A parallel hybrid genetic algorithm for protein structure prediction on the computational grid

A.-A. Tantar\textsuperscript{a}, N. Melab\textsuperscript{a,}\textsuperscript{,*}, E.-G. Talbi\textsuperscript{a}, B. Parent\textsuperscript{b}, D. Horvath\textsuperscript{b}

\textsuperscript{a}Laboratoire d’Informatique Fondamentale de Lille, LIFL/CNRS UMR 8022, DOLPHIN Project - INRIA Futurs, Cité Scientifique, 59655 - Villeneuve d’Ascq Cedex, France

\textsuperscript{b}CNRS UMR8576, Université des Sciences et Technologies de Lille, Bâtiment C9, Cité Scientifique 59655 - Villeneuve d’Ascq Cedex, France

Received 2 February 2006; received in revised form 5 August 2006; accepted 7 September 2006
Available online 1 November 2006

Abstract

Solving the structure prediction problem for complex proteins is difficult and computationally expensive. In this paper, we propose a bicriterion parallel hybrid genetic algorithm (GA) in order to efficiently deal with the problem using the computational grid. The use of a near-optimal metaheuristic, such as a GA, allows a significant reduction in the number of explored potential structures. However, the complexity of the problem remains prohibitive as far as large proteins are concerned, making the use of parallel computing on the computational grid essential for its efficient resolution. A conjugated gradient-based Hill Climbing local search is combined with the GA in order to intensify the search in the neighborhood of its provided configurations. In this paper we consider two molecular complexes: the tryptophan-cage protein (Brookhaven Protein Data Bank ID 1L2Y) and \textit{α}-cyclodextrin. The experimentation results obtained on a computational grid show the effectiveness of the approach.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Protein structure prediction; Genetic algorithm; Hill climbing; Parallel computing; Grid computing

1. Introduction

Nowadays, grid computing is admitted as a powerful way to achieve high performance on computational-intensive applications. The protein structure prediction problem, further referred to as PSP, is one of the particularly interesting challenges of parallel computing on the computational grid. The problem consists in determining the ground-state conformation of a specified protein, given its amino acid sequence—the primary structure. In this context, the ground-state conformation term designates the associated tridimensional native form, known as zero energy tertiary structure. Addressing the mathematical model, paradigms based on quantum mechanics and the Schrödinger equation were developed in the literature, as well as empirical techniques based on classical dynamics—to be further discussed in the following sections.

Although there exist laboratory methods addressing the herein described problem, prohibitory costs and the long experimentation time required make them unfeasible for large scale application. As a consequence, computational protein structure prediction represents an interesting alternative, though complexity matters impose strong limitations. For a reduced size molecule composed of 40 residues, a number of $10^{40}$ conformations must be taken into account when considering, on average, 10 conformations per residue. Furthermore, if a number of $10^{14}$ conformations per second is explored, a time of more than $10^{13}$ years is needed for finding the native-state conformation. For example, for the \textit{met}-enkephalin pentapeptide, composed of 75 atoms and having five amino acids, \textit{Tyr-Gly-Gly-Phe-Met}, and 22 variable backbone dihedral angles, a number of $10^{11}$ local optima is estimated. Detailed...
aspects concerning complexity matters were discussed in [20, 21], leading to the mention of the Levinthal’s paradox [6] which states that, despite numerous pathways, in vivo molecular folding, for example, has a time scale magnitude of several milliseconds.

Notes on molecular structure prediction complexity may be found in [19]. Although it may not be possible to construct a general mathematical model for describing molecular structures, it may be inferred that no polynomial time resolution is possible if no or less a priori knowledge is employed. As a consequence, no simulation or resolution is possible unless extensive computational power is applied. Thus, a distributed grid approach is required.

Genetic algorithms are population-based metaheuristics that allow a powerful exploration of the conformational space. However, they have limited search intensification capabilities, which are essential for neighborhood-based improvement (the neighborhood of a solution refers to part of the problem’s landscape). Therefore, the GA is combined with a conjugated gradient-based Hill Climbing local search method, in order to improve both the exploration and the intensification capabilities of the two techniques. In addition, the GA is parallelized in a hierarchical manner. Firstly, several GAs cooperate by exchanging their genetic material (parallel island model [3]). Secondly, as the fitness function of each GA is time-intensive the fitness evaluation phase of the GA is parallelized (parallel evaluation of the population model [3]). These two models are provided in a transparent way through the ParadisEO-CMW framework [1], dedicated to the reusable design of parallel hybrid metaheuristics on computational grids.

The interest in multicriterion structure prediction resides in result optimality and problem simplification. It can be argued that the native structure of a molecule should not be described through one unique conformation but through an ensemble of conformations, as in statistical mechanics [8]. As per environment interactions and the non-rigidity of a molecule’s conformation, structural description may be performed by using a set of potentially transitory conformations. In this case, the transitory conformations are distributed at the base of a funnel-like energy landscape. As a consequence, relating to mesoscopic and macroscopic realm aspects, multicriterion analytical and computational models are extremely important for the complete in silico characterization of molecular systems.

The latter argument, concerning problem simplification, refers to molecular processes complexity, in terms of number of local optima—as mentioned above, a number of $10^{11}$ local optima is estimated for the [met]-enkephalin pentapeptide. The reduction of the number of local optima may be attained by transforming a monocriterion optimization problem into a multicriterion problem, experimental results in this respect being furnished in [7]. It should be mentioned that, at this time, the existing approaches focus on monocriterion definition terms for problem resolution.

The importance of the PSP problem is reinforced by the ubiquitousness of proteins in the living organisms, applications of computational protein structure prediction directed to computer assisted drug design and computer assisted molecular design. From a structural point of view, proteins are complex organic compounds composed of amino acid residues chains joined by peptide bonds. Proteins are involved in immune response mechanisms, enzymatic activity, signal transduction etc. Due to the intrinsic relation between the structure of a molecule and its functionality, the problem implies important consequences in medicine and biology related fields.

An extended referential resource for protein structural data may be accessed through the Brookhaven Protein Data Bank1 [26]. For a comprehensive introductory article on protein structure, consult [9]. Also, for a glossary of terms, see [29].

In this paper, we propose a bicriterion genetic algorithm (GA), based on Newton’s classical mechanics for performing molecular energy calculations. The proposed approach has been applied for two molecular complexes: the tryptophan cage protein (Brookhaven Protein Data Bank ID 1L2Y) and α-cyclodextrin. The experimental results obtained on a computational grid demonstrate the effectiveness of the approach.

The remainder of the paper is organized as follows: a brief review on the related work is proposed in Section 2 indicating the main directions for solving the problem. Section 3 presents the basis for constructing the parallel GA approach—elementary theoretical elements are also presented. In Section 4, the ParadisEO-CMW framework is described, along with the subsidiary underlying middleware, Condor-MW, the final part of the corresponding section sketching the general implementation aspects. In Section 5, experimentation results are given with an introductory presentation of the GRID5000 computational grid. Section 6 comprises the conclusions.

2. Related work for the protein structure prediction problem (PSP)

In order to address the PSP problem, by analytical and computational means, a mathematical model that describes inter-atomic interactions must be constructed. The interactions to be considered are a resultant of electrostatic forces, entropy, hydrophobic characteristics, hydrogen bonding, etc. The interactions are quantified in terms of energy levels, relating to the internal energy of the molecule. Precise energy determination also relies on the solvent effect enclosed in the dielectric constant $\epsilon$ and in a continuum model based term.

A trade-off is accepted, opposing accuracy against the approximation level, varying from exact, physically correct mathematical formalisms to purely empirical approaches. The main categories to be mentioned are de novo, ab initio electronic structure calculations, semi-empirical methods and molecular mechanics based models. Hybrid and layered approaches were also designed, in order to reduce the amount of performed calculus to the detriment of accuracy.

The mathematical model describing molecular systems is formulated upon the Schrödinger equation, which makes use of molecular wavefunctions for modeling the spatio-temporal

---

1 http://www.rcsb.org—Brookhaven Protein Data Bank; offers geometrical structural data for a large number of proteins.
probability distribution of constituent particles [10]. It should be noted that, though offering the most accurate approximation, the Schrödinger equation cannot be accurately solved for more than two interacting particles. For resolution related aspects, please consult [27,28]. Extended explanations for the herein exposed directions are available via [10–12,9].

Ab initio (first principles) calculations rely on quantum mechanics for determining different molecular characteristics, comprising no approximations and with no a priori required experimental data. Molecular orbital methods make use of basis functions for solving the Schrödinger equation. The high computational complexity of the formalism restricts their application area to systems composed of tens of atoms.

Semi-empirical methods substitute computationally expensive segments by approximating ab initio techniques. A decrease in the time required for calculus is obtained by employing simplified models for electron–electron interactions: extended Hückel model, neglect of differential overlap, neglect of diatomic differential overlap, etc.

Empirical methods rely upon molecular dynamics (classical mechanics based methods), and were introduced by Alder and Wainwright [16,17]. After more than a decade protein simulations were initiated on bovine pancreatic trypsin inhibitor—BPTI [18]. Empirical methods often represent the only applicable methods for large molecular systems, namely, proteins and polymers. Empirical methods do not make use of the quantum mechanics formalism, relying solely upon classical Newtonian mechanics, i.e. Newton’s second law—the equation of motion. As to the basis of the considered approach, we should mention that, according to recent results [22,23], empirical methods exceed ab initio methods. Conceptually, molecular dynamics models do not dissociate atoms into electrons and nuclei but regard them as indivisible entities. The following list offers a few examples of molecular mechanics force fields:

- AMBER—Assisted Model Building with Energy Refinement;
- CHARMM—Chemistry at HARvard Molecular Mechanics;
- OPLS—Optimized Potentials for Liquid Simulations.

Also, hybrid and layered methods exist [13–15], connecting several methods through various computing architectures, in an attempt to obtain accurate results at low computational costs, and, consequently, in a reduced period of time.

3. A parallel hybrid metaheuristic for solving PSP

3.1. Multicriterion evolutionary algorithm basis

Evolutionary algorithms are stochastic search iterative techniques, with a large area of application—epistatic, multimodal, multicriterion and highly constrained problems [1]. Stochastic operators are applied for evolving the initial randomly generated population, in an iterative manner. Each generation undergoes a selection process, the individuals being evaluated by employing a problem specific fitness function.

Algorithm 3.1. EA pseudo-code.

```
Generate(P(0));
t := 0;
while not Termination_Criterion(P(t)) do
    Evaluate(P(t));
    P'(t) := Selection(P(t));
    P'(t) := Apply_Reproduction_Ops(P'(t));
    P(t + 1) := Replace(P(t), P'(t));
t := t + 1;
endwhile
```

The pseudo-code above shows the generic components of an EA. The main subclasses of EAs are the genetic algorithms, evolutionary programming, evolution strategies, etc.

Due to the nontriviality of the addressed problems, requiring extensive processing time, different approaches were designed in order to reduce the computational costs. Complexity is also addressed by developing specialized operators or hybrid and parallel algorithms. We have to note that the parallel affinity of the EAs represents a feature determined by their intrinsic population-based nature. More precisely, the main parallel models are the island synchronous cooperative model, the parallel evaluation of the population and the distributed evaluation of a single solution. For a complete overview on parallel and grid specific metaheuristics refer to [1–4].

3.2. Multicriterion optimization context

A basic introduction to multicriterion theoretical tools is now presented. A succinct overview of existing research directions in multicriterion optimization may be found in [30].

The solution of a multicriterion optimization problem is represented by a multitude of individual feasible solutions—a Pareto-optimal front, to be defined in the following lines. A solution, identified as a composing point of a Pareto front, is designated as a Pareto point.

Definition 1. Let \( x_1, x_2 \in A \) be two feasible solutions for a multicriterion problem \( P \), and \( f : A \to B \), a cost function. We say that solution \( x_1 \) dominates solution \( x_2 \), denoted as \( x_1 < x_2 \), if the following are simultaneously true:

\[
\forall i \in [1, \ldots, t], \quad f_i(x_1) \leq f_i(x_2); \\
\exists i \in [1, \ldots, t], \quad f_i(x_1) < f_i(x_2).
\]

The solutions \( x_1, x_2 \) are said to be non-dominated with respect to each other if neither of the \( x_1 < x_2, x_2 < x_1 \) relations are true, i.e. neither solution dominates the other.

Definition 2. Let \( F \) be a set of solutions for a multicriterion problem \( P \), \( F \subseteq A \). It is said that \( F \) is a Pareto-optimal set (or front) if \( \forall x \in F \) and \( \forall x' \in A - F, x < x' \).

Examples of domination relations may be found in Fig. 1, while Fig. 2 illustrates a Pareto front example.
sequence of atoms. More specifically, each individual is coded as a vector of torsion angle values—Fig. 3.

The defined number of torsion angles represents the degree of flexibility. Apart from torsion angles which move less than a specified parameter, all torsions are rotatable. Rotations are performed in integer increments, energy quantification of covalent bonds and non-bonded atom interactions being used as optimality evaluation criterion.

3.4. A parallel genetic algorithm for solving PSP

Genetic Algorithms (GAs) represent Darwinian-evolution inspired methods, a random population of individuals evolving in generations through different strategies in order for convergence to be achieved, with respect to optimality criteria. The genotype represents the raw encoding of individuals while the phenotype encloses the coded features. For each generation, individuals are selected on a fitness basis, genotype alteration being performed by means of crossover and mutation operators. Applying the genetic operators has as a result the modification of the population’s structure as to intensify exploration inside a delimited segment or for diversification purposes.

The herein described algorithm comes as the result of a meta optimization process [5], experiments being performed for identifying optimal parametrization. A parallel design is considered, the general sustaining architecture of the developed algorithm conceptually following the generic parallel metaheuristic sketch, previously presented.

The granularity of the problem, as a counterpart for the computationally expensive fitness evaluations, biased the resolution pattern towards a parallel, island-model approach. As a consequence, several populations evolve on a master machine, fitness function estimations being distributed on remotely available computing units. We have to note that the evaluation of the fitness function consists of several stages, including the calculation of Cartesian atomic coordinates, inter-atomic distances determination etc. A distributed fitness calculation does not represent an option, incurring a significant synchronization overhead. Common one-point and two-point crossover and mutation operators were used.

3.5. Fitness function

The function to be optimized, under the bicriterion auspices, is computed by making use of bonded atom energy and non-bonded atom energy, as distinct entities. The result obtained is
compared with a Pareto front of solutions, the feasibility of a given individual being related to the dominance concept.

An intuitive reasoning leads to the fact that the bonded and the non-bonded energy terms are antagonist (verified through the performed experiments), although no formal demonstration exists in the literature. Hence, it may be stated that the problem qualifies for multicriteria optimization. The quantification of energy is performed by using empirical molecular mechanics, under the CHARMM realm as follows:

\[ E_{\text{bonded}} = \sum \text{bonds } K_b(b - b_0)^2 + \sum \text{bondangle } K_\theta(\theta - \theta_0)^2 \]

\[ + \sum \text{torsion } K_\phi(1 - \cos n(\phi - \phi_0)) \]

\[ E_{\text{non-bonded}} = \sum \text{VdW } \frac{K_{ij}^{\text{VdW}}}{d_{ij}^{12}} - \frac{K_{ij}^{\text{VdW}}}{d_{ij}^6} + \sum \text{Coulomb } \frac{q_i q_j}{4\pi\varepsilon d_{ij}} \]

\[ + \sum \text{desolvation } \frac{K q_i^2 V_j + q_j^2 V_i}{d_{ij}^4} \]

where \( E_{\text{bonded}} \) and \( E_{\text{non-bonded}} \) represent the energy of the bonded and non-bonded contributions respectively.

The involved factors model oscillating entities, the interatomic forces being conceptually simulated by considering interconnecting springs between atoms. At this point, a specific constant is associated with each type of interaction, notationally denoted by \( K_{\text{inter}} \). An optimal value for the considered entity (bond, angle, torsion) is introduced in the equation as reference for the variance magnitude—\((T - T_0)\). \( T \) stands for the experimentation value, while \( T_0 \) specifies the natural, experimentally observed value, when the entity is pulled out of its context.

In more specific terms, \( b \) represents the bond length, \( \theta \) the bond angle, \( \phi \) the torsion angle and \( q \) the electrostatic charge associated with a given atom, the distance between the \( i \) and the \( j \) atoms and a volumetric measure for the \( p \) atom respectively.

An example of \( \alpha \)-cyclodextrin energy surface is given in Fig. 4. The set of corresponding molecular conformations was obtained by modifying a random initial conformation. More specifically, an arbitrary conformation has been generated, subsequently, two torsional angles being chosen at random. For each of the two torsional angles, values between 0 and 360 have been considered, in \( 10^\circ \) increments, all the other torsional angles being maintained rigid. Thus, 1225 conformations were obtained—the lighter areas on the obtained surface correspond to high-energy conformations. Furthermore, an energy-map representation is given, in the \( XY \)-plane—only the dark regions are meaningful.

Although smooth, the obtained surface is the result of only two torsional angles variation. The hyper-surface, generated by varying the entire set of torsional angles, has an extremely rough landscape, with a large number of local optima.

Fig. 5 depicts the bonded and non-bonded atom derived energies, corresponding to the previous energy surface, shown in 4. The energy surfaces are computed as given by the previously exposed force field.

As can be seen from the figure, the non-bonded atom derived energy component has large values in comparison with the bonded atom derived energy component. The high-energy values for the non-bonded component are determined by the large number of non-bonded interactions, as pairs of atoms are considered.
3.6. Hybridization with a Hill Climbing local search

The developed method has as backbone structure a hybrid architecture, combining a genetic algorithm with a conjugated gradient-based Hill Climbing local search method—a Lamarckian optimization technique.

The exploration and the intensification capabilities of the genetic algorithm do not suffice as paradigm, when addressing rough molecular energy function landscapes. Small variations of the torsion angle values may generate extremely different individuals, with respect to the fitness function. As a consequence, a nearly optimal configuration, considering the torsion angle values, may have a very high energy value, and thus it may not be taken into account for the next generations.

In order to correct the above exposed problem, a conjugated-gradient based method is applied for local search, alleviating the drawbacks determined by the conformation of the landscape. Fig. 6 was obtained by applying the local search technique for each of the conformations that were used for the $\alpha$-cyclodextrin energy surface in Fig. 4.

Although reducing both energies, for the bonded and non-bonded type interactions, the non-bonded energy component still represents the major part of the total energy, as can be seen in Fig. 7.

4. ParadisEO-CMW based implementation

4.1. The ParadisEO framework

The ParadisEO\(^2\) framework is dedicated to the reusable design of parallel hybrid meta-heuristics by providing a broad range of features, including EAs, local search methods, parallel and distributed models, different hybridization mechanisms, etc. The rich content and utility of ParadisEO increases its usefulness.

ParadisEO is a C++ LGPL white-box open source framework, based on a clear conceptual separation of the meta-heuristics from the problems they are intended to solve. This separation, and the large variety of implemented optimization features, allow a maximum code and design reuse. Changing existing components and adding new ones can be easily done, without impacting the rest of the application.

ParadisEO is one of the rare frameworks that provide the most common parallel and distributed models, portable on distributed-memory machines and shared-memory multiprocessors, as they are implemented using standard libraries such as MPI, PVM and PThreads. The models can be exploited in a transparent way—one has just to instantiate its associated ParadisEO components. The user has the possibility of choosing, by a simple instantiation, the MPI or the PVM for the communication layer. The models have been validated on academic and industrial problems, and the experimental results demonstrate their efficiency [4].

4.2. The ParadisEO-CMW framework

The ParadisEO-CMW framework targets non-dedicated environments, having as sustaining structure the ParadisEO framework and the Condor-MW middleware.

The Condor\(^3\) system [33,34] is a high-throughput computing (HTC) system that deals with heterogeneous computing

---

\(^2\) http://www.lifl.fr/~cahon/paradisEO/common.

\(^3\) http://www.cs.wisc.edu/condor/condorg.
resources and multiple users. It allows the management of non-dedicated and volatile resources, by deciding their availability, using both the average CPU load and the information about the recent use of some peripherals, like the keyboard and the mouse. An environment including such resources is said to be adaptive, since tasks are scheduled among idle resources, and dynamically migrated when some resources get used or failed. In addition, Condor-PVM uses some sophisticated
techniques [31] like matchmaking and checkpointing. These allow us, respectively, to associate job requirements and policies on resources owners, and to periodically save/restart the state of/from running jobs.

MW [32] is a software framework allowing an easy development of Master–Worker applications for computational grids. MW is a set of C++ abstract classes including interfaces to application programmers and Grid-infrastructure developers. Grid-enabling an application with MW, or porting MW to a new grid software toolkit, consists in re-implementing a small number of virtual functions. In MW, the infrastructure interface provides access to communication and resource management. The communication is performed between the master and the workers. The resource management encompasses: available resource request and detection, infrastructure querying to get information about resources, fault detection, and remote execution. These basic resource management services are provided by Condor-PVM.

One of the major design goals of MW is to ensure a maximum programmability, meaning that the users should easily be able to interface an existing code with the system. Therefore, porting ParadisEO to Condor-MW can be easily done through the use of the infrastructure and application programming interfaces provided by MW. Moreover, the coupling is facilitated by the fact that the two frameworks are written in C++. The architecture of ParadisEO-CMW is layered as is illustrated in Fig. 8.

From a top-down view, the first level supplies the optimization problems to be solved using the framework. The second level represents the ParadisEO framework, including optimization solvers, embedding single and multicriterion meta-heuristics (evolutionary algorithms and local searches). The third level provides interfaces for Grid-enabled programming and for access to the Condor infrastructure. The fourth and lowest level supplies communication and resource management services.

An important issue to deal with in Grid computing is the fault-tolerance. MW automatically reschedules unfinished tasks as they were running on processors that failed. This cannot be applied to the master process that launches and controls tasks on worker nodes. Nevertheless, a couple of primitives are provided to fold up or unfold the whole application, enabling the user to save/restart the state of/from a file stream. Dealing with meta-heuristics, these functionalities are easily investigated. Checkpointing most of the meta-heuristics is straightforward. It consists at least in saving the current solution(s), the best one found since the beginning of the search, the continuation criterion (e.g. the current iteration for a generational counter) and then some additional parameters controlling the behavior of the heuristic. In ParadisEO-CMW, default checkpoint policies are initially associated to the deployed meta-heuristics.

4.3. Implementation

The implementation relies on invariant elements provided by the ParadisEO-CMW framework, providing support for the insular model approach, as well as for distributed and parallel aspects concerning the parallel population evaluation. In this context, deployment related aspects are transparent, the focus being oriented on the application-specific elements.

The main steps to be performed, in order to configure the environment and to deploy the algorithm, consist in specifying the individual’s encoding, the specific operators and the fitness function. Furthermore, elements concerning selection mechanisms and replacement strategies must be specified, along with configuration parameters (number of individuals, number of generations etc).

5. Experiments and results

For the developed application, the deployment has been performed on a layered framework design, the composing elements being the following: Condor, MW—Master–Worker, ParadisEO-CMW.

The underlying support for performing the experiments was GRID5000, a French nationwide experimental grid, connecting several sites which host clusters of PCs interconnected by RENATER\(^4\) (the French academic network). The GRID5000 is promoted by CNRS, INRIA and several universities.\(^5\)

By the end of 2006 the GRID should gather 2500 processors with 2.5 TB of cumulated memory and 100 TB of non-volatile storage capacity. Inter-connections sustain communications of


2.5 Gbps (10 Gbps soon). The GRID5000 infrastructure offers several tools for controlling, manipulating and supervising activities, Fig. 9 representing a real-time snapshot of the GRID.

The target point to be achieved is a marker-stone of 5000 processors for 2007—at this moment there are almost 2000 processors at this time being, regrouping nine centers: Bordeaux, Grenoble, Lille, Lyon, Nancy, Orsay, Rennes, Sophia-Antipolis, Toulouse. The following results were obtained by performing deployments on the Lille cluster of GRID5000.

The addressed molecular complexes for grid deployment tests were tryptophan-cage (Protein Data Bank ID 1L2Y) and α-cyclodextrin. The trp-cage miniproteins present particular fast folding characteristics, while cyclodextrins, in α, β or γ conformations, are important for drug-stability applications, being used as drug protectors against micro-environment interactions or as homogeneous distribution stabilizers etc.

Structural profile of the tryptophan-cage protein: an α-helical N-terminal region, a short helix and a polyproline II helix at the C-terminus wrapping around for packing the Trp residue within a compact hydrophobic core [35]. Cyclodextrins, as non-reducing macrocyclic oligosaccharides, are constituted as D-glucopyranosyl units interconnected through α – (1, 4) glycosidic links. The ensemble builds as a toroidal structure with hydrophobic interior.

Table 1 offers information regarding the number of active elements used when executing the algorithm—determining the degree of flexibility considered for each of the molecules and, consequently, the dimension of the conformational space. The complexity of the model augments in concordance with the number of active elements—the table lists the considered active elements for each of the molecules under study. The last two lines offer the initial, respectively the final, number of interactions between non-bonded atoms. A cut-off is performed in order to reduce complexity, having as basis inter-atomic distances (interactions between atoms too far apart are ignored, as they cannot contribute significantly, in energy terms). We have to note that energy calculations for the non-bonded atoms set represent the main computational factor, as pairs of non-bonded atoms must be considered. In conjunction with the initial discussion on computational complexity, present in the introduction, the presented data confirm once more the need for a massive parallel computing environment.

In the followings lines, preliminary results are given, execution times for several performed tests being listed in Table 2. For each deployment, identical biprocessor machines were used, the number of computing units being listed on the left outer column. At the same time, the speed-up is depicted in Fig. 10—we are to remember that biprocessor machines were used, the enclosed data relating to distribution aspects.

Figs. 11 and 12 graphically represent the obtained Pareto fronts for the two above mentioned molecular systems—the Pareto points are marked by the blue triangles.
intermediary conformations might exist, though the sampling mechanism might have missed their associated region of space. The previous effect is driven by the granularity of the conformational sampling mechanism.

At the opposite extreme, as can be seen from Fig. 13, there are cases in which the degradation of one energy function does not incur improvements in the complementary energy function. In such cases, both energy functions undergo a degradation, in energy level terms. As a consequence, several neighboring conformations, having almost identical structure, might be separated by high potential barriers, in which case, no Pareto solutions exist between the conformations. This latter case is also determined by how close the mentioned conformations are to local optima, and the granularity considered when representing the torsional angles, with respect to energy variations.

6. Conclusions and future work

Multicriterion problems in general, and protein structure prediction under bicriterion aspects in particular, remain an open research field due to complexity matters, and of extreme importance in multiple domains. Mesoscopic and macroscopic characteristics represent the product of statistical interaction of an ensemble of near-optimal molecular conformations, a more complete description being achievable by defining not only the ground-state energy conformation of a molecule but also the ensemble of potential low-energy conformations.

The reported grid-enabled method offers a proof of feasibility, distributed techniques sustaining complex simulations. Multicriterion approaches, though potentially inducing augmented complexity, provide more accurate solutions for real-life problems, overcoming in particular cases the limitations of monocriterion resolution patterns. At this moment, experimentation and research are underway for specialized operators, exploiting directed mutation operators and approximative models as well as novelty force fields. We also plan to tackle larger molecular complexes using parallel hybrid GAs on a larger computational grid. In this case, the exploitation of the two
parallel models of GAs in a hierarchical way requires several thousands of processors.

References


A.-A. Tantar received the Master’s degree from the Faculty of Computer Science, “A.I. Cuza” University of Iasi, Romania. He is currently a Ph.D. student within the OPAC team at Laboratoire d’Informatique Fondamentale de Lille (LIFL, Université de Lille 1). He is involved in the DOLPHIN project of INRIA Futurs. His major research interests include parallel and grid computing, and combinatorial optimization algorithms and applications.

N. Melab received the Master’s, Ph.D. and HDR degrees in computer science, both from the Laboratoire d’Informatique Fondamentale de Lille (LIFL, Université de Lille 1). He is a Professor at Université de Lille 1 and a member of the OPAC team at LIFL. He is involved in the DOLPHIN project of INRIA Futurs. He is particularly a member of the Steering Committee of the French Nation-Wide project Grid5000. His major research interests include parallel and grid computing, combinatorial optimization algorithms and applications and software frameworks.

E.-G. Talbi received the Master’s and Ph.D. degrees in computer science, both from the Institut National Polytechnique de Grenoble. He is presently Professor in computer science at Polytech’Lille (Université de Lille 1), and researcher in Laboratoire d’Informatique Fondamentale de Lille. He is the leader of OPAC team at LIFL and the DOLPHIN project of INRIA Futurs. He took part to several CEC Esprit and national research projects. His current research interests are mainly parallel and grid computing, combinatorial optimization algorithms and applications and software frameworks.
B. Parent is an engineer from the “Institut Supérieur d’Electronique et du Numerique” (Lille) and got his Master’s degree in cybernetics and computer science from the “Ecole Centrale de Lille”. Currently doing his Ph.D. in Biology and Biophysics, his main research interests involve the study and development of analysis and optimization algorithms for highly-dimensional, non-linear problems.