Analysis of docking algorithms by HPC methods generated in bioinformatics studies

Anton Stoilov[1,2*], *Borislav* Yurukov [3], *Peter* Milanov [3,4]

[1] South-West University, Faculty of Engineering, Department of Electrotechnics, Electronics and Automation, Bulgaria

[2]American University in Bulgaria, Department of Computer Science, Bulgaria

[3]South-West University, Faculty of Mathematics and Natural Science, Department of Informatics, Bulgaria

[4]Bulgarian Academy of Sciences, Institute of Mathematics and Informatics, Bulgaria

Abstract. High-performance computing (HPC) is an important domain of the computer science field. For more than 30 years, it has allowed finding solutions to problems and enhanced progress in many scientific areas such as bioinformatics and drug design. The binding of small molecule ligands to large protein targets is central to numerous biological processes. The accurate prediction of the binding modes between the ligand and protein (the docking problem) is of fundamental importance in modern structure-based drug design. The interactions between the receptor and ligand are quantum mechanical in nature, but due to the complexity of biological systems, quantum theory cannot be applied directly. Consequently, most methods used in docking and computational drug discovery are more empirical in nature and usually lack generality.

1 Introduction

Quantum mechanical phenomena, such as the formation of a covalent bond between the protein and the ligand upon binding during the transition state of the reaction, cannot be predicted and/or evaluated using these empirical methods. In the field of molecular modeling, docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using, for example, scoring functions. Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to in turn predict the affinity and activity of the small molecule. Hence docking plays an important role in the rational design of drugs. Given the biological and pharmaceutical significance of molecular docking, considerable efforts have been directed towards improving the methods used to predict docking. Each docking program makes use of one or more specific search algorithms, which are the methods used to predict the possible conformations of a binary complex. The present benchmark is made from an existing test set (CCDC/Astex Validation Set) on typical HPC system. Selected examples were docked with GOLD software.

2 Molecular docking

Molecular docking is a computer simulation procedure to predict the conformation of a receptor-ligand complex, where the receptor is usually a protein or a nucleic acid molecule (DNA or RNA) and the ligand is either a small molecule or another protein. It can also be defined as a simulation process where a ligand position is estimated in a predicted or pre-defined binding site. Molecular docking research focusses on computationally simulating the molecular recognition process. It aims to achieve an optimized conformation for both the protein and ligand and relative orientation between protein and ligand such that the free energy of the overall system is minimized. Computational docking of a small molecule to a biological target involves efficient sampling of possible poses of the former in the specified binding pocket of the latter in order to identify the optimal binding geometry, as measured by a user-defined fitness or score function. X-ray crystallography and NMR spectroscopy continue to be the primary source of 3-dimensional structural data for protein and nucleic acid targets. In favourable cases where proteins of unknown structure have high sequence homology to known structures, homology modelling can provide a viable alternative by generating a suitable starting point for “in silico” discovery of high affinity ligands. Potential energy of molecular field model is a function of a atomic position (x,y,z) normally in Cartesian space. The equation of the potential energy of the system of atoms in the molecular force field, commonly used in molecular modelling is presented below:

[*] Corresponding author: antonstoilov@swu.bg

$$E=\frac{1}{2}\sum_{i=1}^{m}k_{\theta i}(\theta_i-\theta_{0,i})^2+\frac{1}{2}\sum_{i=1}^{n}k_{bi}(b_i-b_{0,i})^2+\frac{1}{2}\sum_{i=1}^{k}v_i\left[1+\cos\left(n_i\omega_i-\gamma_i\right)\right]+\sum_{i,j}4\varepsilon_{i,j}\left[\left(\frac{\sigma_i}{r_{ij}}\right)^{12}-\left(\frac{\sigma_j}{r_{ij}}\right)^{6}\right]+\sum_{i,j}\frac{1}{4\pi\varepsilon_0}\frac{q_iq_j}{r_{ij}^2} \quad (1)$$

The complexity of computational docking increases in the following order:
(a) rigid body docking, where both the receptor and small molecule are treated as rigid. (b) flexible ligand docking, where the receptor is held rigid, but the ligand is treated as flexible; (c) flexible docking, where both receptor and ligand flexibility is considered.
Rigid-body docking simulation has been employed for virtual-screening initiatives, this method has been used as the fastest way to perform an initial screening of a small molecule database. It has a relatively high accuracy, when compared against crystallographic structures. This accuracy is even higher if we introduced an analysis of the best results using an empirical scoring function for the best results obtained using rigid-body docking simulations. Usually, flexible docking or/and scoring functions have been used for applying a more specific refinement and lead optimization after initial rigid body docking procedure, since these methods demand for computational power and CPU time. Flexible docking methods can consider several possible conformations of ligand or receptor, as well as for both molecules at the same time, at a higher computational time cost. The root-mean-square deviation (RMSD) is calculated between two sets of atomic coordinates, in this case, one for the crystallographic structure (xc, yc, zc) and another for the atomic coordinates obtained from the docking simulations (xd, yd, zd), the summation is taken over all N atoms being compared, the equation is as follows:

$$RMSD=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ci}-x_{di})^2}+(y_{ci}-y_{di})^2+(z_{ci}-z_{di})^2 \quad (2)$$

Before ligands can be docked against a receptor, generally the binding site has to be identified first. This is done to limit the search space on the receptor surface and thus minimize the degrees of freedom that have to be searched. The active site is often known from crystal structures of ligand-bound receptors, but it can also be predicted. The largest cavity n a protein surface is frequently the active site, but this is not always the case and different active site prediction and analysis methods have been developed.

The genetic algorithm (GA) adopted by GOLD algorithm requires as input the approximate size and location of the receptor active site and also the coordinates of protein and a ligand conformation. The active site may be defined by several techniques. GA is also implemented in the program DOCK, which is able to dock either whole ligand inside active site or a rigid fragment of the ligand.“Lamarckian” GA (LGA) is also implemented in docking algorithms. The LGA switches between “genotypic space” and “phenotypic space.” Mutation and crossover occur in genotypic space, while phenotypic space is determined by the energy function to be optimized. Energy minimization (local sampling) is performed after genotypic changes have been made to the population (global sampling) in phenotypic space, which is conceptually similar to MC minimization. After successful docking procedure most important parameters are:

- Binding Energy - (BE), kcal/mol
- Intermol Energy – (ImolE), kcal/mol
- Internal Energy – (IE), kcal/mol
- Torsional Energy – (TE), kcal/mol
- Unbound Energy (UE), kcal/mol

Binding energy calculations can be performed either concurrently with ligand docking or separately for a predetermined protein-ligand structure, obtained from experimental data or other molecular modeling studies. The Binding energy is calculated by sum from Intermol Energy, Internal Energy, Torsional Energy and Unbound Energy.

(3) BE = ImolE + IE + TE + UE

3 Methods

GOLD software uses a Genetic Algorithm (GA) for protein ligand docking which works as follows:

1. Selecting a Protein
2. Adding Hydrogen Atoms
3. Deleting Waters
4. Defining the Protein Binding Site
5. Specifying the Ligand
6. Selecting a Fitness Function
7. Starting the Docking Run
8. Analysis of Output

The population of chromosomes is iteratively optimised. At each step, a point mutation may occur in a chromosome, or two chromosomes may mate to give a child. The selection of parent chromosomes is biased towards fitter members of the population, i.e. chromosomes corresponding to ligand dockings with good fitness scores. The GOLD validation test set is one of the most comprehensive of all of the docking methods reviewed, and achieved a 71% success rate based primarily on a visual inspection of the docked structures. 66 of the complexes had an RMSD of 2.0 Å or less, while 71 had an RMSD of 3.0 Å or less. The omission of hydrophobic interactions and a solvent model may explain some of the docking failures which included highly flexible, hydrophobic ligands, and those complexes containing poorly resolved active sites. However, recent extensions to GOLD include the addition of hydrophobic fitting points that are used in the least squares fitting algorithm to generate the ligand orientation.

In this paper the benchmark test set is based on CCDC/Astex Validation Set developed by Cambridge Crystallographic Data Centre (CCDC) for docking software GOLD. There are 60 entries and protonation states have been set in all cases.

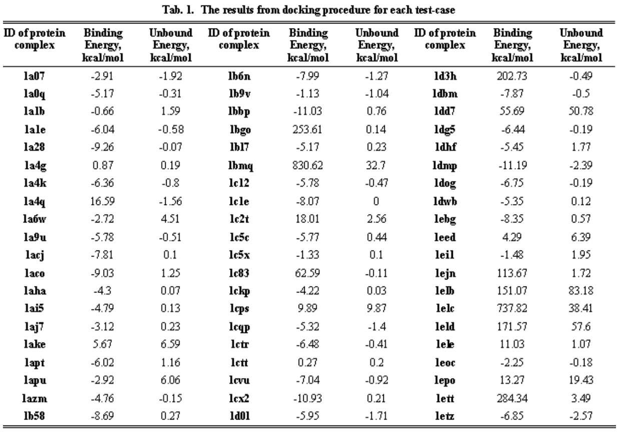
The equipment for experiments was provided by CENTER FOR ADVANCED BIOINFORMATICS RESEARCH, South-West University “Neofit Rilski”, Blagoevgrad, BULGARIA. This equipment include two different computational power with this characteristics:
IBM x3650 M2 – Processor: 2 x Xeon Quad-Core Intel Xeon 4C Model E5520 4C 2.26GHz with EM64T/1066MHz, 8MB L3 Cache, RAM: 2 x2GB 4096MB ECC PC3-10600 DDR3, HDD: 2x146GB 10K SAS Hot Swap RAID controller - Integrated RAID MR10i

IBM BladeCenter HS22 – Processor: 2 x Xeon Quad-Core Intel Xeon 4C Model E5504 80W 2.00GHz/800MHz/4MB L2, 2x2GB, O/Bay 2.5in SAS 1, RAM: 4 x 2GB Single Rank PC3-10600 CL9 ECC DDR3, HDD: 2 x IBM 146 GB 2.5in SFF Slim-HS 10K 6Gbps SAS HDD

4 Results

After the execution of 60 docking procedure were obtained 10 conformation for each case. After that from those 10 confirmation is chosen most energy-favorable i.e. that have the smallest binding energy (BE) so this is the main criteria for present results in Table 1. All confirmation are saved and observed. The summary of these most energy-favorable conformation for each test case are presented in Table 1.

Table 1. The results from docking procedure for each test case.



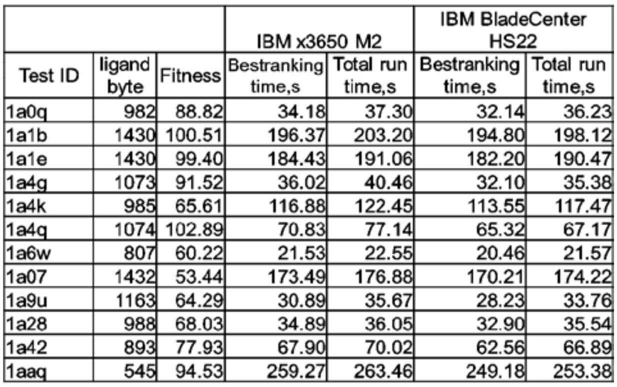
Tab. 1. The results from docking procedure for each test-case

ID of protein complex	Binding Energy, kcal/mol	Unbound Energy, kcal/mol	ID of protein complex	Binding Energy, kcal/mol	Unbound Energy, kcal/mol	ID of protein complex	Binding Energy, kcal/mol	Unbound Energy, kcal/mol
la07	-2.91	-1.92	**lb6n**	-7.99	-1.27	**ld3h**	202.73	-0.49
la0q	-5.17	-0.31	**lb9v**	-1.13	-1.04	**ldbm**	-7.87	-0.5
lalb	-0.66	1.59	**lbbp**	-11.03	0.76	**ldd7**	55.69	50.78
lale	-6.04	-0.58	**lbgo**	253.61	0.14	**ldg5**	-6.44	-0.19
la28	-9.26	-0.07	**lbl7**	-5.17	0.23	**ldhf**	-5.45	1.77
la4g	0.87	0.19	**lbmq**	830.62	32.7	**ldmp**	-11.19	-2.39
la4k	-6.36	-0.8	**lc12**	-5.78	-0.47	**ldog**	-6.75	-0.19
la4q	16.59	-1.56	**lcle**	-8.07	0	**ldwb**	-5.35	0.12
la6w	-2.72	4.51	**lc2t**	18.01	2.56	**lebg**	-8.35	0.57
la9u	-5.78	-0.51	**lc5c**	-5.77	0.44	**leed**	4.29	6.39
lacj	-7.81	0.1	**lc5x**	-1.33	0.1	**leil**	-1.48	1.95
laco	-9.03	1.25	**lc83**	62.59	-0.11	**lejn**	113.67	1.72
laha	-4.3	0.07	**lckp**	-4.22	0.03	**lelb**	151.07	83.18
lai5	-4.79	0.13	**lcps**	9.89	9.87	**lelc**	737.82	38.41
laj7	-3.12	0.23	**lcqp**	-5.32	-1.4	**leld**	171.57	57.6
lake	5.67	6.59	**lctr**	-6.48	-0.41	**lele**	11.03	1.07
lapt	-6.02	1.16	**lctt**	0.27	0.2	**leoc**	-2.25	-0.18
lapu	-2.92	6.06	**lcvu**	-7.04	-0.92	**lepo**	13.27	19.43
lazm	-4.76	-0.15	**lcx2**	-10.93	0.21	**lett**	284.34	3.49
lb58	-8.69	0.27	**ld01**	-5.95	-1.71	**letz**	-6.85	-2.57

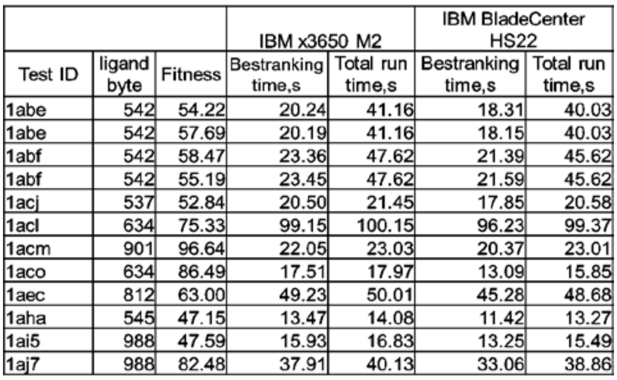
The ligand will have been docked a number of times so a set of files will have been written to the output directory, each containing the results of a separate docking attempt. The result of each docking attempt is written out as gold_soln_ligand_m1_n.mol2, where n is the number of the docking solution 1,2,3 ... and m1 is an index to the ligand (in this example, only one ligand was docked). Note that the file gold_soln_ligand_m1_1.mol2 is not the best GOLD prediction, it is just the solution found in the first docking attempt. However, as GOLD proceeds, symbolic links are created: ranked_ligand_m1_1.mol2 will point to the current top-ranked solution, ranked_ligand_m1_2.mol2 will point to the second-best solution, and so on. With the Hermes 3D view is possible to inspect the solutions predicted by GOLD. The docking solutions are given in their docked order with their corresponding fitness score. If required the solutions can be ordered and Fitness to determine which is the highest scoring. A simple test of the effectiveness of a docking program is to take a protein-ligand complex from the PDB and extract the ligand. The docking program can then be used to predict the binding mode of the ligand and a comparison made with the crystallographically observed position. The crystallographically observed conformation of the docked ligand is stored in the ligand we extracted from the protein, that was subsequently re-loaded. Compare this with the solution predicted by GOLD.

After successful docking procedure for each test-case is observed by Fitness (scoring function), Best ranking time and Total run time . The results for 22 faster cases for docking from initial 60 experiments was present on Table 2. In the fields of molecular modelling, scoring functions are fast approximate mathematical methods used to predict the strength of the non-covalent interaction (also referred to as binding affinity) between two molecules after they have been docked. Most commonly one of the molecules is a small organic compound such as a drug and the second is the drug's biological target such as a protein receptor.

Table 2. 22 specific cases from CCDC/ASTEX validation test



			IBM x3650 M2		IBM BladeCenter HS22	
Test ID	ligand byte	Fitness	Bestranking time,s	Total run time,s	Bestranking time,s	Total run time,s
1a0q	982	88.82	34.18	37.30	32.14	36.23
1a1b	1430	100.51	196.37	203.20	194.80	198.12
1a1e	1430	99.40	184.43	191.06	182.20	190.47
1a4g	1073	91.52	36.02	40.46	32.10	35.38
1a4k	985	65.61	116.88	122.45	113.55	117.47
1a4q	1074	102.89	70.83	77.14	65.32	67.17
1a6w	807	60.22	21.53	22.55	20.46	21.57
1a07	1432	53.44	173.49	176.88	170.21	174.22
1a9u	1163	64.29	30.89	35.67	28.23	33.76
1a28	988	68.03	34.89	36.05	32.90	35.54
1a42	893	77.93	67.90	70.02	62.56	66.89
1aaq	545	94.53	259.27	263.46	249.18	253.38



			IBM x3650 M2		IBM BladeCenter HS22	
Test ID	ligand byte	Fitness	Bestranking time,s	Total run time,s	Bestranking time,s	Total run time,s
1abe	542	54.22	20.24	41.16	18.31	40.03
1abe	542	57.69	20.19	41.16	18.15	40.03
1abf	542	58.47	23.36	47.62	21.39	45.62
1abf	542	55.19	23.45	47.62	21.59	45.62
1acj	537	52.84	20.50	21.45	17.85	20.58
1acl	634	75.33	99.15	100.15	96.23	99.37
1acm	901	96.64	22.05	23.03	20.37	23.01
1aco	634	86.49	17.51	17.97	13.09	15.85
1aec	812	63.00	49.23	50.01	45.28	48.68
1aha	545	47.15	13.47	14.08	11.42	13.27
1ai5	988	47.59	15.93	16.83	13.25	15.49
1aj7	988	82.48	37.91	40.13	33.06	38.86

5 Discussion

Comparisons suggest that the best algorithm for docking is probably a hybrid of various types of algorithm encompassing novel search and scoring strategies. The most useful docking method will not only perform well, but will be easy to use and parametrise, and sufficiently adaptable such that different functionality may be selected, depending on the number of structures to be docked, the available computational resources, and the complexity of the problem. If the parameters cannot be generated quickly then although the algorithm may be computationally efficient, from a practical point of view it is limited. Conversely, a rapid scoring function may not necessarily be able to model some specific interactions. Moreover, although current docking methods show great promise, fast and accurate discrimination between different ligands based on binding affinity, once the binding mode is generated, is still a significant problem.

Acknowledgment

This research was supported by project „Bioinformatic research: protein folding, docking and predicting biological activity" NSF I02/16 12.12.14 funding by National Science Fund - Ministry of Education and Science of Bulgaria.

References

1. Halperin I, Ma B, Wolfson H, Nussinov R.Proteins **47** (4): 409–443 (June 2002).
2. Mustard D, Ritchie DW. Proteins **60** (2): 269–274 (August 2005).
3. Shoichet BK, Stroud RM, Santi DV, Kuntz ID, Perry KM. Science **259** (5100): 1445–50 (March 1993).
4. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK. Biopolymers **68** (1): 76–90 (January 2003).
5. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. J. Med. Chem. **47** (7): 1739–1749 (March 2004).
6. Jain AN. J. Med. Chem. **46** (4): 499–511 (February 2003).
7. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP. J. Mol. Graph. Model. **26** (1): 198–212 (July 2007).
8. Jones G, Willett P, Glen RC, Leach AR, Taylor R. J. Mol. Biol. **267** (3): 727–748 (April 1997).