STUDIA UNIVERSITATIS BABEŞ-BOLYAI, CHEMIA, LIII, 2, 2008

GRID-BASED CONFORMATIONAL SAMPLING

BENJAMIN PARENT[a], ALEXANDRU TANTAR[b], NOUREDINE MELAB[b], EL-GHAZALI TALBI[b], DRAGOS HORVATH[c,*]

ABSTRACT. Computational simulations of conformational sampling in general, and of macromolecular folding in particular represent one of the most important and yet one of the most challenging applications of computer science in biology and medicinal chemistry.

This paper presents a massively parallel GRID-based conformational sampling strategy, exploring the extremely rugged energy response surface in function of molecular geometry, in search of low energy zones through phase spaces of hundreds of degrees of freedom.

We have generalized the classical island model deployment of Genetic Algorithms (GA) to a "planetary" model where each node of the GRID is assimilated to a "planet" harboring quasi-independent multi-island simulations using a hybrid GA.

Although different "planets" do not communicate to each other -thus minimizing inter-CPU exchanges on the GRID- each new simulation will benefit from the preliminary knowledge of already visited geometries, located on the dispatcher machine, and which are disseminated to any new "planet". This "panspermic" strategy allows new simulations to be conducted such as to either be attracted towards an apparently promising phase space zone (biasing strategies, intensification procedures) or to avoid already in-depth sampled (tabu) areas.

Successful all-atom folding of mini-proteins typically used in benchmarks (the Trp cage 1L2Y and the Trp zipper 1LE1) has been observed, although the reproducibility of these highly stochastic simulations in huge problem spaces is still in need of improvement.

Keywords: *molecular modeling, conformational sampling, protein folding simulations, genetic algorithms, massively parallel computing, molecular force fields*

INTRODUCTION

The prediction of 3D shapes of molecules on hand of their connectivity (the so-called *Conformational Sampling, CS*) is a widely addressed, central problem in structural biology and drug design[1]. There are yet no general approaches able to enumerate, for an arbitrary (macro)molecule, the most

[a] *Institut Suprieur d'Electronique et du Numérique; 41, Bd. Vauban, 59000 Lille, FR*
[b] *Laboratoire d'Informatique Fondamentale, Cité Scientifique, 59655 Villeneuve d'Ascq, FR*
[c] *Laboratoire d'InfoChimie, Univ. Louis Pasteur, 4, rue Blaise Pascal, 6700 Strasbourg, FR*
* *Corresponding author: horvath@chimie.u-strasbg.fr*

stable molecular geometries adopted in solution. Several proofs of the NP-completeness of such a problem have been proposed on hand of different models [2, 3], while protein chemists are well aware of the Levinthal paradox [4]. The reformulation in terms of an energy landscape [5], where the energy is a force field-based [6, 7] function of internal coordinates (in this case, dihedral angles around the considered rotatable bonds), enables to attack the problem in the framework of function optimization. Energy minima then correspond to the populated geometries of the molecule. The extreme ruggedness of the response hypersurface causes any deterministic optimization attempt to get stuck in local, most likely irrelevant optima and imposes the use of stochastic sampling procedures.

The ability of GAs to deal with a set of solutions while deriving profit of an intrinsic stochastic behavior in addition to the recombination principle, makes them a suited tool for challenging highly multimodal and highly dimensional problems [8]. The high computational costs, on one hand, and the straightforwardness of parallel deployment strategies for genetic algorithms, on the other, make this problem an ideal candidate for GRID computing. In the following we report, after a short introduction of the hybrid island model [9], a first successful deployment strategy on the parallel GRID[1] context, the "planetary'" model.

The hybrid GA deployed on the "planets" (nodes) of the GRID operates on the degrees of freedom associated to the rotations around interatomic single bonds, so that a chromosome represents the vector of torsional angles associated to rotatable bonds. Certain peculiarities of the sampling problem ask for hybridizations [9] of the GA with other optimization procedures (conducting "Lamarckian" local optimizations to repair local clashes in what would otherwise represent stable conformers, allow for "directed" mutations, permitting the other degrees of freedom to adjust in response to the random shift applied to the mutated chromosome locus, introduce population diversity management criteria, biased random distributions for each degree of freedom, etc).

In the planetary approach, a dispatcher script attempts to deploy island models on as many nodes (planets) as requested, if it can find the resources on the GRID. Once an island model is completed according to the locally specified termination criteria, the island model pilot script sends the locally sampled results back to the dispatcher, which will join them to the "Universal" pool of solutions. Liberation of a node will prompt the dispatcher to restart an island model there, until a total (user-specified) number of sets of results were successfully retrieved, or until the latest (user-defined) N retrieved results failed to contain any fitter solutions. The behavior of each island

[1] supported by the French GRID5000 initiative (www.grid5000.fr) and the Agence Nationale de la Recherche

model is controlled by a set of operational parameters dictated by the dispatcher, which actively tries to optimize these in order to achieve better sampling capacity of the further runs.

A key element of our deployment strategy is “panspermia”, so entitled after the hypothesis that life on Earth might have been seeded by microorganisms from space: the dispatcher may randomly pick a subset of the already visited solutions from the “Universal” pool and “seed” any newly started CS run. The latter may use the provided sample to specify these as tabu zones [10] -forcing the exploration of other phase space zones (default exploratory runs)- or to replace the random initialization of chromosomes by cross-over products of these “ancestors”, thus allowing an in-depth exploration of promising phase space regions (intensification runs).

Unfortunately, the ruggedness of the energy landscape is such that near-native structures (according to geometric criteria) may nevertheless display high energies and fail to rank among the populated states. A specific “intensification” scheme for the GA, [11] allowing the fine exploration of limited phase space zones has therefore been designed. Its initial populations are not random, but loaded with previously sampled geometries representing a same global fold, in search for states of similar overall geometry but lower energy.

Heavily visited phase space zones, where the deepest zonal optimum has been (expectedly) sampled, will be declared tabu areas in future exploratory runs. Any solution close, according to a given similarity metric and similarity cut-off, to a tabu chromosome, will be assigned a low fitness score. The choice of the similarity metric and cut-off (see [11] for details) is paramount: too broad taboo areas may block the access to unexplored deeper local minima in the neighborhood.

The key challenge of an optimal panspermic strategy is to decide at which point chromosomes have served enough in intensification searches, and therefore should be declared tabu. A too early decision in this sense may prematurely block the discovery of deep energy wells, while a too late one translates in wasted computer time. Common sense might suggest that intensification should be applied only to chromosomes of reasonably low energies, but in practice it is unclear what “reasonably low” is supposed to mean: intensification may dramatically lower intramolecular strain.

RESULTS AND DISCUSSION

Up-to-date attempts to use the planetary model led to successful folding experiments of the Trp cage 1L2Y [12] (α-helix) and Trp zipper 1LE1 [13] (β-sheet), in a matter of few days, using only a small subset (20-30 nodes) of GRID5000.

α-helices are structural elements that fold quickly in solution, being stabilized by local, energetically favorable hydrogen bonds. This situation is well suited for GA-based sampling: a helix turn is controlled by 6 degrees of freedom only, i.e. may quite easily emerge by hazard in a chromosome. Being stabilized by internal hydrogen bonds, this structural element is readily inherited, until a favorable cross-over may couple two spontaneously emerged helix loops together. Accordingly, the planetary model has successfully and reproducibly discovered, as shown in Figure 1, geometries that are very close to the native 1L2Y fold reported in literature.

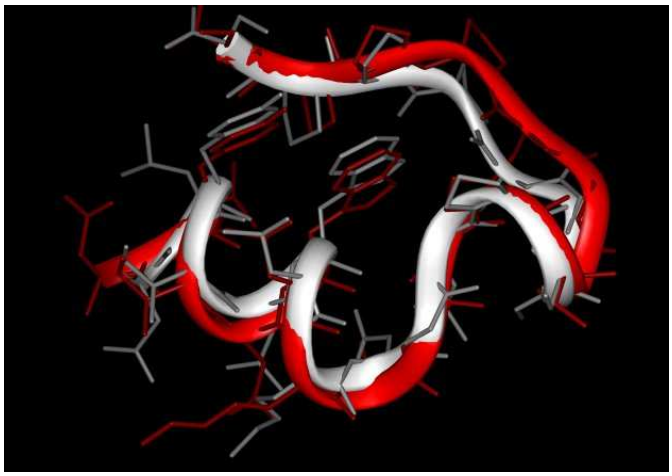

Figure 1. Best ranked (lowest energy) conformer found by the planetary model for 1L2Y (red) overlayed to native 1L2Y geometry (white). Heavy atom RMSD is of 1.8 Å

By contrast, although the Trp zipper has only 53 degrees of freedom, it is nevertheless more difficult to fold computationally than 1L2Y. The main reason is its β-hairpin structure, where stabilizing hydrogen bonds stem from topologically remote pairs of aminoacids. The β-sheet zipper is a cooperative element: it gains stability only when fully structured: chromosomes displaying partly folded sheets will not benefit from stabilization, *i.e.* do not have any obvious evolutionary advantage. This notwithstanding, correctly folded protein backbones have been reproducibly obtained by planetary model-based simulations. In rare cases (2 out of several tens), the simulation actually returned a perfect replica of the experimental fold, both in terms of backbone and side chain orientations. This calculated geometry (Figure 2) was also shown to be the most stable of all the ever visited 1LE1 conformers.

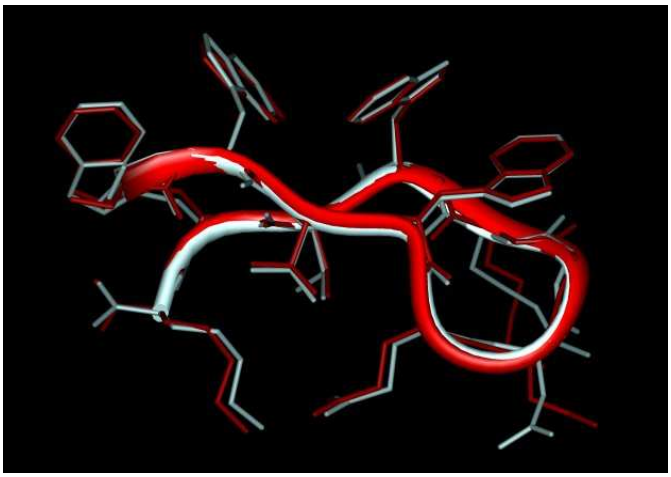

Figure 2. Some simulations lead to the discovery of the exact native geometry of 1LE1

Typical simulations, however, return geometries like in Figure 3, where the backbone is correctly folded but side chains are misplaced. The alternative side chain interactions make physico-chemical sense: (aromatic stacking, like in the native geometry). The computed conformer is not obviously wrong: it may actually correspond to some less populated species which escapes detection by state-of-the-art experimental methods. However, its computed energy is significantly higher than the one of the native state and, unfortunately, also higher than the one of misfolded structures: it was ranked as the 79th most stable geometry out of several hundreds of thousands.

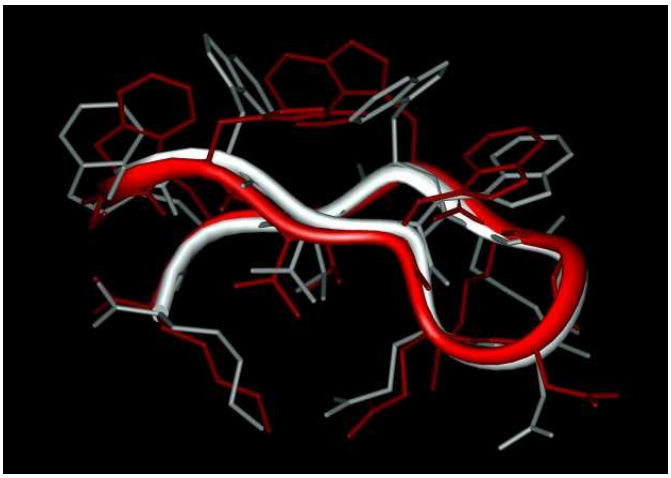

Figure 3. Typical 1LE1 folding behaviour: *almost* native structures differing from the latter only with respect to their side chain placement, are readily found.

CONCLUSIONS

The herein presented “planetary” conformational sampling approach showed some promising first results – correct folding of helical peptides, and occasional convergence towards the native geometry of β-sheet peptides. However, the “panspermic” strategy at the core of the planetary approach needs further refinement, for the failure to systematically converge towards the native 1LE1 geometry may be ascribed to an overhasty setting of tabu areas in the neighborhood of almost native geometries, with a correct backbone fold but wrong side chain orientations. Such tabu zones are responsible, in addition to the intrinsic ruggedness of the energy landscape, for the failure of the method to successfully move from these relatively energy-rich almost correct folds to the low energy native structure.

REFERENCES

1. J. N. Onuchic, P.G. Wolynes, *Curr. Op. Struct. Biol.*, **2004**, *14*, 70.
2. P. Crescenzi, D. Goldman, C.H. Papadimitriou, A. Piccolboni, M. Yannakakis, *J. Comp.Biol.*, **1998**, *5*, 423.
3. R. Unger, J. Moult, *J. Mol. Biol.*, **1993**, *231*, 75.
4. C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems.*, University of Illinois Press, Chicago, Editon edn., 1969, pp. 22.
5. D. J. Wales, T.V. Bogdan, *J. Phys. Chem.*, **2006**, *110*, 20765.
6. A. T. Hagler, E. Huler, S. Lifson, *J. Am. Chem. Soc.*, **1974**, *96*, 5319.
7. A. T. Hagler, S. Lifson, *J. Am. Chem. Soc.*, **1974**, *96*, 5327.
8. J. Holland, *Adaptation in Natural and Artificial Systems.*, University of Michigan Press, Ann Arbor, 1975.
9. B. Parent, A. Kökösy, D. Horvath, *Soft Computing*, **2007**, *11*, 63.
10. F. Glover, J.P. Kelly, M. Laguna, *Computers and Operations Research*, **1995**, *22*, 111.
11. B. Parent, A. Tantar, N. Melab, E.-G. Talbi, D. Horvath, IEEE Congress on Evolutionary Computation, CEC 2007, Singapore, http://www.ieeexplore.ieee.org/iel5/4424445/4424446/04424484.pdf?tp=&arnumber=4424484&isnumber=4424446, 2007.
12. J. W. Neidigh, R. M. Fesinmeyer, N. H. Andersen, *Nature Struct. Biol.*, **2002**, *9*, 425.
13. A. G. Cochran, N.J. Skelton, M.A. Starovasnik, *Proc. Natl. Acad. Sci. USA*, **2001**, *98*, 5578.