Grid-based Evolutionary Strategies Applied to the Conformational Sampling Problem.

Benjamin Parent, Alexandru Tantar, Nouredine Melab, El-Ghazali Talbi, Dragos Horvath

***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. The advent of GRID computing may trigger some major progress in this field. This paper presents our first attempts to design GRID-based conformational sampling strategies, 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 ($\mathcal{GA}$) to a "planetary" model where each node of the grid is assimilated to a "planet" harboring quasi-independent multi-island simulations based on a hybrid GA-driven sampling approach. 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 extracted from the centralized pool of already visited geometries, located on the dispatcher machine, and which is 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 folding of mini-proteins typically used in benchmarks for all-atoms protein simulations has been observed, although the reproducibility of these highly stochastic simulations in huge problem spaces is still in need of improvement. Work on two structured peptides (the "tryptophane cage" 1L2Y and the "tryptophane zipper" 1LE1) used as benchmarks for all-atom protein folding simulations has shown that the planetary model is able to reproducibly sample conformers from the neighborhood of the native geometries. However, within these neighborhoods (within ensembles of conformers similar to models published on hand of experimental geometry determinations), the energy landscapes are still extremely rugged. Therefore, simulations in general produce "correct" geometries (similar enough to experimental model for any practical purposes) which sometimes unfortunately correspond to relatively high energy levels and therefore are less stable than the most stable among misfolded conformers. The method thus reproducibly visits the native phase space zone, but fails to reproducibly hit the bottom of its rugged energy well. Intensifications of local sampling may in principle solve this problematic behavior, but is limited by computational ressources. The quest for the optimal time point at which a phase space zone should stop being intensively searched and declared tabu, a very difficult problem, is still awaiting for a practically useful solution.**

I. Introduction

The prediction of three-dimensional shapes of molecules on hand of their connectivity (the so-called *Conformational Sampling* task or simply $\mathcal{CS}$) 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 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] that frustrate computationalists and illustrate the Levinthal paradox [4]. The reformulation in terms of an energy landscape [5] where the energy, expressed as a function of geometry, is to be minimized, enables to attack the problem in the framework of function optimization. The energy minima then correspond to the populated geometries of the molecule; however entropic effects embedded in the widths of the wells, and which play an important role in determining the *free energy* are very difficult to estimate.

The huge problem size (hundreds of degrees of freedom), is actually not the major challenge: the extreme ruggedness of the response hypersurface (molecular energy as a function of internal coordinates: dihedral angles around the considered rotatable bonds, in this case) causes any deterministic optimization attempt to get stuck in local, most likely irrelevant optima and imposes the use of stochastic sampling procedures. However, the probability of discovering the very narrow low energy zones of phase space by randomly drawing the correct coordinates is virtually null.

A. Conformational sampling task in all-atom description

The estimation (according to a classical force field) of the internal energy of a given structure, in function of the relative positions of the atoms, offers an objective score, allowing to reformulate the question in terms of optimization theory: Boltzmann's equation (1) provides the population level of each state.

$$\Pr(\text{system in state of energy } E) \propto \exp\left(-\frac{E}{k_B T}\right) \quad (1)$$

where T is the absolute temperature and k_B, the Boltzmann constant.

This equation stresses that, no matter how numerous, all the low-energy minima within a few $k_B T$ from the absolute bottom of the energy hypersurface will be populated and are, therefore, important. Every conformational sampling algorithm must therefore address the (highly) multimodal aspect of the optimization.

Since the herein described software is aimed at docking problems and affinity estimation of small ligands with protein binding sites, an all-atom level of description is required. The empirical force field used to estimate the molecular energy as a function of geometry has been derived from the Consistent Valence Force Field [6], [7] (CVFF), enhanced by the addition of a continuum solvent model [8]. Although intrinsically

1-4244-1340-0/07$25.00

inaccurate, the force field-based energy estimation allows a far simpler, Newtonian, description of the problem compared to the correct quantum mechanical formalism.

Whereas molecular dynamics and/or Monte Carlo simulations, proceeding by small perturbations of a local geometry, may successfully avoid visiting the ubiquitous high-energy regions of phase space (provided a low-energy starting geometry is available!), they tend to spend too much time in exploring the local neighborhoods rather than pushing forward to yet uncharted phase space regions. The $\mathcal{GA}$ ability to deal with a set of solutions while deriving profit of both an intrinsic stochastic behavior in addition to the recombination principle, made them, in our opinion, the most suited tool for challenging highly multimodal / highly dimensional problems [9]. Our previous experience [10] showed that hybrid genetic algorithms, relying on the synergy between random exploration, selection and local calls to specific optimization procedures (tailor-made to respond to the peculiarities of the molecular energy landscape), have the ability to successfully cope with the challenges of conformational sampling. Nevertheless, this software would require weeks to month on a typical two-processor workstation in order to complete the successful folding (discovery of the experimentally known energy minimum) of peptides typically used in all-atom folding simulations (tryptophane cage, pdb code 1L2Y [11], 20 aminoacids; tryptophane zipper, pdb code 1LE1 [12], 13 aminoacids; the PIN1 WW domain, 34 aminoacids [13], etc.). 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.

Here we report, after a short introduction of the hybrid island model, a first successful deployment strategy on the parallel GRID[1] context. This "planetary" model was so dubbed as it represents a generalization of the classical island strategy, where each node of the grid represents a "planet" on which an island model will be started. It enables the controlled sharing of computational effort between global Darwinian exploration (some "planets" will be charged with the search for novel, different, low energy folds) and intensification (others perform local searches for the absolute energy minimum within the neighborhoods of newly discovered, "raw" geometries, to fine tune structural details - with potentially dramatic decreases in molecular energies).

II. $\mathcal{GA}$ IMPLEMENTATION

A. Genetic Algorithms

The hybrid GA deployed on the "planets" of the GRID operates on the degrees of freedom associated to the rotations around interatomic single bonds (figure 1), so that a chromosome actually represents the list, or vector of torsional angles associated to each of the considered rotatable bonds: $\overrightarrow{\Theta} = (\Theta_i, i = 1...N_{\text{rotBonds}})$.

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

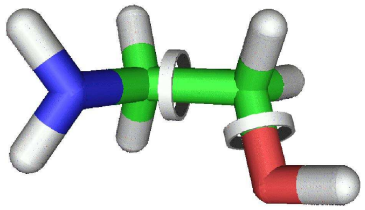

Fig. 1. Torsional angle coding

Certain peculiarities of the sampling problem may ask for hybridizations of the genetic algorithm 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 and "tabu" criteria to block revisiting already sampled phase space zones, bias random distributions for each degree of freedom in order to enhance the probability of drawing values seen to occur in stable conformers, etc. — see below). Moreover, the control parameters inherent to the genetic algorithms (population size, mutation and crossover rates, maximal age, ending condition etc.) have a dramatic impact on the sampling performance. An additional layer of optimization, in search of the optimal operational regime of the $\mathcal{GA}$ for a current sampling problem, was therefore implemented as part of a global sampling strategy involving many successive and/or parallel $\mathcal{GA}$ runs.

B. Hybridizations

a) Parallelism: An island model [14] allows parallel implementations of the core $\mathcal{GA}$ to run independently, but with occasional inter-island migrations of solutions. This basic parallelization scheme favors exploration since each island may in principle harbor a distinct population which may nevertheless be challenged by fitter migrants if it fails to evolve as fast as competing islands. Care should be taken while designing the migration mechanism, to prevent genetic material from spreading to more than one island.

b) Non-uniform probability laws: while $\mathcal{GA}$s usually make use of flat distribution of probability to draw random values for each locus of the chromosome, introducing any knowledge and biasing the search towards peculiar regions of the phase space is possible by modifying these probability laws. The 'knowledge-based' biasing strategy relies on a local energy strain estimation, such that locally more stable staggered conformations will be favored over eclipsed ones. The other, 'tradition-based', strategy exploited here relies on statistics about the preferentially adopted torsional values in the fittest solutions currently available. This latest paradigm

suffers from its self-consistency and it has been shown that extreme caution should be taken to ensure that a sufficiently diverse and relevant pool of precursor solutions is at hand before actively favoring herein encountered torsion angle values. With this reserve, these biasing mechanisms have proven to speed up the overall progression of the populations.

c) Deterministic optimizations: in addition to an occasionally applied conjugated gradient relaxation of individuals (or 'Lamarckian optimization', [15]), a new heuristic has been implemented, taking advantage of both deterministic optimization and stochastic mutations. This search strategy, which actually relies on the 'Torsional Angle Driving' procedures [16], forces one randomly chosen degree of freedom towards a randomly determined target value, by means of an artificial harmonic constraint term added to the energy function to be minimized. A conjugated gradient optimization then allows the torsions to relax in a concerted manner, according to this new fitness landscape, towards the desired torsional value, avoiding the clashes that would have probably arisen if rigid fragments would have been rotated around the given axis (as is the case in classical random mutation). As this deterministic optimization procedure is quite time consuming and would cause serious disruption of the evolutionary loop if run within the islands; it has therefore been programmed under the form of stand-alone 'explorer' processes, started by a $\mathcal{GA}$ run.

III. Meta optimization

The performance of the Conformational Sampling $\mathcal{GA}$ ($\mathcal{CSGA}$) being quite sensitive with respect to the choice of the control parameter values, this choice has been addressed by means of a meta layer of optimization, favoring parameters sets that enhance the search procedure.

The '$\mathcal{CSGA}$ success' optimality criterion (equation 2), took into account both computational time and the so-called '*free energy*' of the sampled conformer ensemble (implicitly accounting for multimodality) at the current operational setup.

$$\mu Fitness = -k_B T \times \ln \left[\sum_{\substack{i \in \text{found} \\ \text{conformers}}} \exp\left(-\frac{E_i}{k_B T}\right) \right] + \alpha \times Time \quad (2)$$

The importances of the meta optimization procedure and the hybridizations was analysed in details elsewhere [10]. This optimized and hybridized tool was able to process bigger molecules (up to a hundred degrees of freedom) at the atomic level in acceptable computing times ($\sim$ one week).

IV. Massively parallel deployment — Planetary model

The above described hybrid Darwinian process is started simultaneously on an arbitrary, user-defined number of planets (nodes): a dispatcher script attempts to deploy island models on as many nodes as requested, if it can find the ressources on the GRID. There is no 'interplanetary' communication at all: fit solutions may only be swapped between islands. Once an island model is completed according to the locally specified termination criteria, or the generic reservation time of that node is about to expire, the pilot script in charge of running the island model will, before termination, send 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 exact behavior of the starting island 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.

Like in the workstation version, the meta-optimization of the operational parameters is performed by learning from previous runs, though a simple genetic algorithm, which runs asynchronously in the planetary model (upon termination of a node, its sampling success is brought in relation to the operational parameters it had used, and this knowledge is stored in a database serving to pick a new operational parameter configuration whenever the next node is due to start).

A. Panspermia

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 planet. The latter may use the provided sample to specify these as 'tabu' zones [17] — forcing the exploration of other phase space zones — 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.

B. Intensification

Although the sampling procedure may rapidly generate structures in the neighborhood of the 'native' (experimentally determined) geometries, the extreme ruggedness of the response surface is such that important energy fluctuations depending on geometry details are certain to occur even within this minimum energy well. As a consequence, many structures that may be regarded as 'correct' according to geometric criteria may nevertheless display high energies and fail to rank among the populated states. In other words, the discovery of the lowest point of the rugged energy well harboring the populated geometries is far from being a trivial problem and may require important intensification efforts. A specific setup scheme for the $\mathcal{GA}$, for fine exploration of limited phase space zones has been designed. It does not start with a random set of chromosomes, but from

previously sampled geometries representing a same global fold, in search for states of similar overall geometry but lower energy. Obviously, intensification runs compete for ressources with the default exploratory runs.

C. Tabu zones

Heavily visited phase space zones where it is 'believed' (see details below) that the deepest local optimum within the zone has already been sampled should be declared tabu areas. This amounts to (i.) eliminating the concerned chromosomes from the pool of 'ancestors' used for intensification and (ii.) defining an exclusion zone around each such chromosome. Any solution close, according to a to-be-defined similarity metric and similarity cut-off, to any tabu chromosome, and of higher energy than the tabu chromosome, will be assigned an abnormally low fitness score in order to force its demise at the next Darwinian selection step. If the new solution is fitter than the tabu chromosome, it will replace the latter. The choice of the similarity metric and cut-off is paramount: a too small cut-off discards only almost-identical pairs of solutions and unnecessarily spare redundant ones. On the opposite, too broad taboo areas may 'block' the access to unexplored deeper local minima in the neighborhood. In the present work we used a weighted block distance score in torsion angle space as a similarity metric of the two torsion angle vectors $\overrightarrow{\Theta}, \overrightarrow{\Theta}^{\text{tabu}}$:

$$\text{DISSIM}(\overrightarrow{\Theta}, \overrightarrow{\Theta}^{\text{tabu}}) = \sum_{i=1}^{N} w_i \times \Delta(\Theta_i, \Theta_i^{tabu}) \qquad (3)$$

where w_i is a weighting factor depending on fragment sizes, in order to tolerate larger variations with respect to terminal torsions, and Δ is the minimal positive rotation angle required to move from one torsional state to the other (e.g. 2 degrees to go from $\Theta_1 = 1$ degree to $\Theta_1^{tabu} = 359$ degrees, for example). Both the way in which torsional weighting factors are calculated with respect to the moving fragment sizes ($w_i = 0$ if fragment size $< MIN_{\text{FRAGSIZE}}$; $w_i = 1$ above MAX_{FRAGSIZE}; linear interpolation between these extremes) and the imposed tabu cut-off MIN_{DISSIM} are key control factors of the shape of the 'ellipsoidal' tabu zone around the tabu chromosome — several working hypotheses have been explored. In particular, all conformers differing only in terms of degrees of freedom associated to terminal fragments of MIN_{FRAGSIZE} and less become tabu.

As soon as regular diversification runs led to the discovery of a tunable minimal number of related geometries (regrouped according to a clustering procedure in torsional space, based on a chromosome dissimilarity score related to equation 3), the next planet will be dedicated to intensification within the phase space zone they populate. The key challenge of an optimal panspermia strategy is to decide at which point a cluster used as attractor in intensification searches has been sufficiently well sampled, in order to declare tabu the area around its cluster 'head' (its representative, most stable of its members). A too early decision in this sense may prematurely block the discovery of deep energy wells, while a too late one will translate in wasted computational time, at a scale proportional to the total number of independent solution clusters (of the order of $10^5 \ldots 10^6$ for a mini-protein like 1LE1 or 1L2Y). Common sense might suggest that intensification should be applied only to clusters of reasonably low energies, but in reality the ruggedness of the energy landscape is such that the energies of the first 'raw' conformers found by the diversification simulations that discovered the new clusters are completely uncorrelated with the final energies of fine-tuned geometries found by intensification in the immediate neighborhood. Restricting intensification to 'promising' solution clusters only is thus risky. The number N_{intens} of maximally tolerated intensification attempts of a cluster (set to 5, by default) is thus a key parameter of the panspermia strategy. Furthermore, the considered clusters are dynamic entities: when the newly added member is more stable than the current cluster head, it will replace the latter and recenter the cluster around the new head. Steadily evolving clusters will not become tabu — the number of maximally tolerated intensification attempts only applies if the cluster head remained unchallenged by the results of these biased searches (details not shown).

V. Results, discussion, prospect

Up-to-date attempts to use the planetary model led to successful folding experiments of the Tryptophane cage (α-helix) and Tryptophane zipper (β-sheet), as well as of key β-sheets and loops of the PIN1 WW domain in a matter of few days, using only a small subset (20-30 nodes) of GRID5000. Simulation results for the two first benchmark molecules will be discussed here.

The tryptophane cage contains an alpha-helical moiety stacked against an extended sequence to which it connects through a loop formed by 4 aminoacids (73 degrees of freedom, including both torsional axes of the protein backbone — except for the rigid peptidic bonds — and sidechains). α-helices are structural elements that fold quickly in solution, being stabilized by local, energetically favorable hydrogen bonds involving a residue and its 3rd successive neighbor. 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 (and perhaps benefit from refinement by "Lamarckian" gradient optimization). Being stabilized by internal hydrogen bonds, this structural element may readily be inherited by the successors until a favorable cross-over may couple two spontaneously emerged helix loops together. Accordingly, the planetary model has successfully and reproducibly discovered geometries as shown in figure 2 that are very close to the native 1L2Y fold reported in literature (white — native geometry; red — typical folded structure). Furthermore, the most stable of all sampled conformers was systematically found to be one of the correctly folded structures.

By contrast, although the tryptophane zipper consists only 53 degrees of freedom, it is nevertheless more difficult to

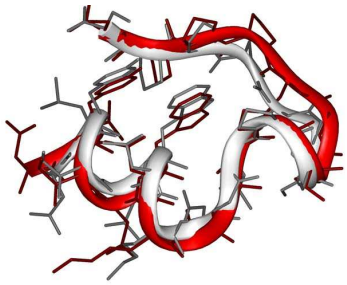

Fig. 2. Native state of 1L2Y, ranked as first among output conformers

fold computationally than 1L2Y. The main reason is the β-hairpin structure it adopts, 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, β-hairpin structures (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 (figure 3), with the native geometry shown in white. This calculated geometry was also shown to be the most stable of all the ever visited 1LE1 conformers.

Typical simulations, however, will return geometries like in figure 4, where the backbone is correctly folded but sidechains are misplaced (are predicted to interact differently with each other). Furthermore, the alternative side chain interactions proposed by the model do make physico-chemical sense: they are aromatic stacking interactions of a same nature as the one seen in the native geometries. The differences between the two structures are subtle, the second is not obviously wrong and it may actually correspond to some less populated species which does exist in solution but escapes detection by state-of-the-art experimental methods. However, the energy of such a conformer is significantly higher than the one of the native state and, unfortunately, also higher than the one of misfolded structures like in figure 5. In that simulation, the almost correct fold 4 was ranked as 79^{th} most stable geometry out of several hundreds of thousands. If the geometry of 1LE1 would not have been known, this simulation would have erroneously predicted the misfolded geometry 5 instead of the almost correct fold 4.

Evolving the latter into the properly folded 3 may require a quite lengthy intensification simulation. An exhaustive search for an optimal 'panspermia' approach (guaranteeing the reproducible discovery of a 'native' geometry at the lowest energy level among the sampled conformers) does however not appear to be feasible: it would require the tuning of at least four parameters (N_{intens}, MIN_{FRAGSIZE}, MAX_{FRAGSIZE} and MIN_{DISSIM}, not mentioning the ones controlling cluster definition). Multiple simulations (of $20\ldots50$ hours each $\times 20\ldots30$ nodes or more for problems larger than 1LE1 or 1L2Y) would be required for due assessment of the reproducibility at each parameter combination. The termination criteria of the method should also be subject to scrutiny would more important simulation efforts ensure the desired reproducibility? If so, which parameter should be first increased: the number of allocated planets or the total physical time? The obtained results show that reproducibility is not solely a matter of allocated ressources: note that the correctly folded 3 differs from the almost correctly folded 4 only by the placement of some low-weight side chains. Depending on the choice of MIN_{FRAGSIZE} and MAX_{FRAGSIZE}, the weighting factors from equation 3 may be such that the correct fold 3 actually falls within the tabu zone instated after the discovery of a structure like 4. If so, it will never be found, no matter for how long time the simulation continues. Renouncing the tabu strategy altogether is not an option, however: the simulations showed — and it makes perfect physical sense — that stable misfolded geometries, representing broader local optima than the native state, are reproducibly the first to be visited during the simulation. This would therefore systematically return to these same attraction pools each time a new run is started, unless tabu zones are declared. The native state owns its stability to more favorable intramolecular contacts. Or, a more compact packing of the protein chain is needed to enable more favorable contacts. This also means that any misplaced terminal fragment is likely to cause heavily penalizing intermolecular clashes, whereas in unfolded geometries side chains are free to move around in solvent. Protein folding amounts to an 'all-or-nothing' situation: the most stable states are achieved if either all degrees of freedom adopt their native values, or none of them do (i.e. all adopt random coil values corresponding to an unstructured peptide chain in solution). Situations in which most of the degrees of freedom are properly set, but a few of them are not, are likely to correspond to highly unfavorable energies due to clashes. The native state is a narrow but deep local minimum surrounded by an 'activation energy' barrier. As mentioned before, 1LE1 expectedly displays a much more marked 'all-or-nothing' behavior intrinsic to β-sheet folds. Therefore, optimal setup of the panspermia strategy is problem-dependent.

An alternative way to address the conformational problem is currently being considered: a thorough search of the maximal phase space volume that may be reproducibly sampled by local intensification procedures will be conducted, using diverse randomly picked phase space zones of different compounds. Phase space will be then divided into cells, optimally defined according to this study, and the overall conformational search will be conducted in this "discretized" problem space, where the fitness score of each phase space cell will be given by the free energy score returned by the local intensification simulation. In a broader perspective, novel deployment strategies using the PARADISEO[2] [18]

[2]http://paradiseo.gforge.inria.fr

core library for genetic algorithm deployment on the GRID will also be explored and compared to the planetary strategy, in search of a procedure optimally exploiting the potential of GRID5000 for solving molecular modeling problems.

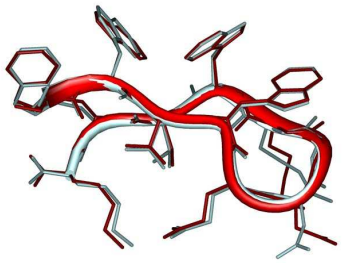

Fig. 3. Almost correctly folded geometry with correctly folded main chain but misplaced side chains, ranked only 79th in terms of stability

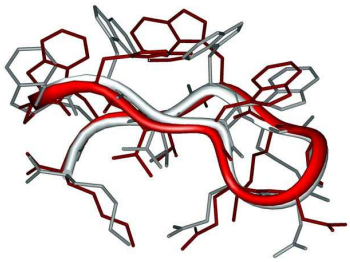

Fig. 4. The almost correct geometry is found among more stable misfolds.

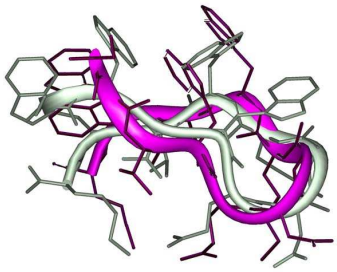

Fig. 5. Top ranked misfolded geometry.

REFERENCES

[1] J. N. Onuchic and P. G. Wolynes, "Theory of protein folding," *Current Opinion in Structural Biology*, vol. 14, pp. 70–75, 2004.

[2] P. Crescenzi, D. Goldman, C. H. Papadimitriou, A. Piccolboni, and M. Yannakakis, "On the complexity of protein folding," *Journal of Computational Biology*, vol. 5, no. 3, pp. 423–466, 1998.

[3] R. Unger and J. Moult, "Genetic algorithms for protein folding simulations," *Journal of Molecular Biology*, vol. 231, no. 1, pp. 75–81, may 1993.

[4] C. Levinthal, "How to fold graciously," in *Mossbauer Spectroscopy in Biological Systems*. University of Illinois Press: Proceedings of a meeting held at Allerton House, Monticello, Illinois, 1969, pp. 22–24.

[5] D. J. Wales and T. V. Bogdan, "Potential energy and free energy landscapes," *J. Phys. Chem.*, vol. 110, no. 42, pp. 20 765—20 776, 2006.

[6] A. T. Hagler, E. Huler, and S. Lifson, "Energy functions for peptides and proteins. i. derivation of a consistent force field including the hydrogen bond from amide crystals," *Journal of American Chemical Society*, vol. 96, no. 17, pp. 5319–5327, aug 1974.

[7] A. T. Hagler and S. Lifson, "Energy functions for peptides and proteins. ii. the amide hydrogen bond and calculation of amide crystal properties," *Journal of American Chemical Society*, vol. 96, no. 17, pp. 5327–5335, aug 1974.

[8] D. Horvath, "A virtual screening approach applied to the search for trypanothione reductase inhibitors," *Journal of Medicinal Chemistry*, vol. 40, no. 15, pp. 2412–2423, 1997.

[9] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, University of Michigan Press, 1975.

[10] B. Parent, A. Kökösy, and D. Horvath, "Optimized evolutionnary strategies in conformational sampling," *Journal of Soft Computing*, vol. 11, no. 1, jan 2007.

[11] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, "Designing a 20-residue protein," *Nature Structural Biology*, vol. 9, pp. 452–430, apr 2002.

[12] A. G. Cochran, N. J. Skelton, and M. A. Starovasnik, "Tryptophan zippers: Stable, monomeric β-hairpins," *Proc Natl Acad Sci USA*, vol. 98, no. 10, pp. 5578–5583, may 2001.

[13] H. Nguyen, M. J. M, J. Kelly, and M. Gruebele, "Engineering a beta-sheet protein toward the folding speed limit," *The Journal of Physical Chemistry B Condens Matter Mater Surf Interfaces Biophys.*, vol. 109, no. 32, pp. 15 182–15 186, aug 2005.

[14] K. Vertanen, "Genetic adventures in parallel: Towards a good island model under pvm," *Oregon State University*, 1998.

[15] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, "Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function," *Journal of Computational Chemistry*, vol. 19, no. 14, pp. 1639–1662, jun 1998.

[16] Accelery, "Accelerys discover simulation package." [Online]. Available: http://www.accelerys.com/insight/discover.html

[17] F. Glover, J. P. Kelly, and M. Laguna, "Genetic algorithms and tabu search: hybrids for optimization," *Computers and Operations Research*, vol. 22, no. 1, pp. 111–134, 1995.

[18] 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, no. 3, pp. 357—380, 2004.