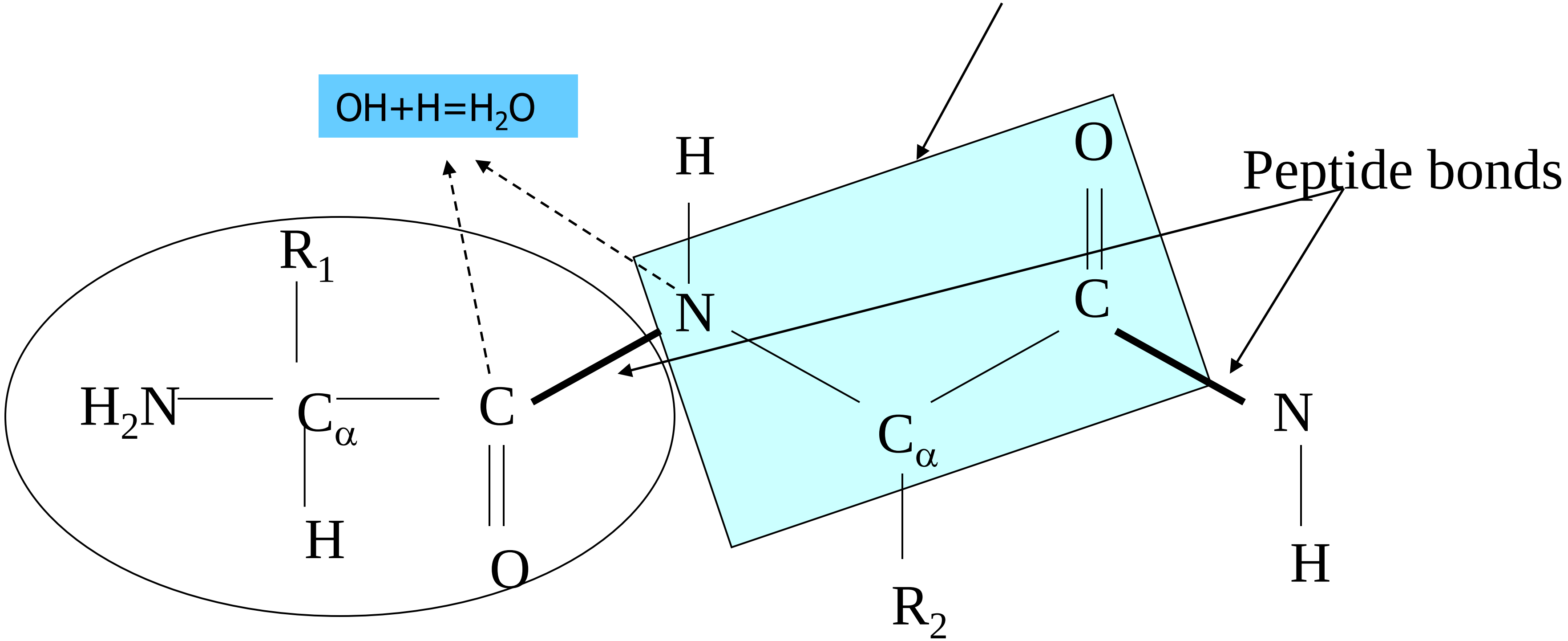
Translation in Ribosome

- rRNA produced in nucleus
- transported to the cytoplasm
- combine with tens of specific proteins
- to form a ribosome



Amino Acid and Proteins

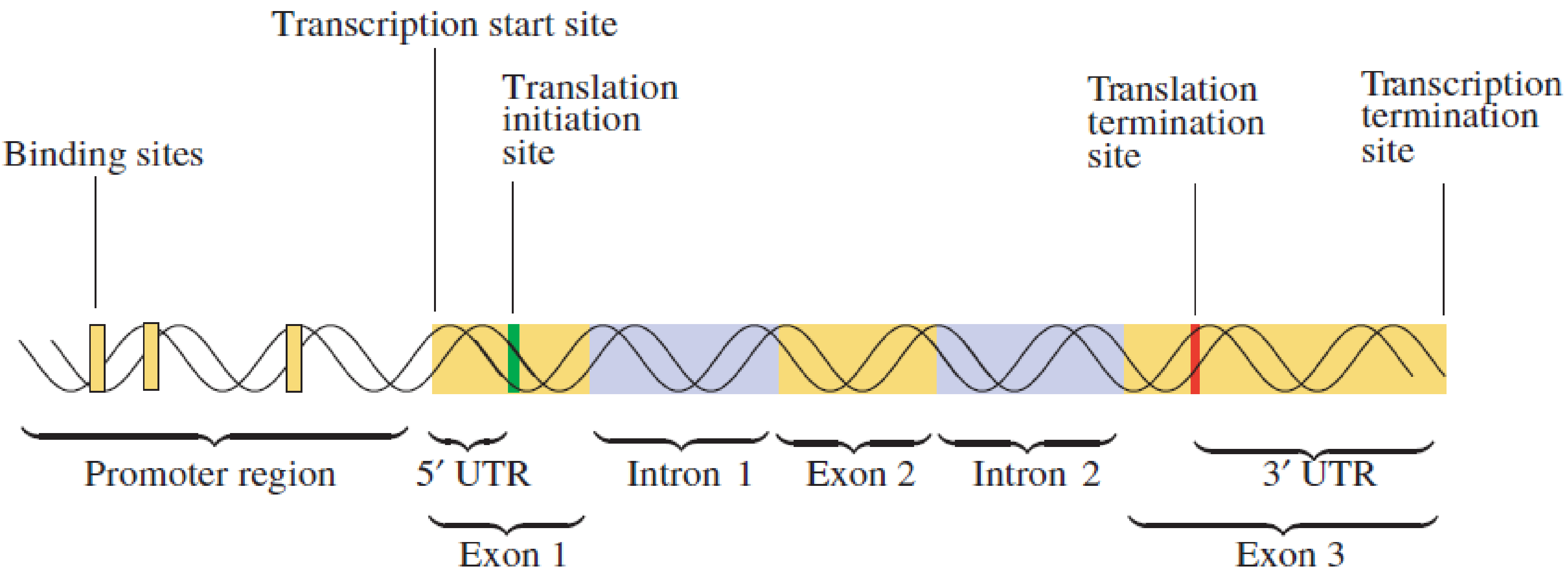




Translation

Translation Video

Snapshot of a Transcriptional Unit



Ack: Zeng et al., Briefings in Bioinformatics, 2009

Types of promoters

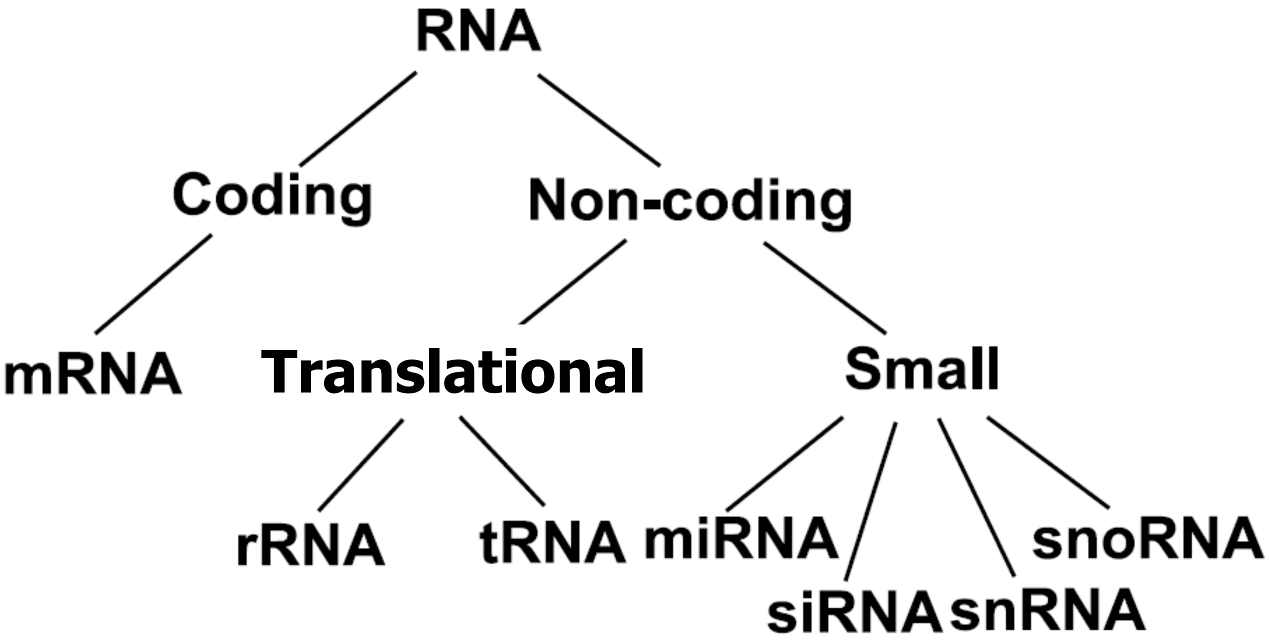
- Core promoter
  - RNA polymerase binding site (within 1 kb from the upstream)
    - Transcription start site

      - Pol I transcribes genes encoding rRNA

      - Pol II transcribes genes encoding mRNA, miRNA, etc.

      - Pol III transcribes genes encoding tRNA, short RNAs, etc.
- Proximal promoter
  - Transcription factor binding site (within 2-3 kb from the upstream)
- Distal promoter
  - Specific transcription factor binding site (within 10kb from the upstream)

Types of RNA


RNA
Coding
Non-coding
mRNA
Translational
Small
rRNA
tRNA
miRNA
siRNA
snRNA
snoRNA

Proteins

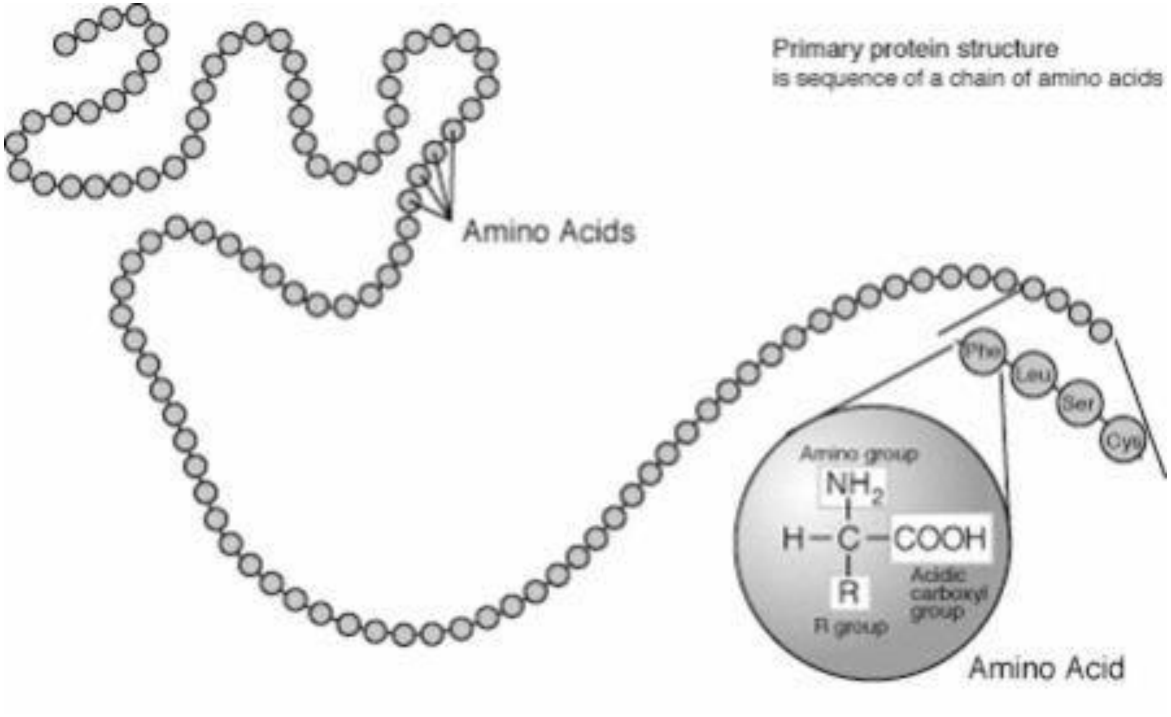
- Protein
  - Polymer of amino acids
  - form a very long chain via peptide linkages
- Functions of Protein
  - enzymes that rearrange chemical bonds
  - carry signals to/from the outside of the cell & within the cell
  - transport small molecules
  - form many of the cellular structures
  - regulate cell process, turn genes on/off and control their rates.
- Protein Structure
  - Primary structure
  - Secondary Structure
  - Tertiary Structure

Primary Structure

- A protein is a linear sequence of amino acids linked together by peptide bonds.
  - covalent bond between the carboxyl group (C) of one amino acid and the amino group (N) of another.
- The peptide bond has particular double bond character and is nearly always in the trans configuration.
  - **trans configuration**: configuration of a geometrical isomer in which two groups are on opposite sides of an imaginary reference line on the molecule.
- Protein can range upto about 5000 amino acids in length, although an average protein is about 350 amino acids length.

Protein chains

- Each protein has a specific sequence of amino acids that are linked together, forming a polypeptide → Primary structure



http://www.mywiseowl.com/articles/Image:Protein-primary-structure.png

Secondary Structure

The driving force behind the formation of a secondary structure is the saturation of backbone hydrogen donors (NH) & acceptors (CO) with intra molecular hydrogen bonds.

There are four types of secondary structural elements

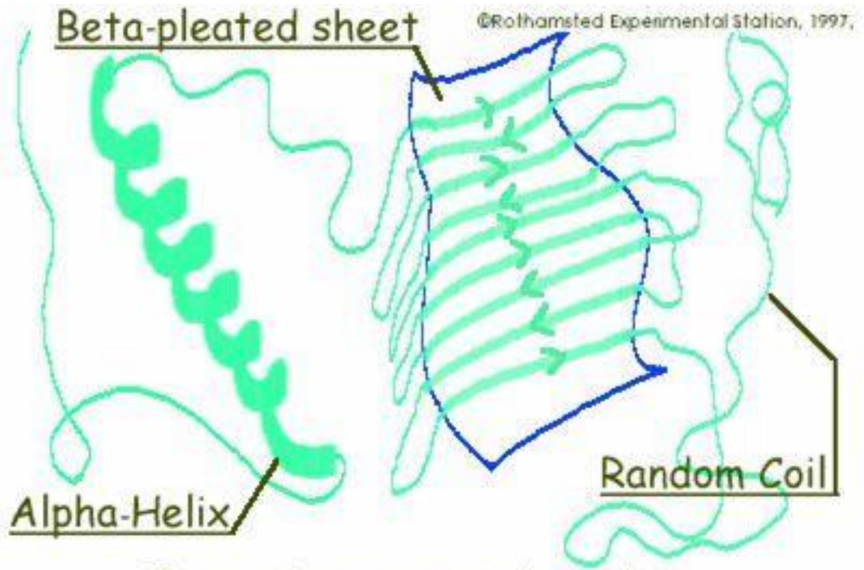
- Alpha (α) Helix.
- Beta (β) Sheet.
- Beta (β) Turn.
- Random coil.

The protein chain folds

Interactions between amino acids in the chain → different secondary structures:

- alpha helices
- beta sheets
- Random coils

} Together usually form the binding and active sites of proteins



The secondary structure is observed in a localised portion of a protein.

Source: http://www.rothamsted.bbsrc.ac.uk/notebook/courses/guide/prot.htm#I

Alpha Helix

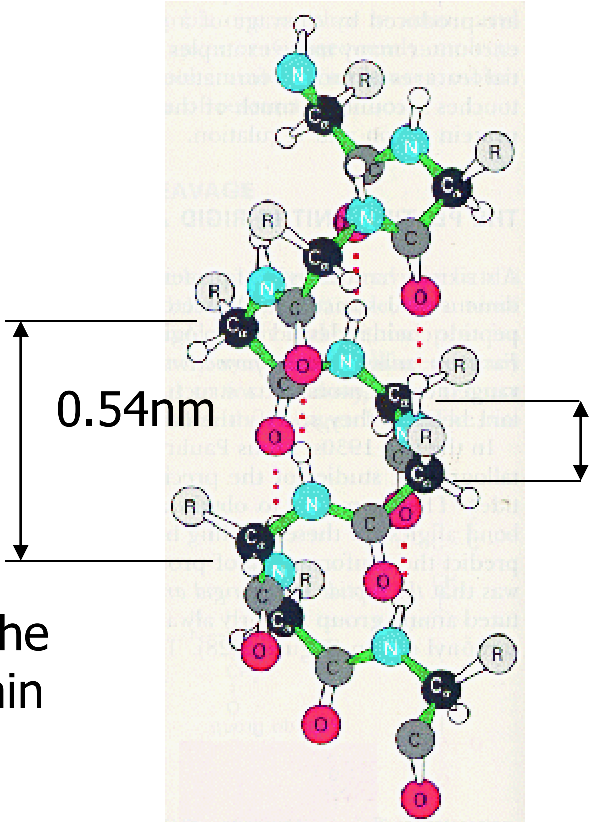
Its main characteristics are:

- Hydrogen bonds between the CO for residue *n* & the NH of residue *n+4.*
- It has 3.6 residues per helical turn covering a distance of 0.54nm
- It is generally a right handed helix

An average alpha helix is 10 residues long, but can range between 4-40 residues in length.

Alpha Helix (contd..)

3.60 amino acids residues per turn



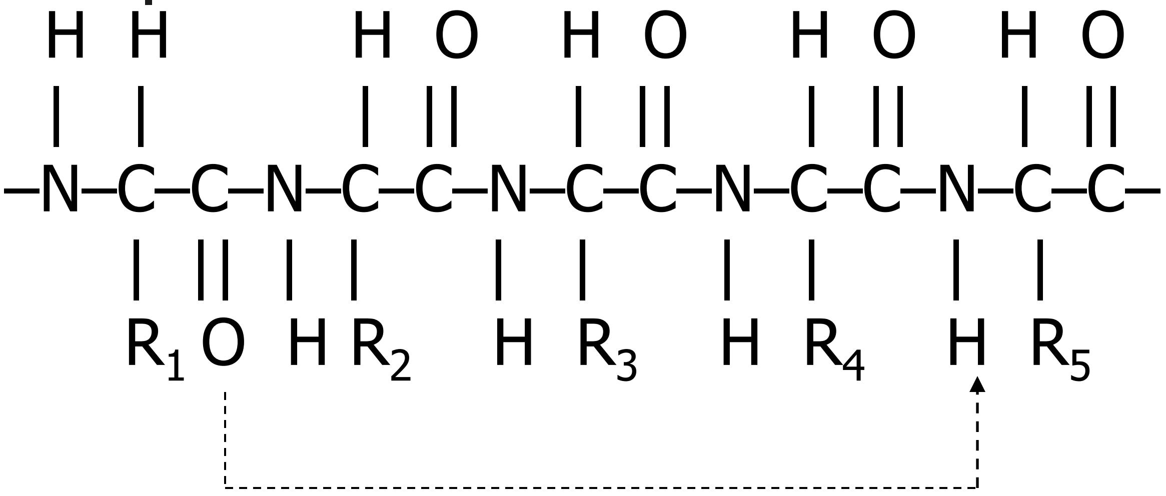
0.54nm

0.15 nm (100^0 rotation per residue)

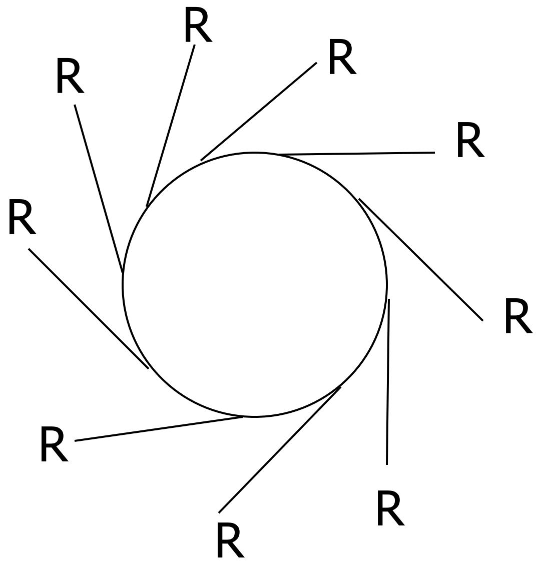
The folding of the polypeptide chain into an α-helix

Ribbon Structure of Alpha helix

Alpha Helix (contd..)



Hydrogen bond

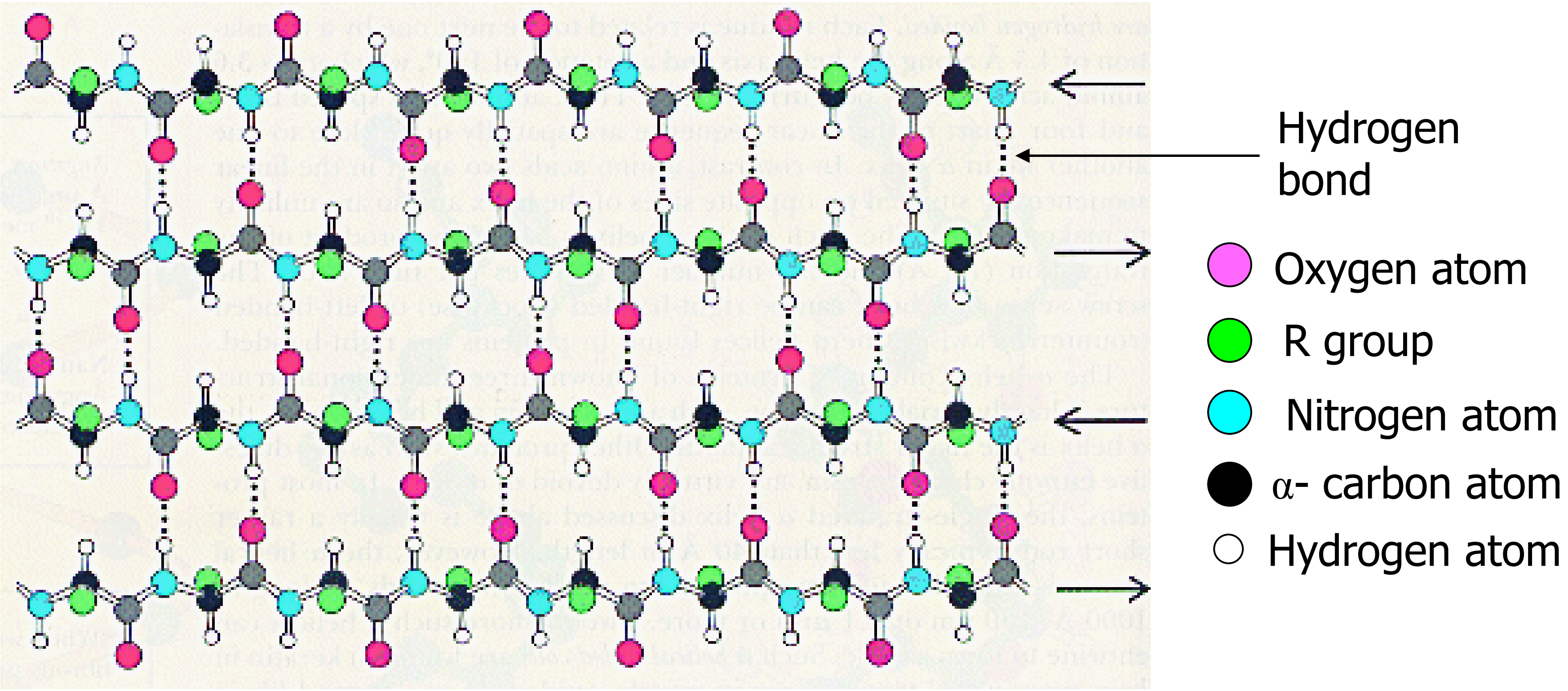


Cross sectional view of an α-helix showing the position of the side chains (R groups) of the amino acids on the outside of the helix

Beta sheet

- Principal component: beta strand
  - sequence of 510 residues in a very extended conformation.
- Beta sheet
  - hydrogen bonding between several beta strands.
- Three ways to form a beta sheet from beta strands.
  - Parallel beta sheet
    - All bonded strands have the same N to C direction
    - separated by long sequence stretches.
    - Hydrogen bonds are equally distanced.
  - Anti parallel beta sheet
    - have alternating sequence directions N to C, C to N etc.
    - can be quite close on the primary sequence
    - The distance between successive bonds is alternating.
  - Mixed beta sheet
    - A mixture of parallel and anti parallel hydrogen bonding
    - About 20% of all beta sheets.

Beta Sheet (contd..)



Hydrogen bond

● Oxygen atom

● R group

● Nitrogen atom

● α- carbon atom

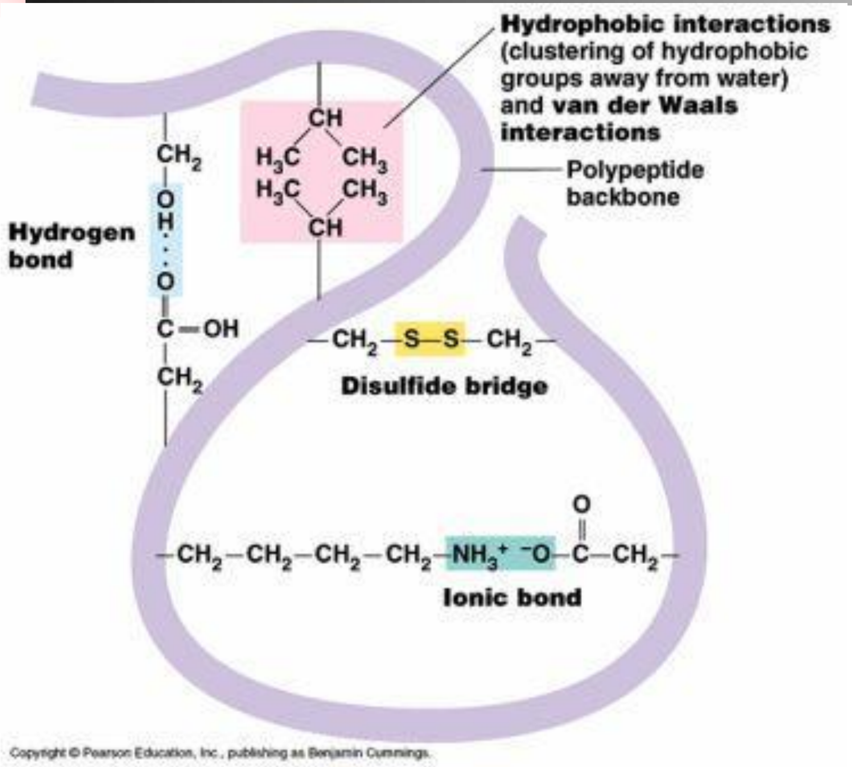
○ Hydrogen atom

Structure of anti parallel beta sheet

Random Coil

- Parts of the protein that are not characterized by any regular hydrogen bonding pattern
- Can be found in the terminal arms loops of the proteins.
- Unstructured regions found between regular secondary structure elements.
- Can be 4 to 20 residues long
  - most loops are not longer than 12 residues.
- Most loops are exposed to the solvent
- Characterized by polar or charged side chains.
- In some cases loops have a functional role, but in many cases they do not.

And folds again!


Hydrophobic interactions (clustering of hydrophobic groups away from water) and van der Waals interactions
Polypeptide backbone
Hydrogen bond
Disulfide bridge
Ionic bond
CH
H3C
CH3
H3C
CH3
CH
CH2
O
H
O
C—OH
CH2
—CH2—S—S—CH2—
—CH2—CH2—CH2—CH2—NH3+ −O—C—CH2—
O

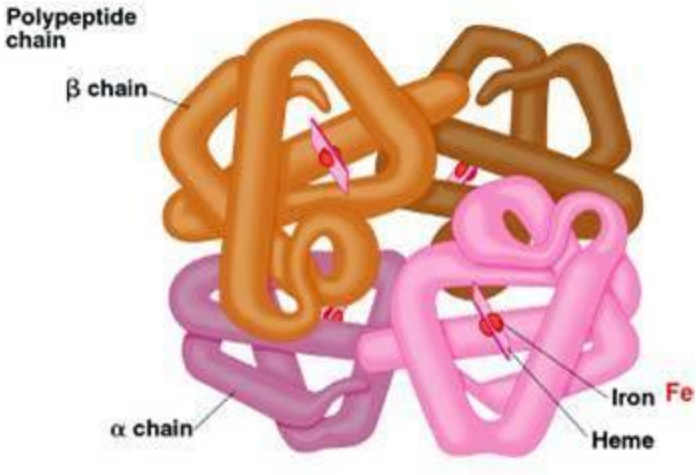
Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

- After folding, amino acids that were distant can become close
- Now the protein chain has a 3D shape that is required for it to function correctly

Source: imuwinnipeg.ca/~simmons/ cm1503/proteins.htm

http://www.gtac.edu.au/site/bioinformatics/bio_task_10/bio_task_10.ppt

The final protein…



(b) Hemoglobin

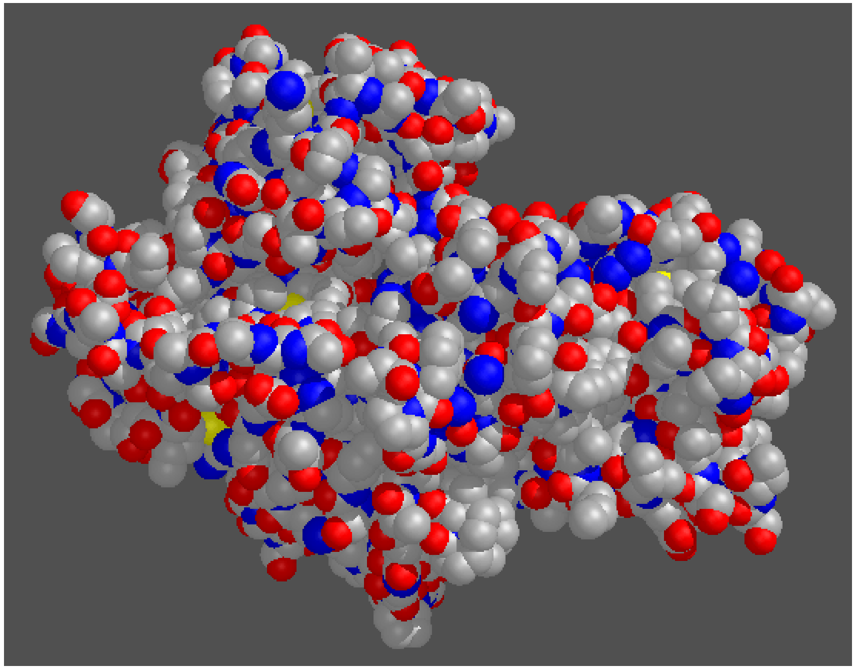
The final protein may be made up of more than one polypeptide chain.

The polypeptide chains may be the same type or different types.

Source: http://fig.cox.miami.edu/~cmallery/150/chemistry/hemoglobin.jpg

Tertiary Structure

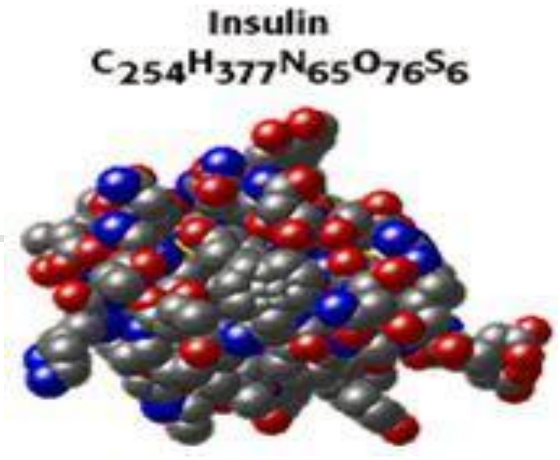
- Full 3-dimensional folded structure of the polypeptide chain.
- Secondary structures of proteins often constitute distinct domains.
- The tertiary structure also describes the relationship between different domains within a protein.
- Interactions are typically governed by several forces, including
  - Hydrogen Bonding
  - Hydrophobic interactions
  - Electrostatic interactions
  - Van der waals forces



Protein Structure

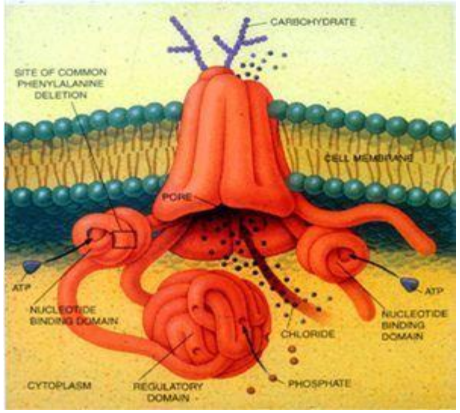
- Function of the protein depends on the structure of the molecule
- Each protein molecule has a characteristics 3D shape
  - That determines its functionality
- Protein folds into different 3d shapes and sizes, depending on the interactions between the component amino acids

Examples of Protein Function



Hormones

Insulin binds to receptors on cell membranes signalling cells to take up glucose from the blood

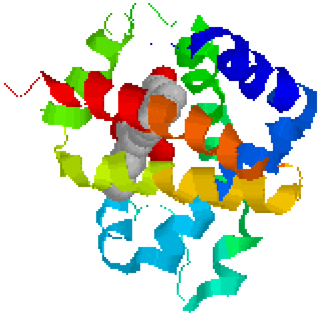

CARBOHYDRATE
SITE OF COMMON PHENYLALANINE DELETION
CELL MEMBRANE
PORE
ATP
ATP
NUCLEOTIDE BINDING DOMAIN
NUCLEOTIDE BINDING DOMAIN
CHLORIDE
CYTOPLASM
REGULATORY DOMAIN
PHOSPHATE

Protein Channels

Regulate movement of substances across the plasma membrane. e.g. The CFTR protein pumps ions across membranes

Transport

Haemoglobin (far right) in red blood cells transports oxygen to cells around the body

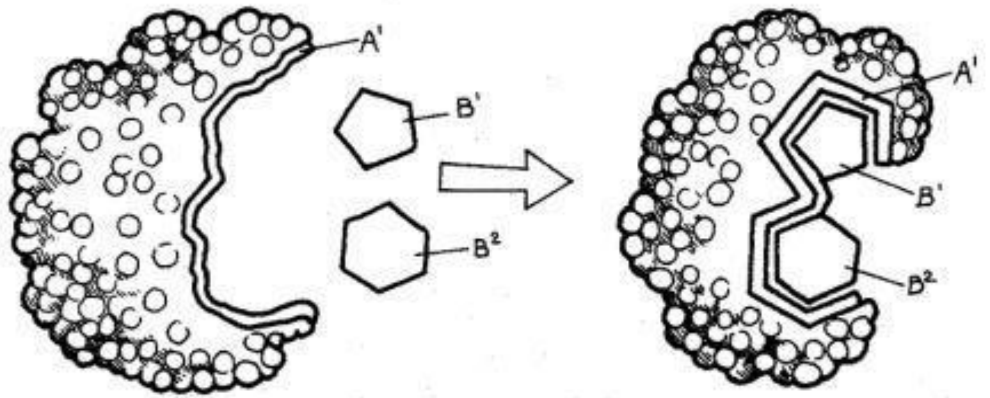


Source: http://www.umass.edu/microbio/chime/

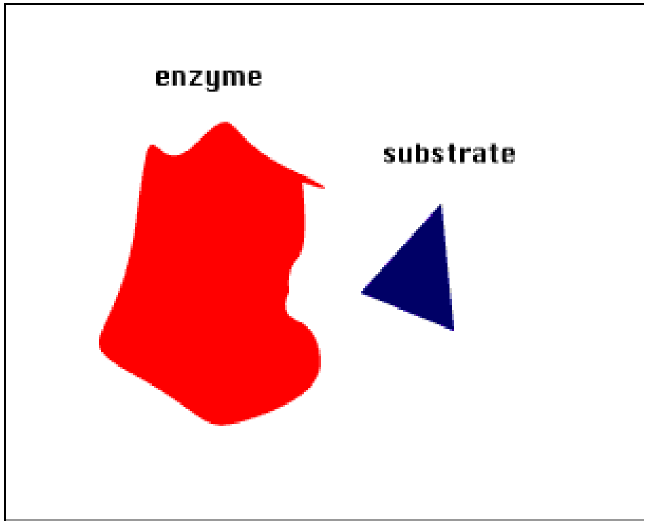
How enzymes do it!

- Enzyme proteins have specific sites where all the action happens. We call this the **active site**. Molecules that need to be ripped apart or put together enter the active site.
- Each protein has a specific shape so it will only perform a specific job.

Joining things together



Ripping things apart



http://chsweb.lr.k12.nj.us/mstanley/outlines/enzymesap/Enzymesap.html
http://academic.brooklyn.cuny.edu/biology/bio4fv/page/active_.html

Computational Biology and Bioinformatics

- Computational biology is an interdisciplinary field that applies the techniques of computer science, applied mathematics, and statistics to address problems inspired by biology.
  - http://en.wikipedia.org/wiki/Computational_biology

- Bioinformatics:refers to the creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data
  - Bioinformatics deals with the applications of algorithms and statistical techniques to biological datasets that typically consist of large numbers of DNA, RNA, or protein sequences.

Biological Data - Sequences

- DNA Sequences
  - TACGAATTGATCCCGCGCGGGTATACAT
    - Genbank: http://www.ncbi.nlm.nih.gov/Genbank, DDBJ, DNA databank of Japan - www.ddbl.nig.ac.jp, EMBL, European Molecular Biology Laboratory – www.ebi.ac.uk/embl, UCSC Genome Browser
- RNA Sequences
  - UACGAAUUGAUCCCGCGCGGGUAUACAU
    - UCSC Genome Browser
- Protein Sequences
  - Atpase superfamily sequence

>
MSVQVKLTKNSFRLEKQKLARLQTYLPTLKLKKALLQAEVQNAVKDAAECDKDYVQAYER
IYAFAELFSIPLCTDCVEKSFEIQSIDNDFENIAGVEVPIVREVTLFPASYSLLGTPIWL
DTMLSASKELVVKKVMAEVSKERLKILEEELRAVSIRVNLFEKKLIPETTKILKKIAVFL
SDRSITDVGQVKMAKKKIELRKARGDECV

  - PIR, Protein Information Resource – pir.georgetown.edu

Biological Data - Structures

- DNA
  - NDB, Nucleic acid database - a repository of three dimensional structural information about nucleic acids, 4585 Structures as on Jan 6, 2010
  - http://ndbserver.rutgers.edu/
- RNA
  - RNA World Website, http://www.imb-jena.de/RNA.html
- Protein
  - PDB: RCSB Protein Databank - *www.rcsb.org/*

```
ATOM   6     C   CG1 . VAL A 1 86   ? 5.241   16.199   -18.127 1.00 70.13   ? ? ? ? ? ? 86
      VAL A CG1 1
ATOM   7     C   CG2 . VAL A 1 86   ? 4.382   18.121   -16.736 1.00 65.59   ? ? ? ? ? ? 86
      VAL A CG2 1
ATOM   8     N   N   . ASP A 1 87   ? 1.404   18.616   -17.326 1.00 82.71   ? ? ? ? ? ? 87
      ASP A N   1
ATOM   9     C   CA  . ASP A 1 87   ? 0.884   19.974   -17.228 1.00 84.68   ? ? ? ? ? ? 87
      ASP A CA  1
```

Biological Data – Expression Profiling Data

- Gene expression values
    - Proportional to the amount of mRNAs produced by a gene
        - Variation over time
        - Variation over tissues
        - Variation over diseases/normal
    - Northern blot
    - RT-PCR
    - Microarray
        - cDNA Microarray
        - Oligonucleotide microarray
- Protein expression
    - Western blot, etc.

	t 1	t 2	t 3
G1	-0.8	-0.3	-0.7
G2	-0.4	-0.8	-0.7
G3	-0.6	-0.8	-0.4
G4	0.9	1.2	1.3
G5	1.3	0.9	-0.6

Important Tasks

- Sequence level tasks
  - Sequencing the genome
  - Fragment assembly
  - Sequence alignment
  - Gene Finding
  - Promoter Identification
  - Phylogenetic tree construction
  - Protein Superfamily Classification

Important Tasks

- Structure level tasks
  - Structure prediction
  - Protein folding
  - Structure based protein classification
  - Molecule design and Docking

Important Tasks

- Expression based tasks
  - Measuring the expression of different bio-molecules
  - Clustering of gene expression data
  - Classification of gene expression data

Important Tasks

- System level tasks : the dynamics of intra and intercellular processes that determine cell function
  - Gene regulatory networks
  - Metabolic pathways
- Related tasks
  - Study of drug response
  - Drug administration schedule optimization
  - Survival prediction
  - Cancer prediction

Challenges

- Huge amount of data
  - Genomic data
  - Expression data of genes
- Lack of data
  - small RNA related data
- Noisy data
  - Difficult to estimate noise and eliminate it
- Missing data
  - Missing value estimation
- Experimental validation

Superfamily Classification of Proteins

- Groups of proteins have similarity in functions and structures and we refer to a group of proteins that share such similarity as a *superfamily*.
- Importance
  - Proper identification of proteins
  - Database maintenance
  - Biological datamining
  - Identification and proper functional assignment of uncharacterized proteins: Drug Discovery and Finding Homologies
- Proteins made up of 20 amino acids

  A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W,Y.
- Example : ….. MKLIPVKTVN…

Problem definition

- Given an unlabeled protein sequence S and a known superfamily F we are to say whether S belongs to F or not.
- Binary Classification of Proteins
  - Input: protein sequence
  - Output: = 1 if the sequence belongs to the target class

    = 0 otherwise

Objective

- Unknown protein extracted from disease D
- By classification of the protein, infer that it belongs to class F
- Drugs existing for F can be considered to be starting point for determination of drugs for D

The need of feature extraction of proteins

- Computational manipulation.
- Evidently a good input representation (extraction of feature) is crucial for proper classification of the proteins.

Existing Feature Extraction Technique (Wang et al, 2001)

- 2-gram encoding: extracts various patterns of two consecutive amino acid residues in a protein sequence and counts the number of occurrences of the extracted residue pairs.
- Example: PVKTNVK is the given protein sequence
  - 1 for PV (indicating PV occurs once),
  - 2 for VK (indicating VK occurs twice),
  - 1 for KT,1 for TN, and 1 for NV.

Contd..

- Feature value x for the 2-gram pattern Y

  x = (# of occurrences of pattern Y in sequence S)
  /(len(S) –1)

- Example:
  - PVKTNVK
  - feature is VK occurring twice
  - the feature value of VK = 2/(7-1) = 0.33.
- Possible 2-gram patterns = 20*20=400

Contd..

- *6*-letter exchange groups
  - e1∈ {H, R, K}, e2∈ *{D, E, N, Q}*, e3 ∈*{C}*, e4 ∈*{S, T, P, A, G}*, e5 ∈*{M, I, L, V}*, e6∈*{F, Y, W}*.
- The 2-gram exchange group encoding for PVKTNVK is
  - 1 for *e4e5 (PV)*
  - 2 for *e5e1* (VK)
  - 1 for *e1e4* (KT), 1 for *e4e2* (TN) and 1 for *e2e5 (NV).*
- Feature definition similar as before.
- *Therefore, 20 X 20 + 6 X 6 = 436* possible features

Contd…

- Selection of Relevant Features

$$D(X) = (m_1 - m_0)^2 / (d_1^2 + d_0^2)$$

where, m_1 and d_1 (m_0 and d_0 respectively) are the mean value and the standard deviation of the feature X in the positive (negative, respectively) training dataset.

Contd…

- Let $X_1, X_2, \ldots, X_{Ng}$, $Ng \ll 436$, be the top *Ng* features with the largest *D(X)* values These are taken as the input features.
- To compensate for the loss of information (of ignoring the other features), a linear correlation coefficient *(LCC)* is used as another input feature value .
- A last input is taken based on the local similarity of protein sequences, which refers to frequently occurring motifs in the target protein sequences.

Classification methodologies

- Classifiers
  - k-NN classifier
  - MLP
- Database used
  - Protein Information Resource(PIR) available at http://pir.georgetown.edu. This contains 172,684 sequences.
  - 3 superfamilies are considered as the target classes:
    - Globin [896]
    - Ras transforming proteins[530]
    - Trypsin homology[521]

Experimental Results

[-in MLP architecture is 62, 30 and 2 nodes in the 3 layers
-in *k*NN, k= 1, no. of inputs = 62 and no. of outputs = 2]

Superfamily	#patterns in training and testing	MLP training	MLP testing	*k*NN testing
Globin	500	98.6	79.0	86.4
	250	98.0	71.0	85.2
Ras	500	99.8	81.0	83.4
	250	97.7	72.2	73.2
Trypsin	500	97.2	79.6	88.4
	250	98.0	69.4	86.2

Summary of Superfamily Classification

- Classification of proteins into superfamilies is an important problem of bioinformatics.
- Traditionally this is done by alignment based methods
- Attempts at extracting features from protein sequences so as employ a classifier for performing the classification exist, e.g., the 2-gram encoding.
- Multilayer perceptron and k-NN classifier as used for classification.
- Extension to multi-class classification.

Gene Expression

- Genome is the same in all the cells
  - Hair, nails, liver, lung, heart
- Then why is the behavior different?
- Not all genes are expressed to the same extent everywhere

↓

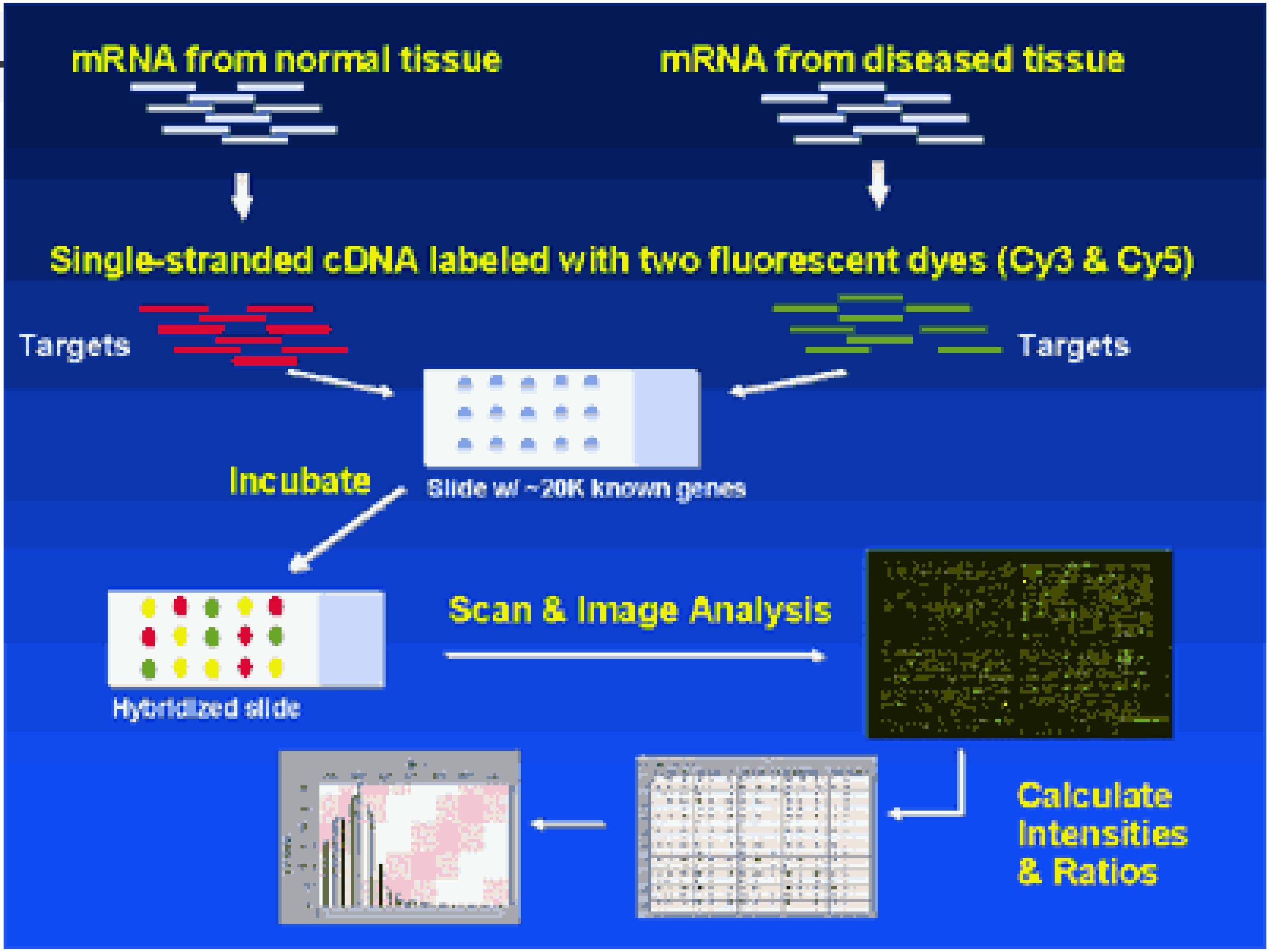
- Differential expression of genes
  - not all mRNAs, and hence their protein products, are generated everywhere
- Expression level of a gene is also dependent on time
  - Amount of mRNA produced varies with time

Microarray

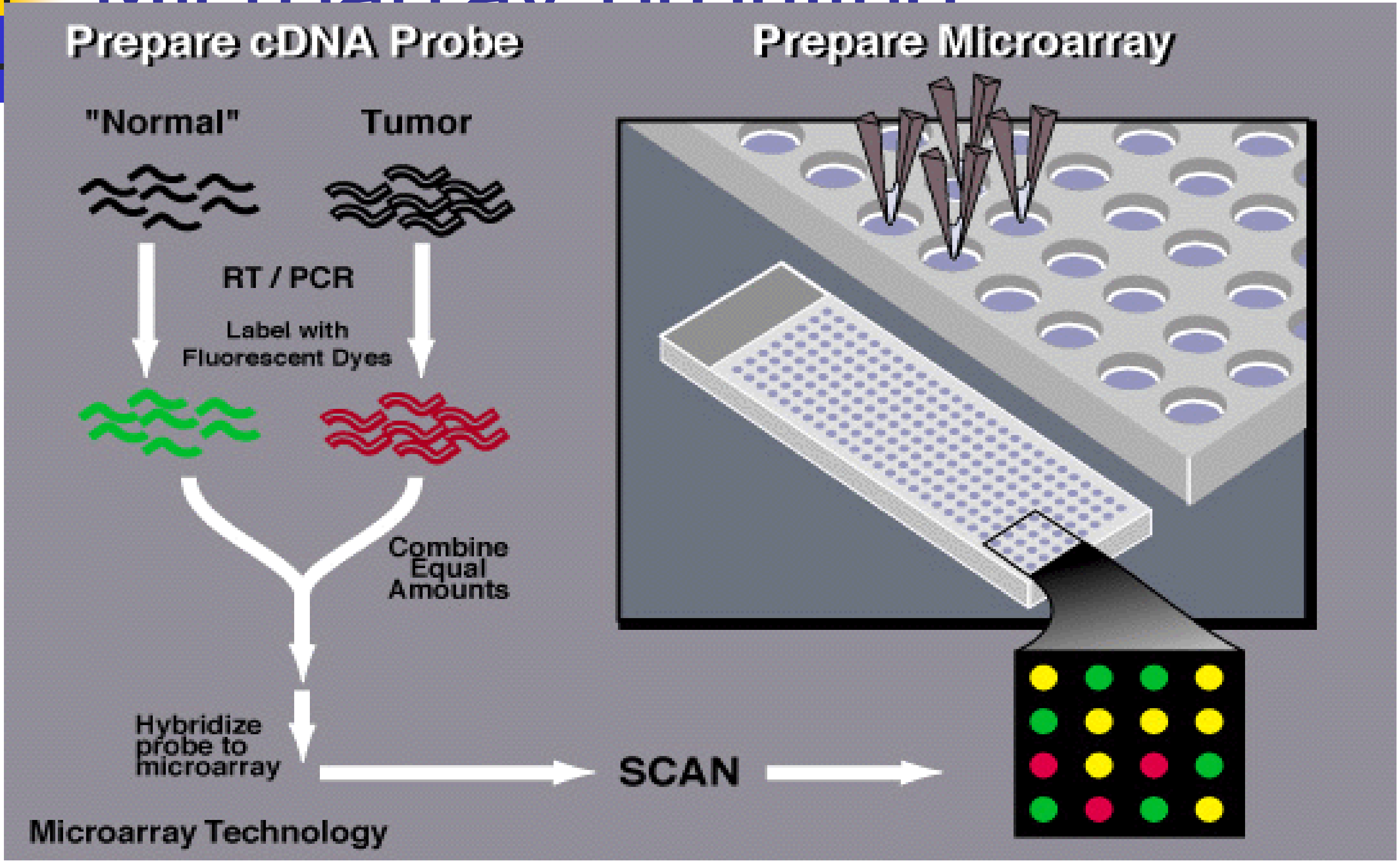
- What is it?
  - Technology to simultaneously monitor the expression levels of a large number of genes
- Typically a glass slide, onto which cDNAs are attached and colored with the green-fluorescent dye Cy3 .
  - Reference/Control sample
- Experimental RNA samples
  - RNA are colored during reverse transcription with the red-fluorescent dye Cy5
- Hybridized with reference sample.
- Separate images acquired for each fluor.

➢ Cy5/Cy3 fluorescence ratio (gene expression) are obtained by measuring the spot intensities with fluorescence scanner

MicroArray


mRNA from normal tissue
mRNA from diseased tissue
Single-stranded cDNA labeled with two fluorescent dyes (Cy3 & Cy5)
Targets
Targets
Incubate
Slide w/ ~20K known genes
Hybridized slide
Scan & Image Analysis
Calculate Intensities & Ratios

Microarray profiling



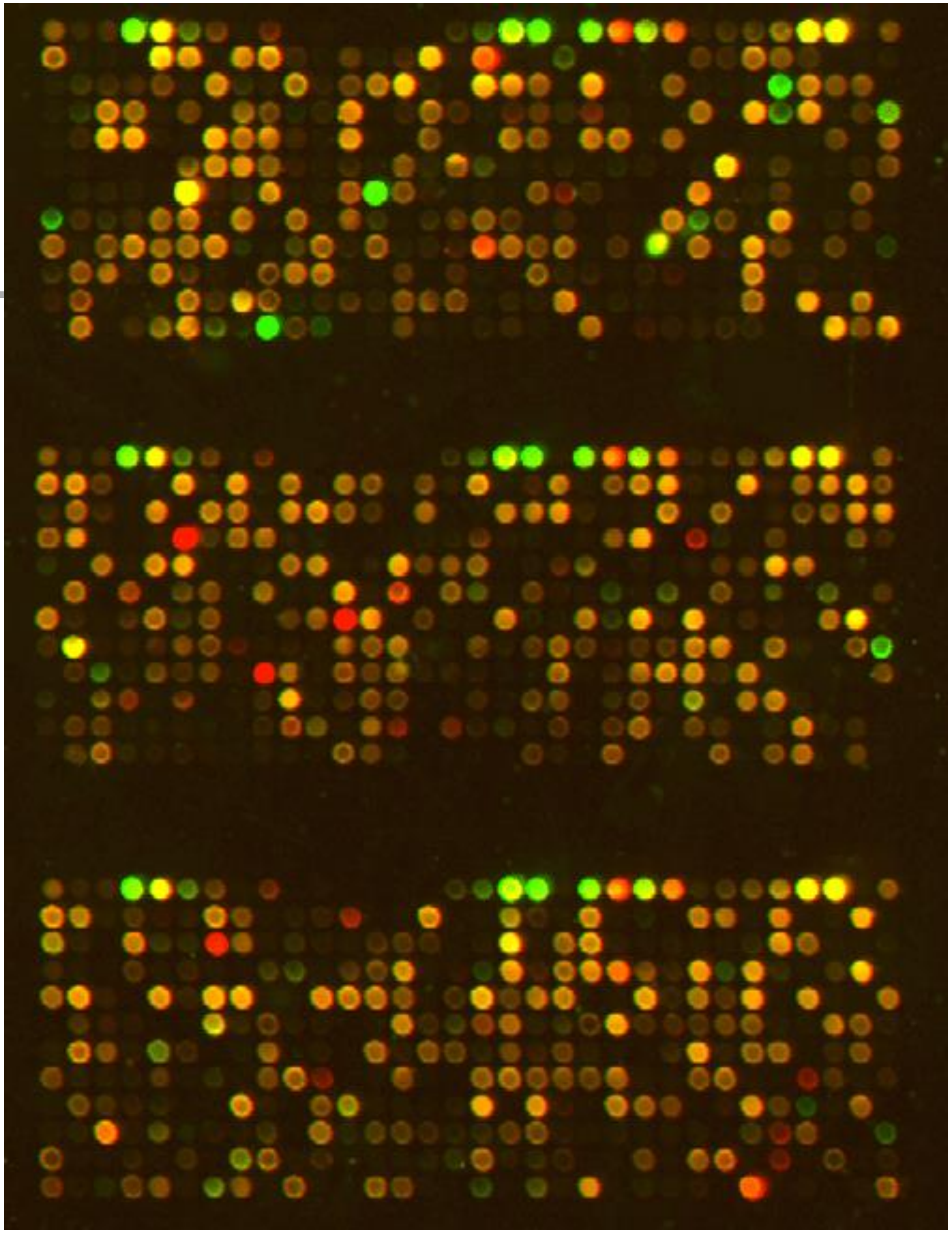
Snapshot of expression data

Gene	I D	t1	t2	t3	t4	t5	t6	t7
G1	…	1.2	1.9	2.4	3.2	1.1	5.7	7.4
G2	…	3.2	3.9	4.4	5.3	3	7.8	9.5
G3	…	1	2.1	3.2	6.2	7.3	8.5	3.7
…	…	…	…	…	…	…	…	…
G100 0	…	2.2	3.1	6.3	5.3	8.2	2.5	4.3

G e	D									n5	n6
G1	…	1.2	1.6	1.8	1.1	1	2	1.3	4	2	1.1
G2	…	1.1	1.5	1.3	1.8	2.1	1.1	1.1	1.1	2.3	1.5
G3	…	1.2	1.7	1.8	1.1	2	1.1	2.1	0.8	1.1	1.9
…	…	…	…	…	…	…	…	…	…	…	…
G100	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Typical Microarray

- Microarray data set:
  - *G X C* matrix *M ,*
    - *G* genes on the rows,
    - *C* conditions/samples on the column



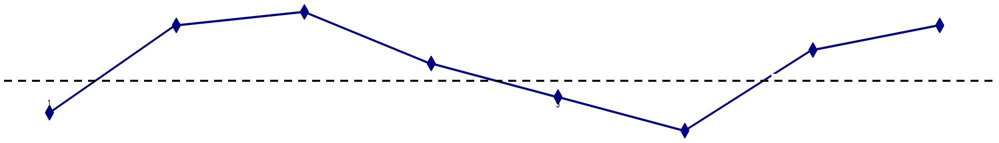
Expression Vectors

Gene Expression Vectors encapsulate the expression of a gene over a set of experimental conditions or sample types.

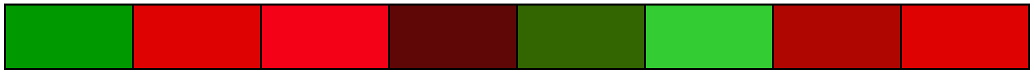
Numeric Vector

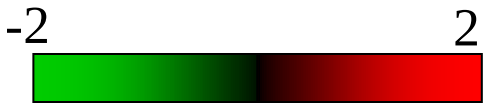
-0.8	1.5	1.8	0.5	-0.4	-1.3	0.8	1.5

Line Graph



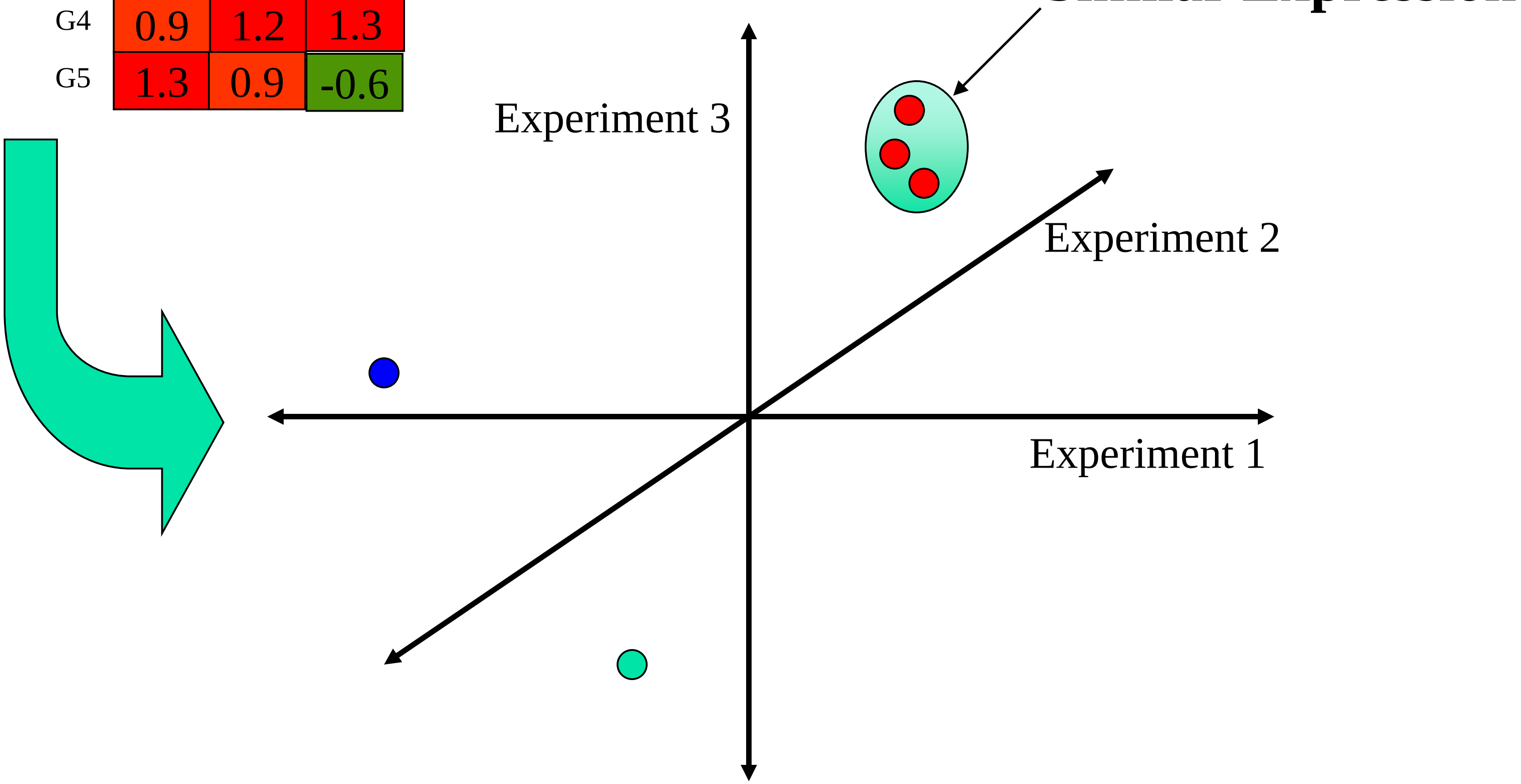
Heatmap





Expression Vectors As Points in ‘Expression Space’

	t 1	t 2	t 3
G1	-0.8	-0.3	-0.7
G2	-0.4	-0.8	-0.7
G3	-0.6	-0.8	-0.4
G4	0.9	1.2	1.3
G5	1.3	0.9	-0.6



Distance and Similarity

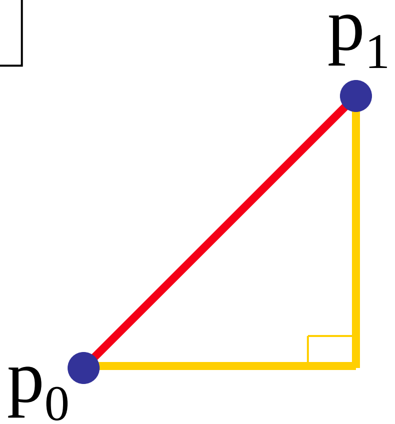
-the ability to calculate a distance (or similarity, it’s inverse) between two expression vectors is fundamental to clustering algorithms

-distance between vectors is the basis upon which decisions are made when grouping similar patterns of expression

-selection of a *distance metric* defines the concept of distance

Distance: a measure of similarity between gene expression.

	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Gene A	x_{1A}	x_{2A}	x_{3A}	x_{4A}	x_{5A}	x_{6A}
Gene B	x_{1B}	x_{2B}	x_{3B}	x_{4B}	x_{5B}	x_{6B}



Some distances: (MeV provides 11 metrics)

1. Euclidean: $\sqrt{\Sigma_{i=1}^{6} (x_{iA} - x_{iB})^2}$
2. Manhattan: $\Sigma_{i=1}^{6} |x_{iA} – x_{iB}|$
3. Pearson correlation

Potential Microarray Applications

- Drug discovery / toxicology studies
- Mutation/polymorphism detection
- Differing expression of genes over:
  - Time
  - Tissues
  - Disease States
- Sub-typing complex genetic diseases

Microarray Data Analysis

- Data analysis consists of several post-quantization steps:
  - Statistics/Metrics Calculations
  - Scaling/Normalization of the Data
  - Differential Expression
  - Coordinated Gene Expression (aka clustering)
- Most software packages perform only a limited number of analysis tasks
- Databases can facilitate the movement of data between packages

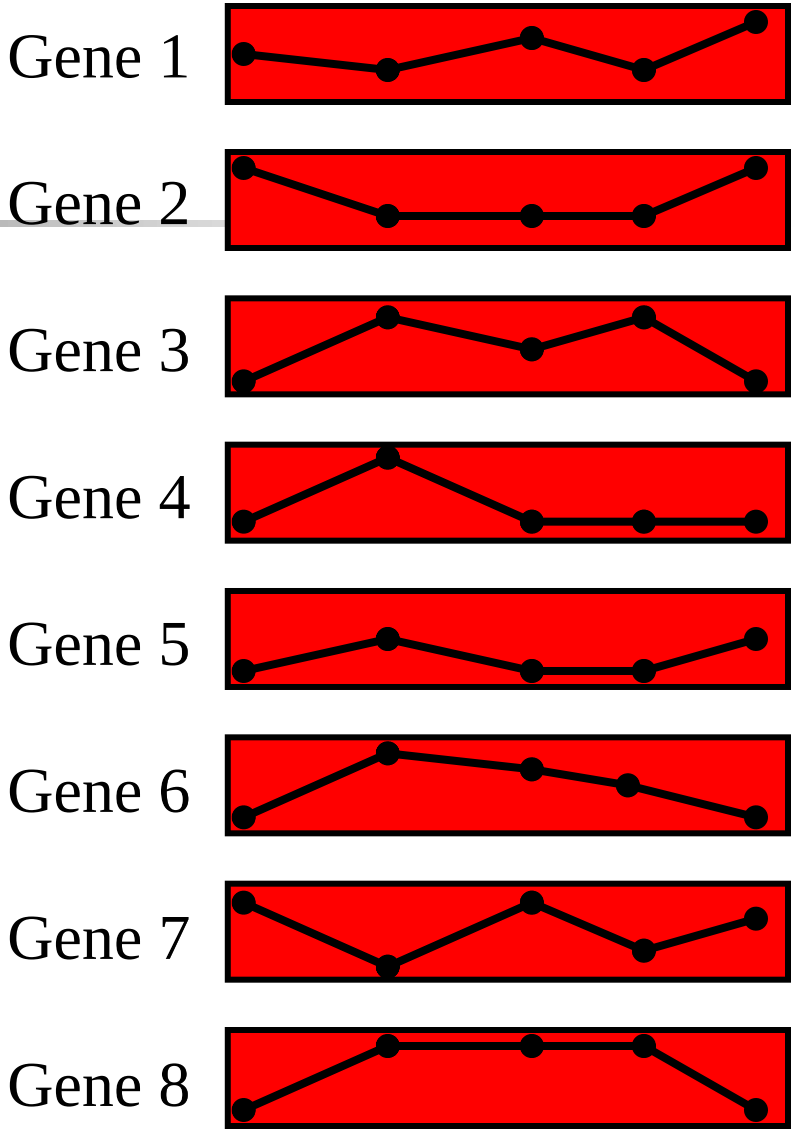
Popular Methods of Clustering of Gene Expression Data

- Hierarchical methods
  - Single link, average link, complete link
    - dendogram
- Self-Organizing Maps
- k-means Clustering

Hierarchical Clustering

- IDEA: Iteratively combines genes into groups based on similar patterns of observed expression
- By combining genes with genes OR genes with groups algorithm produces a dendrogram of the hierarchy of relationships.
- Display the data as a heatmap and dendrogram
- Cluster genes, samples or both

Hierarchical Clustering



Gene 1

Gene 2

Gene 3

Gene 4

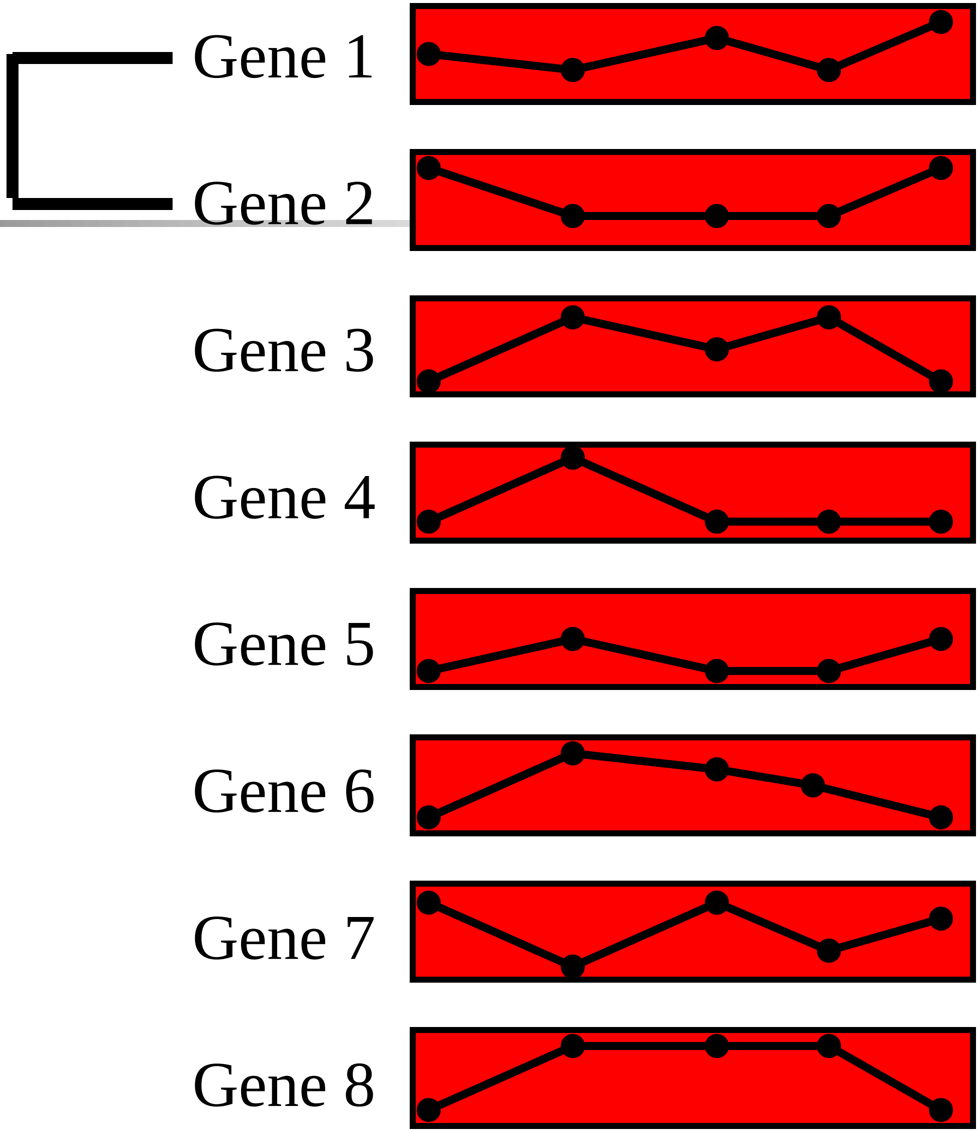
Gene 5

Gene 6

Gene 7

Gene 8

Hierarchical Clustering



Gene 1

Gene 2

Gene 3

Gene 4

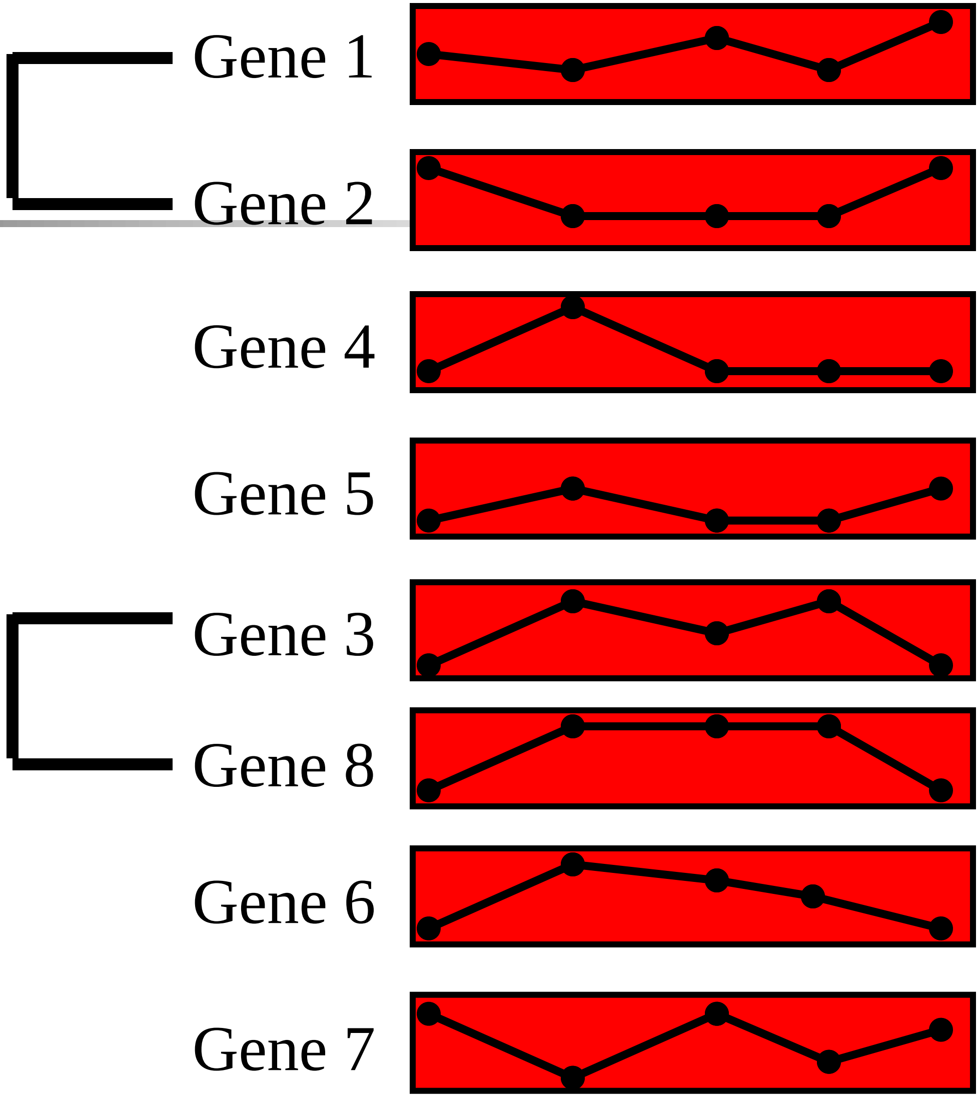
Gene 5

Gene 6

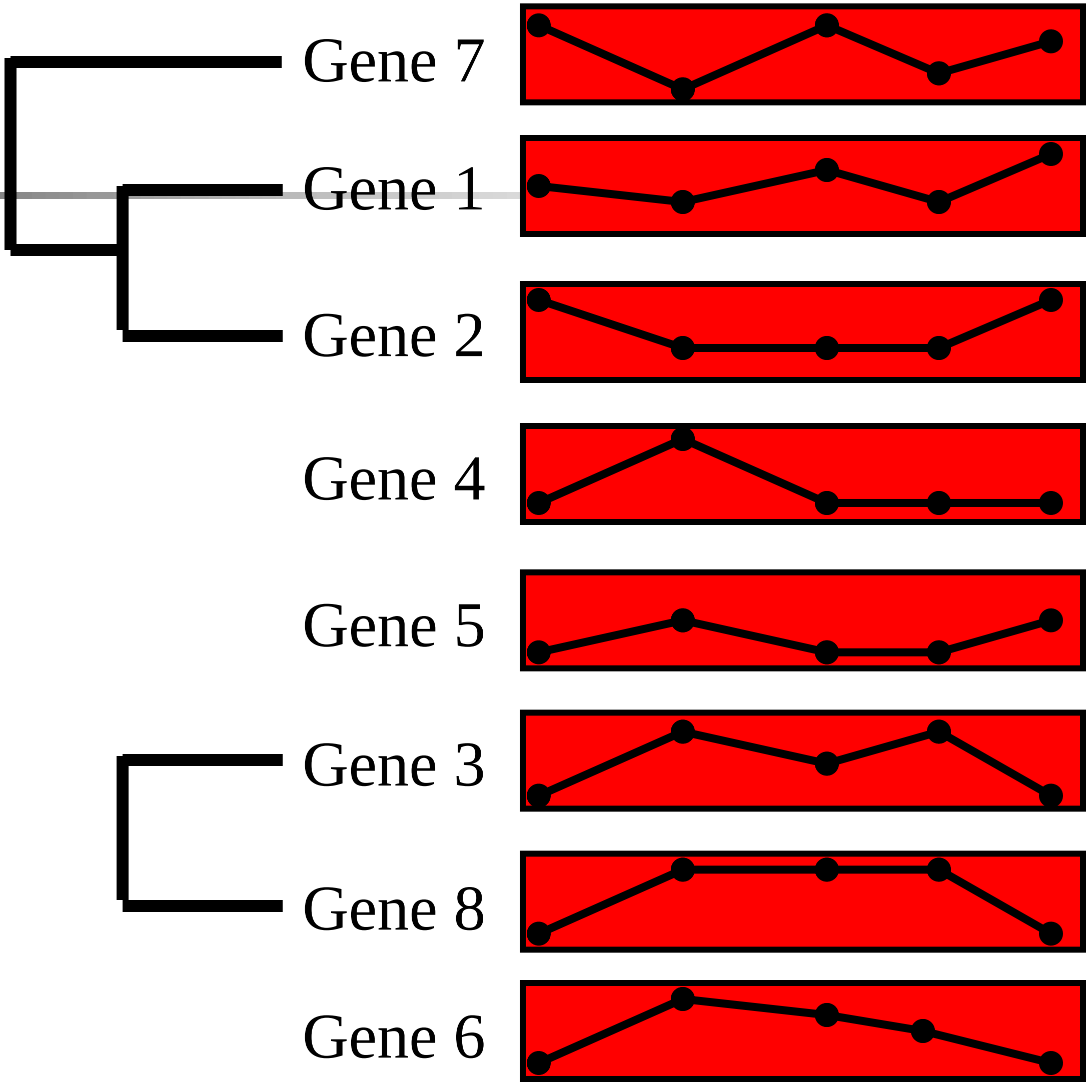
Gene 7

Gene 8

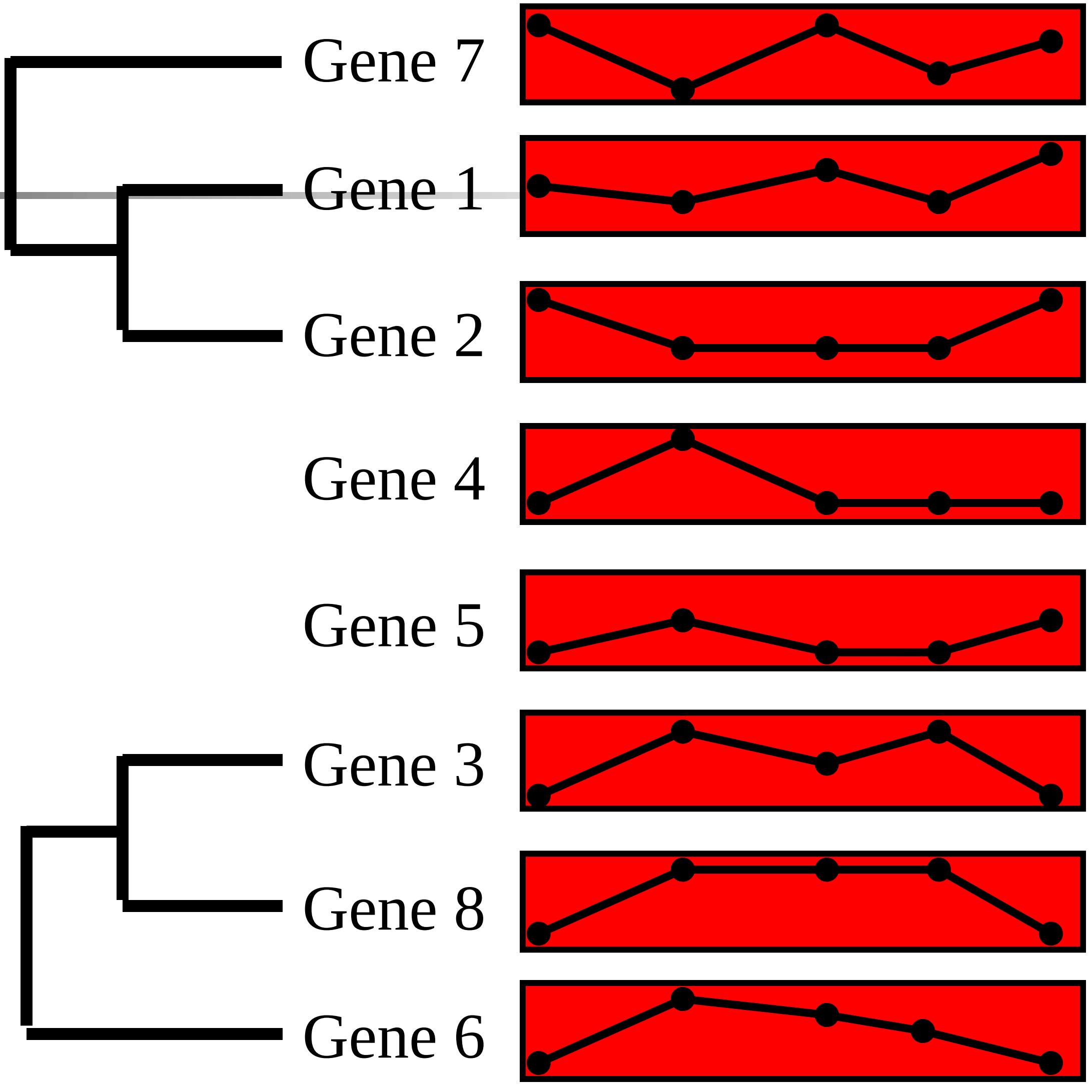
Hierarchical Clustering



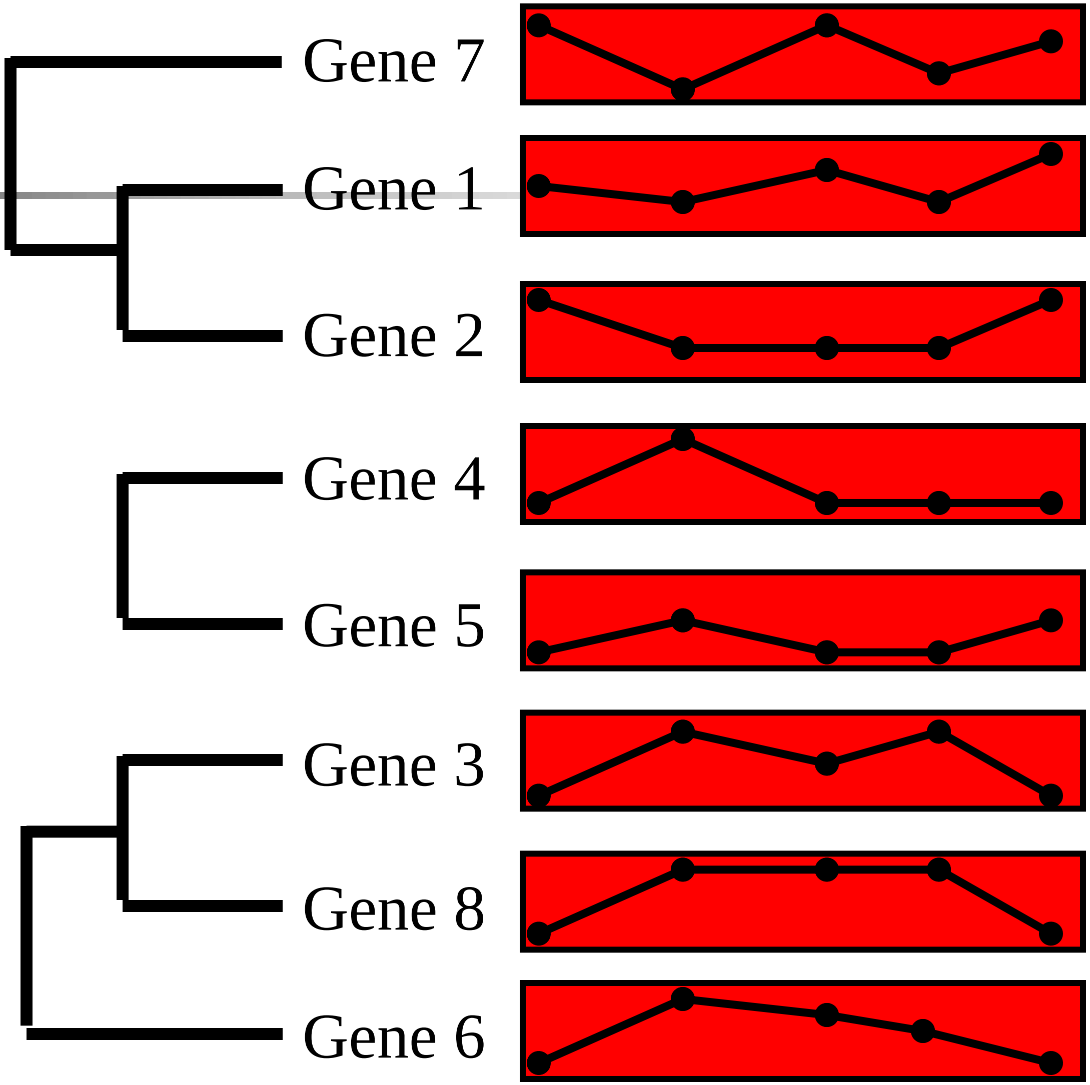
Hierarchical Clustering



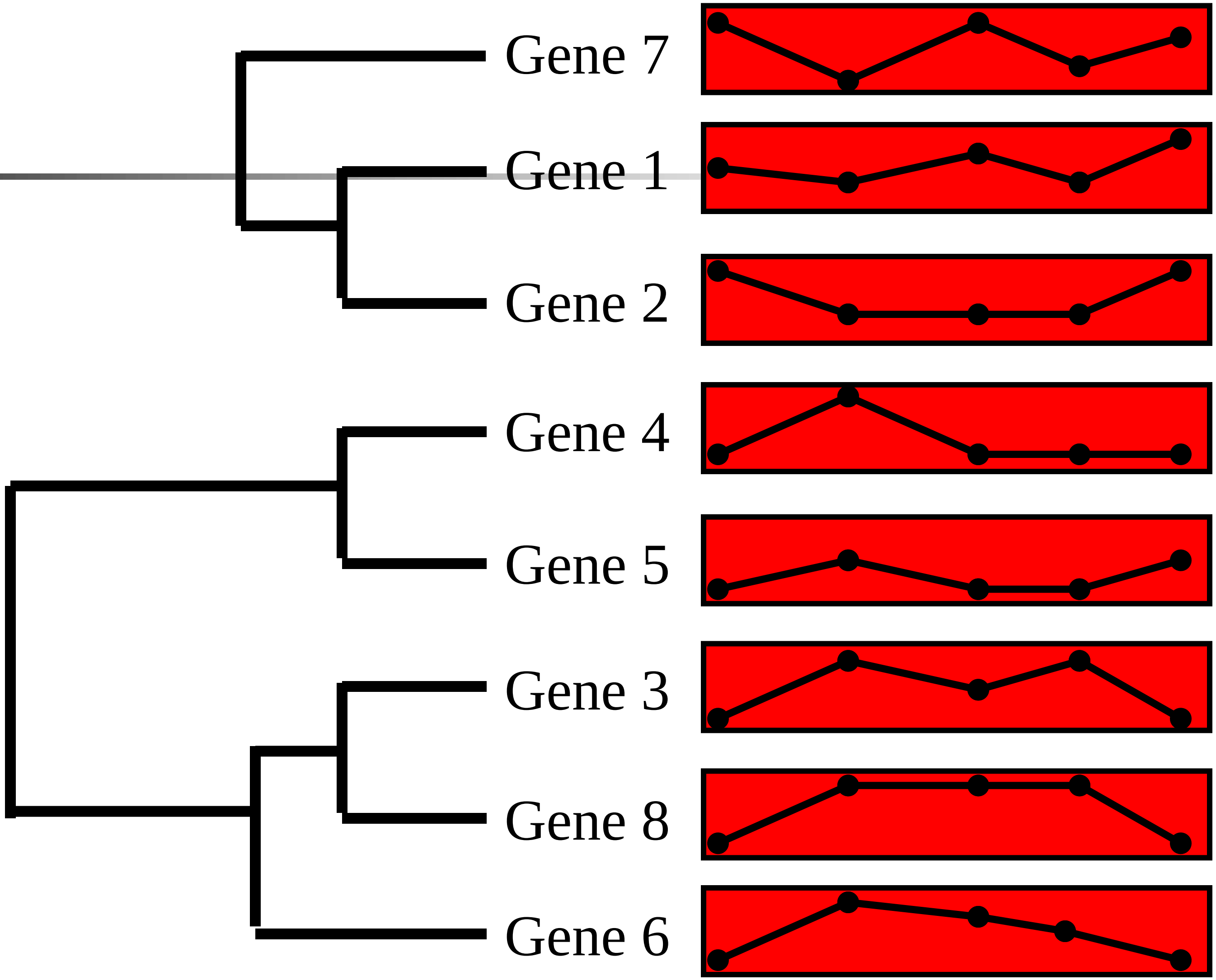
Hierarchical Clustering



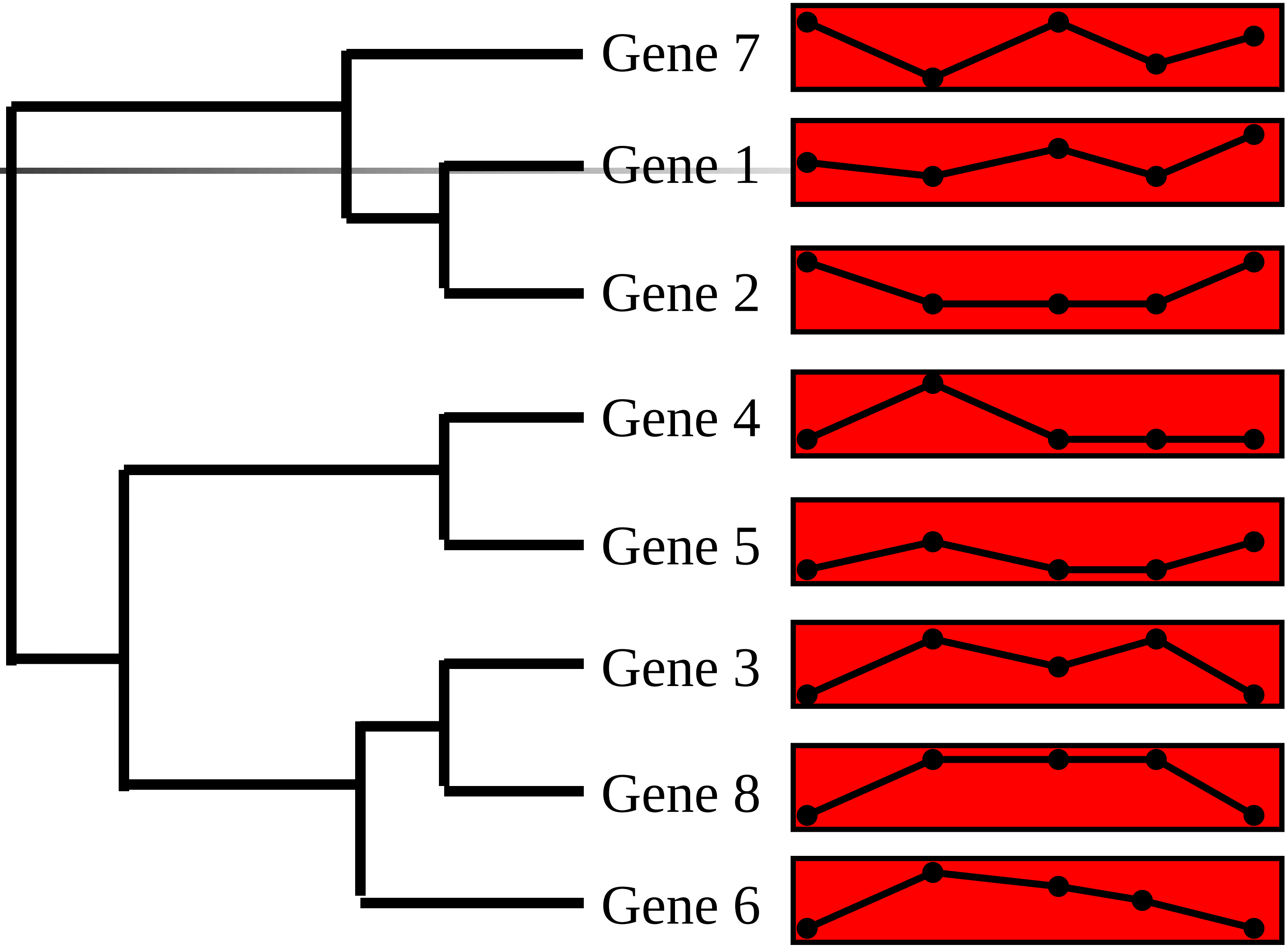
Hierarchical Clustering



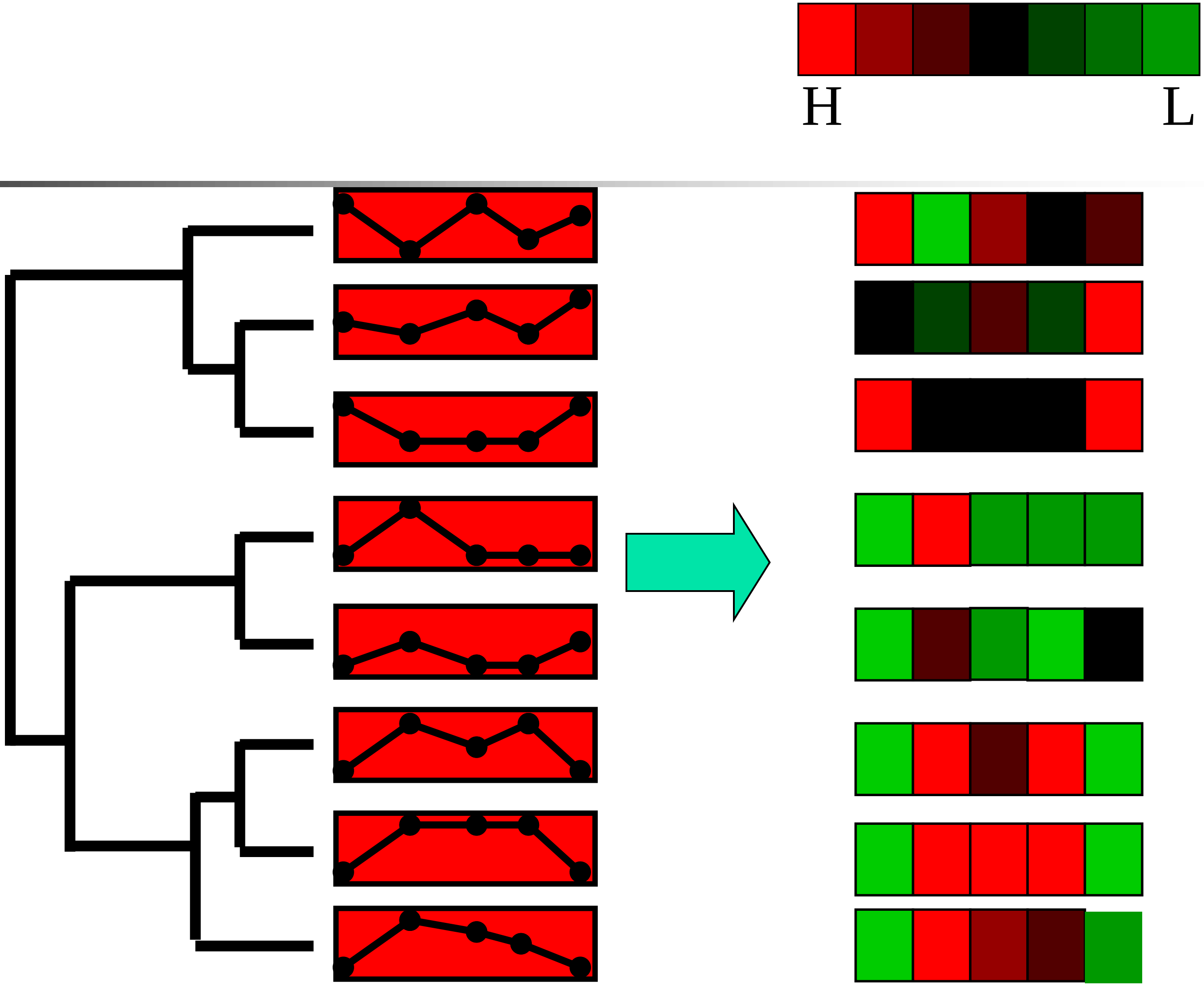
Hierarchical Clustering



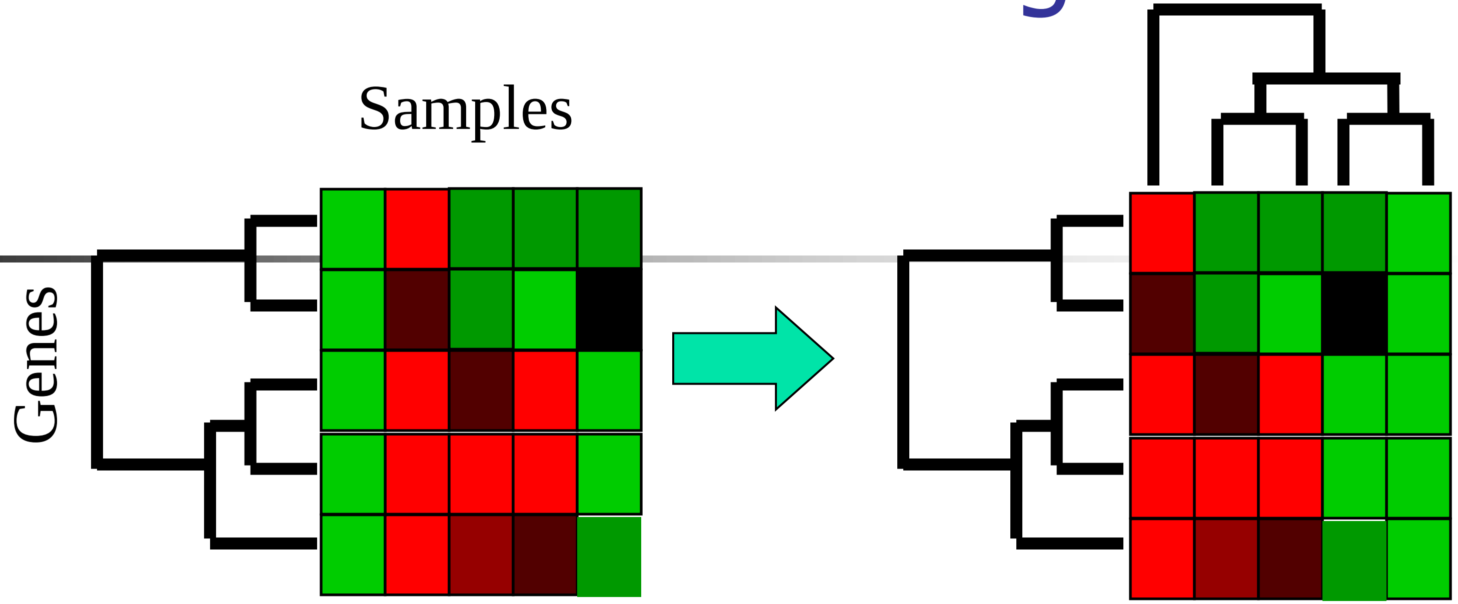
Hierarchical Clustering



Hierarchical Clustering



Hierarchical Clustering

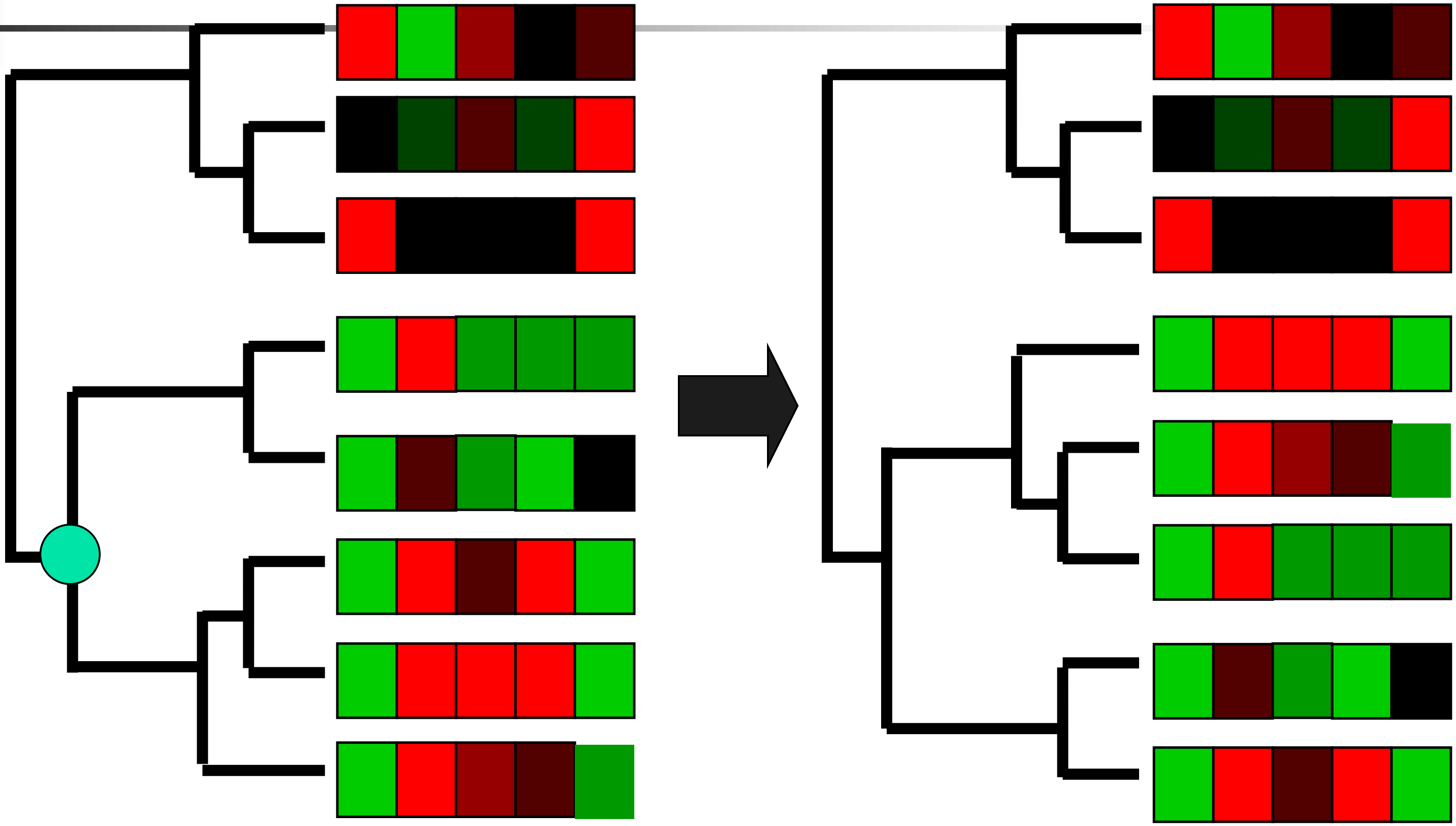


The Leaf Ordering Problem:

- Find ‘optimal’ layout of branches for a given dendrogram architecture
- 2^{N-1} possible orderings of the branches
- For a small microarray dataset of 500 genes there are 1.6*E150 branch configurations

Hierarchical Clustering

The Leaf Ordering Problem:



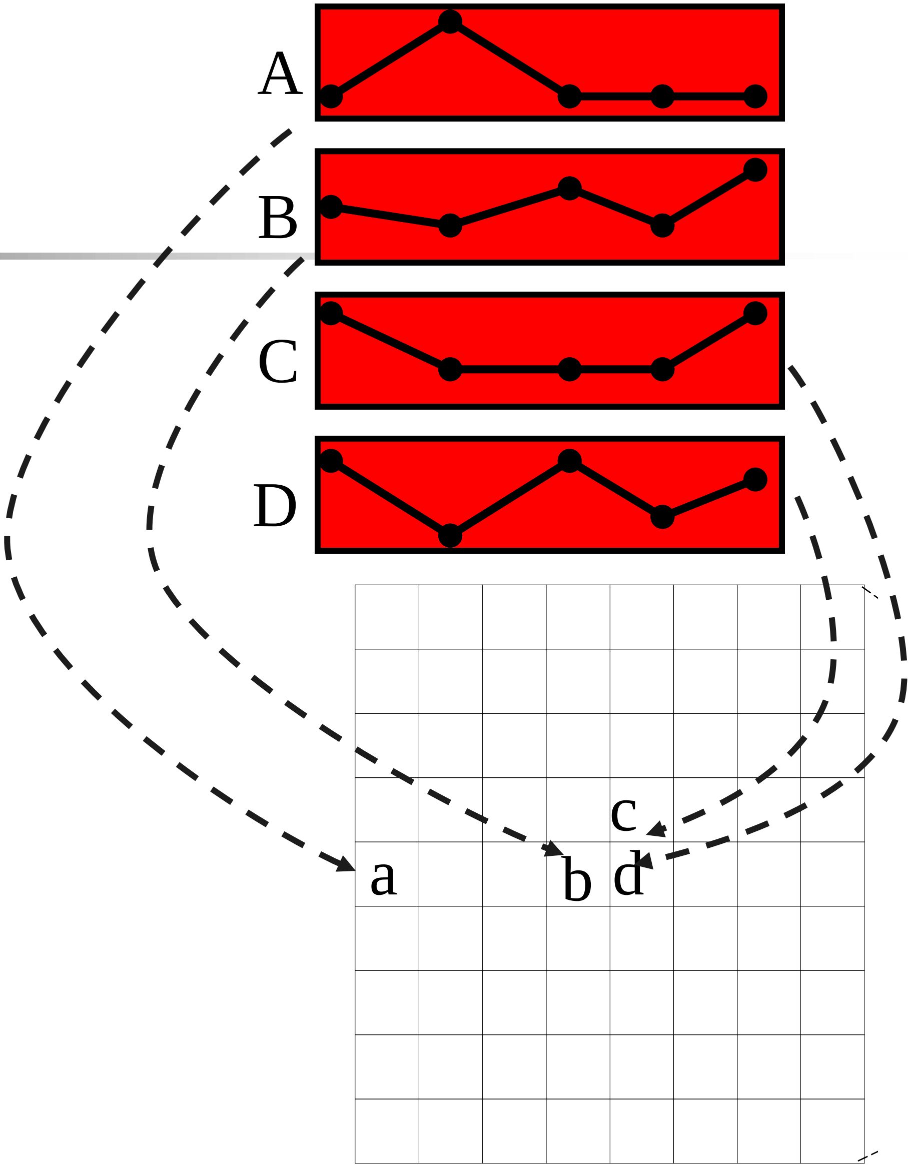
Hierarchical Clustering

- Pros:
  - Commonly used algorithm
  - Simple and quick to calculate
- Cons:
  - Real genes probably do not have a hierarchical organization

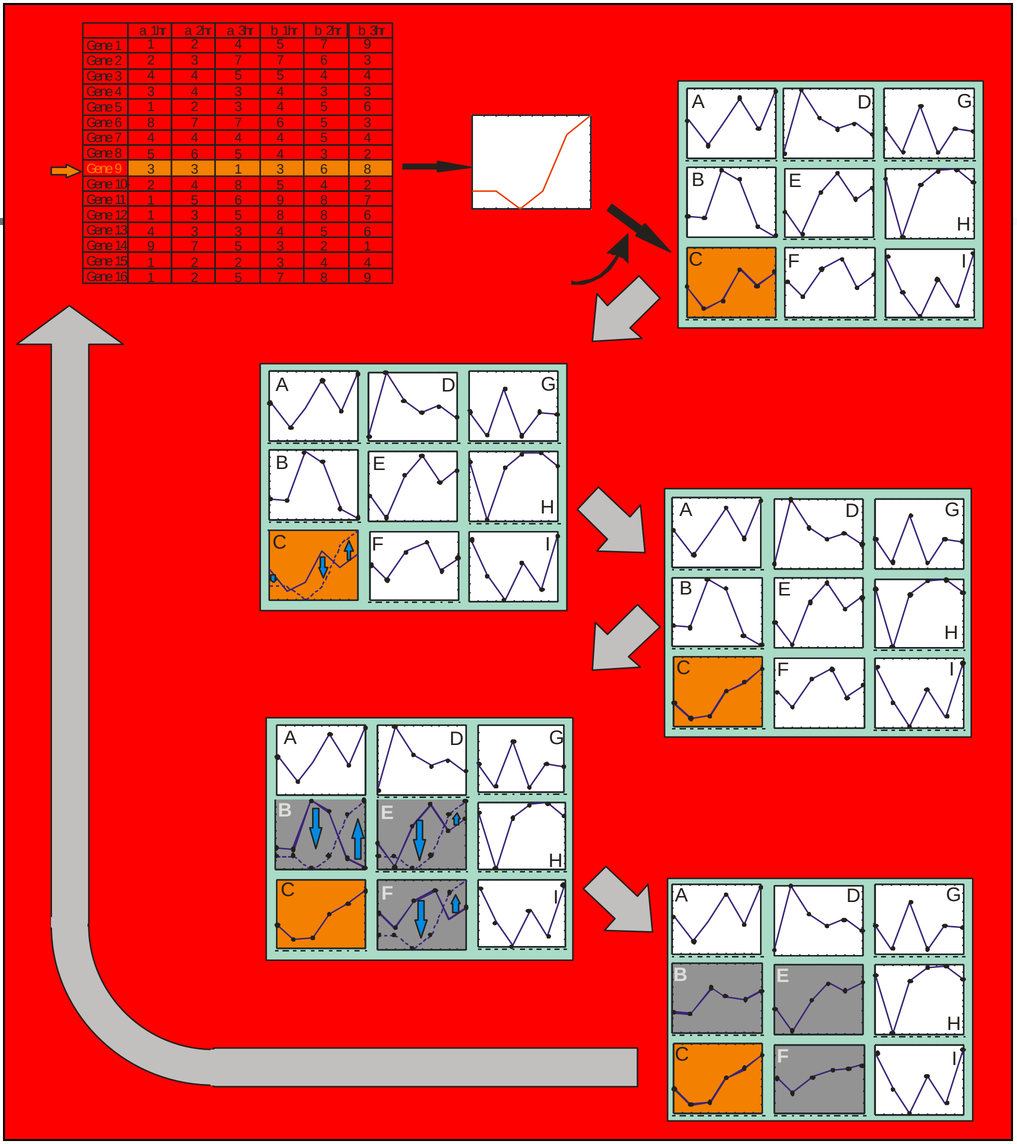
Self-Organizing Maps (SOMs)

Idea:

Place genes onto a grid so that genes with similar patterns of expression are placed on nearby squares.

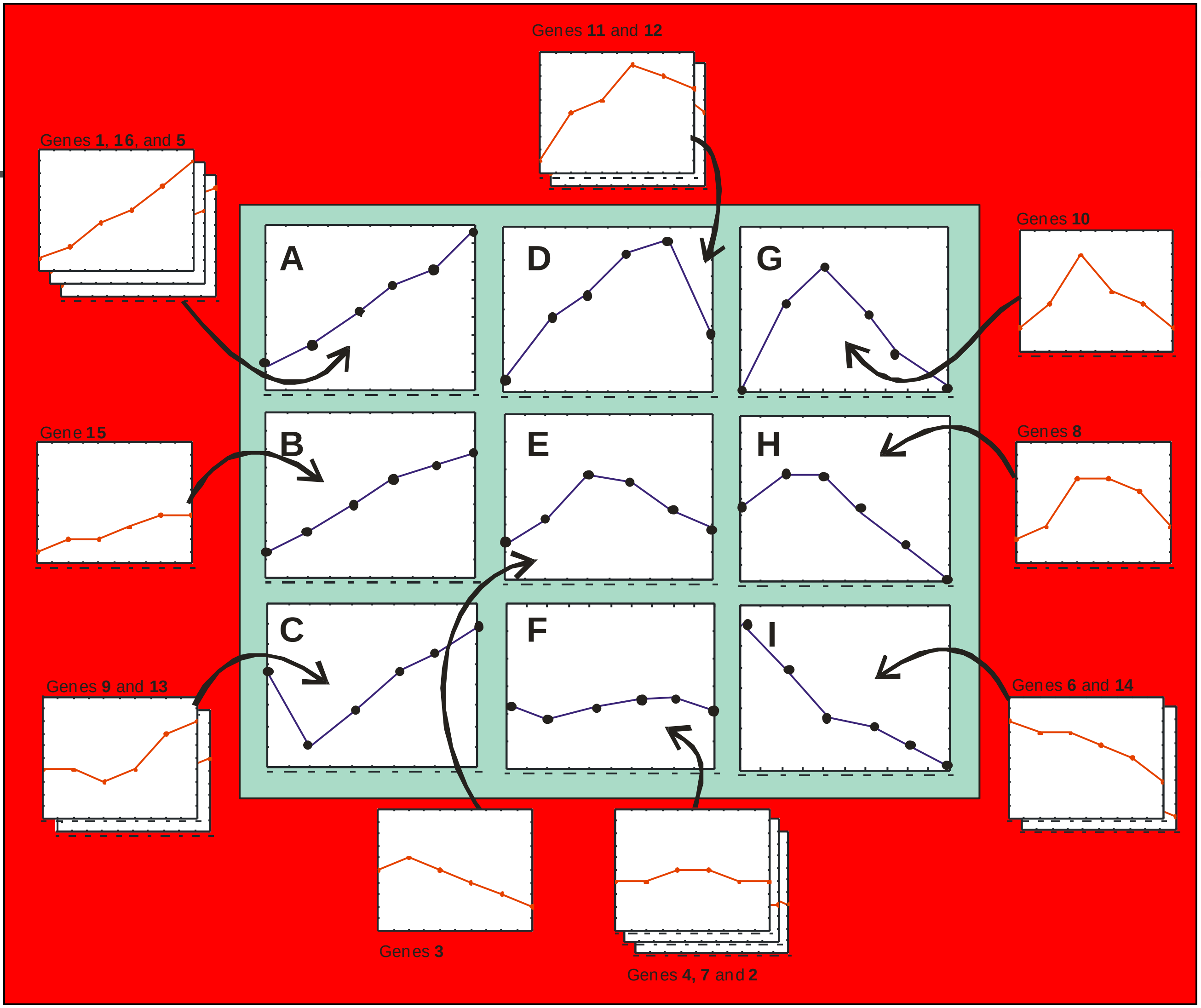


Self-organizing Maps (SOMs)

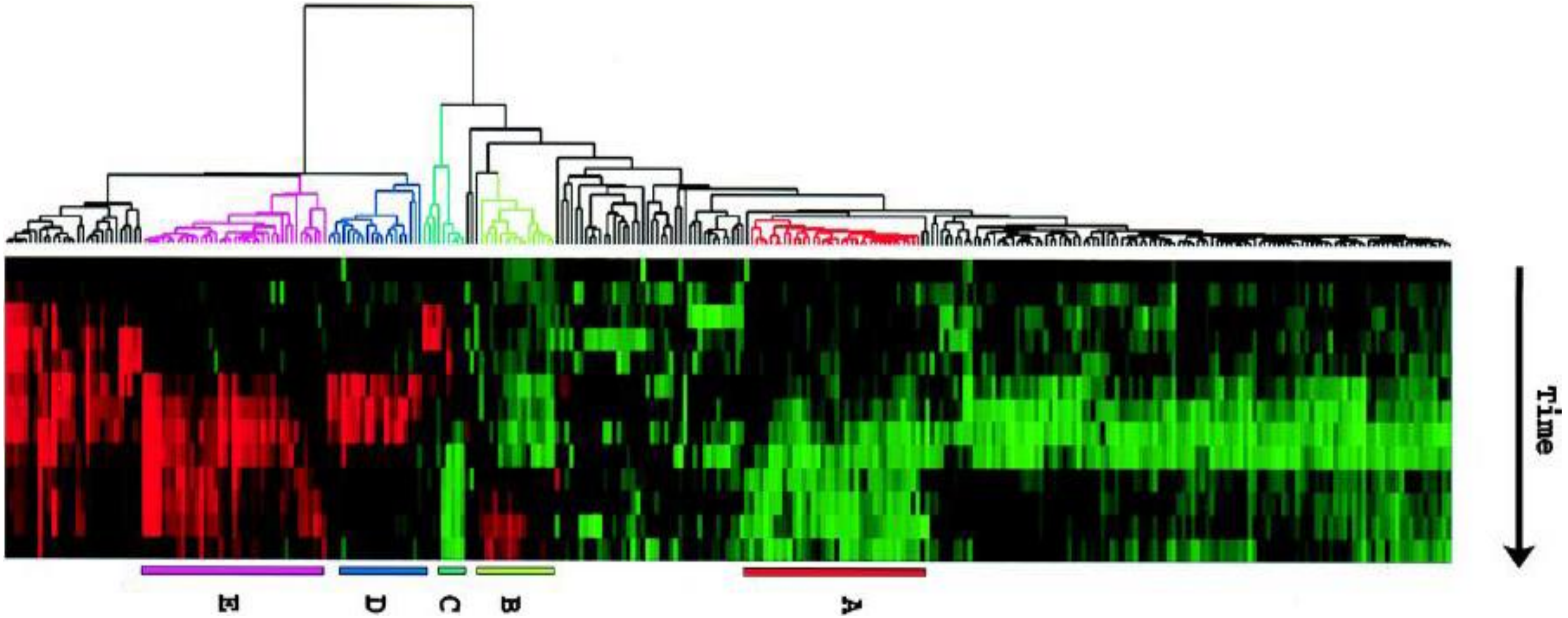


	a 1hr	a 2hr	a 3hr	b 1hr	b 2hr	b 3hr
Gene 1	1	2	4	5	7	9
Gene 2	2	3	7	7	6	3
Gene 3	4	4	5	5	4	4
Gene 4	3	4	3	4	3	3
Gene 5	1	2	3	4	5	6
Gene 6	8	7	7	6	5	3
Gene 7	4	4	4	4	5	4
Gene 8	5	6	5	4	3	2
Gene 9	3	3	1	3	6	8
Gene 10	2	4	8	5	4	2
Gene 11	1	5	6	9	8	7
Gene 12	1	3	5	8	8	6
Gene 13	4	3	3	4	5	6
Gene 14	9	7	5	3	2	1
Gene 15	1	2	2	3	4	4
Gene 16	1	2	5	7	8	9

Self-organizing Maps (SOMS)



Hierarchical Clustering



Experimental results

Data Sets	No. of genes	No. of time points	No. of clusters
Yeast Sporulation	6118	7	7
Human Fibroblasts Serum	517	13	10

Experimental results (Cont.)

- The Sporulation data is filtered to ignore the genes whose expression level didn’t change significantly across different time points. After filtering, 474 prominently expressed genes are found.

- Both the data set is normalized so that each row has mean 0 and variance 1.

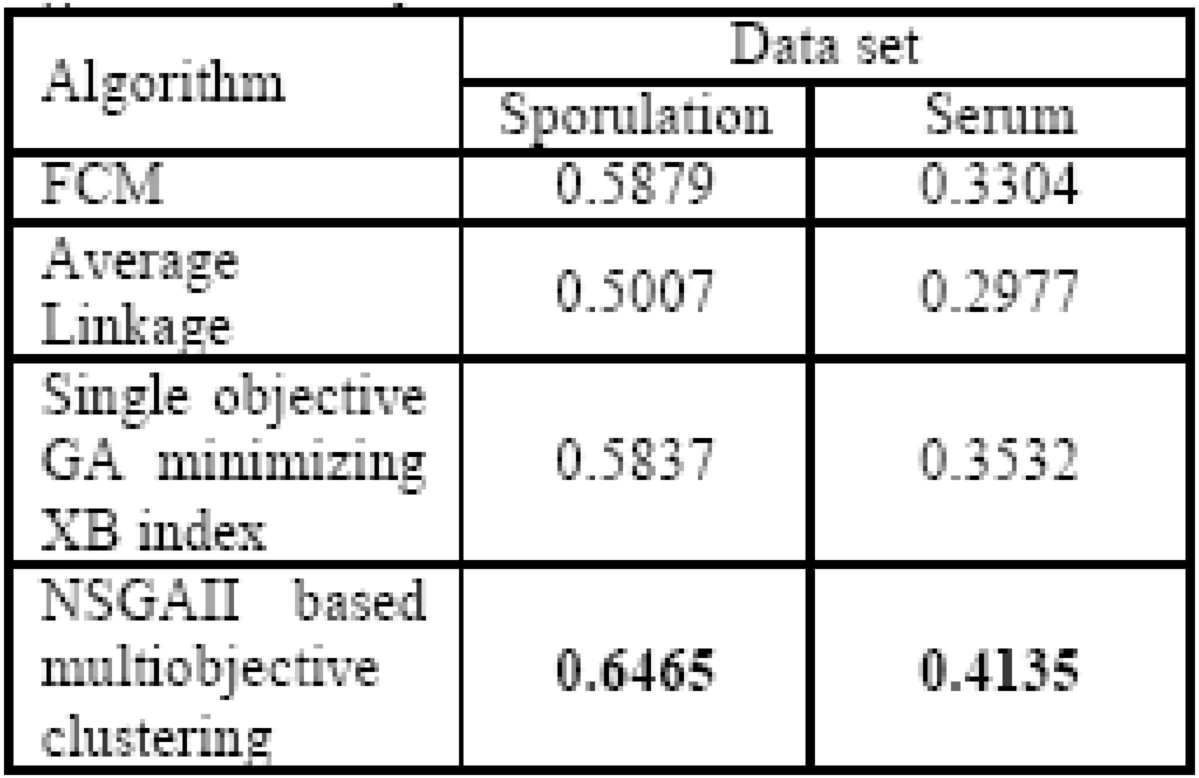
Experimental results (Cont.)

- Performance metric: *Silhouette index*
  - Silhouette width of a point is defined as:

$$s = \frac{b - a}{\max\{a, b\}}.$$

  - **a**: the average distance of the point from the other points of the cluster to which the point is assigned.
  - **b**: the minimum of the average distances of the point from the points of the other clusters.
- Silhouette index is the average silhouette width of all the data points (genes). It ranges between -1 and 1, and larger value indicates better solution.

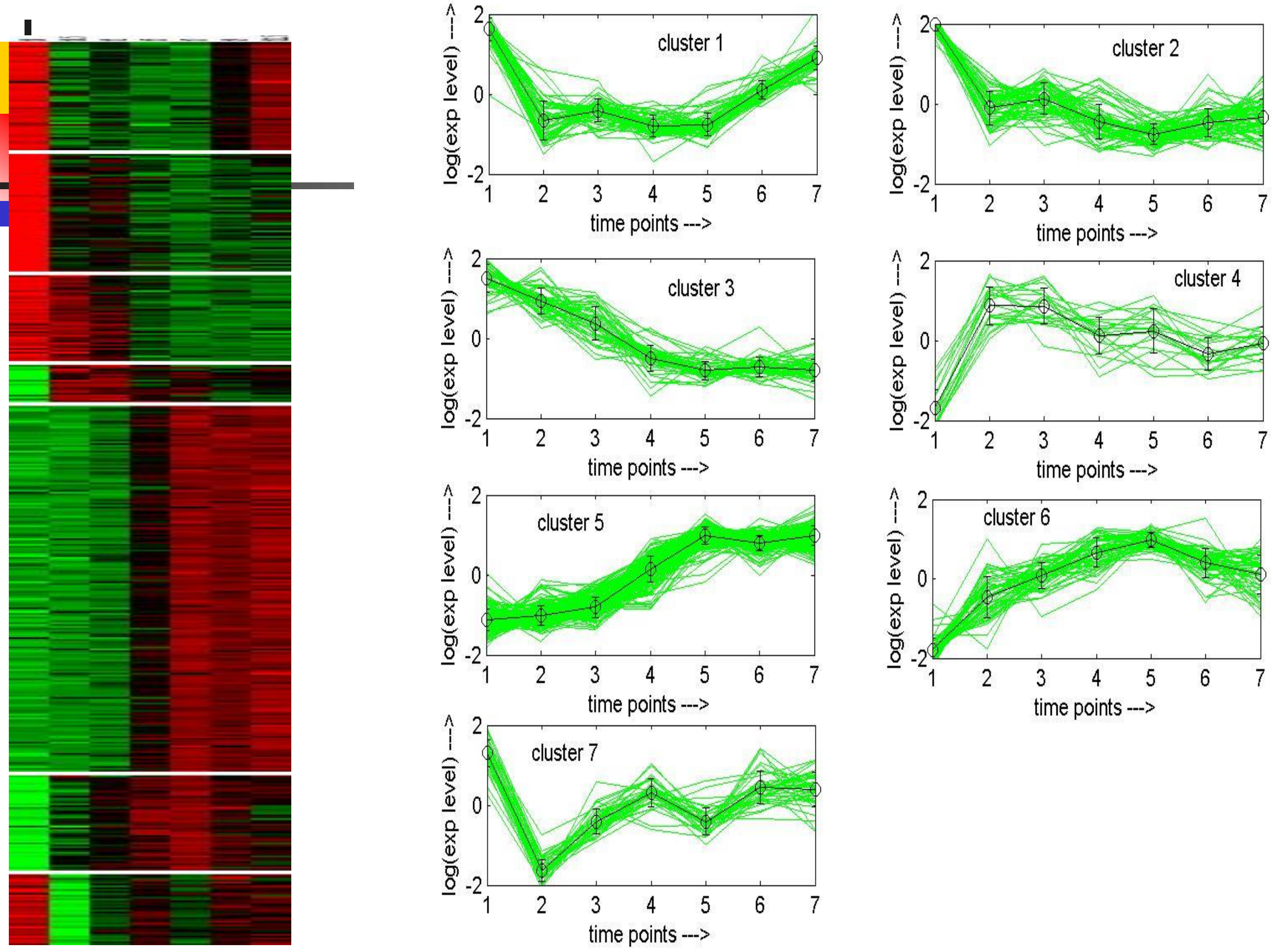
Experimental results (Cont.)



Algorithm	Data set	
	Sporulation	Serum
FCM	0.5879	0.3304
Average Linkage	0.5007	0.2977
Single objective GA minimizing XB index	0.5837	0.3532
NSGAII based multiobjective clustering	**0.6465**	**0.4135**

Silhouette index values for different algorithms on Sporulation and Serum data sets

Visualizing clustering results



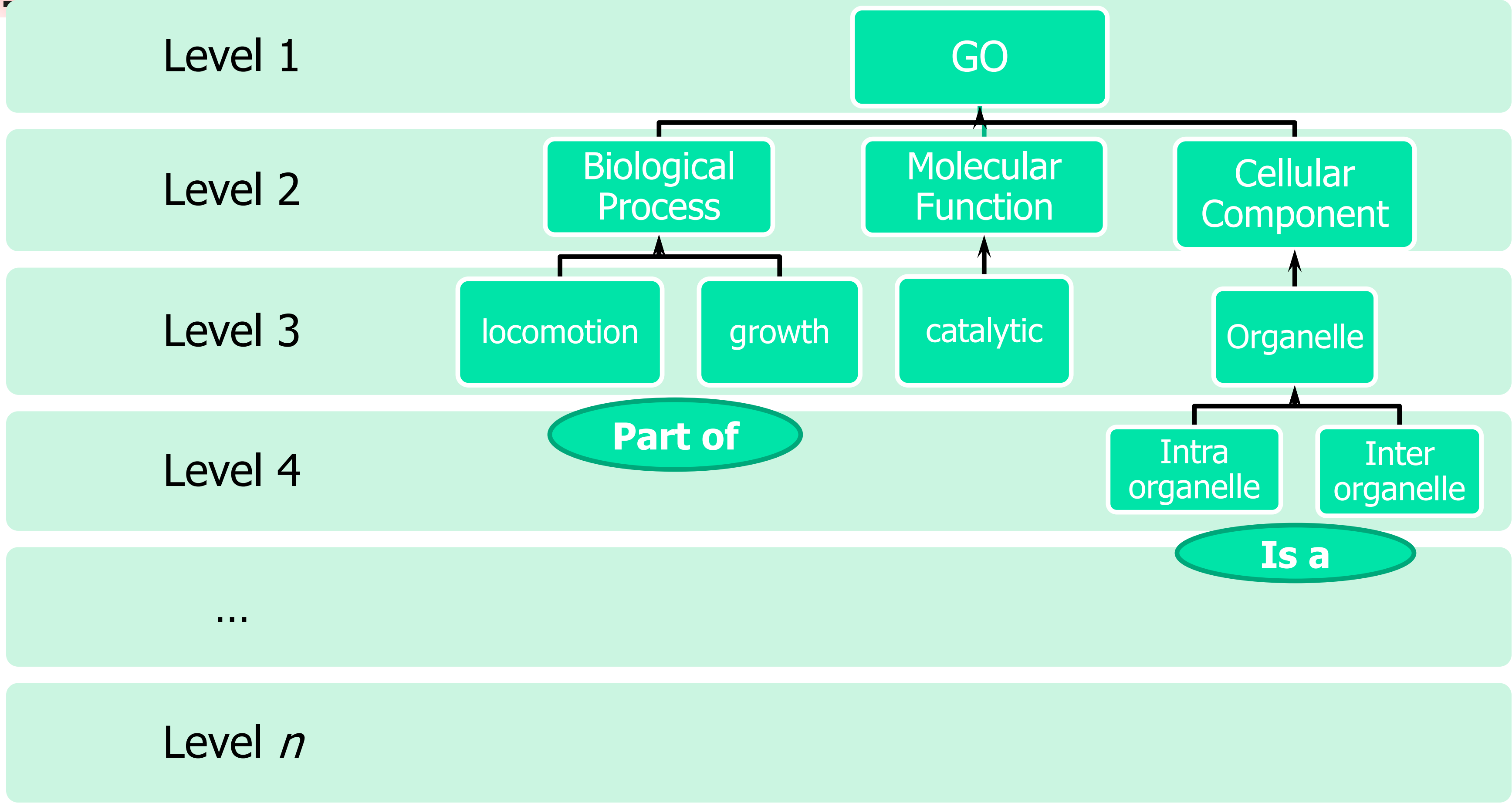
(a)

(b)

Bandyopadhyay, et al, “An Improved Algorithm for Clustering Gene Expression Data", *Bioionf.*, vol. 23, no. 21, pp. 2859-2865, 2007.

Sporulation data clustered using multiobjective clustering (7 clusters): (a) Eisen plot, (b) Cluster profile plots.

Gene ontology



Level 1

GO

Level 2

Biological Process

Molecular Function

Cellular Component

Level 3

locomotion

growth

catalytic

Organelle

Level 4

Part of

Intra organelle

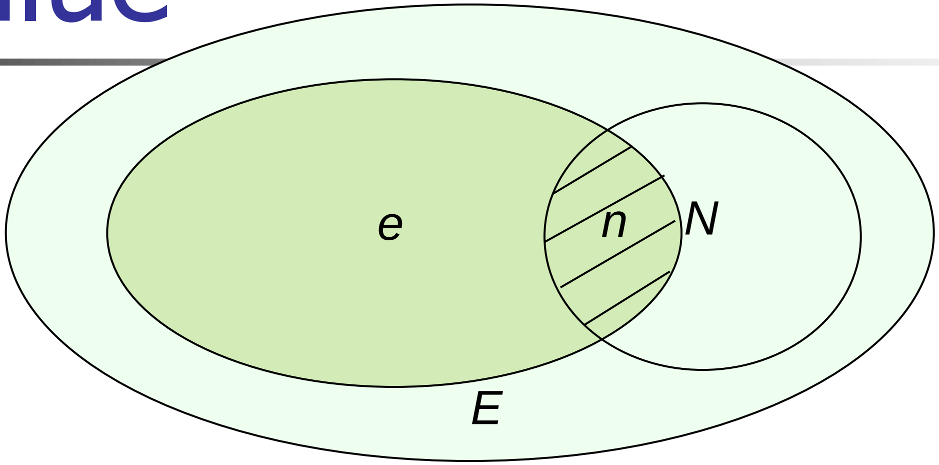
Inter organelle

Is a

…

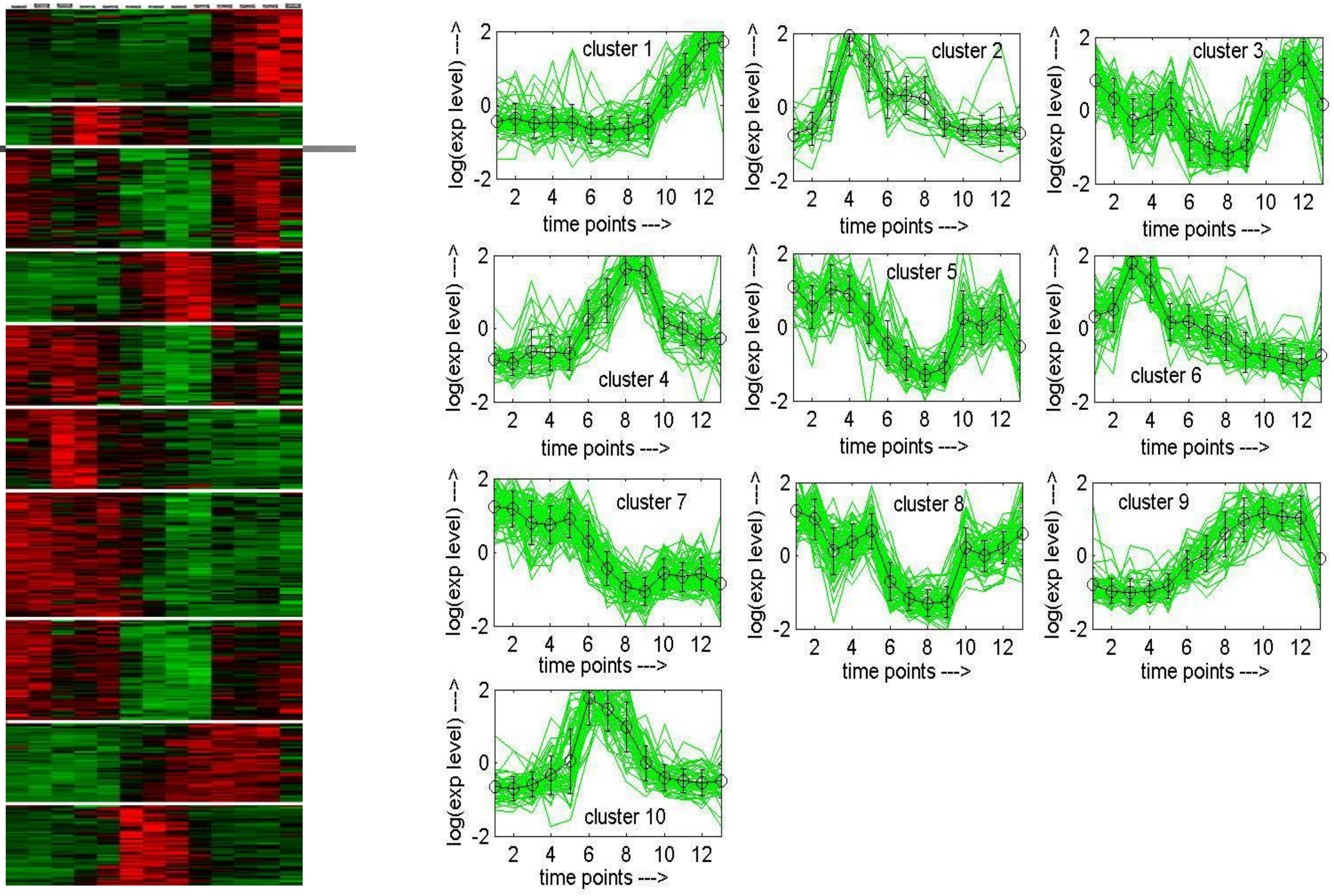
Level *n*

p-value



$$p\text{-value} = \sum_{i=n}^{\min(e,N)} \frac{\binom{e}{i}\binom{E-e}{N-i}}{\binom{E}{N}}$$

Visualizing clustering results (Cont.)



(a) (b)

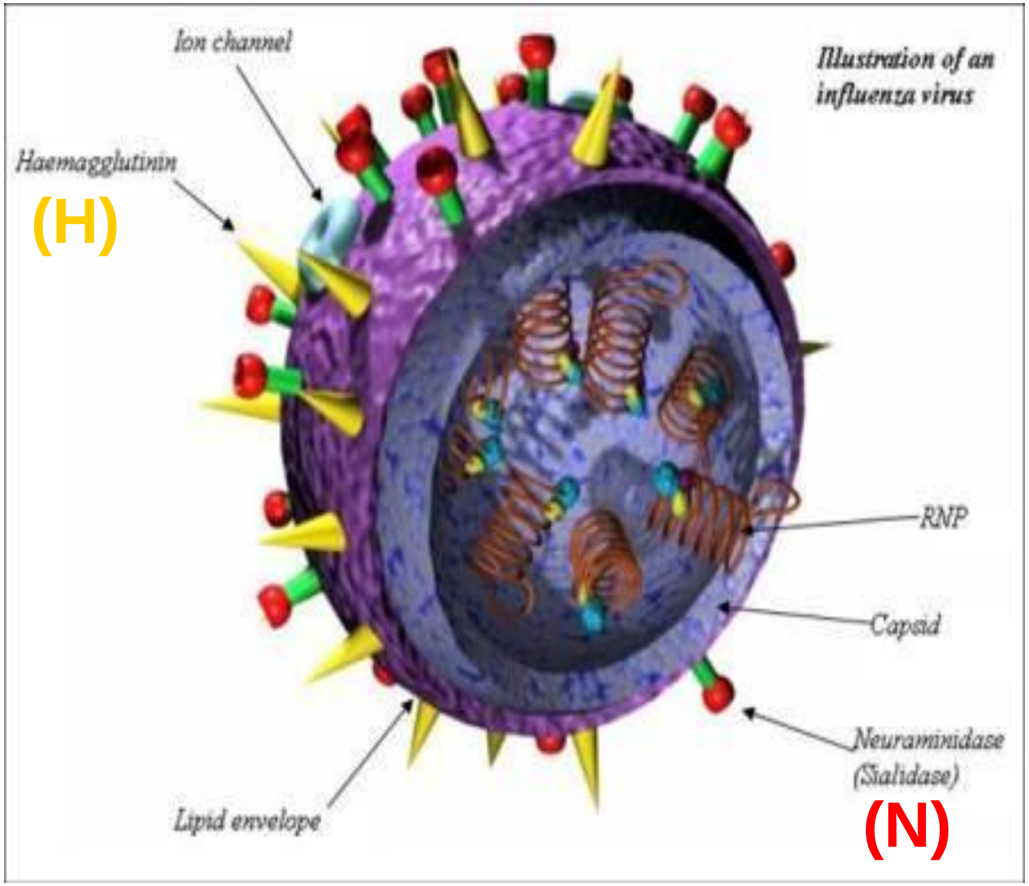
Serum data clustered using multiobjective clustering (10 clusters): (a) Eisen plot, (b) Cluster profile plots.

Rational Drug Design

- Design drugs using the information about the 3D Shape of Proteins
  - To inhibit protein function

- Step 1: Looking for protein targets in the virus
- Step 2: Identify the active site
- Step 3: Design drug for blocking the active site
- Step 4: Analyse the properties of the designed molecules (ADMET properties)
- Step 5: Do further studies with the designed molecule

Designing a Flu Drug
Step 1: looking for protein targets

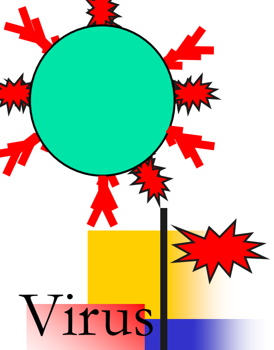


Influenza viruses are named according to the proteins sticking out of their virus coat.

There are two types of protein = N and H.

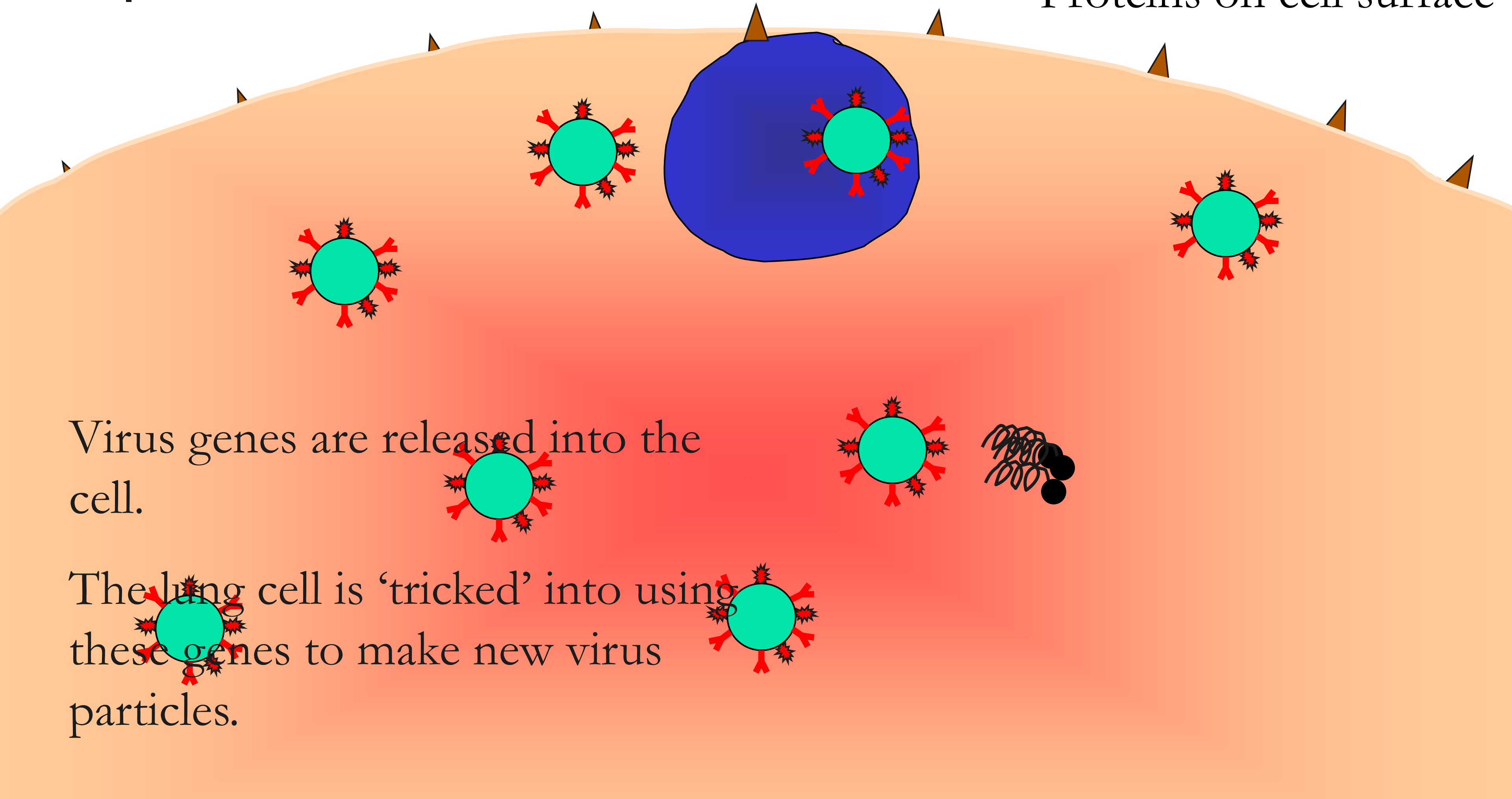
N and H have special shapes to perform specific jobs for the virus.

Source http://www.gtac.edu.au/site/bioinformatics/bio_task_10/bio_task_10.ppt



N cuts the links between the viruses and the cell surface so virus particles are free to go and infect more cells.

H attaches to cell surface proteins so virus can enter



Virus genes are released into the cell.

The lung cell is ‘tricked’ into using these genes to make new virus particles.

Design of Flu Drug

Australian team of scientists headed by Prof Peter Coleman. They designed the flu drug, Relenza

Active site and Drug Design – Relevance of GAs

- Identify/design a suitable ligand which can bind to the active site of a protein to prevent its proliferation.
- Design the ligand using groups from a library of chemical groups
  - Such that interaction energy is minimized
- Drug design problem can be modeled as one of optimization
- Application of GAs becomes relevant.

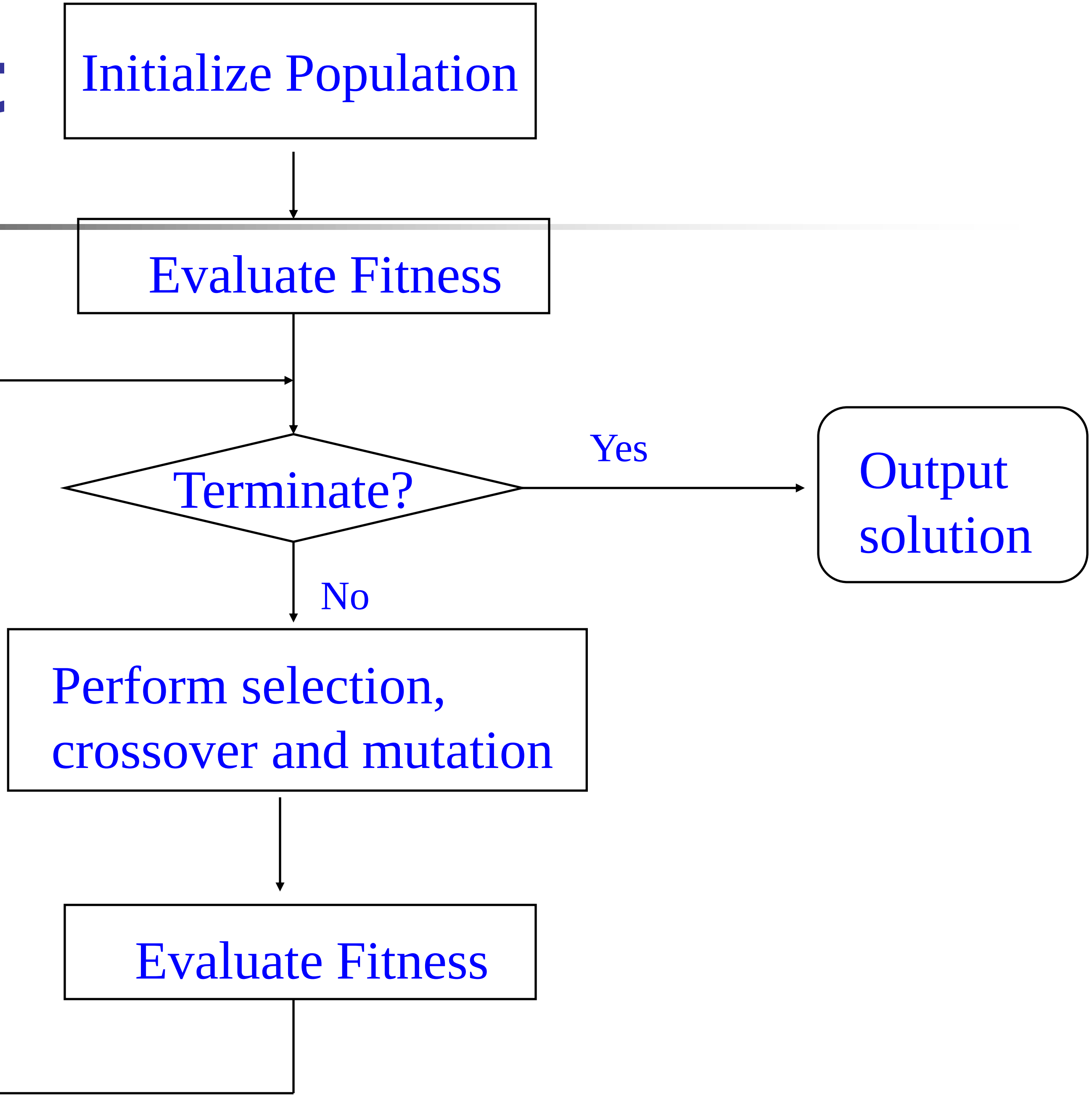
Genetic Algorithmic Classification

Searching for optimum (appropriate) arrangement using GA based searching.

⇩

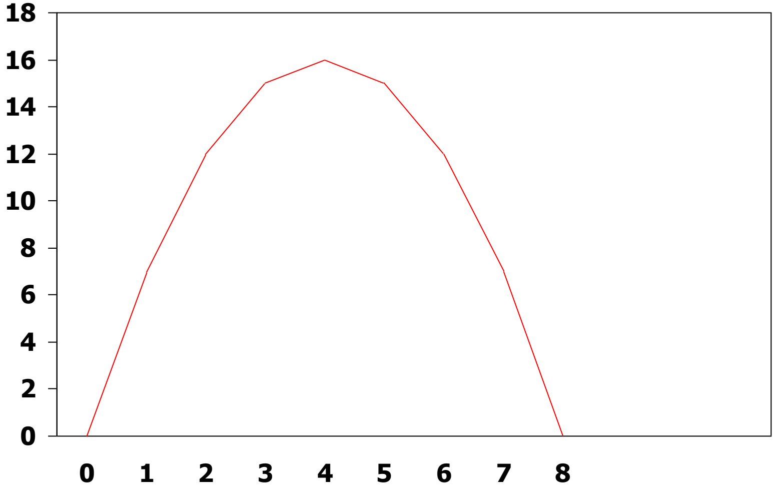
Individual solution	⇒	encoded as chromosomes.
Goodness of a solution	⇒	fitness of the chromosome.
Population	⇒	set of chromosomes.
Primary operations	⇒	selection, crossover, mutation.

GA Flowchart



Encoding and Population - Example

Optimize $f(x) = x(8 - x)$, $x=[0,8]$



1	0	0	1	1	0	1	0

= 154

$$x = 8/255 * 154 + 0 = 4.8313$$

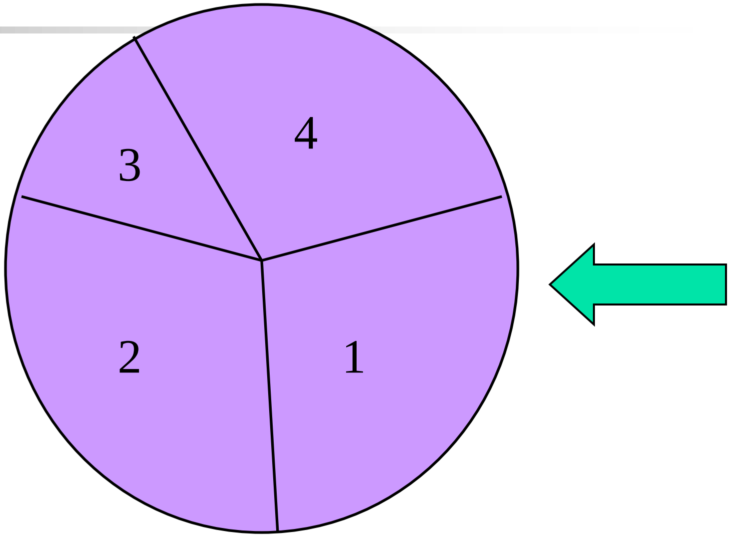
Fitness Evaluation - Example

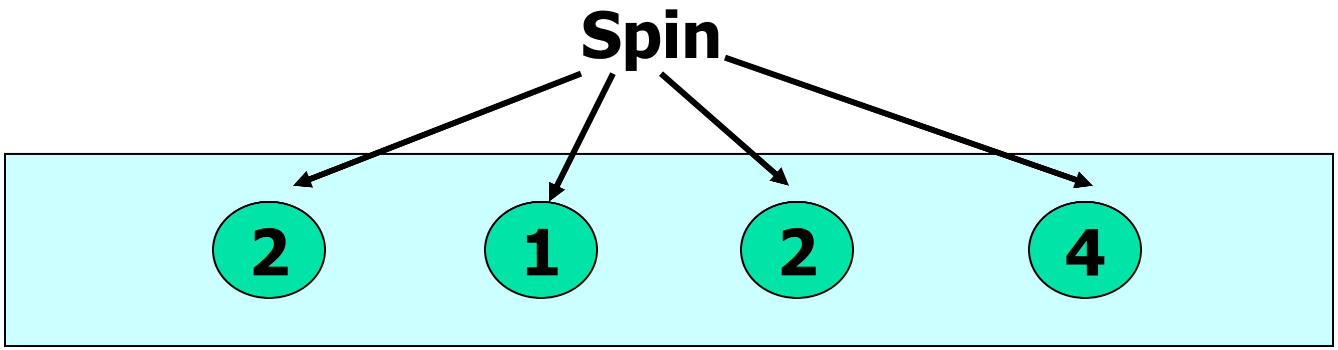
Function $f(x) = x(8-x)$

Chromosome	Corresponding x	Objective/ Fitness fn.
1 0 0 1 1 0 1 0	4.8313	15.3089

Roulette Wheel Selection – Example

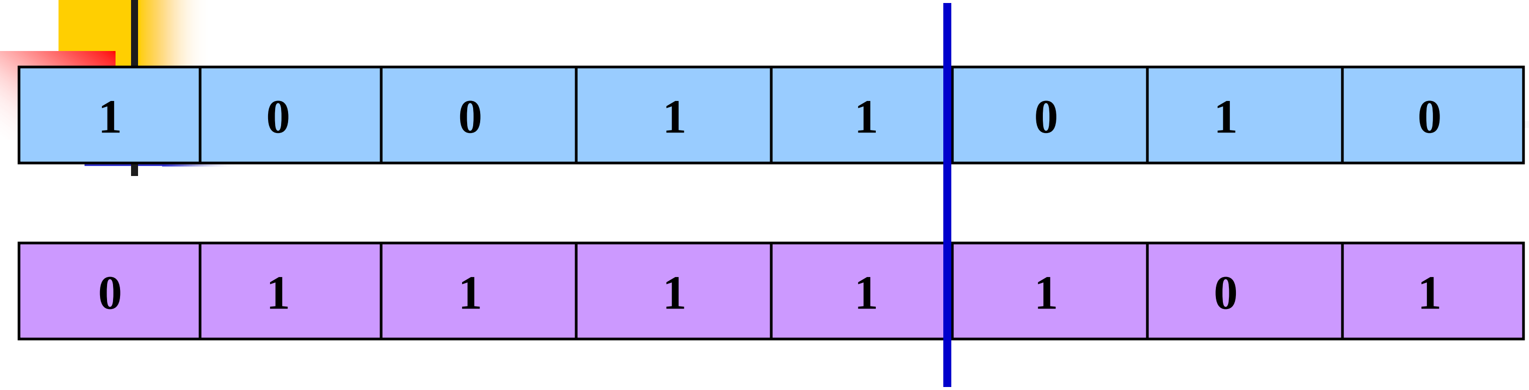
Chromosome #	Fitness
1	15.3089
2	15.4091
3	4.8363
4	12.3975





Mating Pool

Crossover – Example

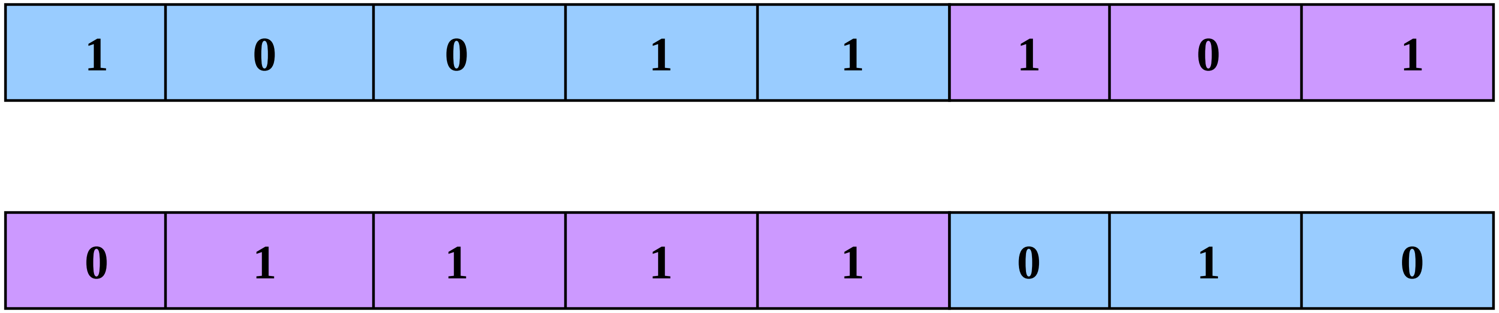


1	0	0	1	1	0	1	0

0	1	1	1	1	1	0	1

Here *l* (string length) = 8. Let *k* (crossover point) = 5

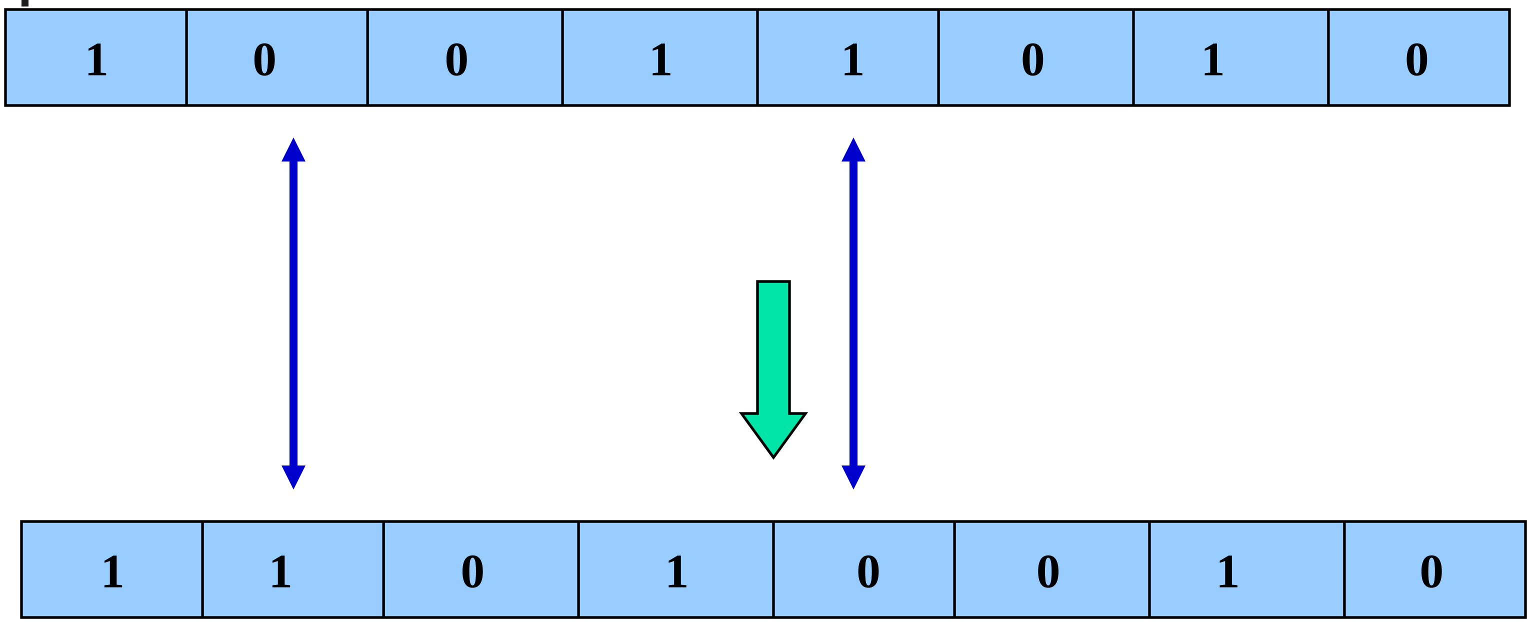
Offspring formed :



1	0	0	1	1	1	0	1

0	1	1	1	1	0	1	0

Mutation- Example



1	0	0	1	1	0	1	0

1	1	0	1	0	0	1	0

Parameters

- Population size – usually fixed
- String length - usually fixed
- Probabilities of crossover, μ_c, and mutation, μ_m
    - μ_c is kept high and μ_m is kept low.
- Termination criteria
- Parameters often manually tuned
- Kept variable or adaptive.

Termination Criterion

- Avg. fitness value of a population more or less constant over several generations,
- Desired objective function value is attained by at least one string in the population,
- Number of generations (or iterations) is greater than some threshold ----- most commonly used.

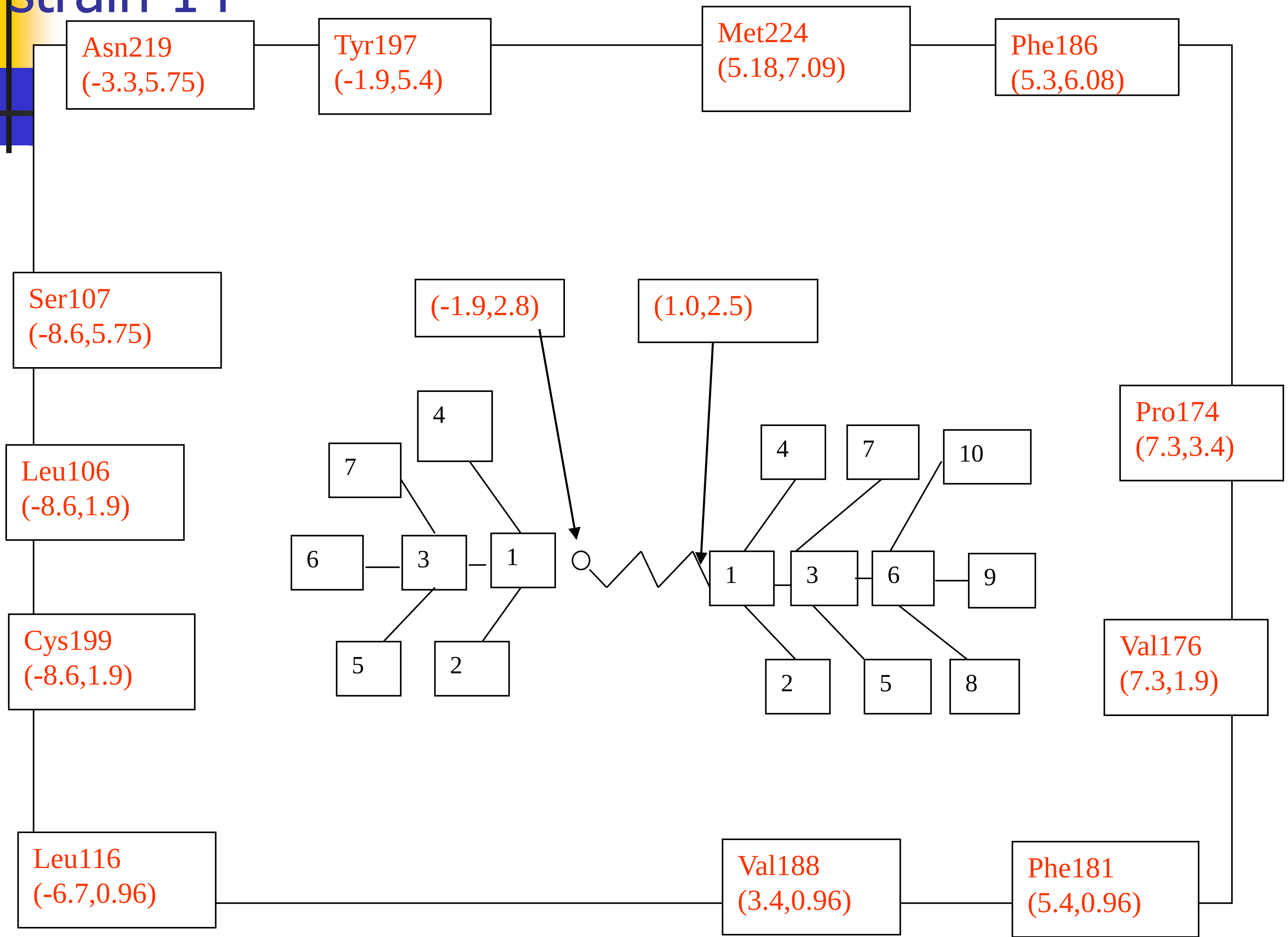
Elitist Model of GAs

The best string seen up to the current generation is preserved in a location either inside or outside the population.

GA for Molecule Design: Problem Objective

- Design of molecules that can bind to the active site of harmful protein (e.g., those crucial for the proliferation of microbial organisms, cancer cells or viruses).
- Such molecules can destroy the action of the target protein
  - thereby nullifying its activity which can be lethal to us.
- Accurate prediction of the structure of the potential inhibitors, while utilizing the knowledge about the structure of a target protein, is important in *drug design*.

Barrel shaped active site of human rhino virus strain 14



Asn219 (-3.3,5.75)

Tyr197 (-1.9,5.4)

Met224 (5.18,7.09)

Phe186 (5.3,6.08)

Ser107 (-8.6,5.75)

Leu106 (-8.6,1.9)

Cys199 (-8.6,1.9)

Leu116 (-6.7,0.96)

Val188 (3.4,0.96)

Phe181 (5.4,0.96)

Pro174 (7.3,3.4)

Val176 (7.3,1.9)

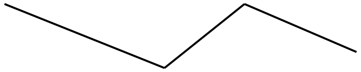
(-1.9,2.8)

(1.0,2.5)

The Design Technique

- The ligand molecule is assumed to have a tree structure on both sides of the *pharmacophore* – the functional part of the molecule.
- The tree is to be filled up by a group from a set of pre-defined 7 groups.
- Van der Waals energy is taken as the minimizing criterion.
- GA is used for minimization

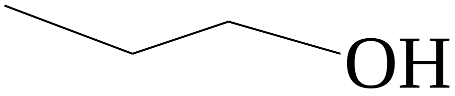
Groups to be taken

- Group 0 Alkyl 1C
  - Bond length ~0.65 along x-axis
- Group 1 Alkyl 3C 
  - Bond length ~ 1.75 along x-axis
- Group 2 Alkyl 1C Polar

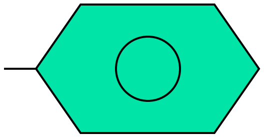
  - Bond length ~ 1.1 along x-axis

Groups To Be Taken

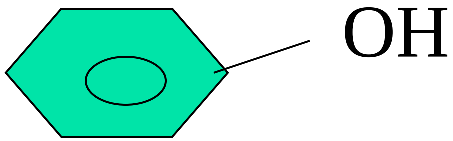
- Group 3 Alkyl 3C Polar



  - Bond length ~ 2.2 along x-axis
- Group 4 Polar —OH
- Group 5 Aromatic



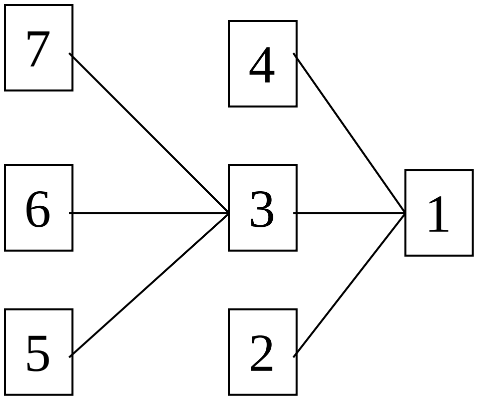
  - Bond length ~1.9 along x-axis
- Group 6 Aromatic polar



  - Bond length ~ 2.7 along x-axis

Encoding Technique

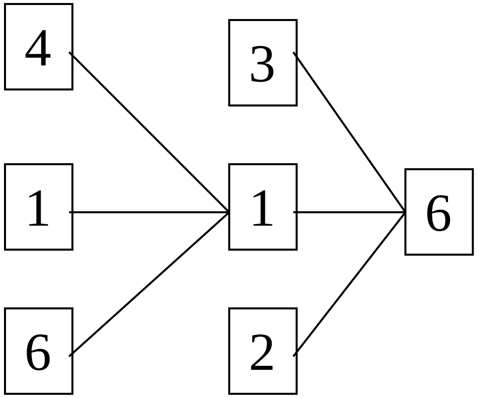
- Chromosome will encode a tree on one side of the pharmacophore.
- The size of the tree is not fixed a priori
  - The ordering of the nodes is fixed.



Node ordering

6 2 1 3 6 1 4

Chromosome



Encoded tree

Fitness Computation

- Based on the proximity of the groups involved
- The distance between the groups & the protein lies between 2.7 & 0.65 Å.
- The interacting groups should be of similar polarity – a polar group should face a polar group & vice versa.
- Van der Waals energy = $[(C_n / r^6) - (C_m / r^{12})],$
  - n and m are integers and C_n and C_m are constant values dependent on the atom pair
  - r is the distance between the atoms
- The total energy is sum of all these energy values.
- fitness value = 1/energy
  - the maximization of the fitness by VGAs leads to the minimization of the energy.

Further Enhancements: Consideration of Nonbonding Interactions

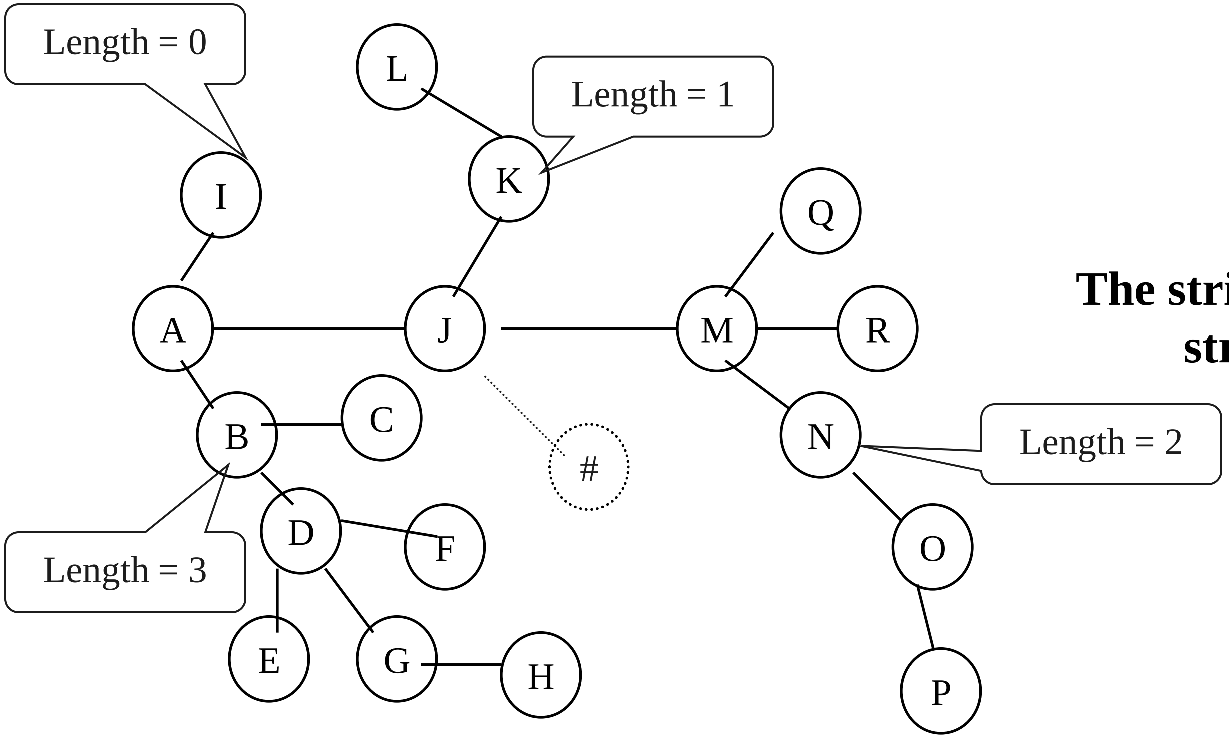
- **Van der Waals energy =** $[(C_n / r^6) - (C_m / r^{12})]$
- **Electrostatic energy =** $(q_1 q_2)/(4\pi\varepsilon_0 r^2)$
  - $\varepsilon_0 = 8.854185\text{x}10^{-12}$ **coulomb^2/(N m^2)**

ENCODING STRATEGY

B	L	LC	LL	LU	LM	LML	LMU	….	U	UC	UL	UU	UM	UL	….	B	L

Lower tree structure

Upper tree structure



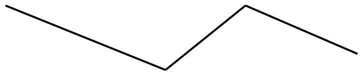
The string representation of this structure is as follows

AB3#CDEFG#H#I0J#0K1#L#MN2##0P##Q0R

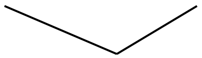
Additional Groups Considered

- Group 0 Alkyl 1C

  - Bond length ~0.65 along x-axis
- Group 1 Alkyl 3C

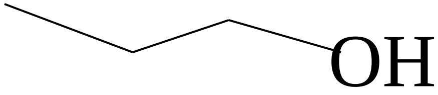


  - Bond length ~ 1.75 along x-axis
- Group 2 Alkyl 1C Polar
  - Bond length ~ 1.1 along x-axis OH

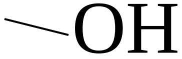


Groups Considered

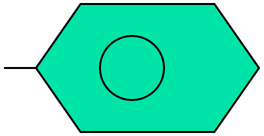
- Group 3 Alkyl 3C Polar
  - Bond length ~ 2.2 along x-axis



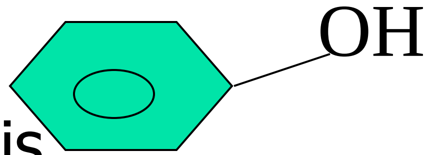
- Group 4 Polar



- Group 5 Aromatic
  - Bond length ~1.9 along x-axis

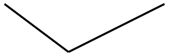


- Group 6 Aromatic polar
  - Bond length ~ 2.7 along x-axis

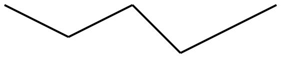


Groups Considered

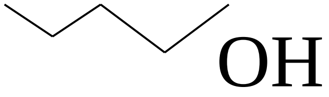
- Group 7 Alkyl 2C
  - Bond length ~ 1.2 along x-axis



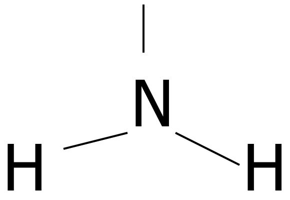
- Group 8 Alkyl 4C
  - Bond length ~ 2.5 along x-axis



- Group 9 Alkyl 4C Polar
  - Bond length ~ 2.9 along x-axis

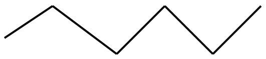


- Group 10 Amine NH_2
  - Bond length ~ 0.5 along x-axis

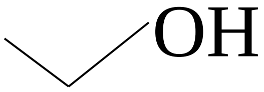


Groups Considered

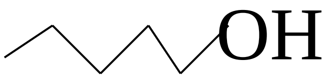
- Group 11Alkyl 5C
  - Bond length ~ 3.1 along x-axis



- Group 12 Alkyl 2C Polar
  - Bond length ~ 1.68 along x-axis



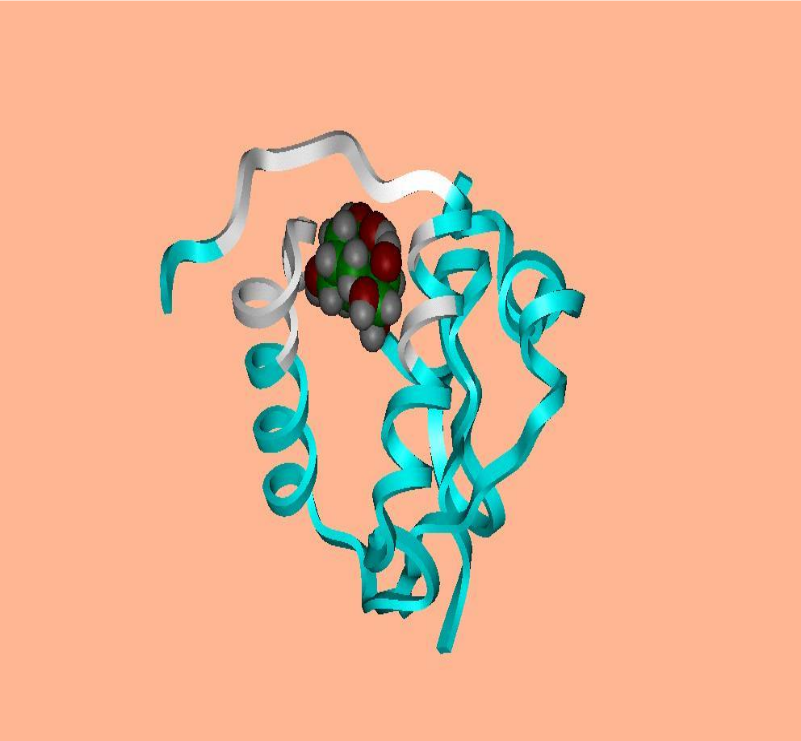
- Group 13 Alkyl 5C Polar
  - Bond length ~ 3.58 along x-axis



Experimental Results

- Experimented with two protein targets
  - HIV-1 Nef Protein
  - HIV Protease
- Two algorithms
  - VGA – An earlier GA based method
  - IVGA – Improved version (present work)
- Real Molecules
  - From Cambridge structural database

HIV-1 Nef protein docked with a molecule designed by IVGA



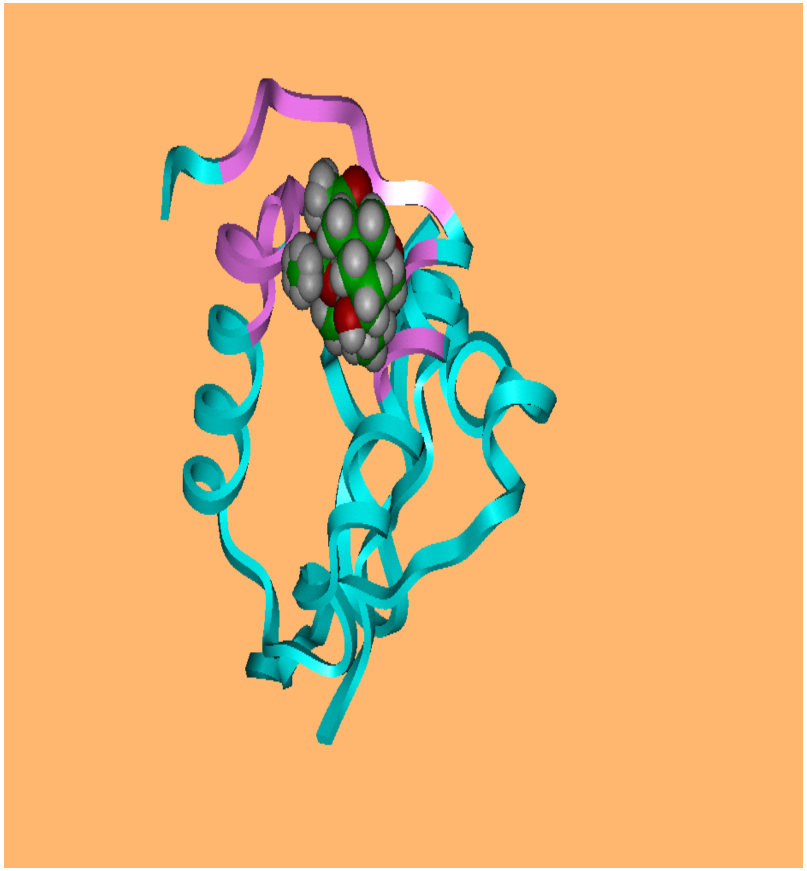
- Color code for HIV-1Nef

  Cyan : protein
  White : Active site

- Color code for ligand

  White : Hydrogen
  Red : Oxygen
  Green : Carbon

HIV-1 Nef protein docked with a molecule from CSD similar to the molecule designed by IVGA



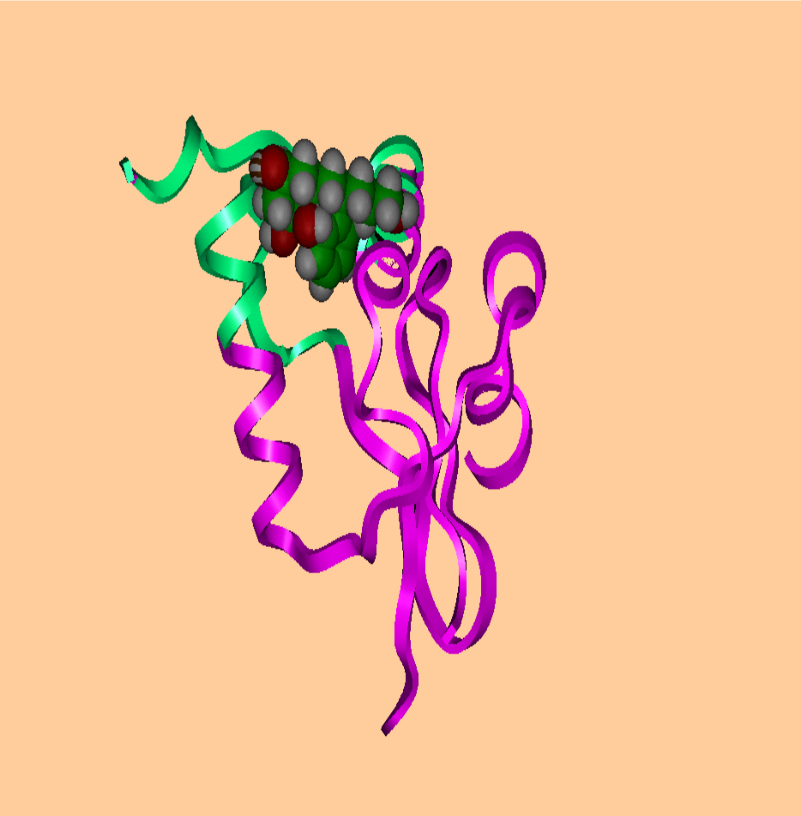
- Color code for HIV-1Nef

  Cyan : protein
  Pink : Active site

- Color code for ligand

  White : Hydrogen
  Red : Oxygen
  Green : Carbon

HIV-1 Nef protein docked with a molecule designed by VGA



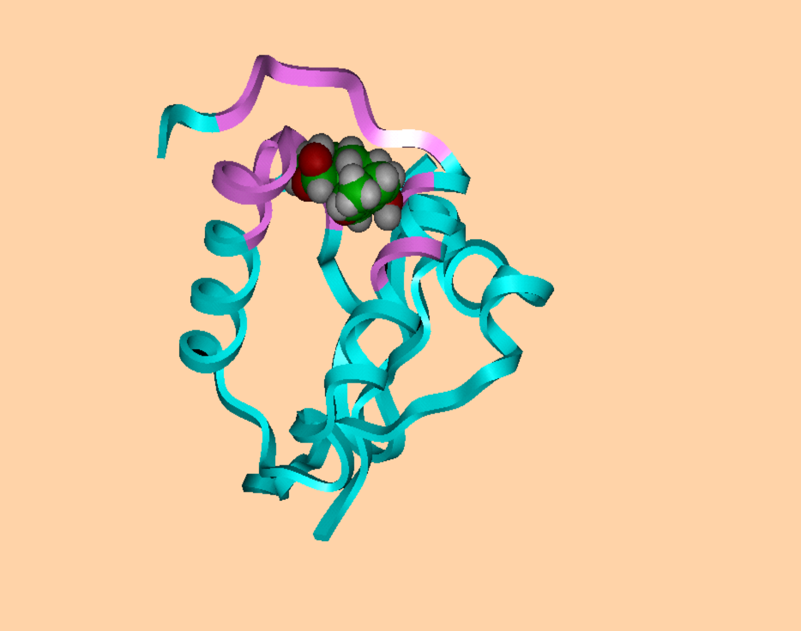
- Color code for HIV-1Nef

Purple	: protein
Green	: Active site

- Color code for ligand

White	: Hydrogen
Red	: Oxygen
Green	: Carbon

HIV-1 Nef protein docked with a molecule from CSD similar to the molecule designed by VGA



- Color code for HIV-1Nef

  Cyan : protein

  Pink : Active site

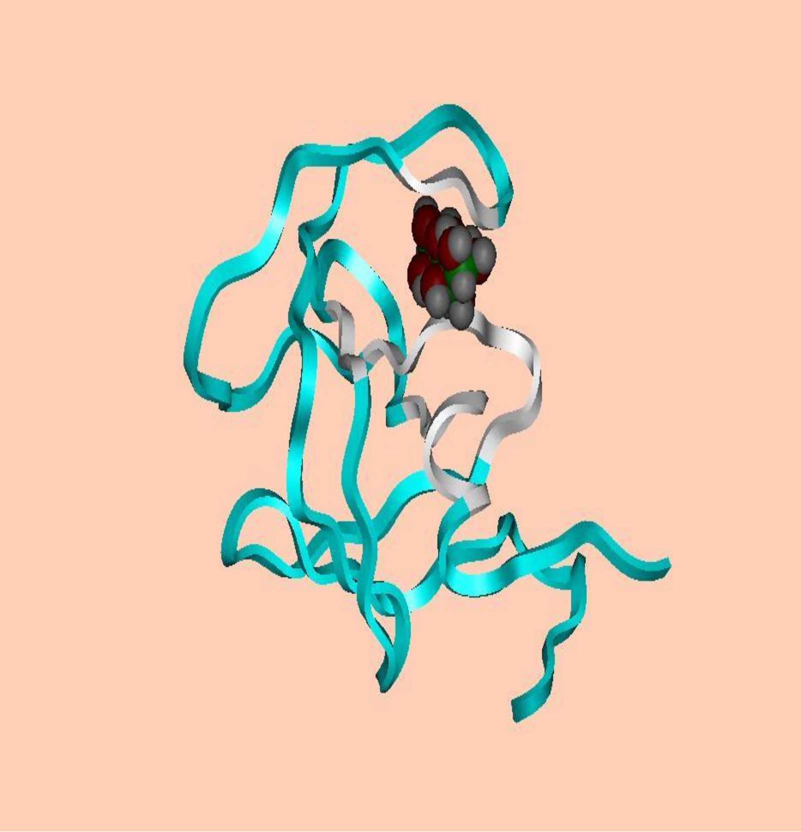
- Color code for ligand

  White : Hydrogen

  Red : Oxygen

  Green : Carbon

HIV Protease docked with a molecule designed by IVGA



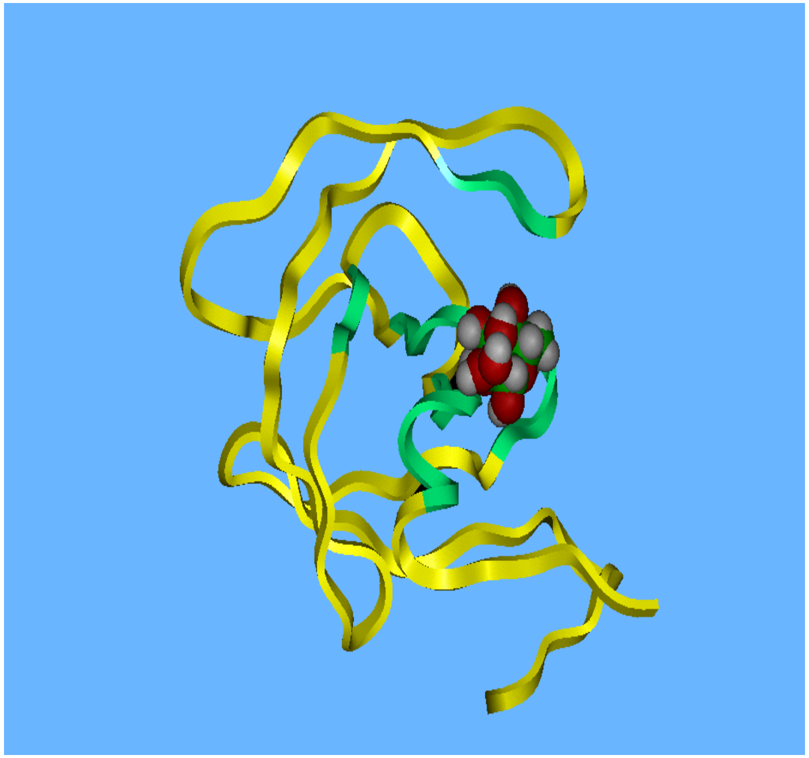
- Color code for HIV-1Nef

  Cyan : protein
  White : Active site

- Color code for ligand

  White : Hydrogen
  Red : Oxygen
  Green : Carbon

HIV Protease docked with a molecule from CSD similar to the molecule designed by IVGA

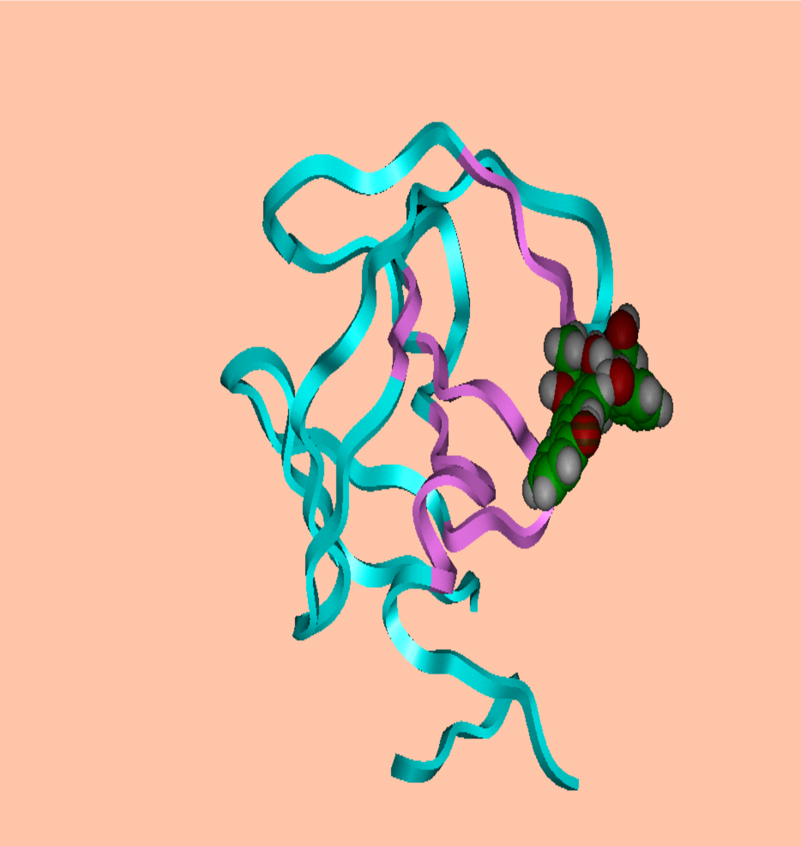


- Color code for HIV-1Nef

  Yellow : protein
  Green : Active site

- Color code for ligand
  White : Hydrogen
  Red : Oxygen
  Green : Carbon

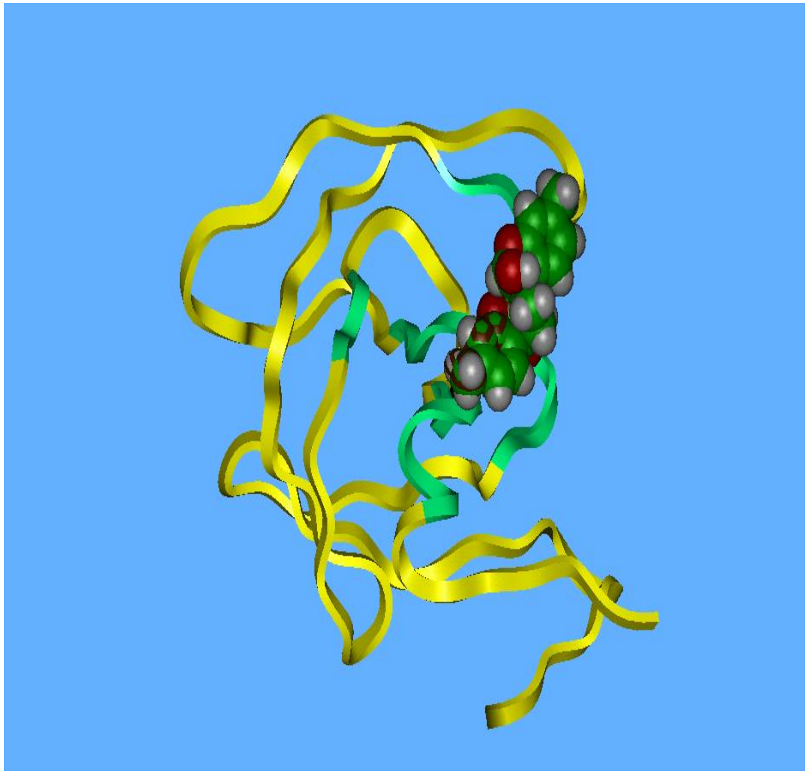
HIV Protease docked with a molecule designed by VGA



- Color code for HIV-1Nef
  Cyan : protein
  Pink : Hydrogen

- Color code for ligand
  White : Hydrogen
  Red : Oxygen
  Green : Carbon

HIV Protease docked with a molecule from CSD similar to the molecule designed by VGA



- Color code for HIV-1Nef

  Yellow : protein
  Green : Active site

- Color code for ligand

  White : Hydrogen
  Red : Oxygen
  Green : Carbon

Comparative Quantitative Results

Energy Values (by InsightII in Kcal/mole)	HIV-1 Protease		HIV-1 Nef Protein	
	VGA	IVGA	VGA	IVGA
Vander Waals Energy	**-9.47589**	**-10.4479**	**-6.80164**	**-6.84964**
Coulombs Energy	**4.15411**	**-2.36619**	**-2.3071**	**-4.30512**
Total Energy	**-5.32178**	**-12.8141**	**-9.10874**	**-11.1546**

Comparison with Real Molecules in CSD

Name of the protein	Method used	CSD Ref code of the molecule	Energy (kcal)
HIV-I-Nef	VGA	IFEFOO	-11.43518
	IVGA	ADAKEW	-26.39
HIV Protease	VGA	VEHMUQ	-17.7638
	IVGA	UNIHII	-35. 0094

Hydrogen Bonds For HIV 1- Nef Protein

- IVGA (Improved variable tree length genetic algorithm)

Donor	Acceptor	Distance(Å)
LigNEF:1C:OH	P_NEF:B83:O	2.32
LigNEF:1I:OH	P_NEF:B120:O	1.87
LigNEF:1I:OH	P_NEF:B124:ONE1	1.91
LigNEF:1K:OH	P_NEF:B79:N	2.80

- VGA (Variable tree length genetic algorithm)

Donor	Acceptor	Distance(Å)
LigNEF:1K:HH	P_NEF:B117:N	2.80

Hydrogen Bonds For HIV Protease Protein

- IVGA (Improved variable tree length genetic algorithm)

Donor	Acceptor	Distance(Å)
P_1AAQ:A48:HN	LigAAQ:1L:OH	2.38
LigAAQ:1L:HH	P_1AAQ :A48:N	2.36
LigAAQ:1L:HH	P_1AAQ :A48:O	2.48

- VGA (variable tree length genetic algorithm)

Donor	Acceptor	Distance(Å)
P_1AAQ:A87:HH11	LigAAQ:1C:OH	2.13
LigAAQ:1C:HH	P_1AAQ:A87:NH1	2.30

Conclusions and Further Work

- An Improved VGA based technique for ligand design is proposed
  - no assumption regarding the size of the tree
  - Modified crossover and mutation operators are used.
- Proposed method found to provide solutions having characteristics amenable to stability
- Lipinski Rule of Five, a drug like compound must not have molecular weight more than 500Da.
  - The new molecule designed is smaller and binds to the given protein to form a more stable complex than the molecules designed by a previous approach.
- Scope for further work
  - Need to analyze in 3 dimensions
  - Consider other optimizing criteria and multi-objective optimization algorithms
  - Consider structures other than tree

Publications

- **Books**
  - S. Bandyopadhyay and S. K. Pal, Classification and Learning Using Genetic Algorithms: Applications in Bioinformatics and Web Intelligence, Springer, Heidelberg, 2007.
  - S. Bandyopadhyay, U. Maulik and J. T. L. Wang, (eds.), Analysis of Biological Data: A Soft Computing Approach, World Scientific, Singapore, 2007.
  - U. Maulik, S. Bandyopadhyay and J. T. L. Wang, Computational Intelligence and Pattern Analysis in Biological Informatics, John Wiley (accepted).

- **Articles**
  - S. Bandyopadhyay, A. Bagchi and U. Maulik, ``Active Site Driven Ligand Design: An Evolutionary Approach", *Journal of Bioinformatics and Computational Biology*, vol. 3, No. 5, pp. 1053-1070, 2005.
  - S. Santra and S. Bandyopadhyay, "Grid Count Tree Based Method For Efficient Outlier Detection", *Proceedings of the International Conference on Emerging Applications of IT*, February 10-11, Kolkata, India, pp. 309-312, 2006.
  - S. Bandyopadhyay, A. Mukhopadhyay and U. Maulik, "An Improved Algorithm for Clustering Gene Expression Data", *Bioinformatics*, Oxford University Press, vol. 23, no. 21, pp. 2859-2865, 2007.
  - S. Bandyopadhyay and S. Santra, "A Genetic Approach for Efficient Outlier Detection in Projected Space", *Pattern Recognition*, vol. 41, no. 4 pp. 1338-1349, 2008.
  - S. Bandyopadhyay, S. Santra, U. Maulik and H. Muehlenbein, "In Silico Design of Ligands Using Properties of Target Active Sites", Analysis of Biological Data: A Soft Computing Approach, World Scientific, pp. 184-201, 2007.

Publications contd…

- **Articles**
  - S. S. Ray, S. Bandyopadhyay, and S. K. Pal, “Genetic Operators for Combinatorial Optimization in TSP and Microarray Gene Ordering", *Applied Intelligence,* vol. 26, no. 3, pp. 183-195, 2007
  - S. Bandyopadhyay, U. Maulik and D. Roy, ``Gene Identification: Classical and Computational Intelligence Approaches", *IEEE Transactions on Systems, Man and Cybernetics,* Part C, vol. 38, no. 1, pp. 55-68, 2008.
  - S. Bandyopadhyay, S. Saha, U. Maulik and K. Deb, ``A Simulated Annealing Based Multi-objective Optimization Algorithm: AMOSA", *IEEE Transaction on Evolutionary Computation,* vol. 12, no. 3, pp. 269-283, 2008.
  - R. Chakraborty, S. Bandyopadhyay and U. Maulik, ``Extracting Features for Protein Sequence Classification", *Intl. Conf. on IT: Prospects and Challenges* (ITPC), 2003.
  - S. Bandyopadhyay, “An Efficient Technique for Superfamily Classication of Amino Acid Sequences: Feature Extraction, Fuzzy Clustering and Prototype Selection", *Fuzzy Sets & Systems*, vol. 152, pp. 5-16, 2005

References

- N. M. Luscombe, D. Greenbaum, M. Gerstein, ``What is Bioinformatics: A Proposed definition and Overview of the field", Methods Inf. Med., vol. 40, pp. 346-358, 2001.
- J. Setubal and J. Meidanis, "Introduction to computational biology", Brooks Cole, 1997.
- Jason T.L. Wang, Qi Cheng. Ma, Dennis Shasha, Cathy H. Wu, ``New Techniques for Extracting Features from Protein Sequences", IBM Systems Journal, Special Issue on Deep Computing for the Life Sciences, vol-40, no-2, pp. 426-441, 2001.
- D. E. Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, New York, 1989.
- Michalewicz, Z. 1996. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag. Third edition.
- Mitchell, Melanie. 1996. *An Introduction to Genetic Algorithms*. Cambridge, MA
- Beyer, Hans-Georg. 2001. *The Theory of Evolution Strategies*. Heidelberg: Springer-Verlag.
- Schwefel, Hans-Paul. 1995. *Evolution and Optimum Seeking*. New York, NY: John Wiley.
- Fogel, David B. 1991. *System Identification through Simulated Evolution*. Needham Heights, MA: Ginn Press.
- Kalyanmoy Deb. Multi-Objective Optimization using Evolutionary Algorithms, John Wiley & Sons, Chichester, UK, 2001, ISBN 0-471-87339-X.
- Carlos A. Coello Coello, David A. Van Veldhuizen and Gary B. Lamont, Evolutionary Algorithms for Solving Multi-Objective Problems, Kluwer Academic Publishers, New York, March 2002, ISBN 0-3064-6762-3.

References

- S. Bandyopadhyay, S. K Pal, and B. Aruna, "Multi-objective GAs, quantitative Indices and Pattern Classification", *IEEE Transactions on Systems, Man and Cybernetics - B,* vol. 34, no. 5, pp. 2088-2099, 2004.
- U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 12, pp. 1650-1654, 2002.
- U. Maulik and S. Bandyopadhyay, "Fuzzy Partitioning Using Real Coded Variable Length Genetic Algorithm for Pixel Classification", *IEEE Transactions on Geosciences and Remote Sensing,* vol. 41, no. 5, pp. 1075-1081, 2003.
- S. Bandyopadhyay, "Simulated Annealing Using Reversible Jump Markov Chain Monte Carlo Algorithm for Fuzzy Clustering", *IEEE Transactions on Knowledge and Data Engineering,* vol. 17, no. 4, pp. 479-490, 2005.

Thank you..