META 2008
International Conference on Metaheuristics and Nature Inspired Computing
October, 29th -31st 2008
Hammamet, Tunisia

Special session on 'Metaheuristics and Structural Biology'

Scientific organisation of this session:	Carsten Baldauf (PICB, Shanghai), Daniel Merkle (University of Southern Denmark, Odense).
Organising committee of the conference:	El Ghazali. Talbi (Conference Chair, INRIA-Lille France), Khaled Mellouli (Co-chair, ISG Tunisie) and their teams.
Edition of this session proceedings:	Laetitia Jourdan (INRIA-Lille France), Sylvaine Roy (CEA – DSV/iRTSV/ CMBA, Grenoble, France).

In the last decades the field of metaheuristics has grown considerably. From the technical point of view as well as from the application-oriented side, these optimization tools have established their value in a remarkable success story. Researchers have demonstrated the ability of these methods to solve hard combinatorial optimization problems of practical sizes within reasonable computational time.

These proceedings contain abstracts of the talks presented at the 'Metaheuristics and Structural Biology' Special Session of the META'08 conference held in October 29th-31th 2008, in Hammamet, Tunisia. These articles are original research works that describe recent advance in the use of metaheuristics to solve problems in structural biology.

Later, for the overall conference, a selection of papers will be published in International Journals such as International Journal of Innovative Computing and Applications, Journal of Mathematical Modelling and Algorithms (JMMA), International Journal of Intelligent Computing and Cybernetics, International Journal of Operations and Quantitative Management,…

META 2008: International Conference on Metaheuristics and Nature Inspired Computing
October, 29th -31st 2008
Hammamet, Tunisia

Special session on 'Metaheuristics and Structural Biology'

organized by Carsten Baldauf (PICB, Shanghai) *and* Daniel Merkle (University of Southern Denmark, Odense) carsten@picb.ac.cn, daniel@imada.sdu.dk

Almost all predictive methods in structural biology induce difficult optimization problems, protein folding and molecular docking being two very prominent examples. Due to the complex nature of these problems, often having several objectives, being inherently dynamic, and having highly irregular fitness landscapes, metaheuristics are a valuable approach to achieve satisfying solutions in an efficient way. From an industrial point of view the importance of this research area is obvious. In the area of life sciences, the approaches satisfy the request for a deeper understanding of the underlying biological processes. From a metaheuristic point of view, it is evident that a thorough analysis of the fitness landscapes and a scientifically founded decision for the appropriate metaheuristic variants is necessary to gain such results.

This special session shows the importance of communication between the communities from computer and life sciences involved in these topics. Teams composed of researchers from interdisciplinary communities have presented their recent research results. Studied biological topics were various and have specially included Molecular docking (Meier *et al.*, Horvath *et al.*, Boisson *et al.*), Protein-protein interactions (Martin and Cornuejols), Structure prediction (Fonseca *et al.*), Conformational sampling (Tantar *et al.*, Horvath *et al.*), Genome rearrangement problems and phylogenetic inference (Lenne *et al.*), DNA renaturalization process for DNA computing (Banos *et al.*).

List of accepted talks:

- Dragos Horvath, El-Ghazali Talbi and Sylvaine Roy. *Force-field-based conformational sampling of proteins within the Docking@GRID project: status, results, issues.*
- Christine Martin and Antoine Cornuéjols. *Relevant features mining on protein-protein interfaces.*
- Renaud Lenne, Solnon Christine, Thomas Stuetzle and Eric Tannier. *Advances on Stochastic Local Search Algorithms for the Genomic Median Problem.*
- Alexandru Tantar, Nouredine Melab and El-Ghazali Talbi. *Locality on Protein-Ligand Docking Optimization.*
- Rasmus Fonseca, Martin Paluszewski and Pawel Winter. *Protein Structure Prediction Using Bee Colony Optimization Metaheuristic.*
- Jean-Charles Boisson, Laetitia Jourdan, El-Ghazali Talbi and Dragos Horvath. *A new tri-objective model for the flexible docking problem.*
- Rafael Baños, Paula Cordero, Angel Goñi and Juan Castellano. *Simulation of a DNA Renaturalization process.*
- René Meier, Carsten Baldauf and Daniel Merkle. *A Modular Framework for the Evaluation of Population-Based Algorithms for Molecular Docking.*

About the Program Chairs of this Special Session :

Daniel Merckle is an associated professor of the Department of Mathematics & Computer Science, at the University of Southern Denmark.

He has been working on Bio-inspired methods of optimization and their applications in Bioinformatics, especially Molecular docking since many years. He is the author of more than 50 scientific publications in international journals, conferences or scientific books to his name. He also brings his competences as the organizer of many European or worldwide conferences and as a referee in more than 20 international journals. With Carsten Baldauf, he is developing a Parallel Docking suite (ParaDocks).

Carsten Baldauf is a trained biochemist and a Feodor-Lynen Fellow of the Alexander-von-Humboldt Foundation at the CAS-MPG Partner Institute for Computational Biology in Shanghai.

Currently, his research deals with the atomic basis of shear-regulation in the blood clotting system. His fields of interest are foldamer research, protein-protein interactions and the development and application of algorithms for structural biology. Together with René Meier (University of Halle-Wittenberg) he is developing PARADOCKS, the Parallel Docking Suite.

Force-field-based conformational sampling of proteins within the Docking@GRID project: status, results, issues

D. Horvath[1], A.-A. Tantar[2], J.C. Boisson[2], N. Melab[2], L. Brillet[3], S. Roy[3], E.-G. Talbi[2]

1.Laboratoire d'Infochimie, UMR 7177 CNRS – Univ. Louis Pasteur, 4, rue Blaise Pascal, 67000 Strasbourg, France ; horvath@chimie.u-strasbg.fr
2. LIFL INRIA,Bât M3,Cité Scientifique,59655 Villeneuve d'Ascq, France
3.CEA, DSV, iRTSV, CMBA, 17 rue des Martyrs, Grenoble, F-38054, France

Keywords : multimodal optimization, molecular simulations, conformational sampling, large-scale distributed genetic algorithms.

1 Introduction

Computer simulations[1, 2] are nowadays a mainstream tool for predicting the properties – including biological activities – of molecules. The ability to compute an estimated value of a molecular property on the basis of its structure (and, perhaps, the one of the biological receptor it is supposed to interact) is of great potential benefit for chemical and pharmaceutical research. *In Silico* prioritization of molecules to be singled out for synthesis and test out of large compound databases is termed "virtual screening". Structure-based virtual screening or "Docking"[3] notably uses a model of the biological receptor to be inhibited by small organic molecules ("ligands") in order to computationally assess the affinity for the latter, by placing them into the active site in a way maximizing favorable site-ligand interactions. Unlike most of the recently reported applications[4-6] of large-scale parallel computing in molecular simulations and drug design – mainly aiming to process large compound libraries in deploying existing docking software on the GRID – the Docking@Grid[7] project advocates the use of massively parallel computational resources not in order to increase virtual screening throughput, but in order to improve the quality of the docking simulations *per se*. Docking is nothing else than a conformational sampling problem involving two different molecules, and should rely on a thorough exploration of all the relevant (low energy) zones of the problem phase space, defined by the – potentially large ($>10^2$) – number of intra-(flexibility-related) and intermolecular degrees of freedom. State-of-the-art docking software notoriously undersample the actual phase space and then need to rely on machine-learned (fitted) empirical equations to (more or less accurately) predict affinities on the basis of the few visited stable poses. On the other hand, statistical physics provides a rigorous definition of binding free energies, however requiring an exhaustive sampling of the ternary receptor-ligand-solvent complex – something well beyond the reach of modern computing technology.

2 The Docking@Grid challenge – more rigorous sampling, less empirical fitting

The Docking@Grid initiative rises the question whether intensive, but "intelligent" conformational sampling on the Grid (*i.e.* wasting a minimum of time in irrelevant high energy phase space zones, while nevertheless avoiding trapping by local minima) may suffice in order to allow the reproducible estimation of some empirical binding free energy "indices" which directly correlate to experimental affinities.

On one hand, this needs powerful conformational strategies, well adapted for GRID deployment, using hybrid heuristics centered on an evolutionary approach[8, 9]. Unlike "screen-saver" strategies, breaking a complex simulation down to very many independent shorter runs, the herein considered deployment approaches favor (a) the cooperation of different optimization heuristics, each specifically in charge of specific problems encountered in molecular modeling, (b) self-adaptive fine-tuning of the evolutionary parameters of the main Darwinian algorithm in charge of phase space exploration and (c) a

continuous monitoring of incoming partial sampling results, to be used in directing further phase space exploring efforts. The “planetary” strategy[9], a GRID-specific generalization of evolutionary island models, has been used to generate millions of conformers of typical 20…30-aminoacid proteins used in folding studies (the most complex of which is the villin headpiece) and sugars (cyclodextrines) – including, in many cases, near-native geometries.

On the other hand, the sampling strategy needs to rely on a properly tuned molecular force field model. According to the consistent force field philosophy, force field parameter refitting might not only compensate for the perturbation induced by the addition of the solvent terms to the original vacuum CVFF force field, but also "smooth out" potential artifacts due to insufficient sampling. The reparameterization concerns both global weighing factors of specific contributions (the van der Waals repulsion weight, the weight of Coulomb contributions, solvation-specific weights, *etc*.) and atom-specific van der Waals parameters. Force field fitting features the following key steps:

1. Test set molecules are submitted to sampling, using the planetary model on the grid, on typically 20..30 nodes, for 24 to 48 hours. Typically, several hundreds of thousands of relatively stable, clash-free conformers are retrieved for each compound. Also, Monte Carlo simulations of the native structure are launched in order enhance exploration of the native neighborhood. This is required since the actual position of the native minimum is expected to fluctuate upon force field parameter change.

3. The joined sets of geometries issued from both extensive sampling and local explorations serve to calculate the "folding free energy index" for each molecule, at current force field setup.

4. The consistency criterion used for force field parameter assessment is to achieve negative free energy index values for all the test set molecules. If this is already the case, including the latest conformers obtained at step (1) – which means that the above-mentioned sampling runs have systematically found near-native geometries and ranked them as the most stable among all the existing conformers – the current force field must then be challenged by application to other sampling problems.

5. Otherwise, a genetic algorithm operating in force field parameter space is run, until discovering a set of parameters fulfilling the above-mentioned condition. Since a single “force field fitness” function evaluation may easily take hours (millions of conformers need to be processed for each compound), the Grid is again the tool of choice for conducting this optimization. The loop then needs to be restarted at step (1), now using this latest force field setup.

This is a long-haul endeavor – the current status of the effort will be presented. Furthermore, the training set will be steadily updated with new molecules - notably receptor-ligand complexes - and the generic consistency criterion of negative free energy scores should then be replaced by the request of obtaining calculated affinity scores which quantitatively correlate to actual experimental values.

At this point, the best-so-far force field parameter set has been used to address docking problems in active sites with flexible loops – problems of complexity well beyond the capacity of classical docking codes which – at best – include only side chain flexibility, or treat the protein site as rigid. For example, in the Caseine Kinase 2 (CK2), the loop (amino acids 117-124) at the border of the active sites changes geometry in response of the structure of the bound ligand – see Figure 1. Conformational sampling based on the “planetary” strategy [9, 10] was challenged to predict the structure of the CK2-emodin [11] complex (not part of the force field parameter fitting set), while considering both (a) the side chains of site aminoacids, (b) all degrees of freedom (back bone and side chains) of amino acids 115-123, (c) internal degrees of freedom of the ligand and (d) translational and rotational degrees of freedom of the ligand – a total of 100 degrees of freedom. After 48 hours, using 20 nodes of the GRID5000 cluster in Lille, the correct experimental geometry was found and ranked as the most stable one out of a pool of $\sim 10^5$ stored geometries, solely based on the force field energy as objective function and with no bias from any experimental input (Figure 2). Note that the same system converges very quickly to the experimental structure if the loop flexibility is ignored (only side chain flexibility being considered). Simulations of the same enzyme in presence of different ligands, and, in perspective, of other flexible enzymes are ongoing. While, in case of success, they will not bring any ultimate proof of the validity of the current force field parameter set, the first failure will, however, represent enough evidence to invalidate it, thus triggering a novel round of force field parameter fitting.

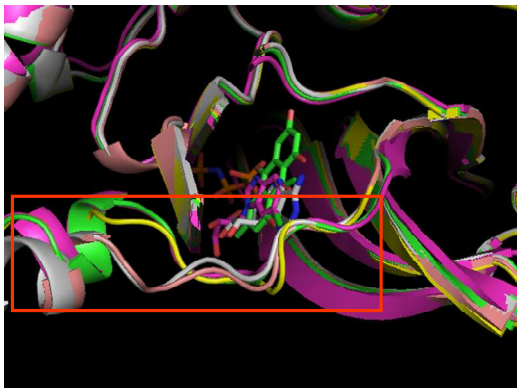

Figure 1: Different loop geometries in human CK2 cocrystallized with different ligands

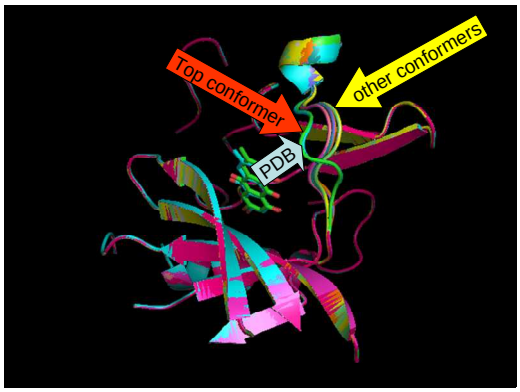

Figure 2 : The top sampled conformer of the CK2-emotilin complex coincides with the experimental structure, both in tems of the ligand pose and the loop geometry

References

[1] Jorgensen, W. L., (1991), Science, 254, 954-955.
[2] Neumaier, A., (1997), SIAM Review, 39, 407-460.
[3] Wang, C., Bradley, P., Baker, D., (2007), J. Mol. Biol., 373, 503-519.
[4] Pande, V., (2008), *Folding@Home*, http://folding.stanford.edu/
[5] Oxford, U. o., (2008), *Screensaver Lifesaver - searching for anti-cancer drugs by distributed computational chemistry*, http://www.chem.ox.ac.uk/curecancer.html
[6] Legrand, Y., Reichstadt, M., Jacq, F., Zimmermann, M., Maas, A., Sridhar, M., Vinod-Kusam, K., Schwichtenberg, H., Hofmann, M., Breton, V., Jacq, N., Salzemann, J., (2006) HealthGrid 2006, Valencia.
[7] Horvath, D., Tantar, A., Boisson, J.C., Melab, N., Roy, S., Brillet, L., Talbi, E.-G., (2008), *ANR Dock - Molecular Docking on Grids*, http://www2.lifl.fr/~talbi/docking/
[8] Parent, B., Kökösy, A., Horvath, D., (2007), Soft Computing, 11, 63-79.
[9] Parent, B., Tantar, A., Melab, N., Talbi, E.-G., Horvath, D., (2007) IEEE Congress on Evolutionary Computation, CEC 2007, Singapore.
[10] Tantar, A.-A., Conilleau, S., Parent, B., Melab, N., Brillet, L., Roy, S., Talbi, E.-G., Horvath, D., (2008), Current Computer-Aided Drug Design, 4, in press.
[11] Raaf, J., Klopffleisch, K., Issinger, O.G., Niefind, K., (2008), J. Mol. Biol., 14, 1-8.

Relevant features mining on protein-protein interfaces

C. Martin[1,2] and A. Cornuéjols[2]

[1] LIMSI CNRS,Bt 508,Université d'Orsay Paris Sud BP 133,91403 Orsay cedex
christine.martin@limsi.fr
[2] UMR AgroParisTech/INRA 518, AgroParisTech, 16 rue Claude Bernard, 75231 Paris cedex 05
antoine.cornuejols@agroparistech.fr

Keywords: protein-protein docking, frequent itemset mining.

1 Introduction

Many important biological processes involve protein-protein interactions. These correspond to contacts, also called dockings, on an interface area (from 900 A^2 to 2000 A^2). Numerous research efforts have been aimed at predicting these interactions, using either automatic or immersive methods. Neither of these two families of techniques is completely satisfactory however [1]. On one side, *automatic methods* are very costly in terms of computation because the large number of degrees of freedom in the representation of the phenomenon leads to a gigantic search space, even when using discretized attributes. Moreover, the current knowledge of the forces involved in these interactions is still scarce and uncertain and, as a result, evaluation functions used to guide the search process are ill-informed, leading to the exploration of a very rugged search landscape and thus yielding many false positive solutions. On the other side, *immersive methods*, that seek to take advantage of the expert's "intuition" about protein docking, specially with regards to geometric configurations, are limited because some feedback devices, like haptic ones, require real-time computations of forces that are difficult to obtain, while, at the same time, identifying the most useful display of additional information needed by the expert is still a matter of research in virtual reality environments.

This is why we have proposed a new hybrid approach (combining immersive and automatic context) named HOSMoS (Human Oriented Selection of Molecular Specimen) [1]. The idea is to rely as much as possible on the expert's knowledge in order to search and evaluate the possible docking situations while easing his/her task by providing predictive features that can be computed in real-time about tentative dockings. The problem is therefore to find relevant and easy to compute "abstractions" that help in the evaluation of the possible solutions. This paper presents a method for the identification of such features in the complex context of protein-protein interactions.

2 The identification of relevant features

The central problem we face in the study of protein-protein interactions is the identification of predictive features. One method, in this context, is to use supervised learning in order to select the features that allow learning algorithms to discriminate between positive and negative instances. However, in the case of protein-protein interactions, only positive instances are known, and in limited amount at that. Putative negative instances may possibly be generated, but their information content would be of very diverse and uncontrolled value. Another line of attack is then to look for conjunctions of descriptors (called *itemsets*) of which the rate of appearance in the known positive instances significantly differs from the rate one would expect given no information about the class. For instance, one could find that the conjunction of descriptors a_1 & a_6 & a_{23} is present in 20% of the positive instances, while one would expect it *a priori* in a proportion of only 1%.

This approach requires solving two problems. First, a dictionary of descriptors must be identified, which is both informative about the phenomenon at hand, but also limited enough so as to enable one to get statistically significant counts of conjunctions of these in the positive examples. Second, a technique must be found in order to compute *a priori* expected rates of itemsets, something also known as a "null hypothesis". Before describing our adaptation of the frequent item set method to our problem, we first present the data representation scheme we use.

3 Data selection and representation

Data about macromolecular structures like protein-protein complexes are getting increasingly available. However, quality and redundancy issues require one to be selective. In our case, 459 protein-protein complexes from the Dockground [3] database were retained. Biological considerations suggest that the contacts between the amino acids involved in the interfaces are a determining factor. We have therefore resorted to a model [4] based on a small set of geometrical patterns, namely, *edges*, *triangles* and *tetrahedra* (Figure 1) coding the geometry of a set of spheres, each one representing an amino acid. Edges represent the contact of two amino acids, triangles and tetrahedra the contact of three (resp. four) amino acids. Using properties of triangulation in a three dimensionnal space, these three geometrical items are sufficient to describe a set of spheres in a unique way.

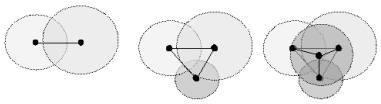

Fig. 1. Elementary objects of the used protein representation

Because there are twenty different amino acids in nature, the number of possible items (approx. 10'600) is too large compared to the number of available examples (459). Hence, we have decided to group the amino acids with respect to their physico-chemical properties. Taking five groups thus leads to a repertoire of 120 distinct descriptive items. Using this representation, one typical interface between proteins involves between 15 to 50 items, some of them possibly repeated.

4 Experiments and perspectives

It is then easy to compute the number of times each item appears in all positive instances. It remains to be decided which items (if any) are predictive of a potential docking. One difficulty is that classes of items (edges, triangles, tetrahedra) appear naturally with different frequencies (e.g. tetrahedra are less likely to appear than edges), requiring that specific decision thresholds be determined. This is a rather well-known problem in frequent item sets mining methods [5]. In our case, we decided to compute a base probability for each item, corresponding to the probability of the item to be part of an arbitrary interface (positive or not). The resulting analysis, still in progress, has, on one hand, filtered out items that were known to be good candidates for complementary roles in docking, and, on the other hand, pointed out unexpected patterns that could be useful indices of potential dockings. For instance, some tetrahedron comprising the same four groups of amino acids is more frequently found than triangles of the same groups!

It therefore appears that our approach for finding relevant and easy to compute description features can both help understanding the domain at hand and provide useful components for the definition of the evaluation functions required to guide the search in an otherwise gigantic and complex state space.

References

1. N. Ferey, J. Nelson, G. Bouyer, C. Martin, P. Bourdot, and J.-M. Burkhardt, User needs analysis to design a 3d multimodal protein-docking interface, In IEEE 3DUI 2008, 8-9th March, Reno, Nevada, USA, 2008.
2. P. Chakrabarti and J. Janin, Dissecting protein-protein recognition sites Proteins: Structure, Function, and Bioinformatics, 47 (3), 334-343, 2002.
3. D. Douguet, H.-C. Chen, A. Tovchigrechko and I. A. Vakser, DOCKGROUND resource for studying protein-protein interfaces, Bioinformatics, 22(21), 2612-2618, 2006.
4. F. Cazals, J. Giesen, M. Pauly and A. Zomorodian, Conformal Alpha Shapes, Eurographics Symposium on Point-Based Graphics, 2005.
5. B. Liu, W. Hsu and Y. Ma, Mining Association Rules with Multiple Minimum Supports, ACM SIGKDD, August 15-18, San Diego, CA, USA, 1999.

Advances on Stochastic Local Search Algorithms for the genomic median problem

Renaud Lenne[1,2], Christine Solnon[2], Thomas Stützle[1], and Eric Tannier[3]

[1] IRIDIA, CoDE, Université Libre de Bruxelles (ULB)
CP 194/6, Av. F. Roosevelt, 1050 Bruxelles, Belgium
{rlenne, stuetzle}@ulb.ac.be
[2] LIRIS, UMR CNRS 5205, Université de Lyon 1, Lyon, France
csolnon@liris.cnrs.fr
[3] INRIA Rhône-Alpes, LBBE, UMR CNRS 5558, Université de Lyon 1, France
eric.tannier@inria.fr

Abstract. The genomic median problem aims at finding the organization of the chromosome of a common ancestor to multiple living species. The problem is often formulated as searching for a genome that minimizes some distance measure among given genomes. In our previous research we have proposed a new stochastic local search algorithm for this problem that combined elements of tabu search, iterated local search, and a reactive search mechanism for adapting crucial parameters while solving an instance. In this talk, we will report on several improvements upon the initial algorithm and additional computational results including comparisons to current state-of-the-art algorithms for the genomic median problem.

1 Introduction

The genomic median problem (GMP) is an important problem arising in the context of genome rearrangement problems, and it is also of interest for phylogenetic inference [3, 4]. Genome rearrangements are evolutionary events that change the organization of genomes, for example, through fissions and fusions or translocations of segments in chromosomes. The GMP asks, given a number of genomes, to find a possible ancestor genome that can be constructed by the minimum number of rearrangements. For three genomes, it consists in searching for a fourth genome that minimizes the sum of the distances to three given genomes in terms of the number of rearrangements.

The problem we handle here, which is called cycle median problem in [6], is a GMP using the number of "double-cut-and-join" operations, as defined in [5, 2], as a distance measure. This problem is NP-hard. This is because the proof of NP-hardness of the reversal median problem [6], which is a GMP using the number of reversals as a distance metric, also shows that for unichromosomal genomes the problem of minimizing our distance formula is NP-hard.

Several algorithms have been designed and implemented for tackling the GMP. Exact algorithms have been applied to the special case where only one chromosome is available; they have been quite successful for rather small instances [6, 7]. Approximate algorithms ranging from heuristics [7, 8] to more complex local search algorithms [9, 1] have been proposed.

In our research, we have developed an algorithm called `MedITaS` (for Median solver by Iterated Tabu Search). It was targeted for tackling the multi-chromosomal case of the GMP; it is based on the "double-cut-and-join" operation described in [5, 2] and embeds an effective neighborhood search, initially proposed by Interian [1], into a tabu search (TS) algorithm. The TS algorithm is then further hybridized with an iterated local search (ILS) algorithm. Both main algorithm components, TS and ILS, use a reactive search mechanism to adjust at computation time the tabu tenure and the perturbation strength, respectively. More recently, `MedITaS` has been extended by using a new neighbourhood structure as well as a new algorithm for calculating an upper-bound for the median distance. In addition, we have now adapted the algorithm to the uni-chromosomal case, which resulted in a somewhat simpler algorithm.

2 Results

We now give preliminary results of a comparison between our `MedITaS` algorithm and the most recent algorithm that has been added to `GRAPPA`; this algorithm has been proposed by Arndt and

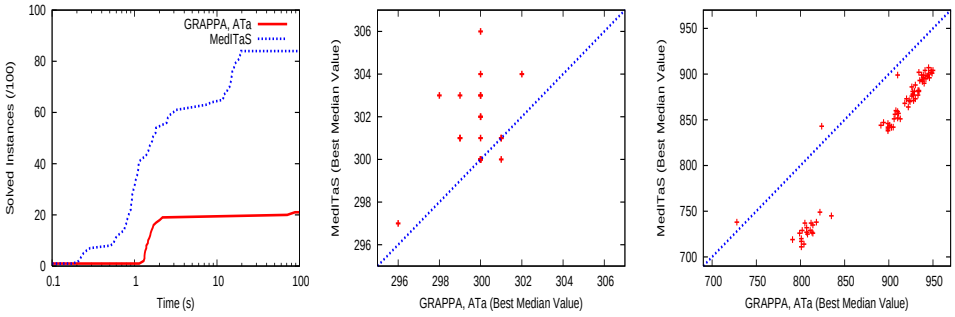

Fig. 1. Comparison between `MedITaS` and `ATa` (implemented in `GRAPPA` based on hardness distribution (left) and median values on $\rho = 20$ (middle) and $\rho >= 50$ (right) instances.

Tang (`ATa`) [8] and it is currently the most performing algorithm for the uni-chromosomal case of the GMP from a solution quality point of view. The computational results shown in Fig. 1 were measured across 100 random instances of a same size (100 markers) that differ in the ratio ρ/n, where ρ is the number of random reversals applied to the identity genome, and n is the number of markers (for each value of $\rho \in \{20, 50, 70, 80, 90\}$ we have generated 20 random instances). The left plot of Fig. 1 gives the cumulative distribution of the number of solved instances, that is, when a given algorithm reaches the best solution found between the two tested algorithms. It shows that `MedITaS` is clearly superior to `ATa`. Taking a closer look on instances with $\rho = 20$, `ATa` reaches slightly better solutions than `MedITaS` for the same computation time. (`MedITaS` is superior in the left plot of Fig. 1 since for some instance with large ρ/n ratio, it finds the best known solution very quickly.) However, on the instances with larger values of ρ, `MedITaS` reaches better quality solutions than `ATa`, in part by a rather large margin. For several instances (which are not shown here), `ATa` didn't return any answer within 24 hours of CPU time on a Dual-Core AMD Opteron2216 HE 2.4GHz and 4GB of RAM (only one core is used since the code is purely sequential), while `MedITaS` did return its best solutions for all instances in less than 30 seconds.

Acknowledgements. Renaud Lenne and Thomas Stützle acknowledge support from the F.R.S.-FNRS of which they are a FRIA fellow and a Research Associate, respectively. Eric Tannier is funded by the Agence Nationale pour la Recherche (ANR), projects REGLIS and GENOMICRO.

References

1. Interian, Y., Durrett, R.: Genomic midpoints: Computation and evolutionary implications (2007) Submitted.
2. Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. In: Proceedings of WABI 2006. Volume 4175 of LNBI. (2006) 163–173
3. Moret, B., Tang, J., Warnow, T.: Reconstructing phylogenies from gene content and gene-order data. In Gascuel, O., ed.: Mathematics of Evolution and Phylogeny. Oxford Univ. Press (2005) 321–352
4. Bernt, M., Merkle, D., Middendorf, M.: Using median sets for inferring phylogenetic trees. Bioinformatics **23** (2007) e129–e135
5. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics **21**(16) (2005) 3340–3346
6. Caprara, A.: The reversal median problem. INFORMS Journal on Computing **15** (2003) 93 – 113
7. Moret, B., Siepel, A., Tang, J., Liu, T.: Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In: Proceedings of the Second International Workshop on Algorithms in Bioinformatics. Volume 2452 of LNCS., Springer-Verlag (2002) 521–536
8. Arndt, W., Tang, J.: Improving inversion median computation using commuting reversals and cycle information. In: Comparative Genomics. Volume 4751 of LNCS., Springer Verlag (2007) 30–44
9. Bourque, G., Pevzner, P.: Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Research **12(1)** (2002) 26–36

A Study on Diversification Operators for Conformational Sampling on Computational Grids

Alexandru-Adrian Tantar, Nouredine Melab, El-Ghazali Talbi
LIFL/CNRS UMR 8022, DOLPHIN Project Team - INRIA Lille - Nord Europe,
Cité Scientifique, Villeneuve d'Ascq Cedex - 59655, France
E-mail: {Alexandru-Adrian.Tantar, Nouredine.Melab, El-Ghazali.Talbi}@inria.fr

With the continuous evolution of distributed high-performance and high-throughput computing and with the support of crystallography and nuclear magnetic resonance data, we are at the dawns of a new era in molecular research and pharmaceutical drug design. A focus is set by current research on protein structure prediction (PSP), molecular folding and molecular docking - the former area of study is here considered, as a particular case of conformational sampling under an evolutionary approach. The PSP problem consists in determining the ground-state conformation of a specified protein, given its amino-acids sequence - the *primary structure*.

As the diversification characteristics of evolutionary algorithms are significantly determined by mutation operators, important information can be derived by analyzing the behavior of EAs as influenced by these components. Classical operators are included, as the Swap, Mutation, Gaussian and the Cauchy ones, a new Pearson-system distribution based mutation being introduced. Further, annealing variance control schemes are employed as to induce a dynamic behavior along the exploration process. An overall of 30 operators is hence studied, employing different annealing schemes and multiple Pearson-system derived distributions. While a larger set of benchmarks has been considered for this study, only resulting conclusions are here presented, having the **Tryptophan-Cagge (PDB ID: 1L2Y)** protein as benchmark. No analogous large scale analysis and experimentation has been previously carried out in this area of study, although similar works exist on artificial academic problems. Namely, refer to the surveys of F. Herrera *et al.* [1], [2].

A first issue consists in identifying the operators which comport the best minimization characteristics. A second question to answer relates to operators which induce a bias in the exploration process. Statistical selection procedures are employed in order to address these matters, with an aim in identifying the best operators, out of a finite set of alternatives [3]. In this context, *best* is inferred in terms of *minimum mean output* for a series of independent samplings of the compared operators. A parallel construction of the OCBA procedure, initially developed by Chun-Hung Chen [4], [5] has been adopted for sustaining the operators analysis.

Evolutionary Algorithm Setup. A basic EA is used as embedding environment for testing the mutation operators. The algorithm is set to evolve an initial population of 150 solutions for 150 generations. At each iteration of the algorithm, the current population undergoes a stochastic tournament selection process, 150 individuals being chosen out of the population. At each selection step, the stochastic tournament component is set to return, out of two (uniform) randomly chosen solutions, the best individual, with a probability of 0.75. The resulting solutions become further subject to mutation, no crossover operator being employed. The embedded mutation operator is applied with a probability of 0.1, resulting that, for a complete execution, 2250 mutations should occur, in average. Further, a generational replacement is used, the entire population being replaced by the individuals resulting out of the mutation phase. A weak elitism scheme ensures that the best individual in the population to be replaced survives the replacement phase.
Statistical Selection Procedure - OCBA. The parameters of the selection procedure can be regrouped as $\tau = (\tau_{init}, \tau_{adt}, \tau_{sys}, \tau_{conf})$ where τ_{init} defines the number of initial simulations to be performed, τ_{adt} represents the number of additional samplings allocated for τ_{sys} heuristically selected systems. The desired confidence level to be attained is specified by the τ_{conf} factor. Additionally, a δ *indifference threshold* and a maximum number of iterations are specified. The *Probability of Correct Selection (PCS)* is estimated for $\delta = 0$ while. For the herein analysis, 30 initial simulations (τ_{init}) were executed for each of the resulting EAs, in case of non-convergence, 25 additional replications (τ_{adt}) being requested for the 5 most promising mutation operators (τ_{sys}), based on the results of the afferent EAs. In addition, a 1.0 indifference threshold (δ) has been set, *i.e.* no difference is

considered for two mutation operators which offer results comparable within a maximum absolute difference of 1.0. The confidence level (τ_{conf}) has been set to 0.1 - a ranking of the operators is considered significant if a PGS value equal or superior to 0.9 is obtained.

Additionally, in order to confirm and sustain the obtained operator rankings, a series of Wilcoxon rank sum non-parametric tests have been conducted. For each of the best three ranked mutation operators, a comparison considering the entire set of operators, except the one under testing, has been performed. Subsequently the determined rankings have been confirmed, the obtained results, for the operators under study, being significantly different as opposed to the rest of the operators. For all the conducted Wilcoxon tests a 0.99 confidence levels has been considered.

For exemplification purposes, a graphical depiction of the results is included in Fig. 1, only the **Tryptophan-Cagge (1L2Y) - Energy** case being here considered. Note that only the histograms corresponding for the best three ranked operators are included. Analyzing the complete set of results it follows that the Pearson types III, IV and VI based operators attain the best (energy) rank in most of the studied cases. Notwithstanding, none of these operators ranked in the first three mutations for the Root Mean Square Deviation (RMSD) section. Instead the Pearson types I, II and VII based operators were classed among the firsts in all the RMSD analysis tests. Consequently, no coherent inference can be made for accepting or rejecting a specific operator, as designated in the energy analysis section. Nevertheless, relying on the heuristic nature of the hybrid approach to be constructed, a straightforward design would consist in using a combined mutation operator. For example, the Pearson types III, IV and VI, no annealing scheme, set of operators may be employed.

In addition to relying on dynamic resolution schemes, adaptive approaches can be addressed. Note that the three mentioned operators, based on the Pearson types III, IV and VI distributions, result by modifying the mean, variance, skewness and kurtosis factors of a unique system. Thus, a polymorphic mutation would seem of interest, capable of standing as a substitute for the defined operators - the Gaussian and the Cauchy distributions can be seen as particular cases, depending on the specified parameters. Moreover, highly flexible exploration strategies can be designed, allowing, for example, for a smooth transition from a Pearson type III based mutation, in the early stages of the search, to a Pearson type VI derived operator for the final part of the exploration.

All experimentations were carried using an MPICH2 based version of ParadisEO, a framework dedicated to the reusable design of parallel hybrid meta-heuristics and which provides a broad range of features, including EAs support, local search methods, parallel and distributed models, hybridization mechanisms, etc. The analysis algorithms were executed on Grid5000[1], a French nation-wide experimental grid, connecting several sites and gathering more than 4000 processors with more than 100 TB of non-volatile storage capacity.

[1]https://www.grid5000.fr

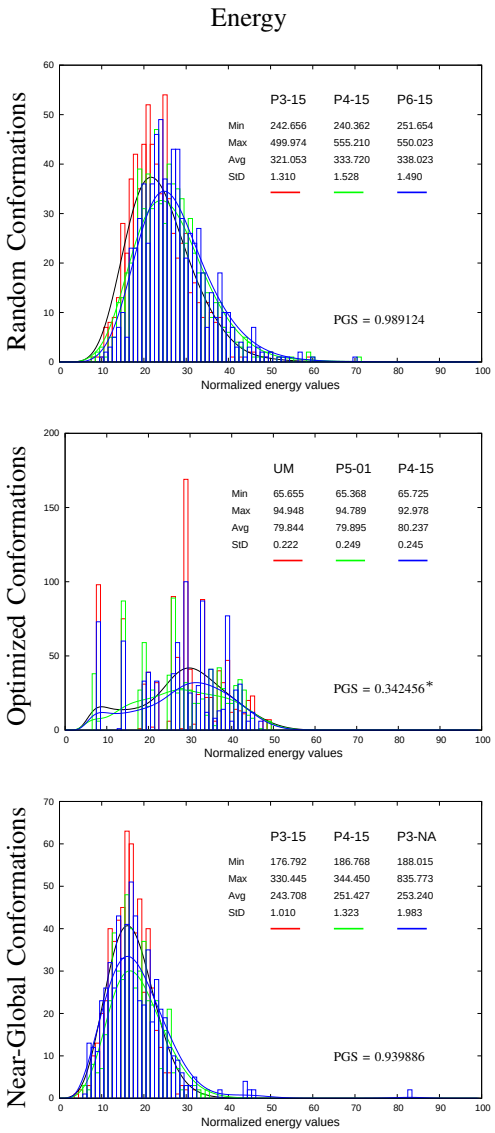

Fig. 1. The best three ranked operators are depicted with the corresponding PGS value.

REFERENCES

[1] F. Herrera, M. Lozano, and J. L. Verdegay. Fuzzy connectives based crossover operators to model genetic algorithms population diversity. Fuzzy Sets Syst., 92(1):21-30, 1997.

[2] F. Herrera, M. Lozano, and A. M. Sanchez. A taxonomy for the crossover operator for real-coded genetic algorithms: An experimental study. International Journal of Intelligent Systems, 18(3):309-338, 2003.

[3] Jürgen Branke, Stephen E. Chick, and Christian Schmidt. Selecting a Selection Procedure. Management Science, 53(12):1916-1932, 2007.

[4] C. H. Chen. A Lower Bound for the Correct Subset-Selection Probability and Its Application to Discrete Event System Simulations. IEEE Transactions on Automatic Control, 41(8):1227-1231, August 1996.

[5] J. Lin E. Yucesan Chen, C. H. and S. E. Chick. Simulation Budget Allocation for Further Enhancing the Efficiency of Ordinal Optimization. Journal of Discrete Event Dynamic Systems: Theory and Applications, 10:251-270, July 2000.

Protein Structure Prediction Using Bee Colony Optimization Metaheuristic: Extended Abstract

R. Fonseca[1], M. Paluszewski, and P. Winter[1]

Dept. of Computer Science, Uni. of Copenhagen (DIKU) Universitetsparken 1, 2100 Copenhagen Ø
{ hite, palu, pawel }@diku.dk

1 Introduction

Proteins are the primary building blocks in all living organisms. They are made of amino acid chains bound together by peptide bonds. Depending on the sequence of amino acids, the proteins fold in three dimensions so that the Gibbs free energy is minimized. The shape determines the function of the protein. *Protein structure prediction* (PSP) is the problem of predicting this three-dimensional structure from the amino acid sequence and is considered one of the most important open problems of theoretical molecular biology. The PSP has applications in medicine within areas like drug- and enzyme design.

The PSP proves to be a very difficult optimization problem. Solving it exactly is still far from realistic. Use of heuristics and less complex models proves to be an absolute necessity. However, even in simplified scenarios, many computational problems arise. One of these problems is the belief that free energy landscapes tend to have many local minima [1]

The *Bee Colony Optimization* (BCO) metaheuristic is a relatively new approach based on swarm-intelligence for solving complex optimization problems. It mimics the foraging behavior of honey-bees searching for nectar in a flower field. The algorithm, like real honey-bees, performs a wide search for good solutions and has a flexible method for allocating resources to intensify the local searches. This seems like a good strategy in the PSP to avoid getting stuck in the local minima of the energy landscape.

Hesham et al. [2] previously used the *Bees Algorithm* [3] to find the native state of the 5-residue peptide 'met-enkalphin' (PDB-ID: 1PLW) using a full resolution torsion angle-based representation. In our work, we apply the BCO metaheuristic to the PSP problem using a simplified representation and generate good quality solutions in terms of the RMSD similarity measure. These decoy solutions can be used as starting solutions for more advanced methods (protein structure refinement algorithms). Since we use a coarser representation, real-sized protein structures can be attacked by our BCO metaheuristic. To our knowledge this is the first time a bee heuristic has been used to predict the structure of proteins. We do not claim to solve the PSP or even compete with state-of-the-art PSP algorithms like Rosetta[4] or I-Tasser [5], however the BCO metaheuristic has nice properties that we believe makes it suitable for the PSP.

2 Model

Proteins usually consist of thousands of atoms, and their full description must contain the coordinates of all atoms. By considering the geometry of the backbone, this representation can be simplified to an average of 5 degrees of freedom per amino acid. However, even for small proteins, this conformational space is still very large and difficult to search. Here, we therefore apply predictions of secondary structure to reduce the degrees of freedom even further by regarding a protein as a sequence of connected segments.

3 Algorithm

In nature, a foraging bee can be said to be in one of three states: A scout bee, a worker bee or an onlooker. Scout bees fly around a flowerfield at random and when a flowerbed is found they return

[1] Partially supported by a grant from the Danish Research Counsil (51-00-0336)

to the hive and perform a waggle dance. The dance indicates the estimated amount of nectar, direction and distance to the flowerbed. Onlooker-bees present in the hive watch different waggle dances, choose one and fly to the selected flowerbeds to collect nectar. Worker bees act like scout bees except that when they have performed the waggle dance they return to their old flowerbed to retrieve more nectar.

In our adaptation of the BCO metaheuristic, each bee corresponds to a solution, and the nectar amount corresponds to an objective value in the energy landscape. Sending out scout bees corresponds to finding a random feasible solution and sending out onlookers corresponds to finding a neighborhood solution. The onlookers choose sites for neighborhood search based on the objective value of scouts and workers in previous iterations. This method is largely the *Bees Algorithm* proposed in [3]. In a non-changing solution space a solution does not deplete in the same way a real life flowerbed depletes of nectar. Exhaustion is therefore forced when a solution cannot be improved. This idea is somewhat similar to the idea of pruning parts of the searchspace as described in [6]. The process of exhausting a local search is proposed as part of the *Artificial Bee Colony* algorithm described in [7]. Our adaptation of the BCO metaheuristic is a synthesis of these approaches.

4 Dataset

To test our BCO metaheuristic, we try both simple and complex proteins with respect to both residue-length and the number of secondary structure segments. Six proteins are from [9] and all have less than 12 segments and from 54 to 76 residues. Six different proteins are chosen from CASP7 [10] which all have more than 76 residues and more than 12 segments.

5 Results and perspective

Simulated Annealing (SA) and Monte Carlo are often used in the PSP [8], so for comparison both BCO and SA were used to minimize the energy of the 12 selected proteins. Despite the fact that SA is so frequently used for the PSP, BCO outperforms SA by finding lower energy structures for the 6 smaller proteins. Partial results show promising predictions for the 6 larger ones as well.

BCO seems to differ from SA in its wide search and good prioritizing of local searches. Furthermore, the algorithm seems extremely flexible. The local search performed by onlookers and the random solutions found by scouts can be implemented using any of the well-known algorithms. SA, Monte Carlo or hill-climbing can for instance be used for local search and genetic algorithms for generating random solutions. Different strategies can even be combined.

References

1. Li, Z. and Scheraga, H. A. Monte carlo-minimization approach to the multiple-minima problem in protein folding. Proceedings of the NationalAcademy of Sciences 84(19):6611-6615,October, 1987.
2. Hesham Awadh Abdallah Bahamish, Rosni Abdullah, Rosalina Abdul Salam, "Protein Conformational Search Using Bees Algorithm," ams , pp. 911-916, 2008.
3. Pham DT, Ghanbarzadeh A, Koc E, Otri S, Rahim S and Zaidi M. The Bees Algorithm. Technical Note, Manufacturing Engineering Centre, Cardiff University, UK, 2005
4. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. Protein structure prediction using Rosetta. Methods in Enzymology 383:66-93, 2004.
5. Y Zhang: I-TASSER server for protein 3D structure prediction. BMC Bioinformatics **9**:40 (Jan 2008)
6. M. Paluszewski, T. Hamelryck, P. Winter. Reconstructing protein structure from solvent exposure using tabu search, Algorithms for Molecular Biology, 2006.
7. D. Karaboga, An Idea Based On Honey Bee Swarm for Numerical Optimization, Technical Report-TR06,Erciyes University, Engineering Faculty, Computer Engineering Department 2005.
8. G. Helles. A comparative study of the reported performance of ab initio protein structure prediction algorithms. J R Soc Interface. 2008 Apr 6;5(21):387-96.
9. Hamelryck, T., Kent, J. T., Krogh, A.: Sampling realistic protein conformations using local structural bias. PLoS Computational Biology **2**(9) (September 2006) e131
10. Critical Assessment of Techniques for Protein Structure Prediction, Asilomar Conference Center, Pacific Grove, CA November 26-30, 2006. http://predictioncenter.org/casp7/Casp7.html

A new tri-objective model for the flexible docking problem

J-C. Boisson[1], L. Jourdan[1], E-G. Talbi[1] et D. Horvath[2]

[1] Equipe Projet INRIA DOLPHIN, 40 avenue Halley, 59650 Villeneuve d'Ascq Cedex France
{Jean-Charles.Boisson, Laetitia.Jourdan, talbi}@lifl.fr
[2] Laboratoire de Glycobiologie Structurale et Fonctionnelle UMR 8576, Université des Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq Cedex, France
dragos.horvath@univ-lille1.fr

The prediction of the best binding mode between two proteins is a critical problem for drug design. Throw the ANR project Dock, we have worked on possible new multi-objective models for the flexible docking problem. So we propose a new tri-objective model combining an energy term, a surface term but also a robustness term. The aim of the surface term is to guide the ligand into the site. The energy term is used to gain complex of low energy. The robustness term is based on the hypothesis that the best binding mode is not sensible to small perturbations and thus the energetic landscape around it is large. The aim of our approach is to propose a more realistic docking process to gain high quality results.

1 Docking problem

For drug design, it is essential to find which molecules can interact with other bigger molecules. The docking problem consists in finding how a small molecule, the ligand, can interact with another one, the receptor which has one or more site for the ligand. Nevertheless, experimental docking studies cost time and resources. There generally exist more than one hundred thousand ligands and the sites are not necessary known. In this situation, automatic docking methods to screen large ligand databases allow to speed up drugs design. Since the 90's, metaheuristics have been used to solve the molecular docking problem. Originally, single solution metaheuristics as Metropolis Monte-Carlo algorithm or Simulated Annealing were used to solve this problem. Later, population based metaheuristics like Genetic Algorithms have been used [4]. Recently, new docking methods have been also proposed using Particle Swarm Optimization [3] or Ant Colony based metaheuristics [5].

2 Tri-objective model

2.1 Energy

The standard criterion to estimate the stability of a molecule is to compute its molecular energy. Smaller the energy is, stabler is the complex. Depending of the force field used and the considered energetic interactions, the computed energy is different. Equation 1 corresponds to our energy evaluation based on the **C**onsistent **V**alence **F**orce **F**ield (CVFF). This energy function has been already used for the **P**rotein **S**tructure **P**rediction problem [7].

$$E = \sum_{bonds} + \sum_{angles} + \sum_{torsions} + \sum_{Van\ der\ Waals} + \sum_{Coulomb} + \sum_{desolvation} \quad (1)$$

All the parameters of the CVFF have been tuned experimentally on a diverse set of molecules.

2.2 Solvent accessible surface

An atom can be represented as a sphere according to its Van der Waals radius. The solvent accessible surface is drawn according to the center of a probe that rolls on the atom spheres. Generally, the probe has a radius of 1.4Å[3] in order to be able to contains a water molecule that is one of the standard solvents. The original algorithm we used was first described in [2] but was recently used in [6]. It is based on the use of look-up tables and Boolean Logic. The solvent accessible surface allows to evaluate the penetration of the ligand into the site. This criterion is essential for simulating realistic flexible docking processes.

[3] $1\ angstrom = 0.1\ nanometer$

2.3 Robustness term

Our robustness term consists in the energy computation of the neighbourhood of a Ligand/Site Complex (LSC). According to a given number of neighbours ($nbNeigh$) with their molecular energy already computed using our first objective function, the robustness term is given by the equation 2:

$$CR = -\frac{1}{\beta} * ln \sum_{i=1}^{nbNeigh} e^{-\beta * Energy[i]} \; with \; \beta = \frac{1}{kT} \tag{2}$$

k is the Boltzmann constant and T the temperature. The definition of this equation maintains the Boltzmann formula. The aim of this objective is to estimate if a LSC is more likely than another one. A complex relatively flexible is in a large valley of potentials. A rigid complex is in a very narrow energy well. Although a complex with a very low energy is generally chosen; in reality, it can be possible that the right complex to choose has a higher energy. The robustness term helps finding this kind of complexes.

3 Method

Using the presented model, a parallel multi-objective genetic algorithm (GA) based on IBEA (*Indicator-Based Evolutionary Algorithm*) has been designed [1]. In this GA, an individual is coded as two vectors of euclidean coordinates: one for the ligand, the other for the receptor. The crossover used is a ligand swap between two receptors. Four mutations has been designed. Two are standard mutations for the rigid docking problem: the rotation and the translation of the ligand. The third adds the possibility to modify the conformation of the molecules (ligand and/or receptor). This mutation allows to make flexible docking. The flexibility of the molecules can be chosen by indicating which atom of each molecule are considered as fixed atoms. The last mutation, called the *reverse* mutation, allows to speed-up the algorithm by making great perturbations on an individual. It is useful to exit from local optima.

In order to test our approach, we have taken data from the CCDC-Astex dataset. This dataset contains a "clean" list of instances dedicated for the docking benchmarking. For each instance, we have extracted the ligand from it crystallographic location in order to obtain a "seed" ligand that will be used to initialise our population of solutions.

We have evaluated our results according to standard indicators of docking benchmark like the RMSD of the final ligand location comparing to its original crystallographic location or the quality of the Pareto front gained. As it is commonly admitted in the literature, a docking is considered as good if the corresponding final RMSD is in [0,2] Å. The majority of the tested instances have produced good docking results. We currently work on behaviour improvement of the GA (speed, operators) to gain good docking results on all the instances of the CCDC-Astex dataset.

References

1. S.Kunzli E. Zitzler. Indicator-based selection in multiobjective search. *Parallel Problem Solving from Nature, PPSN VIII*, 3242:832–842, 2004.
2. S.M. Le Grand and Jr. K.M. Merz. Rapid Approximation to Molecular Surface Area via the Use of Boolean Logic and Look-up Tables. *Journal of Computational Chemistry*, 14:349–352, 1993.
3. S. Janson, D. Merkle, and M. Middendorf. Molecular docking with a multi-objective Particle Swarm Optimization. *Applied Soft Computing*, 2007. doi:10.1016/j.asoc.2007.05.005.
4. G. Jones, P. Willet, and R.C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*, 245(1):43–53, 1995.
5. O. Kord, T. Stützle, and T.E. Exner. An Ant colony optimization approach to flexible protein-ligand docking. *Swarm Intelligence*, 2007. doi= 10.1007/s11721-007-0006-9.
6. A. Leaver-Fay, G.L. Butterfoss, J. Snoeyink, and B. Kuhlman. Maintaining solvent accessible surface area under rotamer substitution for protein design. *Journal of Computational Chemistry*, 28(8):1336–1341, 2007.
7. A-A. Tantar, N. Melab, E-G. Talbi, and B. Toursel. A Parallel Hybrid Genetic Algorithm for Protein Structure Prediction on the Computational Grid. *Elsevier Science, Future Generation Computer Systems*, 23(3):398–409, 2007.

Simulation of a DNA Renaturalization process

R.Baños[1], P. Cordero[2], A.Goñi[3] and J.Castellanos[4]

1. RAFAEL BAÑOS, Facultad de Iinformática, Polytechnic Uiversity of Madrid, Boadilla del Monte, 28660. Madrid, Spain.
2. PAULA CORDERO, Natural Computing Group, Polytechnic Uiversity of Madrid, Boadilla del Monte, 28660. Madrid, Spain. p.cordero@alumnos.upm.es
3.ANGEL GOÑI, Natural Computing Group, Polytechnic Uiversity of Madrid, Boadilla del Monte, 28660. Madrid, Spain. ago@alumnos.upm.es
3.JUAN CASTELLANOS, Artificial Intelligence Department, Facultad de Informática, Polytechnic Uiversity of Madrid, Boadilla del Monte, 28660. Madrid, Spain. jcastellanos@fi.upm.es

Abstract: In this paper is presented a simulation of a DNA renaturalization process with eukaryotic DNA molecules. This simulation is implemented in a computer and represents a useful tool for a virtual laboratory which is oriented to DNA computations. Starting in a theoretical study, in which the simplest parts of the nucleic acids (DNA and RNA) are explained, this simulation model try to give a simple view of the renaturalization process of eukaryotic DNA. The implementation proposed emphasizes the different durations of this process depending on the nature of the DNA molecules involved.

Computing using a DNA molecule is a modern approach to a massive parallel paradigm. Molecular computing consists of representing the information of the problem with organic molecules and to make them react within a test tube in order to solve a problem. The fundamental characteristics of this type of computations are, mainly, the massive parallelism of DNA strands and the Watson-Crick [Watson-Crick, 1953] complementarity. The speed of calculation, the small consumption of energy and the big amount of information which DNA strands are able to store are the best advantages that DNA computing has. Nevertheless one of the problems is the massive calculation space needed, which limits the size of the problems.

When a double stranded DNA molecule (native form) is warmed up, the unions between both strands will break down and, as a result, they will separate. Therefore, a denaturalized DNA molecule is always single stranded. The transition between the native form and the denaturalized form is known as denaturalization. The temperature of Melting (Tm) is necessary to carry out the process of denaturalization. In certain conditions, a monocatenary DNA solution (denaturalized) can return to the native DNA form. This process receives the name of renaturalization of DNA.

The kinetic energy of DNA renaturalization is expressed in the “Cot Equation”. This equation expresses the concentration of single stranded molecules “C” in a certain time “t”. This concentration is based on the initial DNA strands concentration “Co” and a second order constant speed “k”. This constant “k” can be detached based on parameters like: the average number of nucleotides inside a single stranded DNA molecule, the number of units of a non-repetitive sequence of base pairs of a haploid nucleus or a prokaryotic genome, the average density of sites of nucleation in a DNA fragment or a constant of proportionality.

(F1) Cot1/2 = 1/k

One of the applications of the Cot Equation is to calculate the complexity of a DNA molecule, from the observed value of Cot1/2 and comparing it with the DNA of an E. Coli molecule, whose complexity is well-known to be 4200000bp.

$$\textbf{(F3)} \qquad \frac{\textbf{Cot1/2 (DNA from any genome)}}{\textbf{Cot1/2 (E. Coli DNA)}} = \frac{\textbf{Complexity of any genome}}{\textbf{4200000 bp}}$$

The prokaryotic or bacterial DNA doesn't contain repeated sequences; the complexity of the nucleotides in this type of DNA is constant. However, the eukaryotic DNA has different classes of nucleotide sequence, which differ in their complexity and in their repetition rate in the DNA. The DNA's renaturalization speed of an organism is related to its complexity. The renaturalization speed of each family depends on its repetition frequency within the genome.

This work develops a program that simulates the renaturalization process of a series of eukaryotic DNA molecules. The simulation is made on a board of two dimensions; each position of the board can be occupied by a simple DNA chain or be empty. The different types of chains, that compose the total DNA, are also provided. The repetition frequency of the chain and percentage of the total DNA are indicated for each type of that chain. This percentage can be calculated measuring chemically the amount of this type of DNA and comparing it with "Co". Then, the percentage is transformed into a considered number of repetitions of a concrete chain.

Initially all the chains and their complementary ones are separated and arranged at random simulating a laboratory experiment. In proportion as the temperature descends, if two adjacent positions are occupied by a chain and its complementary one, it will be able to take place or not the renaturalization, depending on the simulation conditions in this moment. Three types of denatured chains can be observed initially. In red color some of the fast component chains and their respective complementary chains are identified. In green color the intermediate component chains, and finally in blue color the slow component chains.

One simulation cycle consists of going across every square of the board allowing one movement for each chain, which means moving to an adjacent square. Each cycle of simulation owns its own conditions of temperature and time. While the simulation is running the time increases and the temperature falls.

The main problems of the experiments carried out in-vitro based on the manipulation of DNA are the costs, the time and the space required. Because of that, it exists the need of making these experiments easier by simplify the main bio-operations and bio-molecular processes over DNA molecules. For that reason is so useful and important the possibility of simulating these kind of reactions in a virtual laboratory so that more difficult operations and algorithms based on DNA computations can be done without the difficulties explained above.

This simulation model tries to give a simple view of the renaturalization process of eukaryotic DNA. The implementation proposed emphasizes the different durations of this process depending on the nature of the DNA molecules involved. The use of this simulator can give a result of the reanaturalization process depending on the characteristics of the DNA molecules introduced in a faster way than in-vitro.

A Modular Framework for the Evaluation of Population-Based Algorithms for Molecular Docking

René Meier[1], Carsten Baldauf[2], and Daniel Merkle[3]

[1] Department of Pharmacy, Institute of Pharmaceutical Chemistry
Martin-Luther-University Halle-Wittenberg, Halle, Germany
rene.meier@pharmazie.uni-halle.de

[2] Protein Mechanics and Evolution Group,
CAS-MPG Partner Institute for Computational Biology,
Shanghai, China
carsten@picb.ac.cn

[3] Department of Mathematics and Computer Science
University of Southern Denmark, Odense, Denmark
daniel@imada.sdu.dk

1 Introduction

Molecular docking means to find the correct orientation and position of two molecules (pose) towards eachother. Usually one molecule is a protein (receptor), while the other molecule is a rather small, drug-like compound (ligand). Solving this problem is of great importance in the area of drug discovery. Virtual Screening [5] is the automated evaluation of very large libraries of compounds towards a receptor using computer programs. An immense amount of computation comes along with identifying favorable chemical structures, as i) even after thorough filtering the number of possible compounds is very large, ii) each of these possible compounds has to be analyzed regarding to the binding energy of the ligand and the receptor, and iii) even the computation of the binding free energy for only one conformation of ligand and receptor can be immense, depending on the the chemical model that is used for this computation. Based on these observations we developed a highly flexible and modular program that solves the given problem, using also the possibility of parallel computation. ParaDockS, the *Para*llel *Dock*ing *S*uite, is an open source docking framework featuring a modular design and therewith easy exchangeability of optimizers and fitness functions. Based on this framework we evaluate in this paper the performance of several population based metaheuristics on the molecular docking program.

2 ParaDockS Architecture

The highly modular architecture of ParaDockS can only be sketched here. Its replaceable key components are a fitness function describing the interactions between ligand and receptor and a metaheuristic predicting the lowest energy ligand-receptor pose. Note that a variety of approaches, ranging from additive potentials and grid-based approaches up to quantum chemical methods can be used as fitness function. Additionally, even more technical features like the possibility of parallelization are easily implementable and are in fact implemented. For achieving a high throughput and to support subsequent processing, the communication component of ParaDockS uses XML for input files as well as for output files.

3 Fitness Function

The basis for succesfull prediction of ligand-protein binding is a robust and reliable way to compute the binding free energy. The fitness function of ParaDockS used for this study is derived from X-Score[6], an algorithm to predict the binding affinity of protein ligand complexes. However X-Score itself can not be used as fitness function for molecular docking, because it lacks a description of the ligand conformation. Our newly derived fitness function is based on terms: $E = E_{PL} + E_{vdW} + E_{hb}$. Let $\mathcal{L}$ and $\mathcal{P}$ be the set of ligand and protein atoms. The protein ligand van der Waals interaction

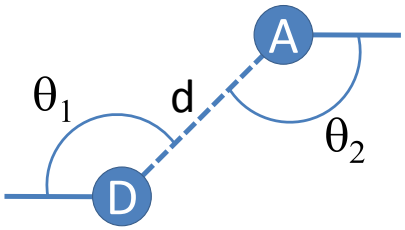

Fig. 1. Illustration of the three parameter used for the characterization of hydrogen bonds; A (acceptor) and D (donor) denote heavy atoms forming the hydrogen bond; d: distance between A and D; Θ_1 and Θ_2: angles characterizing the hydrogen bond

is reflected by $E_{PL} = \sum_{i\in\mathcal{P}}\sum_{j\in\mathcal{L}}\left[\left(\frac{d_{ij0}}{d_{ij}}\right)^8 - 2\left(\frac{d_{ij0}}{d_{ij}}\right)^4\right]$, the energy term for the internal ligand van der Waals clashes is computed by $E_{vdW} = \sum_{i\in\mathcal{L}}\sum_{j\in\mathcal{L}}\left[\left(\frac{d_{ij0}}{d_{ij}}\right)^8 - 2\left(\frac{d_{ij0}}{d_{ij}}\right)^4\right]$, and the protein ligand hydrogen bonding interaction is $E_{hb} = \sum f_1(d)\cdot f_2(\Theta_1)\cdot f_3(\Theta_2)$ (functions f_1, f_2, and f_3 express the deviation from an ideal hydrogen bond conformation, see Figure 1). The parameters of the fitness function were trained based on the well-known Astex-Diverse training set [2].

4 Optimizer

The analysis of conformation with computational approaches for molecular docking is usually performed with metaheuristics by minimizing the binding free energy for ligand and receptor. While in Autodock, a well-known suite of automated docking tools[1], a Lamarckian-GA (LGA) is known to performed best within a set of several metaheuristic approaches, recent research results indicate that metaheuristics that fit better to continuous optimization problems, perform very competitively and usually outperform the LGA [4]. In the ParaDockS framework we included a Particle Swarm Optimizer and a Differential Evolution approach for single objective optimization until now, but the framework can easily handle multiobjective approaches like suggested in [4].

5 Results

ParaDocks was tested on 210 non-redundant complexes and the predicted conformations were compared with the X-ray crystallography conformations based on their RMSD values. Regarding the docking predictions ParaDocks performs at least equally good as the well-known docking program GOLD [3].

References

1. D.S. Goodsell and A.J. Olson. *Automated docking of substrates to proteins by simulated annealing* Proteins: Structure, Function, and Genetics, 8(3):195–202, 1990
2. M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, and C. W. Murray. *Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance* J. Med. Chem. 50:726–741, 2007
3. G. Jones, P. Willett, R.C. Glen, A.R. Leach, and R. Taylor. *Development and validation of a genetic algorithm for flexible docking* J. Mol. Biol. 267:727–748, 1997
4. S. Janson, D. Merkle, and M. Middendorf. *Molecular Docking with Multi-Objective Particle Swarm Optimization* Applied Soft Computing, 8(1):666–675, 2008
5. W.P. Walters, M.T. Stahl, and M.A. Murcko. *Virtual screening an overview* Drug Discov. Today, 3(4):160–178, 1998. 2
6. R. Wang, L. Lai, and S. Wang. *Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction* J. Comput.-Aided Mol. Des., 16:11–26, 2002.