Metaheuristics and bioinformatics: a wonderful life

Positionning

Bioinformatics

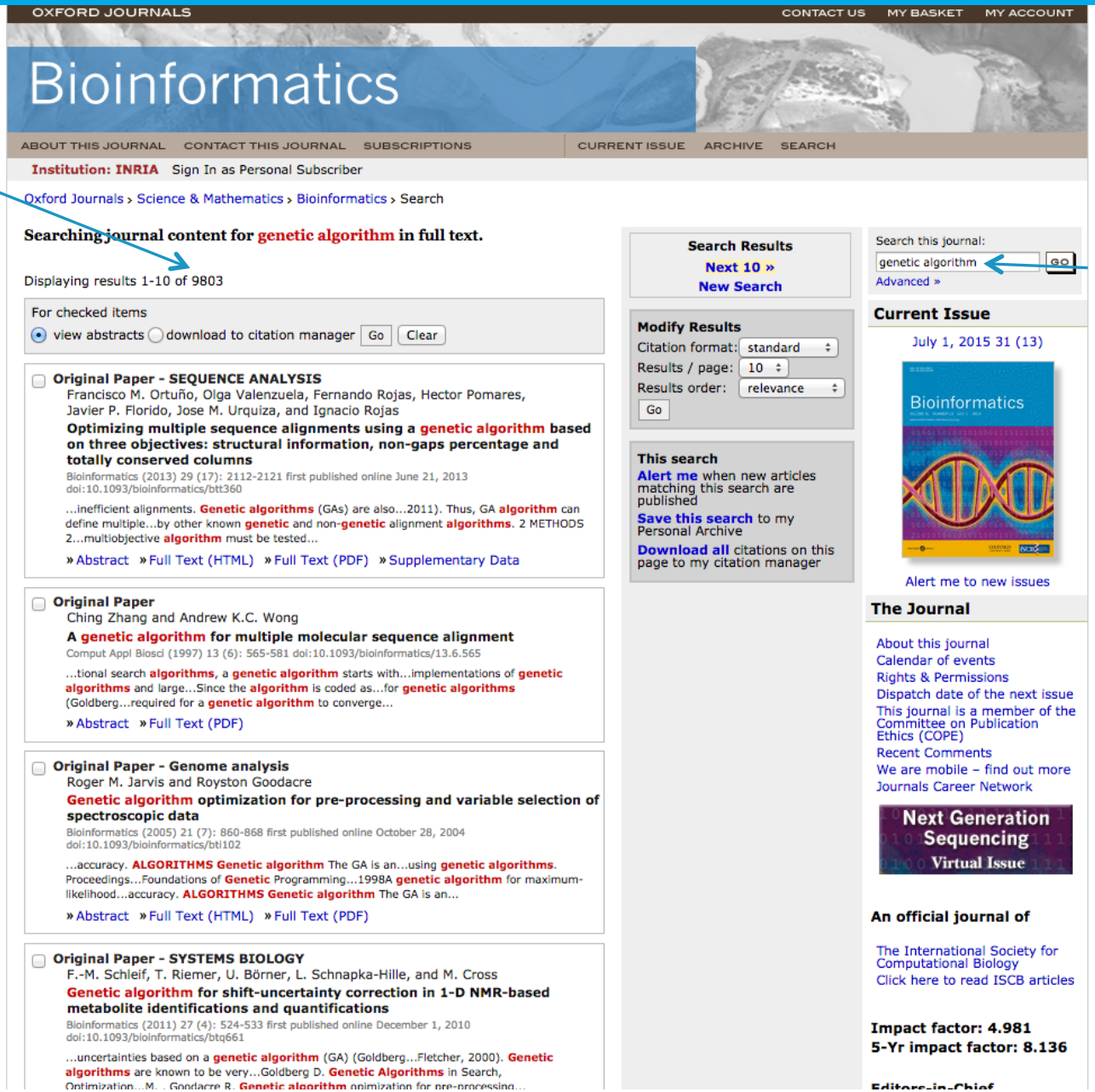
* The creation and development of advanced information and computational techniques for solving problems in biology

Combinatorial optimization

* Solve instances of problems that are believed to be hard in general, by exploring the usually-large solution space of these instances

9803 !!!

Search for genetic algorithm



OXFORD JOURNALS CONTACT US MY BASKET MY ACCOUNT

Bioinformatics

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Institution: INRIA Sign In as Personal Subscriber

Oxford Journals › Science & Mathematics › Bioinformatics › Search

Searching journal content for genetic algorithm in full text.

Displaying results 1-10 of 9803

For checked items
view abstracts download to citation manager Go Clear

Original Paper - SEQUENCE ANALYSIS
Francisco M. Ortuño, Olga Valenzuela, Fernando Rojas, Hector Pomares, Javier P. Florido, Jose M. Urquiza, and Ignacio Rojas
Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: structural information, non-gaps percentage and totally conserved columns
Bioinformatics (2013) 29 (17): 2112-2121 first published online June 21, 2013 doi:10.1093/bioinformatics/btt360
...inefficient alignments. Genetic algorithms (GAs) are also...2011). Thus, GA algorithm can define multiple...by other known genetic and non-genetic alignment algorithms. 2 METHODS 2...multiobjective algorithm must be tested...
» Abstract » Full Text (HTML) » Full Text (PDF) » Supplementary Data

Original Paper
Ching Zhang and Andrew K.C. Wong
A genetic algorithm for multiple molecular sequence alignment
Comput Appl Biosci (1997) 13 (6): 565-581 doi:10.1093/bioinformatics/13.6.565
...tional search algorithms, a genetic algorithm starts with...implementations of genetic algorithms and large...Since the algorithm is coded as...for genetic algorithms (Goldberg...required for a genetic algorithm to converge...
» Abstract » Full Text (PDF)

Original Paper - Genome analysis
Roger M. Jarvis and Royston Goodacre
Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data
Bioinformatics (2005) 21 (7): 860-868 first published online October 28, 2004 doi:10.1093/bioinformatics/bti102
...accuracy. ALGORITHMS Genetic algorithm The GA is an...using genetic algorithms. Proceedings...Foundations of Genetic Programming...1998A genetic algorithm for maximum-likelihood...accuracy. ALGORITHMS Genetic algorithm The GA is an...
» Abstract » Full Text (HTML) » Full Text (PDF)

Original Paper - SYSTEMS BIOLOGY
F.-M. Schleif, T. Riemer, U. Börner, L. Schnapka-Hille, and M. Cross
Genetic algorithm for shift-uncertainty correction in 1-D NMR-based metabolite identifications and quantifications
Bioinformatics (2011) 27 (4): 524-533 first published online December 1, 2010 doi:10.1093/bioinformatics/btq661
...uncertainties based on a genetic algorithm (GA) (Goldberg...Fletcher, 2000). Genetic algorithms are known to be very...Goldberg D. Genetic Algorithms in Search, Optimization...M. . Goodacre R. Genetic algorithm optimization for pre-processing...

Search Results
Next 10 »
New Search

Modify Results
Citation format: standard
Results / page: 10
Results order: relevance
Go

This search
Alert me when new articles matching this search are published
Save this search to my Personal Archive
Download all citations on this page to my citation manager

Search this journal:
genetic algorithm GO
Advanced »

Current Issue
July 1, 2015 31 (13)
Alert me to new issues

The Journal
About this journal
Calendar of events
Rights & Permissions
Dispatch date of the next issue
This journal is a member of the Committee on Publication Ethics (COPE)
Recent Comments
We are mobile – find out more
Journals Career Network
Next Generation Sequencing Virtual Issue

An official journal of
The International Society for Computational Biology
Click here to read ISCB articles

Impact factor: 4.981
5-Yr impact factor: 8.136

Editors-in-Chief

6104 !!!

Search for Tabu Search

OXFORD JOURNALS CONTACT US MY BASKET MY ACCOUNT

Bioinformatics

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS | CURRENT ISSUE ARCHIVE SEARCH

Institution: INRIA Sign In as Personal Subscriber

Oxford Journals › Science & Mathematics › Bioinformatics › Search

Searching journal content for tabu search in full text.

Displaying results 1-10 of 6104

For checked items
view abstracts download to citation manager Go Clear

Original Paper - Structural bioinformatics
Jacek Blazewicz, Marta Szachniuk, and Adam Wojtowicz
RNA tertiary structure determination: NOE pathways construction by tabu search
Bioinformatics (2005) 21 (10): 2356-2361 first published online February 24, 2005 doi:10.1093/bioinformatics/bti351
...solution x i and **searches** its neighborhood...mechanisms, like **tabu** list, prevent...and in their **search** procedure...Since the **tabu search** method is...solution xi and **searches** its neighborhood...mechanisms, like **tabu** list, prevent...and in their **search** procedure...
» Abstract » Full Text (HTML) » Full Text (PDF)

Original Paper
Jacek Błażewicz, Piotr Formanowicz, Frederic Guinand, and Marta Kasprzak
A heuristic managing errors for DNA sequencing
Bioinformatics (2002) 18 (5): 652-660 doi:10.1093/bioinformatics/18.5.652
...known from the literature based on **tabu search** method. Contact: blazewic@sol...known from the literature based on **tabu search** method. | Institute of Computing...known from the literature based on **tabu search** method. Contact: blazewic...
» Abstract » Full Text (PDF)

Original Paper
T J Brunette and Oliver Brock
Improving protein structure prediction with model-based search
Bioinformatics (2005) 21 (suppl 1): i66-i74 doi:10.1093/bioinformatics/bti1029
...A major drawback of beam **search** is that, over times, multiple **searches** tend to converge to a single...it to guide exploration. **Tabu search** (Glover and Laguna, 1997...and Laguna,F. (1997) **Tabu Search**. Kluwer Academic Publisher...
» Abstract » Full Text (PDF)

Original Paper
David Henriques, Miguel Rocha, Julio Saez-Rodriguez, and Julio R. Banga
Reverse engineering of logic-based differential equation models using a mixed-integer dynamic optimization approach
Bioinformatics first published online May 21, 2015 doi:10.1093/bioinformatics/btv314
...employs some elements of scatter **search** and path relinking. MISQP is a...2009) and MITS (Mixed-Integer **Tabu Search**) (Exler et-al., 2008). For...Exler O. , et al. (2008) A **TABU search**-based algorithm for mixed-integer...
» Abstract » Full Text (HTML) » Full Text (PDF) » Supplementary Data
OPEN ACCESS CORRECTED PROOF

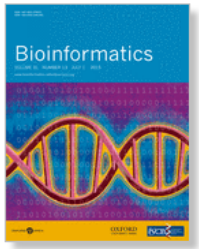
Search Results
Next 10 »
New Search

Modify Results
Citation format: standard
Results / page: 10
Results order: relevance
Go

This search
Alert me when new articles matching this search are published
Save this search to my Personal Archive
Download all citations on this page to my citation manager

Search this journal:
tabu search GO
Advanced »

Current Issue
July 1, 2015 31 (13)

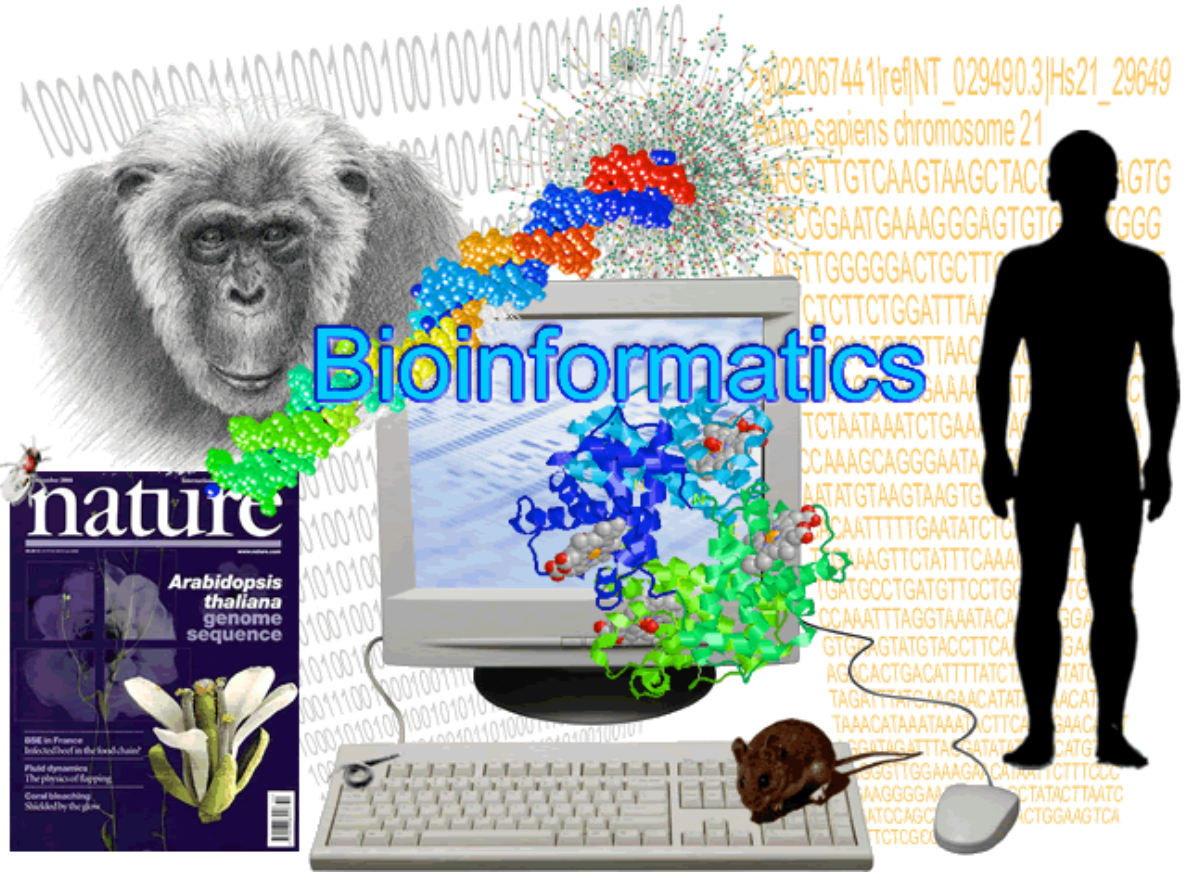


Alert me to new issues

The Journal
About this journal
Calendar of events
Rights & Permissions
Dispatch date of the next issue
This journal is a member of the Committee on Publication Ethics (COPE)
Recent Comments
We are mobile – find out more
Journals Career Network

An official journal of
The International Society for Computational Biology
Click here to read ISCB articles

Impact factor: 4.981
5-Yr impact factor: 8.136

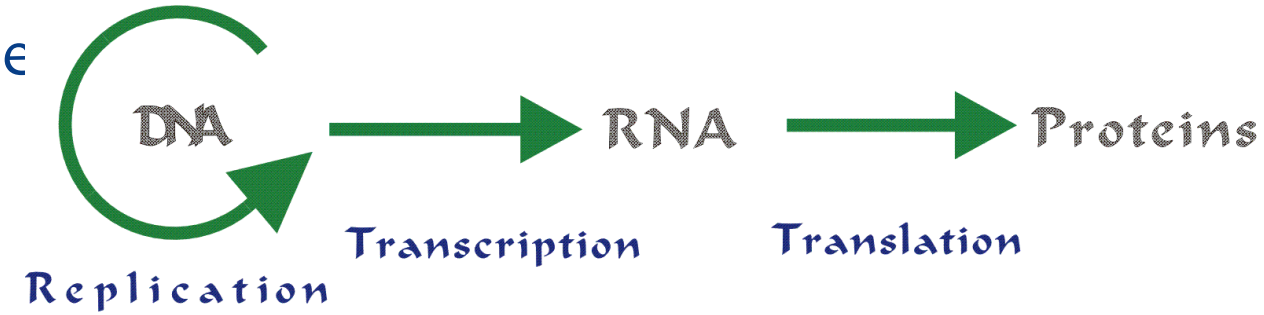

Bioinformatics
nature
Arabidopsis thaliana genome sequence

Definition

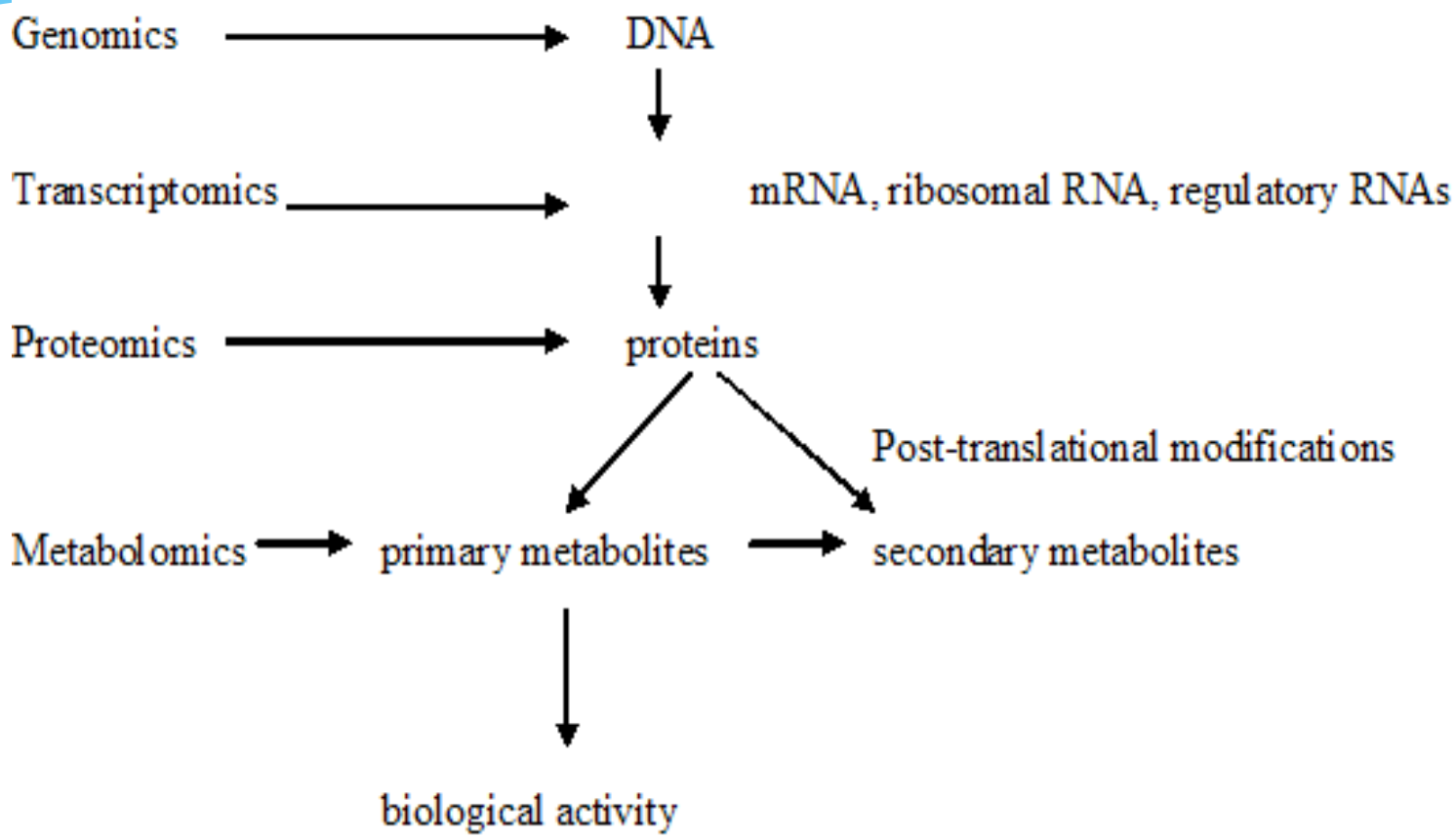
- Can be defined as the **body of tools, algorithms** needed to handle large and complex biological information.
- **Bioinformatics** is a scientific discipline created from the **interaction of biology and computer.**
- The NCBI defines **bioinformatics** as: *"Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline"*

Concepts

- Ce

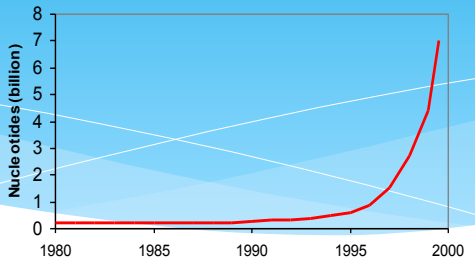


-omics technology ?

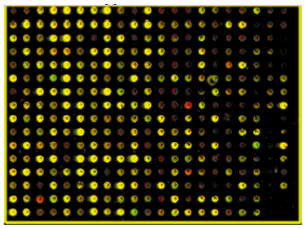


The trend of data growth

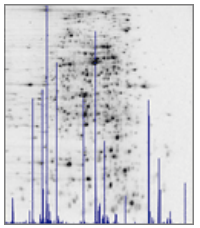
- 21st century is a century of biotechnology:
  - Genomics: *New sequence information is being produced at increasing rates. (The contents of GenBank double every year)*



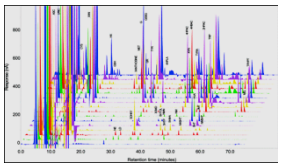
  - Microarray: *Global expression analysis: RNA levels of every gene in the genome analyzed in parallel.*



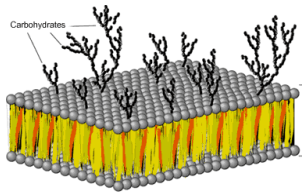
  - Proteomics: *Global protein analysis generates by large mass spectra libraries.*



  - Metabolomics: *Global metabolite analysis: 25,000 secondary metabolites characterized*

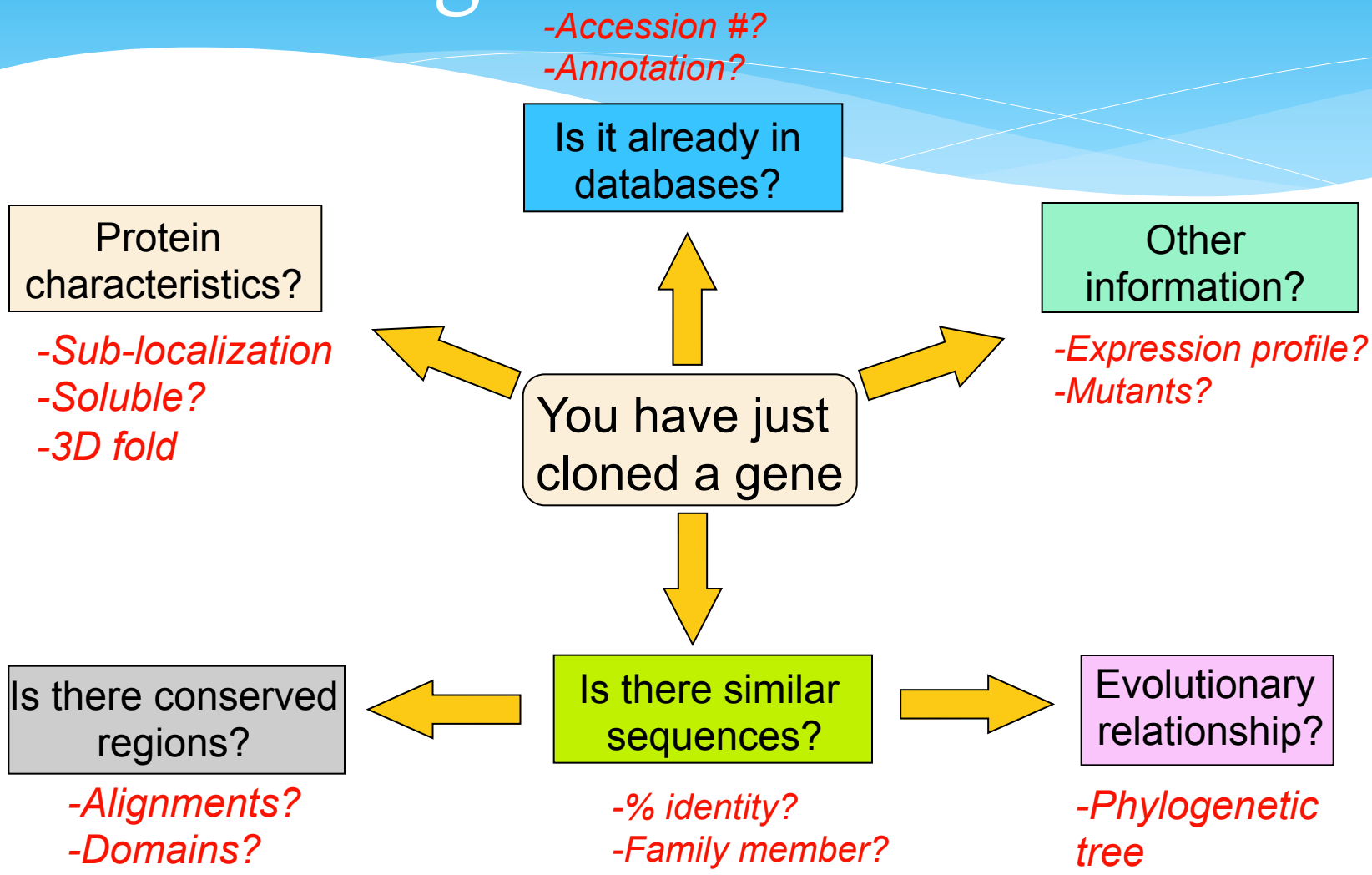


  - Glycomics: *Global sugar metabolism analysis*



Drew Sheneman, New Jersey--The Newark Star Ledger

Applying algorithms to analyze genomics data



Bioinformatics problems

- Gene identification & annotation
  - Identify and classify genes on the genome
- Microarrays & gene expression analysis
  - Use DNA microarray (gene chip) to measure mRNA
- Sequence alignment & multiple alignment
  - Database searches
- Phylogenetic tree induction
- Protein structure determination, modeling, and prediction
- Ligand screening and docking
- Many, many more

Bioinformatics data

- DNA sequence information
  - Genome projects, etc.
- mRNA expression information
  - Microarrays, SAGE
- Metabolite concentrations
  - Mass Spec., NMR Spec., etc.
- Protein sequence information
- Protein structure information
  - X-Ray Crystallography

Concepts

Molecular biology

* **Living organisms (on Earth) require ability to**

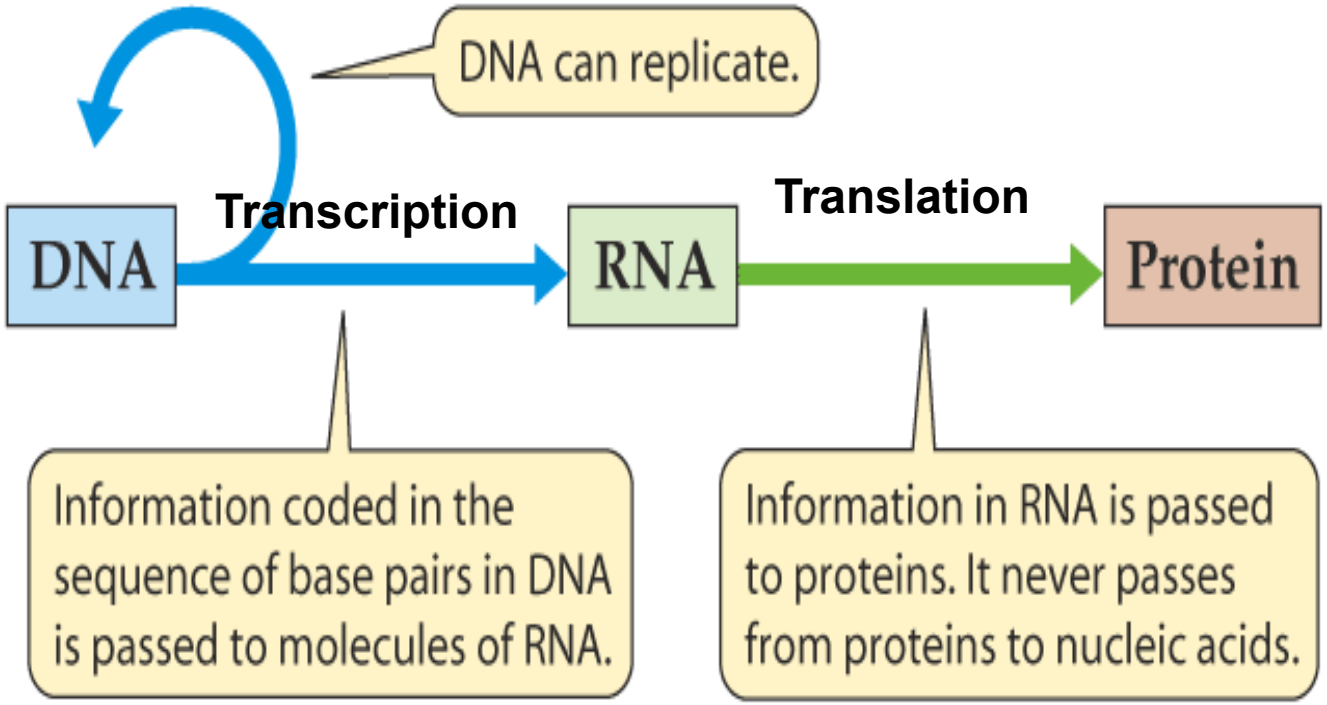
1. Separate inside from outside (lipids)
2. Build 3D machinery to perform biologi functions (proteins)
3. Store information on how to build machinery (DNA)

* **Diagram of a cell**

- Lipid membranes (provide barrier)
- Protein structures (do work)
- DNA nucleus (store info)

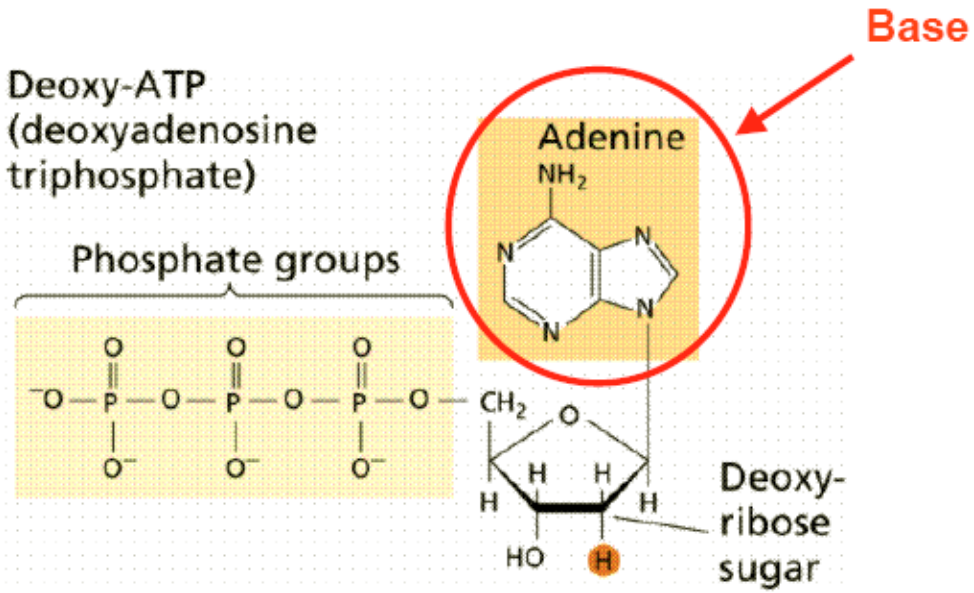
DNA, RNA, and the Flow of Information

Replication

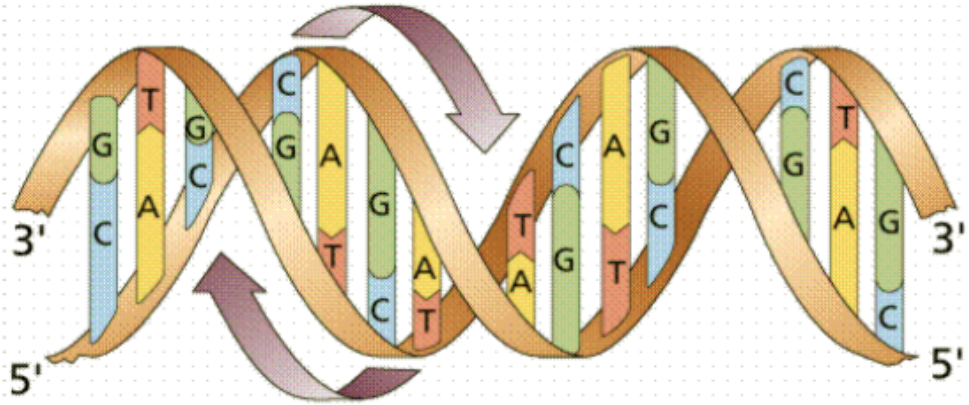


Deoxyribonucleic Acid

* **Composit**
  - Sequence
  - Nucleotic base



DNA



* **Physical structure**
- Double (stranded) helix
- Sugar & phosphate groups form backbone
- Complementary bases (A-T, C-G) connected by hydrogen bond
- 5’ = end w/ free phosphate group
- 3’ = end w/ free hydroxyl group

* Genome – entire collection of DNA for an organism
  * E.Coli chromosome 5 x 106 bases (5 Mbps)
  * Drosophila 8 chromosomes 2 x 108 bases (200 Mbps)
  * Human 48 chromosomes 3 x 109 bases (3 Gbps)

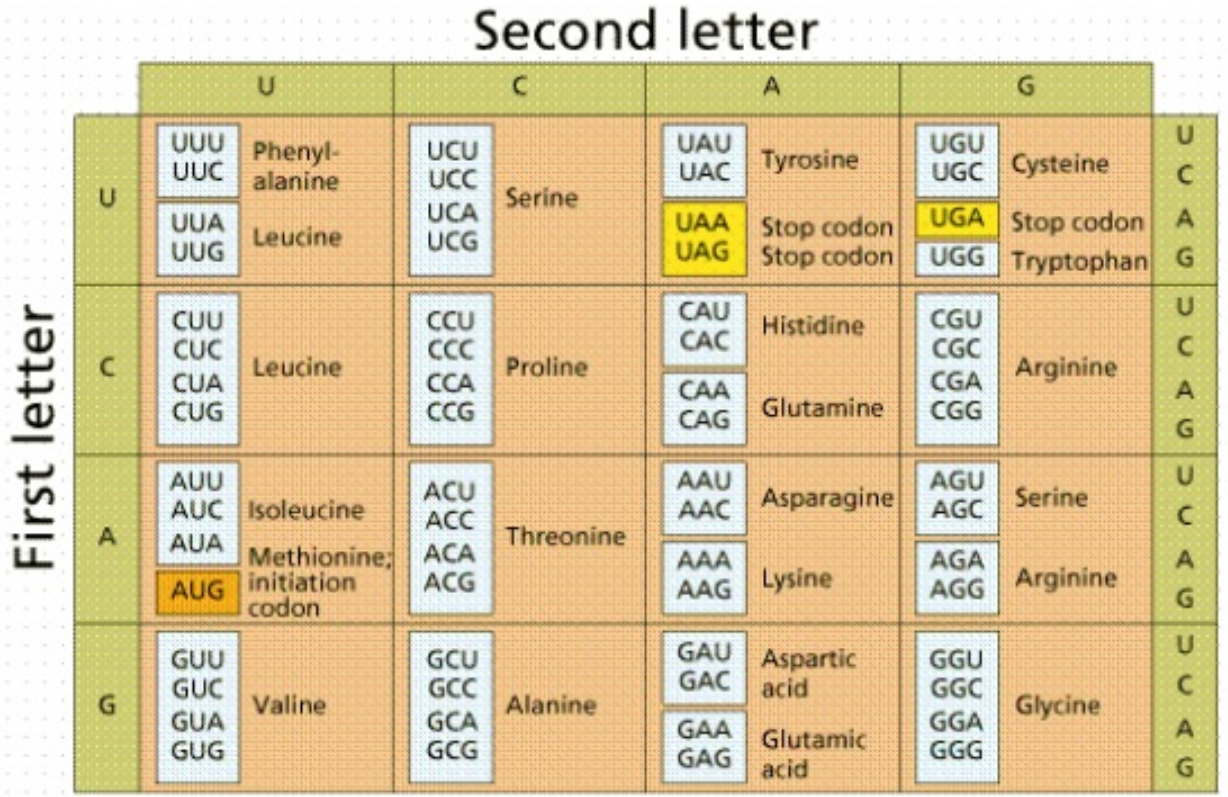
Nucleotide Bases

* **DNA**
* A = Adenine
* T = Thymine
* C = Cytosine
* G = Guanine

* **RNA**
* A = Adenine
* U = Uracil
* C = Cytosine
* G = Guanine

RNA to AA

* **Codon – 3 base RNA sequence**
* 1. (
* 2. S

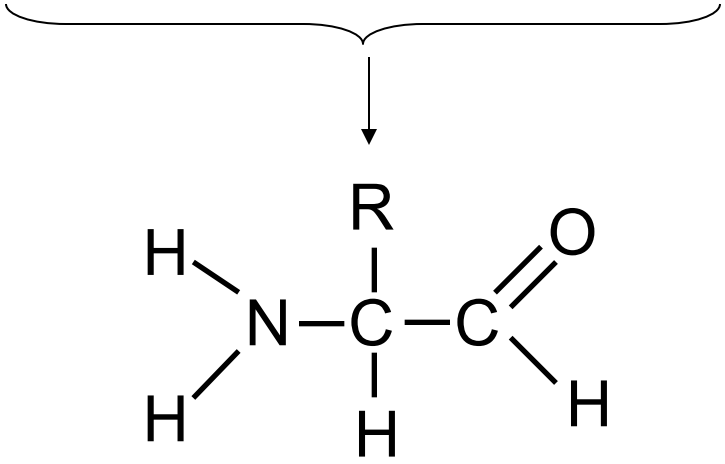


Second letter

First letter	U	C	A	G	
U	UUU, UUC: Phenylalanine UUA, UUG: Leucine	UCU, UCC, UCA, UCG: Serine	UAU, UAC: Tyrosine UAA, UAG: Stop codon Stop codon	UGU, UGC: Cysteine UGA: Stop codon UGG: Tryptophan	U C A G
C	CUU, CUC, CUA, CUG: Leucine	CCU, CCC, CCA, CCG: Proline	CAU, CAC: Histidine CAA, CAG: Glutamine	CGU, CGC, CGA, CGG: Arginine	U C A G
A	AUU, AUC, AUA: Isoleucine AUG: Methionine; initiation codon	ACU, ACC, ACA, ACG: Threonine	AAU, AAC: Asparagine AAA, AAG: Lysine	AGU, AGC: Serine AGA, AGG: Arginine	U C A G
G	GUU, GUC, GUA, GUG: Valine	GCU, GCC, GCA, GCG: Alanine	GAU, GAC: Aspartic acid GAA, GAG: Glutamic acid	GGU, GGC, GGA, GGG: Glycine	U C A G

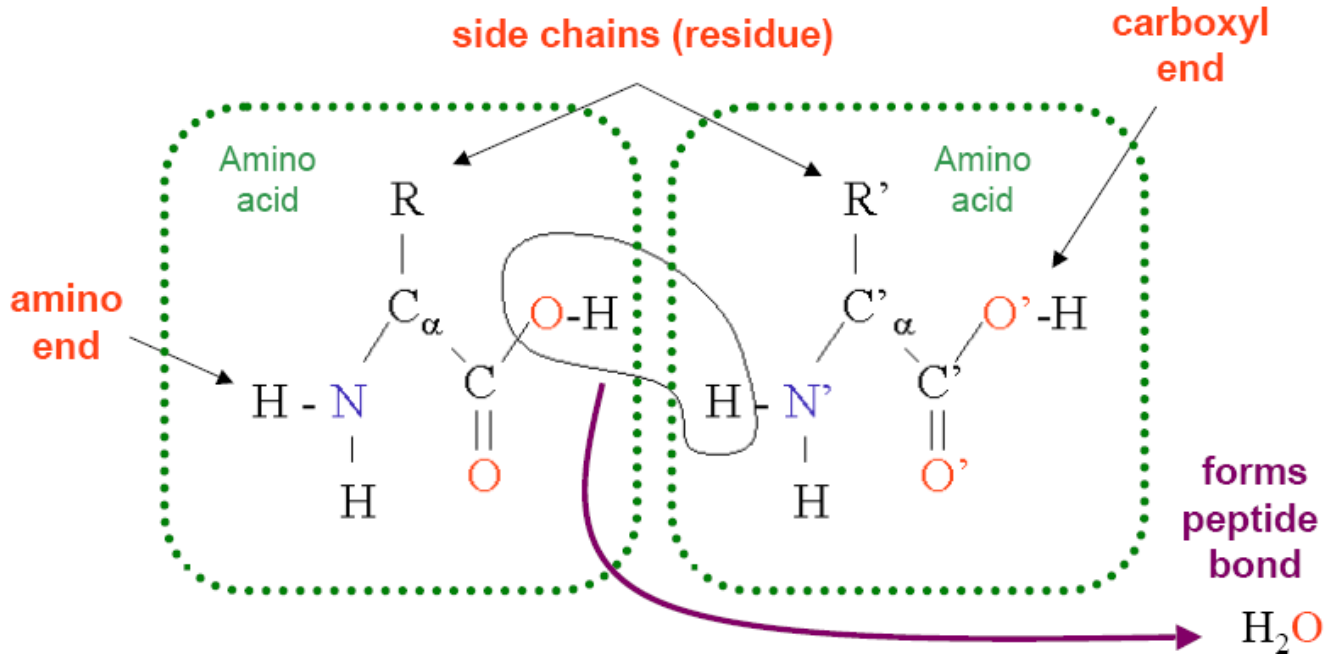
Amino Acids

* **Proteins are composed of amino acids**
* **Composition**
  - Core + side chain (residue)
  - 20 different residues → 20 unique amino acids
* **A C D E F G H I K L M N P Q R S T V W Y**



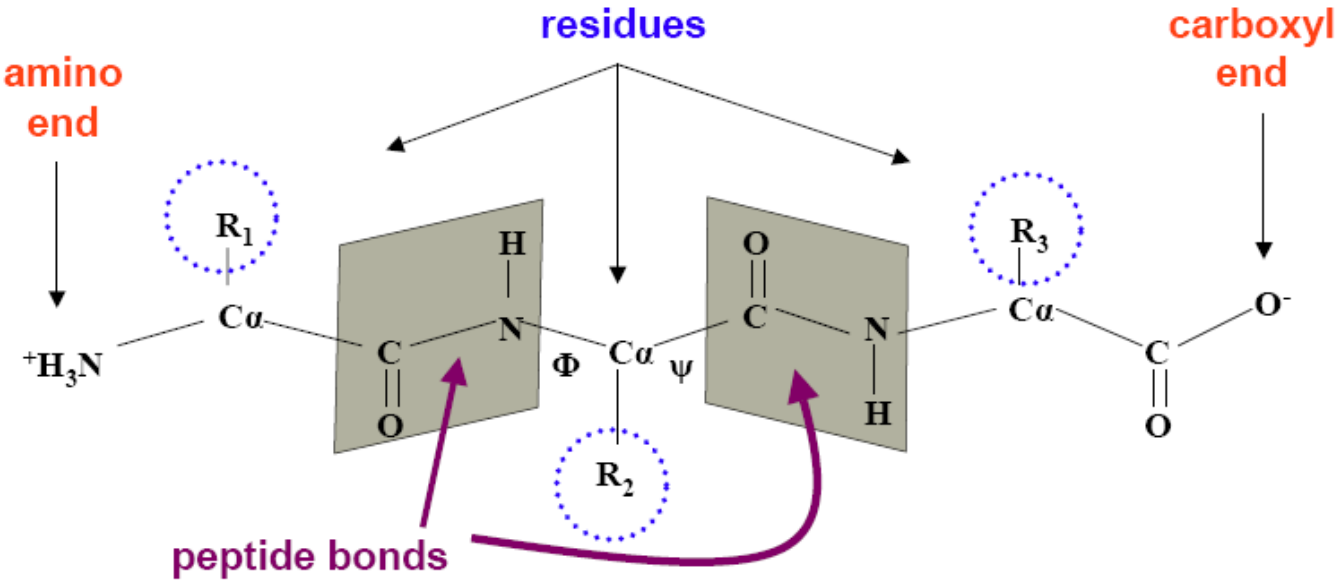
Proteins

* **Composition**
  -
  -



Proteins

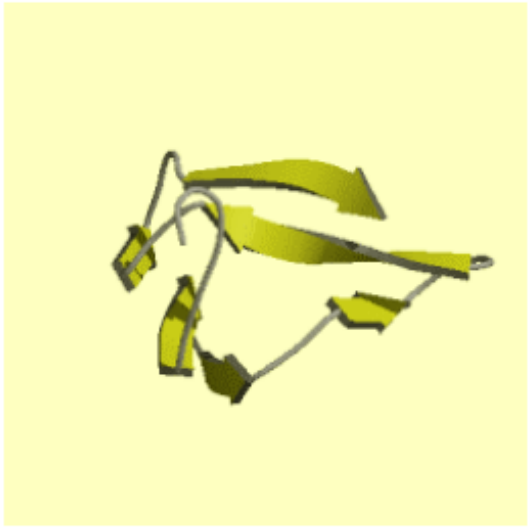
* **Primary structure (1D sequence)**
  - Order of amino acid (residues)



Proteins

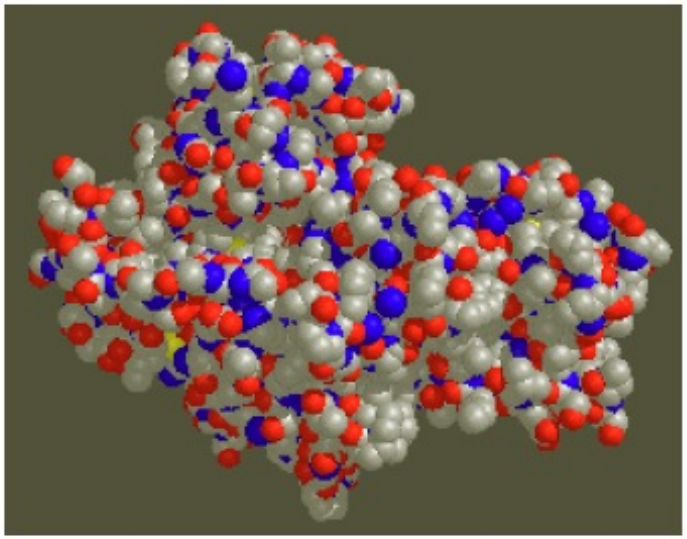
* **Secondary structure (primary substructure)**
  - Alpha
  - Beta
  - Coil

Alpha helix

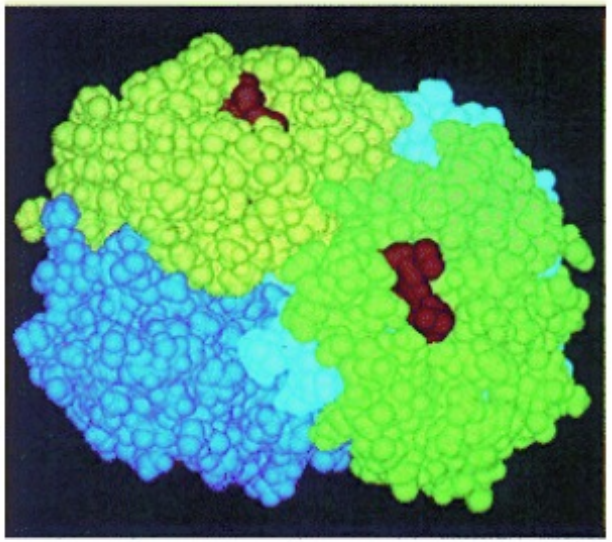


Beta strand

Proteins



Hexokinase



Hemoglobin

Experimental Biochemistry Techniques

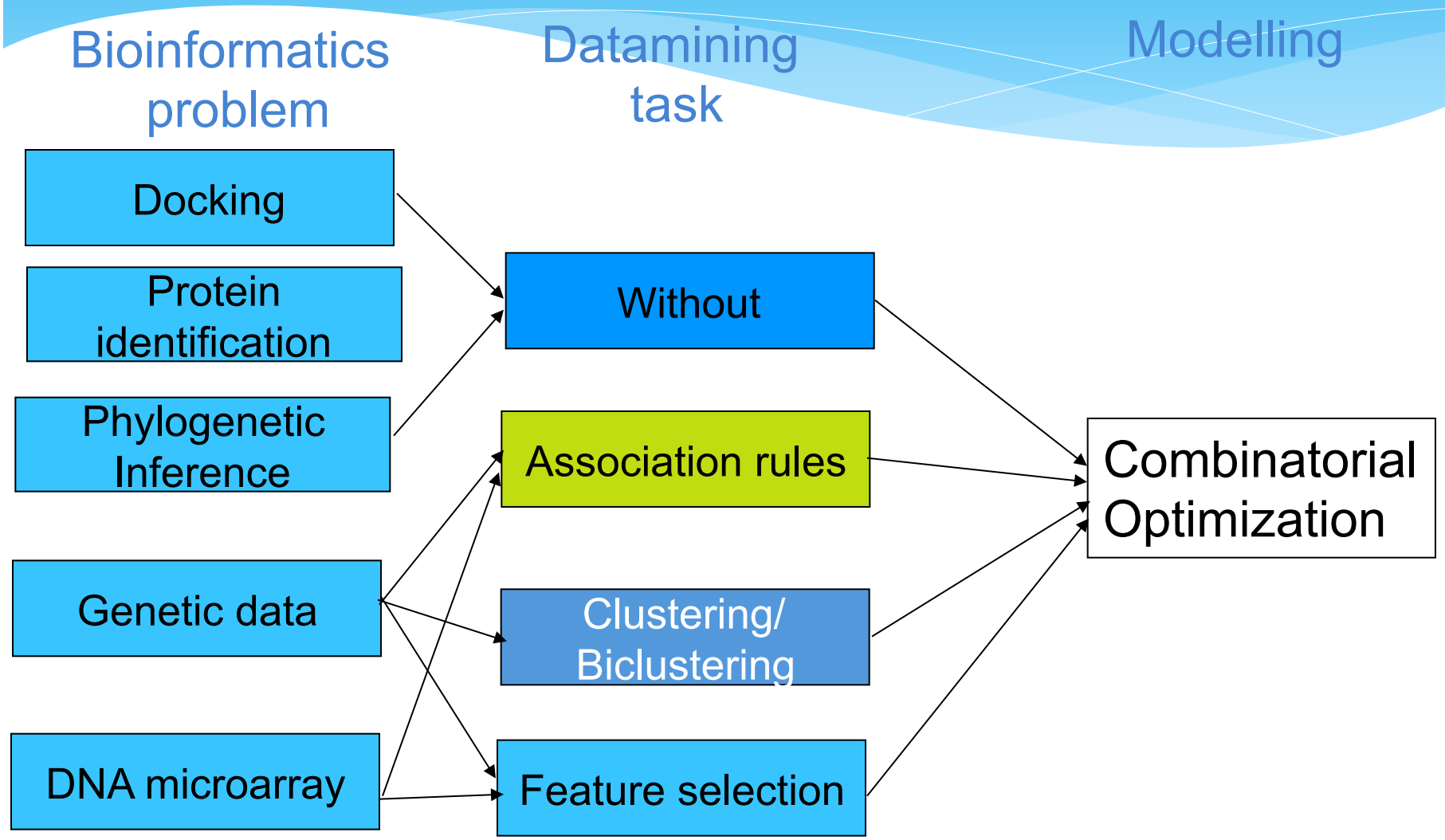
* DNA amplification (PCR)
  - Amplify (copy) DNA sequences
* DNA sequencing
  - Determine sequence of bases in DNA
* Protein sequencing
  - Determine sequence of amino acids in protein
* Protein structure
  - Determine 3D structure of protein

Molecular Biology Summary

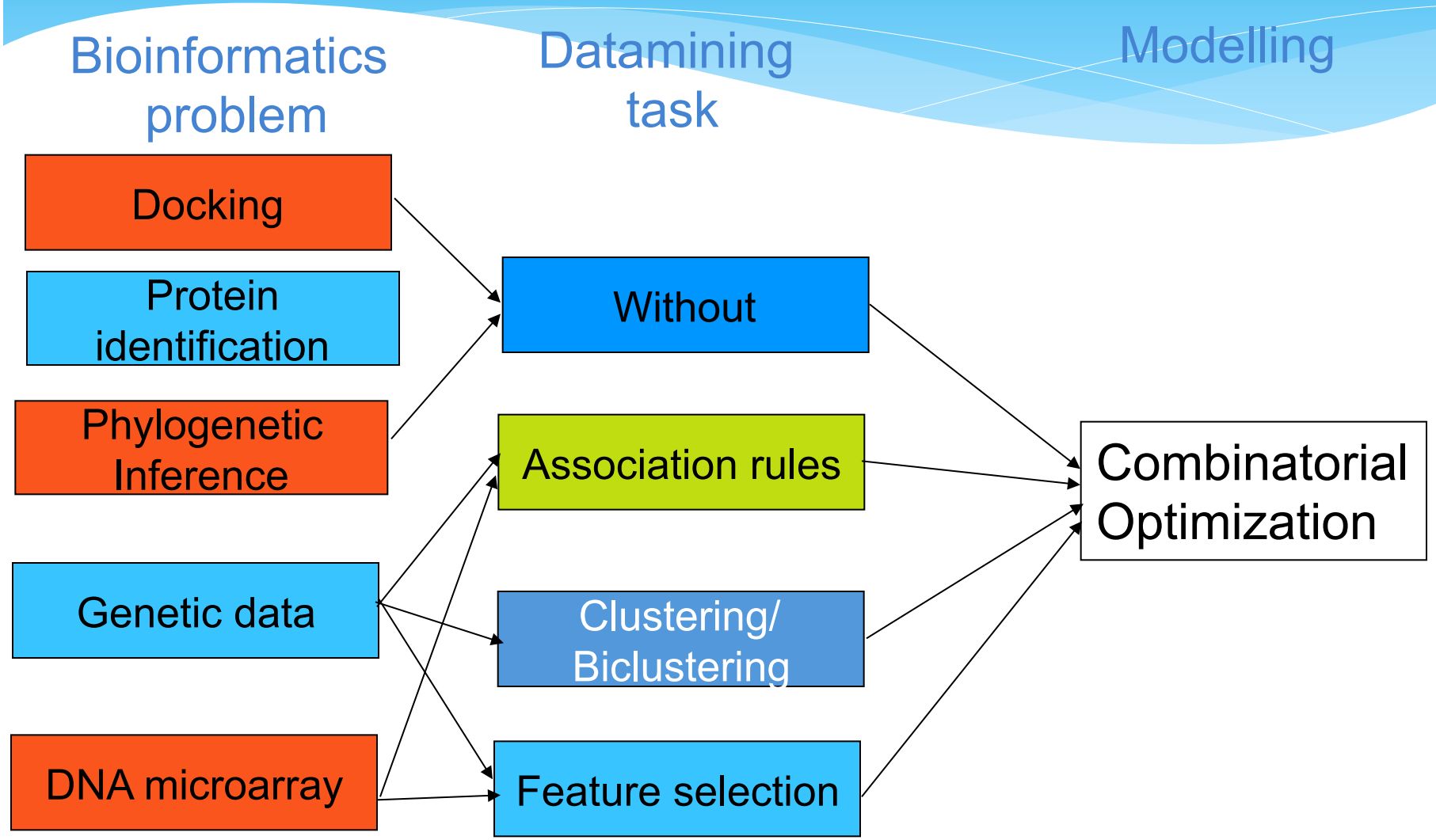
* **Important biological structures**
1. DNA Encodes information in bases
2. RNA Carries information from DNA to ribosome
3. Proteins Sequence of amino acids (3D structure → function)
* **Information transfer via complementary base pairs**
1. DNA → DNA (Replication)
2. DNA → RNA (Transcription / Gene expression)
3. RNA → Protein (Translation)
* **Experimental techniques**
- DNA amplification (PCR), sequencing
- Protein sequencing, structure determination

From bioinformatics problems to metaheuristics

Our strategy



Our strategy



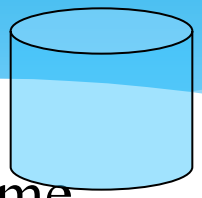
Outlines

* Datamining examples
  - → Modeling datamining tasks as MCOP (Multi-Objective Combinatorial Optimization Problems)
  - → Clustering
  - → Association rules
* Phylogeny
  - → New MO optimization model
  - → Adapted optimization method
* Molecular docking
  - → New optimization model
  - → Efficient optimization methods

Datamining for biological data

Datamining in bioinformatics

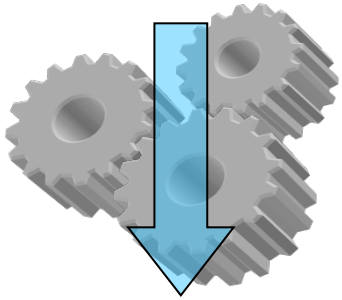
Molecules
Genome
Transcriptome
…



Large Databases

Modeling

Datamining Problem

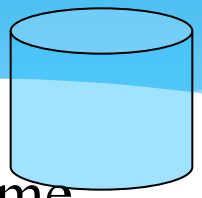


Resolution methods

Results

Datamining in bioinformatics

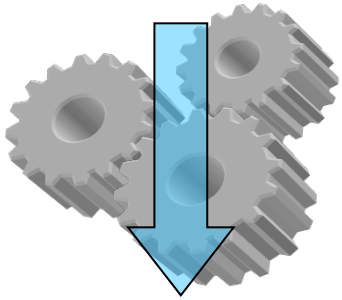
Molecules
Genome
Transcriptome
…



Large Databases

Modeling

Datamining Problem

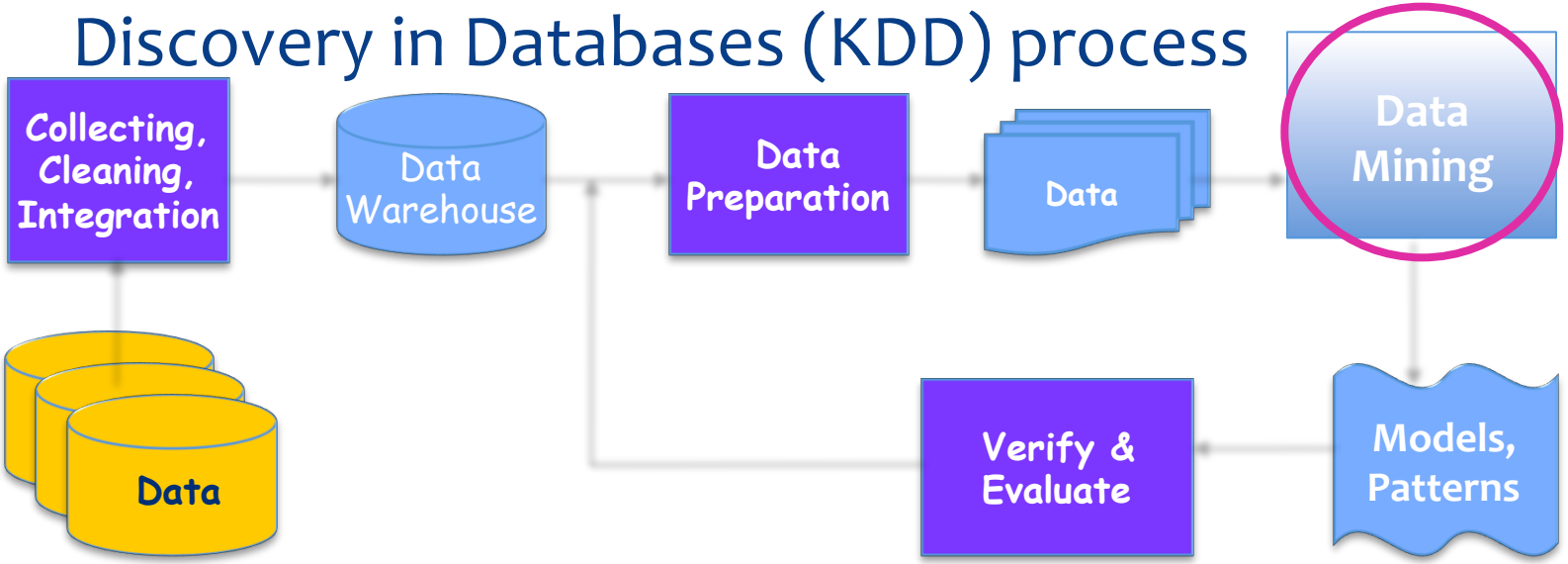


Resolution methods

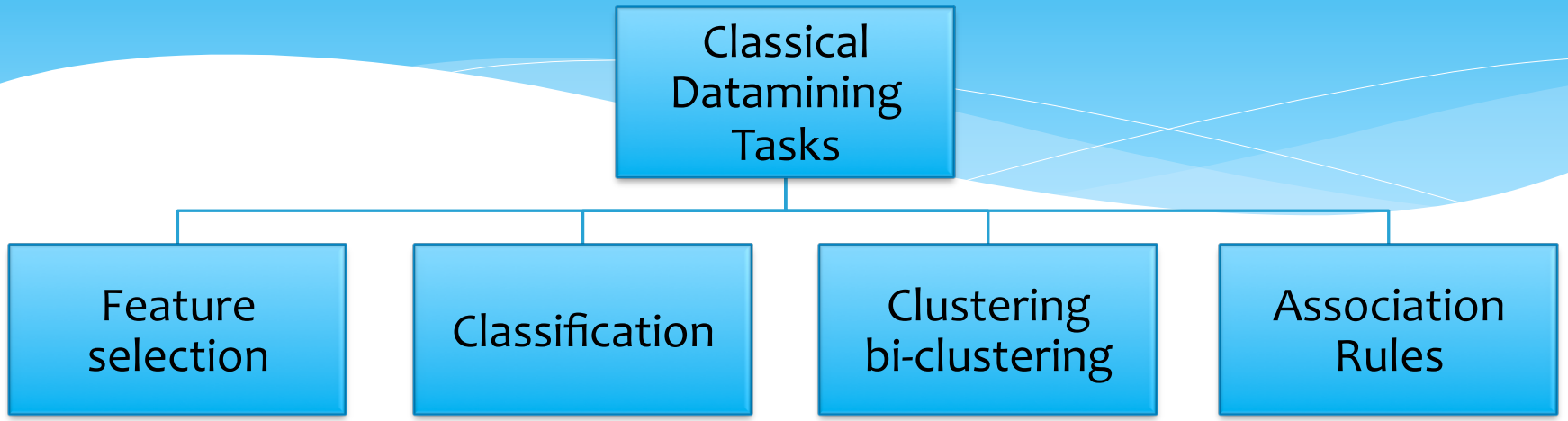
Results

Datamining / machine learning

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

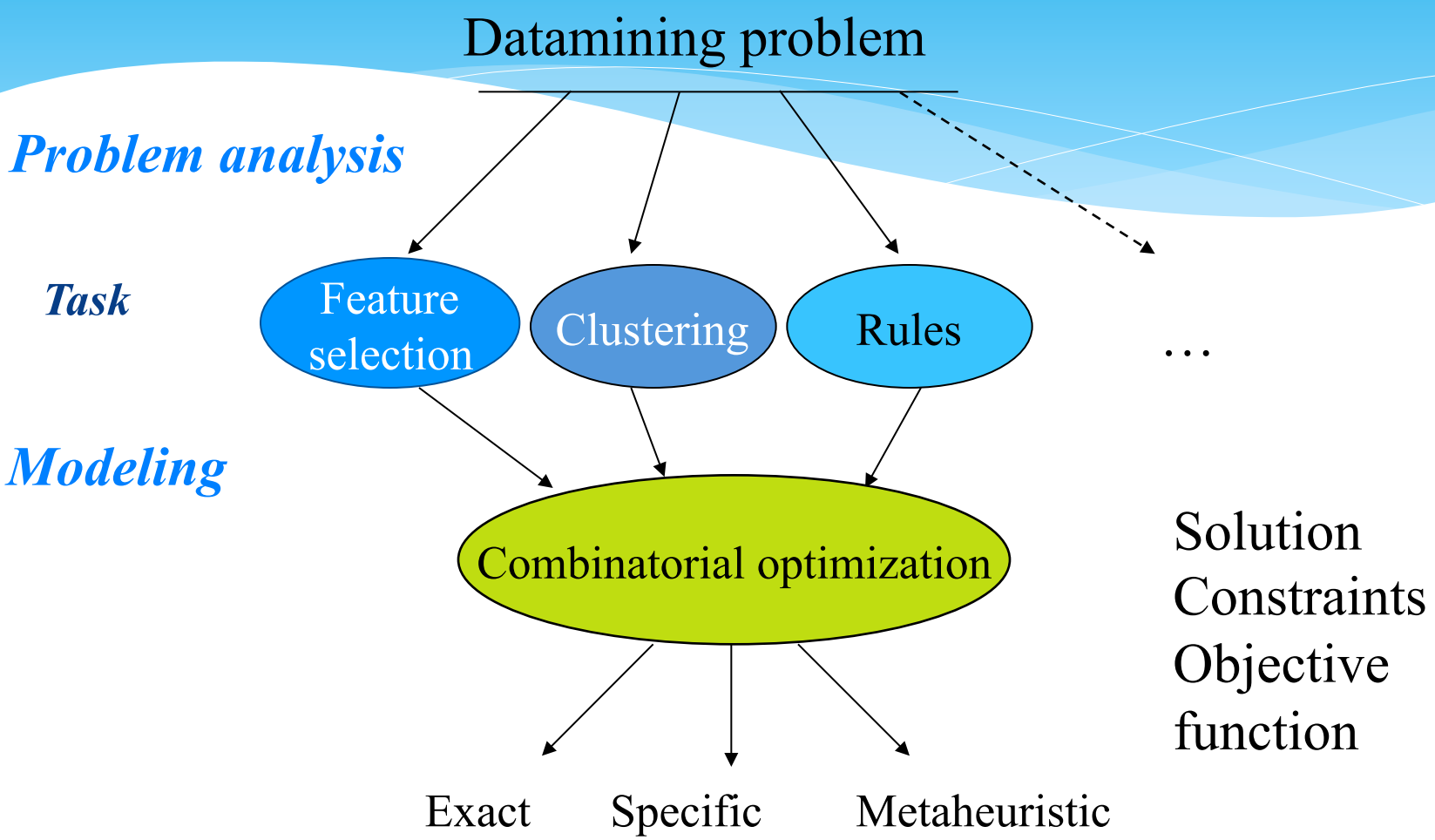


Datamining tasks



* 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

UNIVERSIDAD DE MÁLAGA

Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms

Context: microarray technology

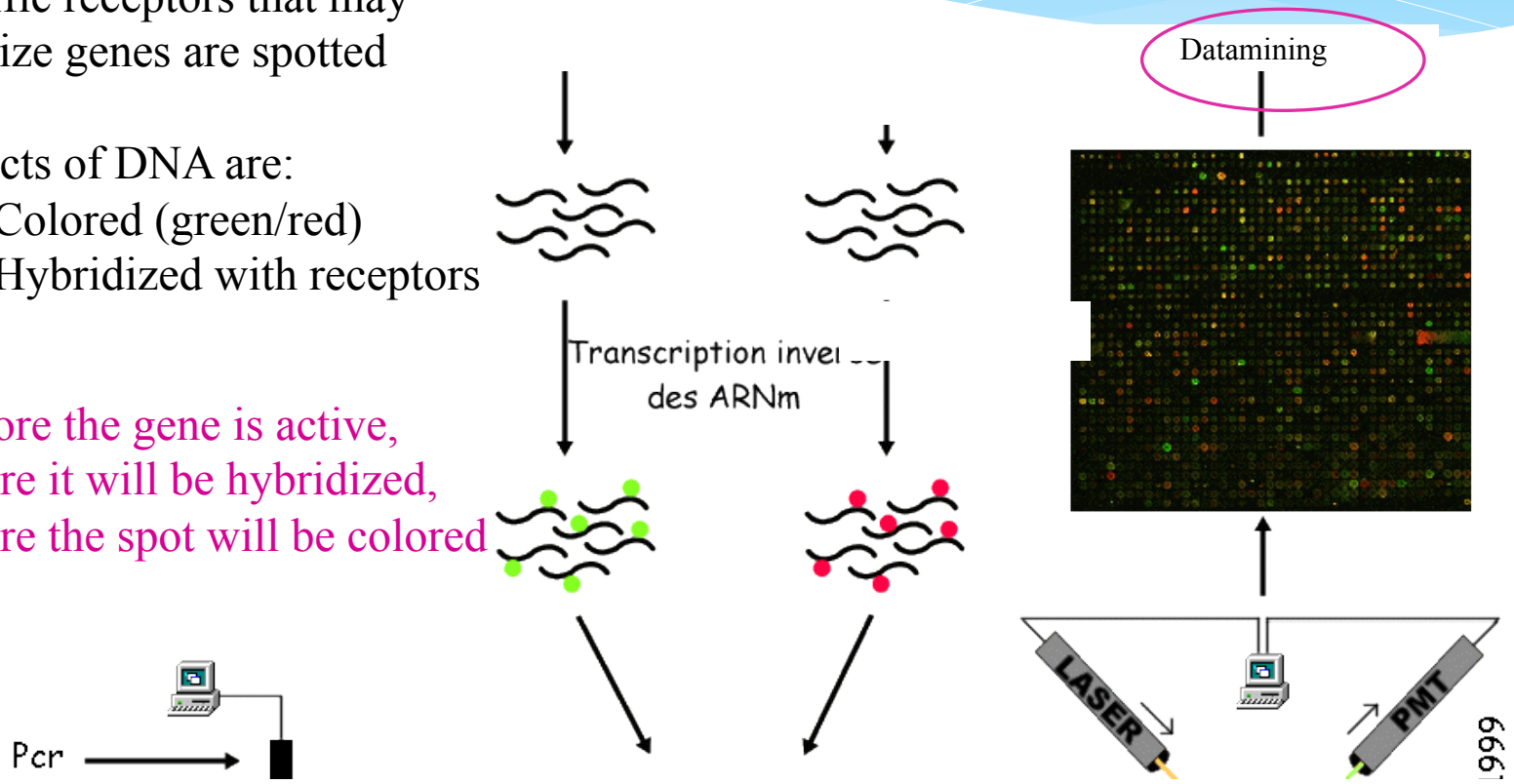
* Microarray experiments
  * Measure the gene expression levels of thousands of genes simultaneously
  * Allow to compare
  * →several conditions: tissue, treatment or time point.
  * Used to
  * Identify genetic factors for some diseases (diabetes, obesity, coronary heart disease,…)
  * Identify function of some genes in genome

Context: microarray experiment

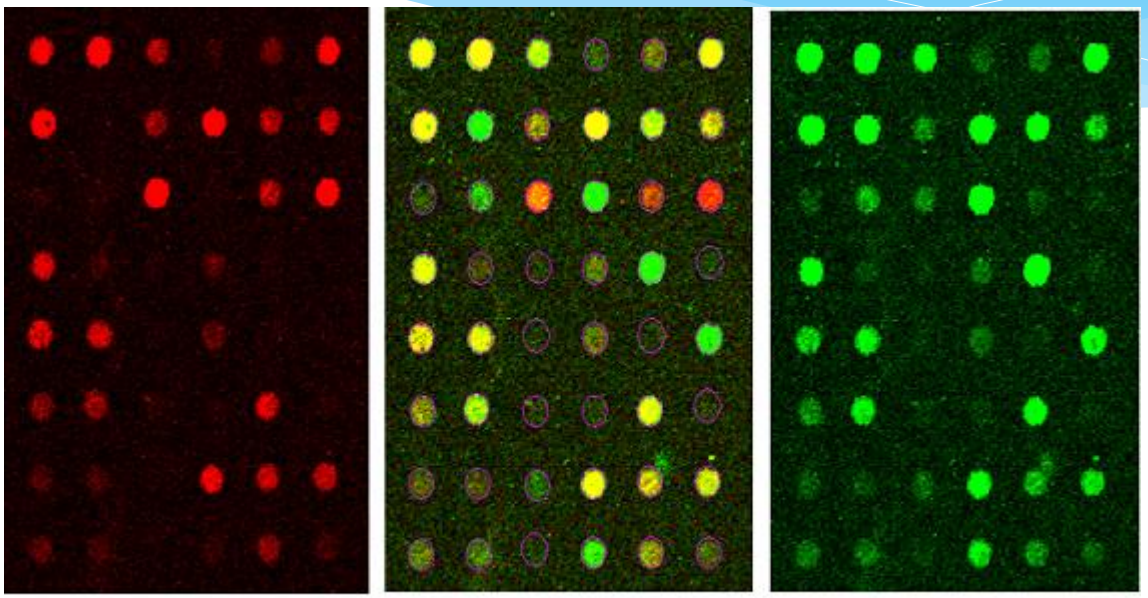
- Specific receptors that may recognize genes are spotted

- 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

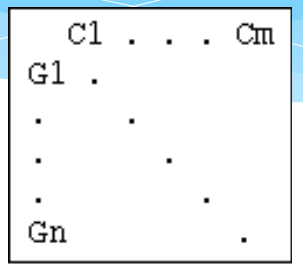


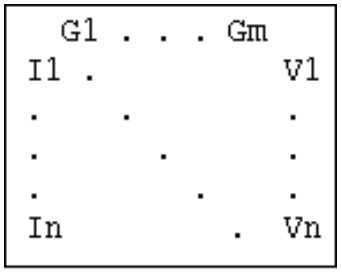
A result example :

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).





* Nature of the data

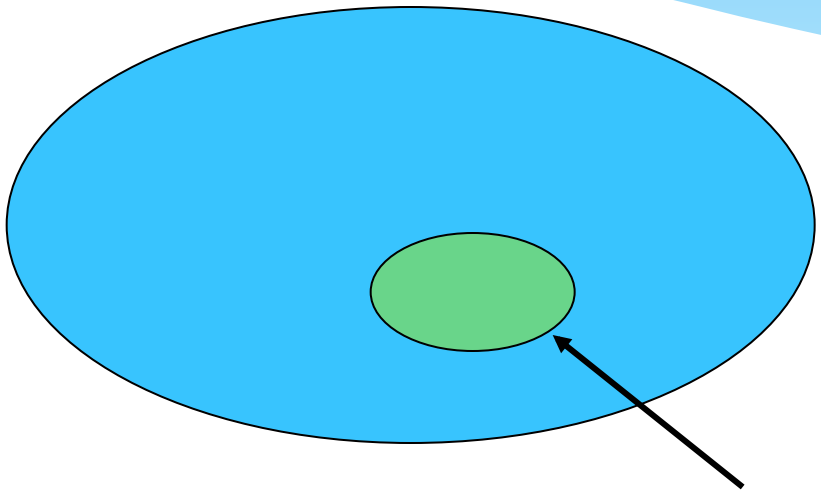
Activity of genes are represented as numerical values

➢ Discretize them into 5 values :

High Increase, Increase, No Change,

Decrease, High Decrease.

Feature selection



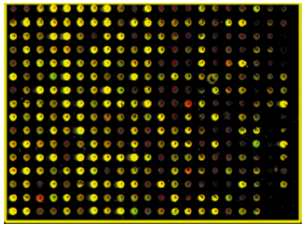
Large set of features
(genes, products, …)

- Redundancy
- Noise

Subset of features
- Significant
- Improve classification

Strategy

Bioinformatics problem	Datamining task	Modelling	Solving
Discover genes Involved in diseases	Feature selection	Combinatorial Optimization	Metaheuristics
DNA microarray		Objective function	PSO GA Operators Encoding



Objectives

* Distinguish (Classify) tumor samples from normal ones (2 classes)
* Discover reduced subsets with informative genes, achieving high accuracies
* Geometric PSO (GPSO) for feature selection
* Classification with Support Vector Machines
* Algorithms comparisons. GPSO vs. GA
* Experimentation 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

Feature Selection (FS) I

Depending on whether the selection is coupled with a learning scheme or not

* FS can reduce the dimensionality of the datasets

Advantageous since the features are selected by optimizing the discriminate power of the induction algorithm used

* Support Vector Machines (SVM), a wrapper method was used in this work.

Given a set of features $F = \{f_1, ..., f_i, ..., f_n\}$, find a subset $F' \subseteq F$, that maximizes a scoring function $\Theta : \Gamma \rightarrow G$ such that

* FS problem definition $F' = argmax_{G \subset \Gamma}\{\Theta(G)\},$

where Γ is the is the space of all possible feature subsets of *F* and *G* a subset of Γ

Feature Selection (FS) II

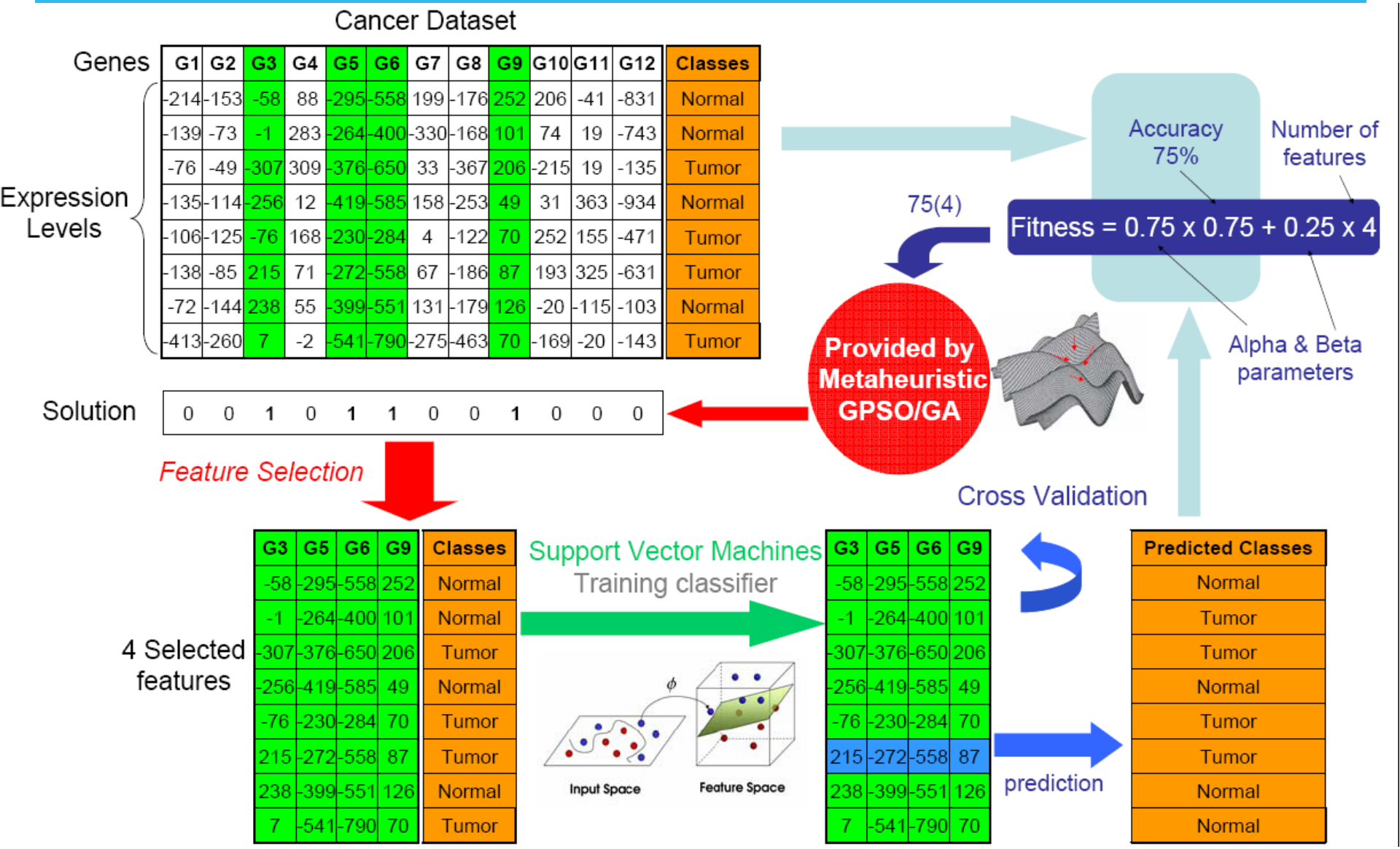
* Evaluation of solutions by means of SVM to assess the quality of the gene subset represented
* After this, 10-Fold Cross Validation is applied to calculate the rate of correct classification
* Fitness Aggregative Function (minimization):

$$fitness(x) = \alpha \cdot (100/accuracy) + \beta \cdot \#features$$

  * The population (swarm) was divided into four subsets of individuals (particles), such that:
    * 10% of individuals → *N* genes (*N* 1's in the solution)
    * 20% of individuals → 2*N genes*
    * 30% of individuals → 3*N* genes
    * 40% of individuals → randomly

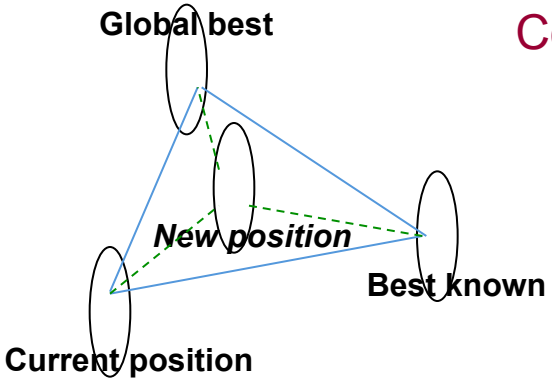
N = 4 in experiments

FS Methodology

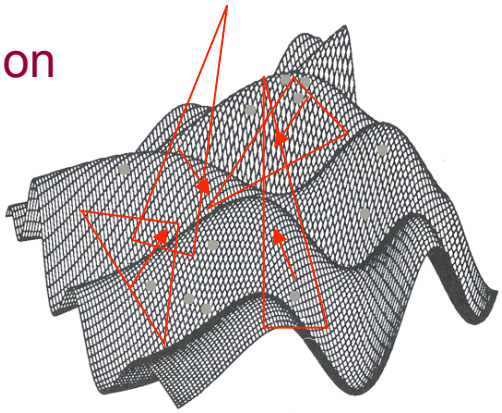


Method 1: Geometric PSO

* Based on Poli & Moraglio 2006, a new binary representation PSO algorithm
* Provide support for more representations: continuous, permutations,…
* Using Metric Space frameworks: Hamming, Euclidean, Manhattan
* Operators
  * Movement by Three Parent Geometric Crossover
  * Without velocity factor
  * Application of Mutation (BitFlip)
  * Adaptation of Three Parent Saving Pattern for FS



Convex Combination
Metric Space



Method 1: Geometric PSO

* Pseudocode

1: $S \leftarrow SwarmInitialization()$
2: **while** not stop condition **do**
3: **for** each particle x_i of the swarm S **do**
4: evaluate(x_i)
5: **if** $fitness(x_i)$ is better than $fitness(h_i)$ **then**
6: $h_i \leftarrow x_i$
7: **end if**
8: **if** $fitness(h_i)$ is better than $fitness(g_i)$ **then**
9: $g_i \leftarrow h_i$
10: **end if**
11: **end for**
12: **for** each particle x_i of the swarm S **do**
13: $x_i \leftarrow 3PMBCX((x_i, w_1), (g_i, w_2), (h_i, w_3))$
14: mutate(x_i)
15: **end for**
16: **end while**
17: **Output:** best solution found

Canonical PSO

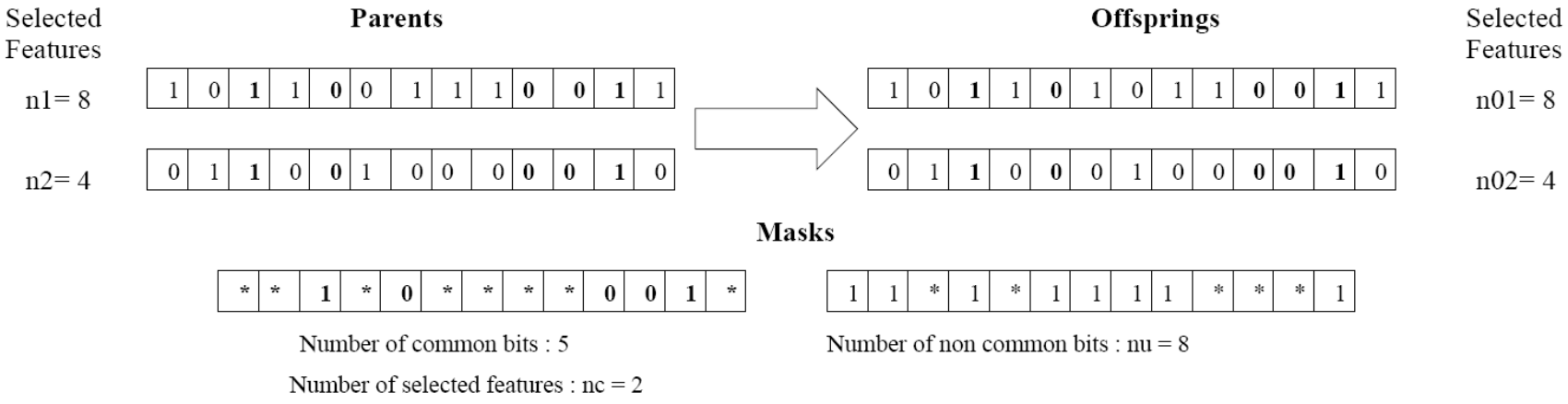
Movement

Three Parent Geometric Crossover

Method 2: Genetic Algorithm

* Generational evolution
* Elitist
* Operators:
  * Deterministic tournament Selection
  * Subset Size-Oriented Common Feature Crossover Operator (SSOCF)

SSOCF



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

Experiments

* Configurations
  * SVM: Linear Kernel using the libsvm library
  * Metaheuristics parameters

PSO		GA	
Parameter	**Value**	**Parameter**	**Value**
Swarm size	40	Population size	40
Number of generations	100	Number of generations	100
Neighborhood size	20	Probability of crossover	0.9
Probability of mutation	0.1	Probability of mutation	0.1
(w1, w2, w3)	(0.33, 0.33, 0.34)	-	-

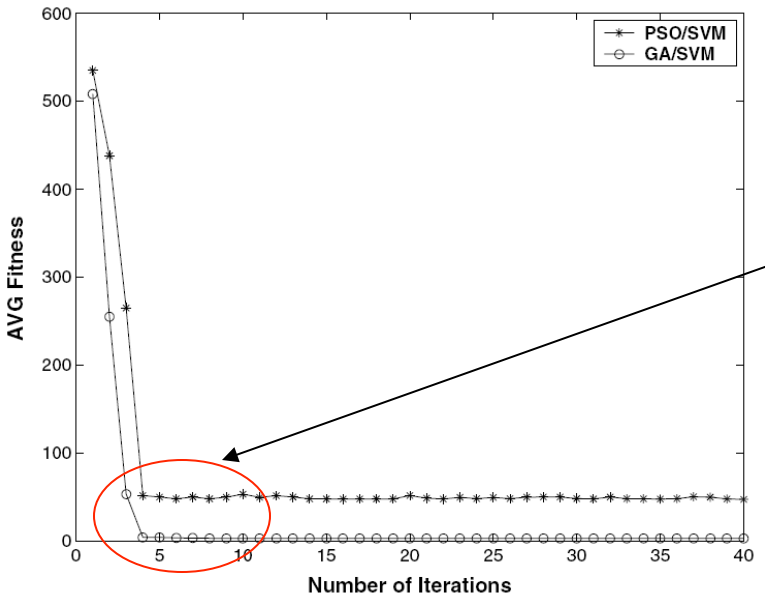
Comparison

- Two algorithms: GPSO (MALLBA Library), GA (Paradiseo Framework) and six datasets

Results

* Performance Analysis
  * Both algorithms obtain acceptable results in few iterations

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)	-	-
				-	-	-	-	99.21(75)	-
				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

* Algorithm Robustness
  * The total accuracy and the number of selected features in all the cases didn't deviate from each other by more than 5.5

Dataset	PSO_{SVM}			GA_{SVM}		
	Best	Mean	Std Dev.	Best	Mean	Std Dev.
Leukemia	100(3)	97.38(3)	3.80	100(4)	97.27(4)	3.82
Breast	90.72(4)	86.35(4)	4.11	100(4)	95.86(4)	5.33
Colon	100(2)	100(2)	0.0000	100(3)	100(3)	0.0000
Lung	99.44(4)	99.00(4)	0.50	100(4)	99.49(4)	0.41
Ovarian	100(4)	99.44(4)	0.38	100(4)	98.83(4)	3.18
Prostate	100(4)	98.66(4)	1.14	100(4)	98.65(4)	3.24

* Examples of Selected Gene Subsets

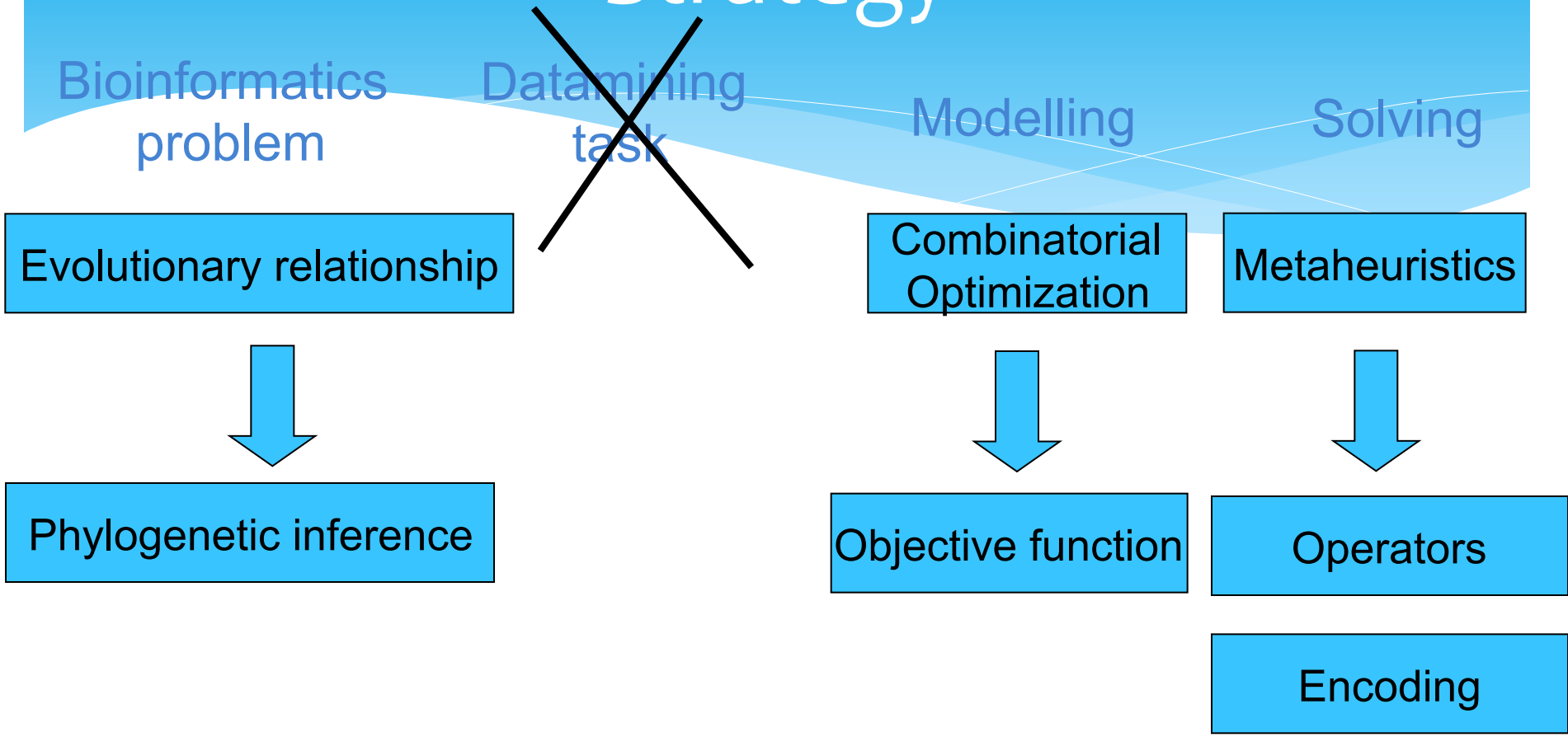
Dataset	PSO_{SVM}		GA_{SVM}	
Leukemia	100(3)	*U39226_at, L12052_at, X99101_at*	100(4)	*Z26634_at, HG870-HT870_at X52005_at, L02840_at*
Breast	90.72(4)	*NM_012269, NM_002850 AL162032, AB022847*	100(4)	*NM_005014, AF060168 NM_021176, NM_013242*
Colon	100(2)	*U29092, M55543*	100(3)	*M90684, M94132 X62025*
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)
* Biological analysis of selected gene subsets

Parallel multi-objective approaches for inferring phylogenies

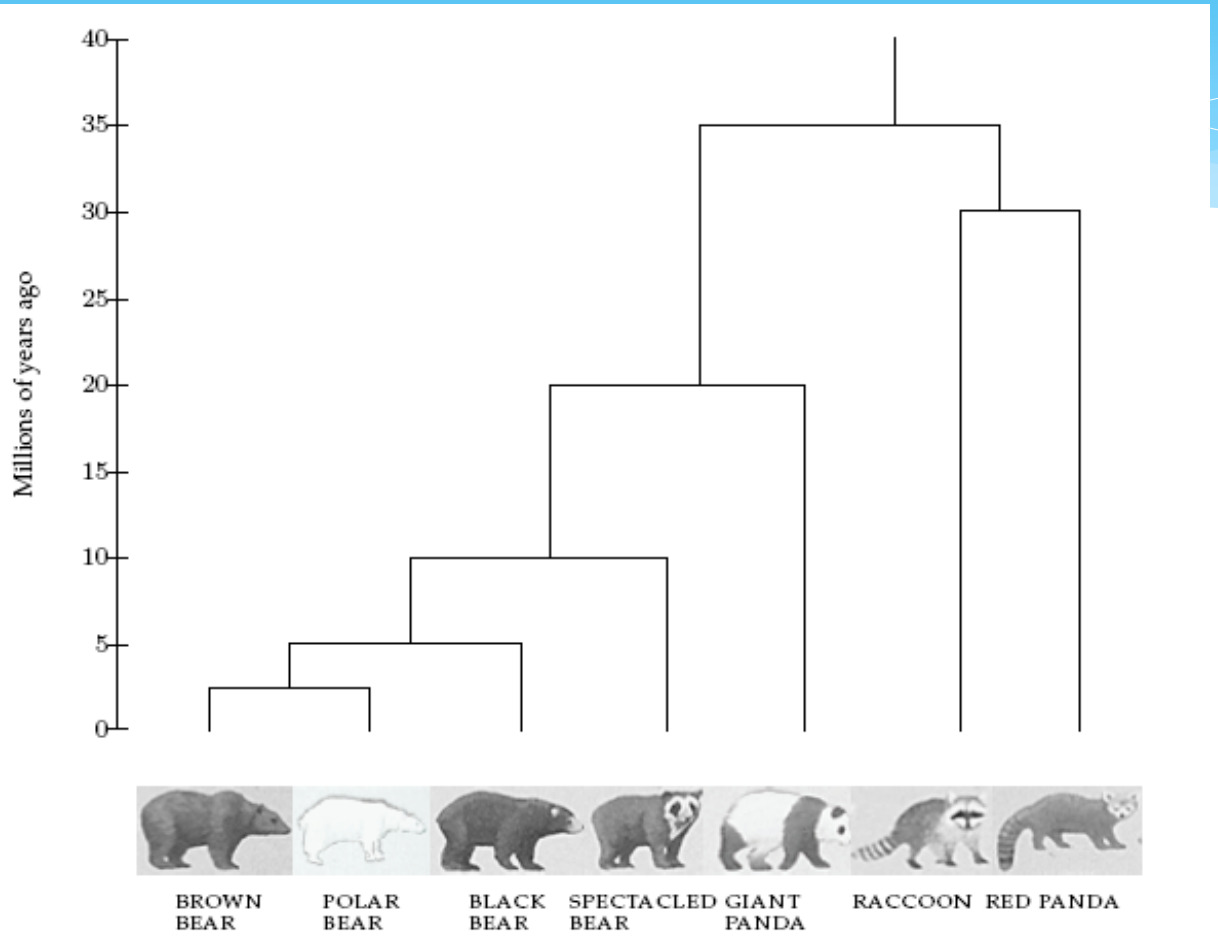
Strategy


Bioinformatics problem
Datamining task
Modelling
Solving
Evolutionary relationship
Combinatorial Optimization
Metaheuristics
Phylogenetic inference
Objective function
Operators
Encoding

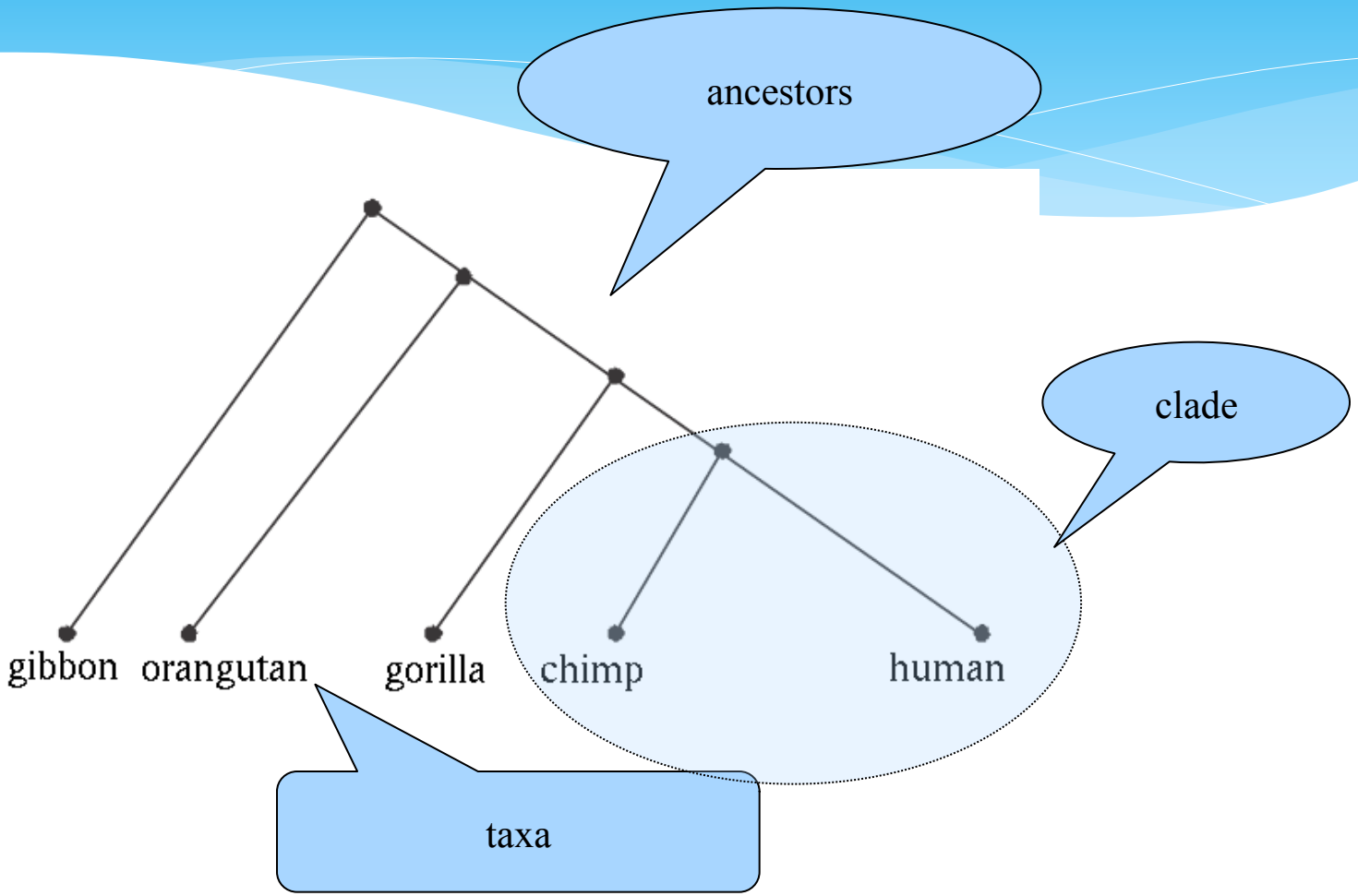
Phylogenetic Inference

* One of the “classic” bioinformatic problems
* Goals:
  * Determine evolutionary relationships among species:
    * Incomplete information
    * Estimation process
  * Estimate divergence times
  * Describe event sequence

Evolutionary Tree of Bears and Raccoons



Phylogenetic Inference



Phylogenetic Inference

* Determination of the best tree

* Phylogenetic reconstruction methods
  * Clustering techniques:
    * NJ [Saitou & Nei, 1987]
    * UPGMA [Michener & Sokal, 1957]
  * Optimality criteria:
    * Maximum parsimony [Fitch, 1971]
    * maximum likelihood [Felsenstein, 1981]
  * Bayesian methods [Ramala & Yang, 1996]

Phylogenetic Inference

* Optimality criteria based methods:
  * Optimization problem
  * Tree evaluation function
  * Search for the best tree topology
  * Huge tree search space

Phylogenetic Inference: Criterion 1

* 1st criterion: Maximum Parsimony
- Parsimony principle
  * Occam's Razor: prefer the simplest hypothesis
  * Associate hypothesis complexity → state changes
  * Counts the number of state changes required
* Maximum parsimony
  * Determine ancestral states such minimize changes for a given topology (AKA. small parsimony problem)
  * Determine the tree topology with the minimum parsimony score (large parsimony problem)

Maximum Parsimony

* input: a multiple sequence alignment of *n* sequences of *m* residues
* algorithm: find the tree with minimum changes
* output: a tree of minimum score, i.e. with minimum changes
* minimization problem

Maximum Parsimony

Number of trees

number of taxa	# of unrooted trees	# of rooted trees
10	2.0e+06	3.4e+07
20	2.2e+20	8.2e+21
30	8.6e+36	4.9e+38
40	1.3e+55	1.0e+57
50	2.8e+74	2.7e+76
80	2.1e+137	3.4e+139
n	$\prod_{i=3}^{n}(2i-5)$	$\prod_{i=2}^{n}(2i-3)$

[Richer'11]

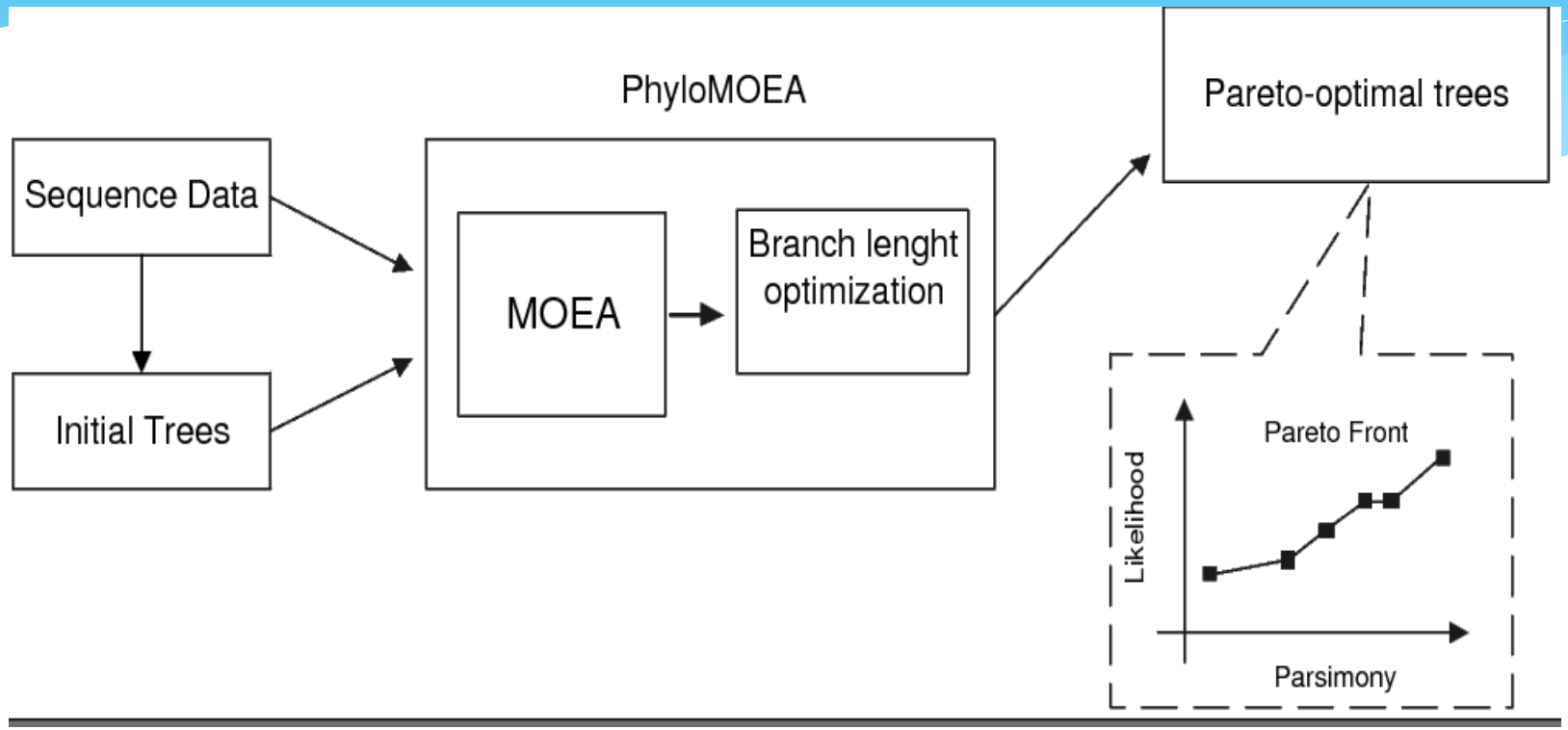
Phylogenetic Inference: criterion 2

* Maximum likelihood
  * Widely-used statistical measurement
  * L = P (D / T,M) conditional probability of the data given the tree and an stochastic model of sequence evolution
  * Assumptions
    * Site independent
    * Lineage independent
  * Must be optimized:
    * Branch lengths
    * Parameters of the sequence evolution model
  * Search for the topology that maximizes L

Phylogenetic Inference

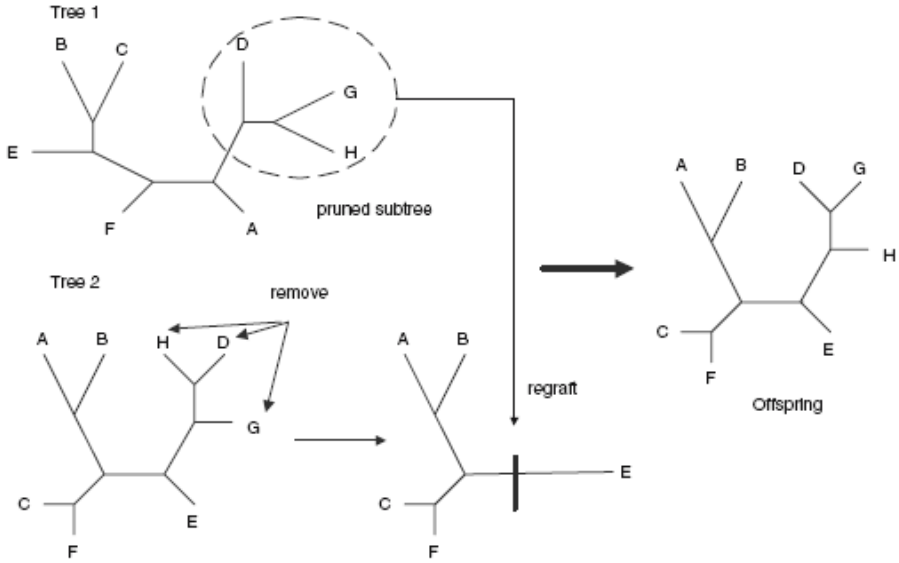
* Multi-Objective Optimization Motivation
  * Use of different inference methods can produce dissimilar trees
  * Data for the same species obtained for different sources (ex. DNA, genes) can even produce different trees using the same inference method
  * Multi-objective approach:
    * combined data
    * combined criteria

PhyloMOEA


PhyloMOEA
Sequence Data
Initial Trees
MOEA
Branch lenght optimization
Pareto-optimal trees
Pareto Front
Likelihood
Parsimony

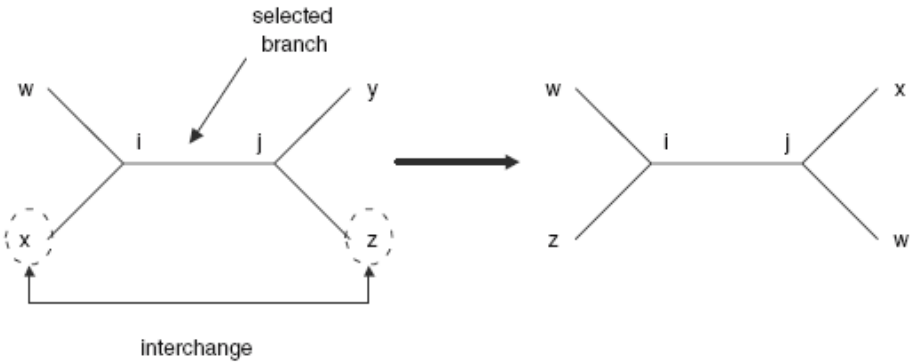
PhyloMOEA

- Encoding: a solution is a possible tree for phylogenetic inference here a graph structure is employed
- Evaluation: MO based on 2 functions
  - Maximum likehood
  - Maximum parsimony
- Crossover



PhyloMOEA (2)

* Mutation operator: NNI mutation (Nearest Neighboor Interchange)
  * 1. Choose an interior branch whose connected nodes *i*, *j* define two pairs of neighbors: *w*, *x* adjacent to *i* (*w*, *x* = *j*) and *y*, *z* adjacent to *j* (*y*, *z* = *i*).
  * 2. Execute a swap of nodes between each pair of neighboors.



Results

Solution	NONA		RAxML-V	
	Parsimony	Likelihood	Parsimony	Likelihood
*rbcL*_55	4874	-24627.8480	4894	-24583.3313
*mtDNA*_186	2438	-41049.7677	2450	-40894.5497
*RDPII*_218	41534	-170831.1213	42631	-156595.8725
*ZILLA*_500	16219	-87361.4841	16276	-86993.8264

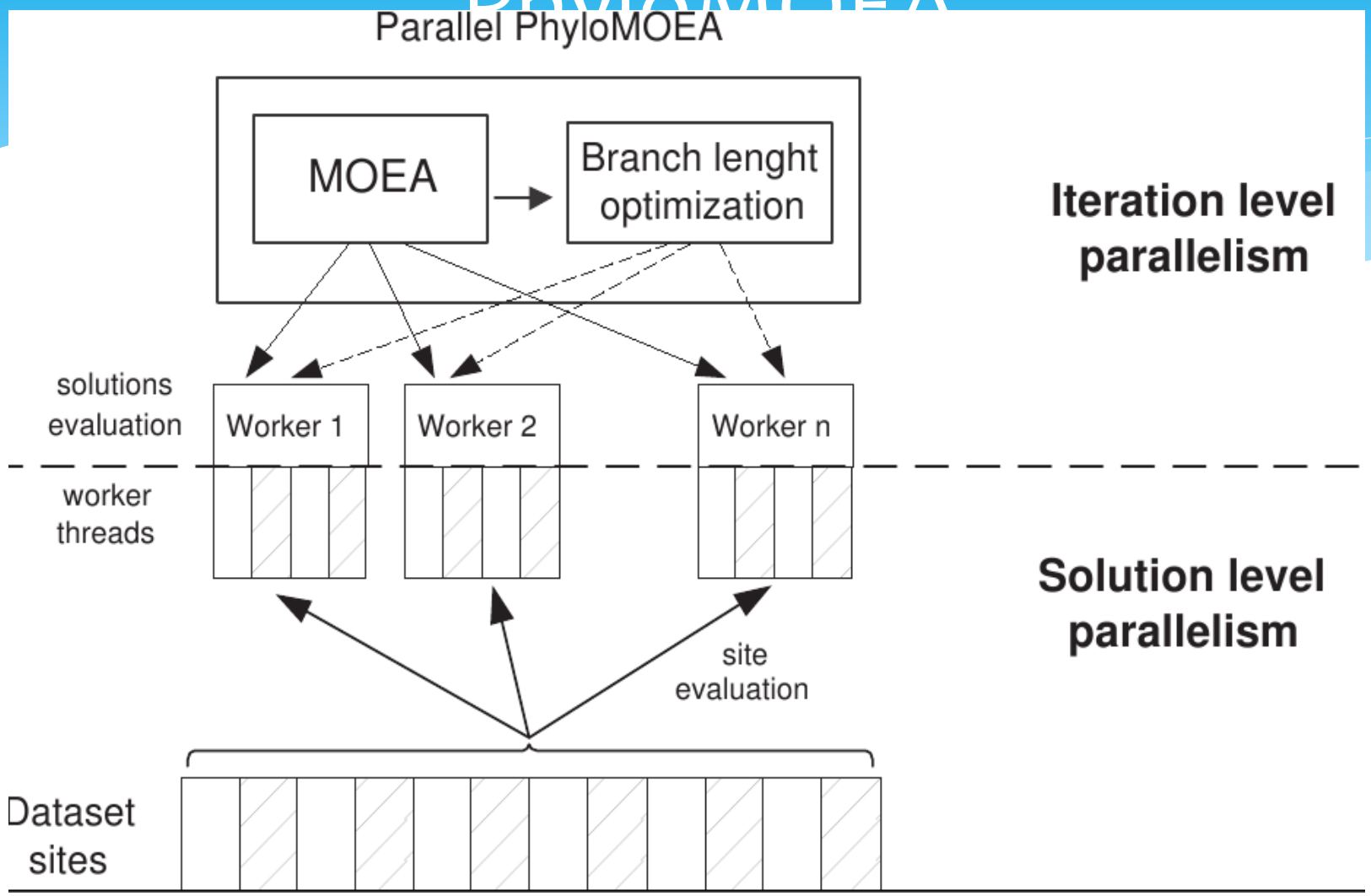
Dataset	Pareto Trees	Final Trees	Best Parsimony Tree Scores		Best Likelihood Tree Scores	
			Parsimony	Likelihood	Parsimony	Likelihood
*rbcL*_55	5	46	4874	-24626.4337	4884	-24583.3297
*mtDNA*_186	12	44	2438	-41004.3018	2450	-40894.3433
218_*RDPII*	21	82	41534	-158724.2803	42631	-156595.8224
500_*ZILLA*	16	107	16219	-87275.2812	16276	-86993.8250

PhyloMOEA

* P-based metaheuristic optimization levels:
  * Algorithmic level
  * Iteration level
  * Solution level

* ParadisEO - PEO

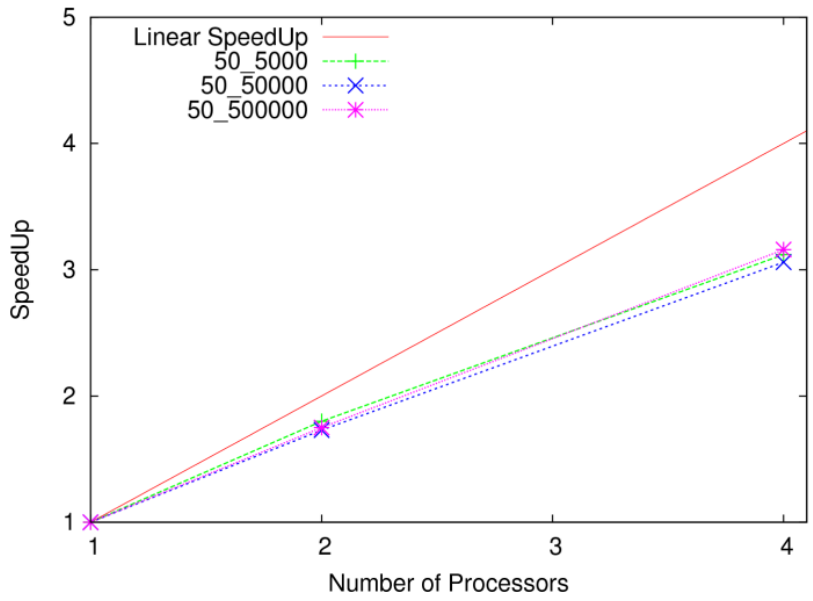
PhyloMOEA

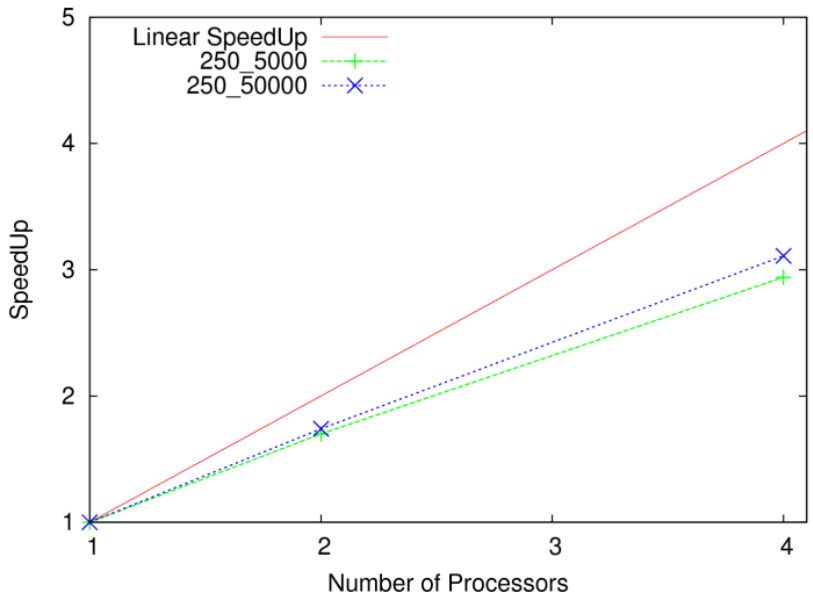

Parallel PhyloMOEA
MOEA
Branch lenght optimization
Iteration level parallelism
solutions evaluation
Worker 1
Worker 2
Worker n
worker threads
Solution level parallelism
site evaluation
Dataset sites

Results

* Likelihood function scalability
* Datasets:
  * d50_5000, d50_50000, d50_500000: 50 species and 5.000, 50.000 & 500.000 sites
  * d250_5000, d250_50000: 250 species and 5.000 & 50.000 sites
  * d500_5000, 500 species and 5.000 sites
* Test site scalability vs. species scalability

Results: number of sites

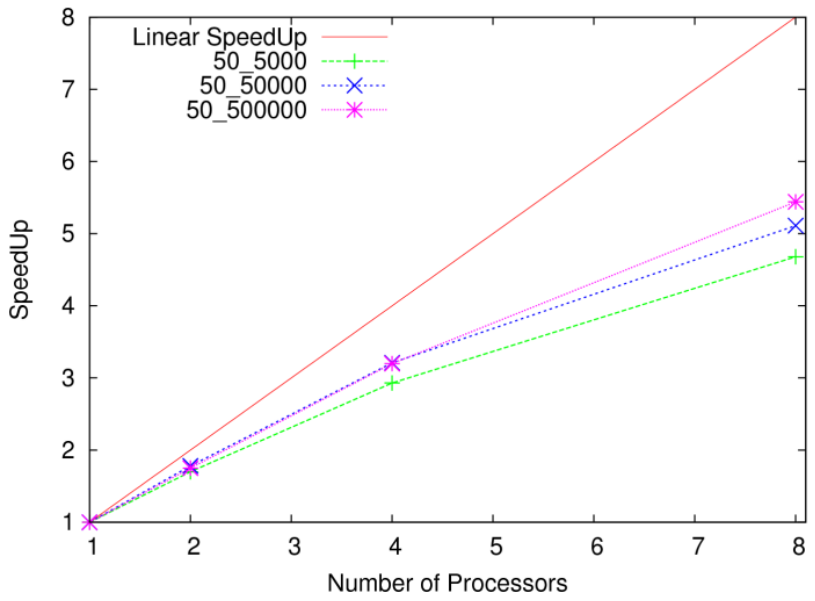




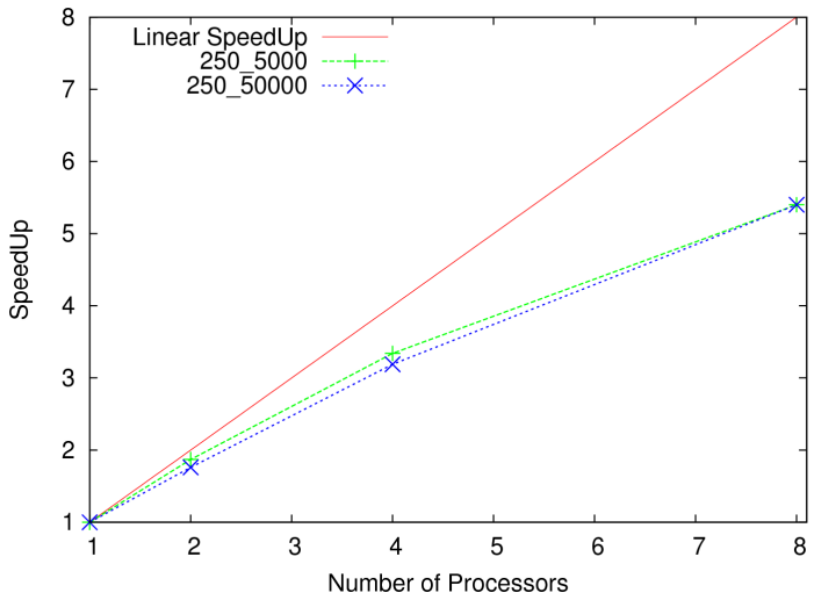
Quad-core Intel Xeon E530: 1, 2, 4 threads

Results: number of sites

50 species



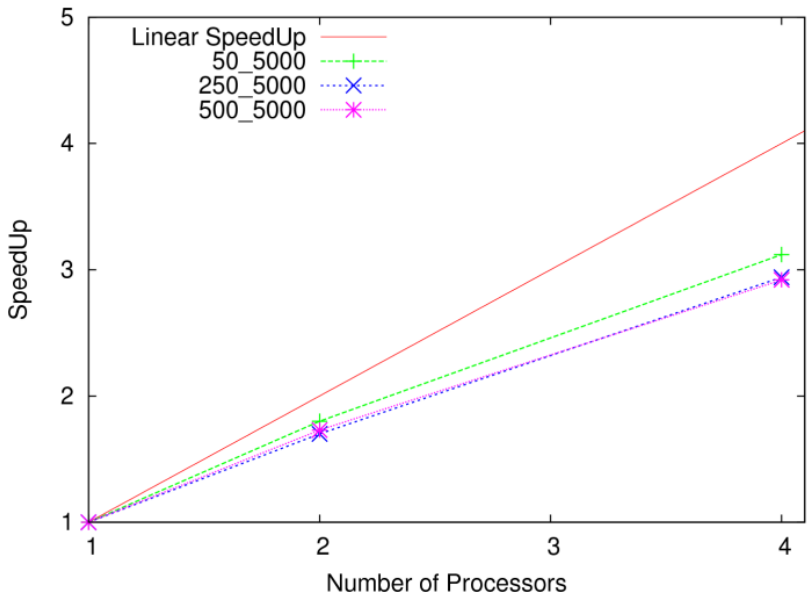
250 species



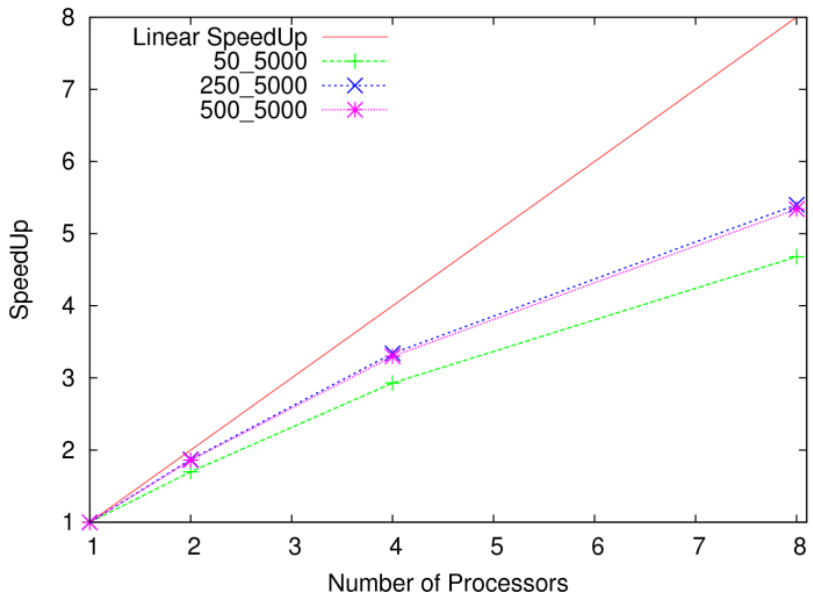
Eight-core Intel Xeon E5420: 1, 2, 4,8 threads

Results: number of species

Quad-core



Eight-core

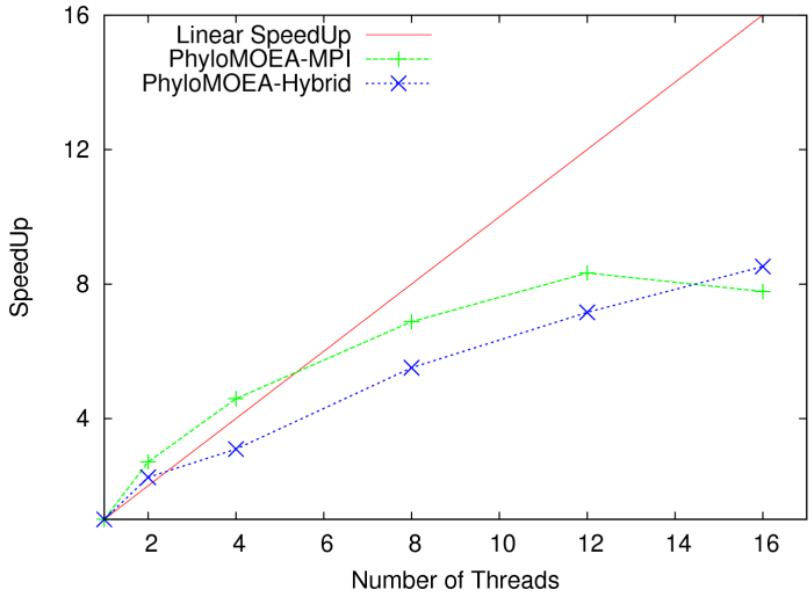


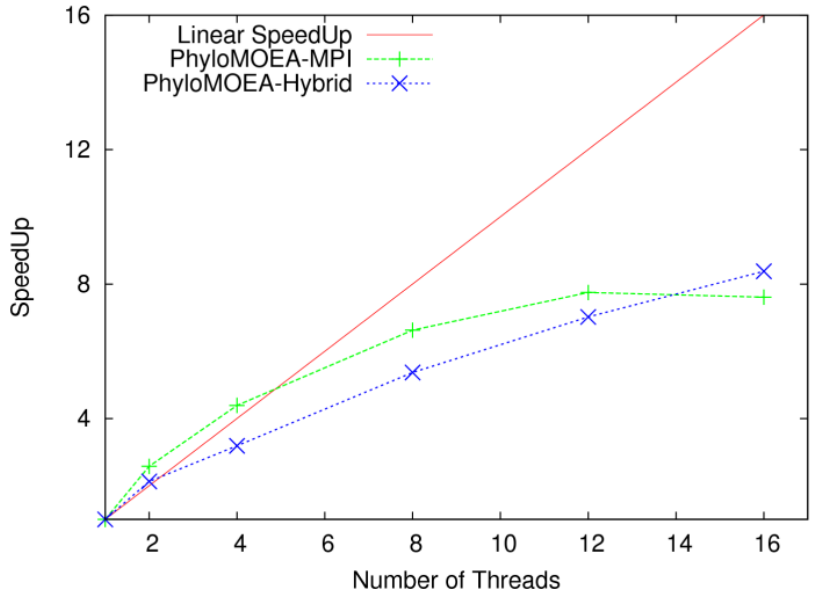
Results: PhyloMOEA scalability

* Serial vs. Parallel
  * PhyloMOEA-MPI
  * PhyloMOEA-Hybrid

* Benchmark datasets
  * rbcL_55: 55 species, 1314 sites
  * mtDNA_186: 186 species, 16608 sites
  * RDPII_218: 218 species, 4182 sites
  * ZILLA_500: 500 species, 1428 sites

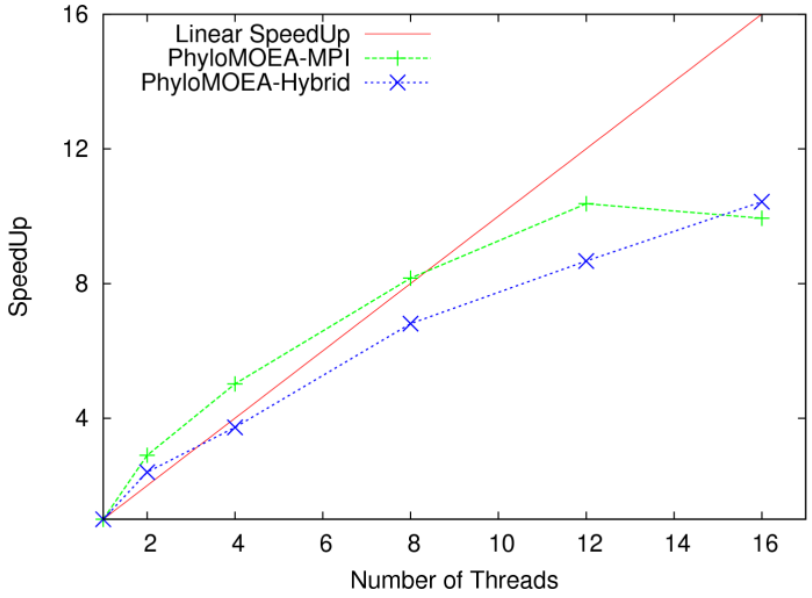
Results: PhyloMOEA scalability



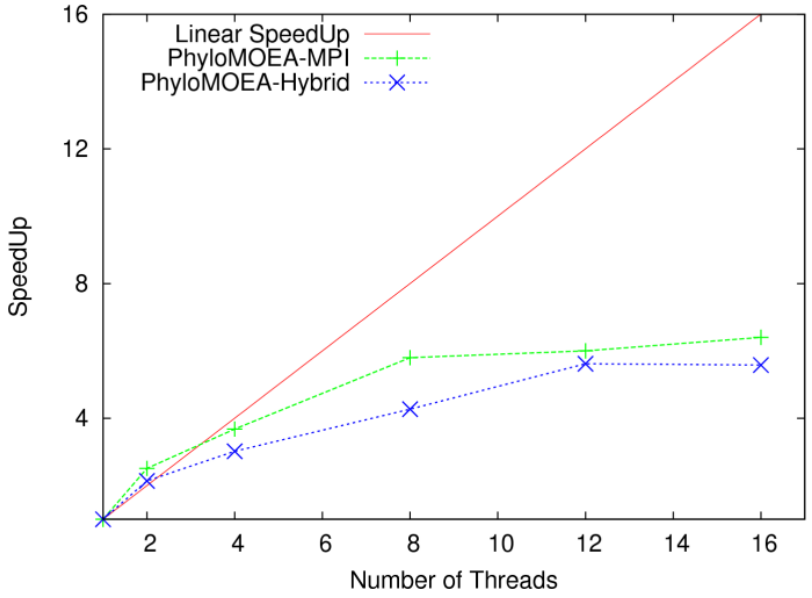


Results: PhyloMOEA scalability

RDPII 218



ZILLA 500



Conclusion

* Superlinear speedup in some cases, sublinear in most of the cases
* Communication and thread synchronization penalize overall speedup
* Total execution time significantly reduced

* Challenges:
  * Data distribution
  * Cooperative search

Parallel multi-objective algorithms for the molecule docking problem

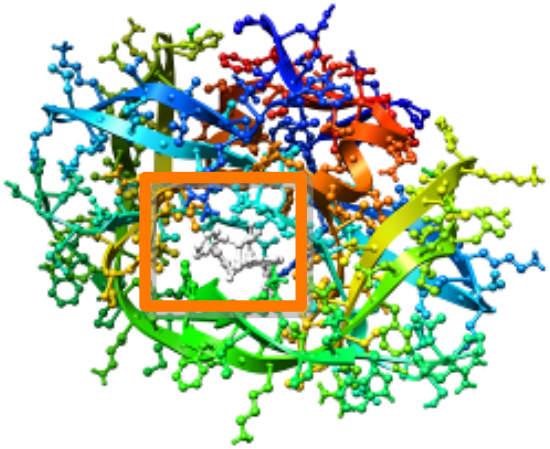
Outline

* Molecular docking.
* ANR Dock project & Docking@GRID.
* A new bi-objective model.
* Multi-objective optimization.
* Algorithm design.
* Comparison results.

Molecular docking

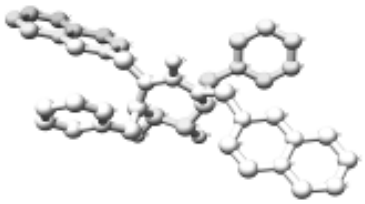
RECEPTOR

LIGAND



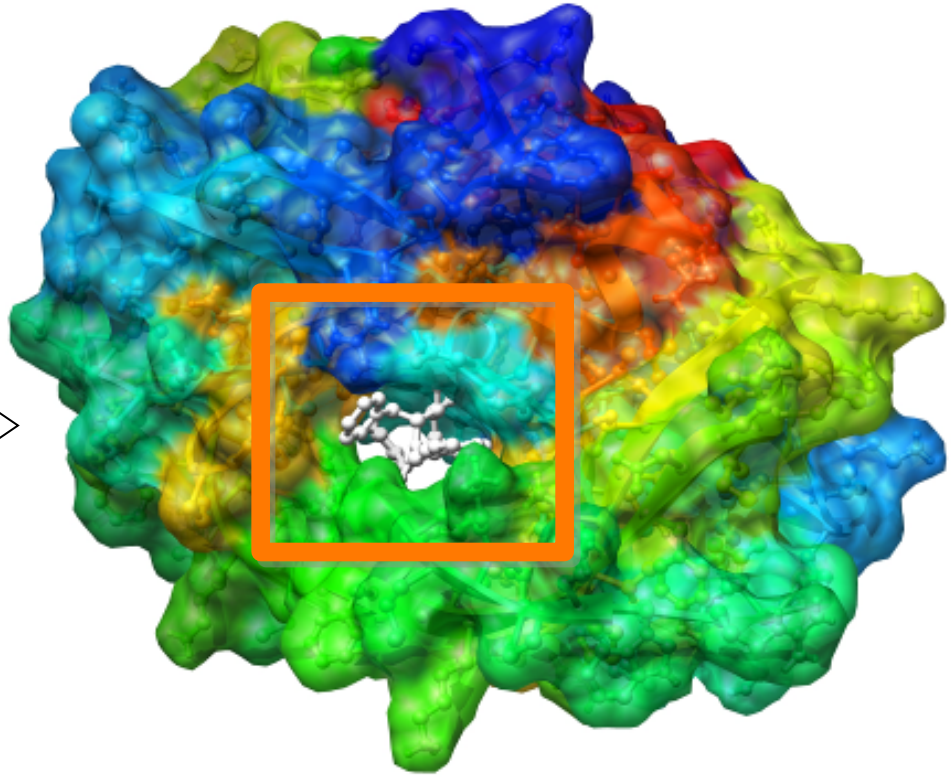
HIV-1 PROTEASE

+



XK263 Inhibitor

HIV-1 PROTEASE + XK263



Molecular docking ⇔ prediction of the optimal complex receptor/ligand according to chemical and geometric properties.

Molecular docking

* Docking simulation :
  * **rigid** ⇔ no conformation modification of the molecules.
  * **semi-flexible** ⇔ one of the two molecules may have its conformation modified during the process (generally the ligand).
  * **flexible** ⇔ conformational modifications for the both molecules
* Several sites can exist for docking the ligand.

Strategy

Bioinformatics problem

~~Datamining task~~

Modelling

Solving

New drug design

Combinatorial Optimization

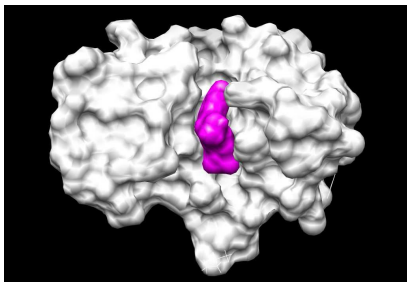
Metaheuristics

Molecular docking

Objective function

Operators

Encoding

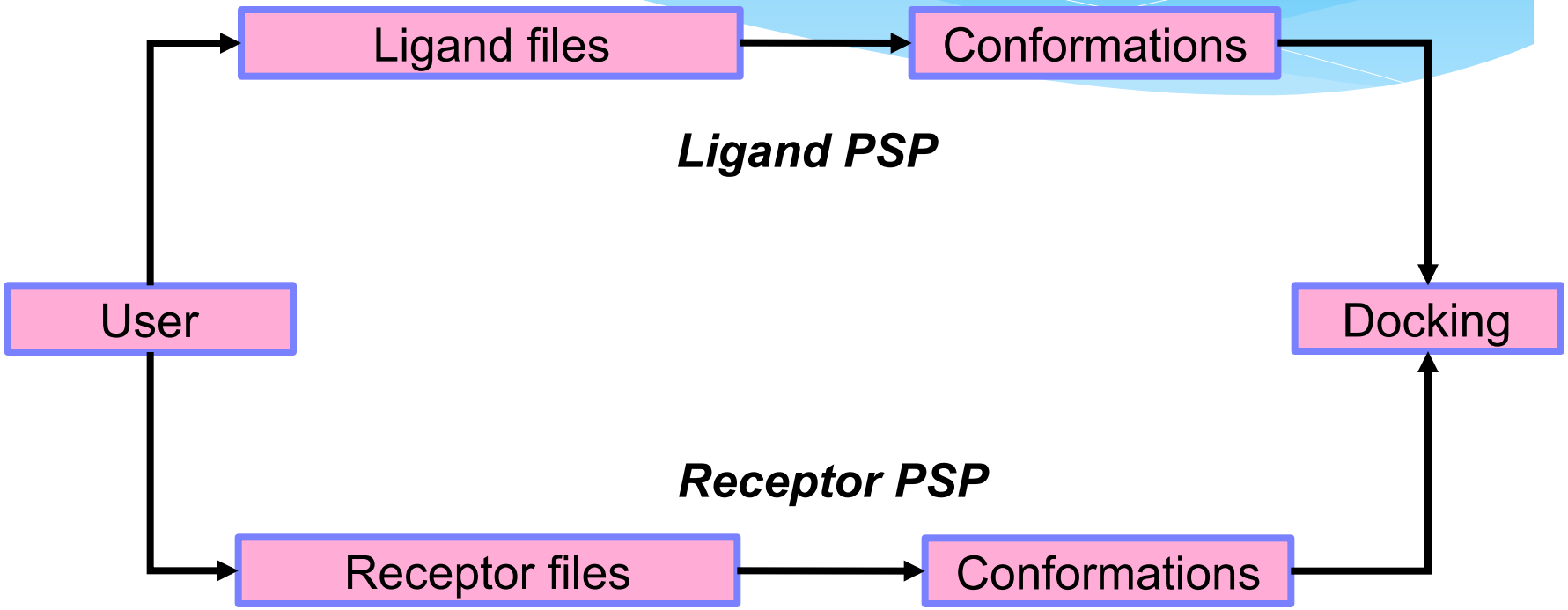


ANR Dock project

* ANR ⇔ french research agency.
* Dock project ⇔ 3-year project about protein structure prediction and docking.
* One of the objectives of the Dock project :
  * Find new multi-objective models for the flexible docking.
  * Implement them with effective (parallel) optimization methods
  * Propose these methods to the community.

Docking@Grid : conformational sampling and docking on grids

PSP = protein structure prediction



More information on :

http://dockinggrid.gforge.inria.fr

A new bi-objective model (1/4)

* Energetic term :
  * Criterium to minimize.
  * Qualify the stability of the complex.

A-A Tantar, N. Melab, E-G. Talbi and B. Toursel. **A Parallel Hybrid Genetic Algorithm for Protein Structure Prediction on the Computational Grid.** Elsevier Science, Future Generation Computer Systems, 23(3):398-409, 2007.

* (Solvent accessible) surface term :
  * Criterium to minimize.
  * Qualify the penetration degree of the ligand into the receptor.

S.M. Le Grand and K.M. Merz, Jr. **Rapid Approximation to Molecular Surface Area via the Use of Boolean Logic and Look-Up Tables**. Journal of Computational Chemistry, 14(3):349-352 (1993).

A new bi-objective model (2/4)

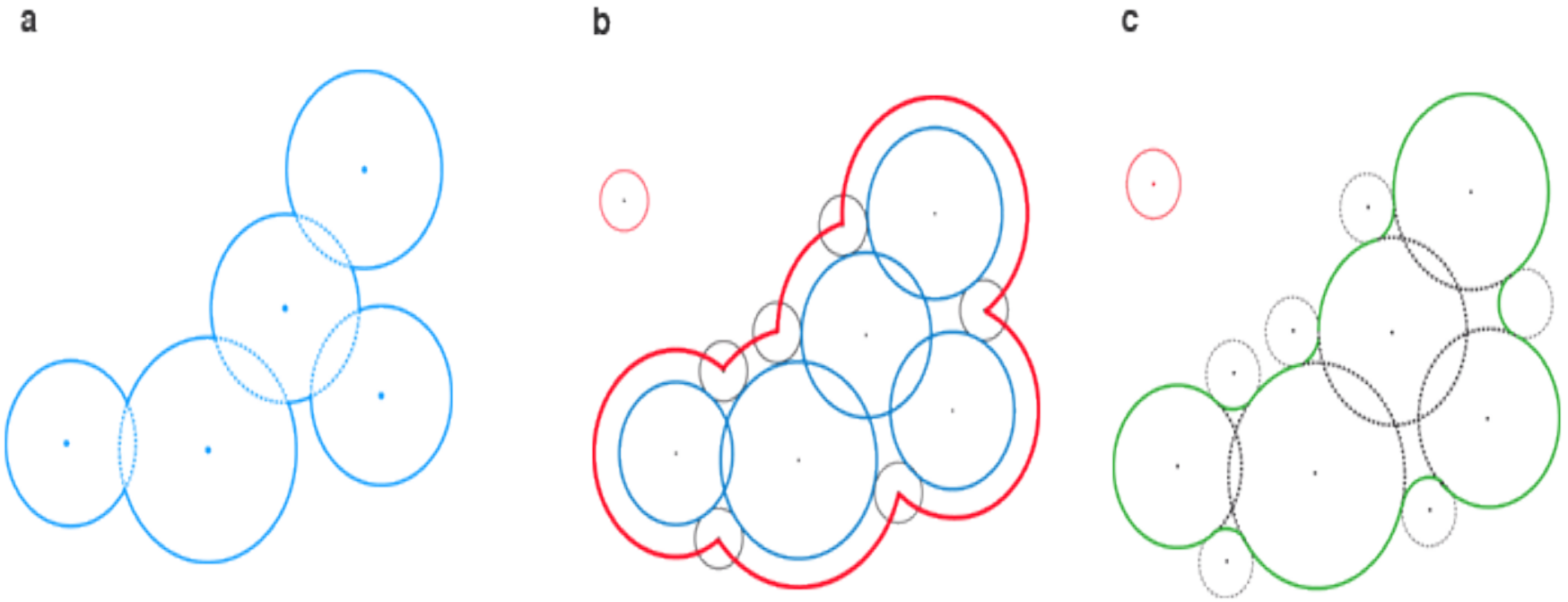
1. Energie of the ligand / receptor complex

$$
\begin{aligned}
E = & \sum_{bonds} K_b(b - b_0)^2 \\
+ & \sum_{angle} K_\theta(\theta - \theta_0)^2 \\
+ & \sum_{torsion} K_\phi(1 - \cos n(\phi - \phi_0)) \\
+ & \sum_{Van\ der\ Waals} \frac{K_{ij}^a}{d_{ij}^{12}} - \frac{K_{ij}^b}{d_{ij}^6} \\
+ & \sum_{Coulomb} \frac{q_i q_j}{4\pi\varepsilon d_{ij}} \\
+ & \sum_{desolvation} \frac{K q_i^2 V_j + q_j^2 V_i}{d_{ij}^4}
\end{aligned}
$$

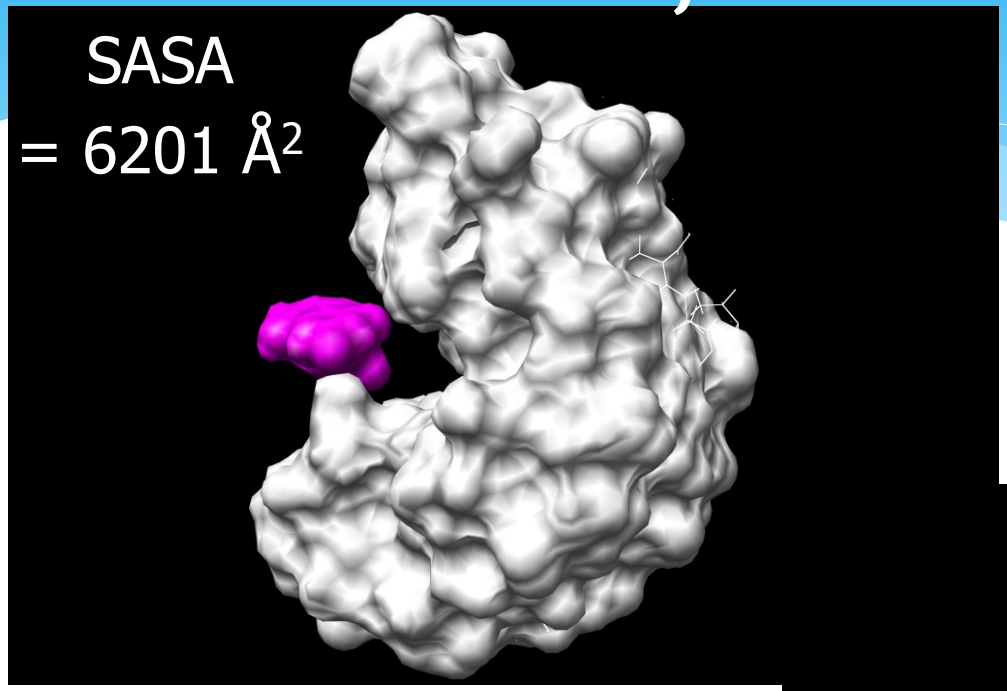
Force field = Consistent Valence Force Field (CVFF)

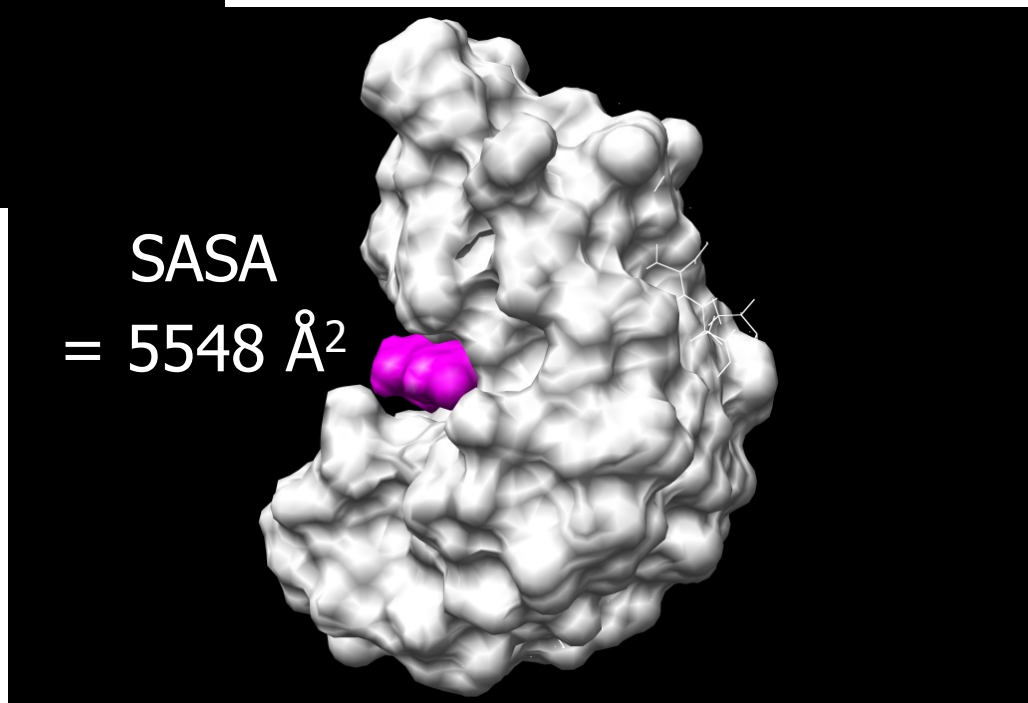
A new bi-objective model (3/4)

* **2. Complex surface**
* Available surfaces :
* ➔ Van Der Waals surface (*a: blue*),
* ➔ Solvent accessible surface (*b: red*),
* ➔ Connoly surface (*c: green*).



A new bi-objective model (4/4)





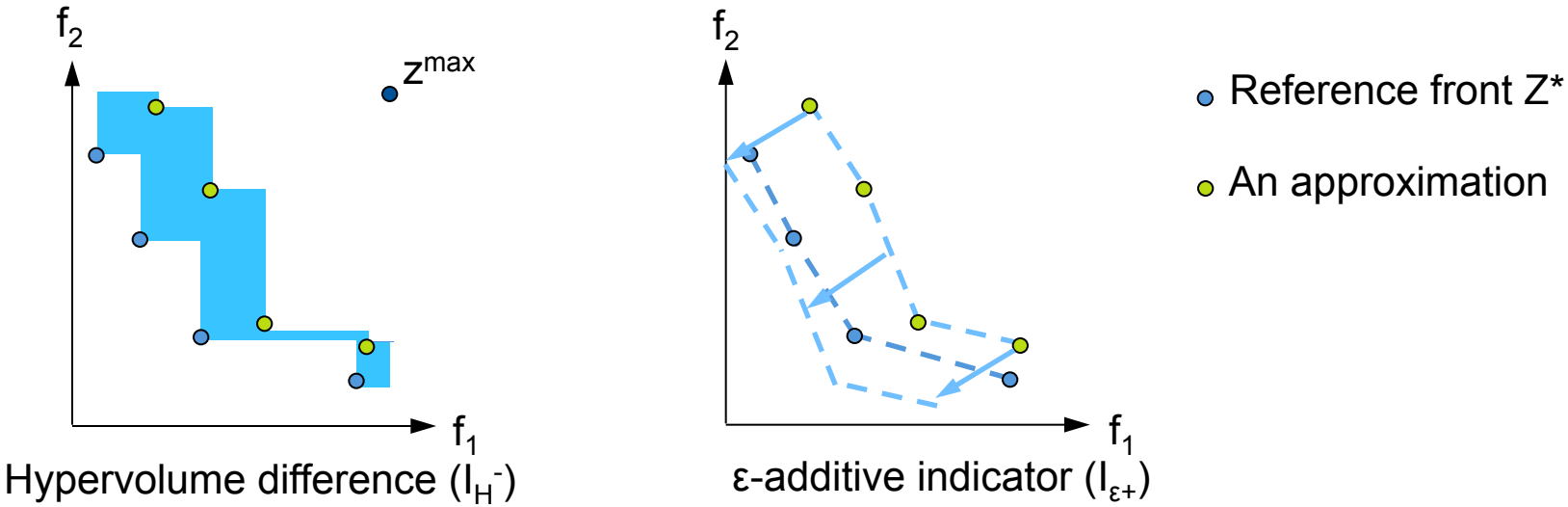
Parallel genetic algorithm : specific parts

* Initial population ⇔ perturbations of “seed” molecules.
* Representation ⇔ two vectors of atomic positions (ligand and receptor).
* Recombination ⇔ ligand swap between receptors.
* Mutation ⇔ molecule rotation and translation (rigid docking) and molecule torsion rotation (semi-flexible and flexible docking).

* Parallel scheme : master/slave paradigm.

Performance assessment (1/3)

- 20 runs per instance and per algorithm.
- Reference front Z*: non-dominated points extracted from all exec.
- Reference point z^{max}:1.05 * upper bound (for all criteria).
- 2 performance metrics.



Performance assessment (2/3)

- For every instances and every search methods, we have 20 I_H^- measures and 20 $I_{\varepsilon+}$ measures.
- For a given instance, how to compare two search methods ?
  - According to the metric under consideration, is a search method A significantly better than a search method B ?
  - Statistical analysis on pairs of search methods.
- Mann-Whitney statistical test :
  - p-value = 0.05.
- Performance assessment achieved using the tool suite of PISA :

  http://www.tik.ee.ethz.ch/pisa/

  [Bleuer et al. 2003]

Performance assessment (3/3)

- Comparison of the final complexes with the crystallographic one.
- Root Mean Square Deviation (RMSD) computation :

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} (dx_i^2 + dy_i^2 + dz_i^2)}{n}}$$

➔with :

➔ n the number of heavy atoms.

➔ dx, dy and dz the deviation between complex and model structures for x, y and z coordinate.

Comparison results (1/2)

Instances from the ccdc astex set

		$I_{\varepsilon+}$		I_h^-	
Instance	**Algorithm**	IBEA	NSGA-II	IBEA	NSGA-II
6rsa	IBEA	-	>	-	>
	NSGA-II	<	-	<	-
1mbi	IBEA	-	>	-	>
	NSGA-II	<	-	<	-
2tsc	IBEA	-	~	-	>
	NSGA-II	~	-	<	-
1htf	IBEA	-	~	-	~
	NSGA-II	~	-	~	-
1dog	IBEA	-	>	-	>
	NSGA-II	<	-	<	-

> ⇔ significantly better

< ⇔ significantly worse

~ ⇔ no significant difference

Comparison results (2/2)

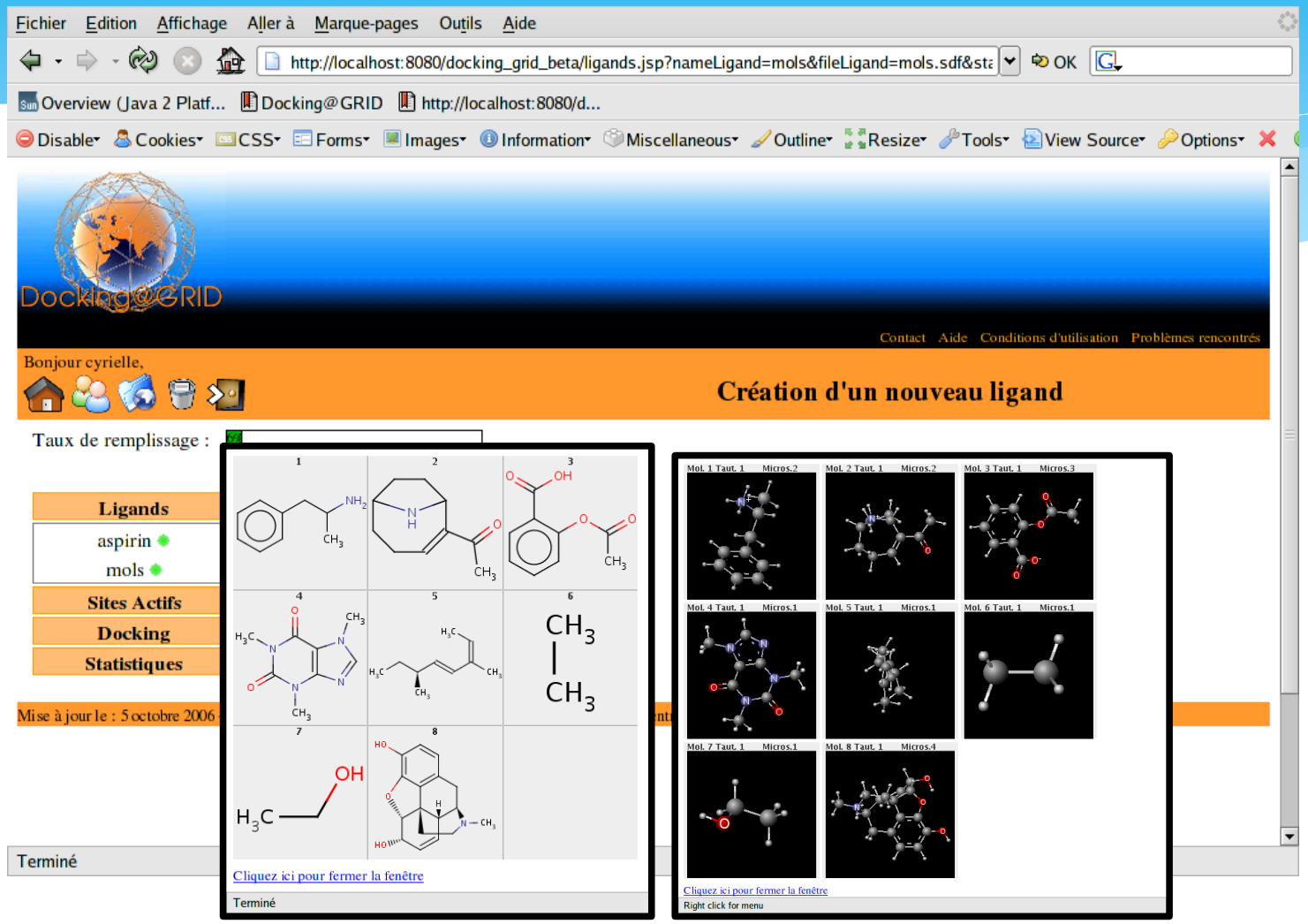
Instances from the ccdc astex set

	NSGA-II		IBEA	
Instance	**RMSD (Å)**	**std**	**RMSA (Å)**	**std**
6rsa	1.66	**1.04**	**1.32**	1.3
1mbi	5.2	**0.4**	**4.16**	0.8
2tsc	**2.19**	2.75	**2.19**	**2.68**
1htf	2.88	2.64	**2.59**	**1.33**
1dog	4.38	0.99	**2.44**	**0.56**

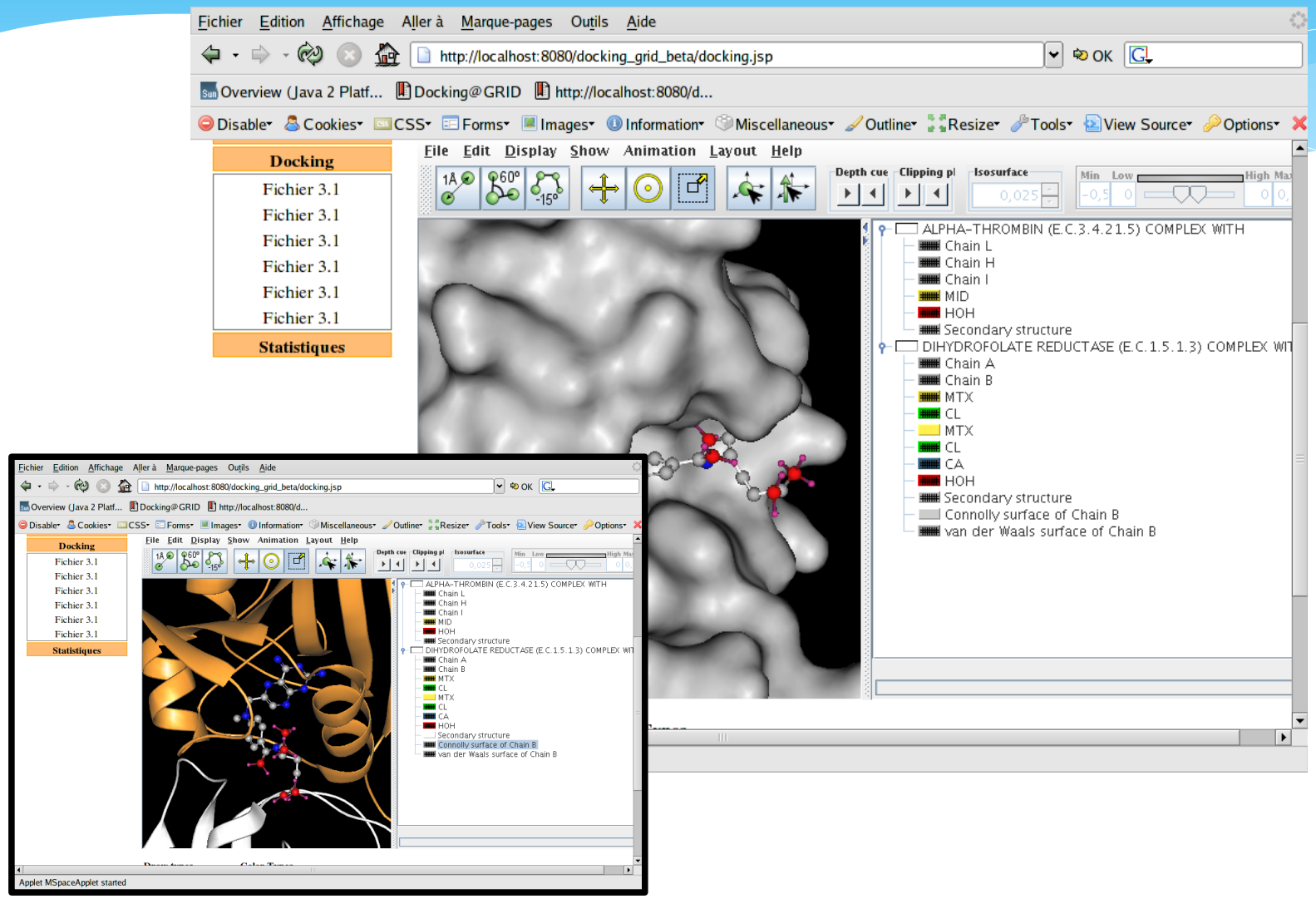
Å ⇔ Angström

std ⇔ standard deviation

Docking @ GRID : WebSite - Ligands



Docking @ GRID : WebSite - Docking



Conclusions and perspectives

* According to the results IBEA is better than NSGA-II for our problem.
* The current model ⇔ the first of 8 models (with CVFF or Autodock 4.0 force field).
* The GA operators ⇔ 8 mutations has been designed (including mono and multi-objective local searches) and are currently tested for each model.
* New instances are currently tested.

Conclusions

* Bioinformatics
  * New model (bi-objective) for the molecular docking
  * In fact 8 models proposed (depending on objective taken into account)
* Computer science
  * According to the results IBEA is better than NSGA-II for this problem
  * GA operators ⇔ 8 mutations has been designed (including mono and multi-objective local searches) and are currently tested for each model

J-C. Boisson, L. Jourdan and E-G. Talbi. "**Single- and multi-objective cooperation for the flexible docking problem**" , Journal of Mathematic modelling and algorithms (JMMA), Vol. 9, n°2, Springer, 2010.

Other problems

Other problems

* Sequence Alignement
* Multiple sequence alignement

} COFFEE and SAGA

→This afternoon during the bibliography study

* Clustering/Biclustering of gene expression profiling
* DNA fragment assembly : build DNA sequence from thousands of overlapping fragments

General conclusion

General conclusion

* **Bioinformatics:**
* Wide area of applications

* **Bioinformatics problems specificities:**
* Huge search space → perfect for Combinatorial optimization
* Problem to define the evaluation of a solution → often an heuristic

* **Metaheuristics:**
* Powerful tools
* Good results compare to other approaches
* Parallelism is a good way to decrease complexity
* BUT must be tailored to the problem