Data Mining in Bioinformatics

Prof. André de Carvalho
ICMC-Universidade de São Paulo

Main topics

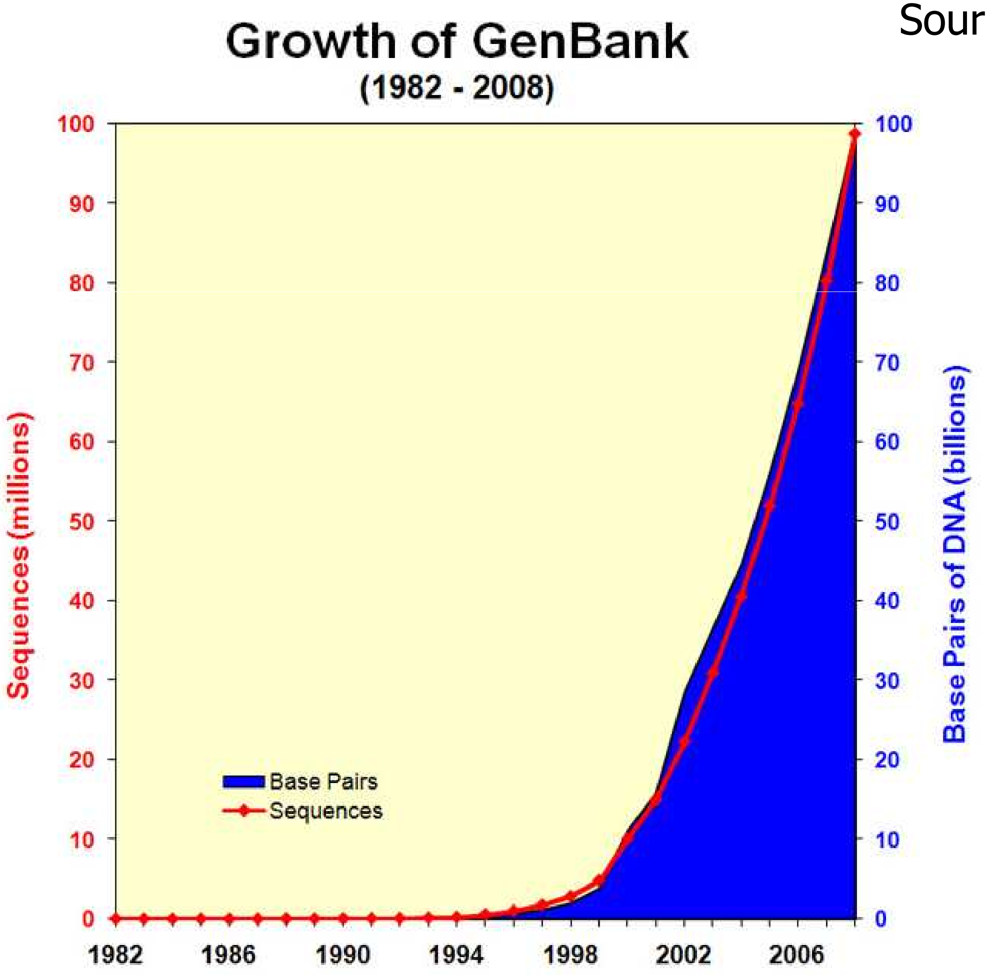
- Motivation
- Data Mining
  - Prediction
- Bioinformatics
  - Molecular Biology
  - Using DM in Molecular Biology
  - Case studies
    - Gene Expression Analysis
    - Protein function prediction

Motivation

- Genome research is producing a very large amount of data
- Exponential growth in the number of stored bp in the last 10 years
    - In the beginning of the decade, doubling every 12-15 months

Motivation

Source: http://www.ncbi.nlm.nih.gov/



GenBank growth

Spring 2009

Motivation

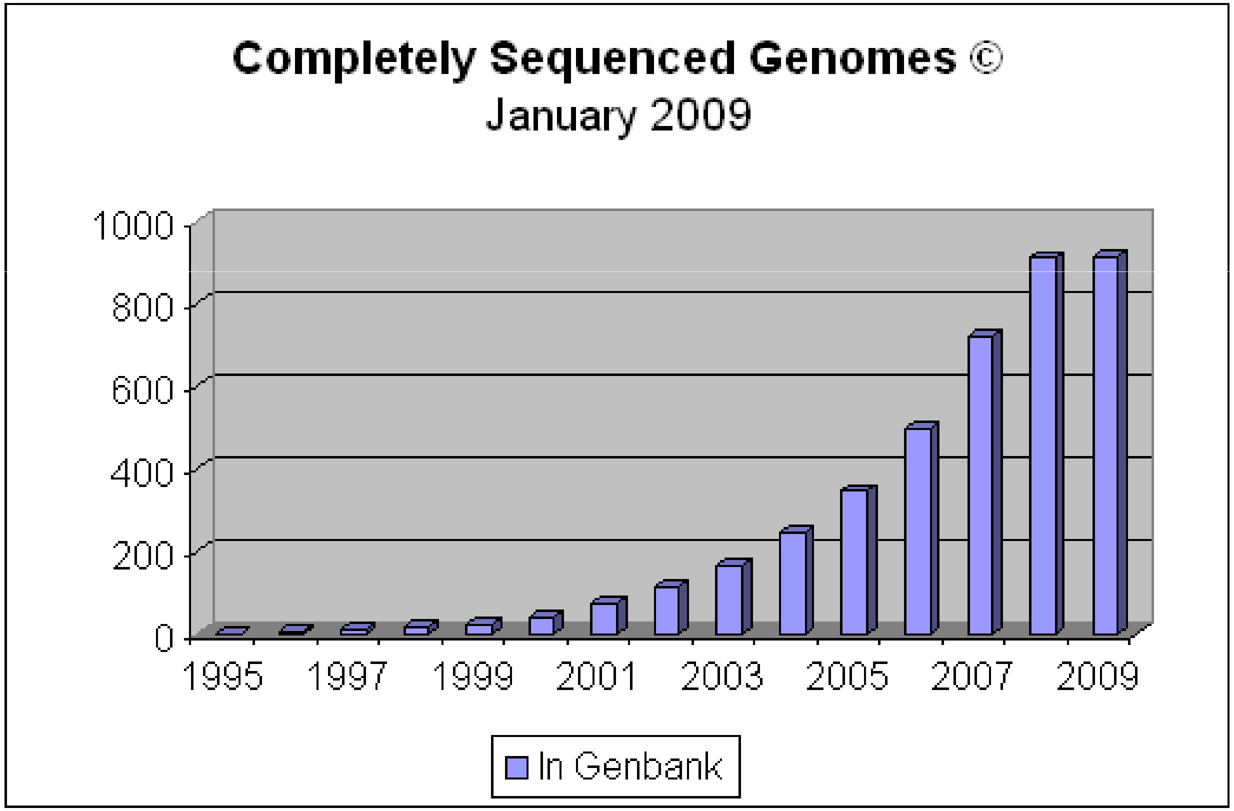
Source: http://www.ncbi.nlm.nih.gov/

Year	Base Pairs	Sequences
2000	11,101,066,288	10,106,023
2001	15,849,921,438	14,976,310
2002	28,507,990,166	22,318,883
2003	36,553,368,485	30,968,418
2004	44,575,745,176	40,604,319
2005	56,037,734,462	52,016,762
2006	69,019,290,705	64,893,747
2007	83,874,179,730	80,388,382
2008	99,116,431,942	98,868,465

Bioinformatics

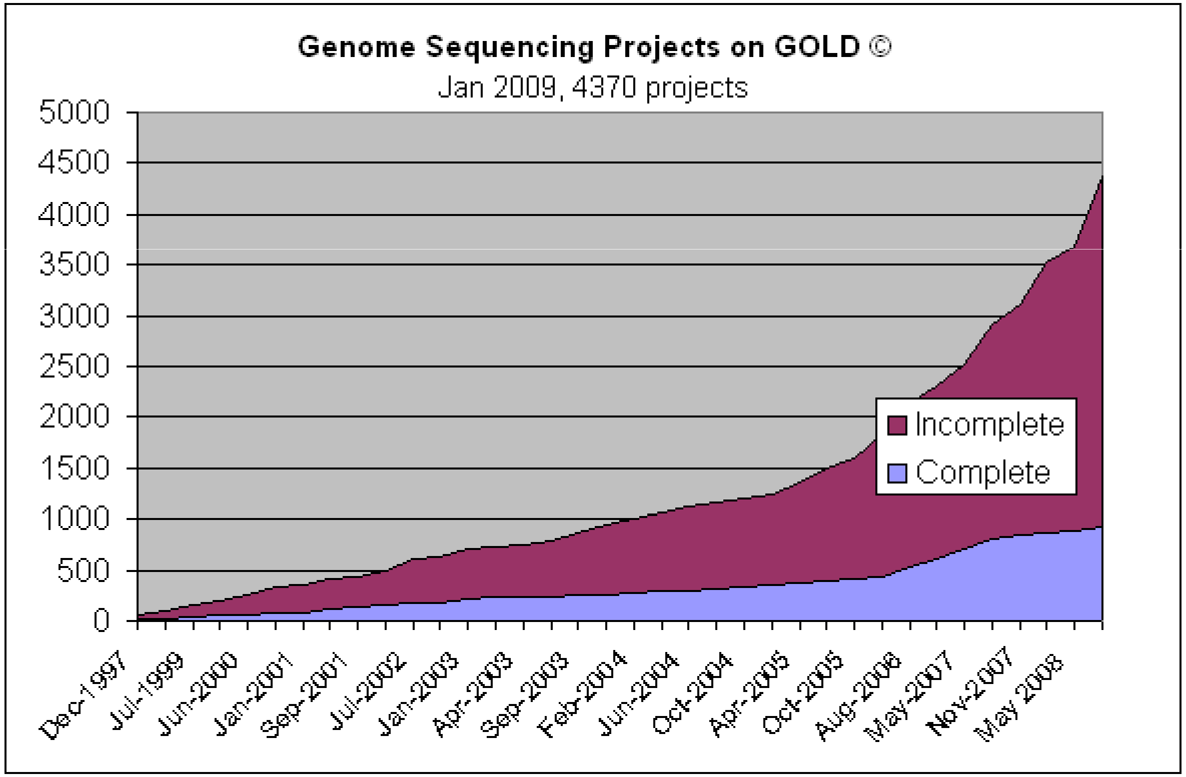
- Several sequencing projects have been concluded lately
  - Producing a large amount of data
  - Until 2009:
    - 4370 projects
      - Almost 1000 completed

Bioinformatics



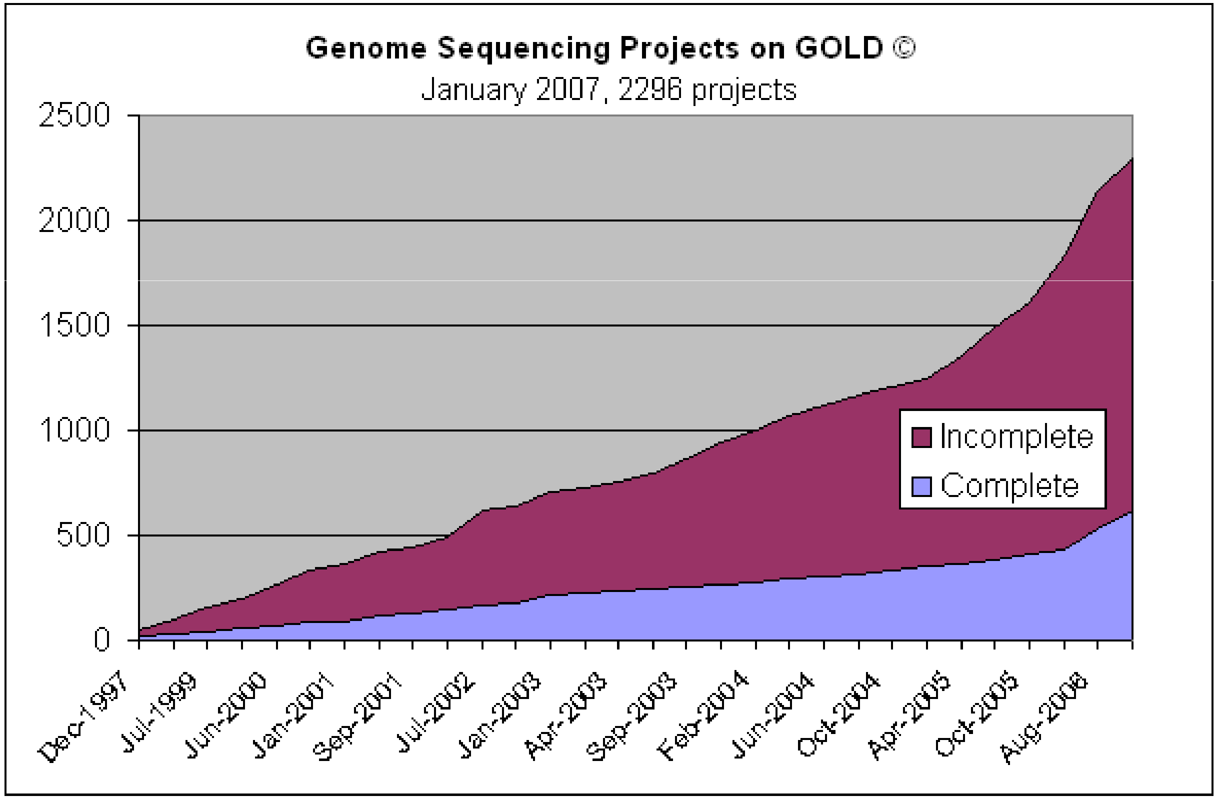
Fonte: http://www.genomesonline.org/gold_statistics.htm

Bioinformatics (2009)



Fonte: http://www.genomesonline.org/gold_statistics.htm

Bioinformatics (2007)

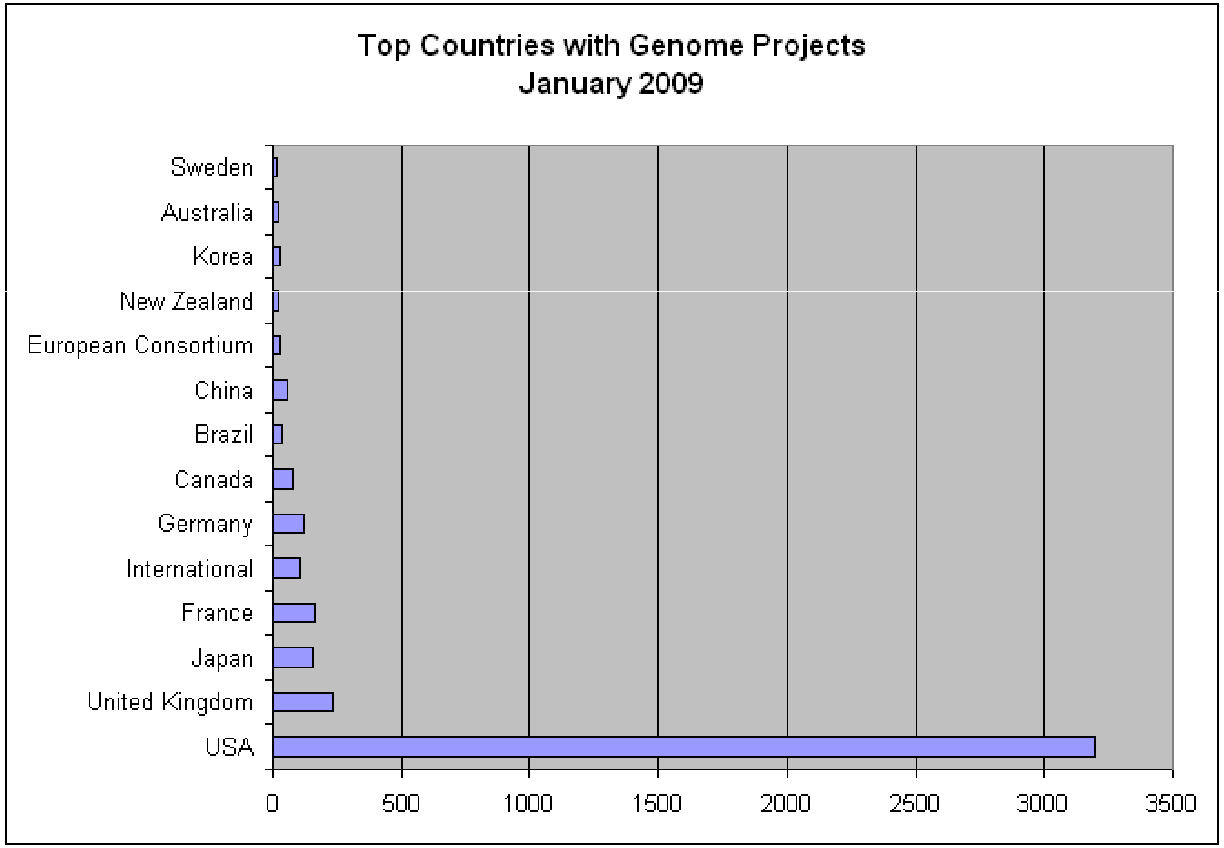


Fonte: http://www.genomesonline.org/gold_statistics.htm

Bioinformatics

- Genome projects
  - Complete genomes published (eukaryote)
    - Human
    - Mouse
    - Drosophila
    - Arabidopsis thaliana
  - Domestic animals
    - Bovine

Bioinformatics



Fonte: http://www.genomesonline.org/gold_statistics.htm

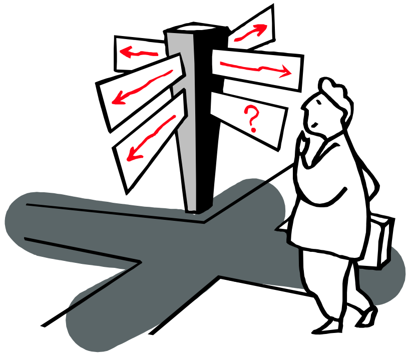
Motivation

- Emphasis is progressively moving from data accumulation to data interpretation
  - Data resulting from sequencing projects
  - These data needs to be analysed
  - Analysis in Laboratories is difficult and expensive
    - Sophisticated computational tools are needed
    - Data mining

DM and Machine Learning

- Most DM methods are based on Machine Learning (ML) techniques
  - Decision Trees
  - Regression
  - Clustering
  - Association rules
  - Artificial Neural Networks
  - Support Vector Machines
  - Evolutionary Computation
  - Hybrid Intelligent Systems

Bioinformatics



- Definition
  - Research and development of computational tools able to solve problems from Biology
    - Molecular biology

Computers are to biology what mathematics is to physics

Harold Morowitz

Bioinformatics

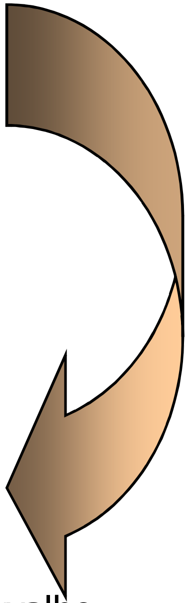
- Several areas may benefit
  - Medicine - Pharmacy - Agriculture
- Molecular Medicine
  - Improve diagnosis of diseases
  - Detect genetic predisposition to pathologies
  - Create drugs based on molecular information
  - Use gene therapy as drugs
  - Design “custom drugs” based on individual genetic profiles

Molecular biology

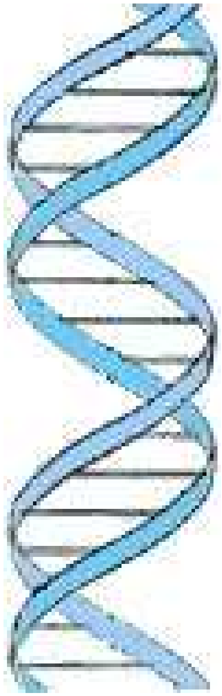
- Study of cells and molecules
  - *Particularly* genome of organisms
- Main structures investigated:
  - Genes
  - Chromosomes
  - DNA
  - RNA

  DNA, RNA → *nucleotides*

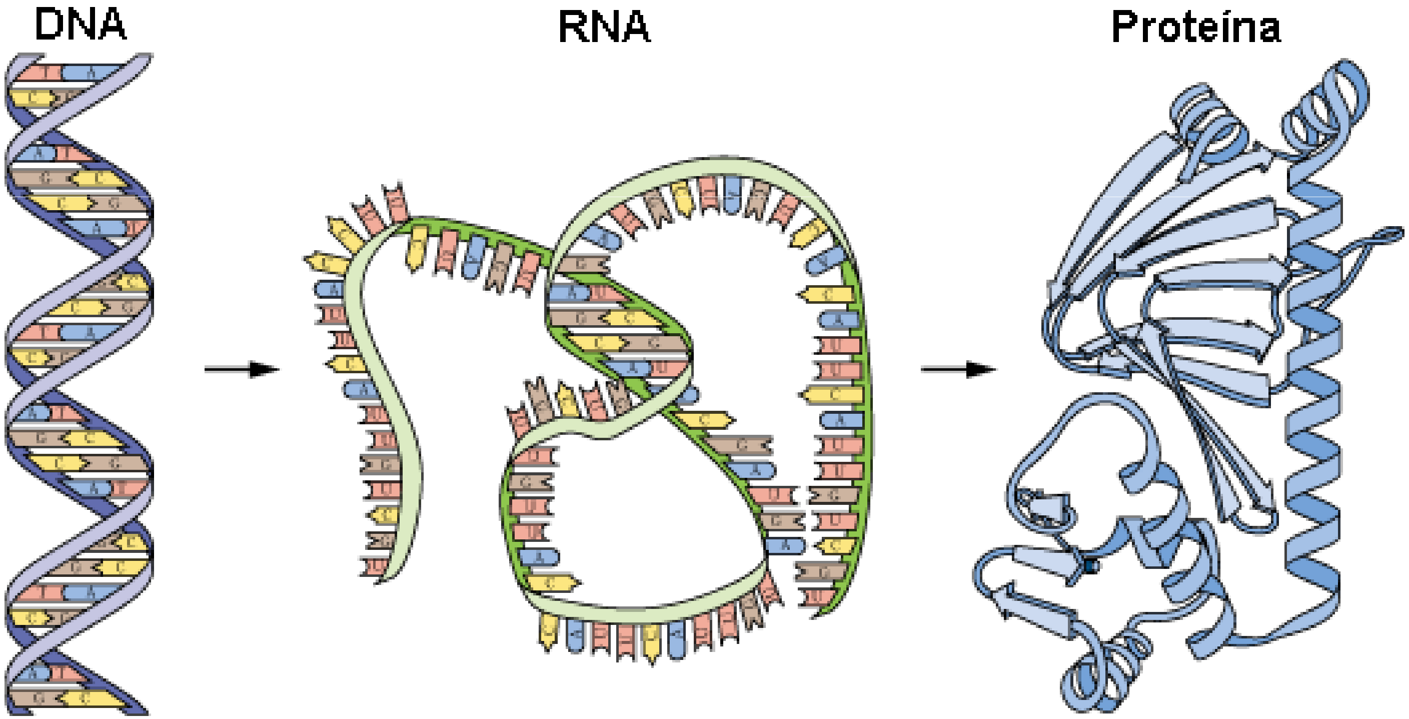
  - Proteins → *Amino acids*



Gene expression



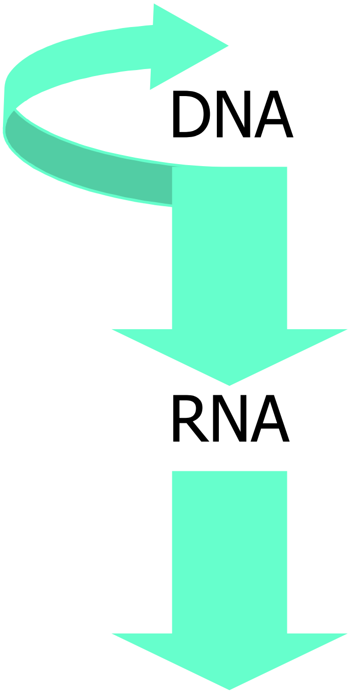
Molecular biology



Molecular biology

- Central Dogma of Molecular Biology
  - Information transference

They are only assumptions



Replication

DNA

Transcription

RNA

Translation

Proteins

Molecular biology

- Recent discoveries contradict this dogma:
  - RNA can suffer replication in some virus and plants
  - Viral RNA, through an enzyme named reverse transcriptase, can be transcribed in DNA
  - DNA can directly produce specific proteins
    - Without going through the transcription process

Molecular biology

- A genome is all the DNA in an organism, including its genes
  - Genes carry information for making all the proteins required by all organisms
  - These proteins determine, among other things:
    - How an organism looks like, how well its body metabolizes food or fights infection, and sometimes even how it behaves

Molecular biology

- DNA (Deoxyribonucleic Acid)
  - Molecule made up of two parallel twisted chains of alternating units of phosphoric acid and deoxyribose sugars
    - Combination of four types of bases
      - A (adenine), C (cytosine), G (guanine) and T (thymine)
    - Chains are held together by links that connect each nucleotide in one chain to its complement in the other chain
      - A connects with T and C with G
      - Gives the double helix appearance

Molecular biology

- RNA (Ribonucleic Acid)
  - Differ from DNA in several aspects:
    - Single stranded molecule
    - Contains ribose sugars
      - Instead of deoxyribose
    - Instead of T (thymine), contains U (uracyl)
    - RNA molecules are smaller than DNA molecules

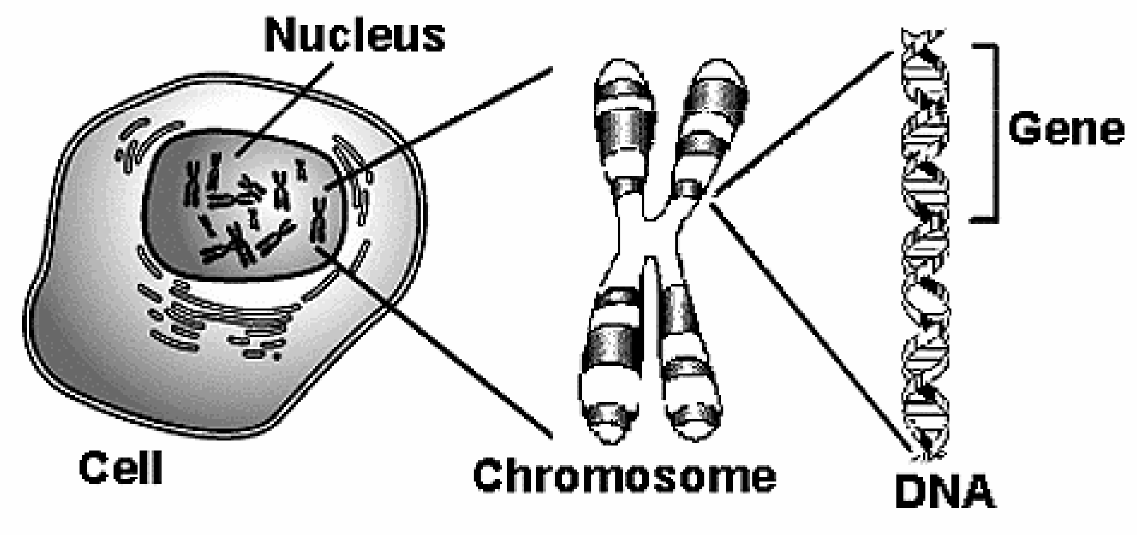
Molecular biology

- Genes
  - Subsequences of DNA
    - Localized in chromosomes
  - Used as mould for the production of proteins
  - There are segments incased between genes named no coding regions

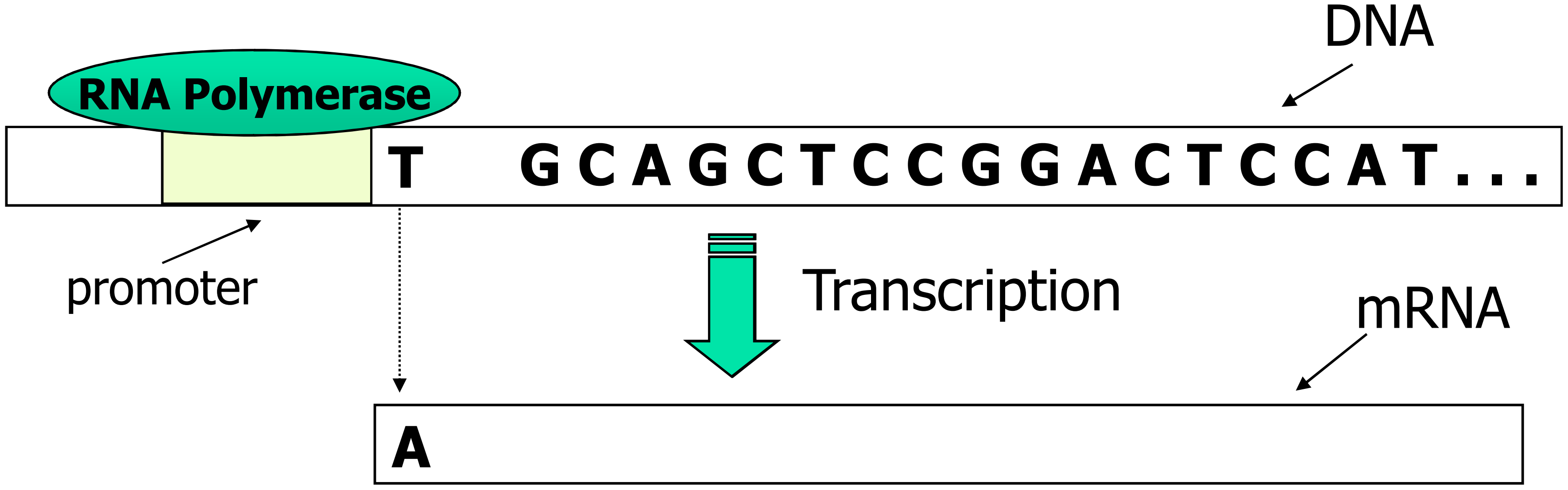
Molecular biology

- Proteins
  - Define structure, function and regulatory mechanisms in the cells
    - Examples of regulatory mechanisms:
      - Cell cycle control, genetic transcription
    - Can be represented by linear sequences
    - Combination of 20 different amino acids
    - Three nucleotides (codon) are mapped to an amino acid

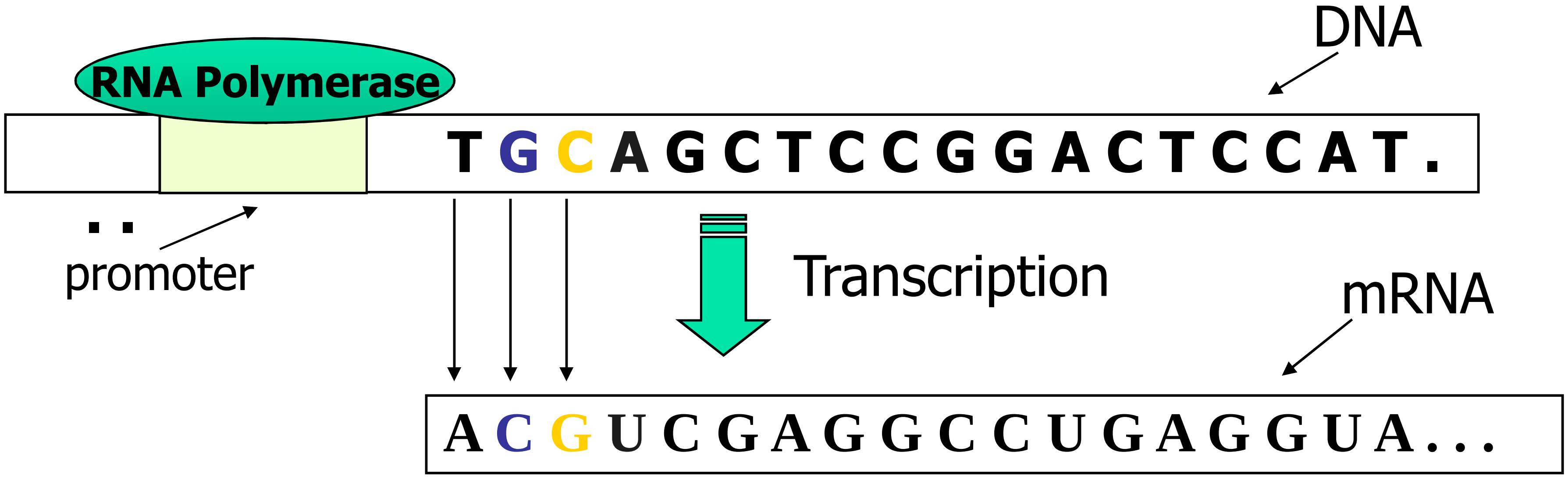
Molecular biology



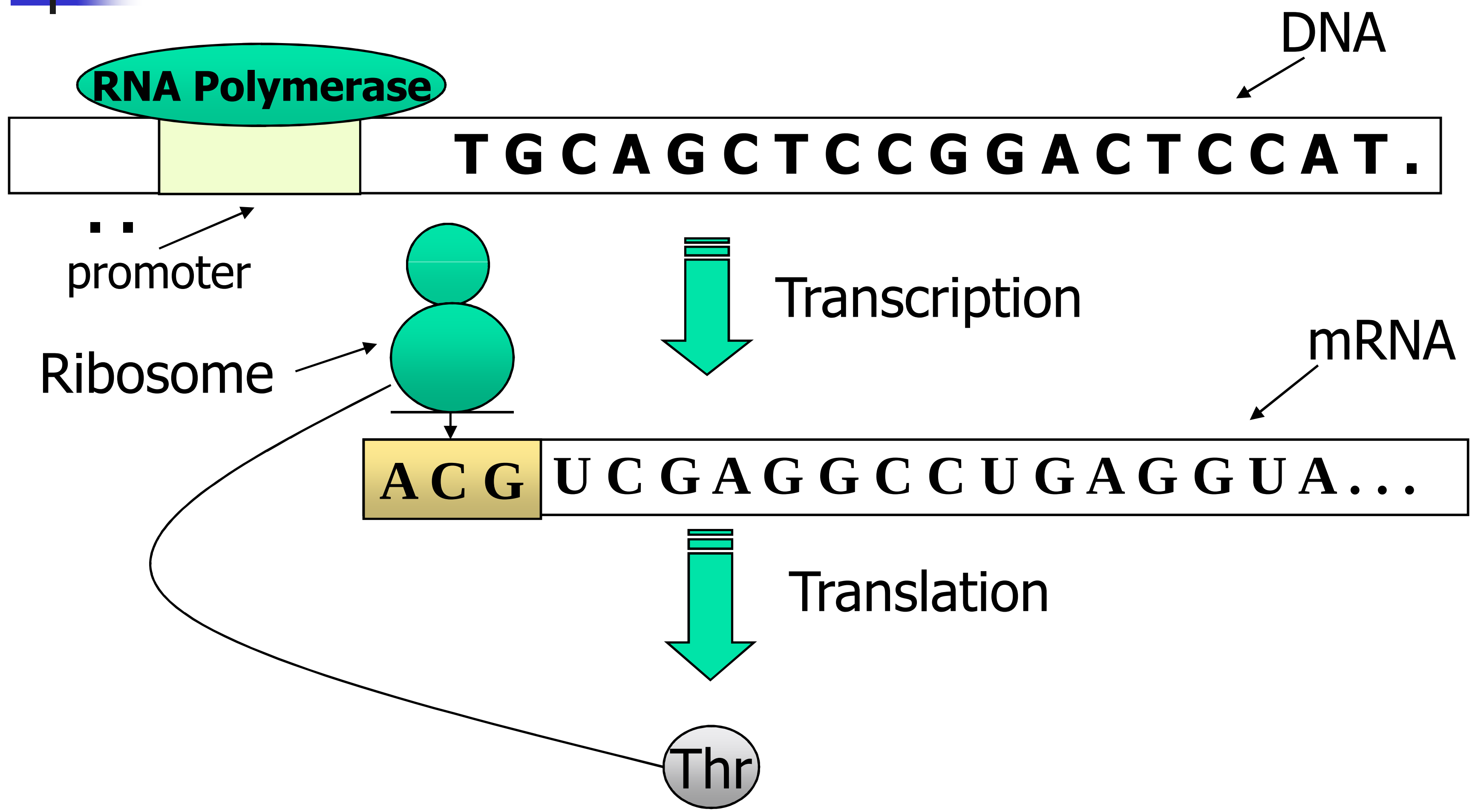
Gene expression process



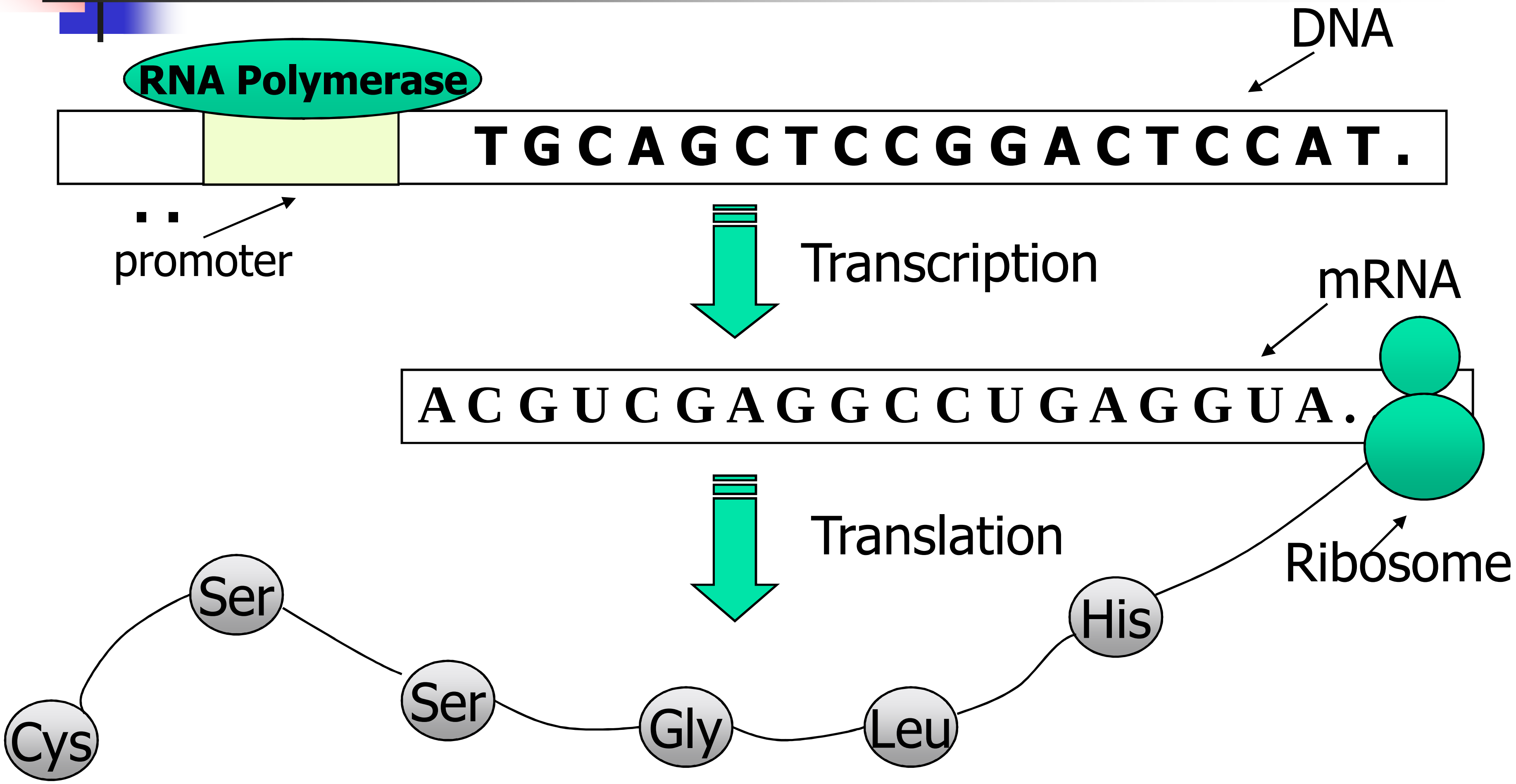
Gene expression process


DNA
RNA Polymerase
T G C A G C T C C G G A C T C C A T .
. .
promoter
Transcription
mRNA
A C G U C G A G G C C U G A G G U A . . .

Gene expression process


DNA
RNA Polymerase
T G C A G C T C C G G A C T C C A T .
. .
promoter
Transcription
Ribosome
mRNA
A C G
U C G A G G C C U G A G G U A . . .
Translation
Thr

Gene expression process



DM and molecular biology

- Problems in Molecular Biology where ML techniques have been used
  - Gene recognition
  - Reconstruction of phylogenies
  - **Gene expression analysis**
  - Protein structure prediction
  - **Protein function prediction**
  - Gene regulation analysis
  - Sequences alignment

Gene expression analysis

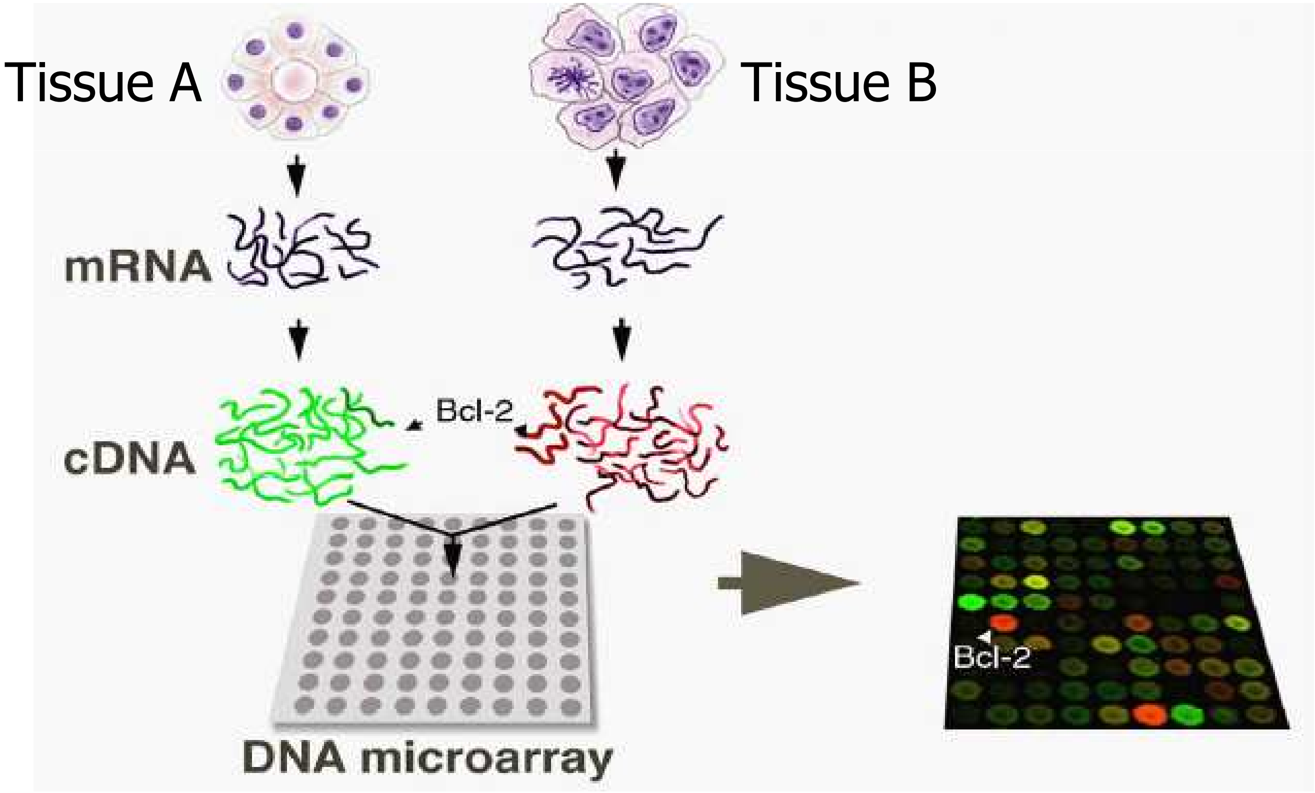
- Concerned with the identification of the function of genes
- Main goals:
  - Reveal patterns in genetic datasets
    - Looks for Patterns of similarity and dissimilarity
    - Analyze expression levels of thousands of genes collected from different tissues

Gene expression analysis

- Several techniques are used to detect gene expression in a tissue
  - Microarrays
  - Sage
  - PCR
  - MPSS

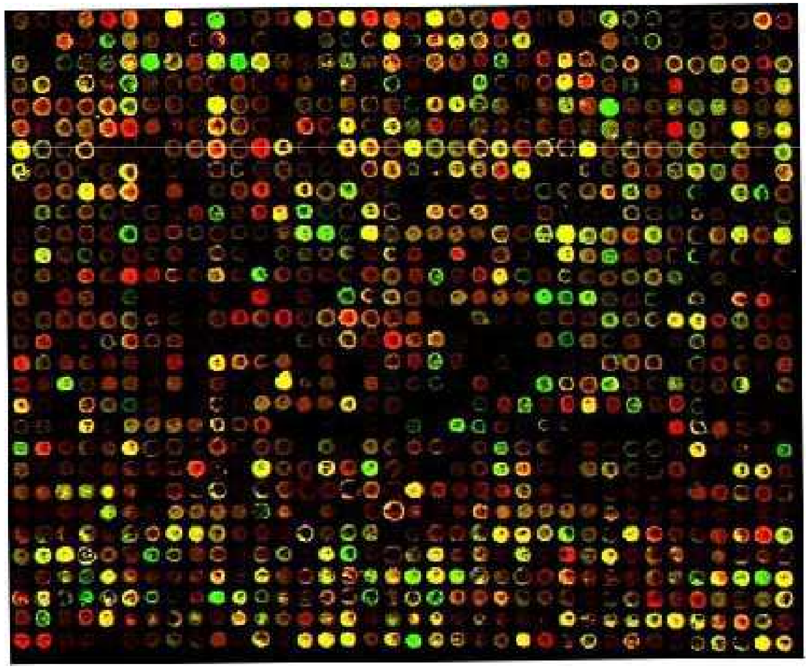
Gene expression analysis

- Microarray



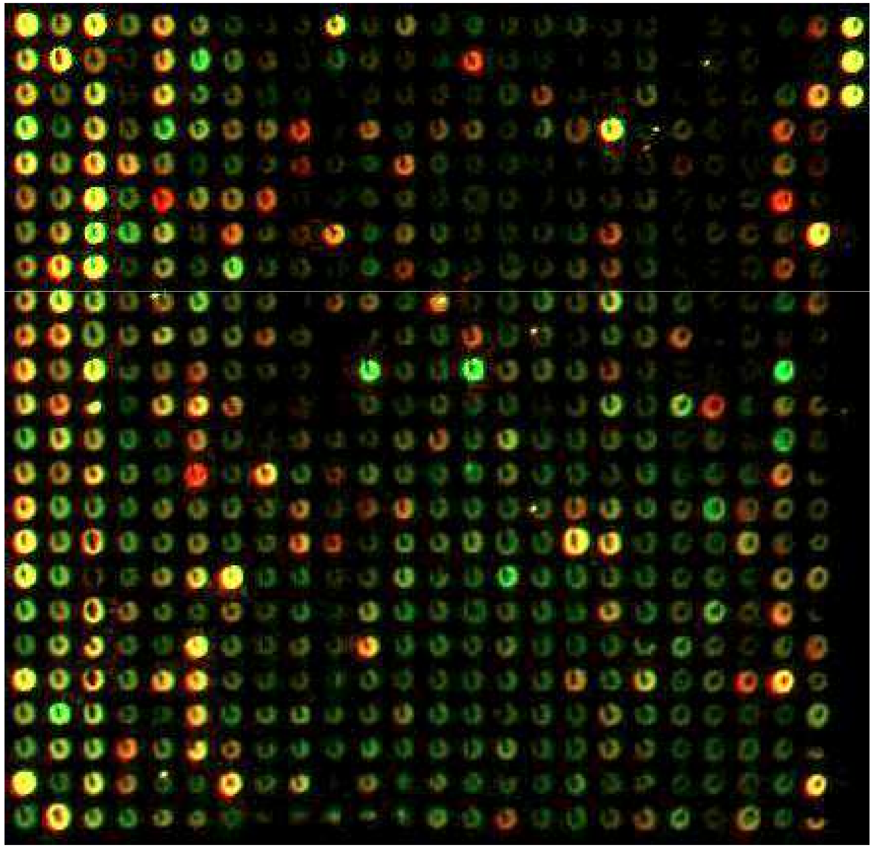
Gene expression analysis

- Microarray



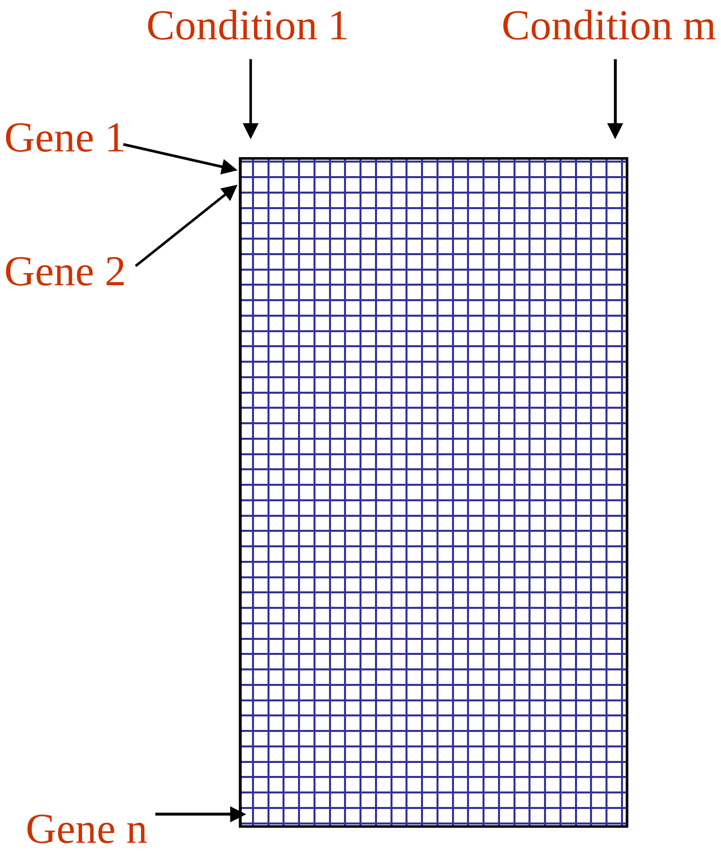
- Green spot: gene is abundant in health state
- Red spot: gene is abundant in disease state
- Yellow spot: gene is abundant in both states
- Black spot: neither health nor disease state express the gene

Gene expression analysis



Gene expression analysis

- Measure expression level under several conditions
  - Normal and cancer tissue
  - Different treatments
    - Before and after using a drug
  - Different periods of treatment
  - Different diseases
- High cost to obtain data



Gene expression analysis

- Data mining techniques have been largely used
  - Classification or clustering
- Microarray data are challenging
  - High dimensionality
  - Irrelevant features
  - Redundant features
  - Noisy data
  - Small number of tissue samples

Gene expression analysis

- Usually works with a subset of genes
  - Identify important genes
  - Improve classification accuracy
  - Minimize effects of noise
  - Make the technology more accessible
    - Become a common clinical tool

Gene expression analysis

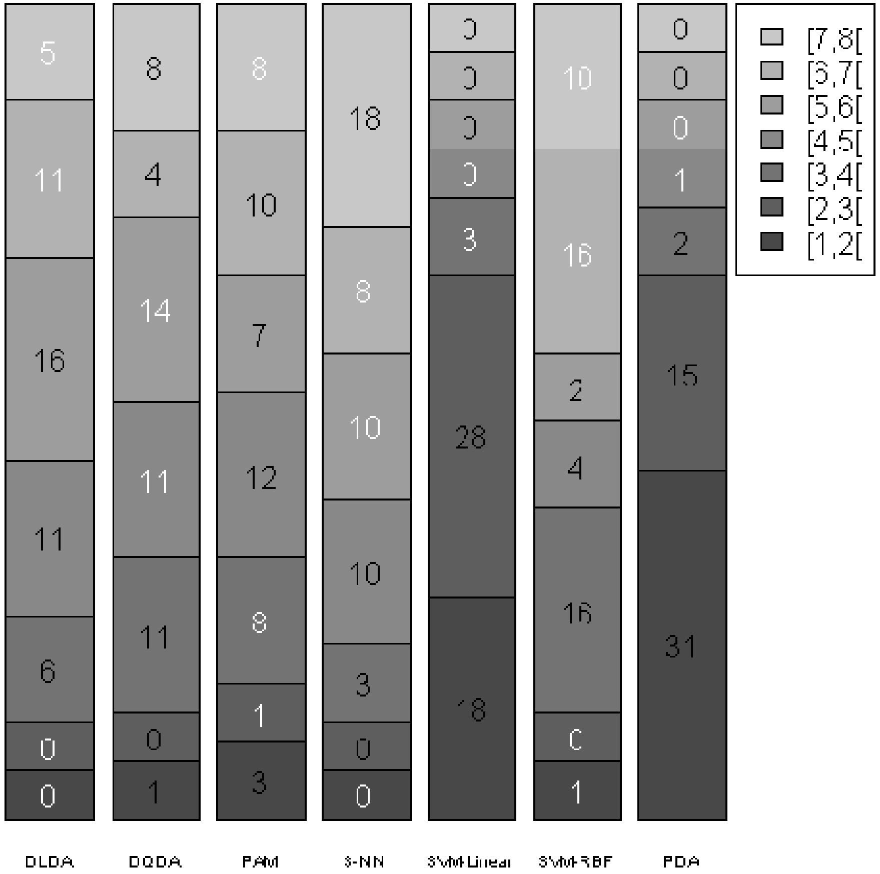
- Gene selection
  - Not all the genes are relevant for tissue classification / clustering
    - Use only the most relevant genes
  - Each gene can be seen as an attribute
  - Problem becomes attribute selection
    - Two approaches are used
      - Ranking of attributes
      - Selection of the best subset of attributes

Experiment 1

- Several ML techniques have been used for gene expression analysis
  - Tissue classification
- Given a gene expression data set, which technique should be used?
  - Trial and error
  - Meta-learning

Experiment 1

- Examples
  - 49 datasets
  - 7 **ML** algorithms
  - Relative performances
  - No clear winner



Metalearning

- Issues with algorithm selection
  - The choice of **ML** algorithm should be data driven
  - Trial-and-error may be very time consuming
- Metalearning
  - Learn from past to predict the future
  - Relate data characteristics with preference for particular algorithms
  - Construct rankings of algorithms
  - It is fast and easy to apply

Metalearning

- 3 main steps
  - Generation of metadata
  - Induction of meta-learning model
  - Application of the metamodel

Metalearning

- Generation of metadata
  - Synthesize data characteristics and algorithms' performances
  - Metaexamples
    - Metafeatures: general, statistical and information-theoretic measures
    - Target: ranking of estimated performances for a set of algorithms
  - Flexible recommendation allows user to try out algorithms in according to his/her preference

Metalearning

- Induction of meta-learning model
  - K-NN ranking method
    - Find nearest metaexamples (Euclidean distance)
    - Combine target rankings (Average rank)

$$\bar{R}_j = \frac{\sum_{i=1}^{k} R_{i,j}}{k}$$

Metalearning

- Application of the metamodel
  - Support user in the algorithm selection process
  - Compute metaexample for new data
    - Metafeatures
    - Target
  - Evaluation of metamodel
    - Leave-one-out (**LOO**)
    - Spearman’s rank correlation for ranking accuracy
    - Default ranking as baseline method
      - All metaexamples are considered

Experimental results

- Data
  - 49 cancer related datasets
    - Mainly disease diagnostics related
    - Diverse data characteristics
    - Usual mean 0, variance 1 transformations
  - Calculation of the meafeatures was preceded by a data reduction step
    - PLS reduced the number of attributes to 3 components

Experimental results

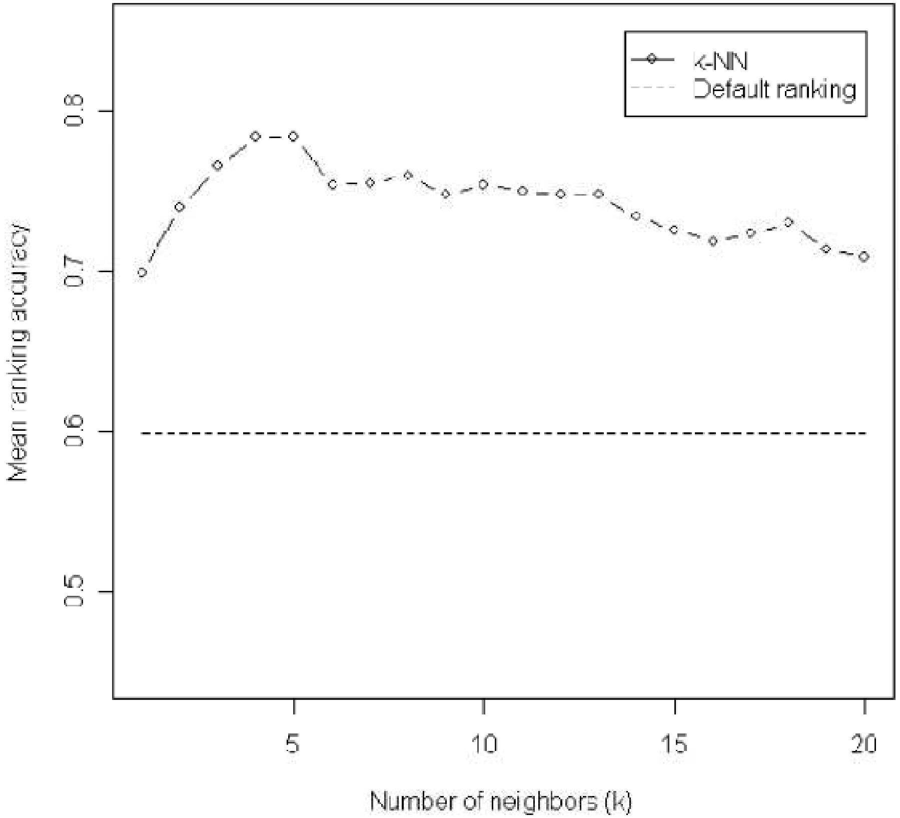
- ML algorithms
  - Common approaches
  - Moderate computational burden
  - Easy availability
  - SVM Linear, SVM RBF, DLDA, DQDA, PAM, 3-NN, PDA
  - Default parameters
  - Performance estimation
    - .632+ estimator with 50 examples

Experimental results

- Metafeatures
  - 10 Statlog continuous measures
    - Log of number of examples
    - Log of number of attributes
    - Log of number of classes
    - Mean absolute skewness
    - Mean kurtosis
    - Geometric mean ratio of the standard deviations of individual populations to the pooled standard deviations
    - First canonical correlation
    - Proportion of total variation explained by the firs canonical correlation
    - Normalized class entropy
    - Average absolute correlation between continuous attributes, per class

Experimental results

- Mean ranking **LOO** accuracies
- Varying k = [1:20]
- Always better than default
- Smooth performance degradation with K
- More homogeneous datasets



Analysis of the results

- Before, metalearning has been applied to general classification domains
- Now, a successful application of metalearning in the gene expression analysis domain is presented
- Future steps
  - Compare metalearning approaches
  - Employ domain specific metafeatures

Protein function prediction

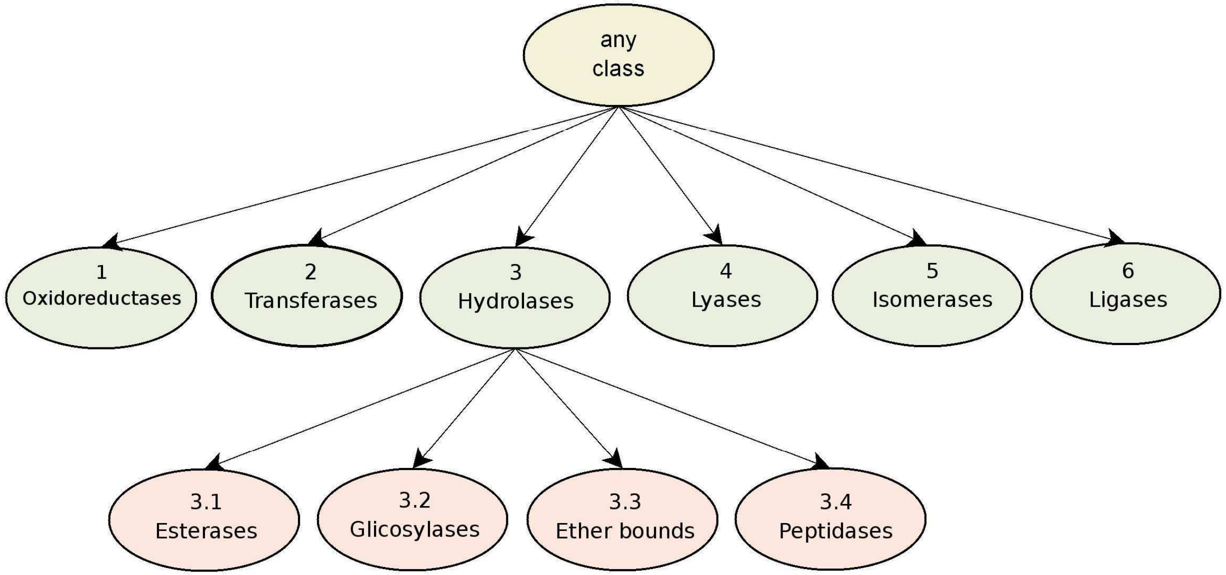
- Allows the assignment of functions to newly discovered proteins
  - Important problem in proteomics
  - Common approach
    - Search for similar frequencies
  - Alternative Approach
    - Induce a classification model

Protein function prediction

- Difficulties associated with the prediction of protein function
  - The same protein may have more than one function
    - Multi-label classification
  - Functions may vary from more generic to more specific
    - Hierarchical classification

Protein function prediction

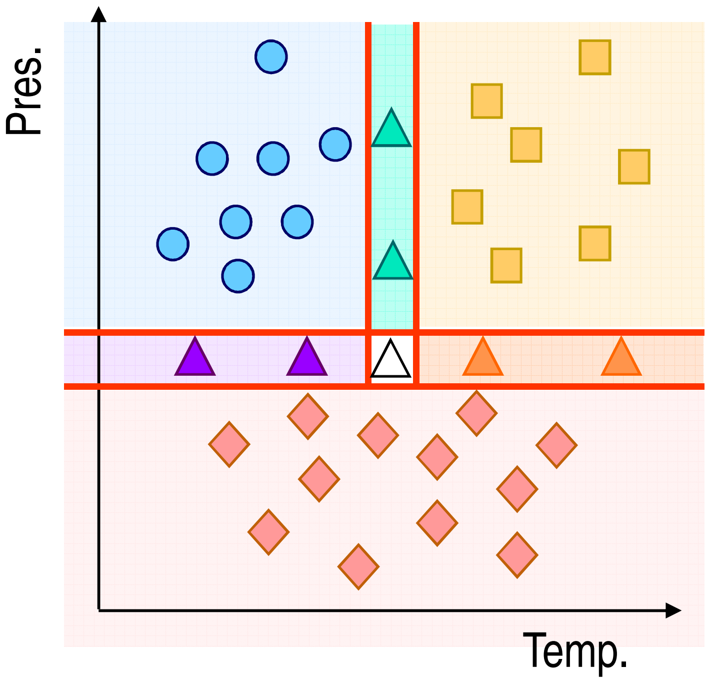
- Class hierarchy of Enzymes



Multi-label classification

- Examples may belong to more than one class
  - Simultaneously

○ Cough	△ Cough e Aches
□ Aches	▲ Cough e Fever
◇ Fever	▲ Aches e Fever
△ Cough, Aches e Fever	



Multi-label classification

- Two main approaches
  - Transformation into a single-label problem
    - Algorithm independent
      - Combination of conventional single label-classifiers
    - Algorithm dependent
      - Modification of single-label classifiers
        - Modification of their internal mechanisms
      - Development of new multi-label classification algorithms
  - Encode multi-label output

Multi-label classification

- Algorithm independent transformation
  - Label-based
  - Instance-based
    - Instance elimination
    - Creation of new labels
    - Label conversion
      - Label elimination
      - Label decomposition

Multi-label classification

- Label-based transformation
  - A classifier is associated to each label / class
    - Binary classification problems

Instance	Classes
1	A and B
2	A
3	A and B
4	C
5	B
6	A

Multi-label Problem

Classifier	Positive	Negative
A	1, 2, 3, 6	4, 5
B	1, 3, 5	2, 4, 6
C	4	1, 2, 3, 5, 6

Single-label problem

Multi-label classification

- Instance-based transformation
  - Instance elimination

Instance	Classes
1	A and B
2	A
3	A and B
4	C
5	B
6	A

Multi-label Problem

Instance	Class
2	A
4	C
5	B
6	A

Single-label problem

Multi-label classification

- Instance-based transformation
  - Label creation (label-powerset)

Instance	Classes
1	A and B
2	A
3	A and B
4	C
5	B
6	A

Multi-label Problem

Instance	Class
1	**D**
2	A
3	**D**
4	C
5	B
6	A

Single-label problem

Multi-label classification

- Instance-based transformation
  - Label elimination

Instance	Classes
1	A and B
2	A
3	A and B
4	C
5	B
6	A

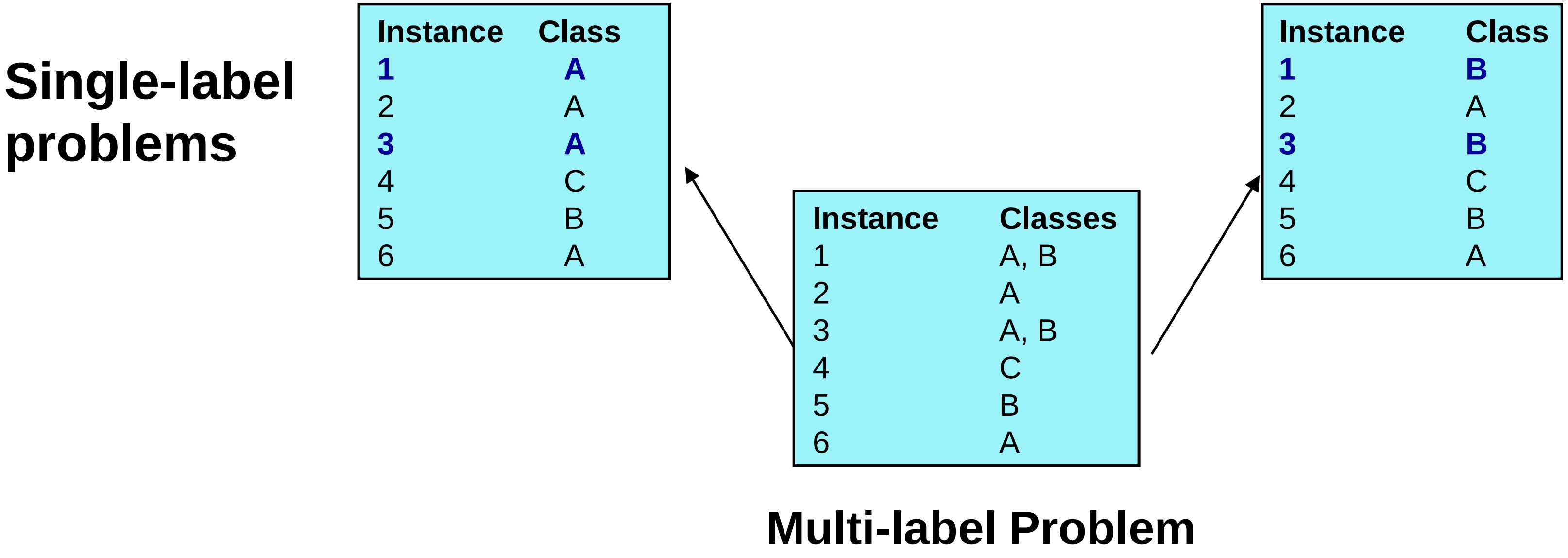
Multi-label Problem

Instance	Class
1	**A**
2	A
3	**B**
4	C
5	B
6	A

Single-label problem

Multi-label classification

- Instance-based transformation
  - Label decomposition (cross-training method)



Single-label problems

Instance	Class
1	**A**
2	A
3	**A**
4	C
5	B
6	A

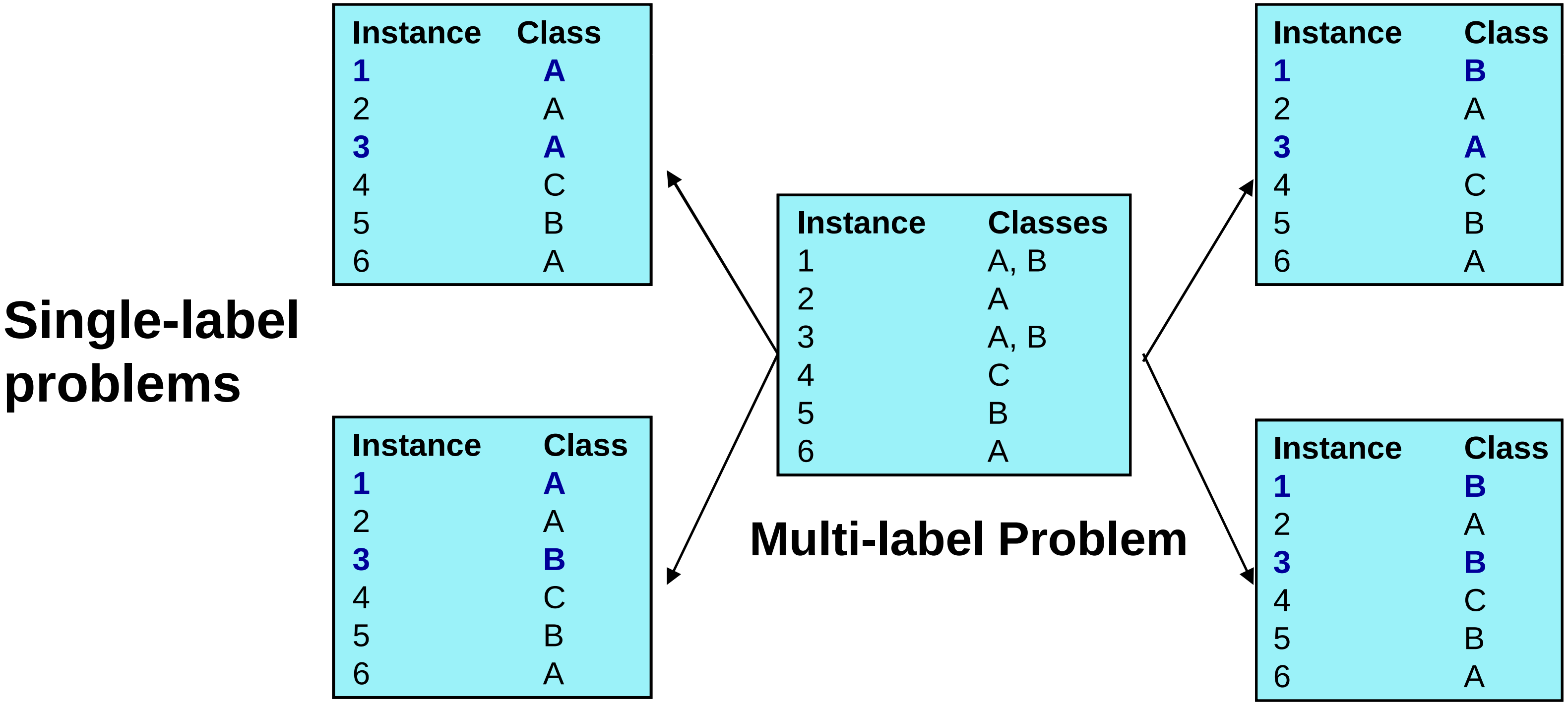
Instance	Class
1	**B**
2	A
3	**B**
4	C
5	B
6	A

Instance	Classes
1	A, B
2	A
3	A, B
4	C
5	B
6	A

Multi-label Problem

Multi-label classification

- Instance-based transformation
  - Label decomposition (multiplicative method)



Single-label problems

Instance	Class
1	**A**
2	A
3	**A**
4	C
5	B
6	A

Instance	Class
1	**A**
2	A
3	**B**
4	C
5	B
6	A

Instance	Classes
1	A, B
2	A
3	A, B
4	C
5	B
6	A

Multi-label Problem

Instance	Class
1	**B**
2	A
3	**A**
4	C
5	B
6	A

Instance	Class
1	**B**
2	A
3	**B**
4	C
5	B
6	A

Output encoding

- Encode desired output by binary vectors
  - Multi-class problem

Instance	Classes
1	A and B
2	A
3	A and B
4	C
5	B
6	A

Multi-label Problem

Class code		
A	**B**	**C**
1	1	0
1	0	0
1	1	0
0	0	1
0	1	0
1	0	0

Single-label problem

Multi-label classification

- Evaluation
  - Require specific measures
  - Examples can be partially correct or partially incorrect classified
  - Classification may use a ranking

Experiments 2

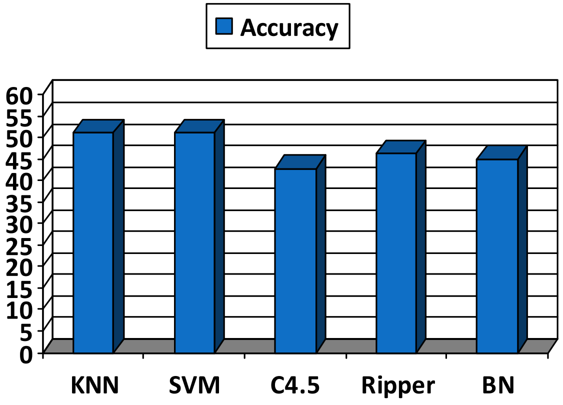
- Comparison of three algorithm independent methods for multi-label classification
  - One-against-all (OAA)
  - Label-Powerset
  - Cross-Training
- Datasets:
  - Yeasts – proteins found in the organism Saccharomyces cerevisiae
  - Sequences – protein sequences classified in structural families

Experiments

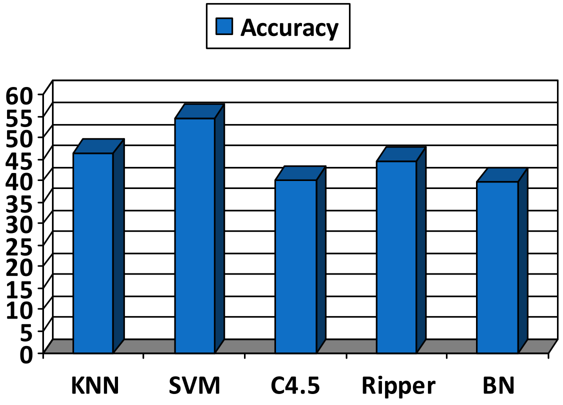
- Datasets
  - Yeasts
    - 2417 examples (2385 multi-label)
    - 103 numerical attributes
    - Distribution (34, 731.5, 1816) and 4.23 classes/example
  - Sequences
    - 662 examples (69 multi-label)
    - 1186 nominal attributes
    - Distribution (16, 54.5, 171) and 1.15 classes/example

Experiments

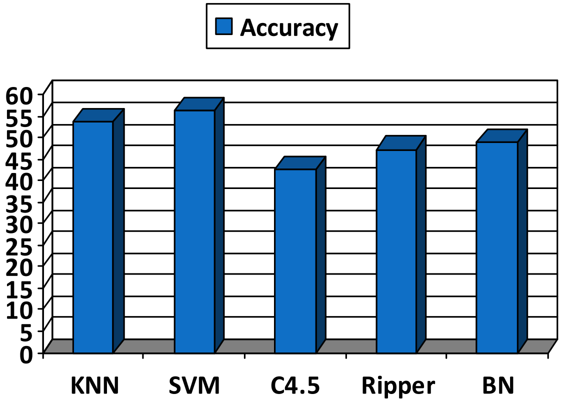
- Yeasts dataset



OAA



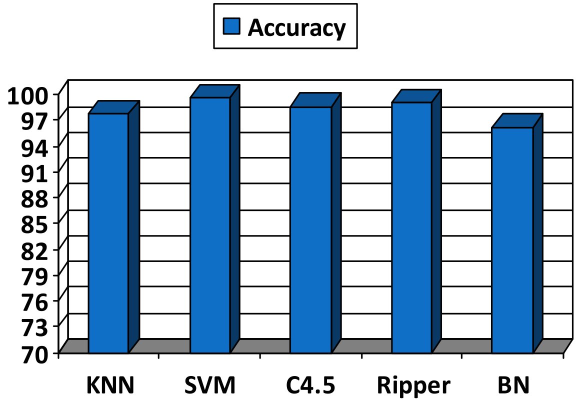
Label-Powerset



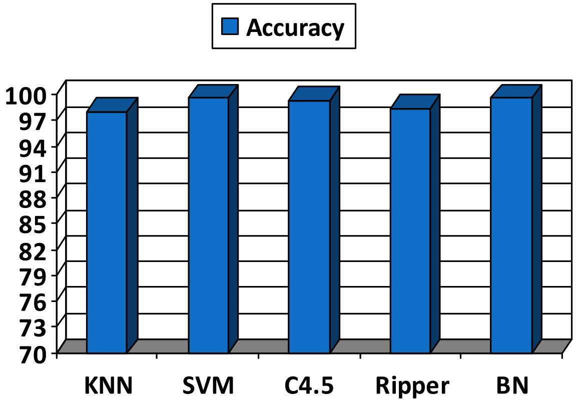
Cross-Training

Experiments

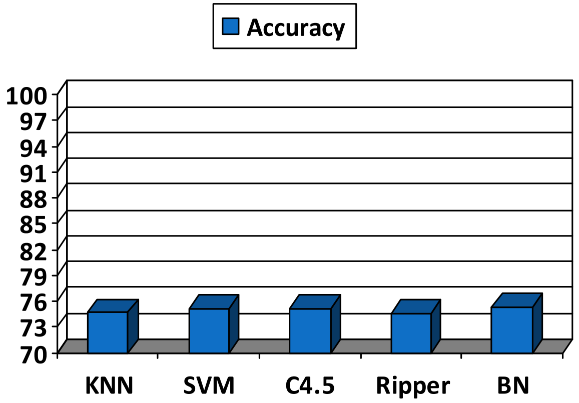
- Sequence dataset



OAA



Label-Powerset



Cross-Training

Analysis of the results

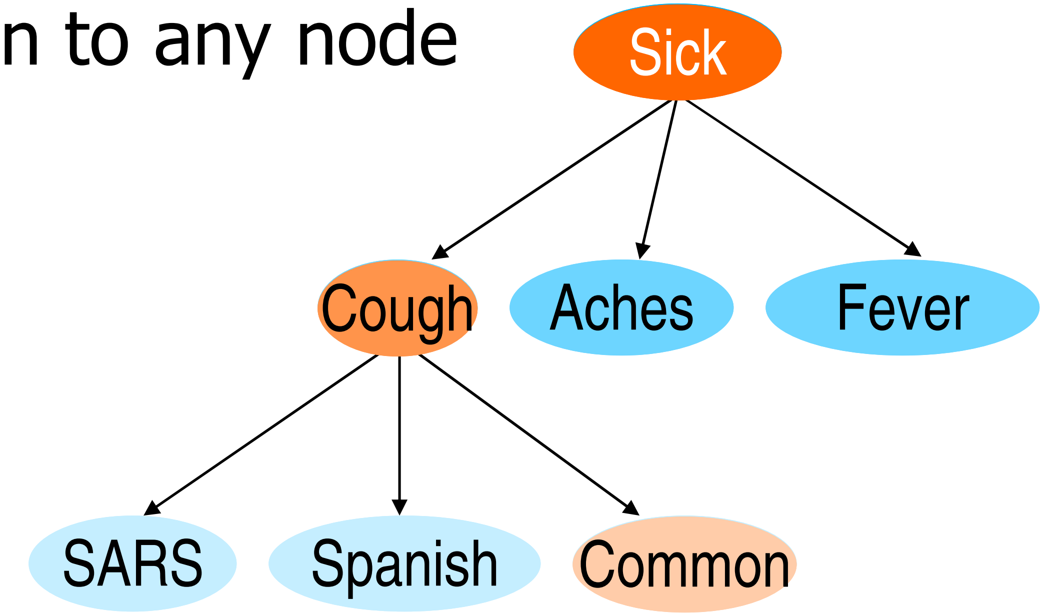
- Each method favours a specific feature of the dataset
  - High / low frequency of multi-label examples
- SVMs usually presents a better predictive accuracy
- Similar results for other datasets and performance metrics

Hierarchical classification

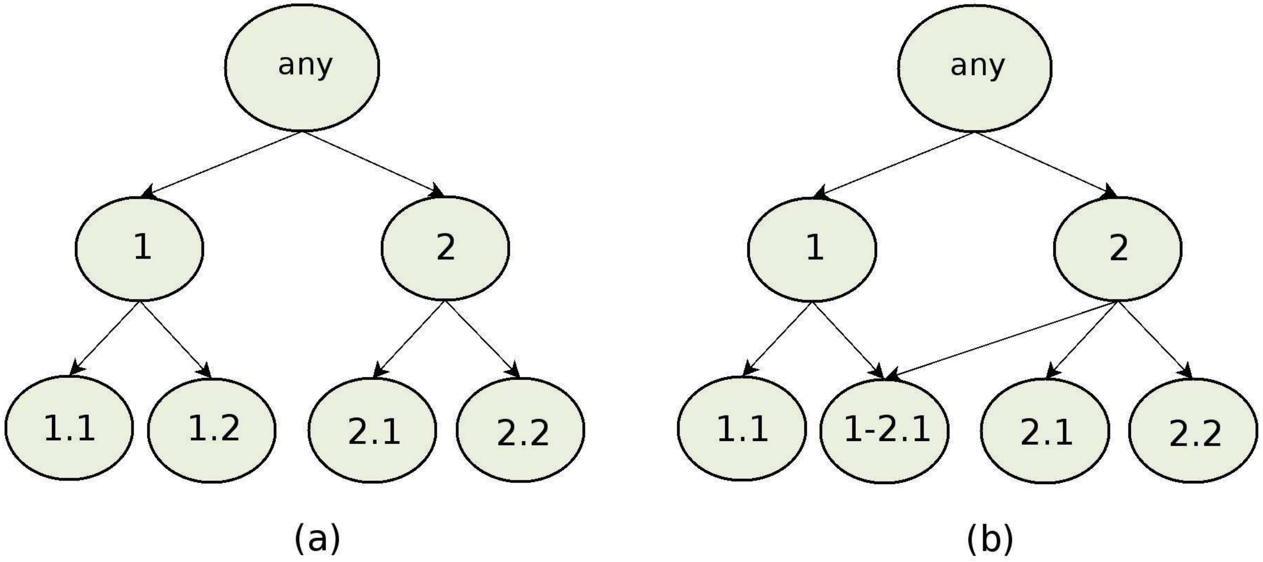
- Classification problems where:
  - Classes can be partitioned into subclasses
  - Classes can be grouped into superclasses
- Data are hierarchically organized
  - {1, 1.1, 1.2, ..., *k*, *k*.1, *k*.2}
  - Classes assume an hierarchical organization

Hierarchical classification

- Types of hierarchical based classification
  - Mandatory prediction to leaf nodes
  - Prediction to any node



Types of hierarchy



(a) Trees

(b) Direct Acyclic Graphs (DAG)

Hierarchical classification

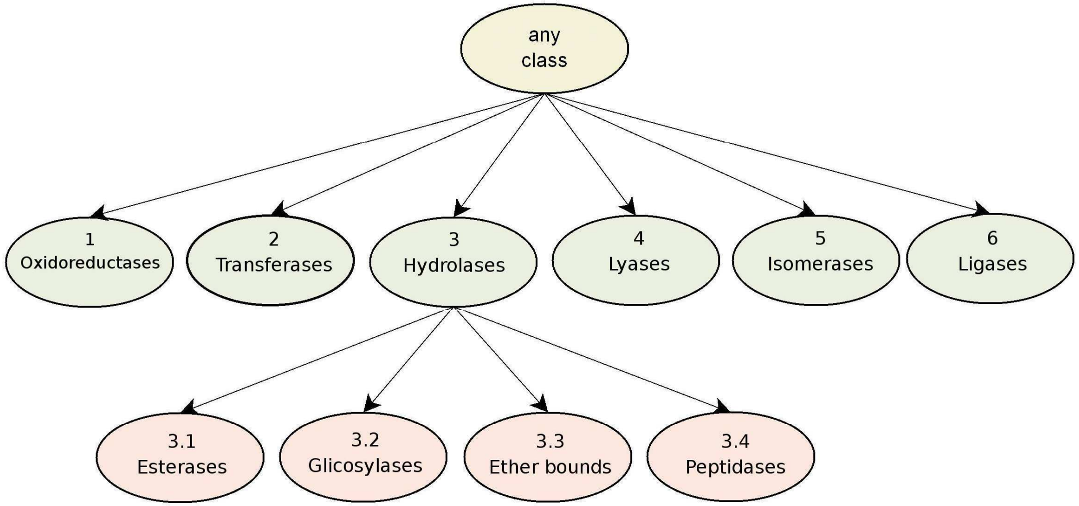
- Main approaches
  - Transformation into a flat classification problem
  - Hierarchical prediction with flat classification algorithms
  - Top-down
  - Big-bang (one-shot)

Hierarchical classification

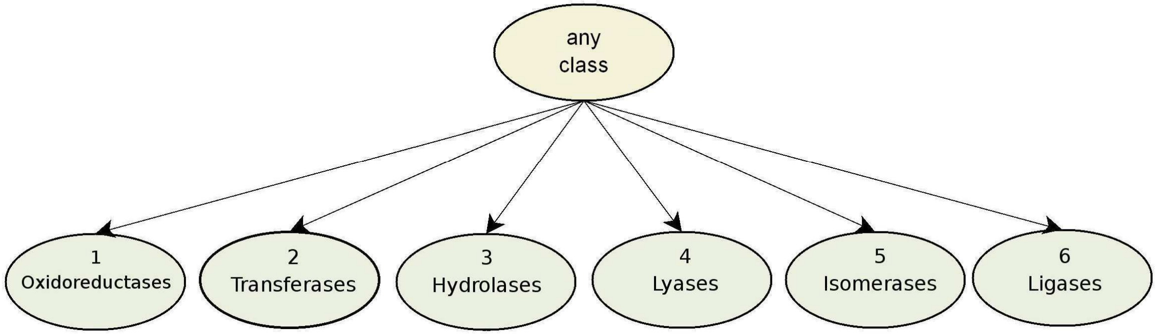
- Transformation into a flat classification problem
  - Reduces problem complexity
  - Most used method:
    - Select a hierarchy level and perform flat classification in this level
  - Advantage: the simplest approach
  - Disadvantage: classification in the other levels of the hierarchy is lost

Hierarchical classification

Original problem



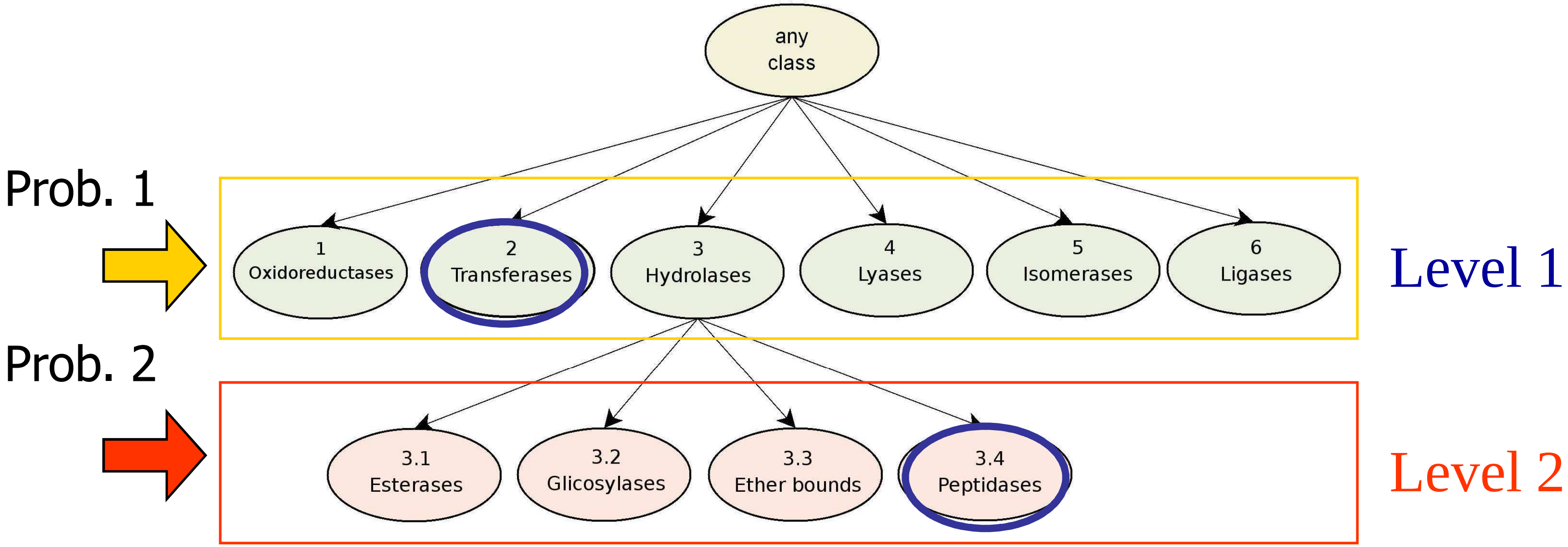
Flat classification problem



Hierarchical classification

- Hierarchical prediction with flat classification algorithms
  - Divides the original problem into a set of flat classification problems
    - A flat classification for each level
  - Advantage: no need to modify flat classification algorithms
  - Disadvantage: classifications in different levels may be inconsistent
    - Ex. class 2 (level 2) and class 3.4 (level 3)

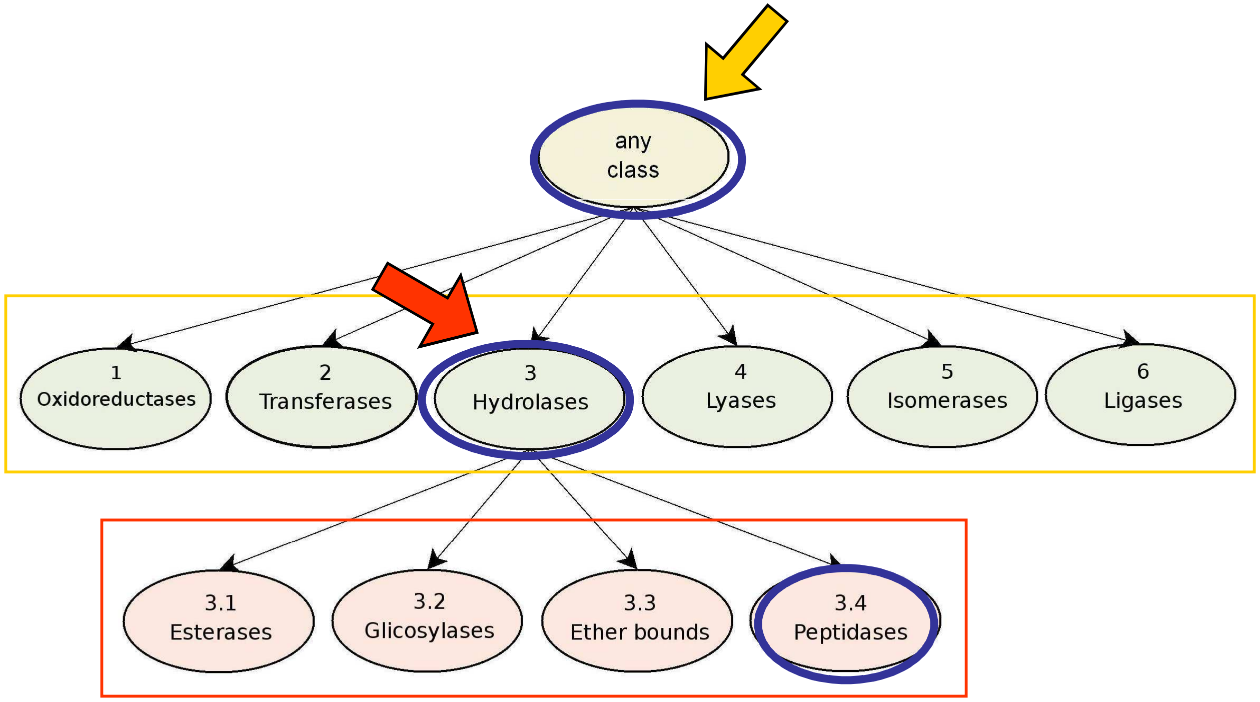
Hierarchical classification



Hierarchical classification

- Top-down
  - Divides original problem into a set of flat classification problems
    - Which are dealt with sequentially, level by level, from the root
    - Classification proceeds in the sub-tree associated with the previously chosen node
  - Advantage: no need to modify flat classification algorithms
  - Disadvantage: risk of classification error propagation

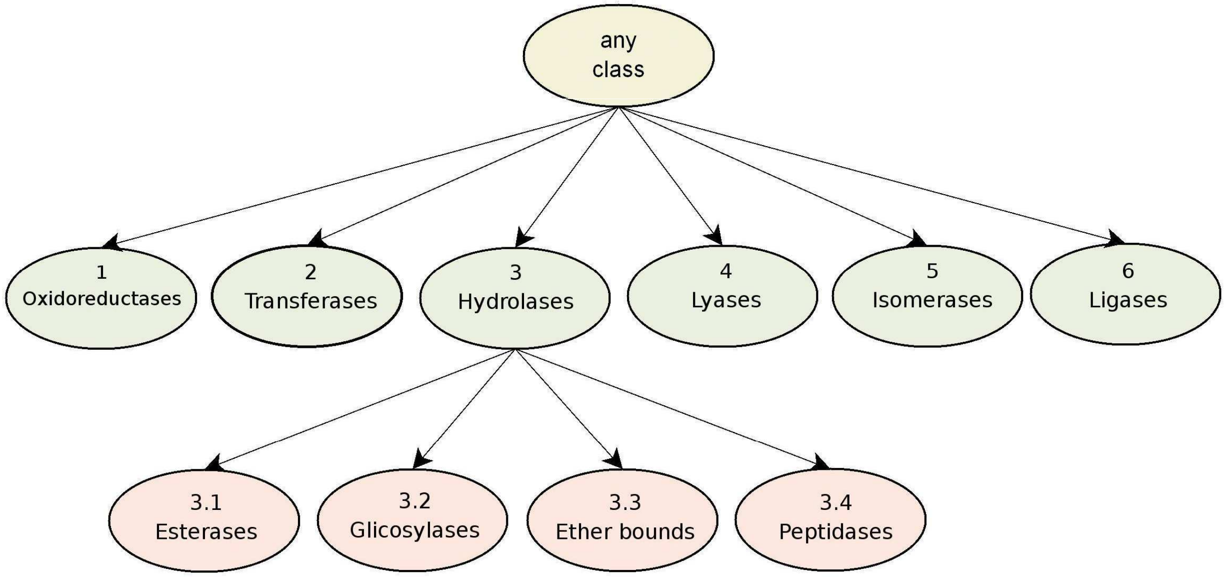
Hierarchical classification



Hierarchical classification

- Big-bang
  - Classification algorithm considers the whole hierarchy
  - Advantage: classification is carried out in just one go
  - Disadvantage: complexity of the algorithms

Hierarchical classification



Hierarchical classification

- Evaluation measures
  - Uniform cost
    - Most used
  - Distance-based cost
    - Based on the distance between the predicted class and the true class
  - Depth dependent
    - Errors at higher levels should have a higher cost
  - Semantic-based cost
    - More similar classes classes have smaller penalizations

Hierarchical classification

- Evaluation measures might
  - Report na accuracy rate for the whole hierarchy
  - Report na accuracy rate for each level
  - Report na accuracy rate for each class

Experiment 3

- Two datasets
  - G-Protein-Coupled Receptors (GPCRs)
  - Enzymes
- Data extracted from UniProt and GPCRDB
- Attributes:
  - Interpro entries, along with molecular weight and sequence length

Data sets

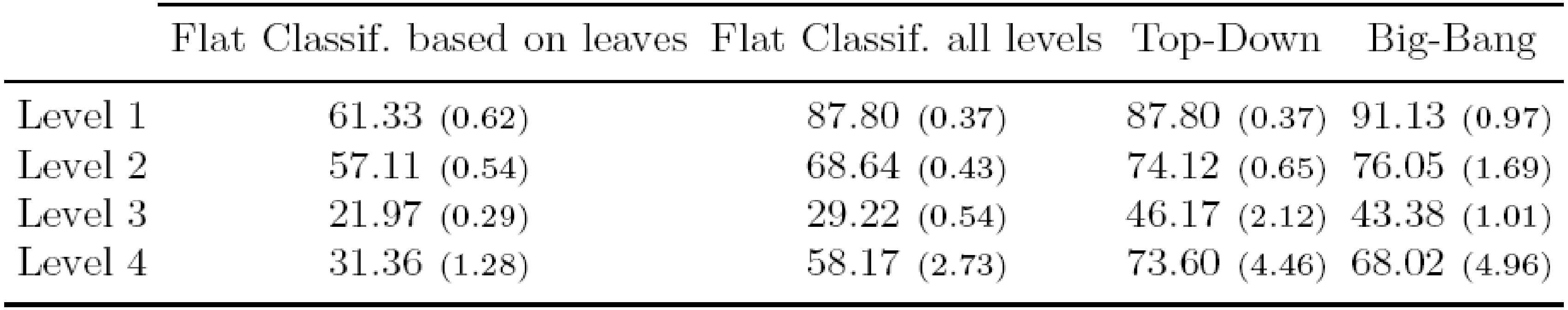
- G-Protein-Coupled Receptors (GPCRs)
  - 40-50% of current drugs target GPCR activity
  - 7461 instances
  - Class hierarchy
    - 12/54/82/50 classes per level
- Enzymes
  - Catalysts which are used to speed up chemical reactions within the cell
  - 6925 instances
  - Class hierarchy
    - 2/21/48/87 classes per level

Investigated approaches

- Classifier technique: decision trees
- Four models were used:
  - Flat – based on leaves
  - Flat – all levels
  - Top-Down
  - Big-Bang
    - (Clare and King, 2003)

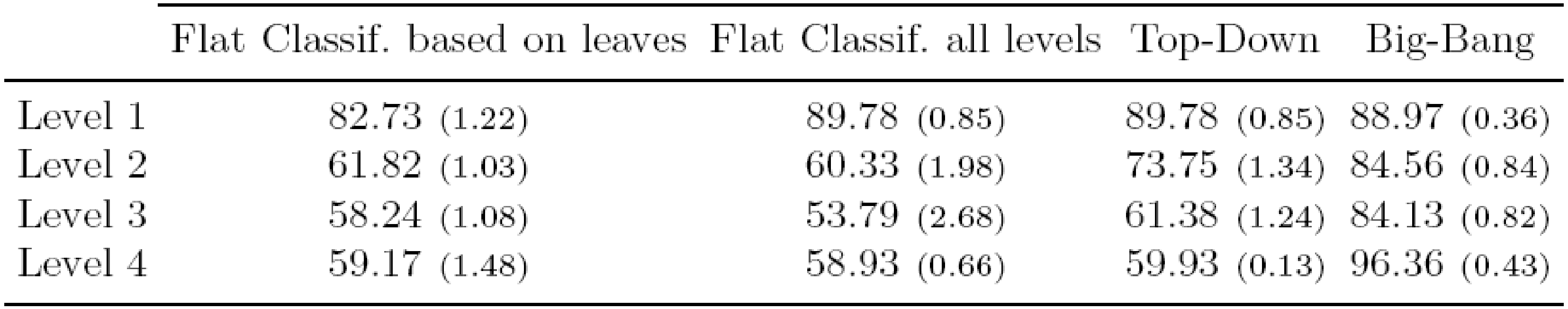
Hierarchical classification of proteins

Table 1. Accuracy results in the GPCR dataset



	Flat Classif. based on leaves	Flat Classif. all levels	Top-Down	Big-Bang
Level 1	61.33 (0.62)	87.80 (0.37)	87.80 (0.37)	91.13 (0.97)
Level 2	57.11 (0.54)	68.64 (0.43)	74.12 (0.65)	76.05 (1.69)
Level 3	21.97 (0.29)	29.22 (0.54)	46.17 (2.12)	43.38 (1.01)
Level 4	31.36 (1.28)	58.17 (2.73)	73.60 (4.46)	68.02 (4.96)

Table 2. Accuracy results in the Enzyme dataset



	Flat Classif. based on leaves	Flat Classif. all levels	Top-Down	Big-Bang
Level 1	82.73 (1.22)	89.78 (0.85)	89.78 (0.85)	88.97 (0.36)
Level 2	61.82 (1.03)	60.33 (1.98)	73.75 (1.34)	84.56 (0.84)
Level 3	58.24 (1.08)	53.79 (2.68)	61.38 (1.24)	84.13 (0.82)
Level 4	59.17 (1.48)	58.93 (0.66)	59.93 (0.13)	96.36 (0.43)

Analysis of the results

- Hierarchical approaches are better than flat approaches
- GPCR
  - Top-Down
- Other classifiers
  - Ensembles
- Different metrics

Conclusion

- Data Mining
- Motivation
- Molecular Biology
- Bioinformatics problems
  - Analysis of Gene Expression
  - Protein function classification
  - DM solutions

Conclusion

- Classification can be a complex tasks
- New types of problems are being investigated
  - And novel demands may arise
- New techniques are needed
  - And metrics to evaluate them in these non-trivial classification problems

Acknowledgements

- Alex Freitas, University of Kent
- Ana Carolina Lorena, UFABC
- André Rossi, ICMC-USP
- Bruno Feres de Souza, ICMC-USP
- Katti Faceli, UFSCar
- Eduardo Costa, UKL
- Eduardo J. Spinosa, ICMC-USP
- Carlos Soares, Universidade do Porto
- João Gama, Universidade do Porto

Acknowledgements

cria

USP

UFABC
Universidade Federal do ABC

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

CNPq
Conselho Nacional de Desenvolvimento
Científico e Tecnológico

CAPES

FAPESP

U.PORTO