High-Performance Metaheuristics

Application to Bioinformatics and Telecommunications

Prof. El-Ghazali TALBI

OPAC (**Parallel Cooperative Optimization**)

INRIA DOLPHIN Project

University of Lille, France

http://www.lifl.fr/~talbi

Motivations

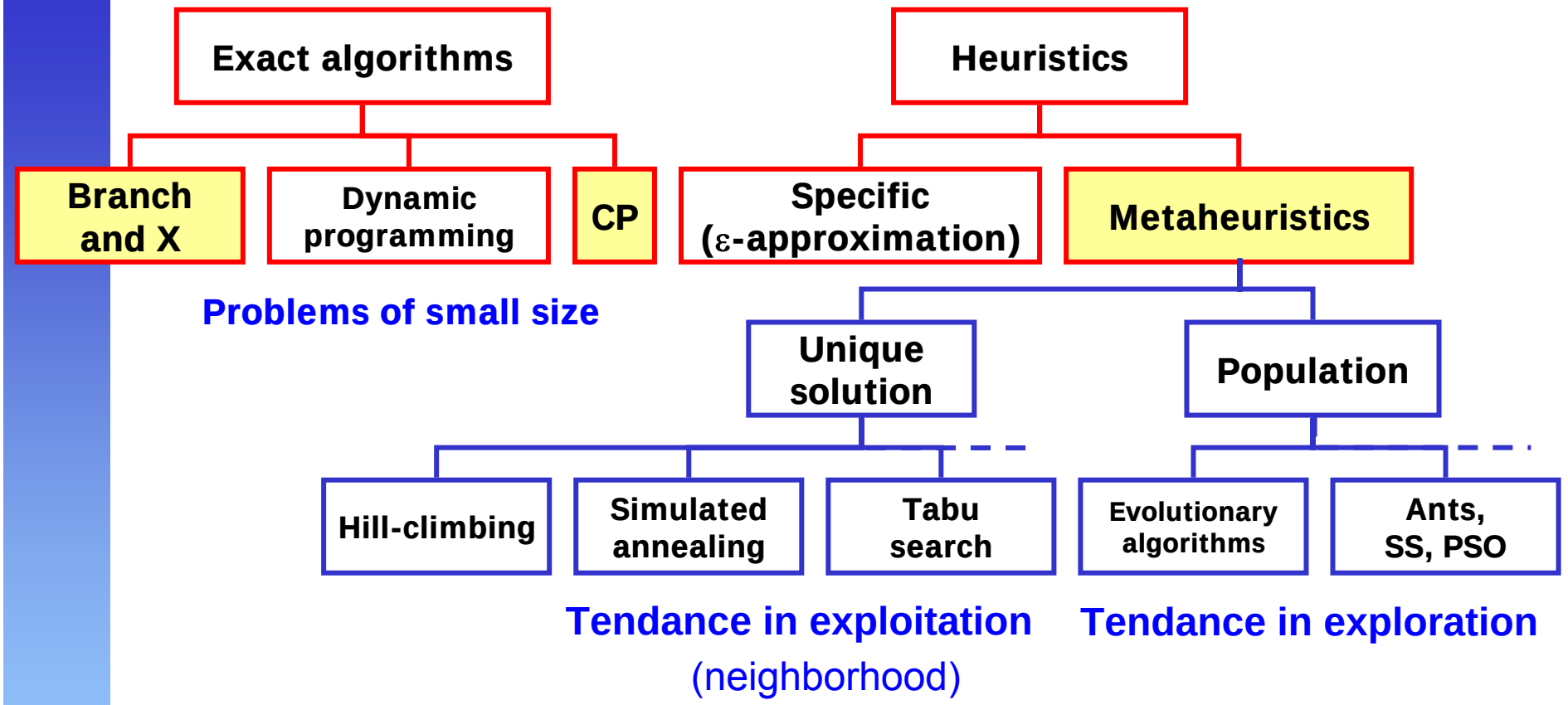
- Large scale multi-objective optimization problems in different industrial domains (Telecommunications, Genomics, Transportation and Logistics, Engineering design, ...).
- Problem size more and more important (combinatorial explosion) et/ou Delays more and more reduced.

$$\textbf{(POC)} \quad \min f(x) \qquad x \in S$$

$$\textbf{(PMO)} \begin{cases} \min f(x) = \left(f_1(x), f_2(x), \ldots, f_n(x)\right) & n \geq 2 \\ \text{s.c. } x \in S \end{cases}$$

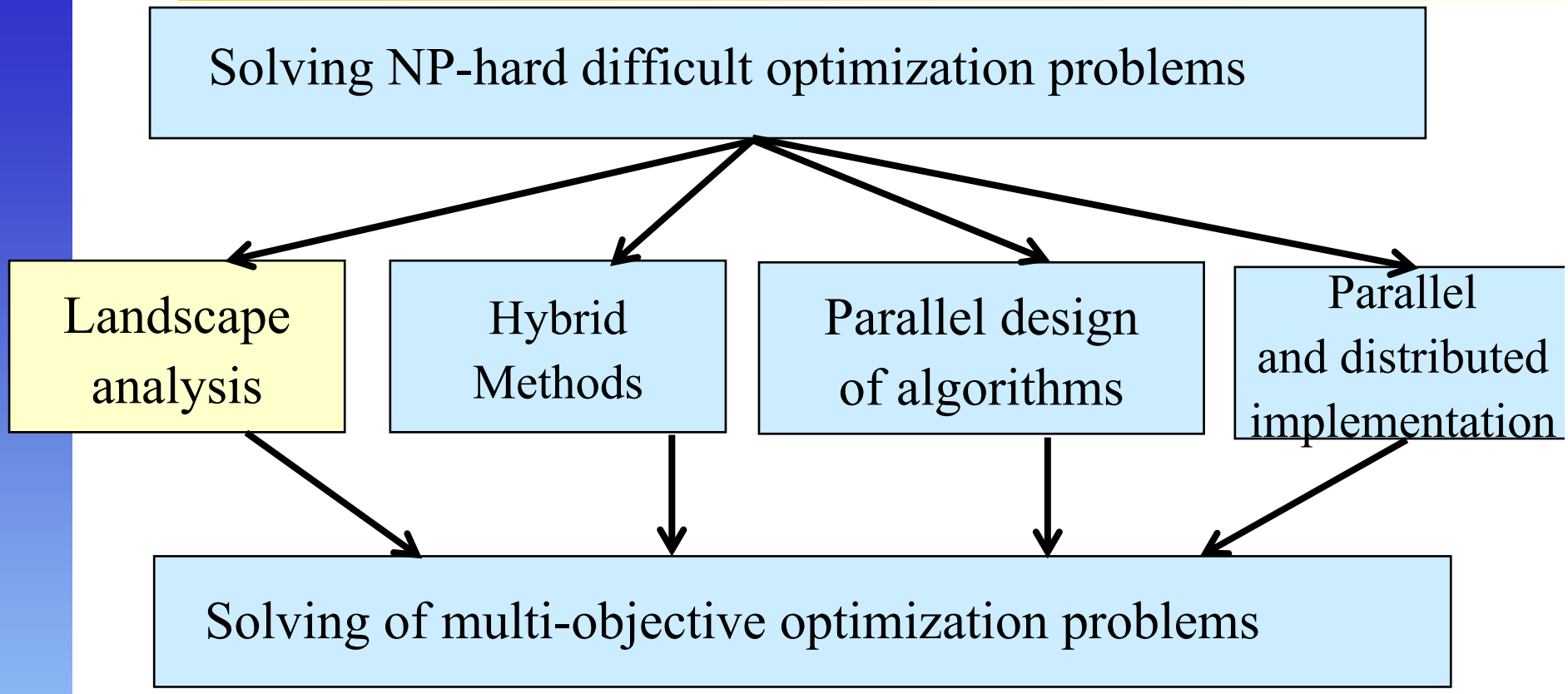
- Objective : Efficient Modelling and Solving of Large (Multi-objective) Combinatorial Optimization Problems

Optimization Algorithms



- **Exact methods are useless** for large problems
- **Metaheuristics are efficient** (lower bound, best known results, …).
- **Lack of theoretical results** (applicable in a real context).

Roadmap



- Motivation : No « super » method

No Free Lunch Theorem : Wolpert - Macready (1995)

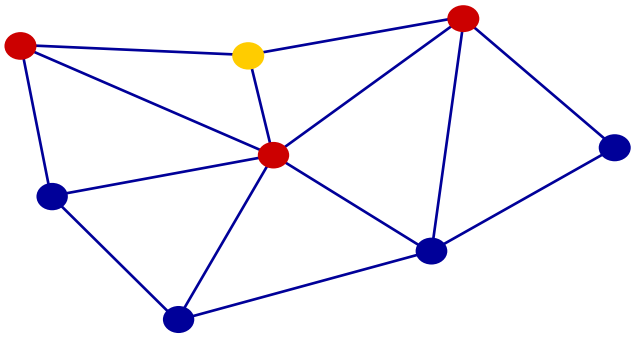
- Better knowledge of problems and instances
- Better knowledge of behavior of optimization methods

Better design of efficient parallel hybrid methods

Landscape analysis

Search space

- Search operator (R)

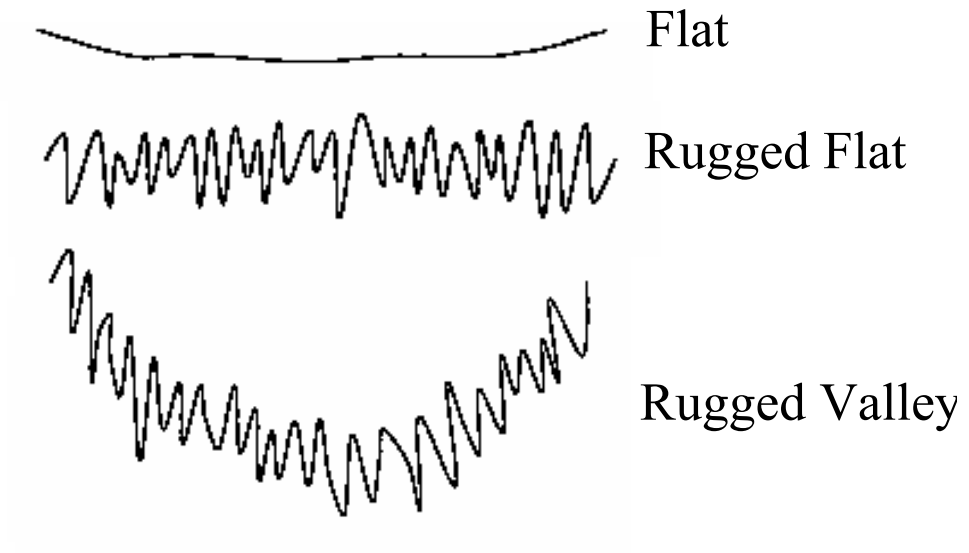
$G_{\mathcal{R}} =$ 

Landscape [Wright 1932, Stadler 1995]

$$L_R = (G_R, f)$$

Geographic Metaphor

(basins, valleys, ridges, plateau...)



- **Global approach** [Weinberger 90] / **Local approach**
  - Characteristics of the landscape shown by heuristic (local search)
- **Statistical indicators** : Distribution of local optima, Rugosity, FDC, Autocorrelation, Entropy, Amplitude, ...

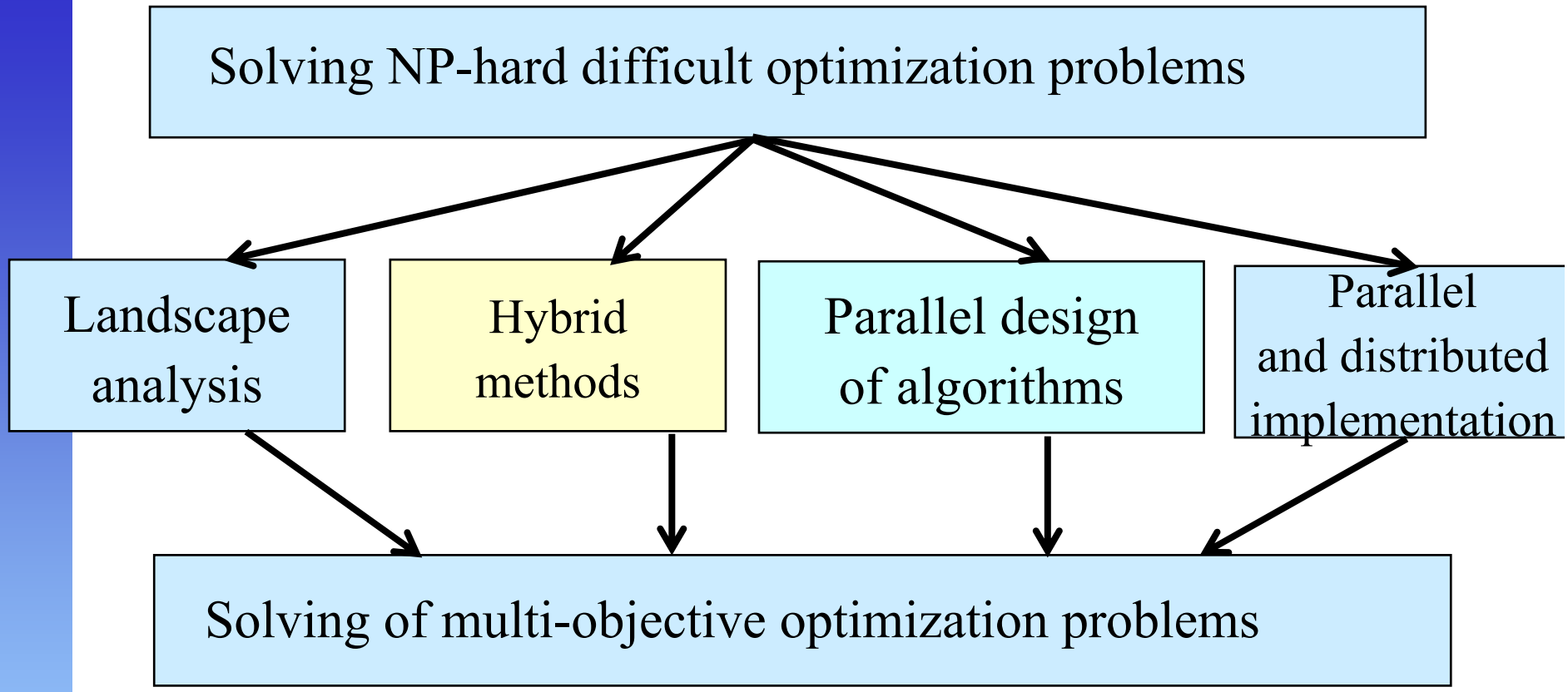
• B. Weinberg, E-G. Talbi, « **Fitness Landscape for metaheuristic design** », *Combinatorial Optimization, 2003.*

Landscape analysis : Some results

- Operators : TSP (city swap / 2-opt), Constrained spanning tree, ...
- Objective function: Scheduling-makespan (flat → smooth), ...
- Encoding: Graph coloring (symmetry problem)
- Types of instances: Quadratic assignment problem (uniform random, structured)
- Distances (Intensification, Diversification): Vehicle routing, ...
- Hybridization-Parallelization: When ?, Which ?
- On-line Adaptive methods : Operators, Methods, ...
- Multi-objective optimization: Improve the model

• *E-G. Talbi et al*, **"Fitness landscape and performance of metaheuristics"**, *Kluwer Academic Press,1999.*

Roadmap



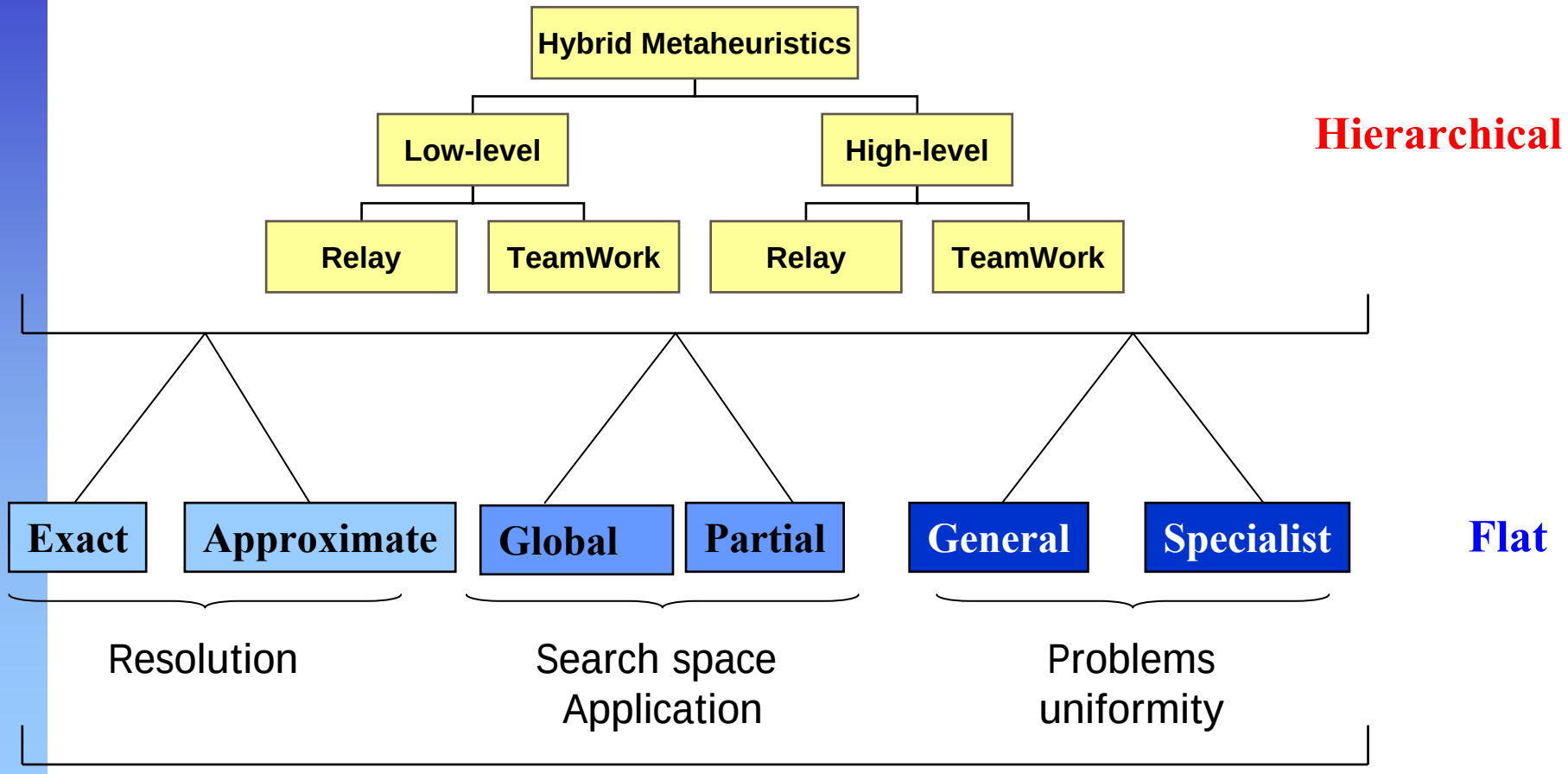
Hybridize = Cooperation of algorithms with complementary behaviour (exploration / exploitation)

Best found solutions for many generic and real-life problems

Hybrid Metaheuristics

➲ Why Hybridation ?

- Equilibrate exploration / exploitation : Gain in performance, Robustness
- Taxonomy : Grammar – Large variety of classified methods



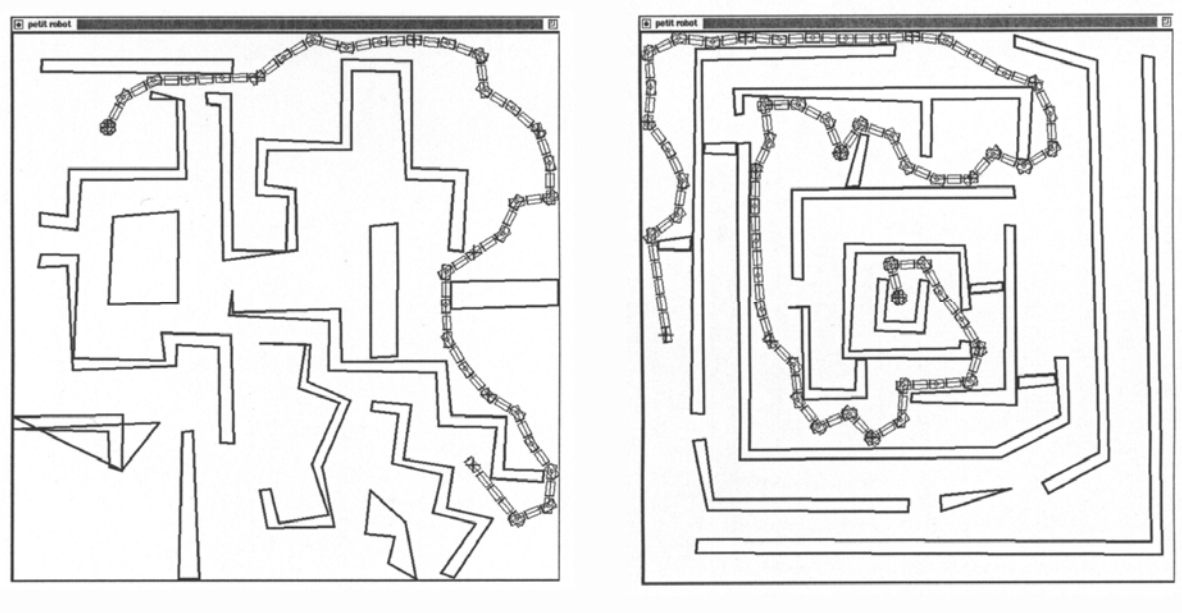
• E-G. Talbi, « **A taxonomy of hybrid metaheuristics** », *Journal of heuristics, 8(5), 2002.*

Robot path planning

[E-G. Talbi, P. Bessière, E. Mazer (France), J-M. Ahuactzin (Mexique), 1994]

(European project PAPAGENA)

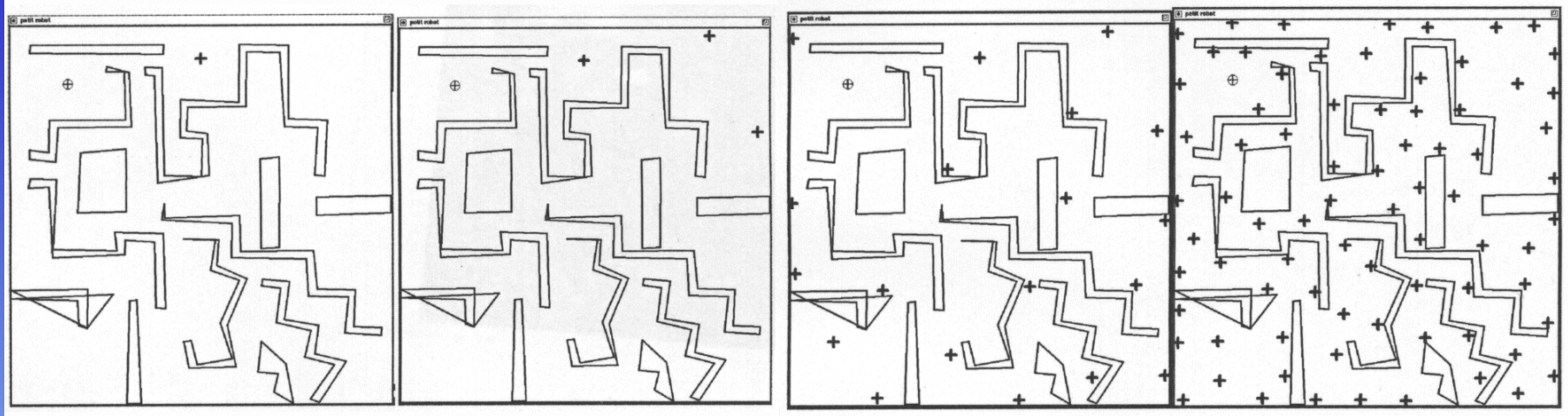
Search Agent SEARCH (local optima)



$$Min_{M \in S_s} F_S(M, \tau) = \begin{cases} 0 & \text{If a direct movement exist from } m_i \text{ to } \tau \ (i<a) \\ Min_{i=,a-1} \|\tau - m_i\| & \text{Else} \end{cases}$$

Robot path planning

Diversification agent EXPLORE



$$Min_{M \in S_e} F_e(M) = Max_{M \in S_e} Min_{\lambda_k \in \Lambda} \|m_a - \lambda_k\|$$

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$$

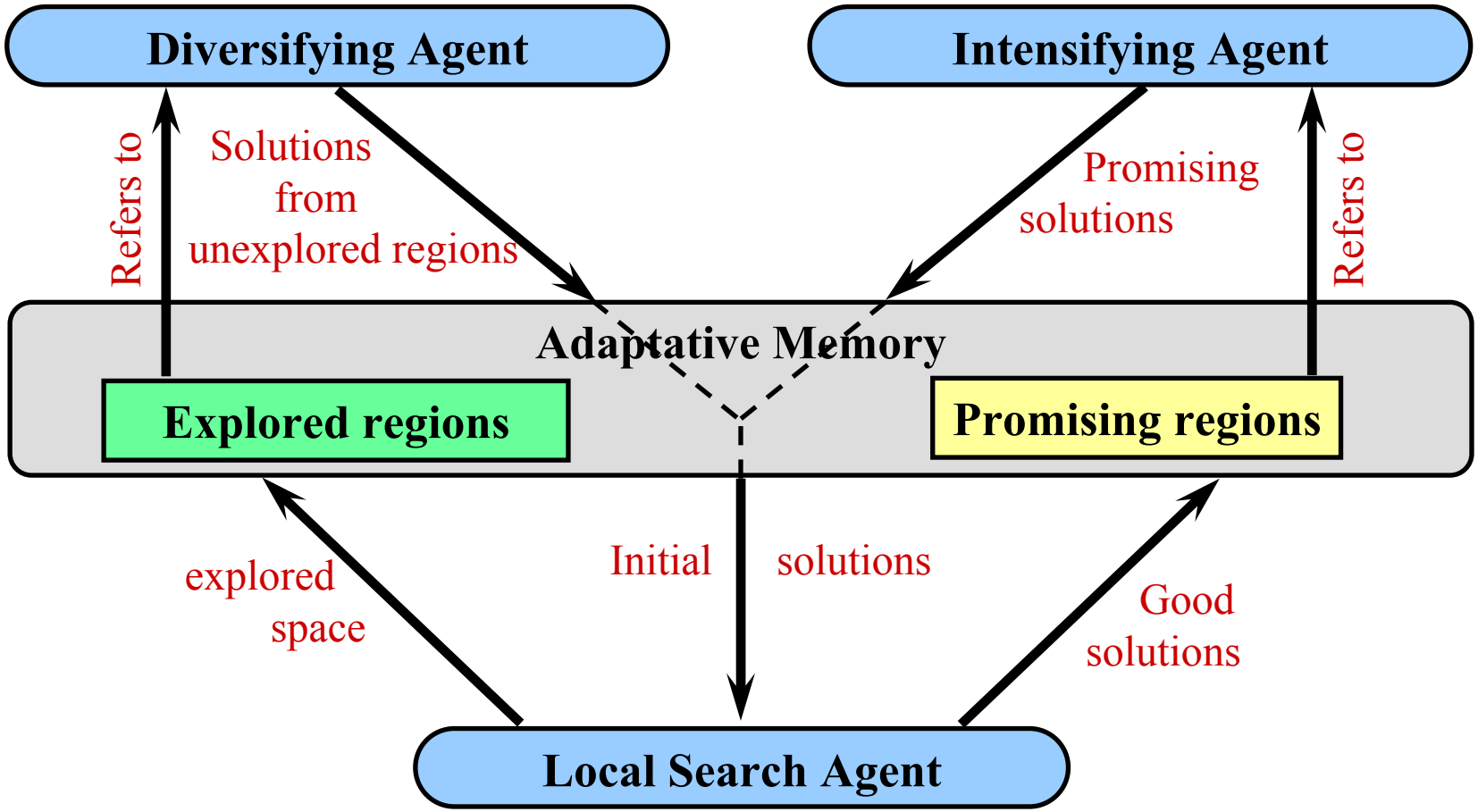
Set of landmarks

SEARCH (Parallel GA) ⟷ **EXPLORE (Parallel GA)**

Fil d'Ariane (Real robots with Aleph Technologies)

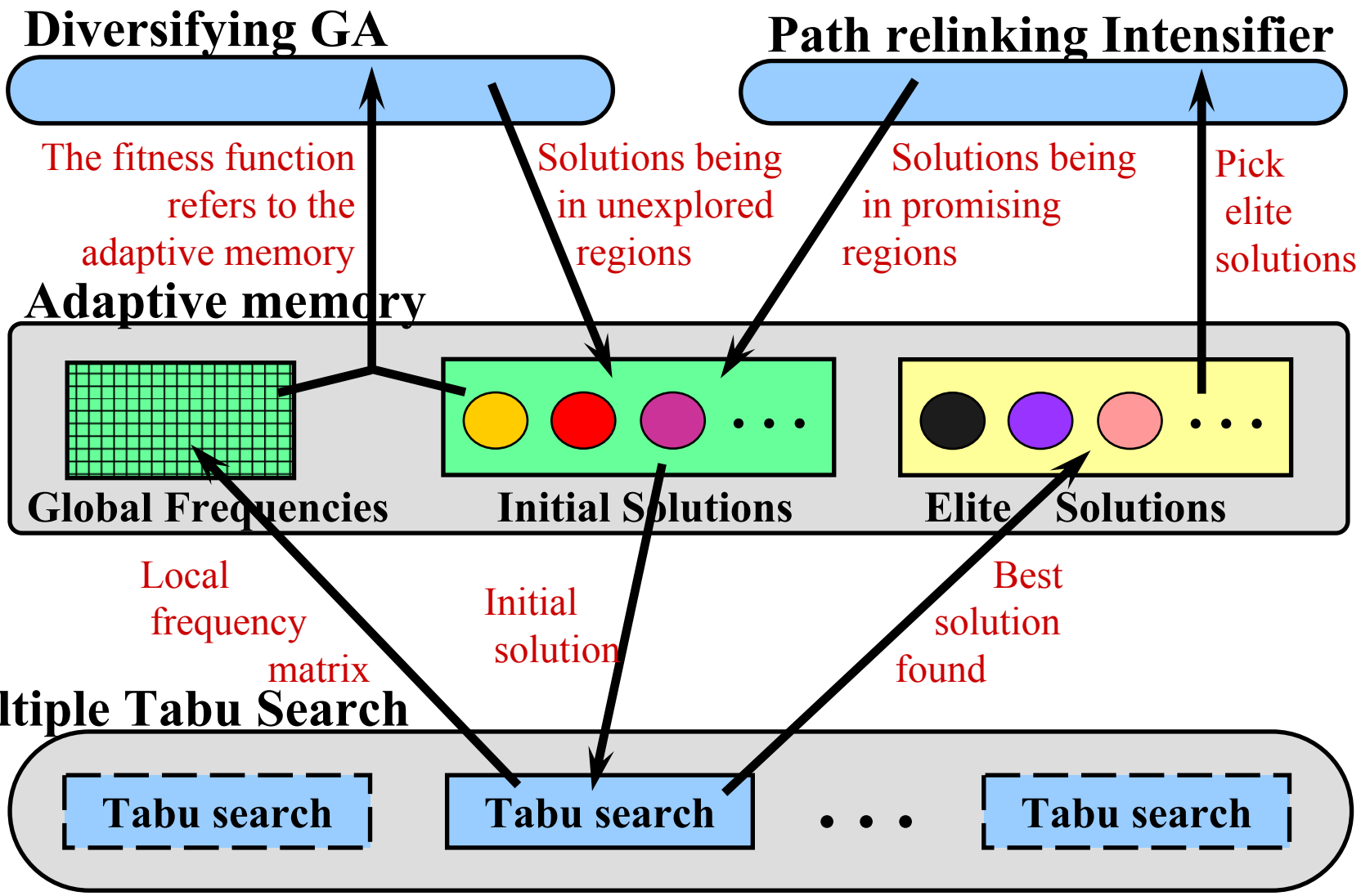
Hybrid Metaheuristics

➲ **COSEARCH : 3 complementary agents cooperate via an adaptive memory**



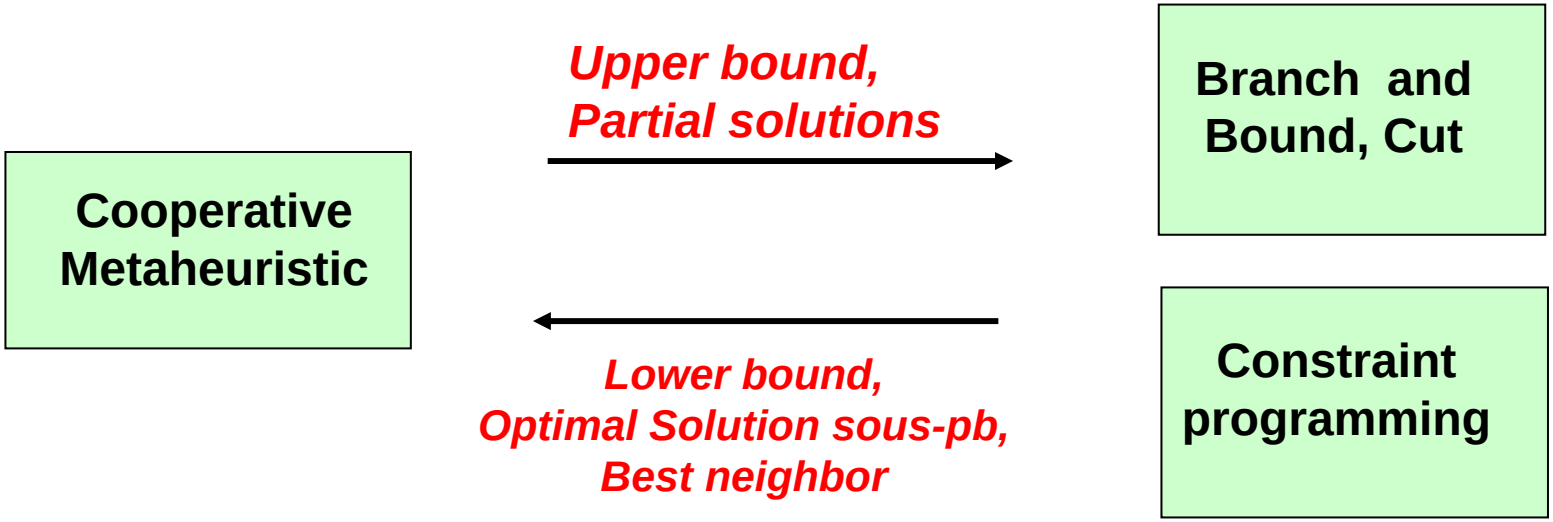
E-G. Talbi et al. **COSEARCH: A parallel cooperativ metaheuristic.** Journal of Mathematical Modeling and algorithms JMMA, Vol.5(2), 2006.

COSEARCH for QAP



Cooperation Meta-Exact

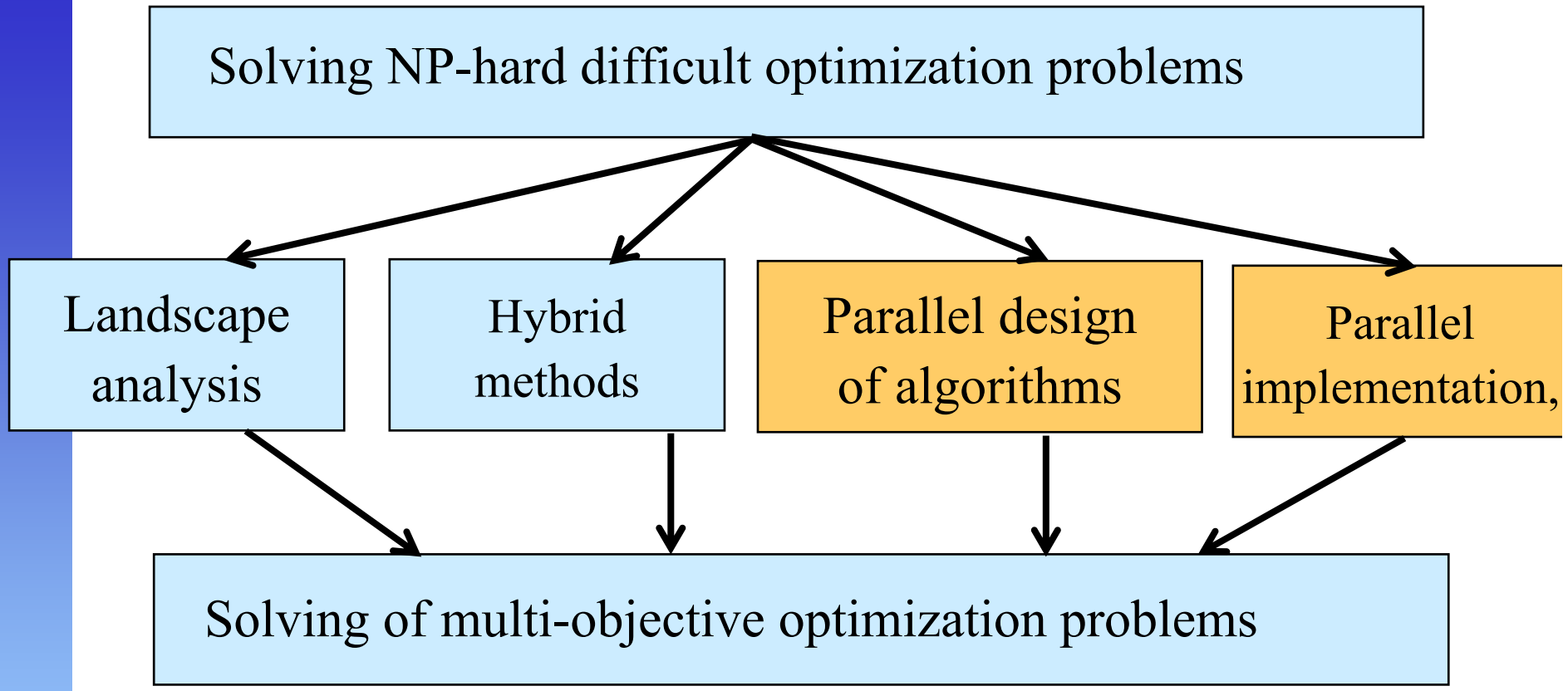
- Heuristic approach :
  - VLNS : Very Large Neighborhood Search (use of an exact method)
  - Generation of different Sub-Problems solved by an exact method
- Exact approach :
  - Good solutions found by metaheuristics to reduce the visited search space for exact methods



• M. Basseur, J. Lemesre, C. Dhaenens, E-G. Talbi, **«Cooperation between branch and bound and evolutionary algorithms to solve a bi-objective flow-shop problem»**, *WEA'2004, LNCS, Springer, 2004.*

• M. Basseur, L. Jourdan E-G. Talbi, **«Cooperation between exact methods and metaheuristics : A survey»**, *EJOR Journal.2007*

Roadmap

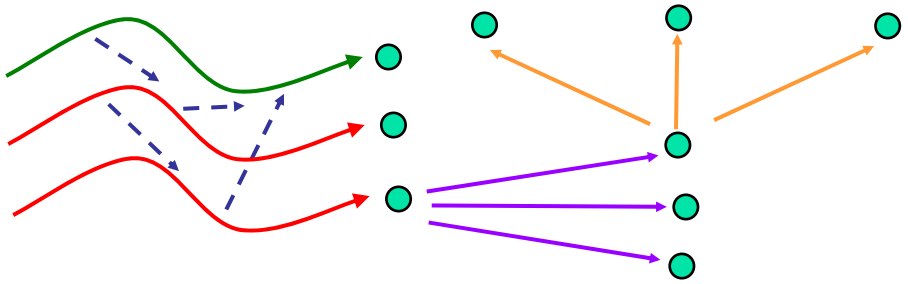


- Speedup the search time
- Solving large and complex problems
- Improve the quality of obtained solutions
- Improve the robustness

Parallel Metaheuristics: Design issues

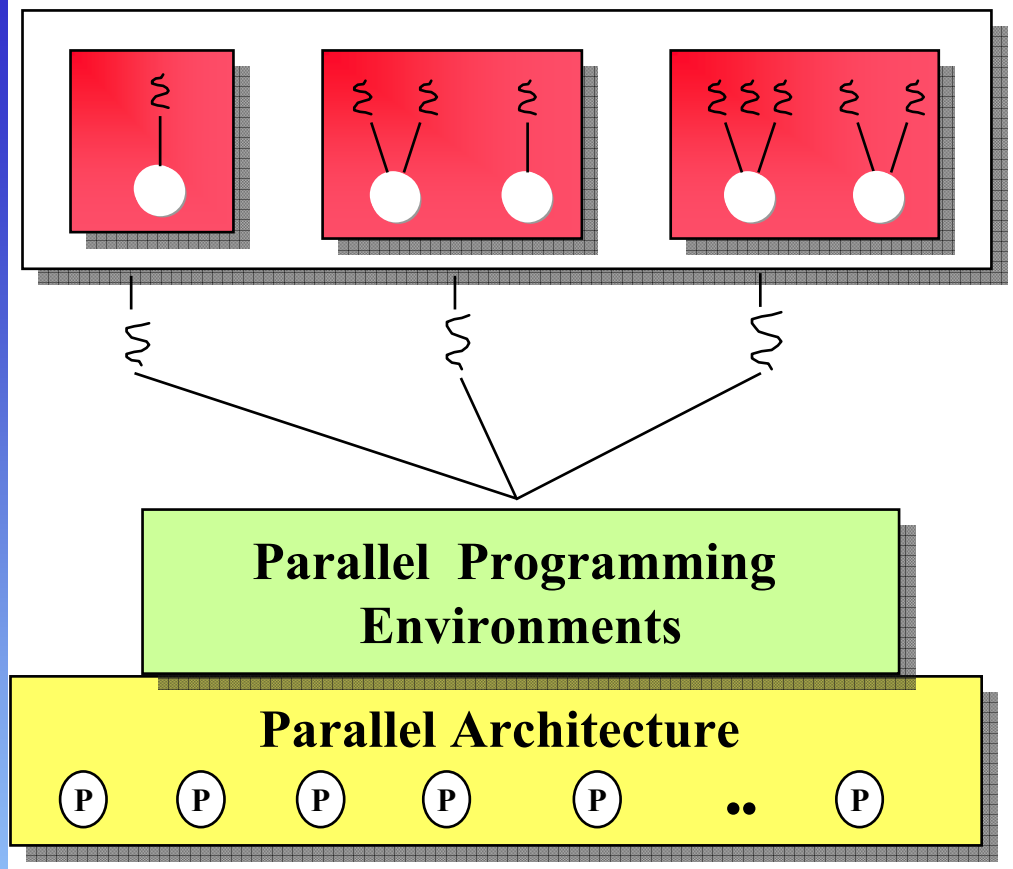
- An unified view for single-based metaheuristics and population based metaheuristics

- Three major hierarchical models:
  - **Algorithm-Level**: Independent/Cooperative self-contained metaheuristics: Problem independent
  - **Iteration-Level**: Parallelization of a single step of the metaheuristic (based on distribution of the handled solutions): Problem independent
  - **Solution-Level**: Parallelization of the processing of a single solution (objective function, constraints, ...): Problem dependent



E. Alba, E-G. Talbi, G. Luque; N. Melab. **Metaheuristics and Parallelism**. In Wiley Book on "Parallel Metaheuristics: A New Class of Algorithms", 2005.

Parallel Metaheuristics: Implementation issues



Design of Parallel Metaheuristics

Programming Paradigms

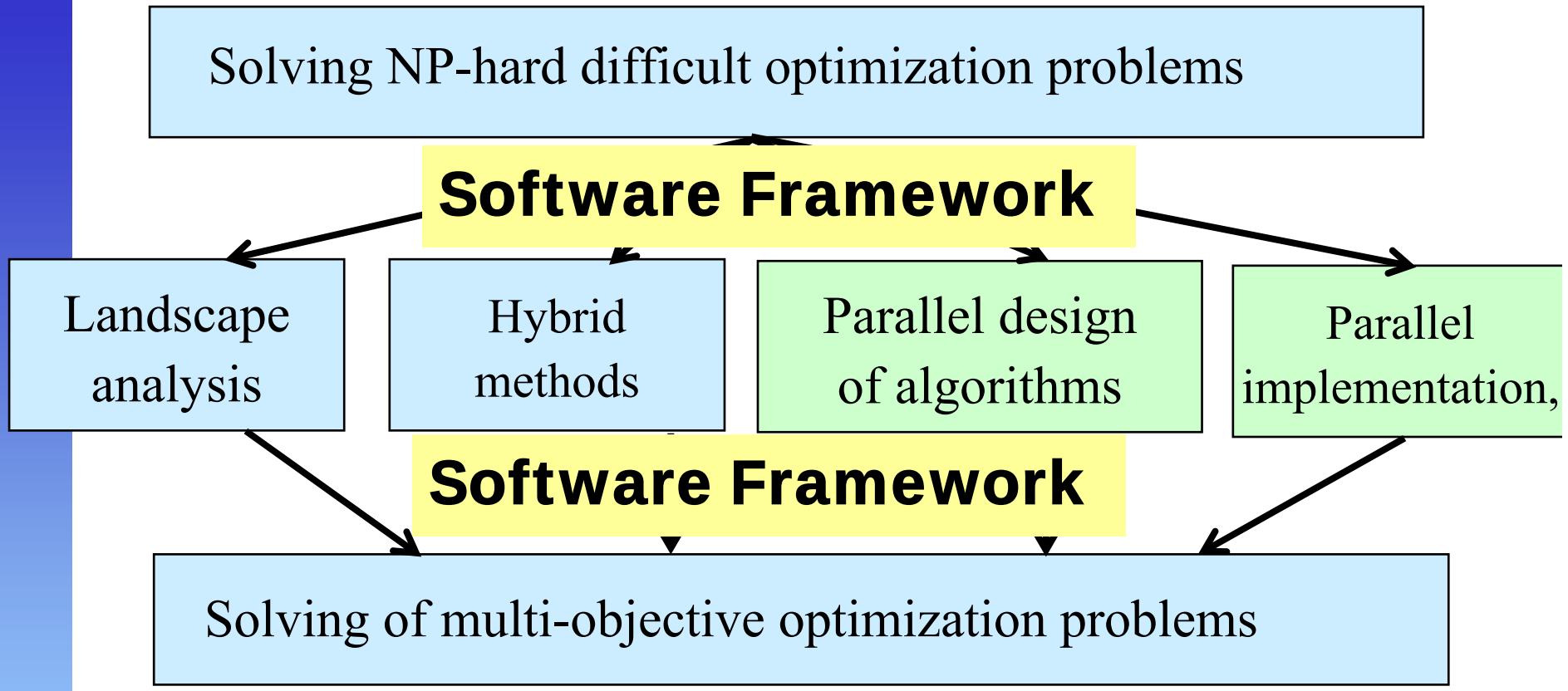
Execution Support

Hardware

Main criteria : Memory sharing, Homogeneity, Dedicated, Scalability, Volatility

Main problems: Load balancing, Fault-tolerance, ...

Roadmap



- From scratch: high development cost, error prone, difficult to maintain, ...
- Code reuse: difficult to reuse, adaptation cost, ...
- Design and code reuse – software components: Hollywood principle « Don't call us, we call you »

Design Objectives

- Maximal Reuse of code and design
  - Separation between resolution methods and target problems
    - Invariant part given
    - Problem specific part specified but to implement
- Flexibility et Adaptability
  - Adding and updating other optimization methods, search mechanisms, operators, encoding, ...
  - ... to solve new problems
- Utility
  - Large panel of methods, hybrids, parallel strategies, ...
- Portability
  - Deployment on different platforms (Standard library)
- Transparent access to performance and robustness
  - Parallel implementation is tranparent to the target hardware platform
- **Open source, Efficiency, Easy to use, ...**

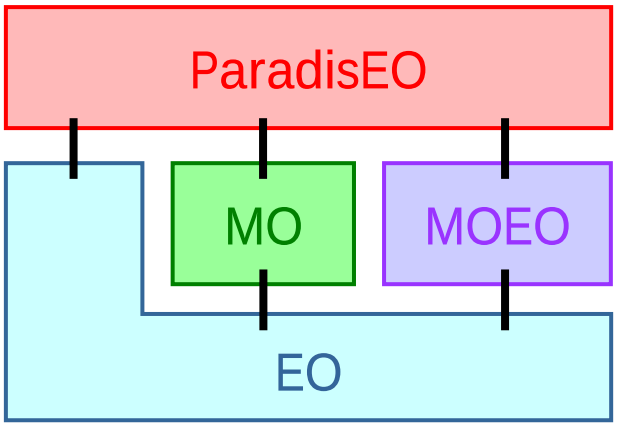
PARADISEO (PARAllel and DIStributed Evolving Objects)

- PARADISEO in some words ...
  http://paradiseo.gforge.inria.fr

- An Open Source C++ framework (STL-Template)
- Paradigm-free, unifying metaheuristics
- Flexible regards the tackled problem
- Generic and reusable components (operators of variation, selection, replacement, criterion of termination, ...)
- Many services (visualization, management of command line parameters, check-pointing, ...)

• S. Cahon, N. Melab, E-G. Talbi, M. Schoenauer **«PARADISEO based design of parallel and distributed evolutionary algorithms »,** *Parallel Computing journal, 2006.*

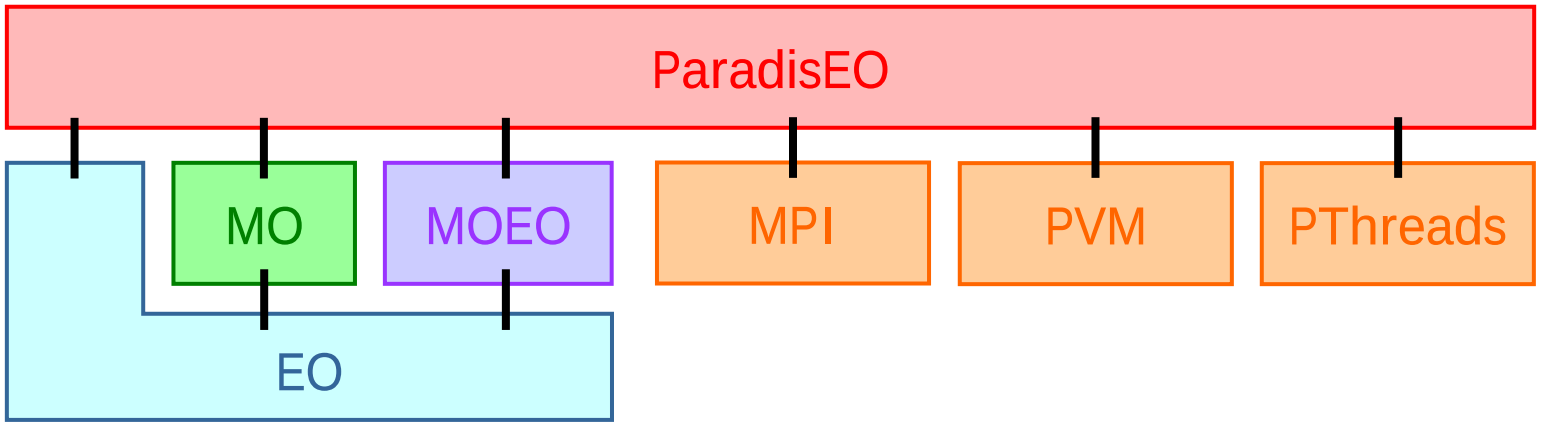
Architecture (level of design)



- Evolving Objects (EO) for the design of population-based metaheuristics *(EA, PSO, EDA, ...)*
- Moving Objects (MO) for the design of solution-based metaheuristics *(LS, Tabu search, Simulated annealing, ILS, ...),*
- Multi-Objective EO (MOEO) embedding features and techniques related to multi-objective optimization *(NSGA-II, IBEA, MOGA, SPEA, ...),* ***[EMO'07]***
- ParadisEO for the parallelization and hybridization of metaheuristics

S. Cahon, N. Melab and E-G. Talbi. **ParadisEO: A Framework for the Reusable Design of Parallel and Distributed Metaheuristics.** Journal of Heuristics, Vol.10(3), 2004.

Architecture (level of execution)
Parallel and distributed platforms



- Parallelism and distribution
  - Communication libraries (MPI, PVM)
    → Deployment on networks/clusters of stations (COWs, NOWs)
  - Multi-threading layer (Posix threads)
    → multi-processors with shared memory (SMPs)
  - Support of parallel and distributed environments
    → Clusters of SMPs (CLUMPS)
  - Transparent to the user

S. Cahon, N. Melab and E-G. Talbi. **Building with ParadisEO reusible parallel and distributed evolutionary Algorithms,** Journal of Parallel Computing, , Vol.30(5), 2004.

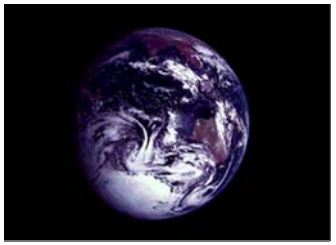
Grid Computing

"Coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations"

High-Performance Computing GRID

- Offer a virtual supercomputer

High-Throughput Computing GRID

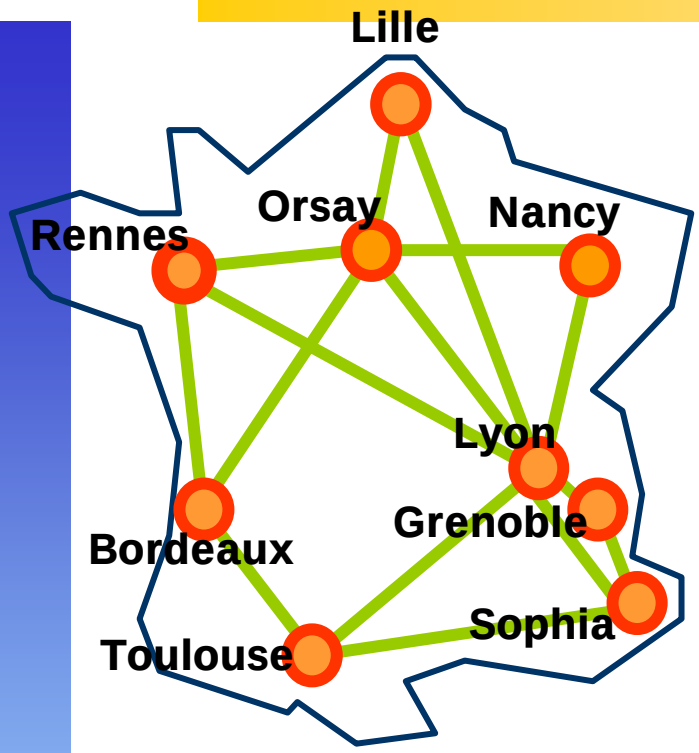


- Billions of idle PCs ...
- Stealing unused CPU cycles of processors (a mean of 47%)

- Inexpensive, potentially very powerful but more difficult to program than traditional parallel computers

N. Melab, S. Cahon, E-G. Talbi, **« Grid computing for parallel bioinspired algorithms»,** *Journal of Parallel and Distributed Computing (JDPC), 66(8), 2006.*

GRID Platforms

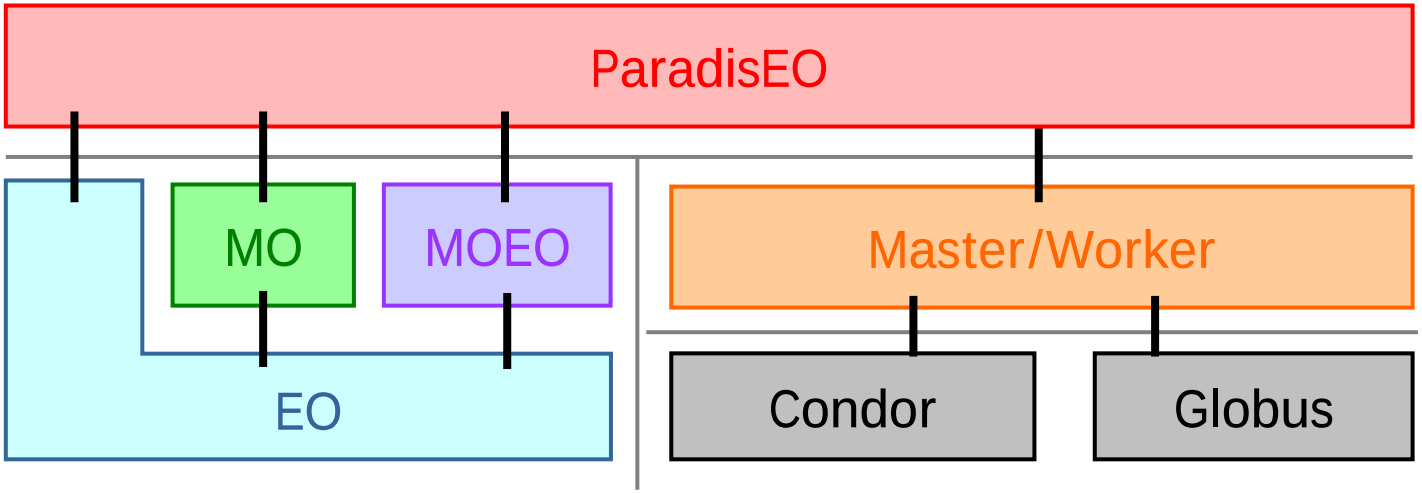


HPC Grid: GRID'5000: 9 sites distributed in France and inter-connected by Renater 5000 proc: between 500 and 1000 CPUs on each site

HTC Grid: PlanetLab: 711 nodes on 338 sites over 25 countries

R. Bolze, ..., E-G. Talbi, ..., **«GRID'5000: A large scale and highly reconfigurable Grid»,** *International Journal of High Performance Computing Applications (IJHPCA), Vol.20(4), 2006.*

Architecture (level of execution) Metacomputing grids



- Gridification
  - Re-visit parallel models taken into account the characterisitics of Grids
  - Coupling of ParadisEO with a Grid middleware (Condor-MW and Globus)
- Transparent volatility & checkpointing
  - Ex : Definition in ParadisEO-CMW of the memory of each metaheuristic and the associated parallel models

N. Melab, S. Cahon and E-G. Talbi. **Grid Computing for Parallel Bioinspired Algorithms.** Journal of Parallel and Distributed Computing (JPDC), 2006.

Open Source Software (http://paradiseo.gforge.inria.fr)

- **School :** June 2006 (67 participants)
  Nov 2006 (120 participants), July 2007, ...
- **Tutorials, Program skeletons, ...**

Diapositive 25

MN10

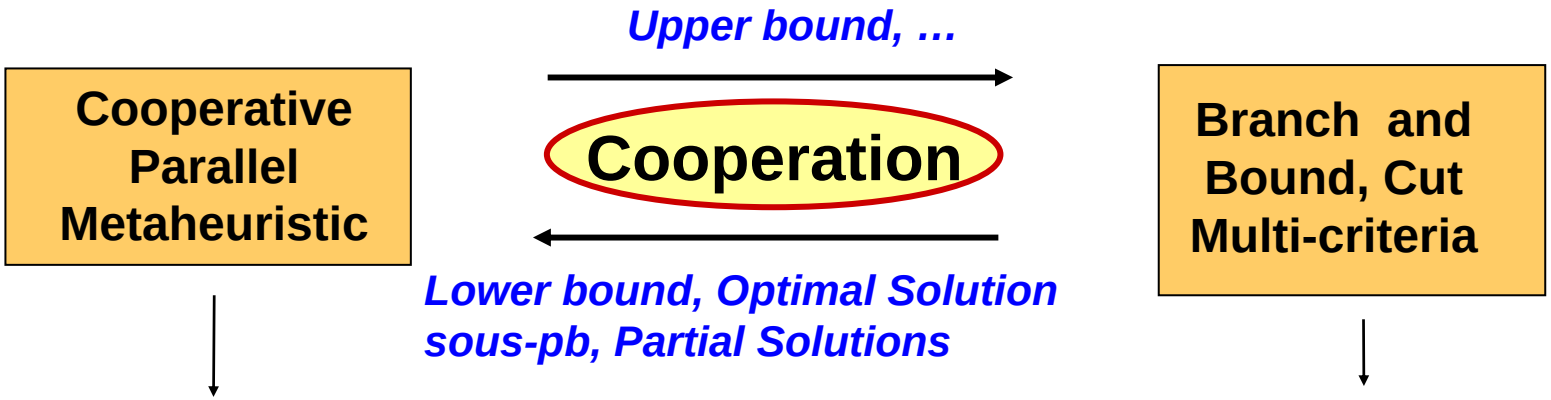
Melab Nouredine 24/11/2005

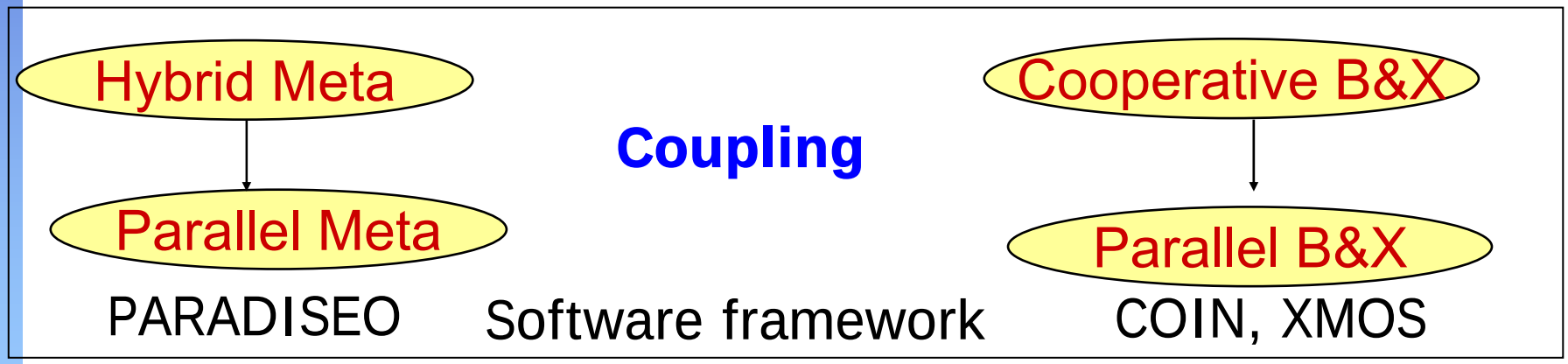
- La plate-forme ParadisEO est un logiciel libre disponible sur le Web.

- Les différents modules peuvent être utilisés de manière indépendante suivant les besoins de l'utilisateur.

Melab Nouredine; 26/01/2006

Cooperation Meta-Exact

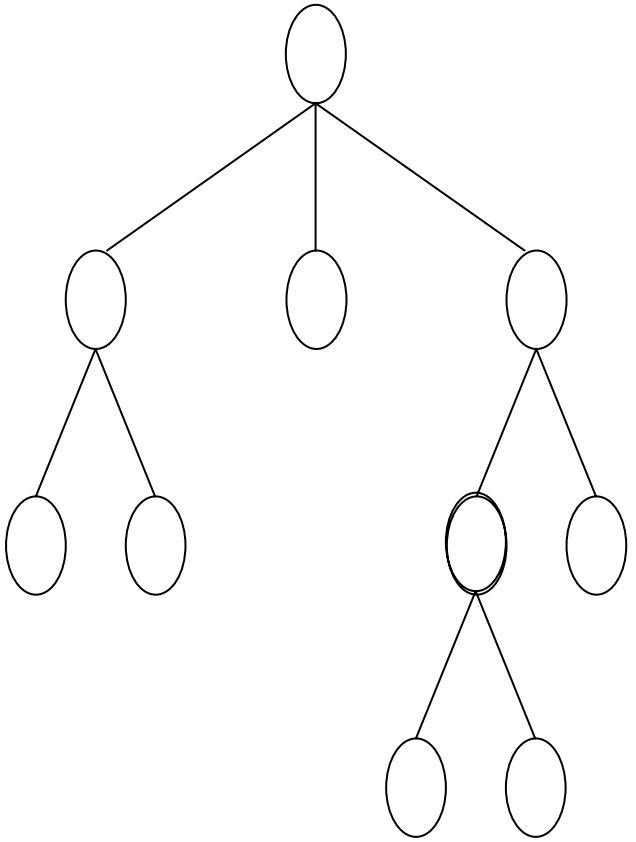




• M. Basseur, J. Lemesre, C. Dhaenens, E-G. Talbi, **«Cooperation between branch and bound and evolutionary algorithms to solve a bi-objective flow-shop problem»**, *WEA'2004, LNCS, 2004.*

Parallel Exact Methods

Branch and Bound



- Dynamic Load Balancing
  - **Irregular** tree in an **heterogeneous volatile** large network of workstations
- Fault Tolerance
  - Searching an **exact** solution in a **volatile** context
- Global solution sharing ...
  - ... Scalability in a **large network**
- Termination detection (**Asynchronism**)

E-G. Talbi, **"Parallel Combinatorial Optimization",** *John Wiley & Sons, USA, 2006.*

Diapositive 27

MN17

Melab Nouredine 23/11/2005

- L'étude de la concpetion parallèle sur grilles des méthodes exactes est inspirée de l'algorithme B&B.

- Quatre modèles ont été identifiés et analysés dans le contexte des grilles. Nous avons en particulier travaillé sur le modèle d'exploration arborescente parallèle car d'une part le parallélisme généré par ce modèle est de nature à justifier fortement l'intérêt des grilles à lui tout seul. D'autre part, il s'agit du modèle le plus utilisé et étudié pour la problématique intéressante qu'il pose :

- le problème de régulation de charge posé par la nature irrégulière de l'arbre exploré dans un contexte hétérogène volatile et à grande échelle.

- le problème de tolérance aux pannes posé par la recherche d'une solution exacte dans un contexte volatile.

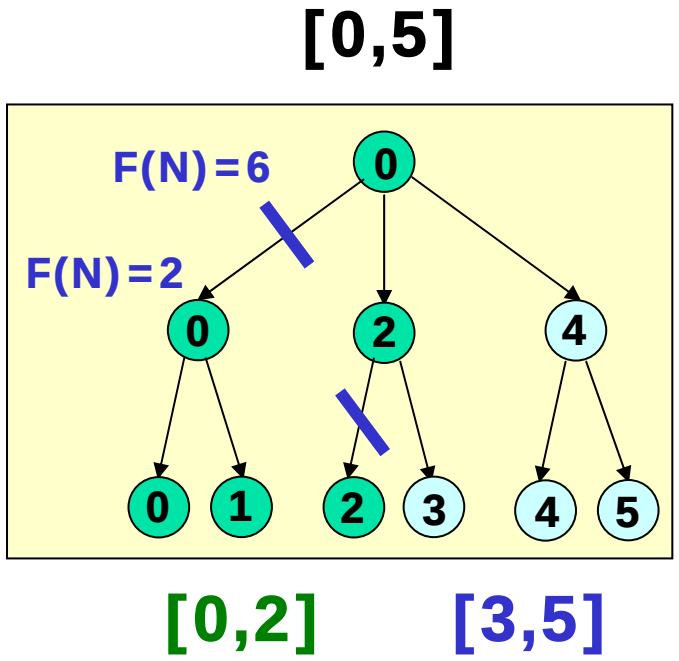
- le partage de la meilleure solution trouvée dans un contexte à grande échelle.

- et la détection de la terminaison en mode asynchrone.

Melab Nouredine; 26/01/2006

Proposed Approach - Gridification

- Specific coding of the tree and units of work (sub-tree)
- Efficient Scheduling mechanism
  - Communications (work migration), heterogeneity (processors)
- Reduced cost of the Checkpointing/Recovery mechanism (memory and communications)
- Implicit and natural detection of termination.



Diapositive 28

MN18

Melab Nouredine 25/11/2005

- Nous avons donc proposé une approche d'exploration parallèle gridifiée avec comme objectifs :

--- d'optimiser le processus de distribution des noeuds de l'arbre exploré en minimisant les coûts de commiuncation liés aux transferts de noeuds et d'autre part en prenant en compte l'hétérogénéité des machines en termes de puissance de calcul.

--- la proposition d'un mécanisme de sauvegarde/restauration à faible coût en termes de stockage et communications.

--- et la détection de la terminaison de manière naturelle et efficace.

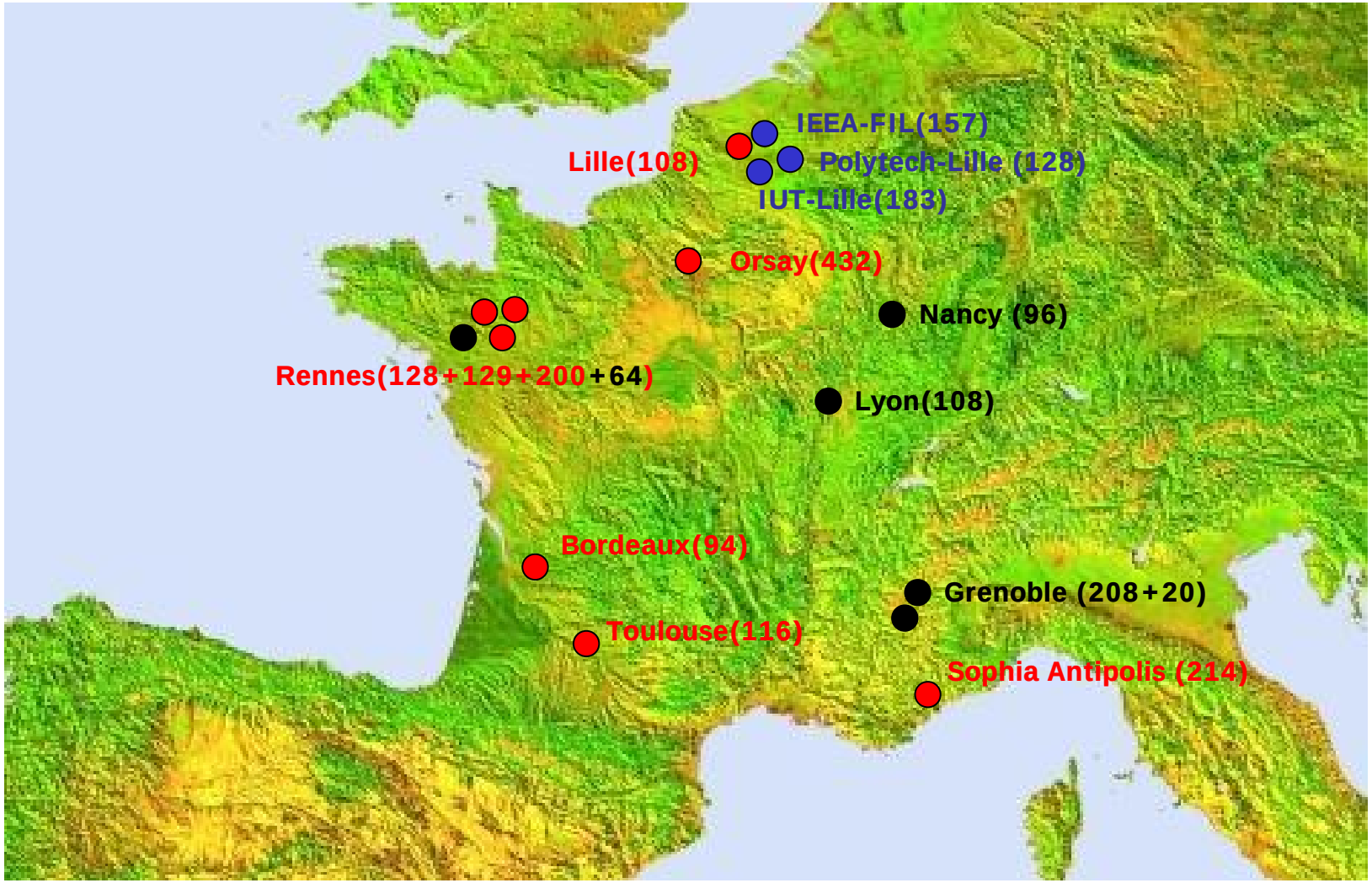
- L'approche proposée est basée sur le codage illustré ici de l'arbre de base, c'est à dire l'arbre obtenu par décomposition du problème traité et sans élagage.

- Pour optimiser les coûts de communication et de stockage, nous utilisons une description minimale des unités de travail : une unité de travail c'est simplement un intervalle. Par exemple, l'intervalle [0,2] décrit les noeuds verts et l'intervalle [3,5] désigne les noeuds bleu clair.

- Se pose alors la question : comment regénérer les noeuds à traiter à partir de leurs descripteurs associés. Il suffit en fait de parcourir l'arbre de base en profondeur d'abord à partir de sa racine en élagant tous les noeuds vérifiant cette règle.

Melab Nouredine; 26/01/2006

Grid of **2235 processors** (GRID'5000 + Campus of Lille)



R. Bolze, ..., N. Melab, ..., E-G. Talbi, ***« GRID'5000: A large scale and highly reconfigurable Grid»****, International Journal of High Performance Computing Applications (IJHPCA), Vol.20, No.4, 2006*

Some numbers : Instance Ta056 Flow-Shop 50-20

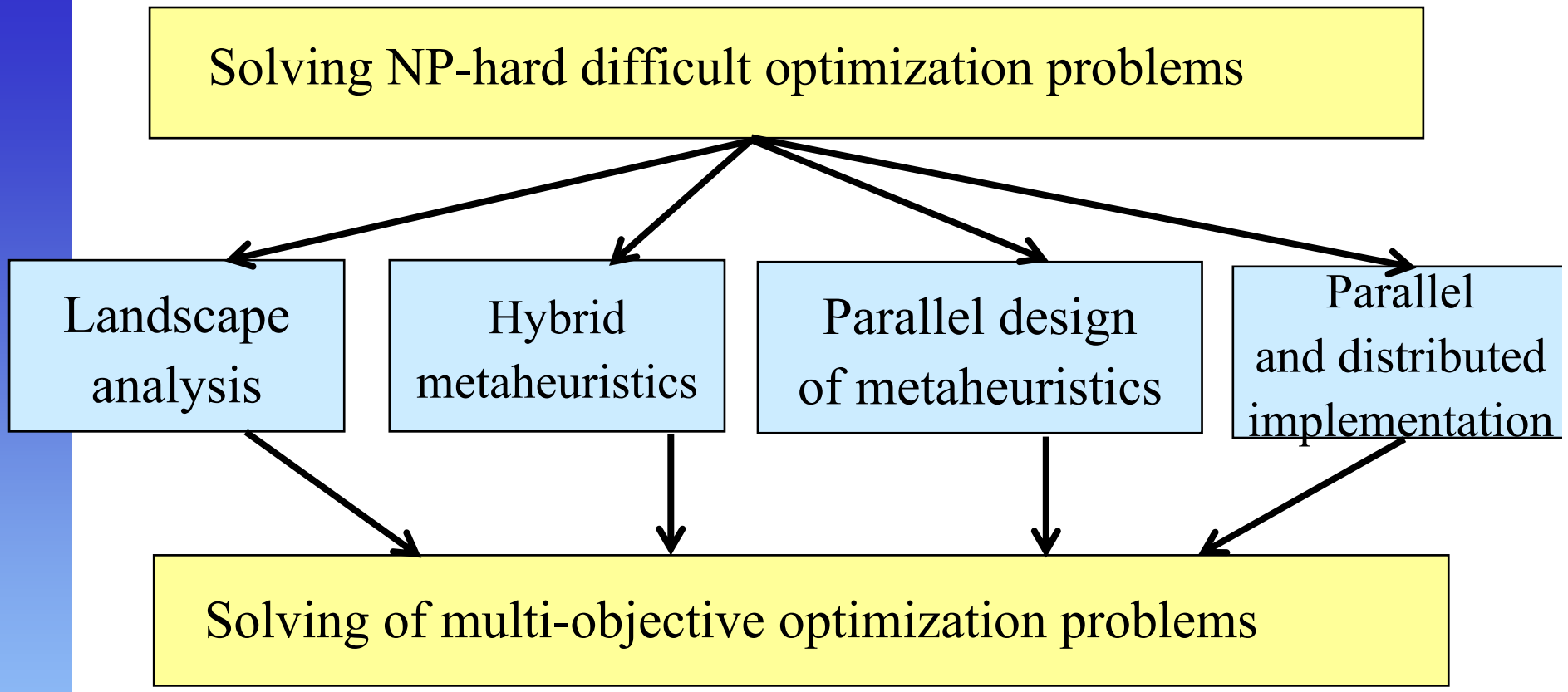
Total execution time	25 days 46 mn
Agregated total time	22 years 185 days 16 h
Mean number of used processors	328
Max number of used processors	1 195
Number of time a processor join the Grid	11 802
Max number of processors / site	Bdx(88), Orsay(360), Sophia(190), Lille(98), Toulouse(112), Rennes(456), Univ.(304)
Number of nodes explored	6,50874 e+12
Number of task units allocated	129 958
Number of checkpointing/recovery	4 094 176

Best known solution
Cmax = 3681
Ruiz & Stutzle, 2004

Exact solution
Cmax = 3679
Mezmaz, Melab & Talbi, 2006

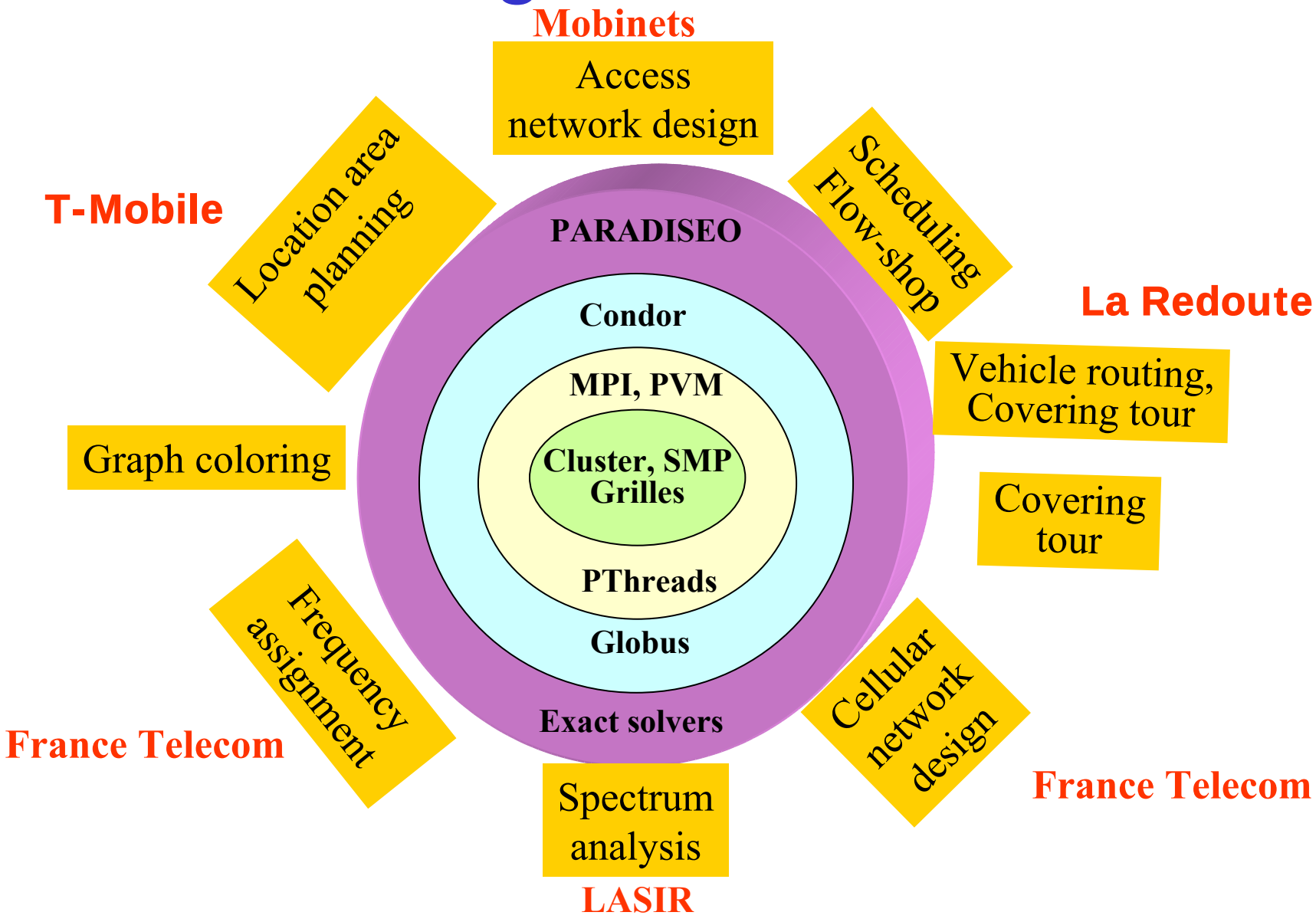
N. Melab, M-S. Mezmaz and E-G. Talbi, "**A Grid hybrid exact approach for solving multi-objective problems**" *Proc. of the 9th IEEE/ACM NIDISC'2006, Apr 2006.*

Applications



- Validate our approaches on **real** and **generic applications**
- Variety of domains : Telecommunications, Genomics.

Applications in Telecom & Logistics



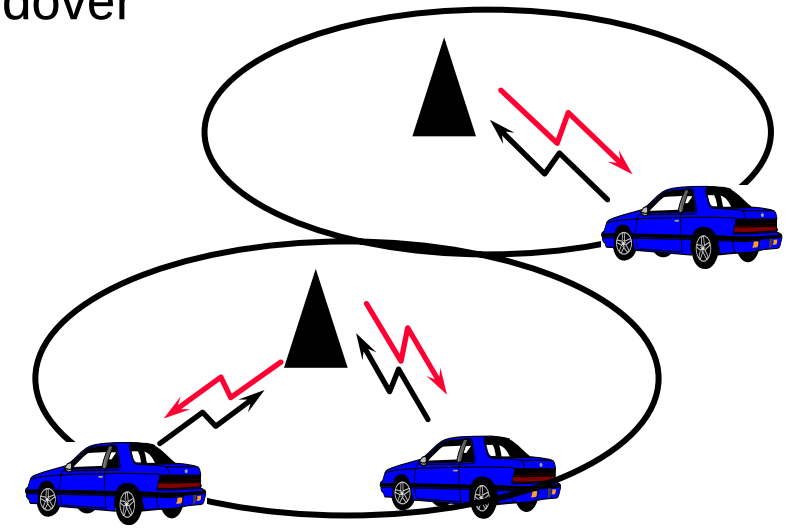
Cellular network design

(contract with France Telecom)

Economical Challenge (Cost, Quality of service)
Network Design (Positioninig of antenna, Configuration of antenna

- Objectives :
  - Min (number of sites)
  - Min (interferences)
  - Min (traffic loss)
  - Max (resource use)

- Constraints :
  - Covering
  - Handover

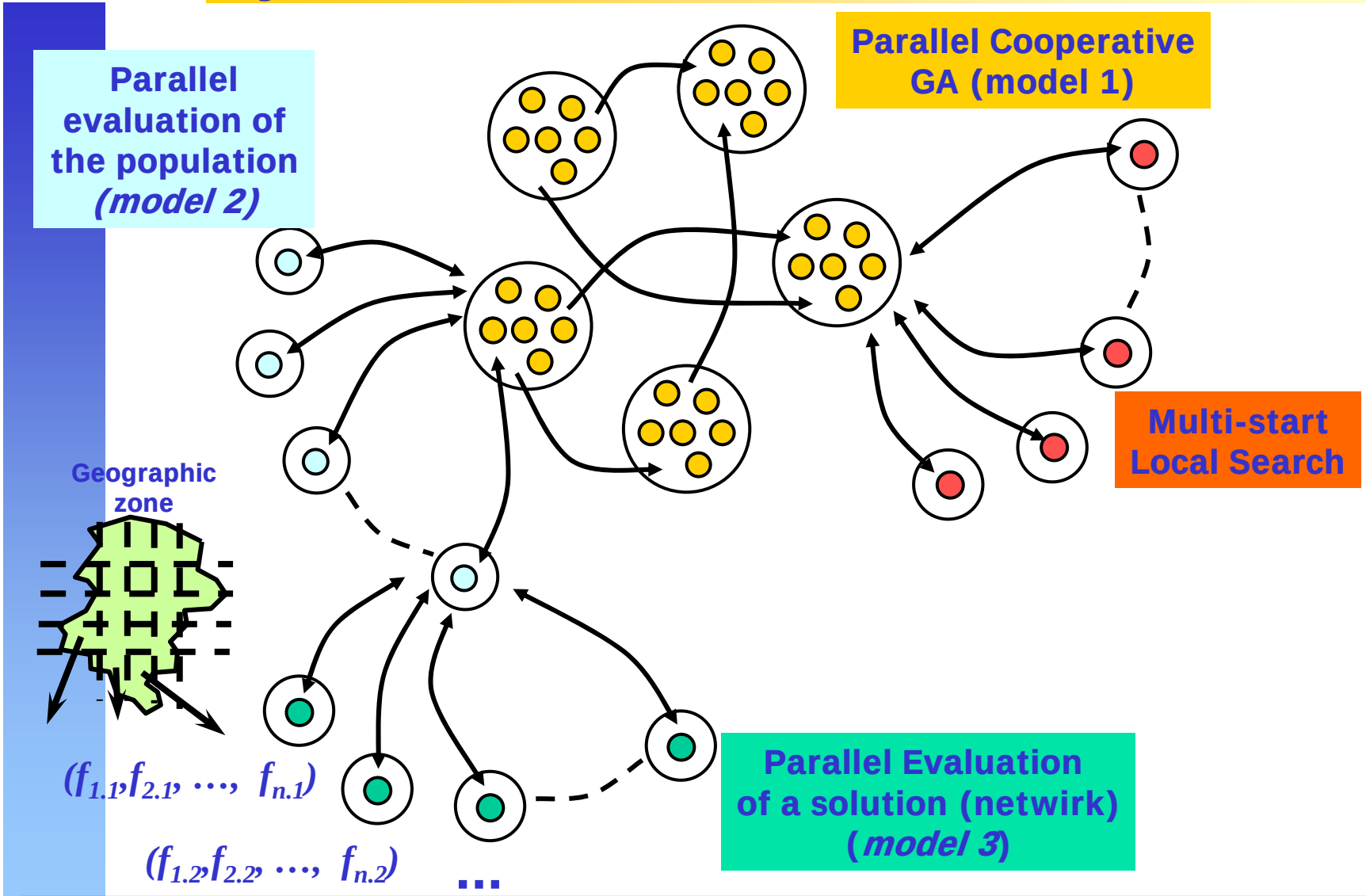


- Search space :
568 sites candidates ➔ $2^{3689160}$ solutions
and ~ 600.10^9 choices for one antenna !

- Cost Evaluation (pop : 100, Gen : 10^5,
cost = more than one year !!)

• E-G. Talbi, H. Meunier, «**A multiobjective algorithm for radio network optimization** », *JPDC Journal of Parallel and Distributed Computing, 2005.*

An Asynchronous Parallel Hierarchical Hybrid Model

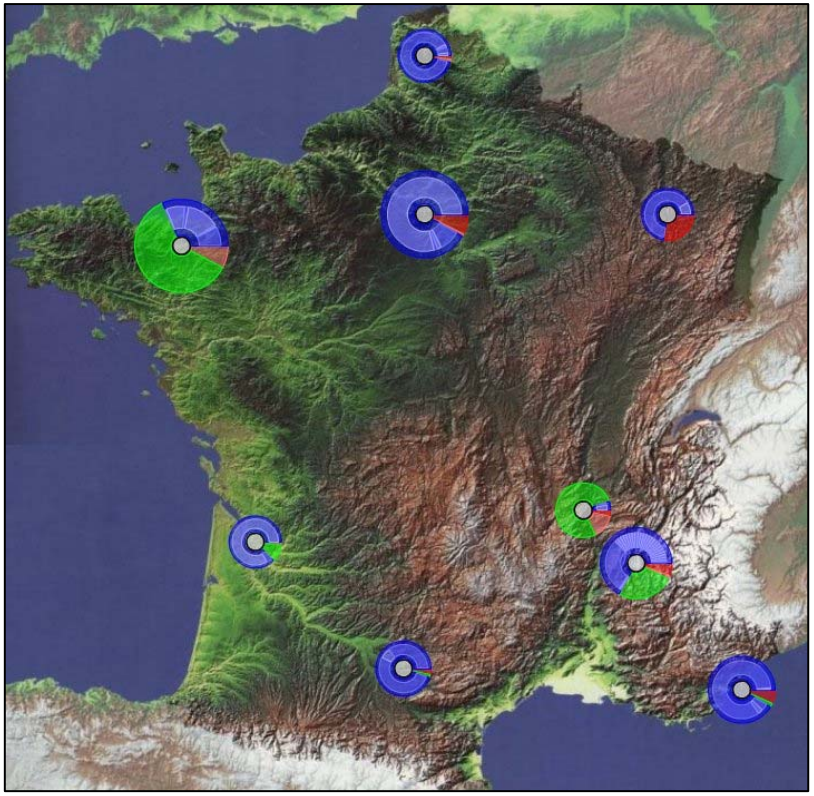


E-G. Talbi, S. Cahon and N. Melab. **Designing Cellular Networks using a Parallel Hybrid Metaheuristic on the Grid.** Journal of Computer Communications, Elsevier Science**,** Accepted**,** To appear in dec. 2005.

High-Performance Computing Grid: GRID'5000 under Globus

- 400 CPUs on 6 sites: Lille, Nice-Sophia Antipolis, Lyon, Nancy, Rennes

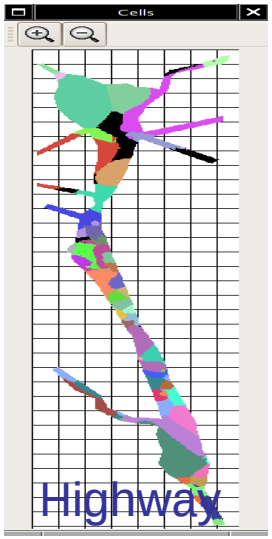
- **Parallel efficiency = 0.92**
- **Best results obtained**
- **More than 1 year of cumulative wall clock time**



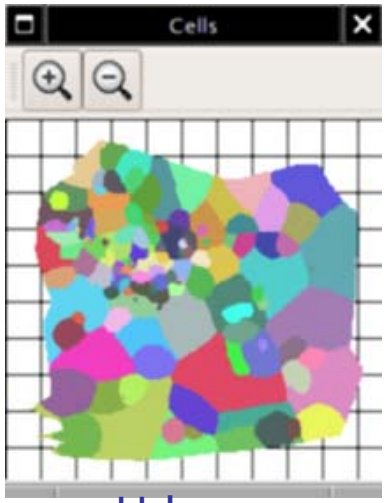
GRID'5000: A fully reconfigurable grid! :
Linux « images » having **Globus** and MPICH-G2 already installed.

High-Throughput Computing Grid: Campus of Lille (3 # administrative domains)

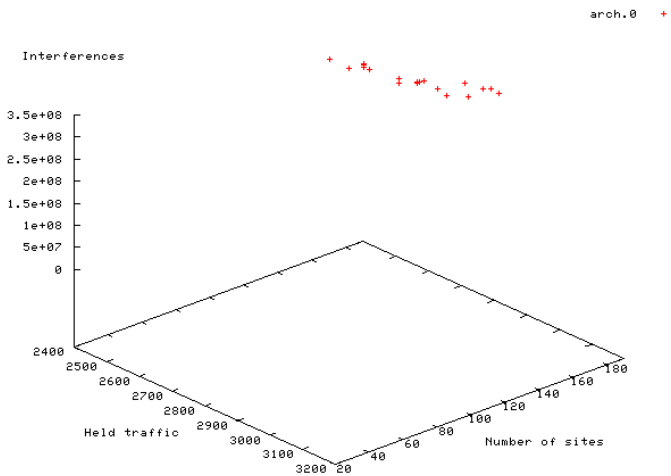
Platform	HTC Grid (Polytech, IUT, LIFL)
Prog. Environment	**Condor**
Number of proc.	100 (heterog. and non dedicated)
Cumulative wall clock time	30681 h.
Wall clock time	Almost 15 days
Parallel efficiency	**0.98**



Highway

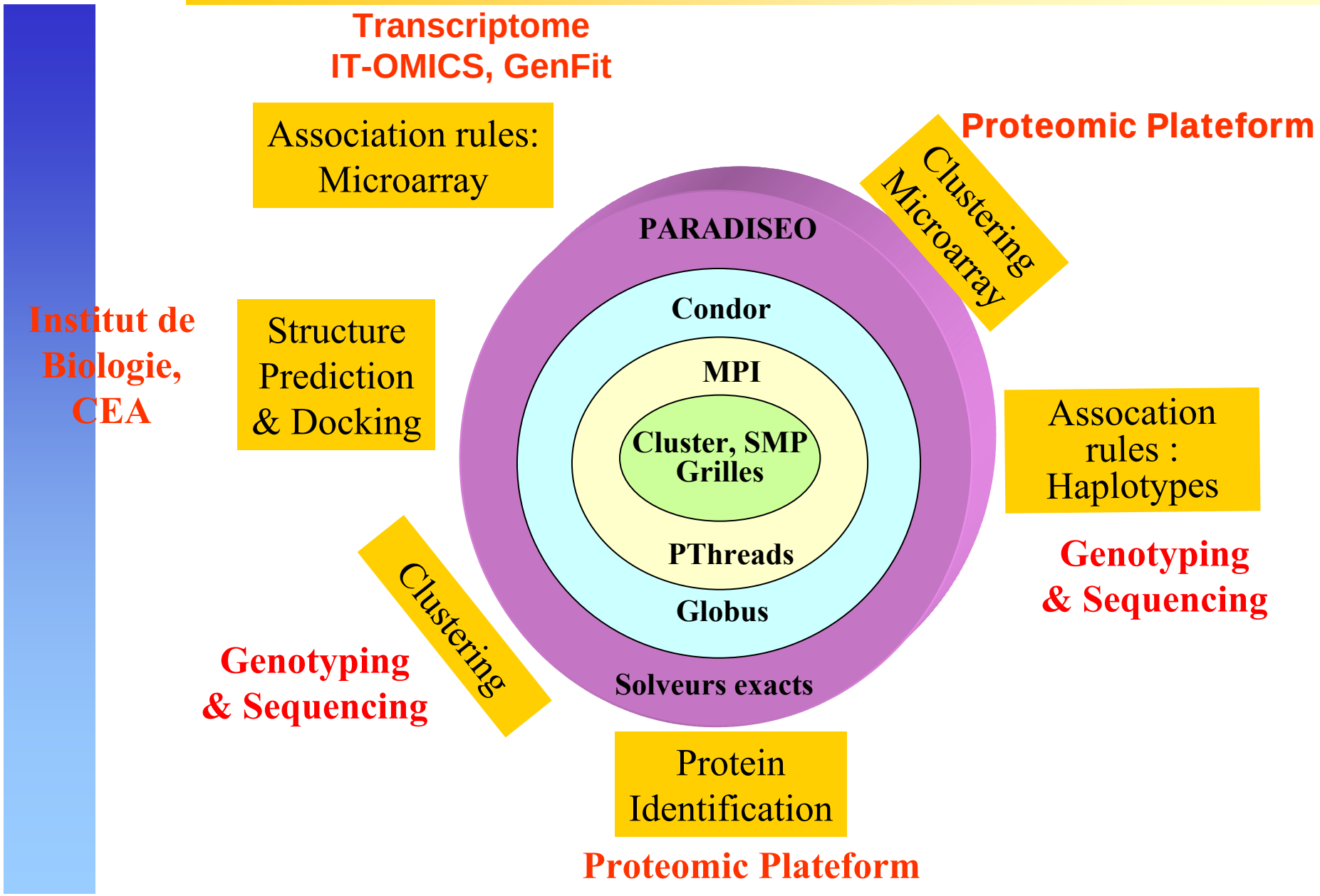


Urban



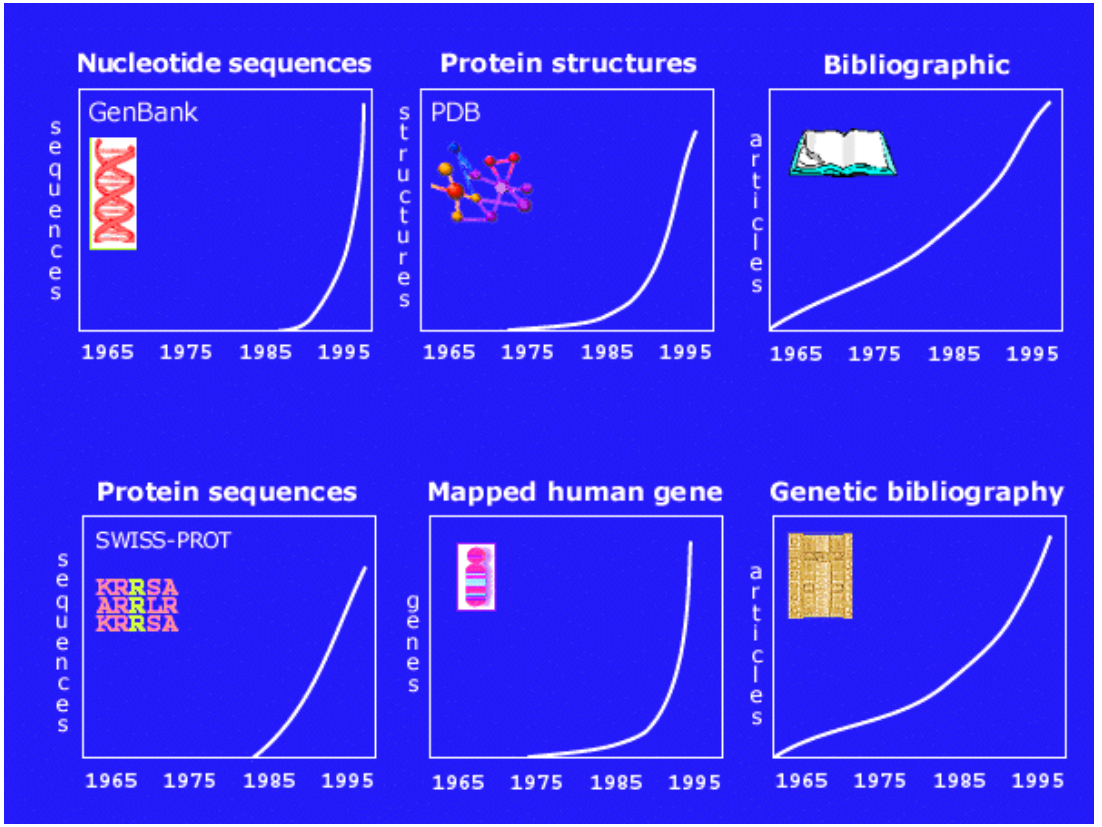
E-G. Talbi, S. Cahon and N. Melab. **Designing cellular networks using a parallel hybrid metaheuristic on the Grid.** Computer Communications Journal, 30(2), 2007.

Applications in Bioinformatics



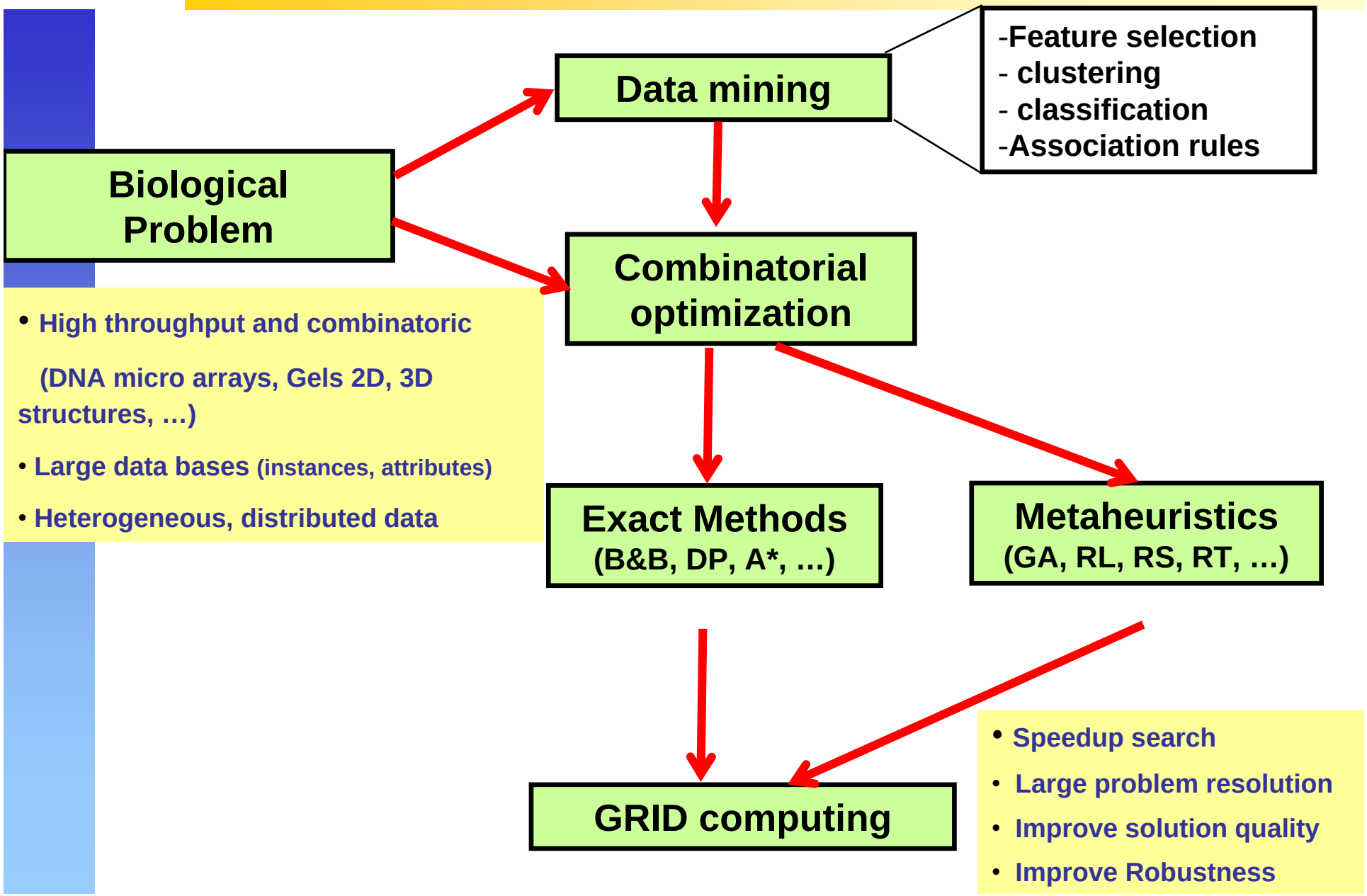
Application in Bioinformatics

- Growth of amount of available data
  - New high throughput technologies (Microarray genomic data, protein sequences, DNA sequences, bibliographic data ...)



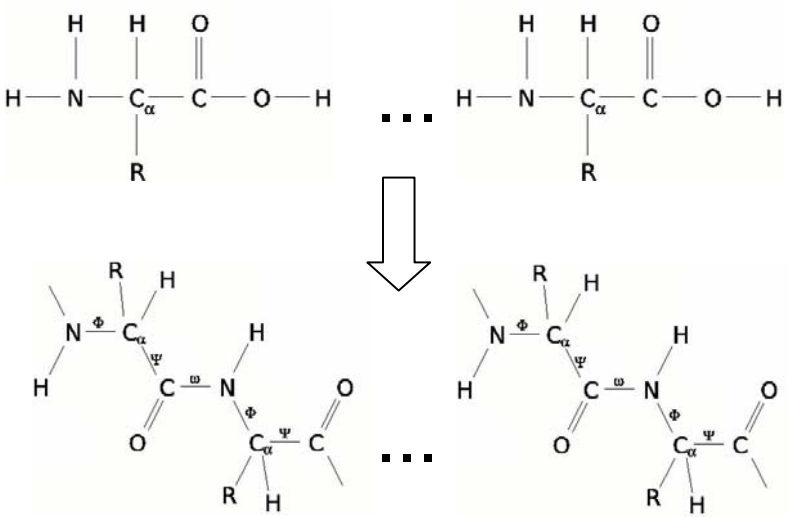
⇨ **Need of knowledge discovery algorithms**

From Bio to Combinatorics & Grids



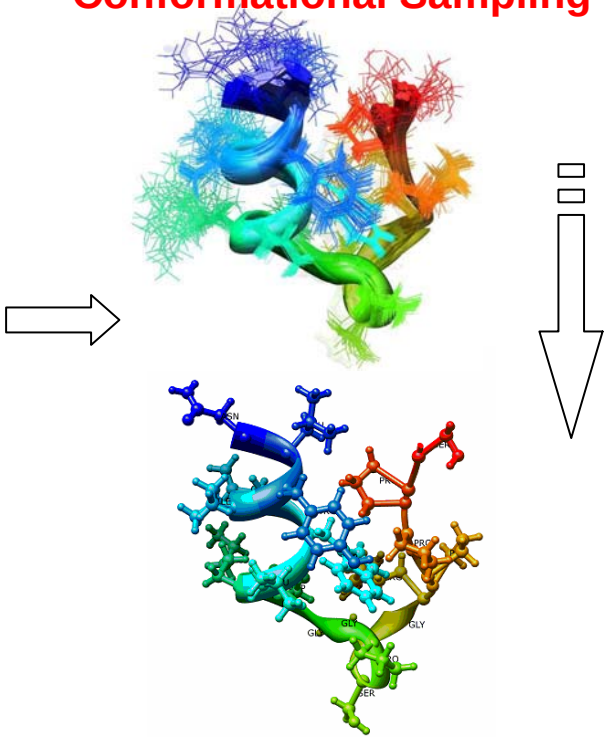
Protein Structure Prediction Problem

Amino-acids



A protein

Conformational Sampling



Native Conformation

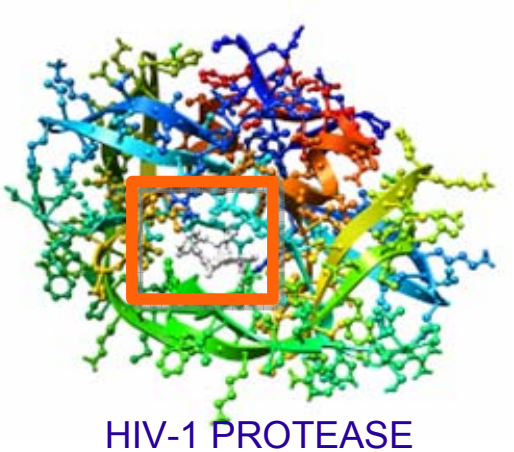
ENERGY MINIMIZATION

Protein Structure Prediction (PSP) ~ finding the ground-state (tertiary structure) conformation of a protein, given its amino-acid sequence - the primary structure

Molecular Docking

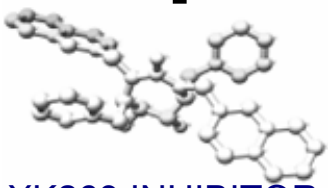
RECEPTOR

LIGAND



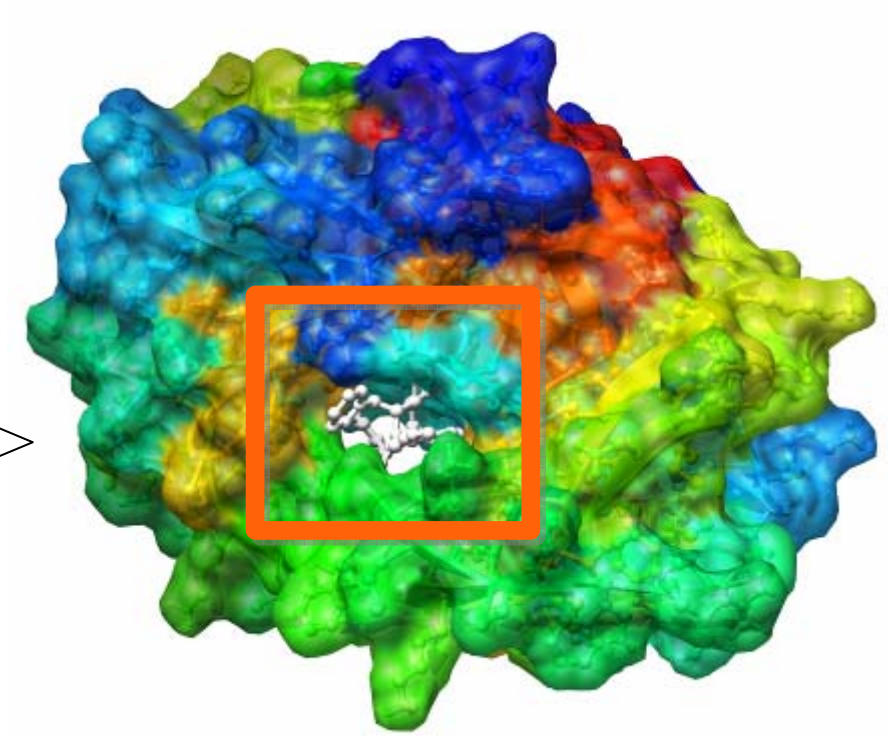
HIV-1 PROTEASE

+



XK263 INHIBITOR

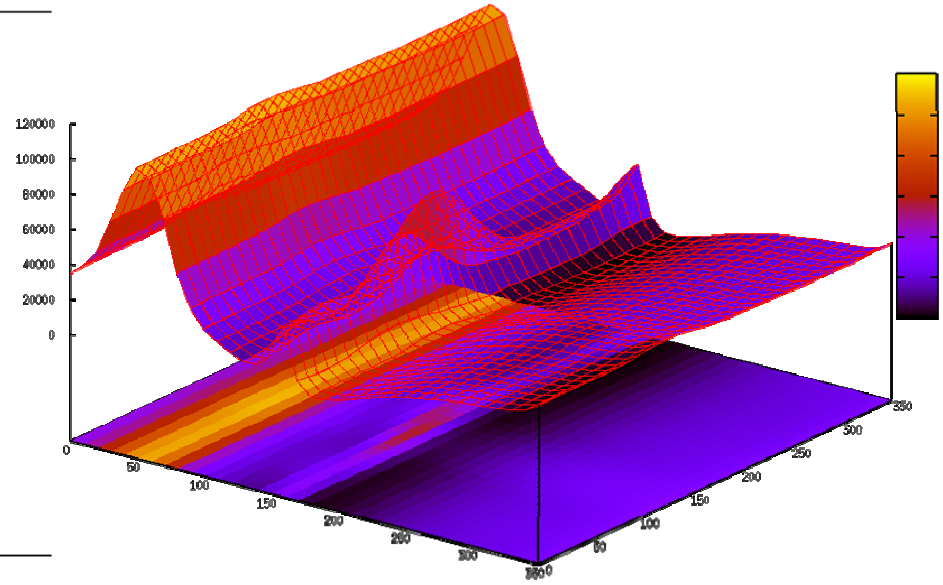
HIV-1 PROTEASE + XK263



Molecular Docking ~ the prediction of the optimal bound conformation of two molecules exerting geometrical and chemical complementarity.

PSP modeling:
Scoring Energy Function

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



- Atom-level energy based on molecular mechanics force fields (CHARMM - Chemistry at HARvard Molecular Mechanics)
- Two major terms: bonded atoms energy & non-bonded atoms energy
- The factors model oscillating entities …
  - … i.e. forces simulated by interconnecting springs between atoms

Complexity Analysis

Levinthal's Paradox

A molecule of 40 residues

10 conformations per residue

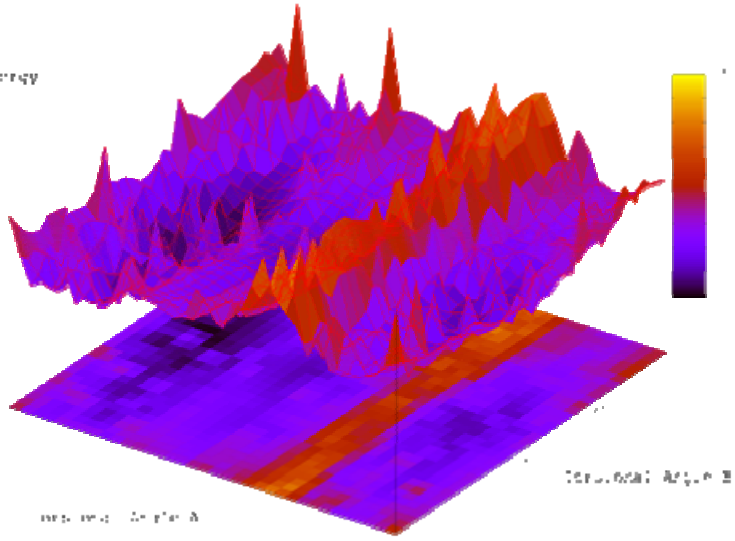
10^{40} **conformations**

10^{14} **conformations per second**

10^{28} **years**

10^{11} local optima for the ***[met]-enkephalin*** **pentapeptide**

- 75 atoms
- Five amino-acids (Tyr-Gly-Gly-Phe-Met)
- 22 variable backbone dihedral angles



Multi-objective model

- **Only based on energetic terms or geometric terms.**
- **Not necessarily describes the real docking mechanisms.**

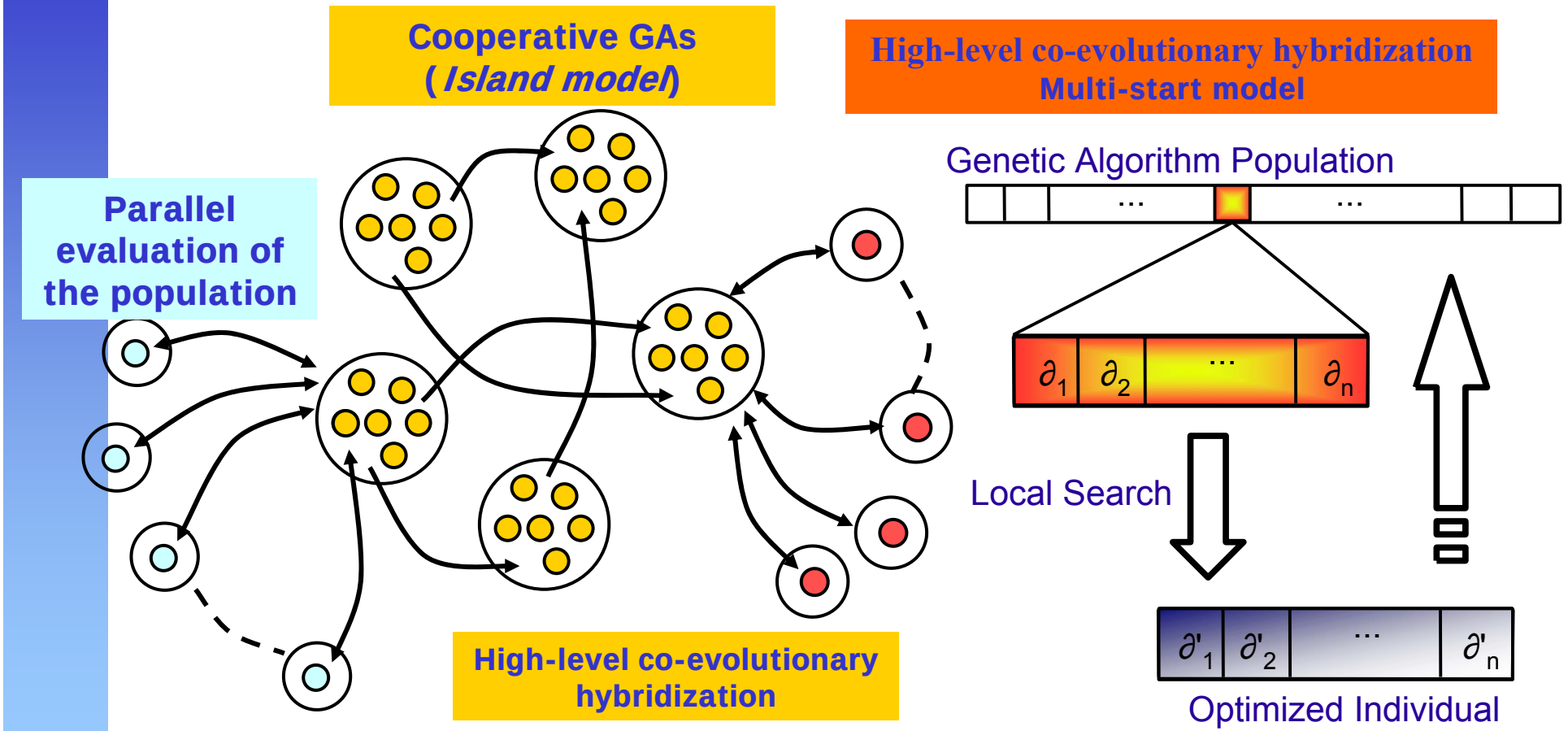
- **The docking combines the difficulties of protein structure prediction, protein folding, geometric complementarity…**

Multi-objective optimization can tackle this problem.

- **An energetic term**: the traditional bonded/non bonded term
- **A surface term** that describes the size of the surface exposed to the solvent ➔ it a good indicator of the entering of the ligand in the site,
- **An entropic term** that insures the robustness of the results (according to the energetic well found).

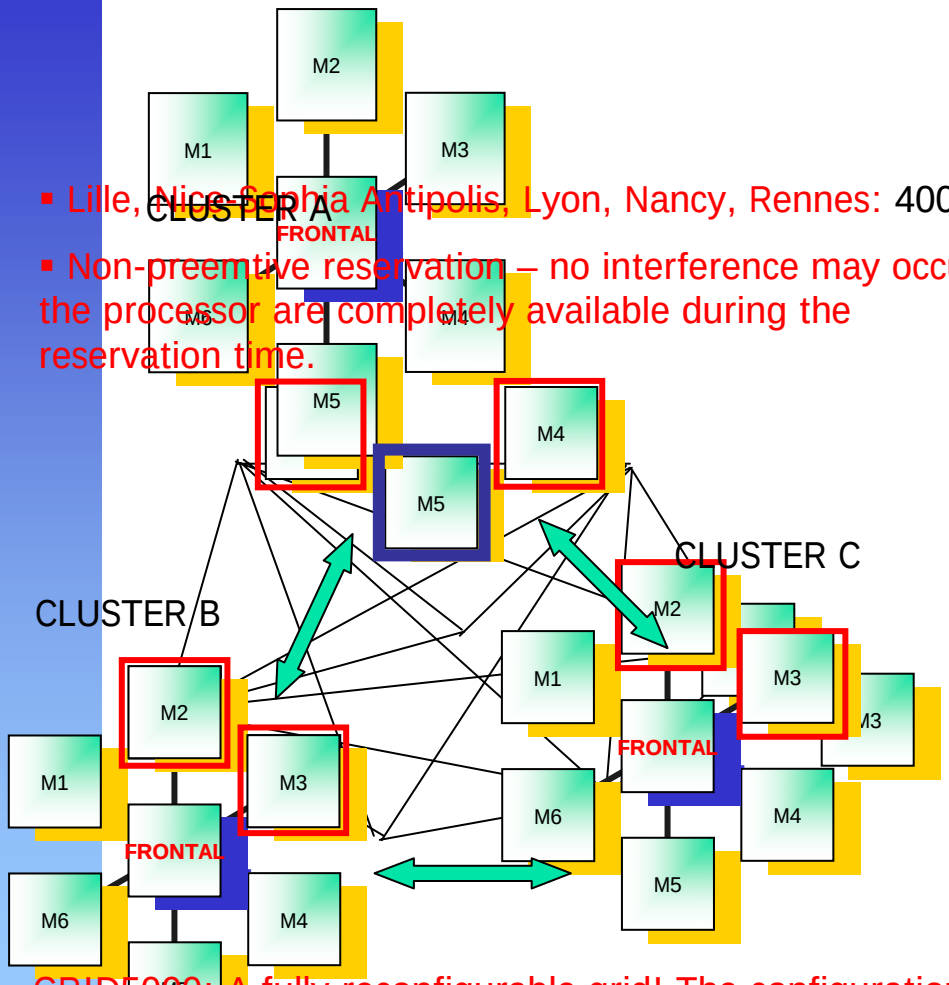
Parallel asynchronous hierarchical Lamarkian GA

Parallel asynchronous hierarchical hybrid meta-heuristic



A-A. Tantar, N. Melab, E-G. Talbi, O. Dragos and B. Parain. **A Parallel Hybrid Genetic Algorithm for Protein Structure Prediction on the Computational Grid**. FGCS, Elsevier Science, Vol.23(3), 398-409, 2007.

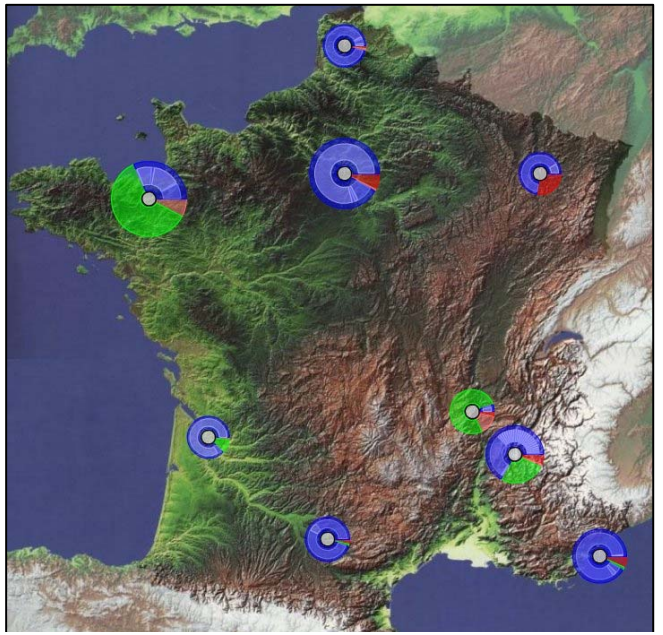
Deployment of ParadisEO-G4

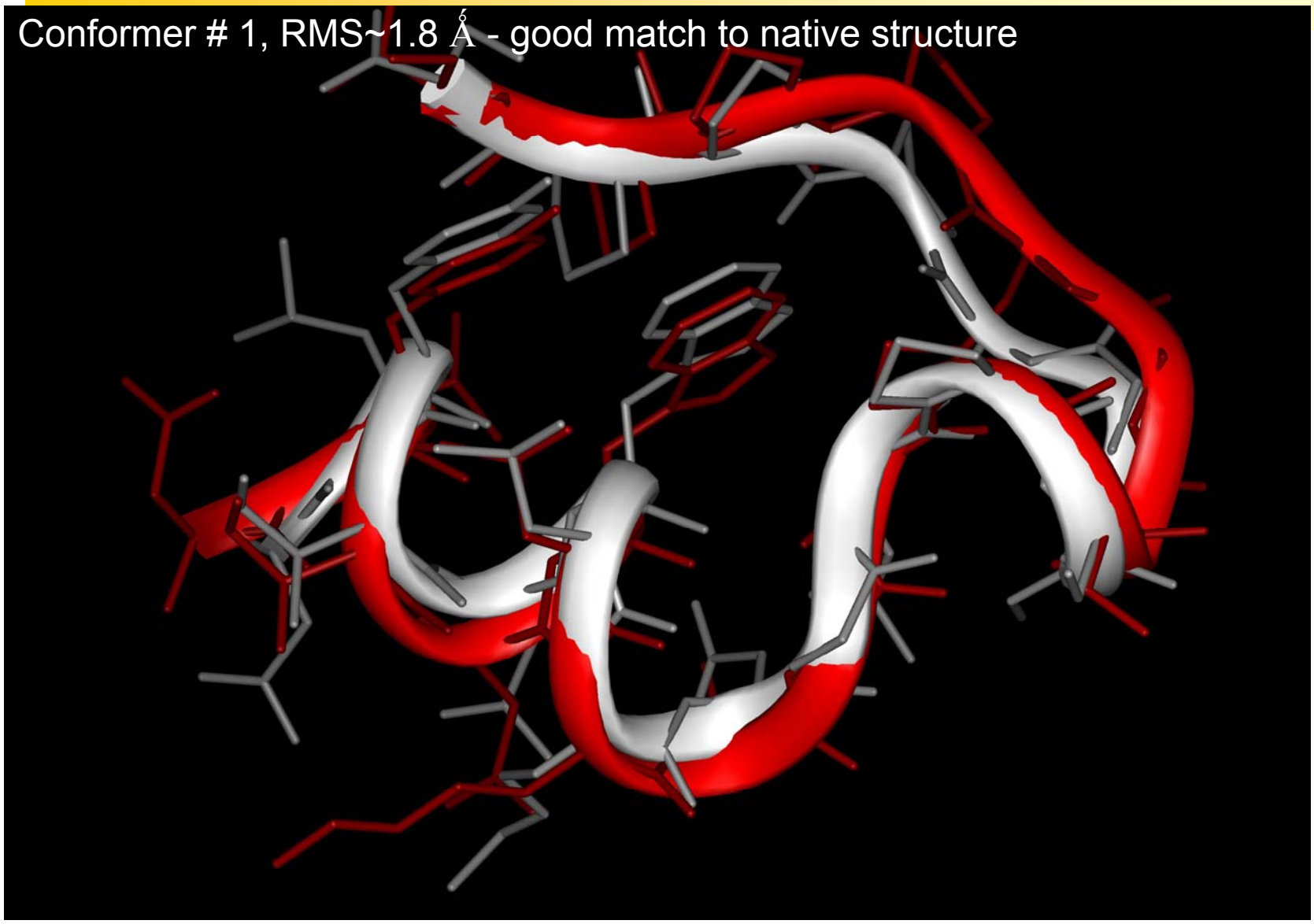


- Lille, Nice-Sophia Antipolis, Lyon, Nancy, Rennes: 400CPUs
- Non-preemtive reservation – no interference may occur; the processor are completely available during the reservation time.

GRID5000: A fully reconfigurable grid! The configuration phase relies on the deployment of pre-built Linux « images » having Globus and MPICH-G2 already installed.

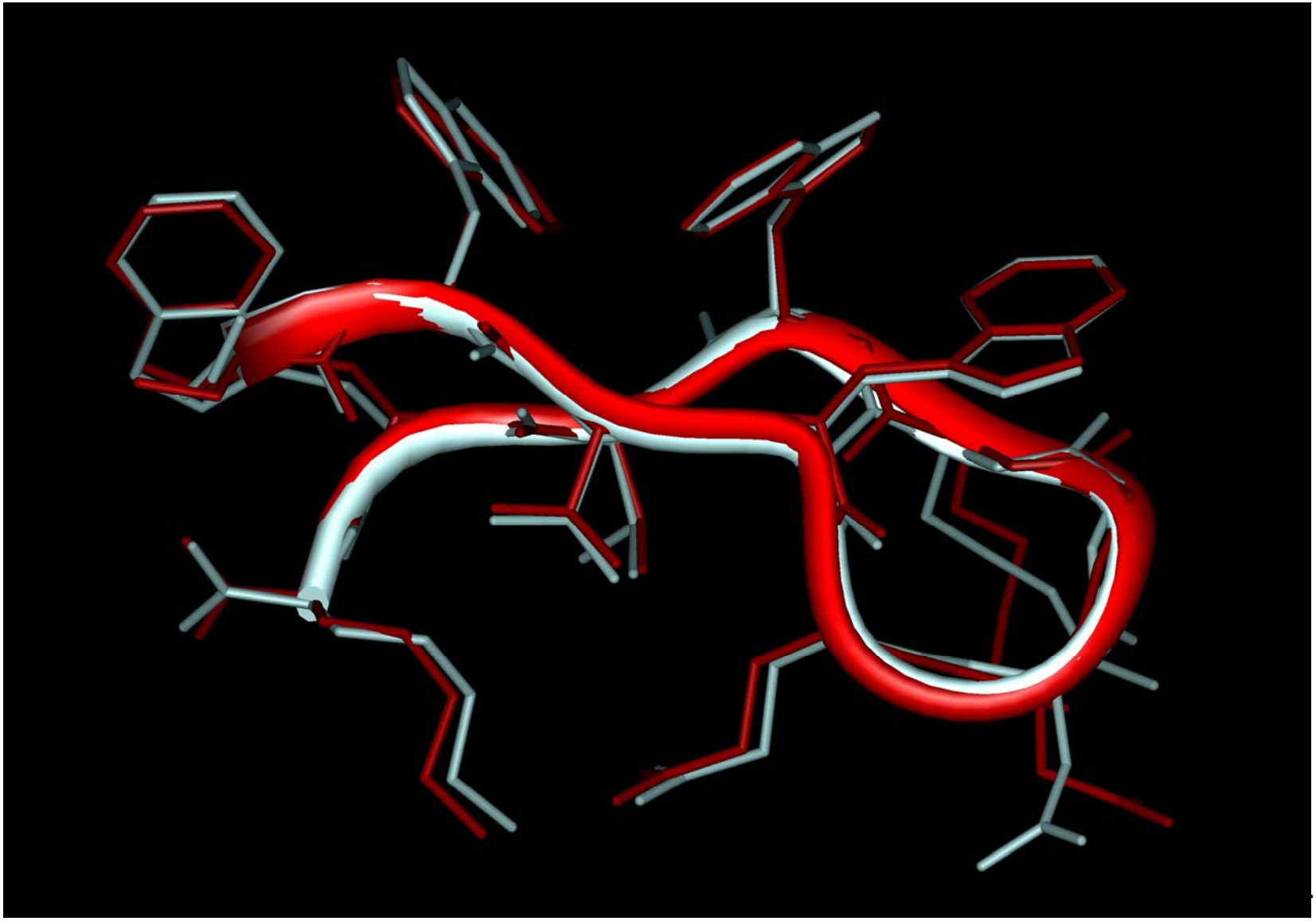
1. Reserve a pool of nodes
2. Select a master node for the Globus GRID
3. Configure the Globus GRID (security and deployment services: certificates, user credentials, xinetd, postgresql, etc.)
4. Deploy and execute – MPICH-G2

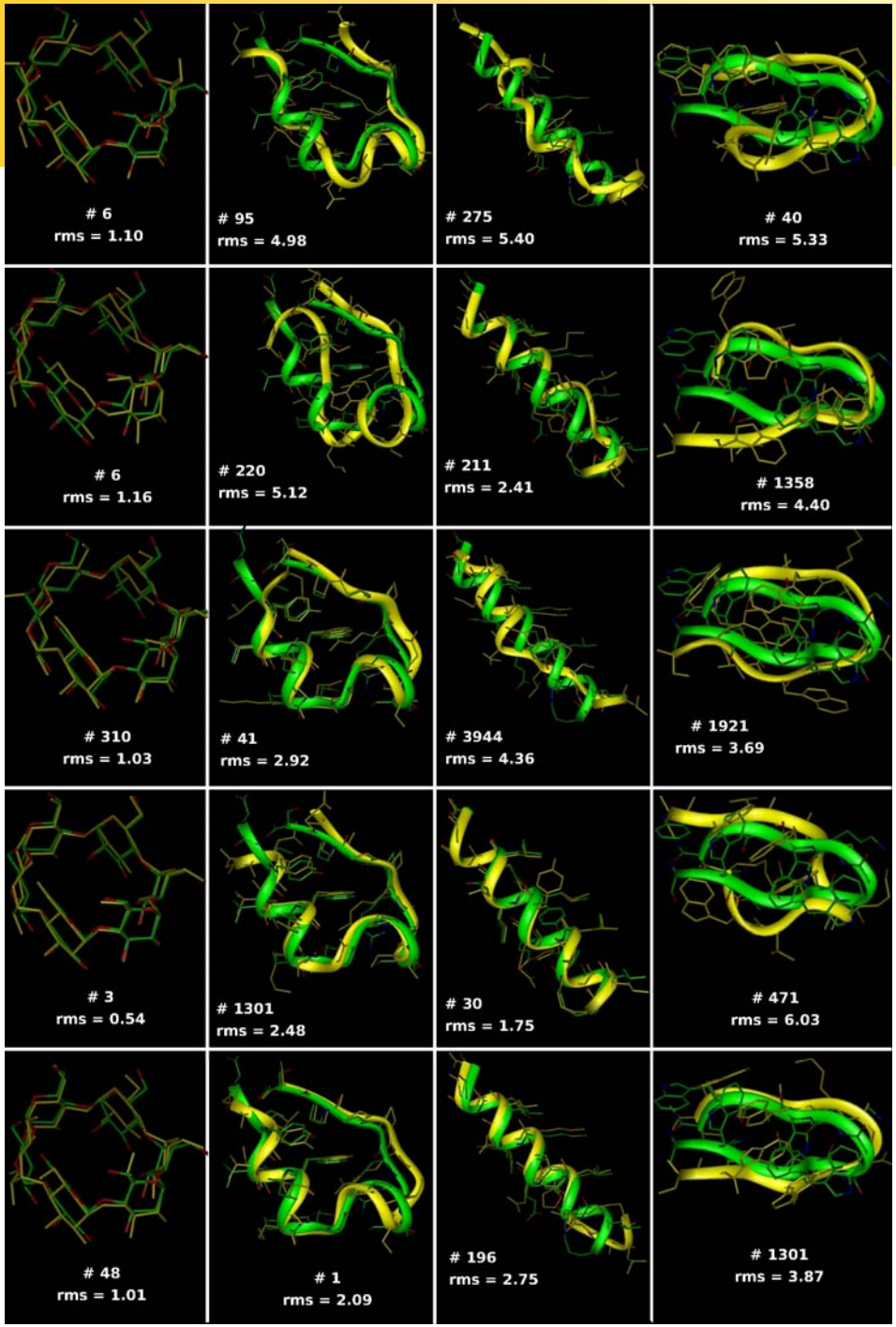




Tryptophan-cage 1L2Y

Conformer # 1, RMS~0.8 Å - perfect match to native structure

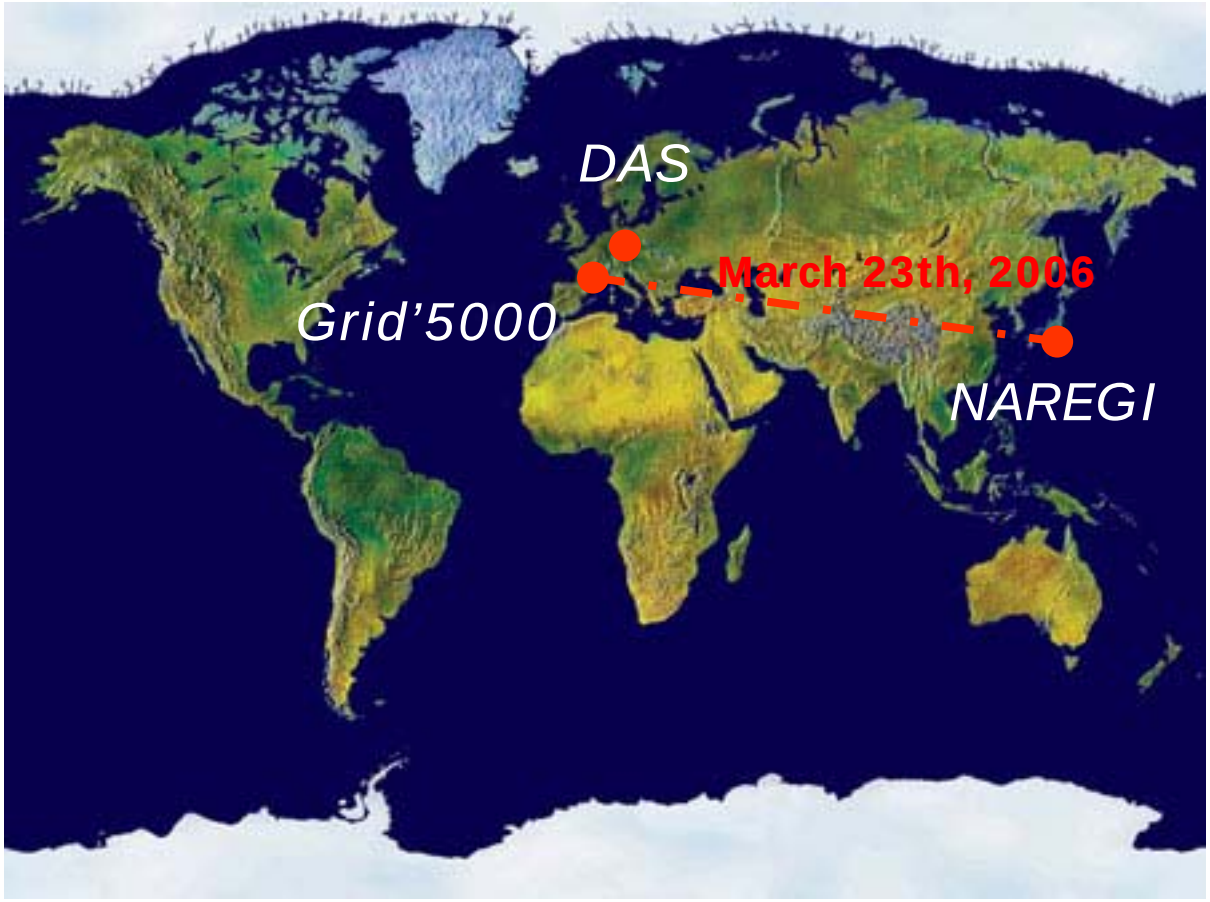



# 6
rms = 1.10
# 95
rms = 4.98
# 275
rms = 5.40
# 40
rms = 5.33
# 6
rms = 1.16
# 220
rms = 5.12
# 211
rms = 2.41
# 1358
rms = 4.40
# 310
rms = 1.03
# 41
rms = 2.92
# 3944
rms = 4.36
# 1921
rms = 3.69
# 3
rms = 0.54
# 1301
rms = 2.48
# 30
rms = 1.75
# 471
rms = 6.03
# 48
rms = 1.01
# 1
rms = 2.09
# 196
rms = 2.75
# 1301
rms = 3.87

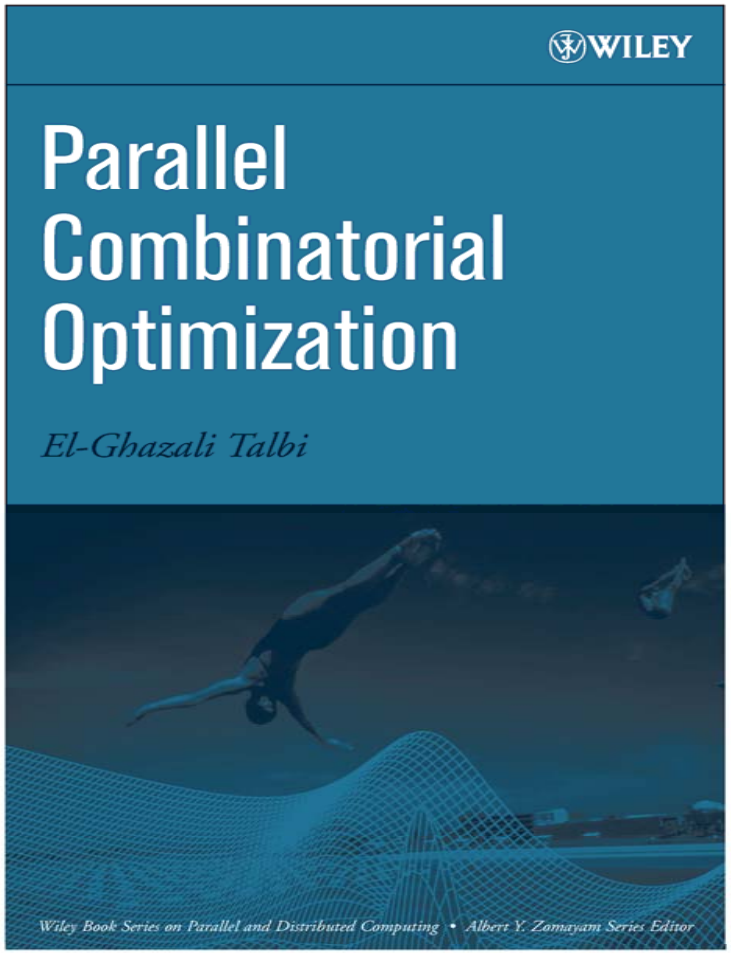
Perspectives

- Parallel models combining Metaheuristics & Exact methods (Algorithms, Coupling of Software, ...)
- Hybrid Parallel models for dynamic optimization problems
- Hybrid Parallel models for multi-objective optimization problems with uncertainty
- Solving challenging problems on Grids (Ex: Molecular biology, Engineering design, ...)

Perpectives: Towards an international Grid



- NAtional REsearch Grid Intiative (NAREGI) – Japan (11 sites)
- Distributed ASCII Supercomputer 2 (DAS2) – Netherlands (5 sites)



- **E-G. Talbi, « Parallel Combinatorial Optimization »,** *Wiley, USA, 2006*

- **E-G. Talbi, A. Zomaya, « Grids for Bioinformatics and Computational Biology »,** *Wiley, USA, 2007*