A Comparative Study of Different Optimization Algorithms for Molecular Docking

Alexander Afanasiev[1] , Igor Oferkin[3], Mikhail Posypkin[1,*] , Anton Rubtsov[1], Alexey Sulimov[3] and Vladimir Sulimov[2,3]

[1]Institute for Systems Analysis RAS, Address 117312, Moscow, pr. 60-letiya Oktyabrya, 9, mposypkin@mail.ru

[2]Research Computer Center at Lomonosov Moscow State University, 119992, Leninskie Gory 1, bldg 4, Moscow 11999 2, Moscow, Russia

[3]Dimonta, Ltd, 117186, Nagornaya Str. 15, build. 8, Moscow, Russia

ABSTRACT

Motivation: Computer modeling of protein-ligand interactions is one of the most important phases in a drug design process. The core part of this modeling is a resolution of a global unconstrained optimization problem. This paper presents a comparative computational experiments aimed at studying the efficiency of the different optimization methods applied to the docking problem. We present experimental results for different optimization algorithms and draw conclusions about their efficiency.

1 INTRODUCTION

Many diseases are caused by foreign protein activity or protein malfunction. For the treatment of these diseases we can try to block these proteins by small organic molecules. These molecules can selectively bind to proteins and thus block their work. This simplified conception allows development of the drugs, using purposeful design of new organic compounds – inhibitors for the given target-protein. The selection of the effective ligands for inhibition of the target enzyme is usually very laborious, long, and expensive process. Contemporary molecular modeling tools can accelerate this process and make it much less expensive. Virtual screening by means of ligands docking is widely recognized as a helpful approach in modern drug design (Kitchen 2004, Zoete 2009).

The goal of docking is to find the positions of interacting ligand and a protein with a minimal energy. The stability of this position is characterized by the energy value. The ligands with minimal values are candidates for further consideration as potential drugs. Thus the docking problem is reduced to the global optimization problem

$$f(x) \to min, x \in P \qquad (1)$$

where x is a tuple that determines the ligand position and P is a bounding box restricting the search area to a reasonable region.

[*]To whom correspondence should be addressed.

Having fast and robust optimization algorithms for solving the problem (1) is crucial for an efficient docking. In this paper we evaluated different optimization algorithms for resolution of the problem (1). We performed numerical experiments, analyzed the results and suggested the most successful combination of optimization algorithms for this problem.

2 RELATED WORK

The comparative efficiency of different optimization algorithms has been studied in various papers (Rosin 1997, Morris 1998, Tavares 2008, Tavares 2009). Authors consider global optimization approaches (genetic algorithms, simulated annealing), local search techniques (L-BFGS, Solis-Wets method) and their combination. The best results were obtained by a combined approach when a genetic algorithm is combined with a local search. In this approach a fraction of all individuals in a generation are further optimized by applying a local search technique.

Two local search techniques addressed in the literature showed the best results. The first method proposed in (Solis 1981) is a direct search method with an adaptive step size, which performs a randomized local minimization of a given candidate solution. Depending on whether a new solution if found or not, a success or a failure is recorded. If several successes occur in a row, the step is increased to move more quickly. If the opposite occurs, the step is decreased. A bias term is applied to drive the search in successful directions. The method terminates when a certain lower-bound threshold for is passed or when a maximum number of steps is reached. The second method is the Broyden-Fletcher-Goldfarb-Shanno method (Nocedal 2006). L-BFGS is a quasi-Newton method, where both the function to minimize and its gradient must be supplied by the user. The method stops as soon as it finds a local optimum or when a threshold number of iterations is exceeded.

The direct evaluation of the protein-ligand interaction energy is computationally expensive. Therefore in modern docking tools the grid of potentials representing protein-ligand interactions is calculated separately before the docking procedure and the energy is approximated using precalculated values. The resulting function is

not differentiable in the grid nodes and thus L-BFGS method is not applicable.

However of the energy is calculated directly local search methods employing gradient information can be advantageous. It is shown in (Tavares 2009) that L-BFGS method (Nocedal 2006) significantly improves the performance of genetic algorithms and gives better results w.r.t. Solis-Wets method.

Though the efficiency of various search methods was addressed in the literature some important methods were missed. For instance semi-local methods e.g. Monotonic Basin Hopping that proved to be very efficient for atomic cluster conformation problem (Leary 2000) were not considered at all. In this paper we classify search methods into three groups: local, semi-local and global and performs the systematic evaluation of the several techniques and their combinations.

3 CALCULATING PROTEIN-LIGAND INTERACTION ENERGY

The protein-ligand interaction energy is the objective function in the problem (1). Its accurate evaluation is crucial for a success of the whole interaction simulation and thus for the validity of the numerical experiments.

Two very popular programs implementing docking algorithms GOLD (Cole 2005) and AutoDock (Morris 1998, Morris 2005) employ too simplified force field, either neglecting electrostatic interaction in GOLD, or too simplified treating of desolvation terms in AutoDock.

For our study we considered the interaction model called SOL proposed in (Romanov 2004, Romanov 2008, Oferkin 2011). The main idea of this model is to describe with maximal possible accuracy the protein-ligand interactions, using the docking procedure based on contemporary molecular mechanics. The main distinctive features of SOL are:

- A rigid target-protein with the active site represented by a set of grids for different type potentials, describing protein-ligand interactions (electrostatic, Van der Waals (VdW) forces) in the frame of Merck Molecular Force Field (MMFF) (Halgren 1996).
- Quite rigorous description of solvation-desolvation effects upon ligand binding process, based on Generalized Born approximation (Ghosh 1998) and included in the set of potential grids.
- The grid of potentials representing protein-ligand interactions are calculated separately before the docking procedure. Electrostatic, VdW and solvation-desolvation potentials were calculated on the 101x101x101 grid inside this cube
- All ligands are considered as fully flexible, i.e. all topologically available torsion degrees of freedom were unfrozen and allowed to rotate freely, directed only by ligand internal energy preferences in the frame of MMFF. Bond lengths and valence angles have been frozen in the course of the docking procedure.

4 OPTIMIZATION METHODS UNDER TEST

We considered three types of optimization methods: local, semi-local and global.

4.1 Local methods

The goal of local methods is to find the local minimum i.e. the point that gives the minimal function value in a some neighborhood. We considered four methods described in table 1.

Table 1. Local optimization methods under test

Name	Reference	Brief description
CG	(Polak 1971)	Conjugate Gradients Method
TNC	(Nash 2000)	Truncated Newton Method
LBFGS	(Nocedal 2006)	The limited memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) method
Powell	(Powel 1964)	Powell's method

The conjugate gradient method is a seminal optimization method that is explained in almost every global optimization textbook. It uses conjugate directions instead of the local gradient for going downhill. We used the Polak-Ribere form for calculating conjugate directions.

The classical Newton method requires fewer iterations than conjugate gradients but each iteration involves the resolution of the system of linear algebraic equations. The truncated Newton methods are based on the idea that an exact solution of the Newton equation at every step is unnecessary and can be computationally wasteful in the framework of a basic descent method. Any descent direction will suffice when the objective function is not well approximated by a convex quadratic and, as a solution to the minimization problem is approached, more effort in solution of the Newton equation may be warranted.

Another very successful approach to local unconstrained optimization is Broyden–Fletcher–Goldfarb–Shanno (BFGS) method. This algorithm belongs to the family of quasi-Newton methods. In these methods the Hessian matrix of second derivatives need not be evaluated directly. Instead, the Hessian matrix is approximated using rank-one updates specified by gradient evaluations (or approximate gradient evaluations). The L-BFGS algorithm studied in this paper stores only a few vectors that represent the approximation implicitly. Due to its moderate memory requirement, L-BFGS method is particularly well suited for optimization problems with a large number of variables.

The three methods outlined above require the evaluation of the objective function gradient. However for docking problem the gradient is either not known symbolically or computationally in-

tractable. Thus one has to approximate the gradient numerically or use methods that don't rely on gradient information. For the Powell's method the objective function need not be differentiable, and no derivatives are taken. The method minimizes the function by a bi-directional search along each search vector. The new position can then be expressed as a linear combination of the search vectors. The new displacement vector becomes a new search vector, and is added to the end of the search vector list. Meanwhile the search vector which contributed most to the new direction, i.e. the one which was most successful, is deleted from the search vector list. The algorithm iterates an arbitrary number of times until no significant improvement is made. The method is useful for calculating the local minimum of a continuous but complex function, especially one without an underlying mathematical definition, because it is not necessary to take derivatives.

4.2 Semi-local methods

For functions with multiple extremes a local search methods get stuck in local minima. The semi-local methods can "escape" from a local minimum by exploring its neighborhood. Such methods have proved their efficiency for molecular conformation problems (Wales 1997). We considered two such methods summarized in the Table 2.

Table 2. Semi-Local optimization methods under test

Name	Reference	Brief description
MBH	(Leary 2000)	Monotonic Sequence Basin Hopping
BP	(Panteleev 2005)	Best Probe Method

The monotonic sequence basin hopping method tries to improve the minimum until the number of attempts exceeds the threshold value M. Starting from some point $x \in P$ MBH performs the following steps ($N = 0$ at the beginning):

1. Select randomly a point $y \in U(x)$.
2. Obtain point z by applying local minimization to $y \in U(x)$: $z = L(y)$.
3. If $f(z) \le f(x)$ then assign $x \coloneqq z, N \coloneqq 0$. Otherwise assign $N \coloneqq N + 1$..
4. If $N \ge M$. then finish, otherwise go to 1.

where $U(x) = \{y \in R^n : |x_i - y_i| \le \rho, i = 1, \dots, n\}$ is a box neighborhood of the given radius ρ .
Thus the MBH method is parameterized by the threshold value M, the neighborhood radius ρ and the local search method L.

The best probe method is based on the same idea as MBH. But unlike MBH it chooses points on the *n*-dimensional sphere rather than in a box and uses adaptive neighborhood size. Starting with a large radius it reduces its size if M attempts didn't lead to an improvement.

4.3 Global methods

The goal of global methods is to diversify the search in order to explore the whole feasible set. At the moment we considered only one globalization strategy: the Monte-Carlo method. It generated a sequence of uniformly distributed random generated points in a feasible box. After that each point is used a starting point for semi-local methods.

5 EXPERIMENTS

We performed several experiments for different protein-ligand pairs. Methods demonstrate the same relative behavior for all variants and thus we present data only for one pair: the target protein is thrombin (PDB code 1O2G) and the ligand is 4-aminopyridine in the protonated form.

5.1 Testing results for local methods

The results for local methods are summarized in the table 3. The average energy is calculated as a mean value of 64 runs with randomly generated initial points. The best energy is a lowest value found throughout these runs. The percentage of errors shows the number of runs that produced a value greater than the initial value.

Table 3. Testing results for local methods

Name	Avrg. Energy	Best Energy	Errors (%)	Avrg. Time(s)
Initial	15385,25	4728,66	0,00	0,00
CG	5790,28	874,80	9,52	1,81
TNC	4222,24	1322,46	0,00	3,90
LBFGS	3119,95	782,21	1,59	5,92
Powell	611,49	28,66	0,00	5,12

As expected the local methods generally improve the energy value. The Powell method remarkably outperforms methods that rely on Taylor formula. This is an expectable result as the objective function obtained as a result of piece-wise linear approximation on a mesh is non-differentiable in minima. Therefore the Taylor series gives a poor approximation for a goal function in a neighborhood of such points. Thus the gradient information can only be used to bias the search to the descending direction but not as a stopping criterion.

5.2 Testing results for semi-local methods

Results of the previous section clearly indicate that the Powell method is the best local search strategy among the considered set. Thus we used this method as a local method L in MBH and BP semi-local methods. After a set of experiments we found that the best results are obtained if MBH uses the radius $\rho = 0.1$ and the BP uses the initial sphere radius $\rho = 0.5$. The accuracy of results for both methods depends on the threshold values: the higher threshold value the lower the minimum. To put both methods in the equal conditions we set the threshold value to 30 and 90 for BP and MBH respectively. With such parameters both methods take approximately the same time.
The Table 4 compares the results of the basin hopping and best probe methods. The results were averaged over 10 runs with different random initial points generated in a box P.

Table 4. Testing results for semi-local methods

Name	Avrg. Energy	Best Energy	Avrg. Time(s)
Initial	16885,3	5628,76	0,00
MBH	-73,44	-134,60	451,37
BP	-27,77	-131,93	491,84

Both methods under test gives approximately the same best results, however the average behavior of MBH is much better.

5.3 Testing results for global methods

At the moment we tried only one globalization strategy: Monte Carlo method that generates random initial points for MBH methods. Points are generated in the bounding box *P*. Table 5 summarizes results obtained from 10 runs of this combination. The threshold value for MBH method was set to 90 and the initial radius was set to 0.1. The Powell method was used for a local search. The number of initial points generated by Monte-Carlo methods was set to 64.

Table 5. Testing results for Monte-Carlo method coupled with MBH method

Name	Avrg. Energy	Best Energy	Avrg. Time (s)
Monte-Carlo	-138,66	-154,47	10913,7

The obtained results demonstrate that even a very simple globalization strategy gives a noticeable improvement over plain semi-local and local methods. But the running time is considerably higher.

6 SCIENCE GATEWAY INTEGRATION

The ultimate goal of our work is to create software environment for molecular simulation. Such environments are useful to isolate the end-user from technical details of the application.

The use of high-performance computing in docking is inevitable because in practice millions of ligands have to be processed independently. Docking is perfectly suitable for the desktop grid computing (Kiss, Greenwel 2010, Kiss 2010). We are going to develop this application and deploy it at ISA RAS desktop grid (boinc.isa.ru/dcsdg) and on a combined infrastructure that connects ISA RAS desktop grid and service grid infrastructure provided by EDGeS VO (www.edges.eu). Such combined infrastructures get much attention in the grid community in recent years (Urbach 2009). This approach provides the transparent seamless integration of desktop and service grids and results in huge consolidated computing power.

The deployment will done on top of the BNB-Grid (Evtushenko 2009) programming environment – a library for solving large-scale optimization problems in the grid and supercomputers. This tool is currently ported to the desktop grid and is validated for running in the EDGeS VO. Docking will be one of the supported applications. Another one – atomic cluster conformation problem (Leary 2000) is already running in production. With such approach we'll reuse the one deployed application for different problems. This is very important for desktop grids and combined infrastructures where the deployment and validation requires lots of efforts.

The docking is only one element of a complex workflow in a drug design process. The good software environment is crucial for making the docking software tools useful for a wide community of researchers. One of such environments is described in (Kim 2008). It flexibly integrates the convenient Web-based user interface and the powerful processing back-end deployed in the Grid. Such software architecture seems to be the standard approach for docking and other applications demanding for the huge computing power. To implement a consistent and flexible software environment for drug design one needs a powerful workflow engine. This is crucial for combining different building blocks in Drug Design in a flexible and configurable way. We plan to use P-PGRADE portal (Farkas 2011) which is capable to construct complex workflows and harness different types of grids and clouds.

7 CONCLUSIONS

The docking problem is basically a global optimization problem (1) where objective f(x) is a protein-ligand interaction energy. Possessing the powerful techniques for resolving this problem is crucial for the efficiency of docking. In this paper we considered several optimization methods for problem (1). Numerical experiments showed that the best results can be achieved by combining some globalization techniques (e.g. Monte-Carlo method) and the monotonic basin hopping semi-local method coupled with the Powell local search algorithm.

We also outlined the future software environment which will provide a consistent and convenient access for a wide range of researchers to the drug-design experimetns.

ACKNOWLEDGEMENTS

This study was supported in part by grant 10-07-00595 from the Russian Foundation for Basic Research and the DEGISCO FP/7 European (contract no. 261561).

REFERENCES

Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discov., 3, 935-949.

Zoete, V.; Grosdidier, A.; Michielin, O. (2009) Docking, virtual high throughput screening and in silico fragment-based drug design. J. Cell. Mol. Med., 13, 238–248.

Rosin, C.; Halliday, S.; Hart, W.; Belew, R. (1997) A comparison of global and local search methods in drug docking. Proc 7th Intl Conf on Genetic Algorithms. San Francisco, CA., 221-228.

Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. (1998) Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. J. Comp. Chem., 19, 1639-1662.

Tavares, J.; Melab, N.; Talbi, E. (2008) An Empirical Study on the Influence of Genetic Operators for Molecular Docking Optimization. Inria Research Report, ISSN 0249-6399 ISRN INRIA/RR--6660--FR+ENG.

Tavares, J.; Melab, N.; Talbi, E. (2009) On the Efficiency of Local Search Methods for the Molecular Docking Problem. EvoBIO. 104-115.

Solis, F.; Wets, R. (1981) Minimization by Random Search Techniques. Mathematical Operations Research, 6: 19-30.

Nocedal, J.; Wright, S.J. (2006), Numerical Optimization (2nd ed.), Berlin, New York: Springer-Verlag, ISBN 978-0-387-30303-1.

Leary, R. H. (2000) Global Optimization on Funneling Landscapes. J. of Global Optimization 18, 4, 367-383.

Cole, J.C.; Nissink, J.W.M.; Taylor, R. (2005) Virtual Screening in Drug Discovery; Alvarez, J.; Shoichet, B. Eds.; CRC Press: Taylor & Francis Group, Boca Raton, London, New York, Singapore; pp. 379-415.

Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J. Comp. Chem., 30, 2785–2791.

Romanov, A.N.; Kondakova, O.A.; Grigoriev, F.V.; Sulimov, A.V.; Luschekina, S.V.; Martynov Ya.B.; Sulimov V.B. (2008) The SOL docking package for computer-aided drug design. Numerical Methods and Programming (Section 1. Numerical methods and applications9(2), 64-84 (in Russian).

Romanov, A.N.; Jabin, S.N.; Martynov, Ya.B.; Sulimov, A.V.; Grigoriev, F.V.; Sulimov, V.B. (2004) Surface Generalized Born method: a simple, fast and precise implicit solvent model beyond the Coulomb approximation. J. Phys. Chem. A, 108, 9323-9327.

Oferkin, I.V.; Sulimov, A.V.; Kondakova, O.A.; Sulimov V.B. (2011) Docking programs SOLGRID & SOL: parallel implementations. Numerical Methods and Programming (Section 2. Programming) 12, 9-23 (in Russian).

Halgren, T.A. (1996) Merck Molecular Force Field. J. Comp. Chem., 17(5&6), 490-641.

Ghosh, A.; Rapp, C.S.; Friesner, R.A. (1998) Generalized Born Model Based on a Surface Integral Formulation. J Phys Chem B, 102, 10983-10990.

Polak, E. (1971) Computational Methods in Optimization. New York: Academic Press.

Nash, S. (2000)A survey of truncated-Newton methods. Journal of Computational and Applied Mathematics. Volume 124 Issue 1-2, Dec. 2000.

Powell M.J.D. (1964) An efficient method for finding the minimum of a function of several variables without calculating derivatives, Computer Journal, 7 :155–162.

Wales, D.; Doye, J. (1997) Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. J. Phys. Chem. A, 101, 5111-5116.

Panteleev, A.V., Letova. T.A. (2005) Optimization methods via examples. Moscow, 544 p.

Yoo,M.S. *et al.* (2003) Oxidative stress regulated genes in nigral dopaminergic neurnol cell: correlation with the known pathology in Parkinson's disease. *Brain Res. Mol. Brain Res.*, **110**(Suppl. 1), 76–84.

Kim, J.B., Nhan, N.D. et al. (2008) DrugScreener-G: Towards an Integrated Environment for Grid-Enabled Large-Scale Virtual Screening and Drug Discovery. Proceedings ESCIENCE '08 Proceedings of the 2008 Fourth IEEE International Conference on eScience, 666 – 671.

Kiss T., Greenwell P., Heindl H. *et al.* (2010) Parameter Sweep Workflows for Modelling Carbohydrate Recognition, Journal of Grid Computing, volume 8, number 4, pp 587-601, DOI: 10.1007/s10723-010-9166-8

Kiss T., Szmetanko G. et al. (2010) Molecular Docking Simulations on a Combined Desktop and Service Grid Infrastructure, Proceedings of the 3rd Grid Experience workshop - Desktop Grid Applications for eScience and eBusiness, March 30, 2010, Almere, The Netherlands, p. 23-28

E. Urbach, P. Kacsuk, Z. Farkas *et al.* (2009) EDGeS: Bridging EGEE to BOINC and Xtrem Web, Journal of Grid Computing, Vol 7, No. 3, pages 335 -354.

Evtushenko, Y., Posypkin, M., Sigal, I. (2009) A framework for parallel large-scale global optimization. Computer Science - Research and Development 23(3), 211-215.

Farkas, Z., Kacsuk, P. (2011) P-GRADE Portal: A generic workflow system to support user communities. Future Gener. Comput. Syst. 27, 5, 454-465. DOI=10.1016/j.future.2010.12.001.

Evtushenko Yu. (1971) Numerical methods for finding global extreme (case of a non-uniform mesh). U.S.S.R. Comput. Maths. Math. Phys., Vol. 11, pp. 1390-1403.

De Jong K. A. (2006) Evolutionary computation: a unified approach. MIT Press, Cambridge MA