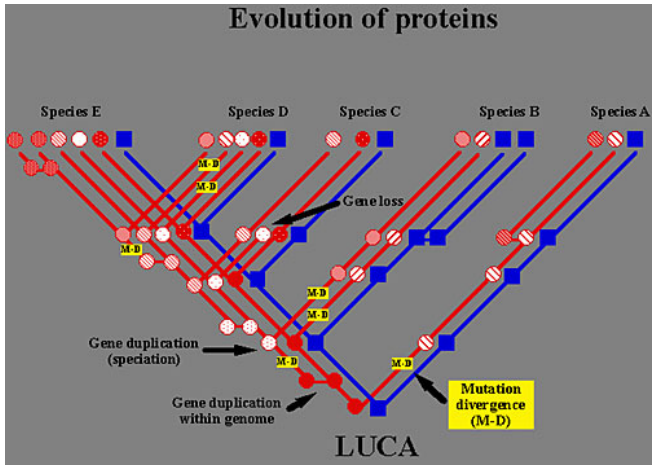
Computational Biology

What for ?

Protein family

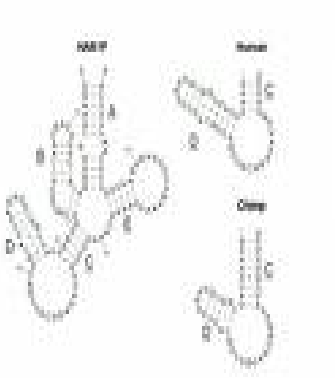


Gene sequence

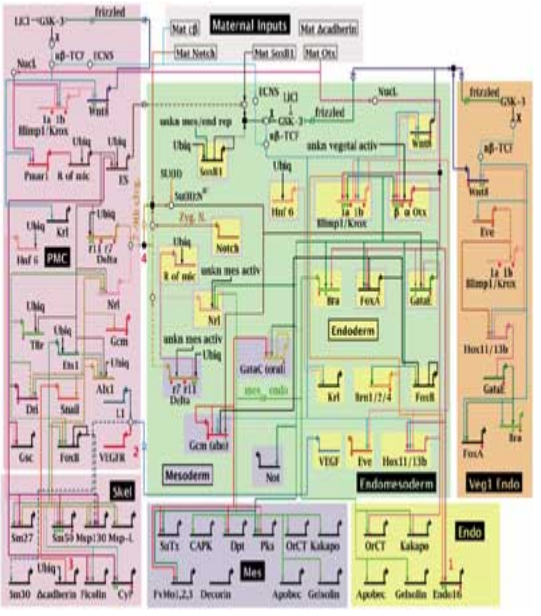
```
                420         *         440         *         460         *
Hs-RPL3-I3 : AATGCGTGTAACTTAA-ATTAACCTTGTGGACCTCTGCTCAGCTCCGCTCGGCTCTGCC : 45
Mm-RPL3-I3 : TATATTTGTGGCTTAT-ATTAACCT-GTGGACCTCTGCTCAGCTCCGCTCGGCTCTGCC : 42
Bt-RPL3-I3 : GAATGTTGTGTTTCATTATTAACCTTGTGGACCTCTGCTCAGCTCCGCTCGGCTCTGCC : 39
              At  tTGTg cTtAt ATTAACCTtGTGGACCTCTGCTCAGCTCCGCTCGGCTCTGCC

                480         *         500         *         520         *
Hs-RPL3-I3 : CGATGAGCTCCATCCAGGCTCCGCTTGCCGGTGGAAAAGGCTCCTTAGAAGCCGGCAAT : 51
Mm-RPL3-I3 : CGATGAGCTTCATCCAGGCTCCGCTTGCCGGTGGAAAAGGCTCCTTAGAAGCCGGCAAT : 47
Bt-RPL3-I3 : CGATGAGCTCCATCCAGGCTCCGCTTGCCGGTGGAAAAGGCTCCTTAGAAGCCGGCAAT : 45
             CGATGAGCTcCATCCAGGCTCCGCTTGCCGGTGGAAAAGGCTCCTTAGAAGCCGGCAAT

                540         *         560         *         580         *
Hs-RPL3-I3 : GAGCTCCATCCCCACGCGGTGCCAGTGTGCCTTCCGCTCACCCCTCGGAGGGGTGATGA : 57
Mm-RPL3-I3 : GAGCCCCATCCCCAAATGGTGCCAGTCTGCCTTCCTCTCACCTGTTGCAAGGGTGATGA : 53
Bt-RPL3-I3 : GAGCCCCGTCCCCACACGGTGCCTGTGTGCCTCCCCCTCACCCGTTGGAGGGCTGATGA : 50
             GAGCcCCaTCCCCAcacGGTGCCaGTgTGCCTtCC CTCACCcgTtGgAgGGgTGATGA

                600         *         620         *
Hs-RPL3-I3 : AGGCCTGCACC-TGGTCCCCTCCCCAACTCTGCTCTGCTCCTGAAG : 619
Mm-RPL3-I3 : AGGCCTGCACC-GGGCCCCTCCCCCAACTGTGCTCTGCTCCTGAAG : 583
Bt-RPL3-I3 : AGGCTGGCACCTGGGCCCCCTCCCCAACTCTGCTCTGCTCCTGAAG : 555
             AGGCctGCACC gGGcCCCctCCCCAACTcTGCTCTGCTCCTGAAG
```

Figure 2. Alignment of human, bovine and mouse RPL3 gene (intron 3) using the Genedoc multiple alignment program. Note the identical positions of nucleotides.

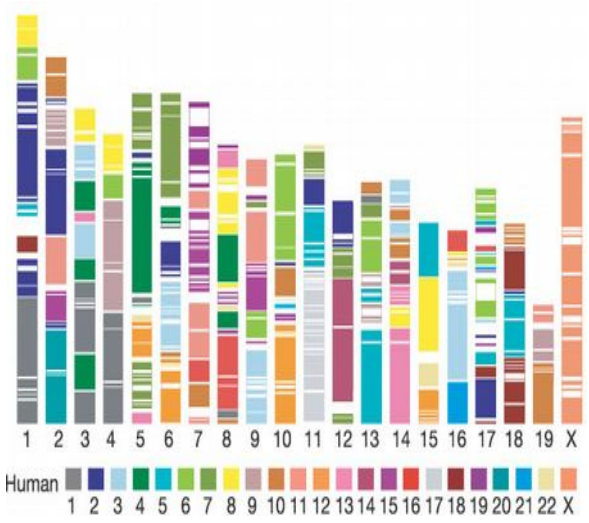
Structure



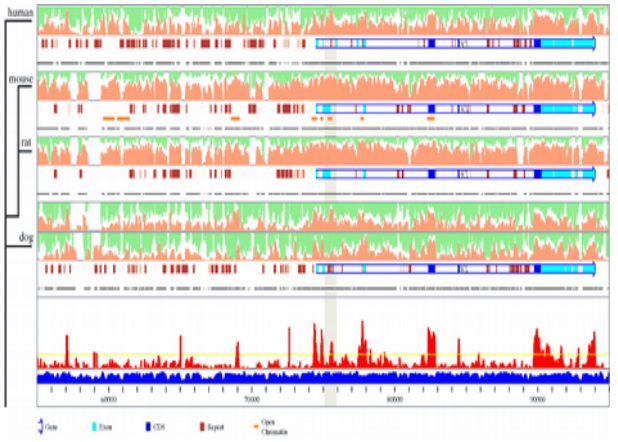
Gene networks



Gene order

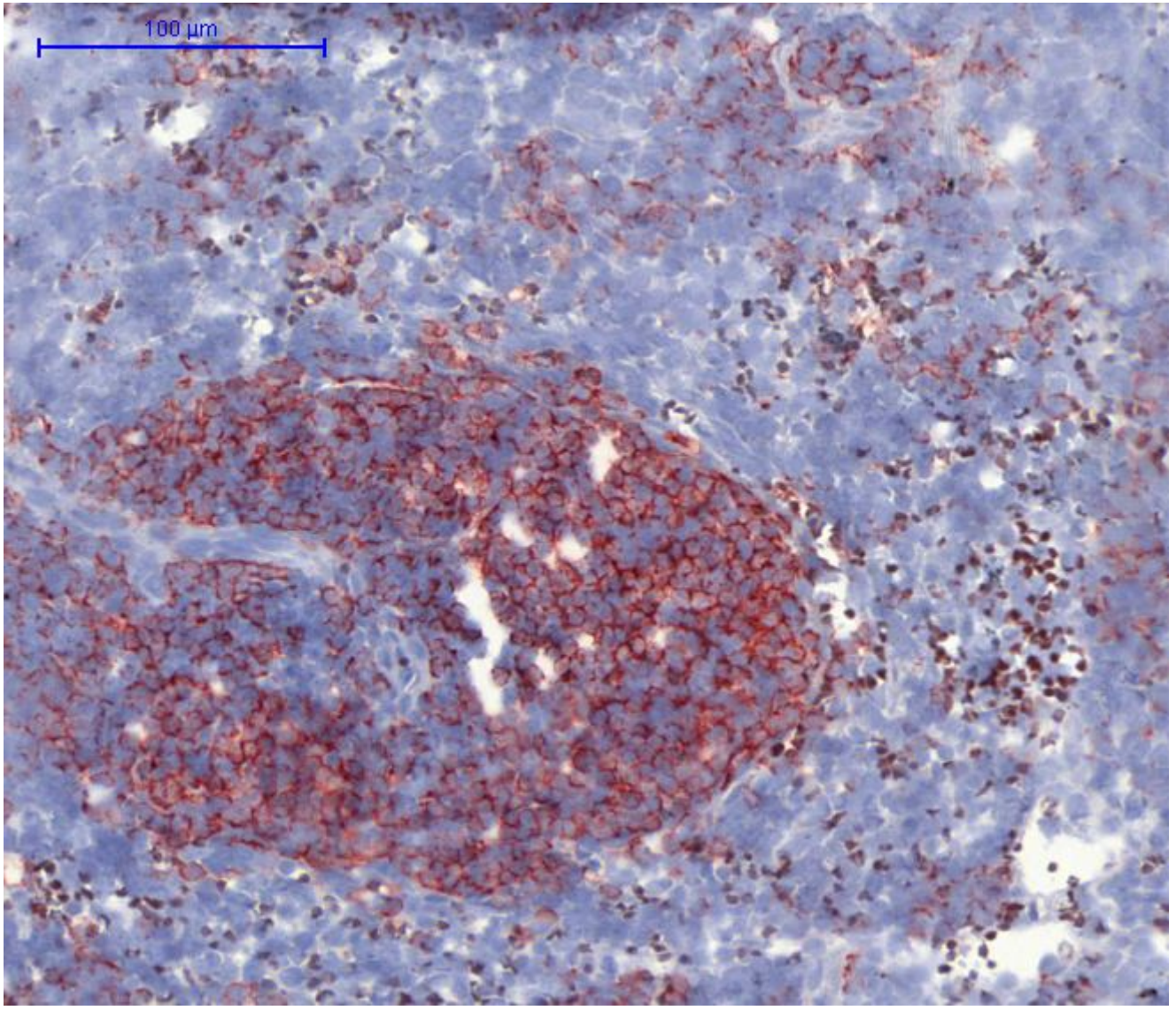


Non-coding regions



Computational Biology
What for ?

And their phenotypic counterpart

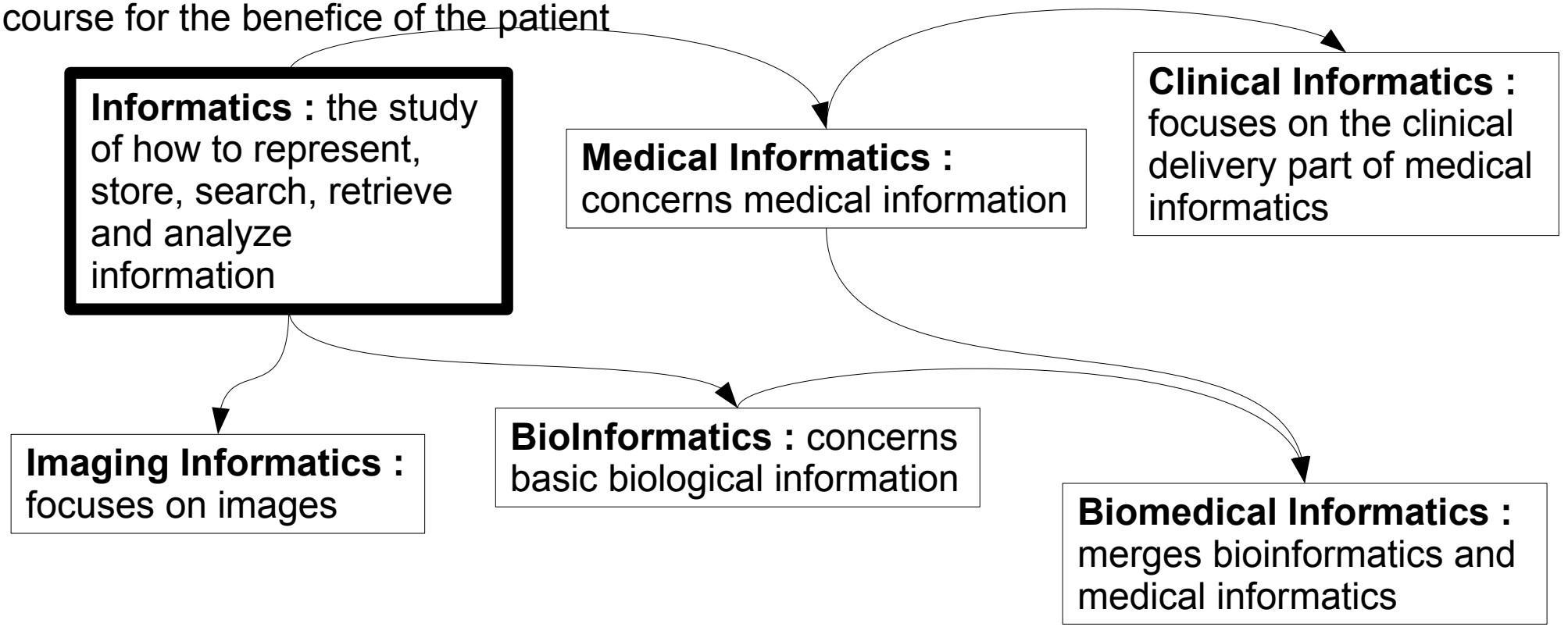


Translational BioInformatics

PLOS Computational Biology
A Peer-Reviewed, Open Access Journal
Translational Bioinformatics : a collection of Education Articles, 2012
http://www.ploscollections.org/article/browseIssue.action?issue=info:doi/10.1371/issue.pcol.v03.i11

Impact of Computational Biology : translational sciences

Integrate huge amount of **heterogeneous** molecular and clinical data for a better understanding of molecular basis of diseases and subsequently **changing** clinical practices of course for the benefice of the patient



Translational :

- How to improve diagnostic, pronostic and patients' care ?
- Small devices
- Molecular dignostic
- Nano-particules based treatment
- Vaccine
  Etc.
- Mastering the huge amount of new knowledge in molecular biology, genetics and genomic.

Double helixoidal structure of ADN → pratical improvement of human health from a technological point of view ?

For sure, we are able to quickly compute/measure :

- DNA sequences (whole genome scale)
- RNA sequences and expression
- protein sequences, structure, expression and modification
- structure, presence and quantity of small molecular metabolites
- generate a lot of data including images

2 playground chapters for this sessions :

- Quantitative Imagery

- Machine Learning / Data Mining

2 important chapters in an ideal world :

- Graph and Network representations

- Knowledge representations : data, database, ontologies

Then technologies/ environments for sotware use/development :

- Java : ImageJ, Weka

- Python : Biopython, Numpy, Scipy, Matplotlib, Pyvis, Enthought Python Distribution and Canopy, Anaconda Pandas

- scripts : Perl, Gawk

- Inkscape, ImageMagik / Sphinx / XML, SBML, BioPax, GPML, JSON, SQL, noSQL, Hadoop

- Clustal → T-coffee, PathwayAPI, BioGRID, PatternHunter ….

Learning code In a biological perspective for being able to get involved in...

- **Gene feature recognition :**
  - → **TIS (Translation Initiation Site)**
  - → **TSS (Transcriptional Start Sites)**
  - → **Feature Generation → Feature Selection → Feature integration**
  - → **Gene finding**

- **Gene expression analysis :**
  - → **Affymetrix Gene Chip Data**
  - → **Gene expression Profile Classification**
  - → **Gene expression Profile Clustering**
  - → **Gene Regulatory Circuits Reconstruction :** Differentially Expressed Genes, Gene Interaction Prediction

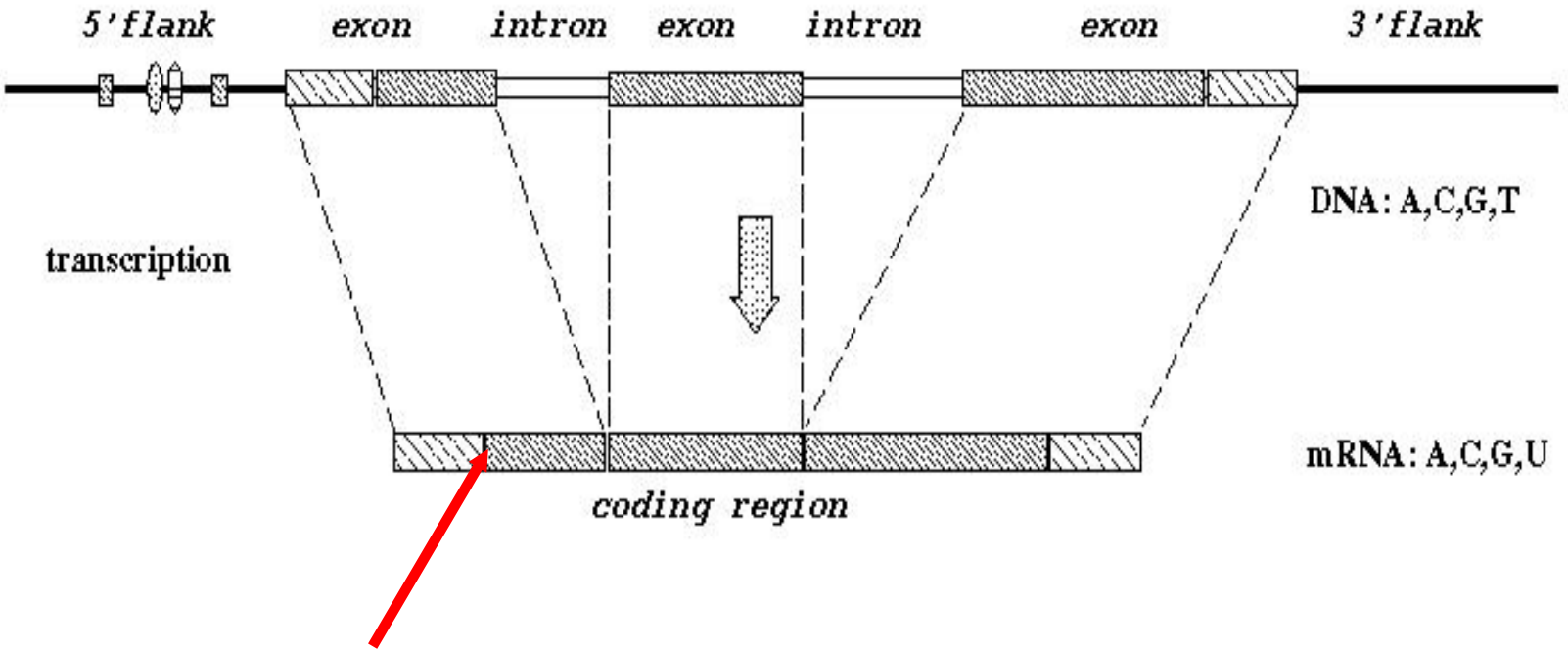
- **Sequence Alignment / Comparaison / Homology :**
  - → **Multiple Sequence Alignment (Dynamic Programming)**
  - → **Function assignment to protein sequence (Guilt by Association)**
  - → **Discovery of Active Site or Domain of a function**
  - → **PPI / Proteomic Profile Analysis**
  - → **key mutation site identification**

- **Phylogenetic tree :**
  - → **Construction**
  - → **Comparison**

- **Image analysis :**
  - → **High-throughput screening**

-**Biological Networks / Graph of Interactions :**
  - → **Natural pathways**
  - → **PPI Networks**
  - → **Protein Complex Prediction**

TIS : Translation Initiation Site Recognition/Prediction

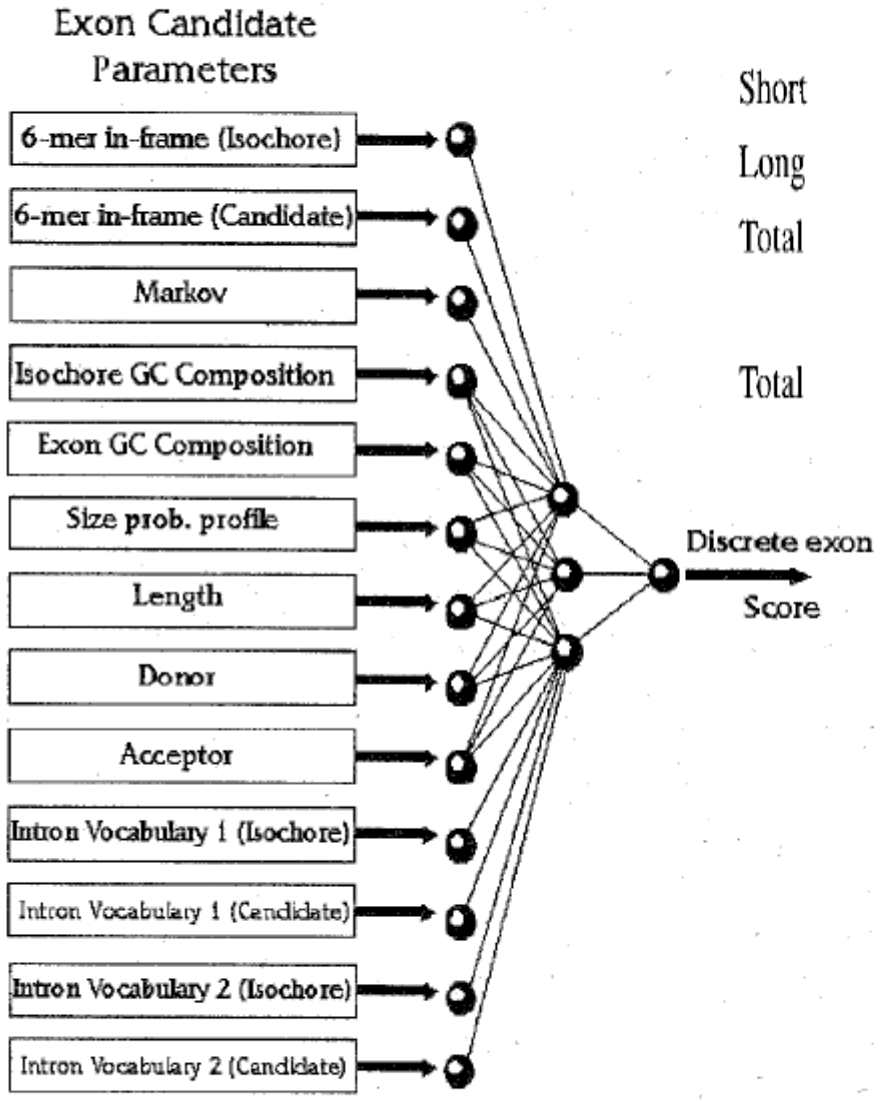


cDNA sample

```
  299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA     160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA     240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
..............................................................      80
..............................iEEEEEEEEEEEEEEEEEEEEEEEEE     160
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE     240
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

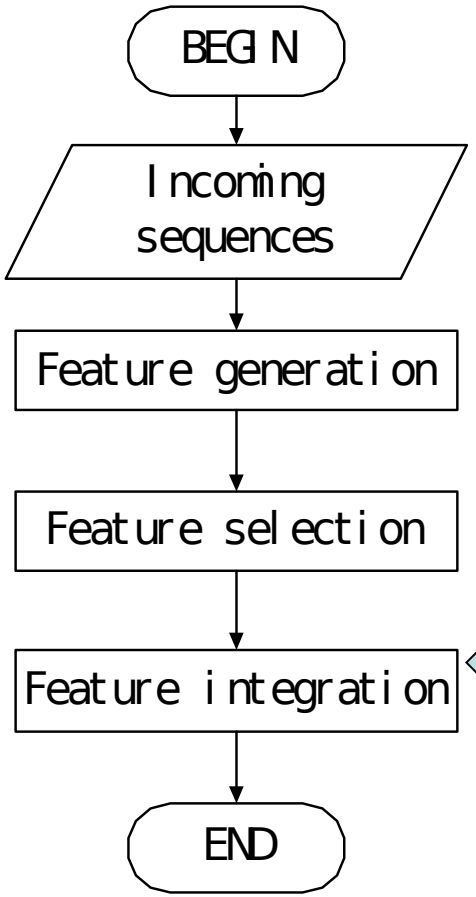
Why the second ATG is a TIS?

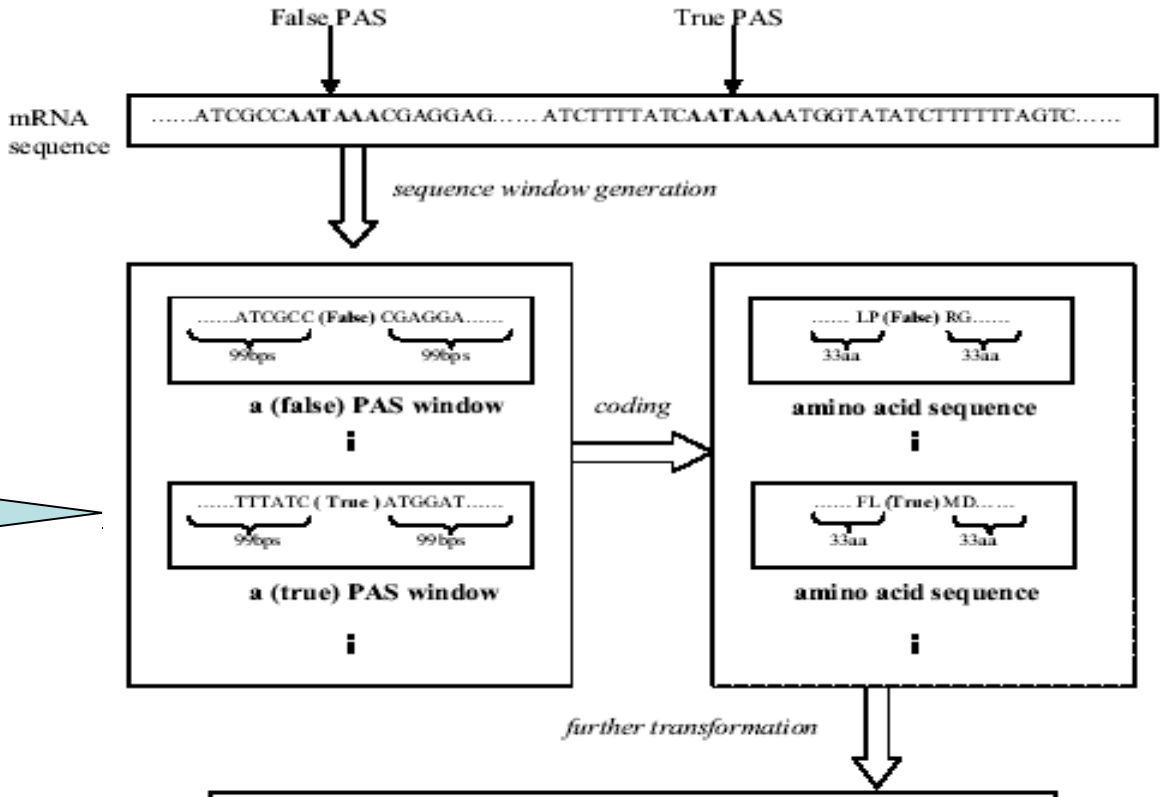
Gene prediction



DNA	Predictions			
# Exons	TP	%	FP	%
Short 229	171	74.7	39	18.6
Long 600	575	95.8	30	4.9
Total 829	746	90.0	69	8.5
# Bases				
Total 134814	122885	91.2	13048	9.6

PAS Prediction





New feature space (total of 925 features + class label)			
42 1-gram amino acid patterns	882 2-gram amino acid patterns	1 bio-know-ledge pattern	class label
UP-A, UP-R, ...,UP-N, DOWN-A, DOWN-R, ..., DOWN-N (numeric type)	UP-AA, UP-AR, ..., UP-NN, DOWN-AA, DOWN-AR , ..., DOWN-NN (numeric type)	UP-T-number (numeric type)	True, False
Frequency as values			
1, 3, 5, 0, 4, ... ! 6, 5, 7, 9, 0, ... ⋮	6, 2, 7, 0, 5, ... ! 2, 0, 3, 10, 0, ... ⋮	10, ! 50, ⋮	False ! True ⋮

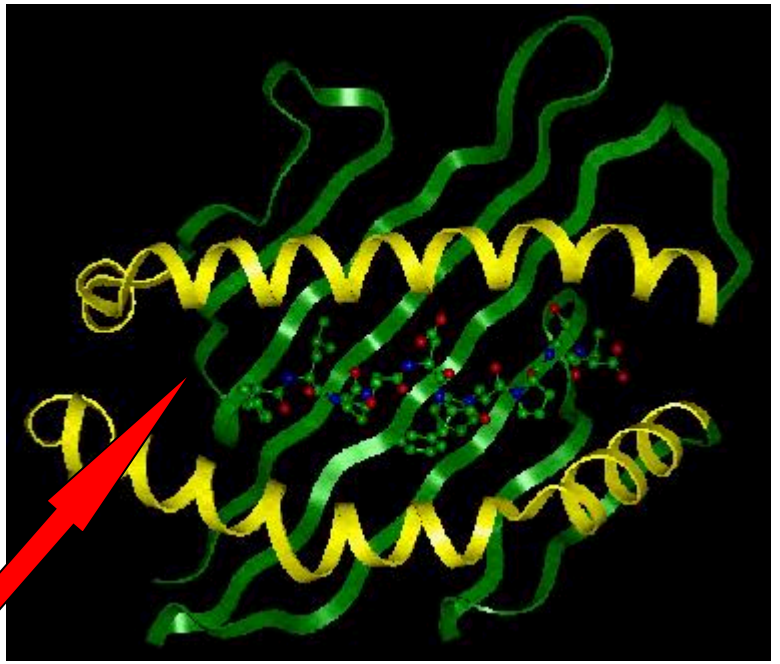
SVM in *Weka*

T-Cell Epitopes Prediction By Artificial Neural Network

- Honeyman et al., *Nature Biotechnology* 16:966-969, 1998

TRAP-559AA

MNHLGNVKYLVIVFLIFFDLFLVNGRDVQNNIVDEIKYSE
EVCNDQVDLYLLMDCSGSIRRHNWVNHAVPLAMKLIQQLN
LNDNAIHLY**VNVFSNNAK**EIIRLHSDASKNKEKALIIIRS
LLSTNLPYGRTNLTDALLQVRKHLNDRINRENANQLVVIL
TDGIPDSIQDSLKESRKLSDRGVKIAVFGIGQGINVAFNR
FLVGCHPSDGKCNLYADSAWENV**KNVIGPFMKAVCVEVEK**
TASCGVWDEWSPCSVTCGKGTRSRKREILHEGCTSEIQEQ
CEEERCPPKWEPLDVPDEPEDDQPRP**RGDNSSVQK**PEENI
IDNNPQEPSPNPEEGKDENPNGFDLDENPENPPNPDIPF
KPNIPEDSEKEVPSDVPKNPEDDREENFDIPKKPENE DN
QNNLPNDKSDRN**IPYSPLPPK**VLDNERKQSDPQSQDNNGN
RHVPNSEDRETRPHGRNNENRSYNRKYNDTPKHEREEHE
KPDNNKKKGESDNKYKIAGGIAGGLAL**LACAGLAYK**FVVP
GAATPYAGEPAPFDETLGEEDKDLDEPEQFRLPEENEWN



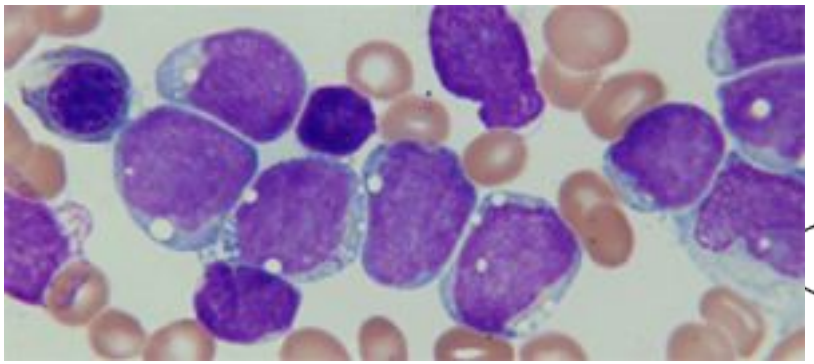
Histone Promoter Recognition Programs

"GENERAL PROMOTERS"			
First Generation			
Name	**Scoring Technique used**	**Search by content/signal**	**Features used**
NNPP	Time delay NN	Signal	TATA box, Inr
Promoter 2	NN	Signal	TATA box, Inr, CAAT box, GC Box
PromFind	Discriminative count	Content	Hexamer frequency
PromoterScan	Discriminative count	Signal	TATA box, TFBS
TSSG/TSSW	Linear discriminant analysis	Content + signal	TATA box, TSS, hexamer frequency upstream TSS, TFBS
Second Generation			
DGSF	NN	Content + signal	CpG island, TSS, DPF output
DPF	NN	Content + signal	Promoter, exon, intron, TSS
Eponine	SVM variant	Content + signal	TATA box, GC rich content, TSS
FirstEF	Quadratic discriminant analysis	Content + signal	First exon, CpG islands
Mcpromoter	NN & Interpolated markov models	Content + signal	TATA box, CAAT box, GC box, nucleosome position
PromoterInspector	Discriminative counts	Content	Oligonuleotides, Exon, Intron, 3'UTR, Promoter genomic context
CpG Promoter	Quadratic discriminant analysis	Content + signal	CpG island, TSS
CpGProD	Generalised linear model	Content	CpG island, AT/GC content
"SUB-CLASS OF PROMOTERS"			
Muscle family	Discriminative counts	Signal	TFBS, relative distance
Globin family	Logical operators AND, OR and NOT	Signal	TFBS, relative distance

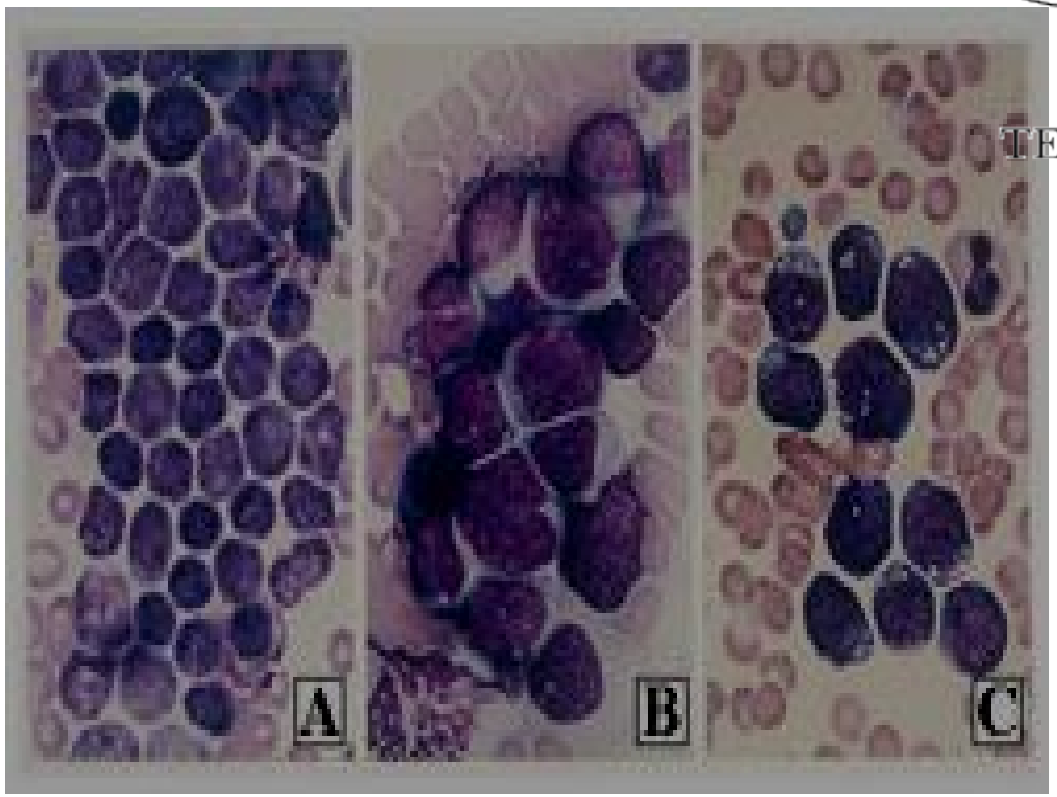
9 Motifs Discovered by MEME algo in Histone Promoter 5' Region [-250,-1] among 127 histone promoters

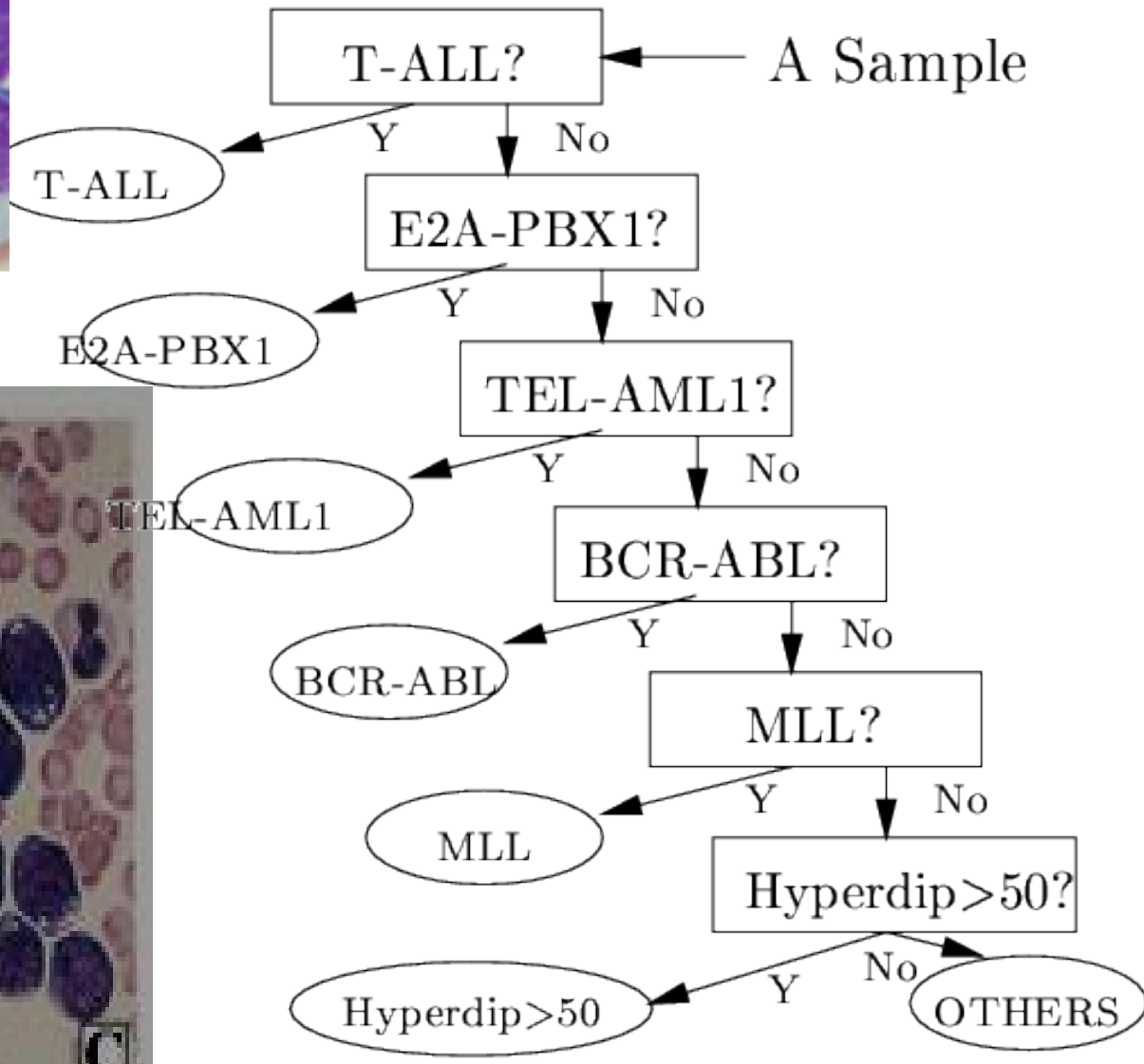
MOTIF NO.	MOTIF DEFINITION	TFBS AND ASSOCIATED FACTORS	TRANSFAC SITE NUMBER
1	TCTGATTGGTTA	CCAAT-box: H1TF2 (La Bella et al. 1989; Martinelli and Heintz 1994; Gallinari et al. 1989), HiNF-B (van Wijnen et al. 1988a,b), NF-Y (Mantovani 1999), HiNF-D (van Wijnen et al 1996; Grimes et al. 2003)	R00660
2	ATGCAAATGAGG	Oct-1: Octamer transcription factor 1 (OTF-1) (Fletcher et al. 1987)	R00662
3	CTATAAAAACC	TATA-box: TBP, TFIID (Nakajima et al. 1988)	R00770
4	TTTTCGCGCCCA	E2F-binding site: E2F-1 factor (Oswald et al. 1996)	R09798
5	CAATCAGGTCCG	H4TF2 binding site: H4TF2 (La Bella and Heintz 1991)	R00681
6	AACAAACACAA	AC-box: H1TF1 (La Bella et al. 1989), HiNF-A (van Wijnen et al. 1988b), HiNF-D (van Wijnen et al. 1996; Grimes et al. 2003)	R00658
7	CAGCCAATCAGA	CCAAT-box: H1TF1 (La Bella et al. 1989), HiNF-B (van Wijnen et al. 1988a,b), NF-Y (Mantovani 1999), HiNF-D (van Wijnen et al. 1996; Grimes et al. 2003), H1TF2 (La Bella et al. 1989; Martinelli and Heintz 1994; Gallinari et al. 1989)	R00659, R00660
8	CCATTGGTTAAA	CCAAT-box: H1TF2 (La Bella et al. 1989; Martinelli and Heintz 1994; Gallinari et al. 1989), HiNF-B (van Wijnen et al. 1988a,b), NF-Y (Mantovani 1999), HiNF-D (van Wijnen et al. 1996; Grimes et al. 2003)	R00660
9	CCCCGCCCCCCG	GC-box: HiNF-C (van Wijnen et al. 1989), Sp1 (Courey and Tjian 1988), Sp3 (Bimbaum et al. 1995; Hagen et al. 1994)	R00684

Diagnosis of Childhood Acute Lymphoblastic Leukemia (ALL) and Optimization of Risk-Benefit Ratio of Therapy

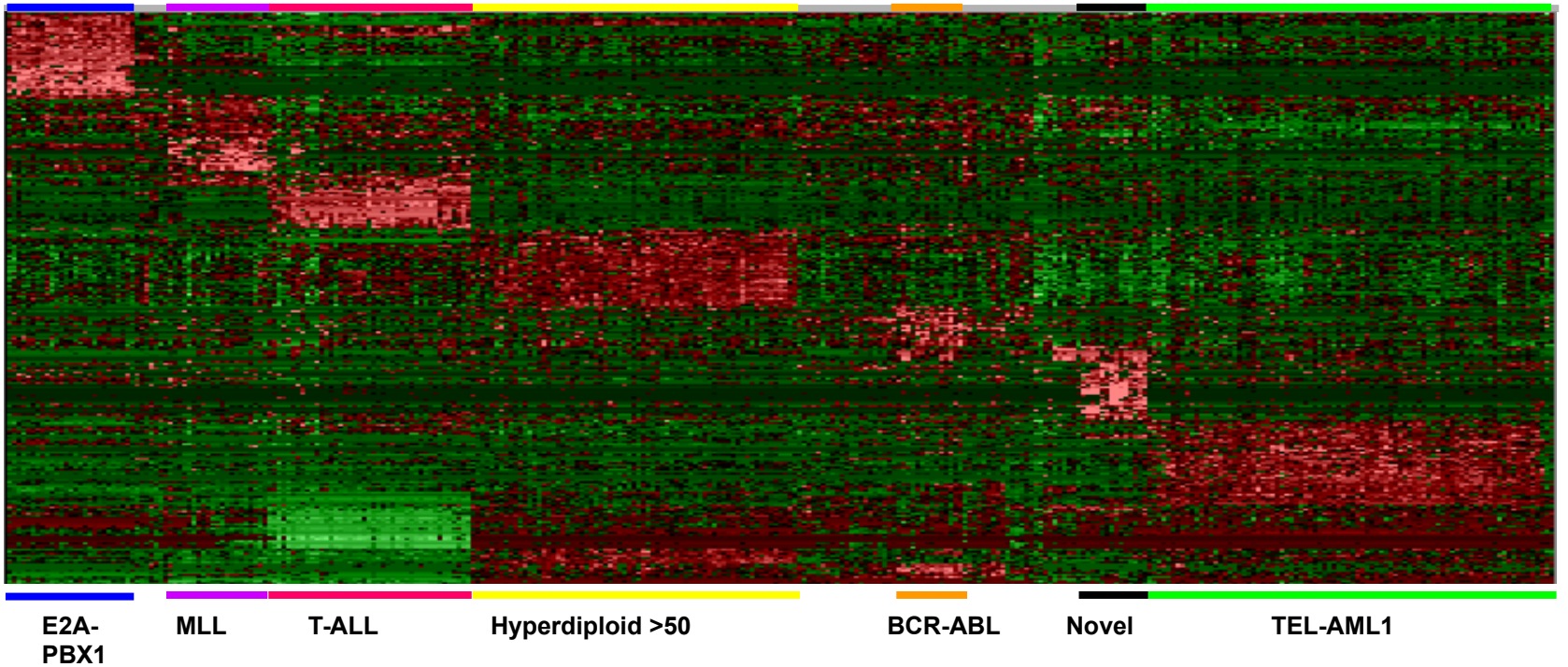


Immuno-phenotyping





Affymetrix GeneChip Micro Array Analysis



Proteomics Data : Guilt-by-Association

Compare *T* with seqs of known function in a db

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

```
>gi|134476732|ref|NP_108301.1|   unknown protein [Mesorhizobium loti]
gi|14027493|dbj|BAB53762.1|      unknown protein [Mesorhizobium loti]
          Length = 105

 Score =  105 bits (262), Expect = 1e-22
 Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
           MK G L  ++        MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between Amicyanin and u[illegible]wn M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

Poor Sequence Alignment

- Poor seq alignment shows few matched positions

⇒ The two proteins are not likely to be homologous

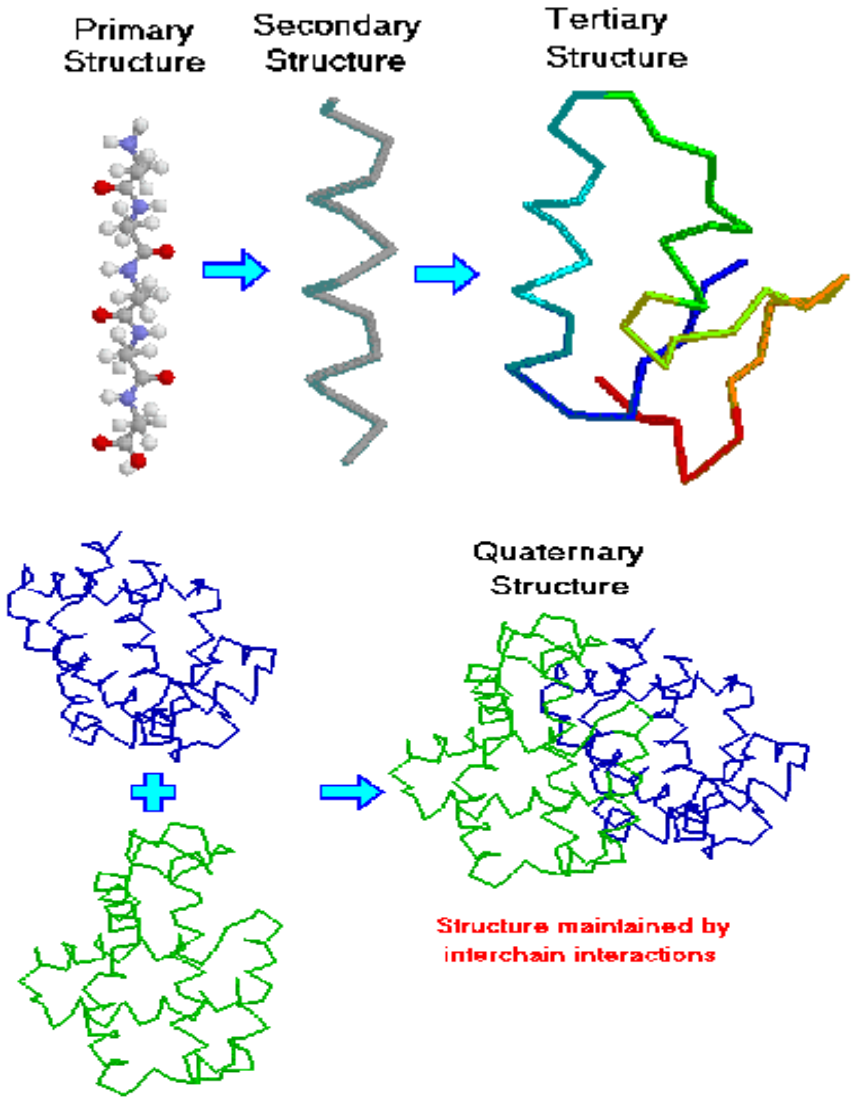
Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```
                      60        70        80        90        100
Amicyanin          MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVV
                                                :..:  . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYG
                     70        80        90        100       110
```

No obvious match between Amicyanin and Ascorbate Oxidase

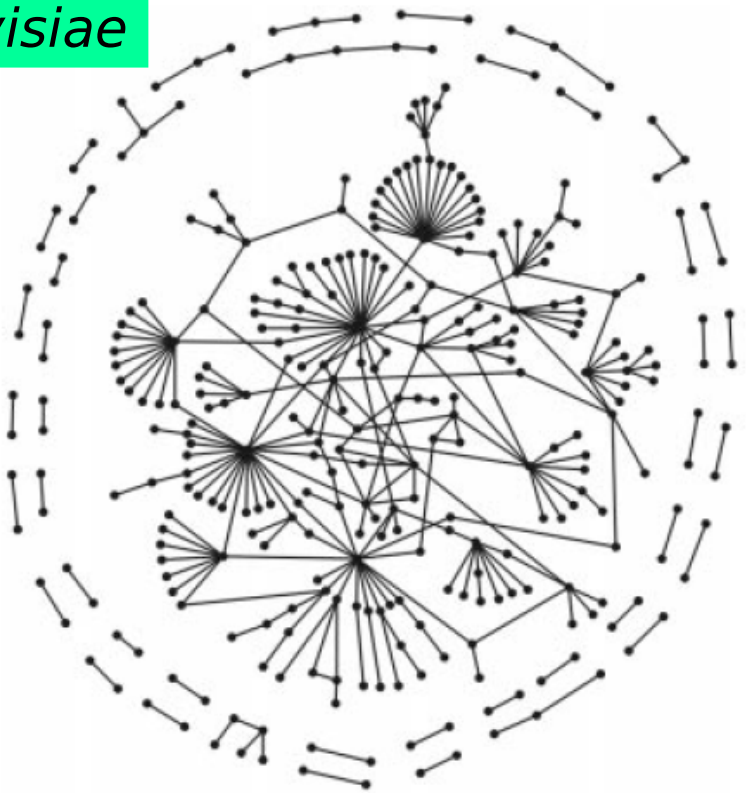
Discard this function as a candidate

Proteomics Data: Subgraphs in Protein Interaction

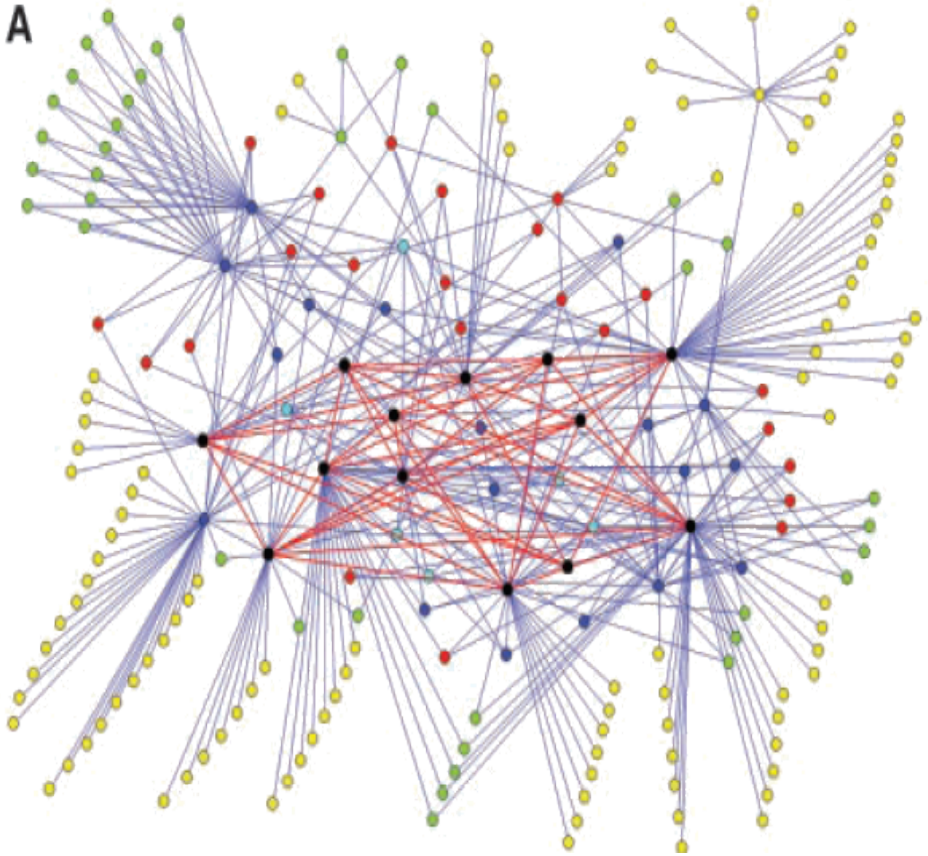


318 edges
329 nodes
In nucleus of
S. cerevisiae

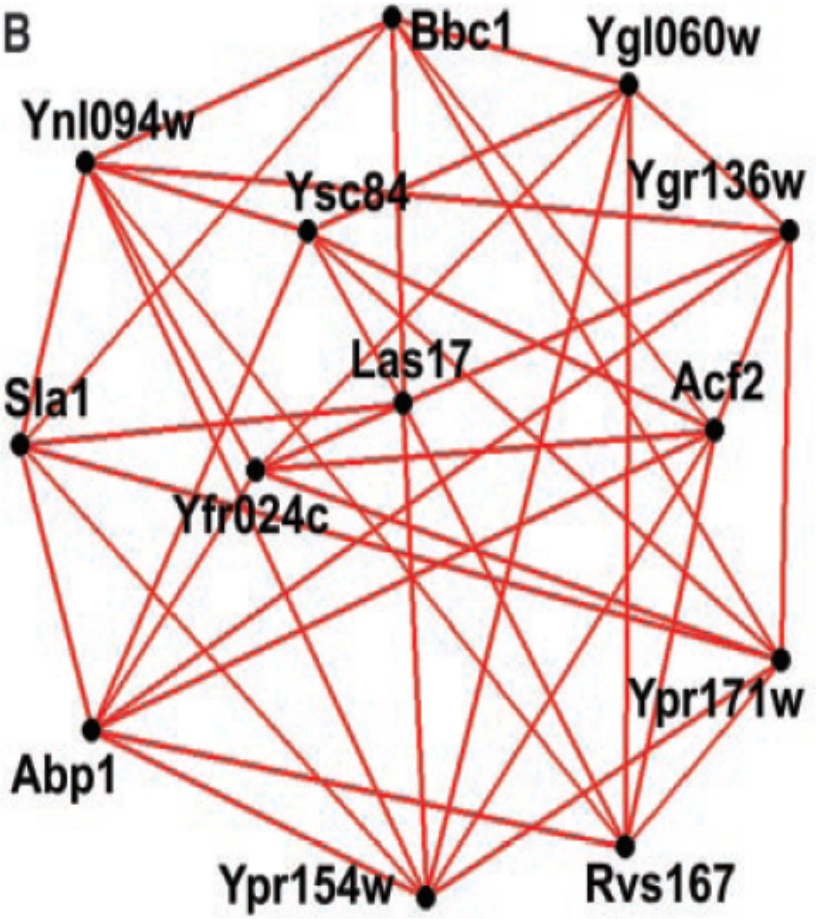
Maslov & Sneppen,
Science, v296, 2002



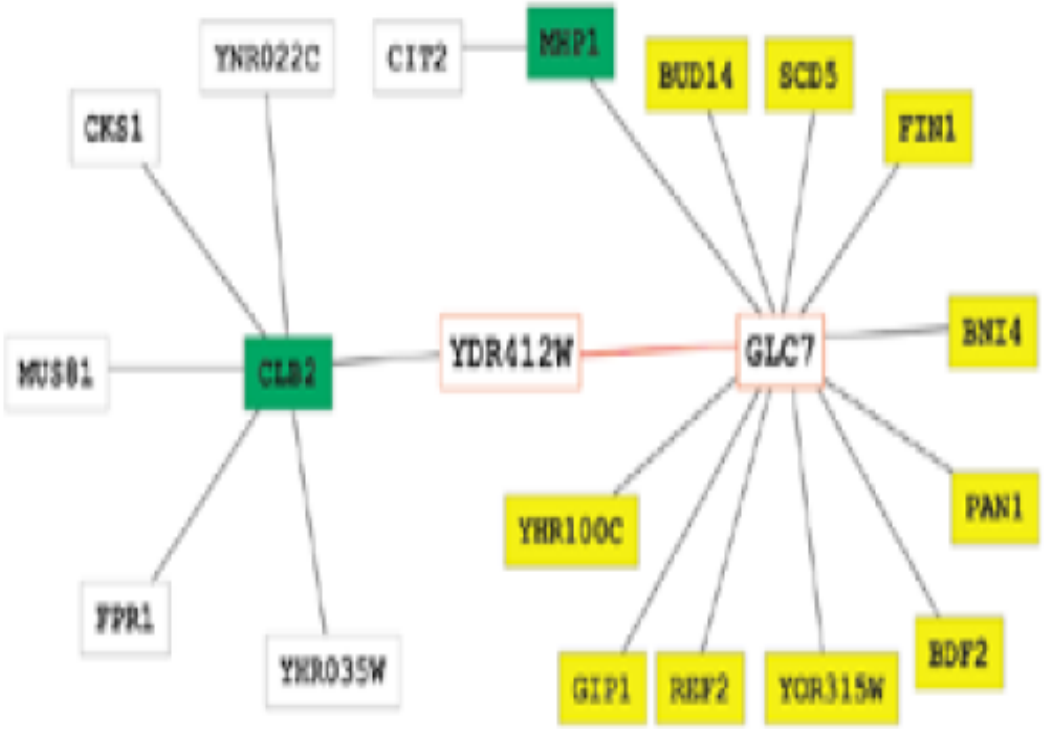
Topology of Protein Interaction Networks:
Hubs, Cores, Bipartites

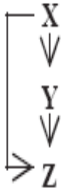


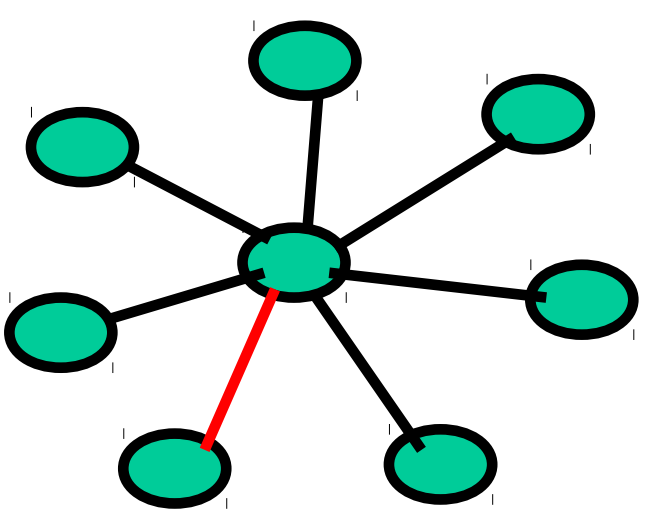
Yeast SH3 domain-domain
Interaction network:
394 edges, 206 nodes
Tong et al. *Science*, v295. 2002



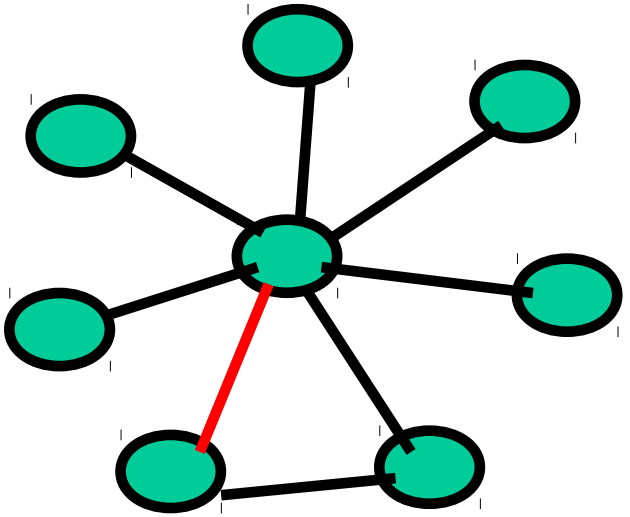
8 proteins containing SH3
5 binding at least 6 of them



Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					Feed-forward loop			Bi-fan
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41



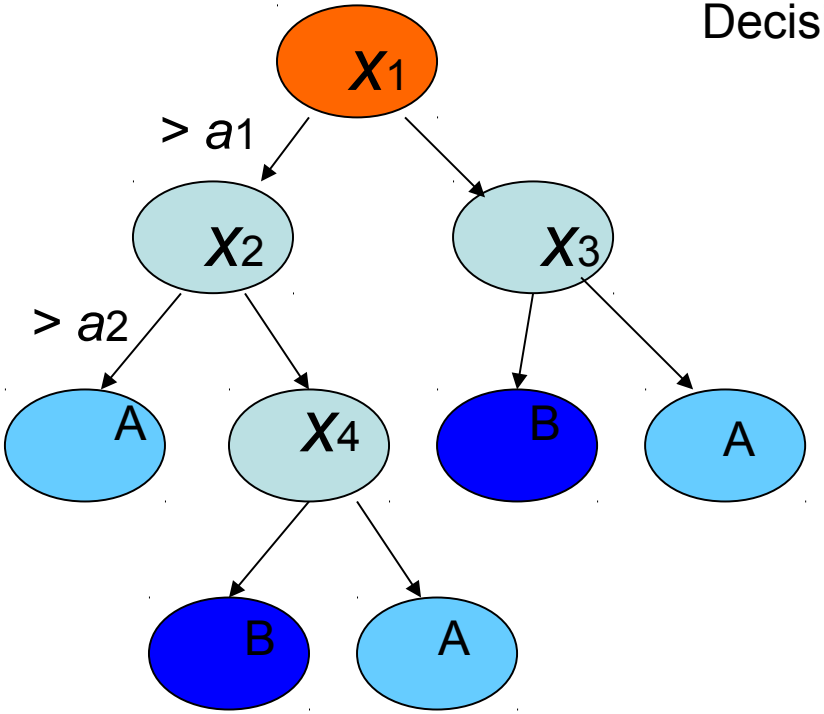
a



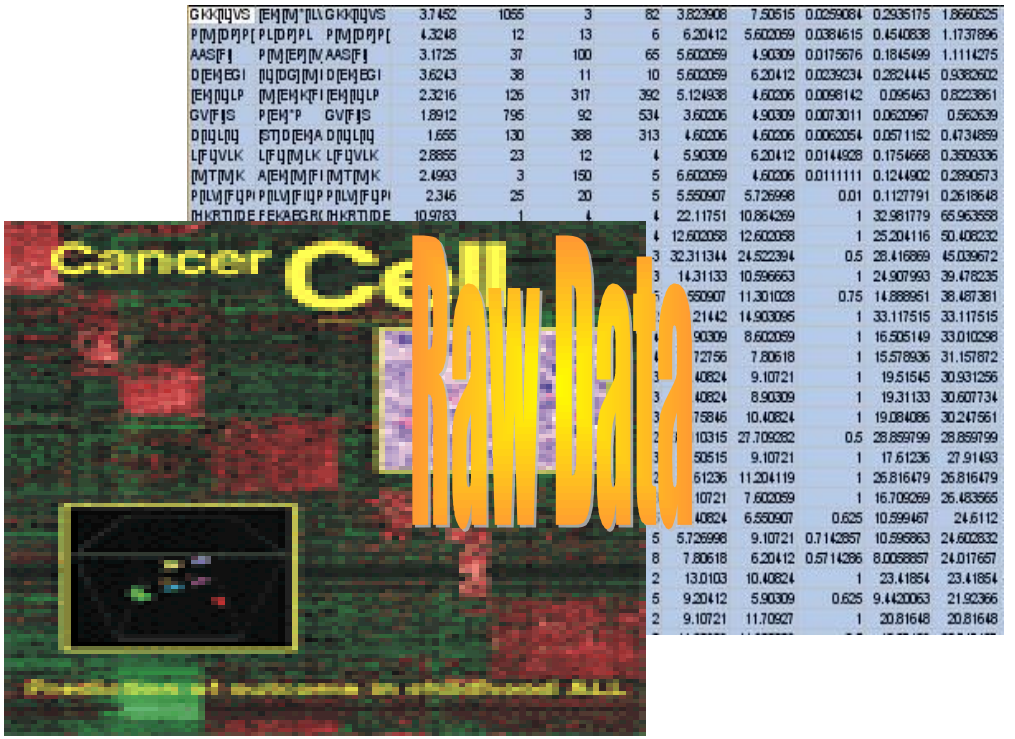
b

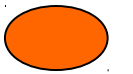
Configuration a is less likely than b in protein interaction networks → Graph/Network Mining

Prognosis based on Gene Expression Profiling

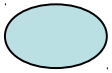


Decision Tree Based In-Silico Cancer Diagnosis

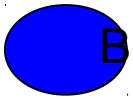




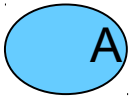
Root node

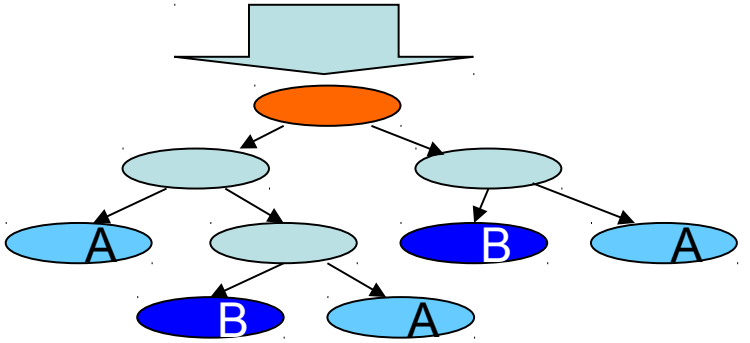


Internal nodes



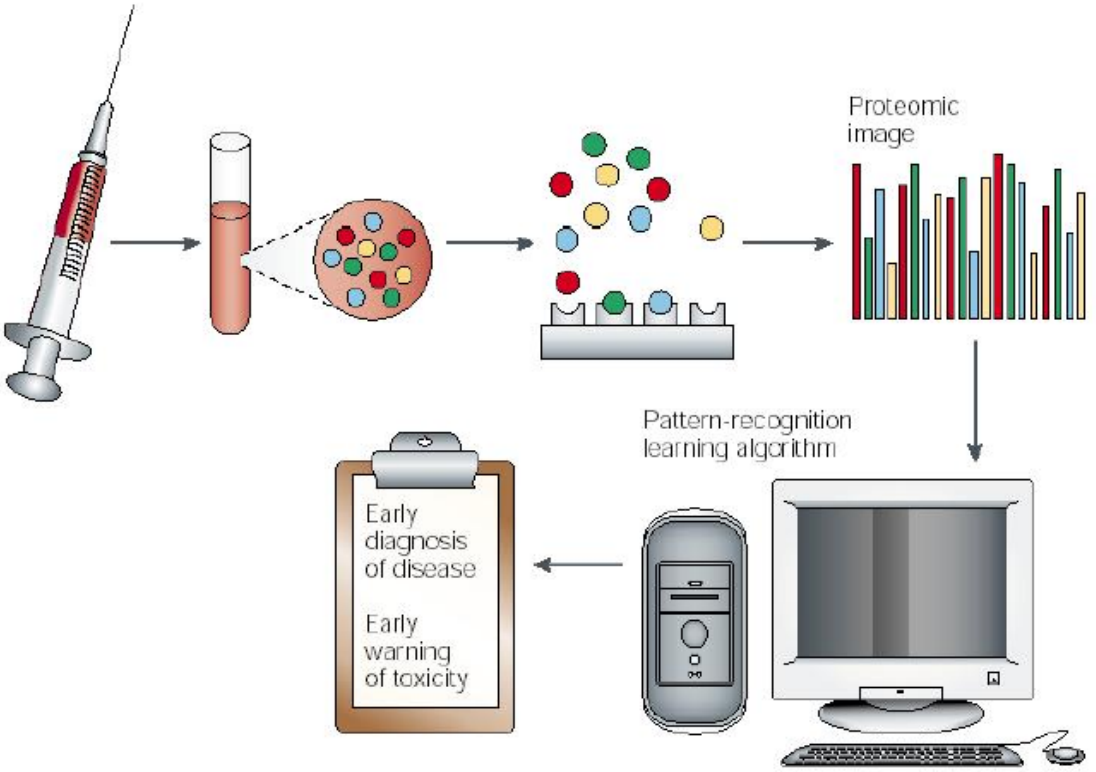
Leaf nodes





Discovery of Diagnostic Biomarkers for Ovarian Cancer

- Motivation: cure rate ~ 95% if correct diagnosis at early stage
- Proteomic profiling data obtained from patients' serum samples
- The first data set by Petricoin et al was published in *Lancet*, 2002
- Data set of June-2002.
- 253 samples: 91 controls and 162 patients suffering from the disease; 15154 features (proteins, peptides, precisely, mass/charge identities)



Methods	CS4	C4.5		
		Single	Bagging	Boosting
Errors	0 (0:0)	10 (4:6)	7 (3:4)	10 (4:6)

SVM: 0 errors; Naïve Bayes: 19 errors; *k*-NN: 15 errors.

Mining Errors from Bio Databases

- ATTRIBUTE
  - Invalid values
    - Uninformative sequences
    - **Undersized sequences**
  - Ambiguity
  - Dubious sequences
  - Vector contaminated sequence
- RECORD
  - Cross-annotation error
  - Annotation error
  - Sequence structure violation
- SINGLE SOURCE DATABASE
  - Sequence redundancy
  - Data provenance flaws
- MULTIPLE SOURCE DATABASE
  - Erroneous data transformation
  - Incompatible schema

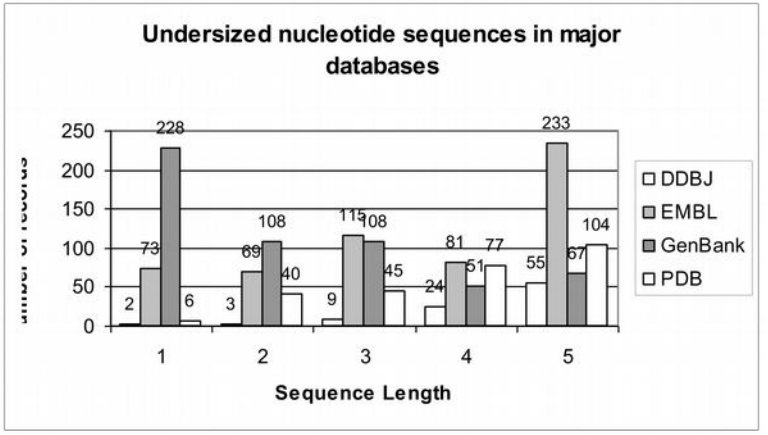
Example Meaningless Seqs

• Among the 5,146,255 protein records queried using Entrez to the major protein or translated nucleotide databases , 3,327 protein sequences are shorter than four residues (as of Sep, 2004).

• In Nov 2004, the total number of undersized protein sequences increases to 3,350.

• Among 43,026,887 nucleotide records queried using Entrez to major nucleotide databases, 1,448 records contain sequences shorter than six bases (as of Sep, 2004).

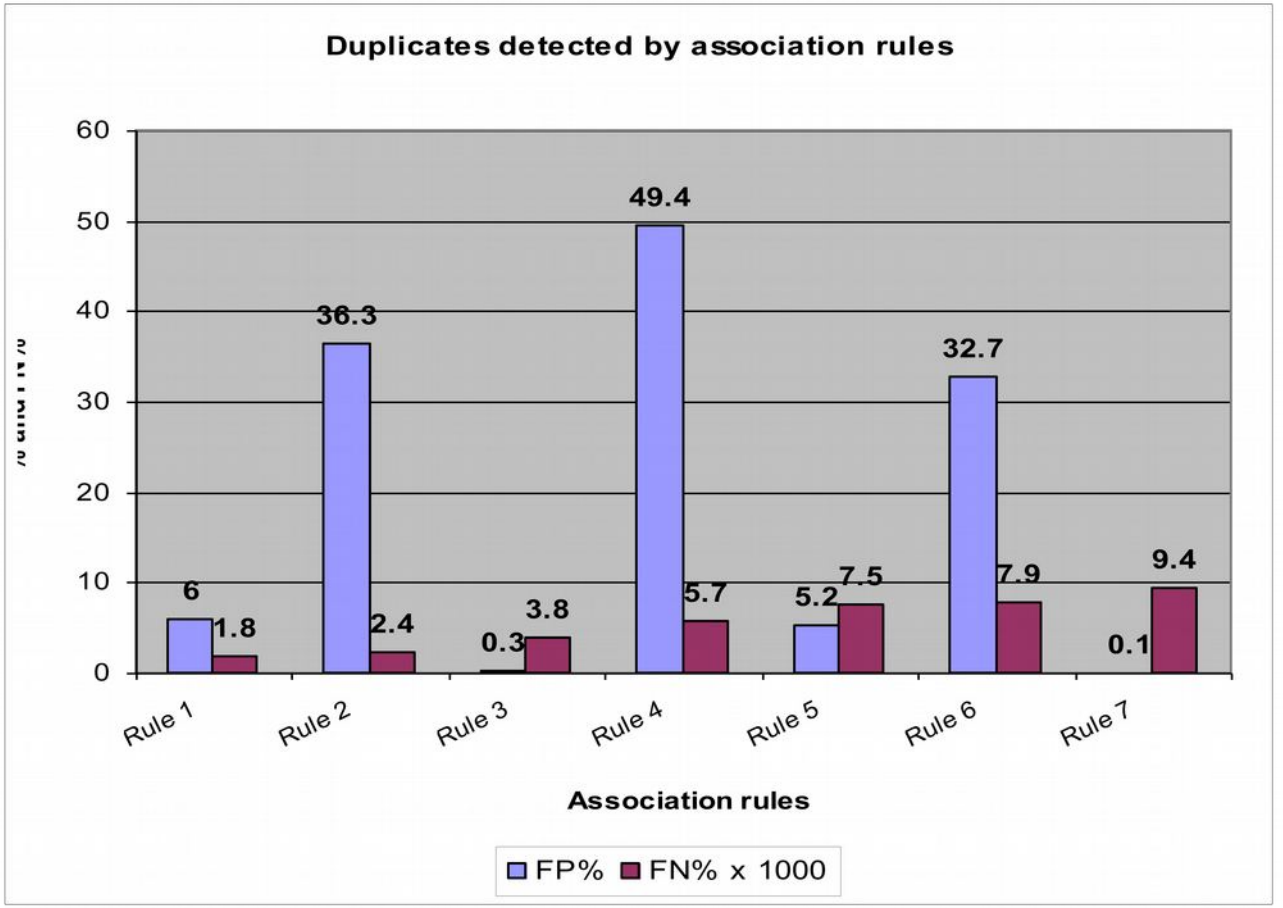
• In Nov 2004, the total number of undersized nucleotide sequences increases to 1,711.



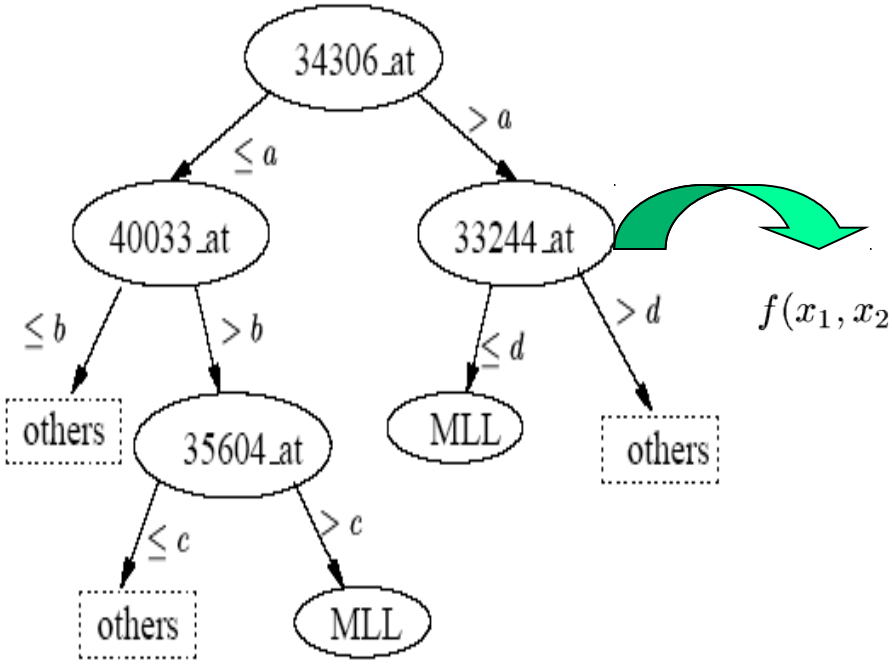
Rule 1. Identical sequences with the same sequence length and not originated from PDB are 99.7% likely to be duplicates.

Rule 2. Identical sequences with the same sequence length and of the same species are 97.1% likely to be duplicates.

Rule 3. Identical sequences with the same sequence length, of the same species and not originated from PDB are 96.8% likely to be duplicates.



Rule 1	S(Seq)=1 ^ N(Seq Length)=1 ^ M(PDB)=0 (99.7%)
Rule 2	S(Seq)=1 ^ M(PDB)=0 ^ M(Species)=1 (97.1%)
Rule 3	S(Seq)=1 ^ N(Seq Length)=1 ^ M(Species)=1 ^ M(PDB)=0 (96.8%)
Rule 4	S(Seq)=1^ M(PDB)=0 ^ M(DB)=0 (93.1%)
Rule 5	S(Seq)=1 ^ M(Seq Length)=1 ^ M(PDB)=0 ^ M(DB)=0 (92.8%)
Rule 6	S(Seq)=1 ^ M(Species)=1 ^ M(PDB)=0 ^ M(DB)=0 (90.4%)
Rule 7	S(Seq)=1 ^ N(Seq Length)=1 ^ M(Species)=1 ^ M(PDB)=0 ^ M(DB)=0 (90.1%)

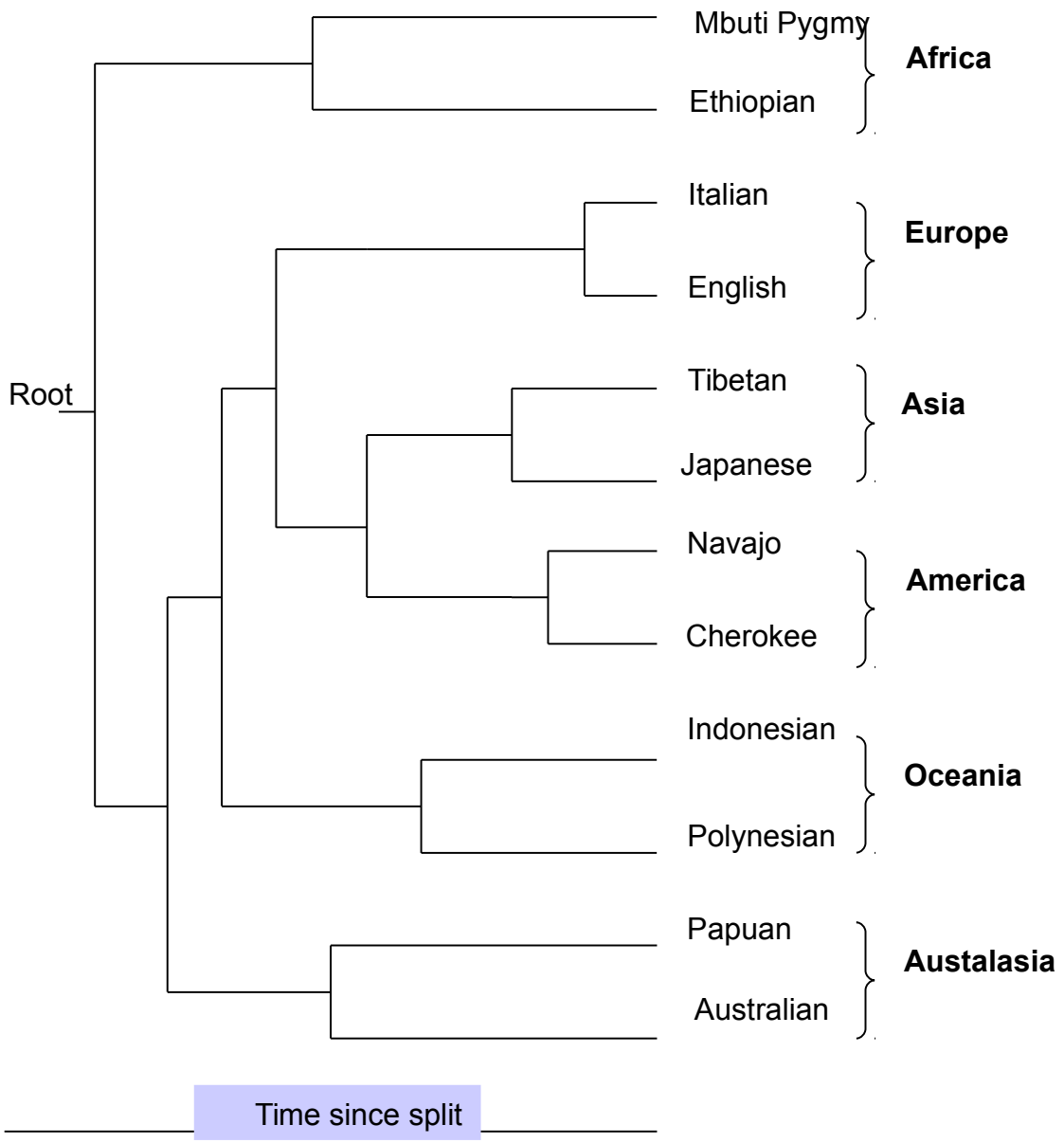


$$f(x_1, x_2, x_3, x_4) = \begin{cases} -1 & \text{if } x_1 \le a,\ x_2 \le b \\ -1 & \text{if } x_1 \le a,\ x_2 > b,\ x_3 \le c \\ 1 & \text{if } x_1 \le a,\ x_2 > b,\ x_3 > c \\ 1 & \text{if } x_1 > a,\ x_4 \le d \\ -1 & \text{if } x_1 > a,\ x_4 > d \end{cases}$$

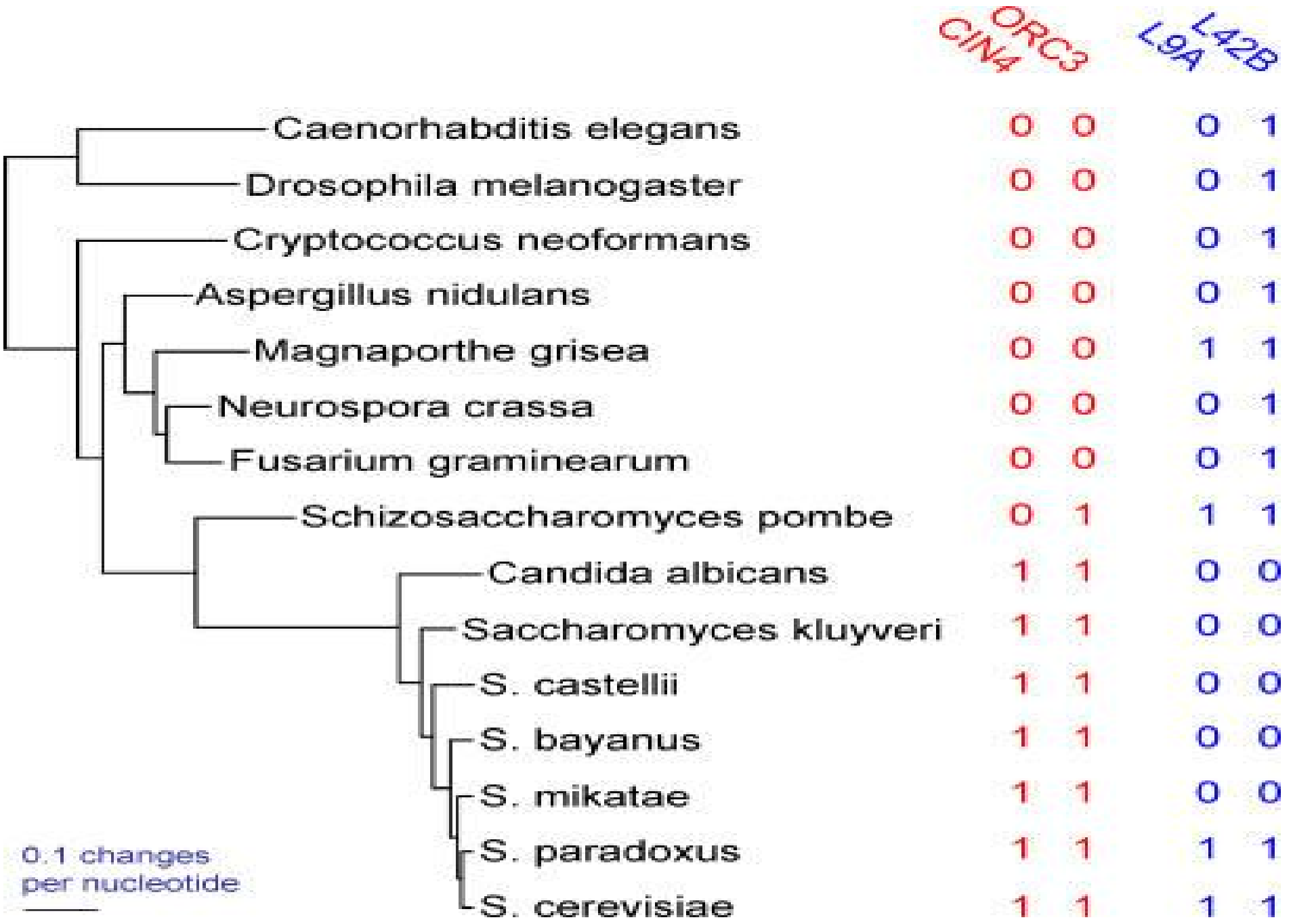
Given a test sample, at most 3 of the 4 genes' expression values are needed to make a decision!

- Yeoh et al., *Cancer Cell* 1:133-143, 2002;
  Differentiating MLL subtype from other subtypes of childhood leukemia
- Training data (14 MLL vs 201 others), Test data (6 MLL vs 106 others), Number of features: 12558

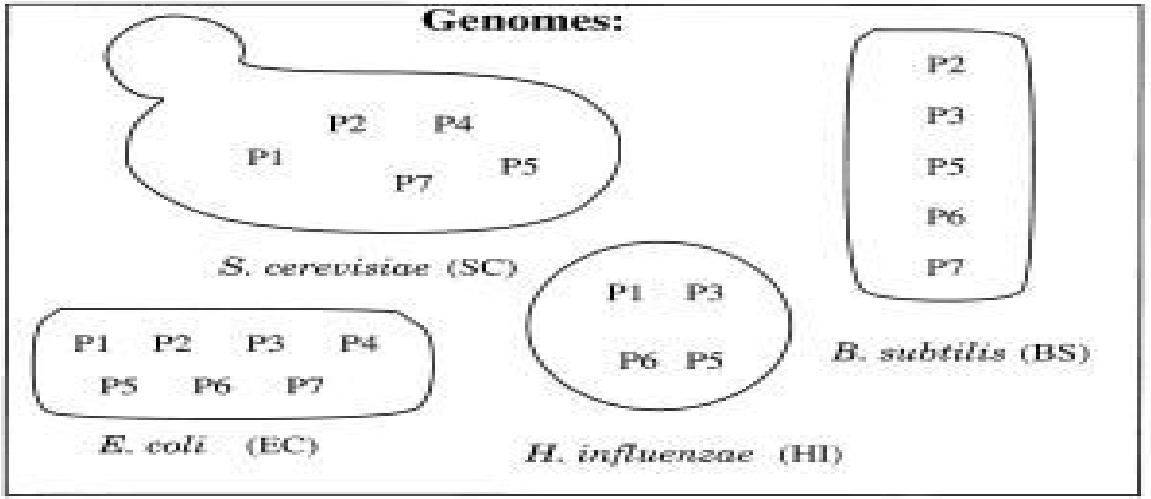
Phylogenetic tree construction



- Estimate order in which "populations" evolved
- Based on assimilated freq of many different genes
- But …
  - is human evolution a succession of population fissions?
  - Is there such thing as a proto-Anglo-Italian population which split, never to meet again, and became inhabitants of England and Italy?

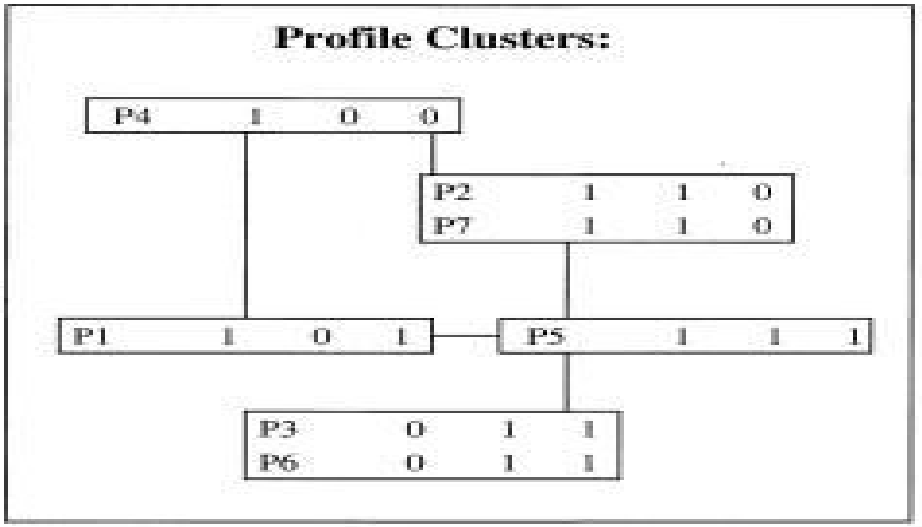

ORC3
CIN4
L42B
L9A
Caenorhabditis elegans 0 0 0 1
Drosophila melanogaster 0 0 0 1
Cryptococcus neoformans 0 0 0 1
Aspergillus nidulans 0 0 0 1
Magnaporthe grisea 0 0 1 1
Neurospora crassa 0 0 0 1
Fusarium graminearum 0 0 0 1
Schizosaccharomyces pombe 0 1 1 1
Candida albicans 1 1 0 0
Saccharomyces kluyveri 1 1 0 0
S. castellii 1 1 0 0
S. bayanus 1 1 0 0
S. mikatae 1 1 0 0
S. paradoxus 1 1 1 1
S. cerevisiae 1 1 1 1
0.1 changes
per nucleotide

Predicting interactions using phylogenetic profile



Phylogenetic Profile:

EC	SC	BS	HI
P1	1	0	1
P2	1	1	0
P3	0	1	1
P4	1	0	0
P5	1	1	1
P6	0	1	1
P7	1	1	0



Conclusion: P2 and P7 are functionally linked ,
P3 and P6 are functionally linked

*Pellegrini et al. PNAS **96**, 4285-4288 (1999)*

Comparative genomics