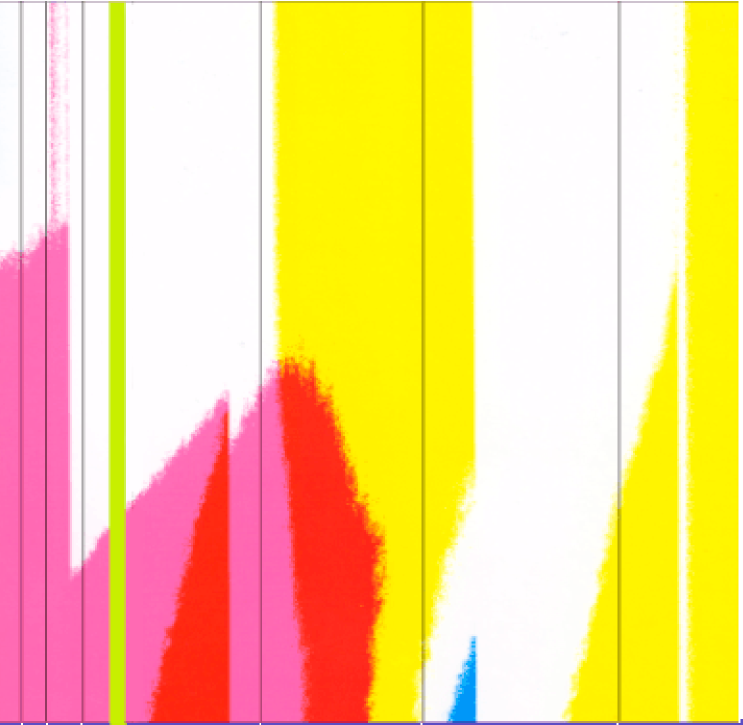


Combinatorial Optimization in Bioinformatics

Clarisse Dhaenens, Laetitia Jourdan
University of Lille - France

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

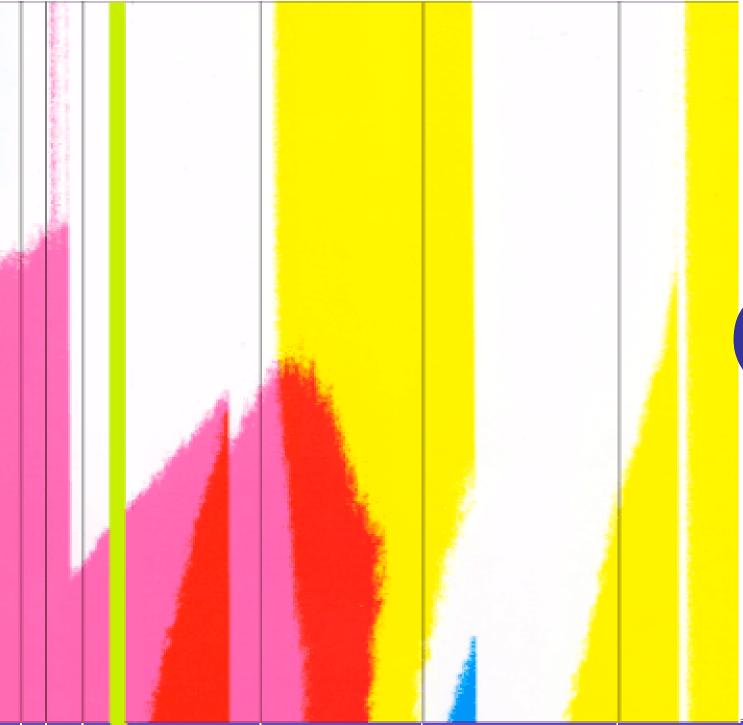
INRIA

centre de recherche FUTURS

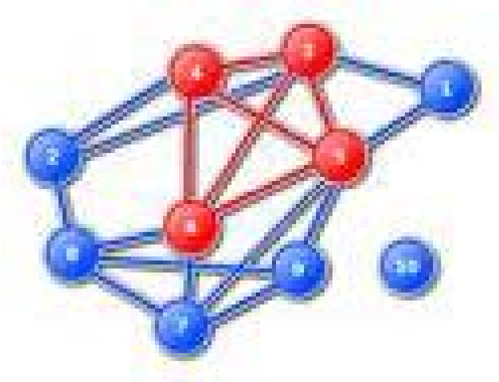
5th IAPR International Conference on
Pattern Recognition in Bioinformatics
22–24 September 2010, Nijmegen, The Netherlands

Outlines

- Optimization
  - Combinatorial optimization
  - MetaHeuristics
  - Multi-objective optimization
- Applications in Bioinformatics
  - Combinatorial optimization for Datamining
    - Feature selection
    - Association rules
  - A Genetic algorithm for Molecular docking



Optimization ??



A small introduction to

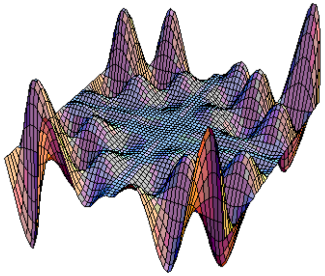
COMBINATORIAL OPTIMIZATION

C. Dhaenens, L. Jourdan – PRIB 2010

Definition

Wikipedia

Combinatorial optimization is a topic in theoretical computer science and applied mathematics that consists of finding the least-cost solution to a mathematical problem in which each solution is associated with a numerical cost.



(P) $\quad Opt\ F(x) \longrightarrow$ Cost = objective function (min/max)

s.c. $\quad x \in C \longrightarrow$ Set of feasible solutions defined using constraints

Combinatorial problem

Model: different elements to be defined

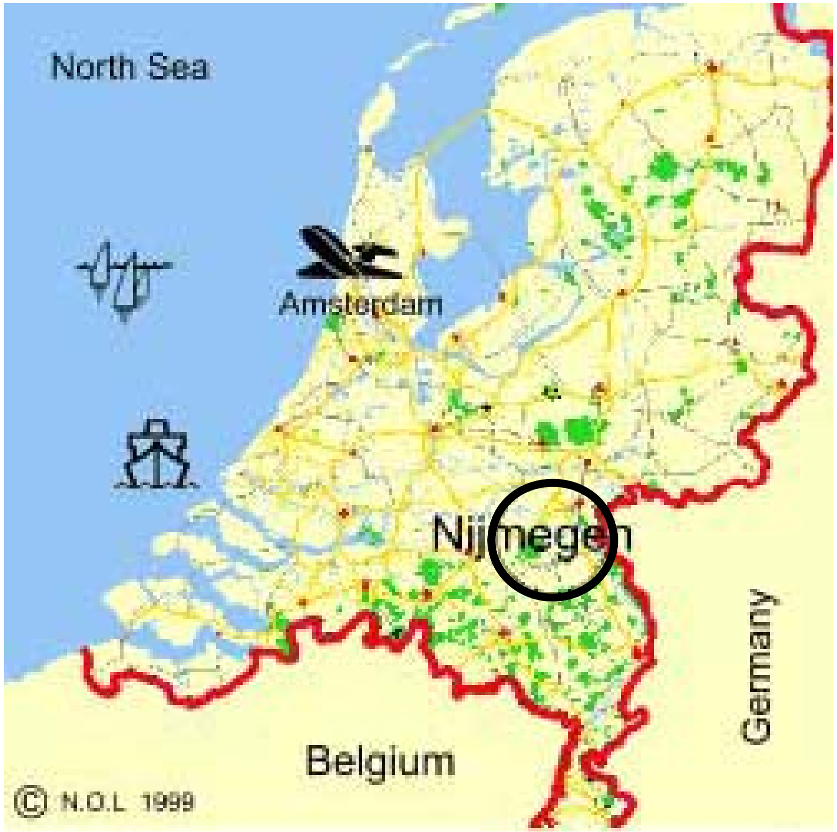
- Solutions
  - How to characterize a solution?
  - How to define feasible solutions?
- Objective function
  - What is the criterion to optimize (cost, duration…)?
  - Is there only one criterion?

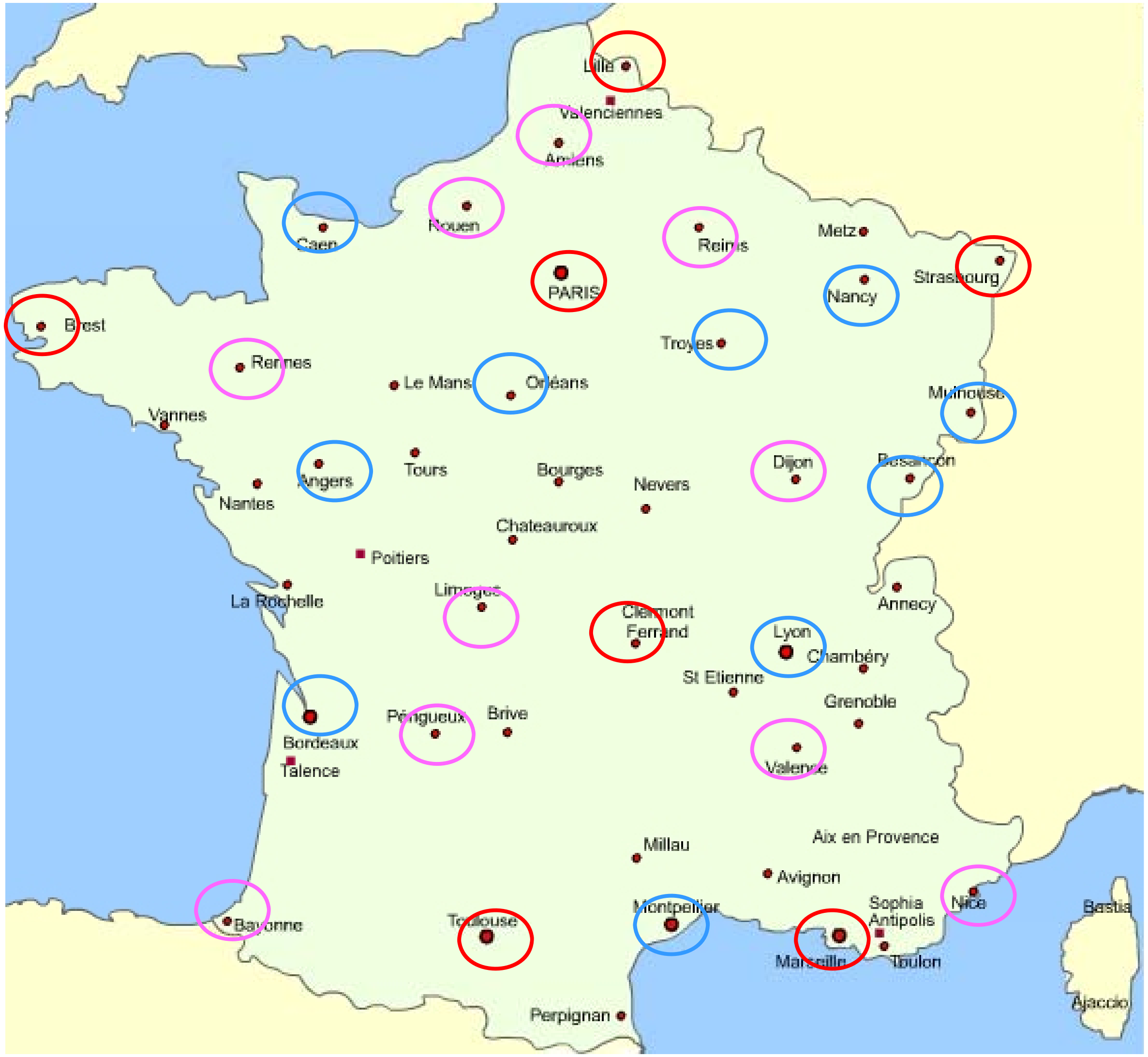
An example:
The traveling salesman problem (TSP)

- First formulated as a mathematical problem in 1930

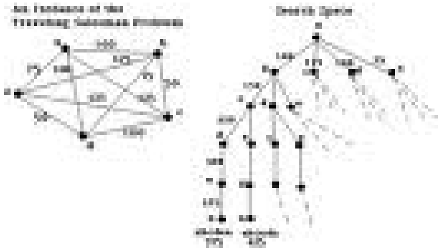
- One of the most intensively studied problems in optimization (Operations Research)

- « Given a list of cities and their pairwise distances, the task is to find a shortest possible tour that visits each city exactly once. »


North Sea
Amsterdam
Nijmegen
Germany
Belgium
© N.O.L 1999


Lille
Valenciennes
Amiens
Caen
Rouen
Reims
Metz
Strasbourg
Nancy
PARIS
Brest
Troyes
Rennes
Le Mans
Orléans
Mulhouse
Vannes
Dijon
Besançon
Tours
Angers
Bourges
Nevers
Nantes
Chateauroux
Poitiers
Limoges
Annecy
La Rochelle
Clermont Ferrand
Lyon
Chambéry
St Etienne
Grenoble
Brive
Périgueux
Bordeaux
Talence
Valence
Aix en Provence
Millau
Avignon
Sophia Antipolis
Nice
Bastia
Montpellier
Bayonne
Toulouse
Marseille
Toulon
Ajaccio
Perpignan

The traveling salesman problem



- **NP-hard problem**
  - **No efficient (polynomial) algorithm**

- **Simple resolution**:
  Exhaustive enumeration of all solutions
  If N cities ➔ (N-1)! Possibilities

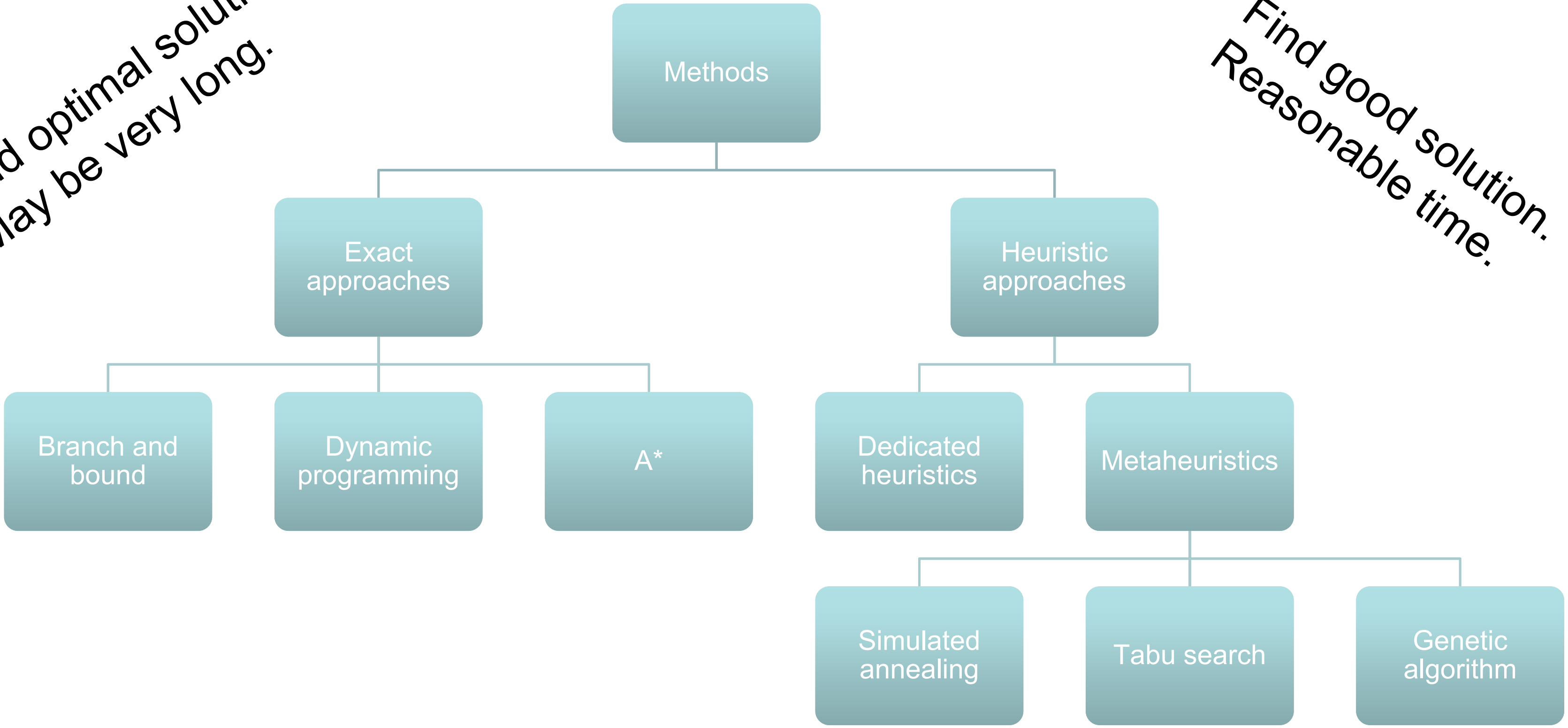
Ex : 5 cities ➔ 12 possibilities	**6 µsec**
10 cities ➔ 181 440 possibilities	**0,09 sec**
20 cities ➔ 60 × 10^{15}	**964 years**

- Let's suppose a computer requires 1/2 microsecond to evaluate a tour.

Need efficient combinatorial optimization methods

Combinatorial optimization methods

Find optimal solution.
May be very long.

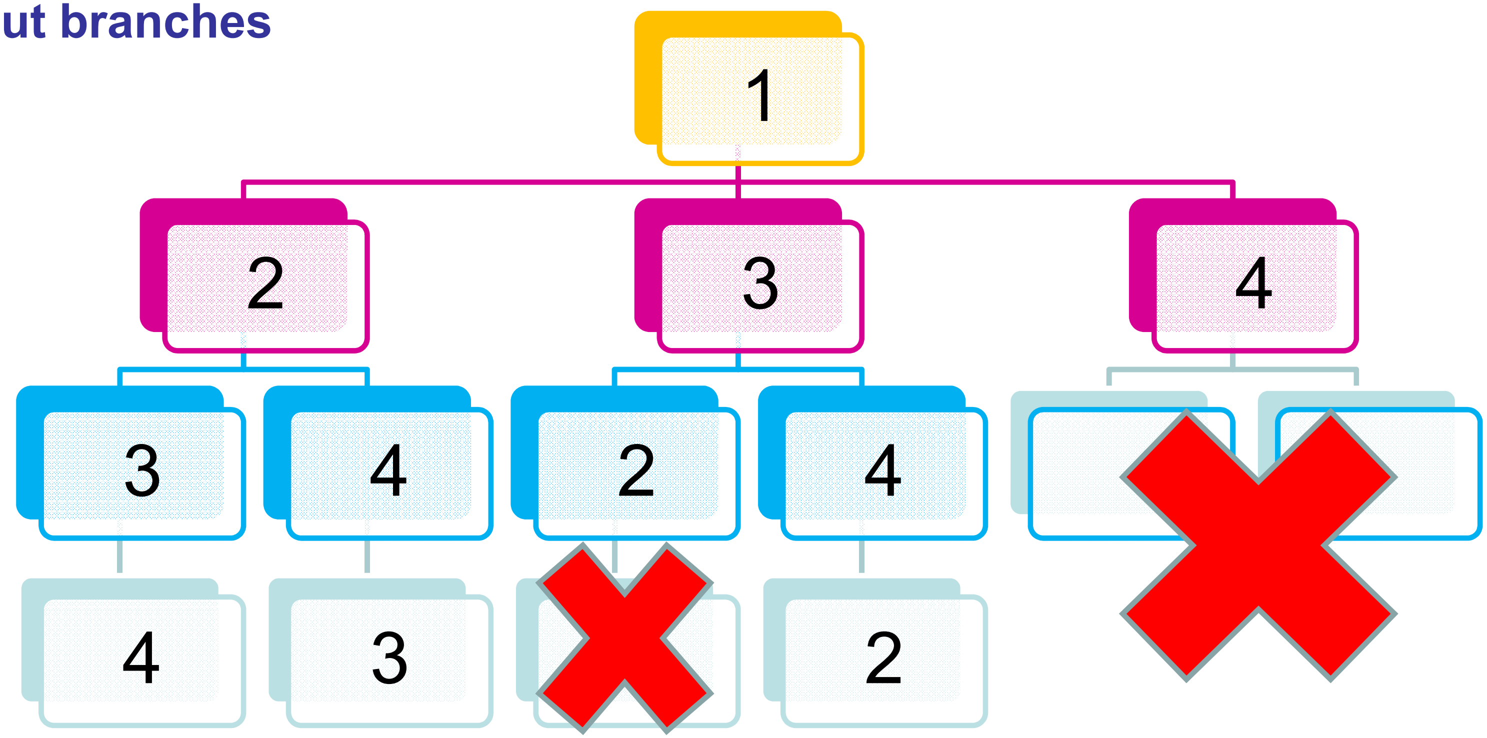


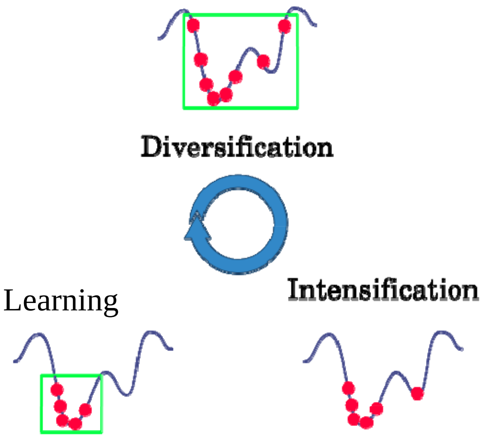
Find good solution.
Reasonable time.

The traveling salesman problem

Exact method

- **Intelligent enumeration**
- **Cut branches**





Presentation

METAHEURISTICS

Definition

Wikipedia

In computer science, **metaheuristic** designates a **computational method** that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality.

Metaheuristics make few or no assumptions about the problem being optimized and **can search very large spaces** of candidate solutions.

However, metaheuristics **do not guarantee an optimal solution** is ever found.

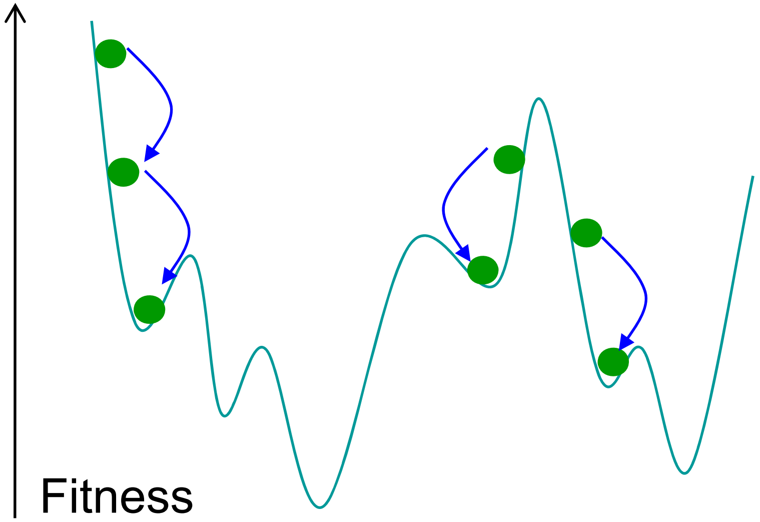
Many metaheuristics implement some form of **stochastic optimization**.

Descent method

Hill climbing

Gradient method

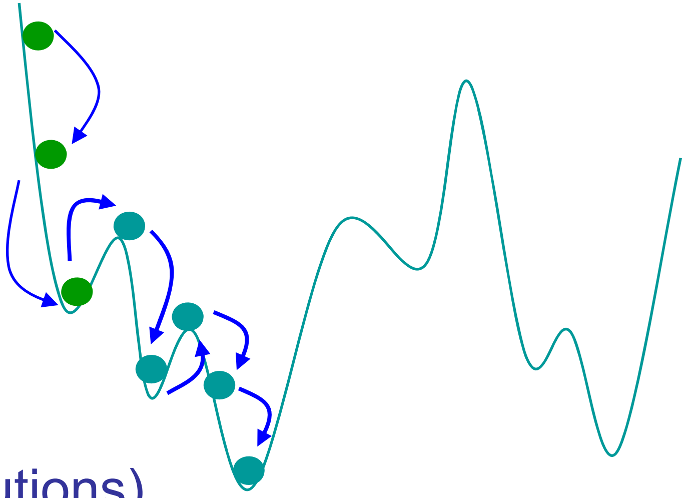
- Neighborhood notion
  - Small modification
  - Local search
- Landscape representation
- From an initial solution
  - Look for a best neighbor
  - Move to this neighbor
  - When no better neighbor → local optimum



Minimization problem

Tabu search [Glover, 1986]

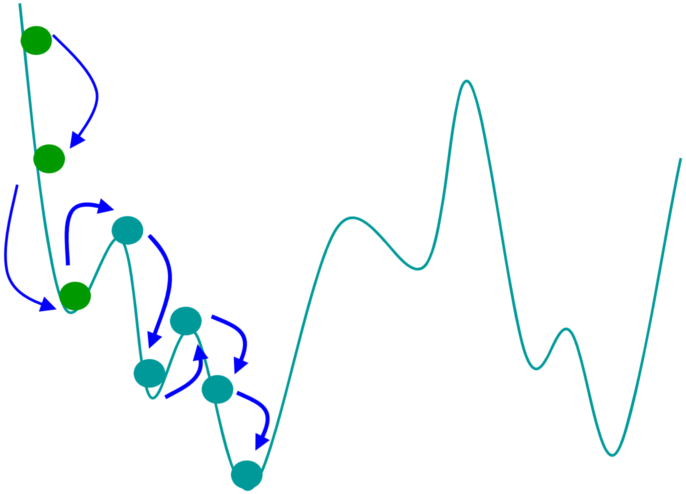
- From an initial solution
  - Look for a best neighbor
  - Move to this neighbor
  - When no better neighbor
    - May degrade the solution
    - Interdiction to come back to recently visited solution (tabu solutions)



  - Parameters:
    - Tabu move
    - Size of the Tabu list (short term memory)

Simulated annealing [Kirkpatrick, 1983]

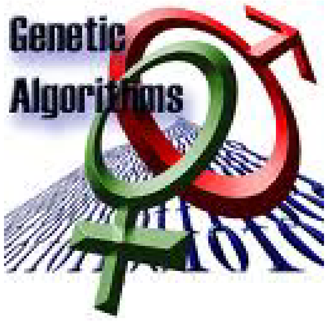
- Name inspired from annealing in metallurgy



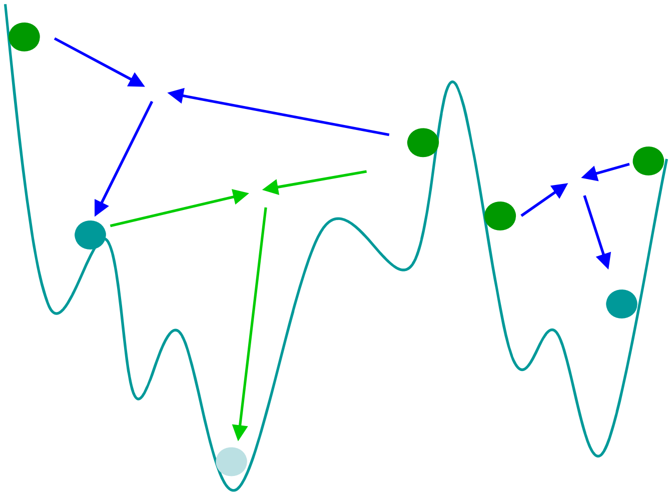
- From an initial solution
  - Look for a neighbor
  - If better solution
    - Move to this neighbor
  - If not
    - Accept to move to this neigbhor according to a probability that depends on a temperature T
  - Parameters:
    - Management of temperature T

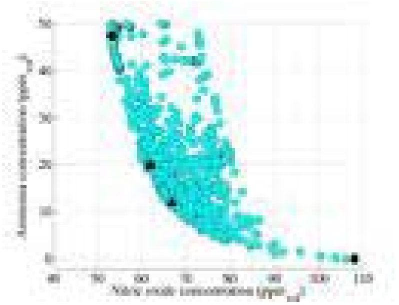
Genetic algorithm

[Holland, 1975]



- Population based (set of solutions)
- Inspired by natural evolution
  - Inheritance
  - Selection
  - Mutation …
- Global improvement
- Parameters:
  - Objective function
  - Population size
  - Operators
  - Selection of parents
  - Replacement





Introduction to

MULTI-OBJECTIVE OPTIMIZATION

C. Dhaenens, L. Jourdan – PRIB 2010

Motivations

- Many real world problems are multi-objective by nature
- Objectives may be in conflict
- Not always possible to construct a single criterion

Main concepts

- Multi-objective Optimization Problem (MOP):

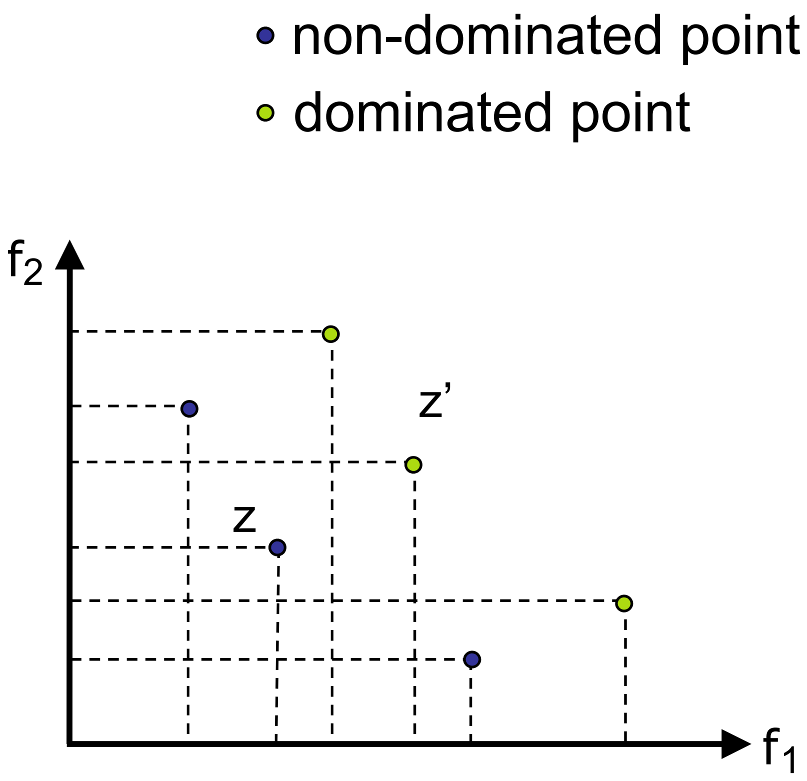
$$(\text{MOP}) = \begin{cases} \text{min (or max) } f(x) = (f_1(x), f_2(x), \ldots, f_n(x)) \\ \text{Subject to } x \in X \end{cases}$$

  - $n \geq 2$ objective functions $f_1(x), f_2(x), \ldots, f_n(x)$
  - $x \in X$ is a decision vector $(x_1, x_2, \ldots, x_k)$
  - X is the set of feasible solutions in the decision space
  - Z is the set of feasible points in the objective space

Dealing with multiple objectives

Definitions:

- $z \in Z$ dominates $z' \in Z$ iff $\forall i \in [1..n]$, $z_i \leq z_i'$ and $\exists j \in [1..n]$, $z_j < z_j'$.
- $z \in Z$ is **a non-dominated vector** if there does not exist another $z' \in Z$ such that z' dominates z.
- The **Pareto frontier** is the set of all non-dominated points.
- The **efficient set** is the set of all efficient solution.



objective space

Difficulties of MOP

- Definition of the optimality: partial order relation, final choice depend on the decision
- Number of Pareto solutions grows with the problem size and the number of criteria
- For non convex MOP, solutions are not all located on the domain boundary but also in the convex hull → difficulty to find them.
- Performance assessment is difficult (comparisons of methods = comparisons of sets of solutions)

Population based algorithms are well fitted to solve Multi-objective problems

Non-dominated Sorting GA (NSGA-II) [Deb et al. 2002]

- **Initialization** of population P
- **Fitness assignment** non-dominated sorting
  - Population divided into fronts
  - Fitness (x) = index of the front x belongs to
- **Diversity** preservation ⇔ crowding distance.
- **Selection** ⇔ Binary tournament
- Recombination and mutation operators
- **Replacement** ⇔ N worst individuals are removed
- **Elitism** ⇔ Archive A of potentially efficient solutions

Pareto based

Indicator-Based EA (IBEA)

[Zitzler et al. 2004]

- **Initialization** of population P
- **Fitness assignment** quality indicator Q_i :
  - Fitness (x) = Q_i (x , P\{x})

Indicator based

- **Diversity preservation** ⇔ none
- **Selection** ⇔ binary tournament
- Recombination and mutation operators
- **Replacement** ⇔ remove the worst individual and update fitness values until |P| = N
- **Elitism** ⇔ Archive A of potentially efficient solutions

Applications in Bioinformatics

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

5th IAPR International Conference on
Pattern Recognition in Bioinformatics
22–24 September 2010, Nijmegen, The Netherlands

Outlines

- Datamining examples

  → Modeling datamining tasks as MCOP (Multi-Objective Combinatorial Optimization Problems)

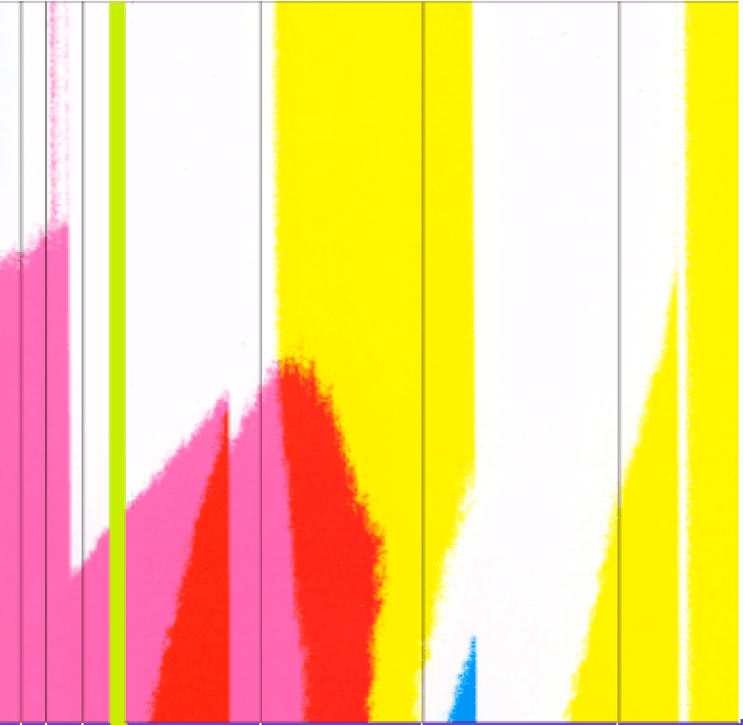
  → Clustering

  → Association rules

- Molecular docking

  → New optimization model

  → Efficient optimization methods



Datamining in Bioinformatics

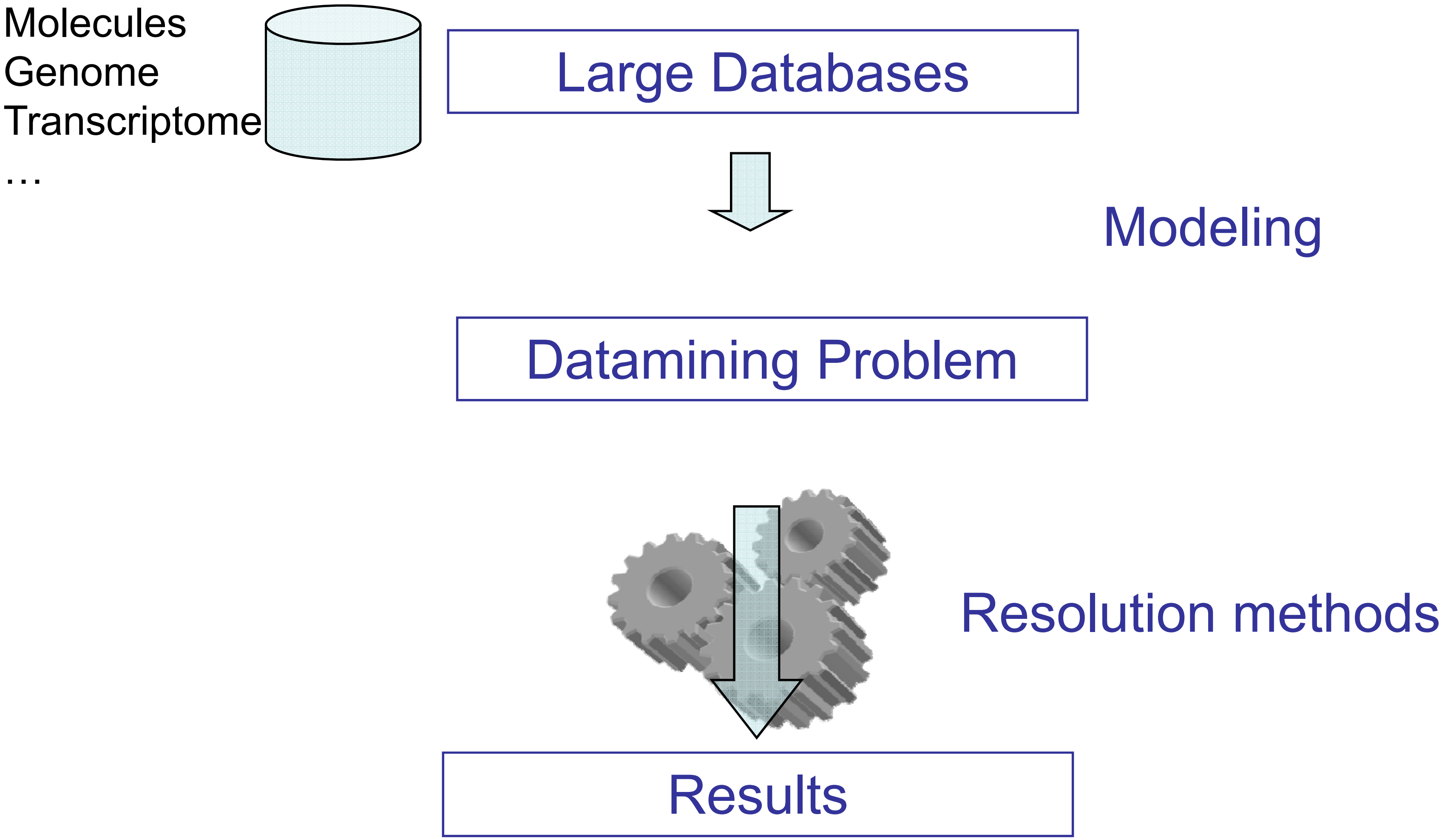
... A combinatorial optimization problem

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

5th IAPR International Conference on
Pattern Recognition in Bioinformatics
22–24 September 2010, Nijmegen, The Netherlands

Datamining in bioinformatics



Molecules
Genome
Transcriptome
…

Large Databases

Modeling

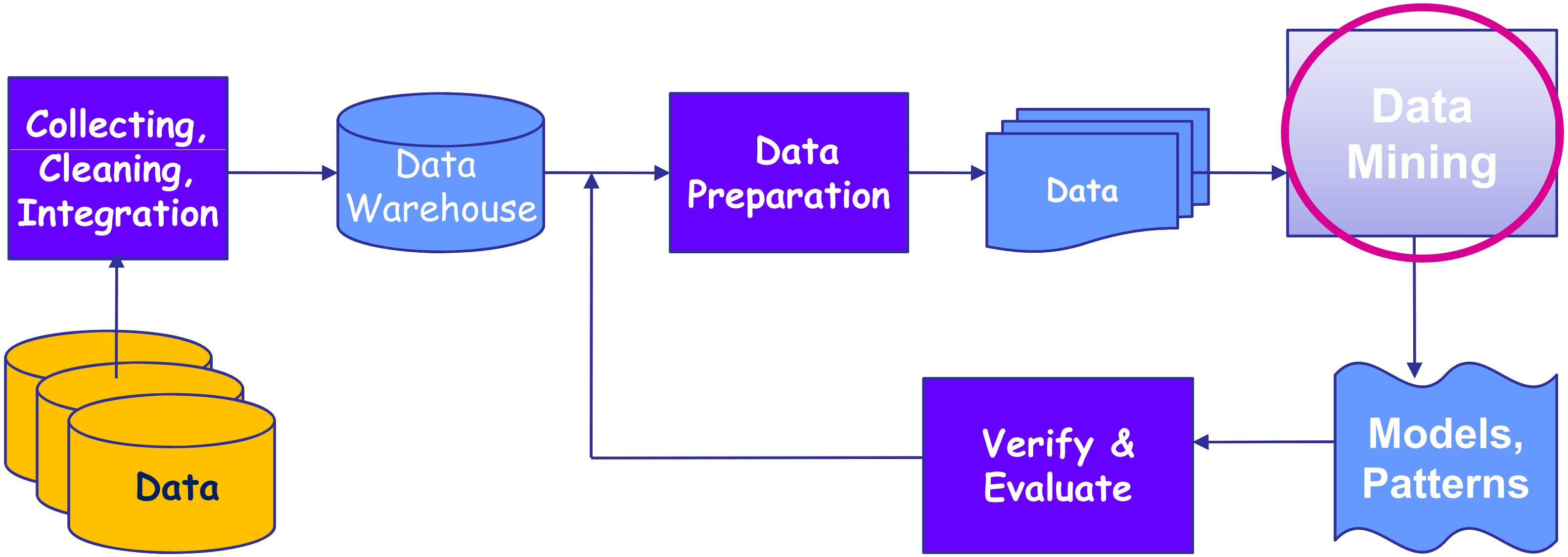
Datamining Problem

Resolution methods

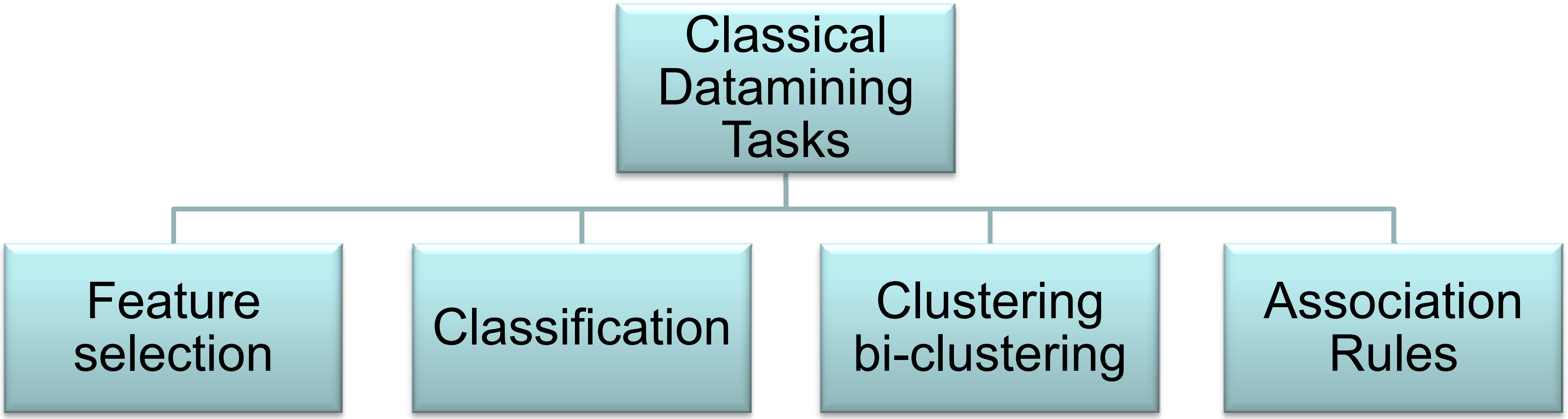
Results

Datamining / machine learning

- One step of the complex Knowledge Discovery in Databases (KDD) process

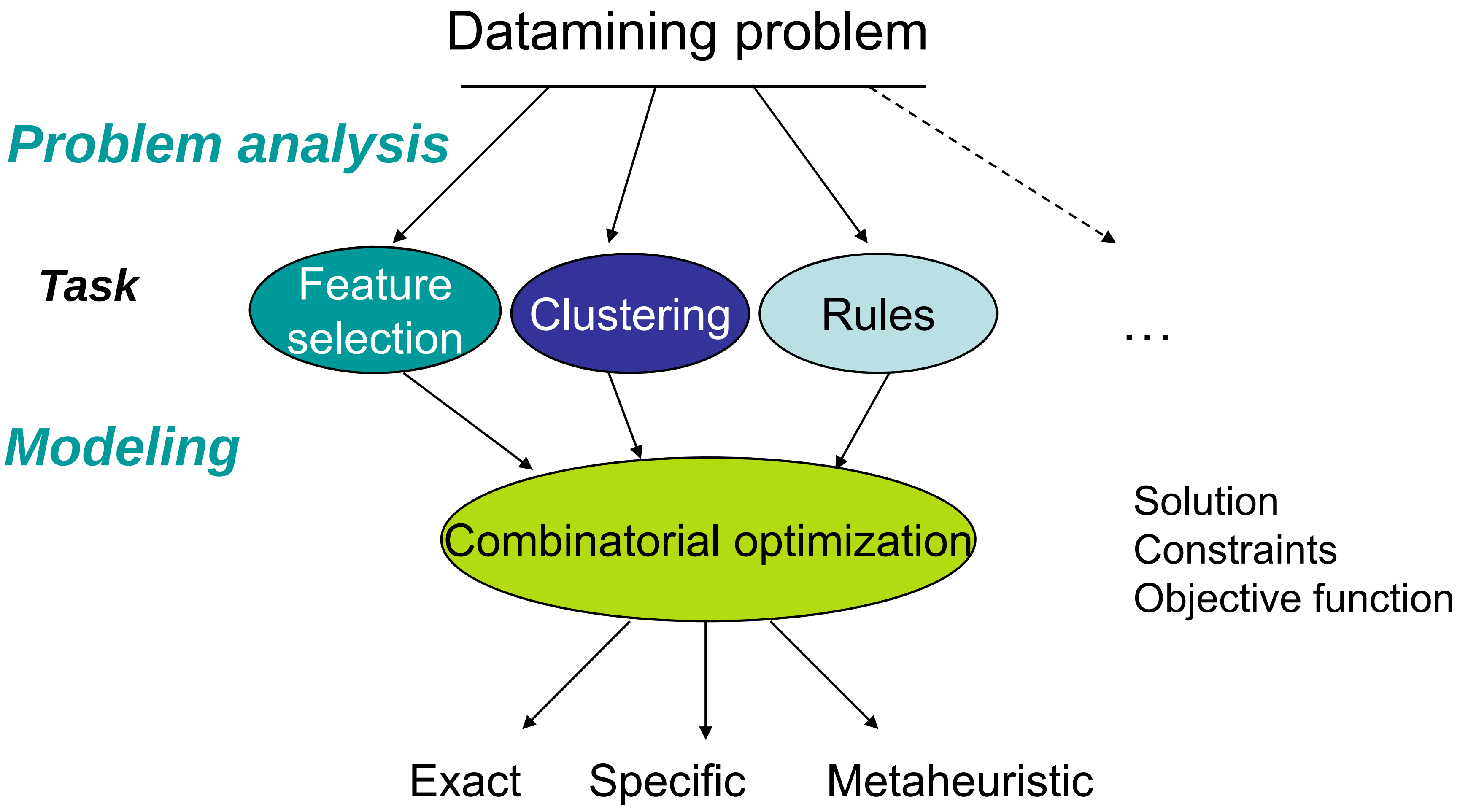


Datamining tasks



- Feature selection: to reduce the complexity of the problem
- Classification: supervised learning
- Clustering: unsupervised classification
- Association rules: represent relation between features

Strategy



Datamining problem

Problem analysis

Task

Feature selection

Clustering

Rules

…

Modeling

Combinatorial optimization

Solution
Constraints
Objective function

Exact

Specific

Metaheuristic

Feature selection for

CANCER DIAGNOSIS

Outline

- Context: Microarray experiment
- Feature selection presentation
- Methodology
- Results

Context: microarray technology

- Microarray experiments
  - Measure the gene expression levels of thousands of genes simultaneously
  - Allow to compare

    →several conditions: tissue, treatment or time point.

  Used to

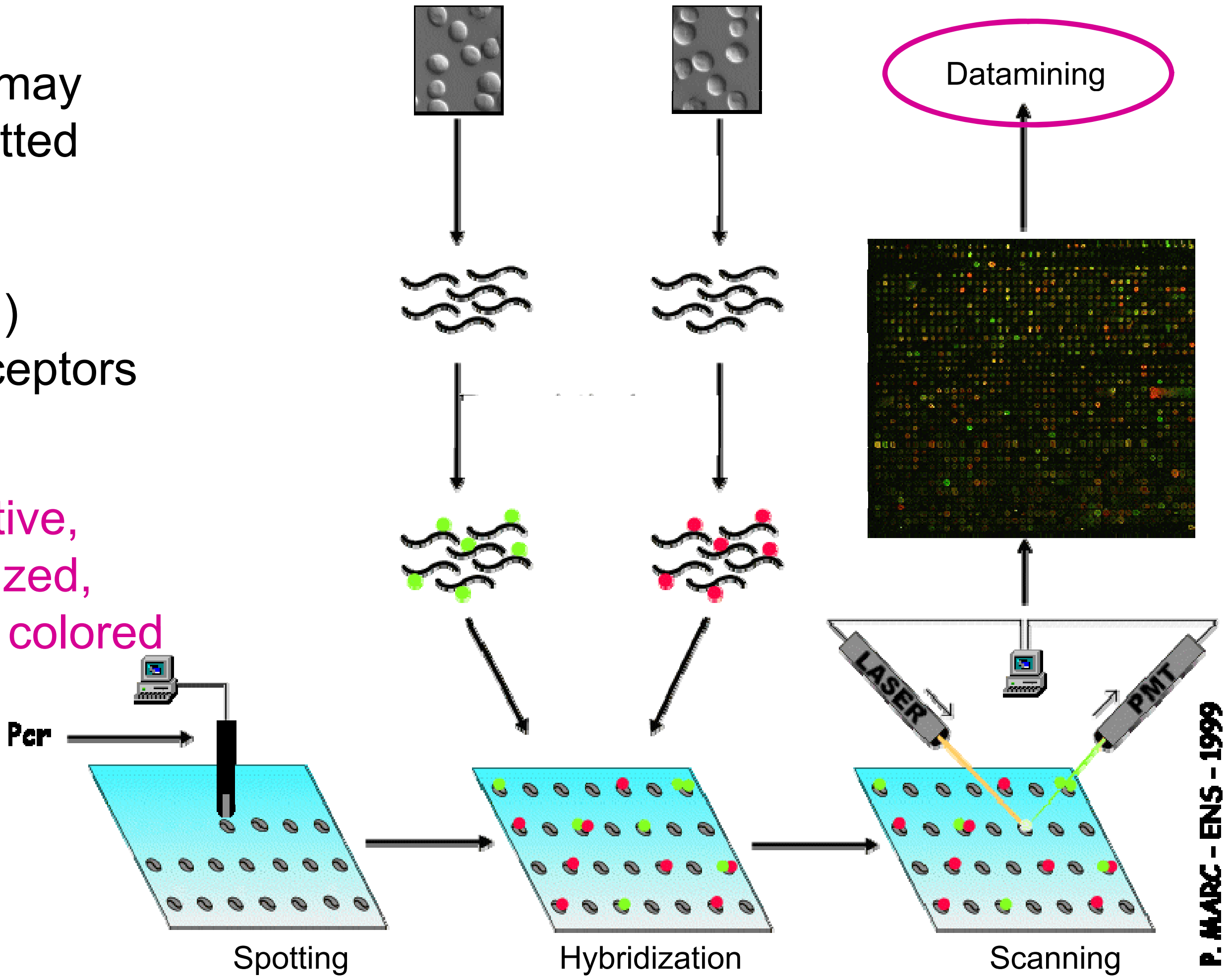
  - Identify genetic factors for some diseases (diabetes, obesity, coronary heart disease,…)
  - Identify function of some genes in genome

Context: microarray experiment

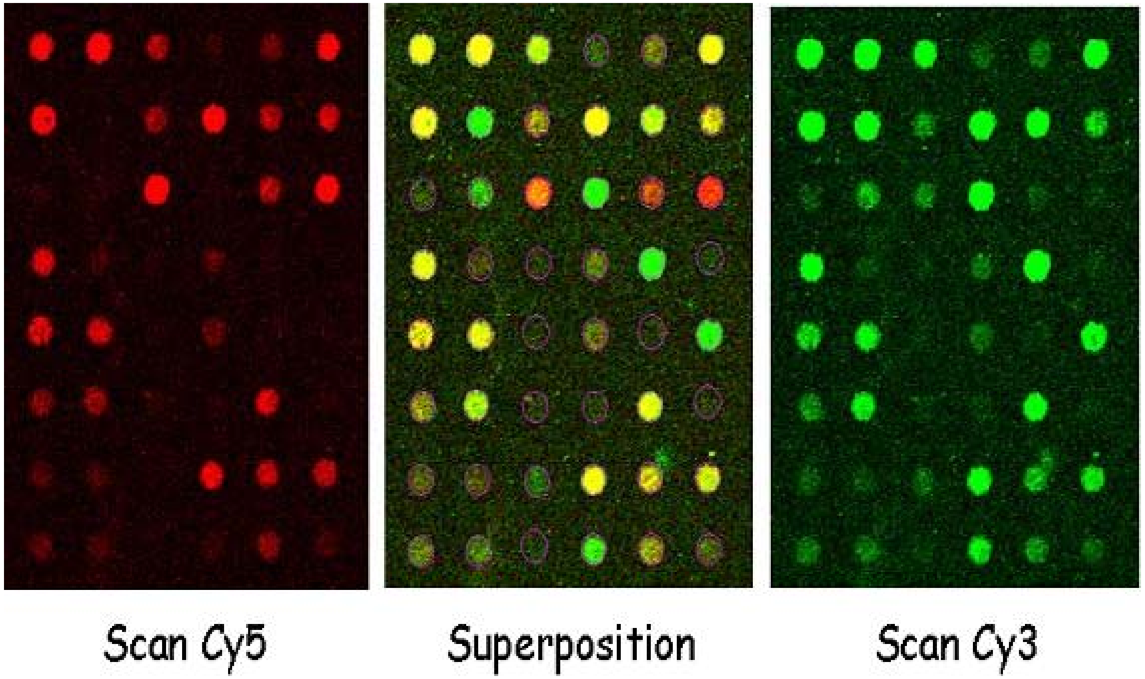
- Specific receptors that may recognize genes are spotted

- Extracts of DNA are:
    - Colored (green/red)
    - Hybridized with receptors

The more the gene is active, the more it will be hybridized, the more the spot will be colored



Context: microarray experiment



A result example :
colors indicate over/under expressed genes

Context: microarray data

- Different data matrices
  - Gene table
    - Rows: genes (G).
    - Columns: conditions (C)
  - Treatment table
    - Rows: Interactions (I).
    - Columns: genes (G).

$$\begin{array}{c|ccccc} & C1 & . & . & . & Cm \\ \hline G1 & . & & & & \\ . & & . & & & \\ . & & & . & & \\ . & & & & . & \\ Gn & & & & & . \end{array}$$

$$\begin{array}{c|cccccc} & G1 & . & . & . & Gm & \\ \hline I1 & . & & & & & V1 \\ . & & . & & & & . \\ . & & & . & & & . \\ . & & & & . & & . \\ In & & & & & . & Vn \end{array}$$

- Nature of the data

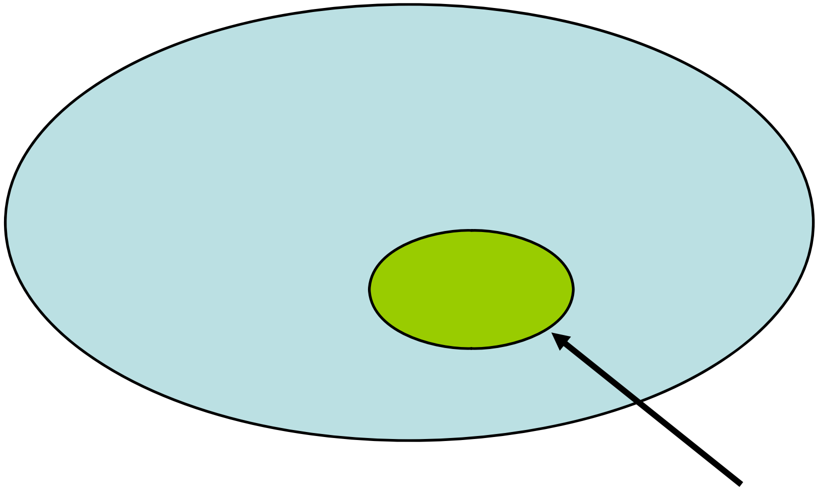
Activity of genes are represented as numerical values

➢ Discretize them into 5 values :

High Increase, Increase, No Change,

Decrease, High Decrease.

Feature selection



Large set of features (genes, products, …)

- Redundancy
- Noise

Subset of features
- Significant
- Improve classification

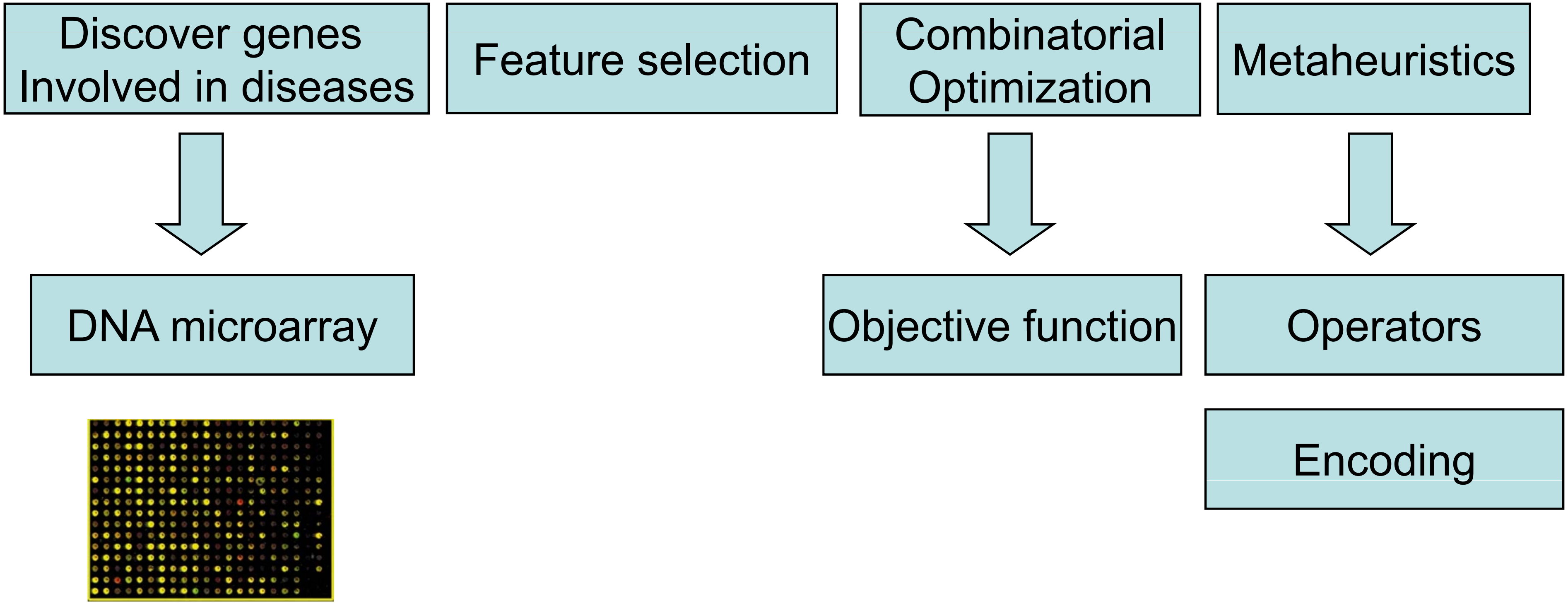
Strategy

Bioinformatics problem

Datamining task

Modeling

Solving



Discover genes Involved in diseases

DNA microarray

Feature selection

Combinatorial Optimization

Objective function

Metaheuristics

Operators

Encoding

Objectives

- Distinguish (Classify) tumor samples from normal ones (2 classes)
- Discover reduced subsets with informative genes, achieving high accuracies
- Classification with Support Vector Machines
- Algorithms comparisons.
  - 2 optimization algorithms (metaheuristics)
    - GPSO - Geometric Particle Swarm Optimizer
    - GA – Genetic Algorithm
- Experimentations using 6 public cancer datasets

E. Alba, J. Garcia-Nieto, L. Jourdan, E-G. Talbi. **Sensitivity and Specifity Based Multiobjective Approach for Feature Selection: Application to Cancer Diagnosis.** Information Processing Letters, Volume 109 (16), p 887-896, 2009

FS Methodology

Cancer Dataset

Genes

Expression Levels

G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	Classes
-214	-153	-58	88	-295	-558	199	-176	252	206	-41	-831	Normal
-139	-73	-1	283	-264	-400	-330	-168	101	74	19	-743	Normal
-76	-49	-307	309	-376	-650	33	-367	206	-215	19	-135	Tumor
-135	-114	-256	12	-419	-585	158	-253	49	31	363	-934	Normal
-106	-125	-76	168	-230	-284	4	-122	70	252	155	-471	Tumor
-138	-85	215	71	-272	-558	67	-186	87	193	325	-631	Tumor
-72	-144	238	55	-399	-551	131	-179	126	-20	-115	-103	Normal
-413	-260	7	-2	-541	-790	-275	-463	70	-169	-20	-143	Tumor

Accuracy 75%

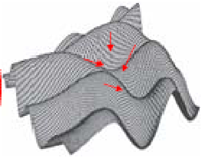
Number of features

75(4)

Fitness = 0.75 x 0.75 + 0.25 x 4

Alpha & Beta parameters

Provided by Metaheuristic GPSO/GA



Solution											
0	0	1	0	1	1	0	0	1	0	0	0

Feature Selection

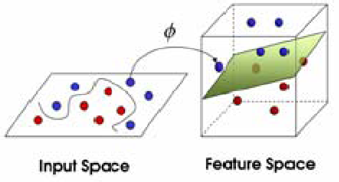
Cross Validation

4 Selected features

G3	G5	G6	G9	Classes
-58	-295	-558	252	Normal
-1	-264	-400	101	Normal
-307	-376	-650	206	Tumor
-256	-419	-585	49	Normal
-76	-230	-284	70	Tumor
215	-272	-558	87	Tumor
238	-399	-551	126	Normal
7	-541	-790	70	Tumor

Support Vector Machines

Training classifier



G3	G5	G6	G9
-58	-295	-558	252
-1	-264	-400	101
-307	-376	-650	206
-256	-419	-585	49
-76	-230	-284	70
215	-272	-558	87
238	-399	-551	126
7	-541	-790	70

prediction

Predicted Classes
Normal
Tumor
Tumor
Normal
Tumor
Tumor
Normal
Normal

Data Sets - Kent Ridge Bio-medical Data Set Repository

http://sdmc.lit.org.sg/GEDatasets/Datasets.html

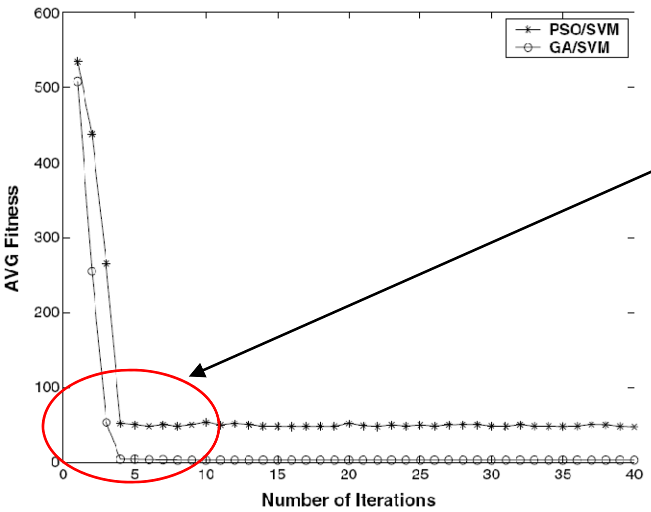
- ALL-AML Leukemia 7129 gene expression levels and 72 samples
- Breast Cancer 24481 gene expression levels and 97 samples
- Colon Tumor 2000 gene expression levels and 62 samples
- Lung Cancer 12533 gene expression levels and 181 samples
- Ovarian Cancer 15154 gene expression levels and 162 samples
- Prostate Cancer 12600 gene expression levels and 136 samples

Few samples / number of genes

Results

Performance Analysis: comparison of the two algorithms

Dataset	GPSO	GA	Huerta et al.	Juliusdotir et al.	Deb et al.	Guyon et al.	Yu et al.	Liu et al.	Shen et al.
Leukemia	97.38(3)	97.27(4)	**100(25)**	-	100(4)	**100(2)**	87.44(4)	-	-
Breast	86.35(4)	**95.86(4)**	-	-	-	-	79.38(67)	-	-
Colon	**100(2)**	**100(3)**	99.41(10)	94.12(37)	97(7)	98(4)	93.55(4)	85.48(-)	94(4)
Lung	99.00(4)	**99.49(4)**	-	-	-	-	98.34(6)	-	-
Ovarian	**99.44(4)**	98.83(4)	-	-	-	-	-	99.21(75)	-
Prostate	**98.66(4)**	98.65(4)	-	88.88(20)	-	-	-	-	-

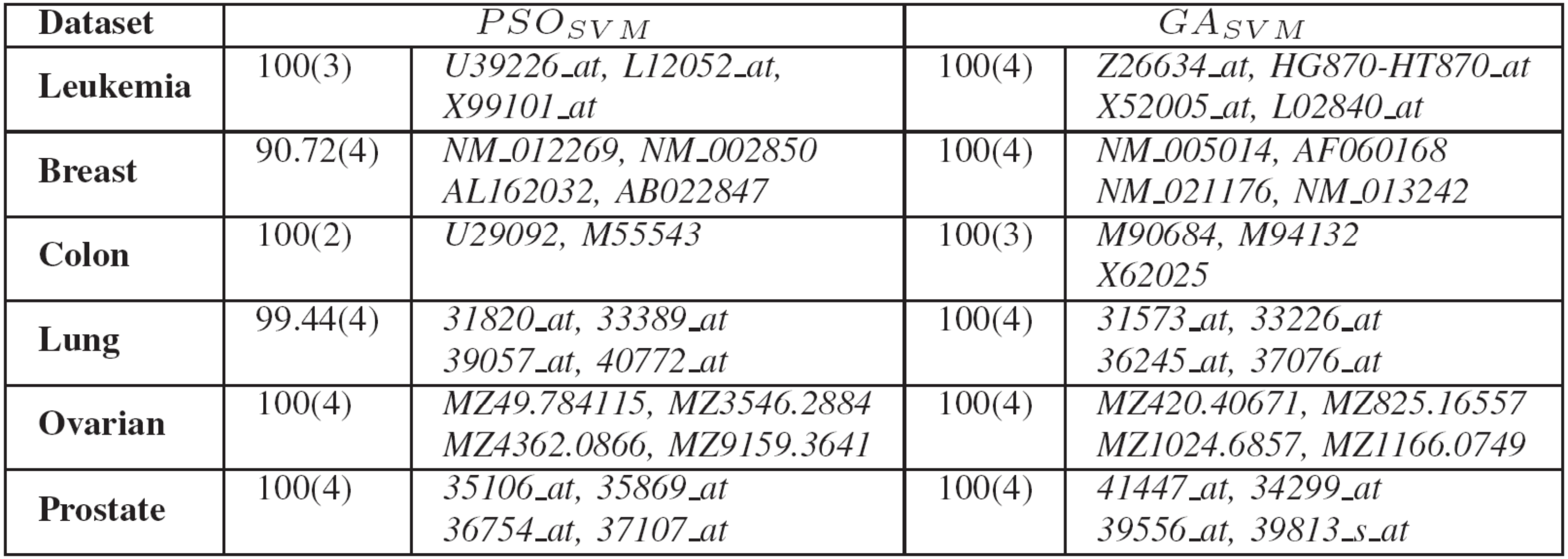


In few iterations the average of fitness Decrease quickly

GAsvm obtains generally lower average than GPSOsvm, whose solutions have in turn higher diversity

Results

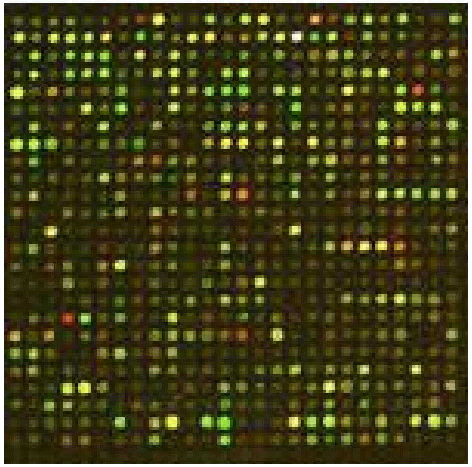
Examples of Selected Gene Subsets



Dataset	PSO_{SVM}		GA_{SVM}	
Leukemia	100(3)	*U39226_at, L12052_at, X99101_at*	100(4)	*Z26634_at, HG870-HT870_at X52005_at, L02840_at*
Breast	90.72(4)	*NM_012269, NM_002850 AL162032, AB022847*	100(4)	*NM_005014, AF060168 NM_021176, NM_013242*
Colon	100(2)	*U29092, M55543*	100(3)	*M90684, M94132 X62025*
Lung	99.44(4)	*31820_at, 33389_at 39057_at, 40772_at*	100(4)	*31573_at, 33226_at 36245_at, 37076_at*
Ovarian	100(4)	*MZ49.784115, MZ3546.2884 MZ4362.0866, MZ9159.3641*	100(4)	*MZ420.40671, MZ825.16557 MZ1024.6857, MZ1166.0749*
Prostate	100(4)	*35106_at, 35869_at 36754_at, 37107_at*	100(4)	*41447_at, 34299_at 39556_at, 39813_s_at*

Conclusions

- Two hybrid algorithms for gene selection and classification of high dimensional DNA Microarray were presented
- New algorithm GPSO for feature selection was applied
- GPSOsvm vs. GAsvm were experimentally assessed on six well-known datasets
- Results of 100% accuracy and few genes per subset (3 and 4)
- Use of adapted initialization method
- Use of adapted operators for FS (3PMBCX & SSOCF)



Association rules for

DNA MICROARRAY

Context: available data

	Gene 1	Gene 2	…	…	Gene 22000
Patient 1 / Control 1					
Patient 1 / Control 2					
…					
Patient 15 / Control 3					

Look for subsets of genes having linked comportments

➔ Association rules

A general approach

Expression data often analyzed thanks to classification/clustering.

But 3 main drawbacks:

1. One gene participating to one relation will be classified in a single group
2. Difficulty to point out relations between genes belonging to a same group
3. Classification made according to the whole set of experiments

Association rules may overcome these drawbacks

Association Rules: Definition

Goal : Discover patterns, associations between items (columns=attributes) of a database.

Form : *if* C *then* P

C = term_1 and term_2 and… and term_n

P = term_{n+1}

term_i = <attribute_j, *op*, value>

Ass. Rules: examples of results

Association rules may produce different results

- Situation ⇒ Expression of particular genes

  Situation x ⇒ {Gene A ↑, Gene B ↓} [Creighton - Hanash, 03]

- Relations between genes (general case)

  {Gene A ↑, Gene B ↓, Gene C ↑} ⇒ Gene D ↑ [Kotala et al, 01]

- Relations between genes (for some situations) [Becquet et al, 02]

  {(Gene A ↑, Gene B ↑) in situation y} ⇒ Gene D ↑ in situation y

- Comportment of genes ⇒ Functional characteristics

  ⇒ Structural characteristics

  {Genes ↑ in situation y} ⇒ Function x [OPAC, IT-Omics, 03]

Ass. Rules: optimization criterion?

Association rules

Classical problem of datamining

Studied by statistic, machine learning, combinatorial optimization,… communities

⇒ a lot of indicators proposed to measure rules quality

[Hilderman et Hamilton, 1999], [Tan et Kumar, 2002], [Adomavicius, 2002], [Lenca et al, 2003],…

How to choose a good indicator?
No universal criterion

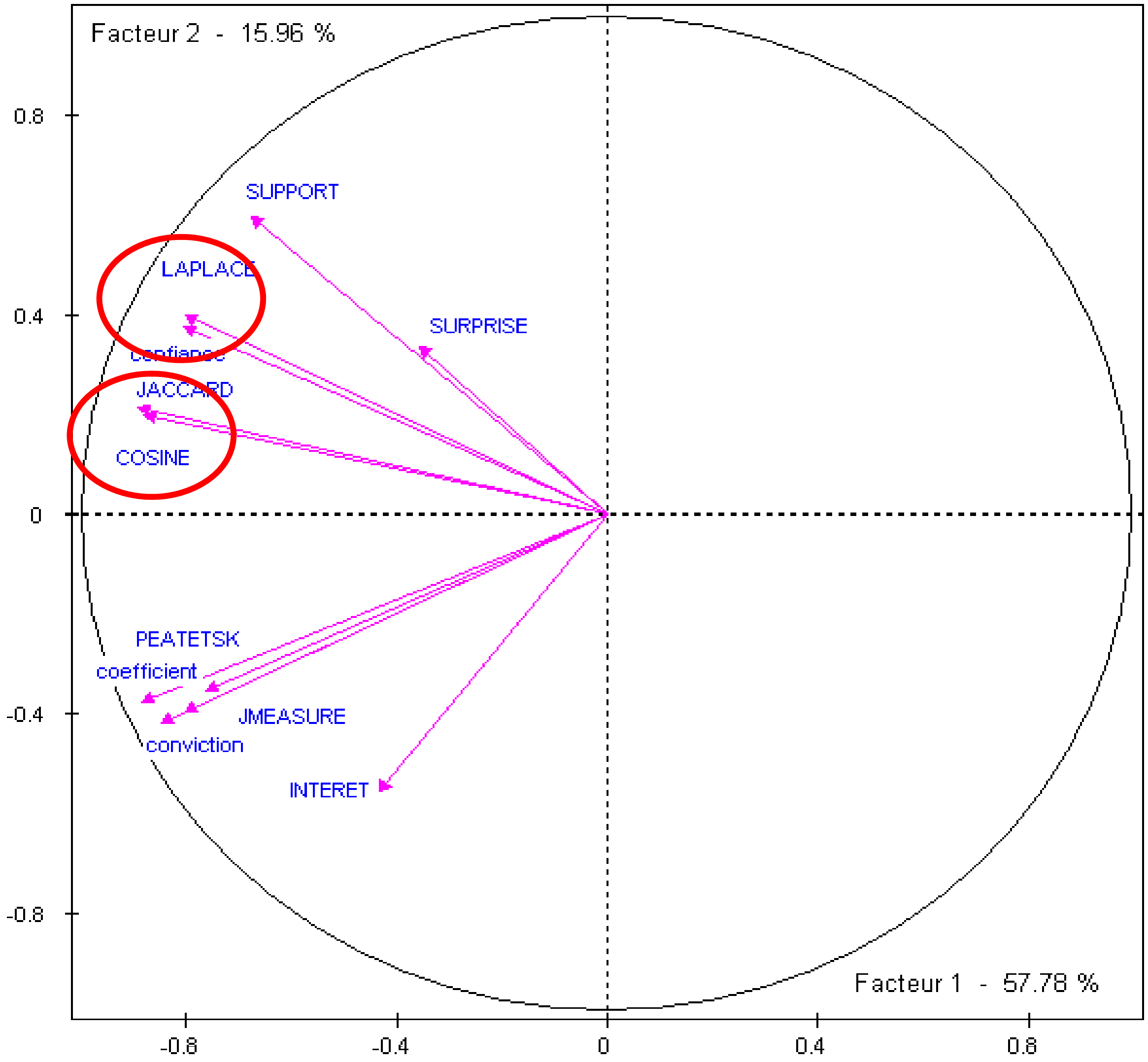
Ass. Rules: examples of criteria

Criteria	Math. definition	Explanation
Support S	$\frac{C \text{ and } P}{N}$	% lines having C and P
Confidence C	$\frac{C \text{ and } P}{C}$	Conditional probability
Interest I	$\frac{C \text{ and } P}{C \times P}$	Favors rare pattern (small support)
Conviction V	$\frac{C \times \overline{P}}{C \text{ and } \overline{P}}$	Measures the weakness of (C, not P)
Piatetsky-Shapiro's PS	$C \text{ and } P - C \times P$	Measure dependency
Surprise R	$\frac{\left(C \text{ and } P - C \text{ and } \overline{P}\right)}{P}$	Look for surprising rules

Statistical analysis: Matrix of linear correlations

	Supp	Conf	Inte	Conv	Surp	Jacc	PhiC	Cos	JMea	Piat	Lapl
Supp	1,00										
Conf	0,62	1,00									
Inte	-0,09	0,20	1,00								
Conv	0,27	0,56	0,47	1,00							
Surp	0,17	0,48	0,07	0,17	1,00						
Jacc	**0,87**	0,62	0,32	0,55	0,20	1,00					
PhiC	0,38	0,50	0,62	0,81	0,26	0,76	1,00				
Cos	**0,86**	0,68	0,34	0,56	0,19	**0,98**	0,76	1,00			
JMea	0,34	0,50	0,40	0,84	0,15	0,64	**0,89**	0,62	1,00		
Piat	0,29	0,49	0,25	0,71	0,15	0,51	0,75	0,51	**0,93**	1,00	
Lapl	0,63	**0,99**	0,18	0,54	0,53	0,61	0,49	0,67	0,50	0,51	1,00

Correlations between criteria



Statistical analysis: results

5 clusters

C1 : Support, Cosine, Jaccard

C2 : Laplace, Confidence

C3 : Phi-Coefficient, Jmeasure, Piatetsky, Conviction

C4 : Interest

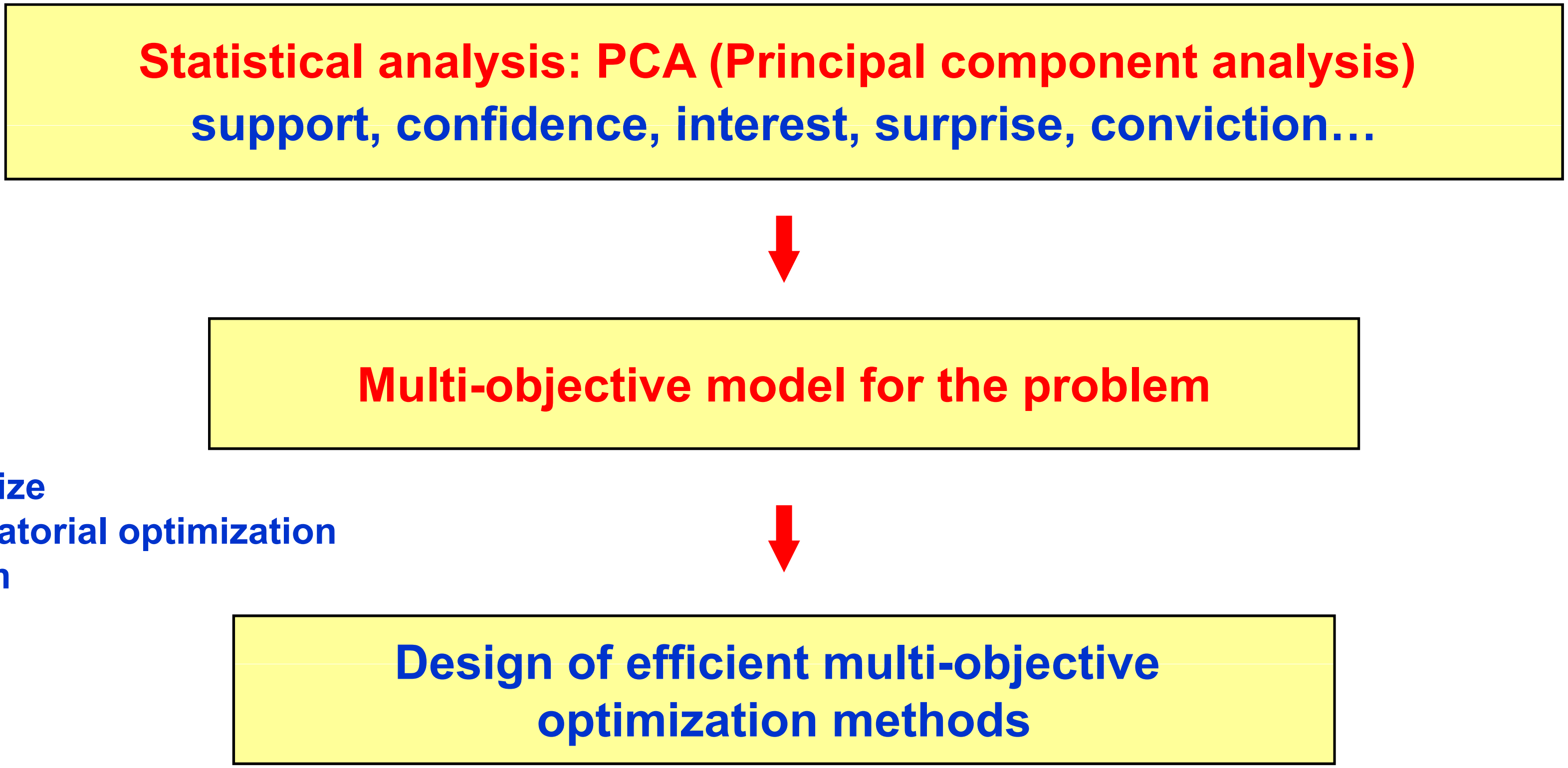
C5 : Surprise

Each cluster groups similar criteria ⇒ Redundant

Choose one criterion per cluster

➔ Optimization with five objectives

General methodology



Statistical analysis: PCA (Principal component analysis)
support, confidence, interest, surprise, conviction…

Multi-objective model for the problem

Design of efficient multi-objective optimization methods

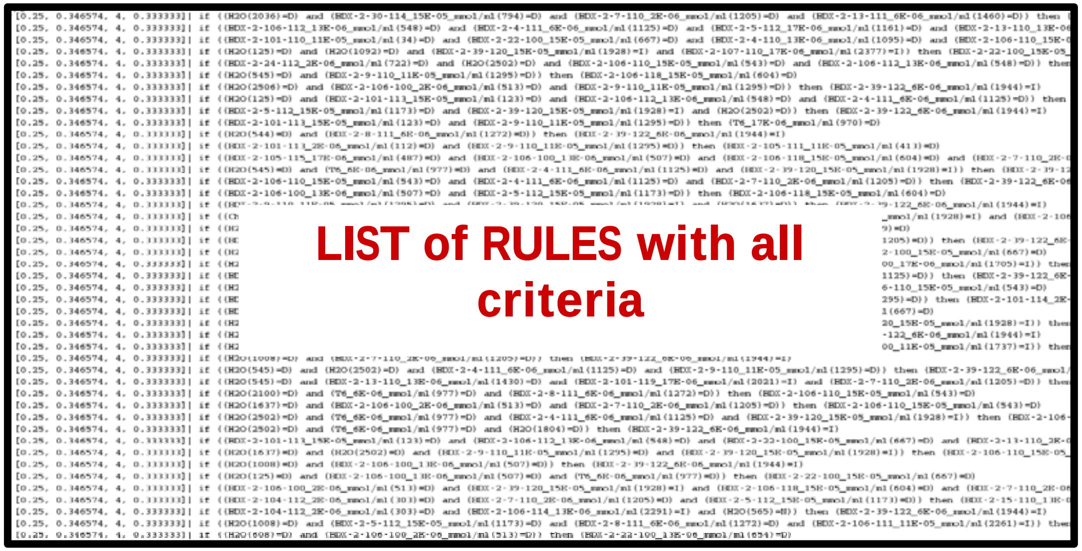
Large size combinatorial optimization problem

Rulemining: Application overview (1/2)

DATAMINING :
Rulemining

Set the genetic algorithm:

- **Number of generations**
- **Population size**
- **Enable/Disable Support, Confidence, J-measure, Interest, Surprise Criteria**
- **…**

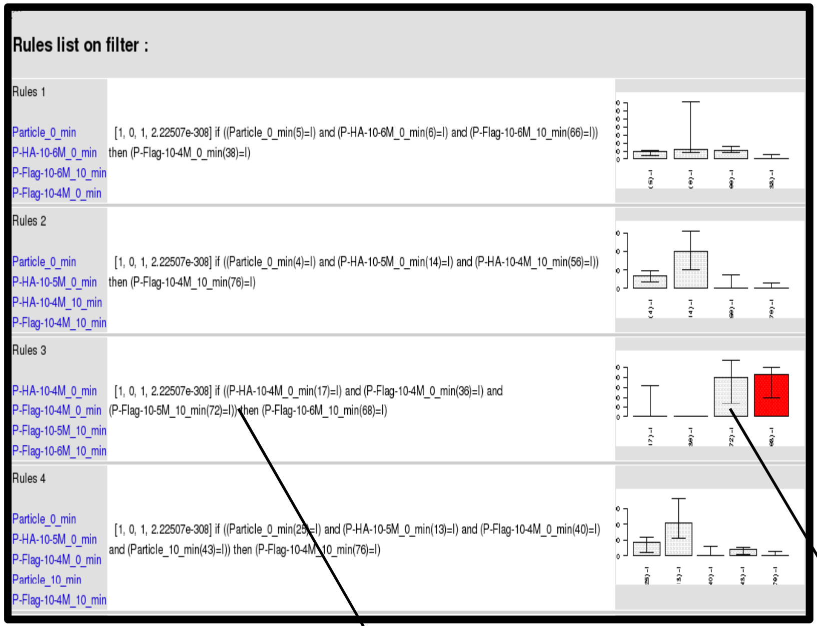


LIST of RULES with all criteria

```
[0.25, 0.346574, 4, 0.333333]| if ((H2O(2036)=D) and (BDX-2-30-114_15E-05_mmol/ml(794)=D) and (BDX-2-7-110_2E-06_mmol/ml(1205)=D) and (BDX-2-13-111_6E-06_mmol/ml(1460)=D)) then (
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-106-112_13E-06_mmol/ml(548)=D) and (BDX-2-4-111_6E-06_mmol/ml(1125)=D) and (BDX-2-5-112_17E-06_mmol/ml(1161)=D) and (BDX-2-13-110_13E-06
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-101-110_11E-05_mmol/ml(34)=D) and (BDX-2-22-100_15E-05_mmol/ml(667)=D) and (BDX-2-4-110_13E-06_mmol/ml(1095)=D) and (BDX-2-106-110_15E-0
[0.25, 0.346574, 4, 0.333333]| if ((H2O(125)=D) and (H2O(1092)=D) and (BDX-2-39-120_15E-05_mmol/ml(1928)=I) and (BDX-2-107-110_17E-06_mmol/ml(2377)=I)) then (BDX-2-22-100_15E-05_
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-24-112_2E-06_mmol/ml(722)=D) and (H2O(2502)=D) and (BDX-2-106-110_15E-05_mmol/ml(543)=D) and (BDX-2-106-112_13E-06_mmol/ml(548)=D)) then
[0.25, 0.346574, 4, 0.333333]| if ((H2O(545)=D) and (BDX-2-9-110_11E-05_mmol/ml(1295)=D)) then (BDX-2-106-118_15E-05_mmol/ml(604)=D)
[0.25, 0.346574, 4, 0.333333]| if ((H2O(2506)=D) and (BDX-2-106-100_2E-06_mmol/ml(513)=D) and (BDX-2-9-110_11E-05_mmol/ml(1295)=D)) then (BDX-2-39-122_6E-06_mmol/ml(1944)=I)
[0.25, 0.346574, 4, 0.333333]| if ((H2O(125)=D) and (BDX-2-101-113_15E-05_mmol/ml(123)=D) and (BDX-2-106-112_13E-06_mmol/ml(548)=D) and (BDX-2-4-111_6E-06_mmol/ml(1125)=D)) then
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-5-112_15E-05_mmol/ml(1173)=D) and (BDX-2-39-120_15E-05_mmol/ml(1928)=I) and (H2O(2502)=D)) then (BDX-2-39-122_6E-06_mmol/ml(1944)=I)
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-101-113_15E-05_mmol/ml(123)=D) and (BDX-2-9-110_11E-05_mmol/ml(1295)=D)) then (T6_17E-06_mmol/ml(970)=D)
[0.25, 0.346574, 4, 0.333333]| if ((H2O(544)=D) and (BDX-2-8-111_6E-06_mmol/ml(1272)=D)) then (BDX-2-39-122_6E-06_mmol/ml(1944)=I)
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-101-113_2E-06_mmol/ml(112)=D) and (BDX-2-9-110_11E-05_mmol/ml(1295)=D)) then (BDX-2-105-111_11E-05_mmol/ml(413)=D)
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-105-115_17E-06_mmol/ml(487)=D) and (BDX-2-106-100_13E-06_mmol/ml(507)=D) and (BDX-2-106-118_15E-05_mmol/ml(604)=D) and (BDX-2-7-110_2E-0
[0.25, 0.346574, 4, 0.333333]| if ((H2O(545)=D) and (T6_6E-06_mmol/ml(977)=D) and (BDX-2-4-111_6E-06_mmol/ml(1125)=D) and (BDX-2-39-120_15E-05_mmol/ml(1928)=I)) then (BDX-2-39-12
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-106-110_15E-05_mmol/ml(543)=D) and (BDX-2-4-111_6E-06_mmol/ml(1125)=D) and (BDX-2-7-110_2E-06_mmol/ml(1205)=D)) then (BDX-2-39-122_6E-06
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-106-100_13E-06_mmol/ml(507)=D) and (BDX-2-5-112_15E-05_mmol/ml(1173)=D)) then (BDX-2-106-118_15E-05_mmol/ml(604)=D)
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-9-110_11E-05_mmol/ml(1295)=D) and (BDX-2-39-120_15E-05_mmol/ml(1928)=I) and (H2O(1637)=D)) then (BDX-2-39-122_6E-06_mmol/ml(1944)=I)
[0.25, 0.346574, 4, 0.333333]| if ((Ch                                                                  _mmol/ml(1928)=I) and (BDX-2-106
[0.25, 0.346574, 4, 0.333333]| if ((H2                                                                  9)=D)
[0.25, 0.346574, 4, 0.333333]| if ((BE                                                                  1205)=D)) then (BDX-2-39-122_6E-
[0.25, 0.346574, 4, 0.333333]| if ((H2                                                                  2-100_15E-05_mmol/ml(667)=D)
[0.25, 0.346574, 4, 0.333333]| if ((H2                                                                  00_17E-06_mmol/ml(1705)=I)) then
[0.25, 0.346574, 4, 0.333333]| if ((BD                                                                  1125)=D)) then (BDX-2-39-122_6E-
[0.25, 0.346574, 4, 0.333333]| if ((H2                                                                  6-110_15E-05_mmol/ml(543)=D)
[0.25, 0.346574, 4, 0.333333]| if ((BD                                                                  295)=D)) then (BDX-2-101-114_2E-
[0.25, 0.346574, 4, 0.333333]| if ((BD                                                                  1(667)=D)
[0.25, 0.346574, 4, 0.333333]| if ((H2                                                                  20_15E-05_mmol/ml(1928)=I)) then
[0.25, 0.346574, 4, 0.333333]| if ((H2                                                                  -122_6E-06_mmol/ml(1944)=I)
[0.25, 0.346574, 4, 0.333333]| if ((H2                                                                  00_11E-05_mmol/ml(1737)=I)) then
[0.25, 0.346574, 4, 0.333333]| if ((H2O(1008)=D) and (BDX-2-7-110_2E-06_mmol/ml(1205)=D)) then (BDX-2-39-122_6E-06_mmol/ml(1944)=I)
[0.25, 0.346574, 4, 0.333333]| if ((H2O(545)=D) and (H2O(2502)=D) and (BDX-2-4-111_6E-06_mmol/ml(1125)=D) and (BDX-2-9-110_11E-05_mmol/ml(1295)=D)) then (BDX-2-39-122_6E-06_mmol/
[0.25, 0.346574, 4, 0.333333]| if ((H2O(545)=D) and (BDX-2-13-110_13E-06_mmol/ml(1430)=D) and (BDX-2-101-119_17E-06_mmol/ml(2021)=I) and (BDX-2-7-110_2E-06_mmol/ml(1205)=D)) then
[0.25, 0.346574, 4, 0.333333]| if ((H2O(2100)=D) and (T6_6E-06_mmol/ml(977)=D) and (BDX-2-8-111_6E-06_mmol/ml(1272)=D)) then (BDX-2-106-110_15E-05_mmol/ml(543)=D)
[0.25, 0.346574, 4, 0.333333]| if ((H2O(1637)=D) and (BDX-2-106-100_2E-06_mmol/ml(513)=D) and (BDX-2-7-110_2E-06_mmol/ml(1205)=D)) then (BDX-2-106-110_15E-05_mmol/ml(543)=D)
[0.25, 0.346574, 4, 0.333333]| if ((H2O(2502)=D) and (T6_6E-06_mmol/ml(977)=D) and (BDX-2-4-111_6E-06_mmol/ml(1125)=D) and (BDX-2-39-120_15E-05_mmol/ml(1928)=I)) then (BDX-2-106-
[0.25, 0.346574, 4, 0.333333]| if ((H2O(2502)=D) and (T6_6E-06_mmol/ml(977)=D) and (H2O(1804)=D)) then (BDX-2-39-122_6E-06_mmol/ml(1944)=I)
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-101-113_15E-05_mmol/ml(123)=D) and (BDX-2-106-112_13E-06_mmol/ml(548)=D) and (BDX-2-22-100_15E-05_mmol/ml(667)=D) and (BDX-2-13-110_2E-0
[0.25, 0.346574, 4, 0.333333]| if ((H2O(1637)=D) and (H2O(2502)=D) and (BDX-2-9-110_11E-05_mmol/ml(1295)=D) and (BDX-2-39-120_15E-05_mmol/ml(1928)=I)) then (BDX-2-106-110_15E-05_
[0.25, 0.346574, 4, 0.333333]| if ((H2O(1008)=D) and (BDX-2-106-100_13E-06_mmol/ml(507)=D)) then (BDX-2-39-122_6E-06_mmol/ml(1944)=I)
[0.25, 0.346574, 4, 0.333333]| if ((H2O(125)=D) and (BDX-2-106-100_13E-06_mmol/ml(507)=D) and (T6_6E-06_mmol/ml(977)=D)) then (BDX-2-22-100_15E-05_mmol/ml(667)=D)
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-106-100_2E-06_mmol/ml(513)=D) and (BDX-2-39-120_15E-05_mmol/ml(1928)=I) and (BDX-2-106-118_15E-05_mmol/ml(604)=D) and (BDX-2-7-110_2E-06
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-104-112_2E-06_mmol/ml(303)=D) and (BDX-2-7-110_2E-06_mmol/ml(1205)=D) and (BDX-2-5-112_15E-05_mmol/ml(1173)=D)) then (BDX-2-15-110_13E-0
[0.25, 0.346574, 4, 0.333333]| if ((BDX-2-104-112_2E-06_mmol/ml(303)=D) and (BDX-2-106-114_13E-06_mmol/ml(2291)=I) and (H2O(565)=D)) then (BDX-2-39-122_6E-06_mmol/ml(1944)=I)
[0.25, 0.346574, 4, 0.333333]| if ((H2O(1008)=D) and (BDX-2-5-112_15E-05_mmol/ml(1173)=D) and (BDX-2-8-111_6E-06_mmol/ml(1272)=D) and (BDX-2-106-111_11E-05_mmol/ml(2261)=I)) then
[0.25, 0.346574, 4, 0.333333]| if ((H2O(698)=D) and (BDX-2-106-100_2E-06_mmol/ml(513)=D)) then (BDX-2-22-100_13E-06_mmol/ml(654)=D)
```

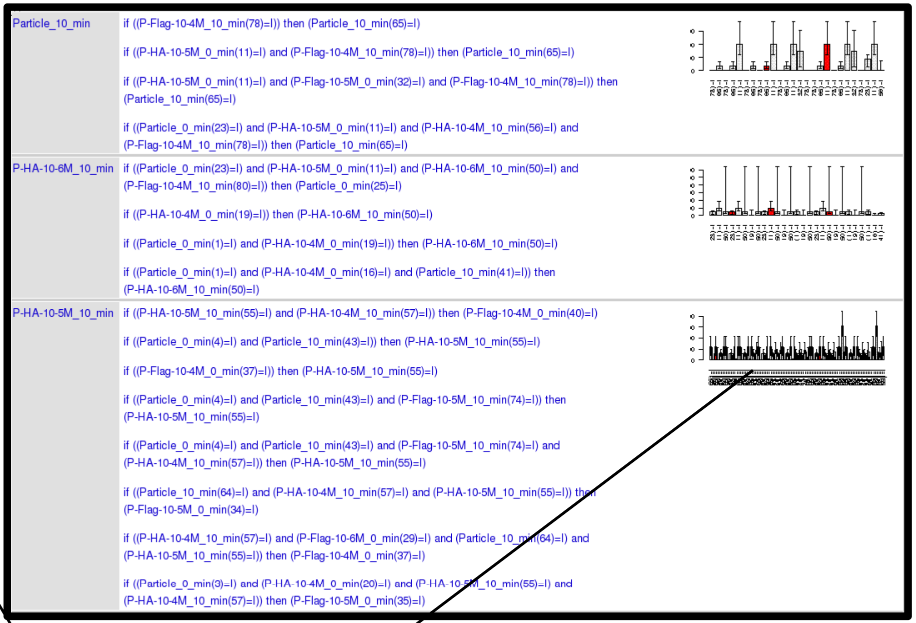
Rulemining: Application overview (2/2)

RESULTS



Rules list on filter :

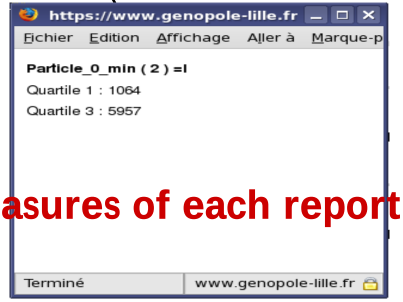
Rules 1 Particle_0_min P-HA-10-6M_0_min P-Flag-10-6M_10_min P-Flag-10-4M_0_min	[1, 0, 1, 2.22507e-308] if ((Particle_0_min(5)=I) and (P-HA-10-6M_0_min(6)=I) and (P-Flag-10-6M_10_min(66)=I)) then (P-Flag-10-4M_0_min(38)=I)	
Rules 2 Particle_0_min P-HA-10-5M_0_min P-HA-10-4M_10_min P-Flag-10-4M_10_min	[1, 0, 1, 2.22507e-308] if ((Particle_0_min(4)=I) and (P-HA-10-5M_0_min(14)=I) and (P-HA-10-4M_10_min(56)=I)) then (P-Flag-10-4M_10_min(76)=I)	
Rules 3 P-HA-10-4M_0_min P-Flag-10-4M_0_min P-Flag-10-5M_10_min P-Flag-10-6M_10_min	[1, 0, 1, 2.22507e-308] if ((P-HA-10-4M_0_min(17)=I) and (P-Flag-10-4M_0_min(36)=I) and (P-Flag-10-5M_10_min(72)=I)) then (P-Flag-10-6M_10_min(68)=I)	
Rules 4 Particle_0_min P-HA-10-5M_0_min P-Flag-10-4M_0_min Particle_10_min P-Flag-10-4M_10_min	[1, 0, 1, 2.22507e-308] if ((Particle_0_min(25)=I) and (P-HA-10-5M_0_min(13)=I) and (P-Flag-10-4M_0_min(40)=I) and (Particle_10_min(43)=I)) then (P-Flag-10-4M_10_min(76)=I)	



Particle_10_min	if ((P-Flag-10-4M_10_min(78)=I)) then (Particle_10_min(65)=I)
	if ((P-HA-10-5M_0_min(11)=I) and (P-Flag-10-4M_10_min(78)=I)) then (Particle_10_min(65)=I)
	if ((P-HA-10-5M_0_min(11)=I) and (P-Flag-10-5M_0_min(32)=I) and (P-Flag-10-4M_10_min(78)=I)) then (Particle_10_min(65)=I)
	if ((Particle_0_min(23)=I) and (P-HA-10-5M_0_min(11)=I) and (P-HA-10-4M_10_min(56)=I) and (P-Flag-10-4M_10_min(78)=I)) then (Particle_10_min(65)=I)	

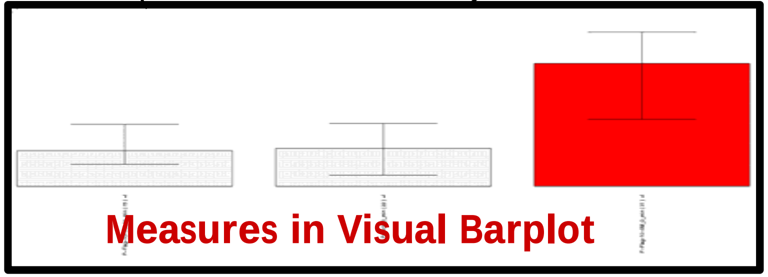
P-HA-10-6M_10_min	if ((Particle_0_min(23)=I) and (P-HA-10-5M_0_min(11)=I) and (P-HA-10-6M_10_min(50)=I) and (P-Flag-10-4M_10_min(80)=I)) then (Particle_0_min(25)=I)
	if ((P-HA-10-4M_0_min(19)=I)) then (P-HA-10-6M_10_min(50)=I)
	if ((Particle_0_min(1)=I) and (P-HA-10-4M_0_min(19)=I)) then (P-HA-10-6M_10_min(50)=I)
	if ((Particle_0_min(1)=I) and (P-HA-10-4M_0_min(16)=I) and (Particle_10_min(41)=I)) then (P-HA-10-6M_10_min(50)=I)	

P-HA-10-5M_10_min	if ((P-HA-10-5M_10_min(55)=I) and (P-HA-10-4M_10_min(57)=I)) then (P-Flag-10-4M_0_min(40)=I)
	if ((Particle_0_min(4)=I) and (Particle_10_min(43)=I)) then (P-HA-10-5M_10_min(55)=I)
	if ((P-Flag-10-4M_0_min(37)=I)) then (P-HA-10-5M_10_min(55)=I)
	if ((Particle_0_min(4)=I) and (Particle_10_min(43)=I) and (P-Flag-10-5M_10_min(74)=I)) then (P-HA-10-5M_10_min(55)=I)
	if ((Particle_0_min(4)=I) and (Particle_10_min(43)=I) and (P-Flag-10-5M_10_min(74)=I) and (P-HA-10-4M_10_min(57)=I)) then (P-HA-10-5M_10_min(55)=I)
	if ((Particle_10_min(64)=I) and (P-HA-10-4M_10_min(57)=I) and (P-HA-10-5M_10_min(55)=I)) then (P-Flag-10-5M_0_min(34)=I)
	if ((P-HA-10-4M_10_min(57)=I) and (P-Flag-10-6M_0_min(29)=I) and (Particle_10_min(64)=I) and (P-HA-10-5M_10_min(55)=I)) then (P-Flag-10-4M_0_min(37)=I)
	if ((Particle_0_min(3)=I) and (P-HA-10-4M_0_min(20)=I) and (P-HA-10-5M_10_min(55)=I) and (P-HA-10-4M_10_min(57)=I)) then (P-Flag-10-5M_0_min(35)=I)	



Particle_0_min (2) =I
Quartile 1 : 1064
Quartile 3 : 5957

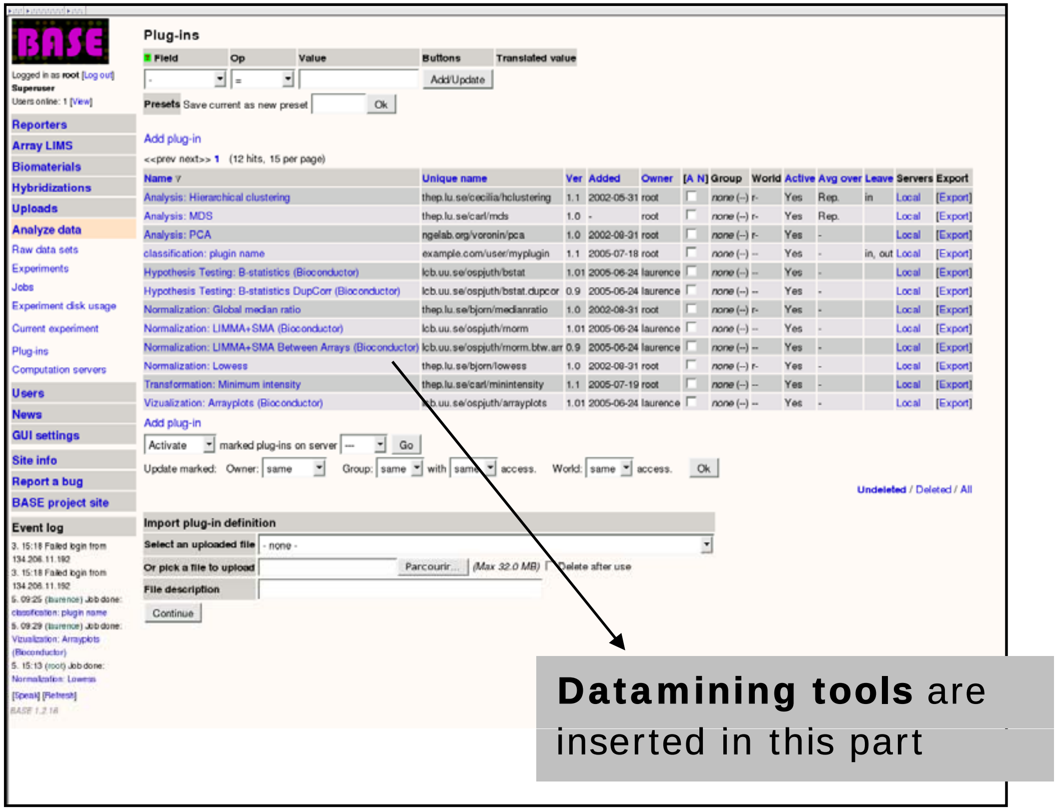
Measures of each reporters



Measures in Visual Barplot

Validation in BASE: Plugins

BioArray Software Environment. BASE is a comprehensive free web-based database solution for the massive amounts of data generated by microarray



BASE

Logged in as **root** [Log out]
Superuser
Users online: 1 [View]

- Reporters
- Array LIMS
- Biomaterials
- Hybridizations
- Uploads
- Analyze data
- Raw data sets
- Experiments
- Jobs
- Experiment disk usage
- Current experiment
- Plug-ins
- Computation servers
- Users
- News
- GUI settings
- Site info
- Report a bug
- BASE project site

Event log

3. 15:18 Failed login from 134.206.11.192
3. 15:18 Failed login from 134.206.11.192
5. 09:25 (laurence) Job done: classification: plugin name
5. 09:29 (laurence) Job done: Vizualization: Arrayplots (Bioconductor)
5. 15:13 (root) Job done: Normalization: Lowess

[Speak] [Refresh]

BASE 1.2.16

Plug-ins

Field	Op	Value	Buttons	Translated value
-	=		Add/Update	

Presets Save current as new preset [Ok]

Add plug-in

<<prev next>> 1 (12 hits, 15 per page)

Name ▽	Unique name	Ver	Added	Owner	[A N]	Group	World	Active	Avg over	Leave	Servers	Export
Analysis: Hierarchical clustering	thep.lu.se/cecilia/hclustering	1.1	2002-05-31	root	□	*none* (--)	r-	Yes	Rep.	in	Local	[Export]
Analysis: MDS	thep.lu.se/carl/mds	1.0	-	root	□	*none* (--)	r-	Yes	Rep.		Local	[Export]
Analysis: PCA	ngelab.org/voronin/pca	1.0	2002-08-31	root	□	*none* (--)	r-	Yes	-		Local	[Export]
classification: plugin name	example.com/user/myplugin	1.1	2005-07-18	root	□	*none* (--)	--	Yes	-	in, out	Local	[Export]
Hypothesis Testing: B-statistics (Bioconductor)	lcb.uu.se/ospjuth/bstat	1.01	2005-06-24	laurence	□	*none* (--)	--	Yes	-		Local	[Export]
Hypothesis Testing: B-statistics DupCorr (Bioconductor)	lcb.uu.se/ospjuth/bstat.dupcor	0.9	2005-06-24	laurence	□	*none* (--)	--	Yes	-		Local	[Export]
Normalization: Global median ratio	thep.lu.se/bjorn/medianratio	1.0	2002-08-31	root	□	*none* (--)	r-	Yes	-		Local	[Export]
Normalization: LIMMA+SMA (Bioconductor)	lcb.uu.se/ospjuth/morm	1.01	2005-06-24	laurence	□	*none* (--)	--	Yes	-		Local	[Export]
Normalization: LIMMA+SMA Between Arrays (Bioconductor)	lcb.uu.se/ospjuth/morm.btw.arr	0.9	2005-06-24	laurence	□	*none* (--)	--	Yes	-		Local	[Export]
Normalization: Lowess	thep.lu.se/bjorn/lowess	1.0	2002-08-31	root	□	*none* (--)	r-	Yes	-		Local	[Export]
Transformation: Minimum intensity	thep.lu.se/carl/minintensity	1.1	2005-07-19	root	□	*none* (--)	--	Yes	-		Local	[Export]
Vizualization: Arrayplots (Bioconductor)	lcb.uu.se/ospjuth/arrayplots	1.01	2005-06-24	laurence	□	*none* (--)	--	Yes	-		Local	[Export]

Add plug-in

Activate marked plug-ins on server --- [Go]

Update marked: Owner: same Group: same with same access. World: same access. [Ok]

Undeleted / Deleted / All

Import plug-in definition	
Select an uploaded file	- none -
Or pick a file to upload	[Parcourir...] *(Max 32.0 MB)* □ Delete after use
File description	

Continue

Datamining tools are inserted in this part

Conclusion

- This plugin has been tested on several classical databases for microarray experiments, it shows very good results

- ***Associated publications:***

*G. Even, P. Laurence, C. Dhaenens and E-G. Talbi. **"Rulemining : A new analysis tool for PASE, the web-based platform for polypeptide chips experiments"**, Poster, JOBIM 2007.*

*G. Even, L. Jourdan, C. Dhaenens and E-G. Talbi. "**Evolutionary feature selection plugin for BASE**", Poster, JOBIM 2007.*

Association rules for

LINKAGE DESEQUILIBRIUM

Linkage disequilibrium study

Objective:

Find set of haplotypes (of size between 3 and 6) which can explain the status of people in the context of the type 2 diabetes

Data:

- For each individual: the value of its SNPs and its status
- For each SNP: allels frequencies
- For each two by two combination of SNPs: their disequilibrium

Constraints:

- Snips of an haplotype must be independent:
  - Difference of frequencies < threshold1.
  - Linkage disequilibrium > threshold2.

Approach

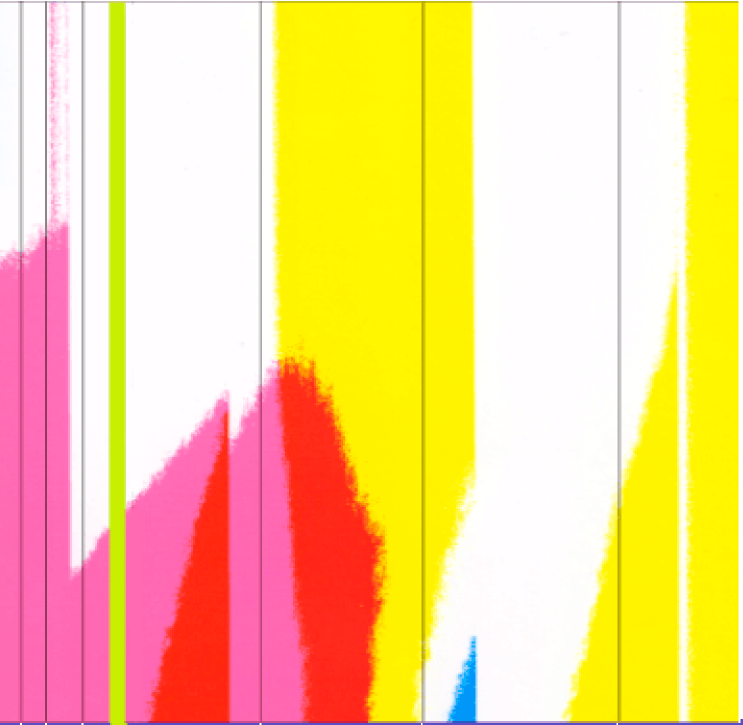
- Search method : An adaptive multi-population genetic algorithm

- A specific evaluation function based on classical biological software : CLUMP and EH-DIALL

Results

Association rules such as :

- SNP_{10}=1-2 and SNP_{20}=2-2 THEN Status=ill
- SNP_{17}=1-1 and SNP_{45}=2-2 THEN Status=ill



Conclusions

Perspectives

5th IAPR International Conference on
Pattern Recognition in Bioinformatics
22–24 September 2010, Nijmegen, The Netherlands

Conclusions

- Many problems in bioinformatics are combinatorial by nature
- Operations research (optimization) may give answers to these problems in term of:
  - Modeling
  - Definition of objective functions (quality of solutions)
    - Possibility to have several quality measures
    - Possibility to use complex evaluation of solutions
  - Provide guidelines to develop efficient optimization methods (Metaheuristics)

Perspectives

- Future researches?
- Still need more knowledge about the domain
- Hybridization of methods of different types:
    - Hybridization with domain specific methods
    - Hybridization with statistical methods

Questions ??

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

5th IAPR International Conference on
Pattern Recognition in Bioinformatics
22–24 September 2010, Nijmegen, The Netherlands

To go further …

PRIB 2010 presentations

Optimization Algorithms for Identification and Genotyping of Copy Number Polymorphisms in Human Populations
Gökhan Yavaş, Mehmet Koyutürk, and Thomas LaFramboise

Iterated Local Search for Biclustering of Microarray Data
Wassim Ayadi, Mourad Elloumi, and Jin-Kao Hao

Pattern Recognition for High Throughput Zebrafish Imaging using Genetic Algorithm Optimization
Alexander E. Nezhinsky and Fons J. Verbeek

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

5th IAPR International Conference on

Pattern Recognition in Bioinformatics

22–24 September 2010, Nijmegen, The Netherlands