JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 00, 2011

Pp. 1–10
DOI: 10.1089/cmb.2009.0111

Research Article

HarmonyDOCK: The Structural Analysis of Poses in Protein-Ligand Docking

*DARIUSZ PLEWCZYNSKI,[1] *ANNA PHILIPS,[2] MARCIN VON GROTTHUSS,[3] LESZEK RYCHLEWSKI,[3] and KRZYSZTOF GINALSKI[1]

ABSTRACT

Molecular docking is a widely used method for lead optimization. However, docking tools often fail to predict how a ligand (the smaller molecule, such as a substrate or drug candidate) binds to a receptor (the accepting part of a protein). We present here the HarmonyDOCK, a novel method for assessing the docking software accuracy, and creating the scoring function which would determine consensus protein-ligand pose among those generated by available docking programs. Conformations for few hundred protein-ligand complexes with known three-dimensional structure were predicted on a benchmark set by set of different docking programs. On the basis of the derived ranking, the point of reference and the lower score limit were determined for subsequent investigations. The focus of the methodology is on the top-ranked poses, with the assumption being that the conformation of the docked molecules is the most accurate. We found out that some docking programs perform considerably better than the others, yet in all cases the proper selection of decoys, namely HarmonyDOCK, is needed for successful docking procedure.

Key words: computational molecular biology, machine learning, protein motifs, protein structure.

1. INTRODUCTION

MOLECULAR DOCKING IS A SIMULATION PROCESS used for predicting how a ligand (the smaller molecule, such as a substrate or drug candidate) binds to a receptor (the accepting part of a protein). When these two molecules form a stable complex, the preferred orientation is found.

The computational solution of the molecular docking problem results in predicting the affinity of the small molecules, i.e., drug candidates, which reduce the number of lead components for a good drug. The computational methods reduce time and cost needed to introduce a new drug into the market.

Until the 1970s, hypothetical activity models dominated the syntheses of new compounds in drug research. The biological activity of these compounds was verified by experiments with isolated organs or animals. Accordingly, the throughput was limited by the speed and the cost of the experimental trials. The development of a new drug by classic methods takes 12–15 years and can cost above 800 million euro

[1]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland.
[2]Molecular Biophysics, Faculty of Physics, UAM, Poznan, Poland.
[3]BioInfoBank Institute, Poznan, Poland.
*These two authors contributed equally.

(Terfloth, 2003). The Boston Consulting Group has come to the conclusion that genomics could reduce the cost of a new drug production from 800 million euro by about 305 million euro in the ideal case (Terfloth, 2003).

Thanks to the Virtual High-Throughput Screening (VHTS) method, three-dimensional (3D) libraries can be rapidly screened (up to approximately 100,000 ligands/day) for virtual hits showing the best complementarity to a defined active site of the target protein. The VHTS procedure is much more cost-efficient than experimental high-throughput screening (EHTS) by robots. The single test for EHTS method with a library of 150,000 compounds costs about US$ 300,000 (Wermuth, 2003). The VHTS method is mostly used to select a small number (several hundreds up to a thousand) of lead compounds for featured laboratory tests. For example, researchers from Hoffman-La Roche pharmaceutical company have discovered new inhibitors of DNA gyrase. The discovering process was based on the 3D structural information of the DNA gyrase binding site. They applied *in silico* screening to find potential leads from the chemical databases. The LUDI software (Bohm, 1992a,b, 1994) was employed to perform docking. The hits of the screening were validated by biophysical methods and finally optimized. The process resulted in a set of DNA gyrase inhibitors. They were up to ten times more potent than another known inhibitor, novobiocin (Bohm et al., 2000).

Molecular docking simulations are of great importance in academic and industrial research, because they provide a deeper insight into biochemical processes and are instrumental in finding potent drugs.

1.1. Molecular docking process

The computational molecular docking is a research technique for predicting whether one molecule will bind to another one or not.

The interaction between two molecules can be evaluated by a scoring function that includes terms describing the inter- and intra-molecular energies (Lodish et al., 2000), such as electrostatic forces, Van der Walls interactions, hydrogen bounds (Rodwell, 1999), or hydrophobic interactions (Goss and Schwarzenbach, 2003). In general, the docking problem is posed as an optimization problem, where the objective is to minimize this interaction energy. The favourably bonded state is the state of the lowest energy. If conformation is complementary and involves favorable biochemical interactions, the ligand should bind to the protein *in vitro* and *in vivo.*

1.2. Docking search algorithms

The search space consists of all possible orientations and conformations of the protein paired with the ligand. However, the number of accessible states grows immensely with the size of the components (a phenomenon known as the combinatorial explosion). The possible number of conformations can be very large. Sophisticated optimization algorithms are needed to explore such search spaces successfully. The search algorithm should generate a set of complexes in which experimentally determined binding modes are included. To evaluate the accuracy of the computationally predicted poses, the root-mean-squared deviation (RMSD) is frequently used. It compares the experimentally observed heavy-atom positions of the ligands and the ones determined by the method. The number of algorithms available to assess and rationalize ligand-protein interactions is large and ever increasing. In addition, combining different docking techniques into a single strategy is a common and useful method to increase the effectiveness of a docking protocol. The most popular searching strategies are brute-force search (Katchalski-Katzir et al., 1992; Ritchie and Kemp, 2000), genetic algorithms (Mitchell, 1997), Monte Carlo simulations (Jorgensen and Tirado-Rives, 2005), simulated annealing, and structure-based methods.

1.3. Scoring functions

Computational docking methods generate a large number of candidate solutions, where a protein and a ligand contact each other in many different orientations. It is critically important to select the near-native poses from among all the others that were explored by the searching algorithm. Scoring functions express the geometric complementarity of the two molecules surfaces in contact and also the strength of the interaction, based on the physico-chemical characteristics of the two molecules. Most procedures handle the geometric and physico-chemical criteria separately. The accuracy of the scoring function can be evaluated on experimentally resolved complexes, where the native conformation is known. There are

databases with experimentally determined affinity data (e.g., freely accessible AffinDB) (Block et al., 2006).

Commonly, to improve scoring confidence, hybrid scoring functions (''consensus scoring'') are created. They combine the terms from two or more sources. Nevertheless, such complex scoring methods require multiple algorithms and are therefore computation-expensive, which could affect the speed of the whole docking process. Most often, scoring schemes are based on force-field based methods (Meng et al., 1992), empirical methods (Eldridge et al., 1997), and knowledge-based methods (Gohlke et al., 2000).

2. METHODS

2.1. Computations

First of all, the RMSD value to the native structure for all poses generated by tested docking methods was calculated.

2.2. Evaluation of docking software

The accuracy of each of the docking programs was estimated using the following two sums:

(1) The sum of the RMSD values: when the best-scored poses (proposed by docking application as the best ones) were chosen for each protein-ligand complex.
(2) The sum of the RMSD values: when the poses with the lowest RMSD were chosen for each protein-ligand complex.

2.3. Poses scoring functions

On the basis of the docking software rankings, the following poses scoring functions were created to choose the consensus pose for each complex among all poses generated by docking programs for this complex.

- Arithmetic - arithmetic mean;
- Arithmetic - harmonic (+0.5) mean;
- Arithmetic - harmonic (+2) mean;
- Arithmetic - harmonic (+5) mean;
- Arithmetic - harmonic (+20) mean;
- Arithmetic - quadric mean;
- Harmonic $(+0,5)$ - arithmetic mean;
- Harmonic $(+1)$ - arithmetic mean;
- Harmonic $(+2)$ - arithmetic mean.

All of the created methods are based on arithmetic, harmonic, and quadric means.

2.4. HarmonyDOCK: algorithm for scoring a set of poses

The computations were performed successively for each protein-ligand pair from the data set. We have chosen the consensus pose for each complex from among all poses generated by docking programs using previously mentioned poses scoring functions. All poses derived for a particular protein-ligand complex are scored according to the following scheme, where the algorithm consists of two main steps.

Step 1. Comparative process. All poses derived for a particular protein-ligand complex are compared to one another. That gives $\frac{(n^2-n)}{2}$ pose-pairs, where n is the number of the poses. Comparative process consists of two steps:

- Calculation of the square-distance between the corresponding heavy atoms (all atoms excluding hydrogen) according to the following equation:

$$d = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2$$

Where d is a square-distance between the corresponding atoms; x_1, y_1, z_1 are the coordinates of the first atom and x_2, y_2, z_2 are the coordinates of the second one.

- Calculation of a distance between two poses as the mean square-distance between all atoms using arithmetic and harmonic means. The names of the above means form the first element of the poses scoring functions' names.

At the end of step 1 (comparative process), the mean distance between all heavy atoms is determined for each of the pose-pair.

Step 2. Poses scoring process. The score of the certain pose is the mean of all the distances between this pose and all the other poses. For these calculations arithmetic, harmonic, and quadric means are used. The names of this mean form the second element of the poses scoring functions' names.

Subsequently, the poses are ranked by the calculated scores and the lowest-scored pose is chosen as the best one.

2.5. Evaluation of poses scoring functions

The accuracy of each of the poses scoring functions was estimated using the sum of the RMSD values, when the best-scored (lowest-scored) poses (proposed by a/the poses scoring function as the best ones) were chosen for each protein-ligand complex.

3. RESULTS

The results of the docking software evaluation are summarized in Table 1. The lower the value, the better the docking result.

In the first test (Table 2), the highest sum of RMSD was received by eHiTS (5074.45), which preformed worst. FlexX, AutoDock, GLIDE, Surflex, and GOLD obtained 2588.44, 2410.16, 2208.39, 2000.56, and 1792.81 points, respectively. LigandFit obtained the best result (1535.86). The docking software accuracy varies significantly between complexes. Some protein–ligand complexes have been accurately predicted by all the docking programs (e.g., 1mdq complex), and some examined docking simulations have been totally unsuccessful (e.g., 1pip and 1qji complexes). A graphical illustration of them is presented in Figures 1–4.

Thanks to the second method mentioned above, the lower possible RMSD-sum limit for a poses scoring function was determined. Again, LigandFit appeared to be the best, with the value of 774.777. The results of the second test confirmed the ranking obtained in the first test: eHiTS, FlexX, AutoDock, GLIDE, Surflex, GOLD, and LigandFit.

3.1. Results of poses scoring functions' evaluation

A few alternative methods based on different kinds of means were proposed to select a consensus protein-ligand pose from among those generated by the docking programs for a certain protein-ligand complex:

- Arithmetic - arithmetic mean;
- Arithmetic - harmonic $(+0.5)$ mean;
- Arithmetic - harmonic $(+2)$ mean;

Table 1. The Summary of Docking Software Results (in A) and Ranking

Rank	*Program*	*First test's results**	*Second test's results***
1	LigandFit	1535.86	774.777
2	GOLD	1792.81	1137.975
3	Surflex	2000.56	1368.599
4	GLIDE	2208.39	1399.091
5	AutoDock	2410.16	1522.193
6	FlexX	2588.44	1986.752
7	eHiTS	5074.45	2402.523

*Sum of RMSD of best scored poses.

**Sum of RMSD of poses with the lowest RMSD.

TABLE 2. THE SUMMARY OF POSES SCORING FUNCTIONS RESULTS (IN A) AND RANKING

Rank	*Poses scoring function*	*Score*[*]
	LigandFit – lower limit	774.77
1	HarmonyDOCK – using arithmetic-harmonic (+5) mean	1480.86
2	Arithmetic-harmonic (+2) mean	1490.09
	LigandFit - the point of reference	1535.86
3	Arithmetic-harmonic (+20) mean	1538.84
4	Arithmetic-harmonic (+0.5) mean	1565.44
5	Arithmetic-arithmetic mean	1673.86
6	Harmonic (+2)-arithmetic mean	1873.23
7	Harmonic (+1)-arithmetic mean	1895.89
8	Harmonic (+0.5)-arithmetic mean	1905.53
9	Arithmetic-quadric mean	1998.97

*Sum of RMSD distances to the native structure.

- Arithmetic - harmonic (+5) mean;
- Arithmetic - harmonic (+20) mean;
- Arithmetic - quadric mean;
- Harmonic (+ 0,5) - arithmetic mean;
- Harmonic (+ 1) - arithmetic mean;
- Harmonic (+ 2) - arithmetic mean.

The aim was to create the scoring method which would perform better than LigandFit, the best performing program. The first result of the docking software test (1535.86) became the point of reference to be outranked.

The second docking software test determined the theoretically possible accuracy of re-ranking the poses generated by one program, because only the best possible results were summarized (the results with the lowest RMSD) for each complex. This value is also the limit on my scoring method. In the ideal case, the accuracy of poses scoring function should reach the range of this value (774.777 for LigandFit).

The summary of poses scoring functions' results is collected in Table 2. Two functions, Arithmetic-harmonic (+5) mean and Arithmetic-harmonic (+2) mean, performed better than LigandFit (Arithmetic-

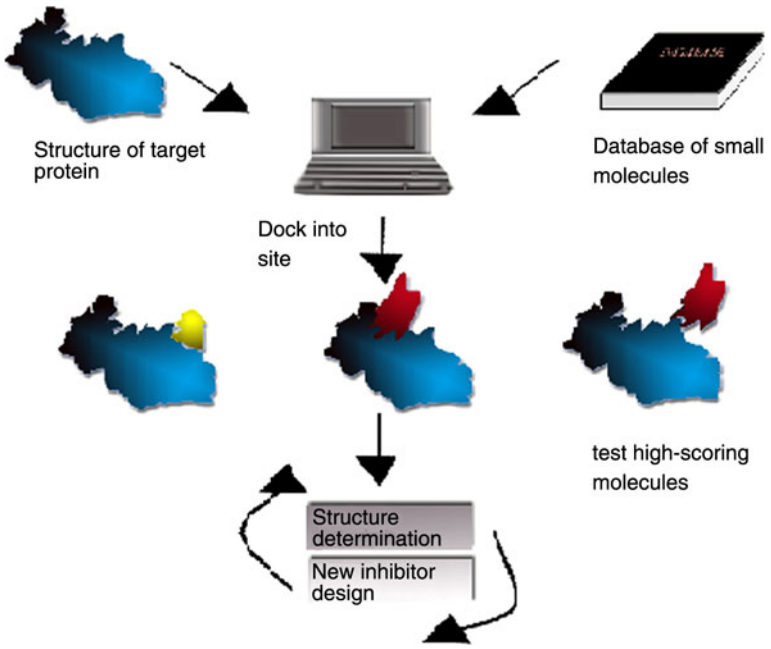

FIG. 1. The *in silico* process of new drug development.

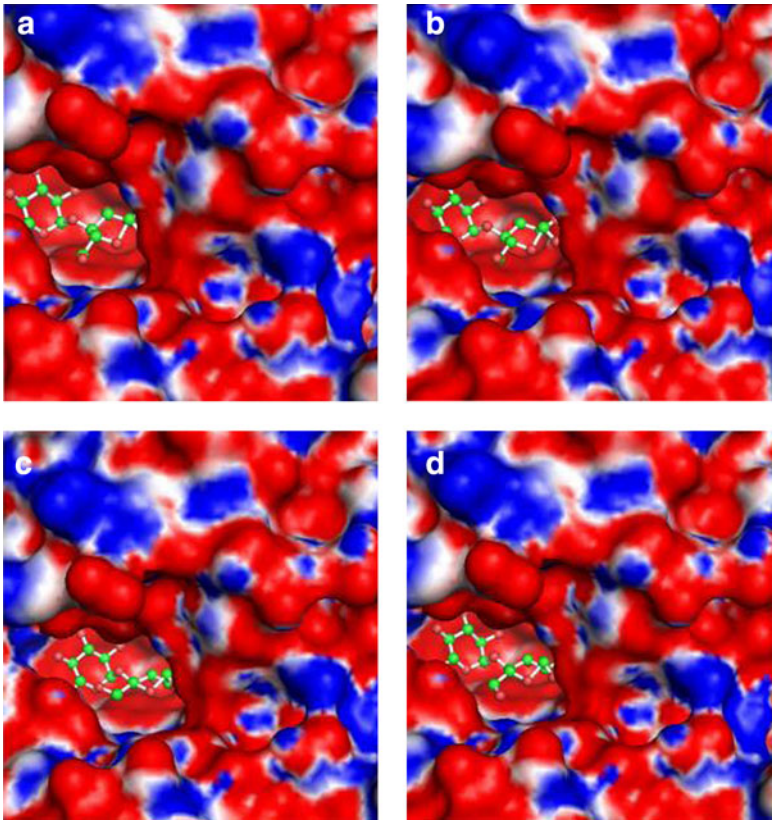

FIG. 2. Example of well–predicted poses (1mdq). **(a)** Native structure. **(b)** GOLD best-scored pose (GOLD 0). **(c)** FlexX best-scored pose (FlexX 0). **(d)** LigandFit best-scored pose (LigandFit 0).

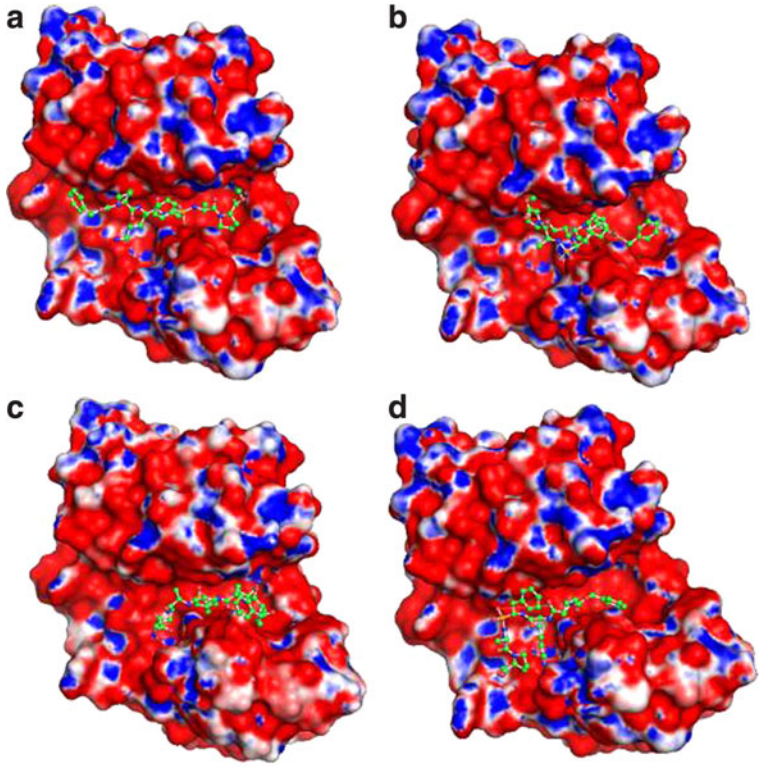

FIG. 3. Example of a poor predicted poses (1qji). **(a)** native structure. **(b)** Surflex best-scored pose (Surflex 0). **(c)** eHiTS best-scored pose (eHiTS 0). **(d)** GLIDE best-scored pose (GLIDE 0).

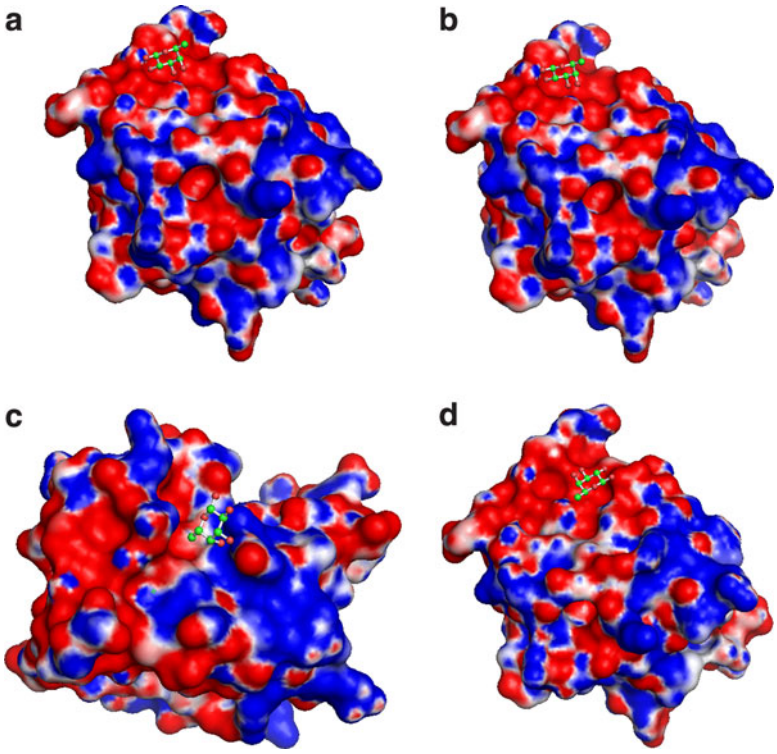

FIG. 4. Example of well-predicted poses by Arithmetic-harmonic (+5) mean (1rdj). **(a)** Native structure. **(b)** Arithmetic-harmonic (+5) mean bestscored pose (Surflex 6). **(c)** eHiTS best-scored pose (eHiTS 0). **(d)** GLIDE best-scored pose (GLIDE 0).

harmonic (+5) mean gained 1480.86 and Arithmetic-harmonic (+2) mean gained 1490.09). Arithmetic-harmonic (+20) mean (1538.84) and Arithmetic-harmonic (+0.5) mean (1565.44) scored similarly to LigandFit; the other functions ranked worse than the reference determined by the best performed software, LigandFit (1535.86). The results of the best performing function, Arithmetic-harmonic (+5) mean, are presented in Tables 3 and 4.

4. DISCUSSION

The docking software accuracy varies significantly between complexes. We find protein–ligand pairs to be completely incorrectly docked by all the docking programs (e.g., 1qji complexes) but there are also accurately predicted poses (e.g., 1mdq complex). The graphic illustration of the examples can be found in the Results section.

The size and diversity of data set seem to be sufficient to correctly assess the average results of docking software.

However, it must be noted that all simulations were carried out under default settings. Hence, the lower accuracy of some docking programs could have been caused by suboptimal choice of the searching and scoring procedures. For example, eHiTS results would have been better, if a search function other than a comprehensive search space one had been used. It is possible to apply pre-processing and determine the areas of the protein active sites. Also, it is worth bearing in mind that, in our docking simulations, the native conformation of the ligand was chosen as the starting conformation. Under such conditions, the methods that do not alter the conformation of the ligand during the docking process (LigandFit) have an advantage over methods that completely rebuild the conformation of the ligand (eHiTS). More natural conditions where the ligand conformation is not known could have resulted in a different ranking of the docking methods.

For poses scoring function, Arithmetic-harmonic (+5) mean produces better results than any other docking software tested in our first experiment based on the sum of RMSD values. However, its achievement differs only slightly from the results gained by LigandFit. The region of the lower score limit, which was determined by LigandFit in the second accuracy test, was not achieved.

Table 3. Comparison of the Success Rates of 17 Scoring Functions When the Cut-off is RMSD <3.0. (The Table is the Modified Figure 3. of Tiejun Cheng Group ("*Comparative Assessment of Scoring Functions on a Diverse Test Set*")

1	GOLD:ASP	89%
2	DS:PLP1	85%
3	GlideScore:SP	85%
4	DrugScorePDB:PairSurf	82%
5	GOLD:GoldScore	81%
6	DS:LigScore2	80%
7	GOLD:ChemScore	79%
8	X-score1.2:HMScore	78%
9	HarmonyDOCK	76%
10	SYBYL:F-Score	74%
11	SYBYL:ChemScore	71%
12	DS:Ludi2	67%
13	DS:Jani	65%
14	SYBYL:PMF-Score	58%
15	SYBYL:G-Score	56%
16	DS:PMF	53%
17	SYBYL:D-Score	48%

If the number of generated poses by each docking program had been increased from 10 to 100, a more precise computation would have been possible and the accuracy of poses scoring functions could have been higher. However, such docking simulations have not been performed because the computations would take too much time.

Arithmetic and harmonic means are simple statistic measures. They determined consensus pose among all the available ones for each complex. However, each set of poses generated by docking software for a certain protein-ligand pair has artefacts, i.e., outlier members, which could obscure the results and lower their quality. The substitution of arithmetic and harmonic means for some methods, which would exclude extreme poses, would possibly improve the accuracy of poses scoring functions. In such a case, the favorable consensus pose would be chosen only among the ''good'' ones. This has not been tested yet.

Table 4. Comparison of the Success Rates of 17 Scoring Functions When the Cut-off is RMSD <2.0. (The Table is the Modified Figure 3. of Tiejun Cheng Group ("*Comparative Assessment of Scoring Functions on a Diverse Test Set*")

1	GOLD:ASP	82%
2	DS:PLP1	75%
3	DrugScorePDB:PairSurf	75%
4	GlideScore:SP	72%
5	DS:LigScore2	71%
6	GOLD:ChemScore	70%
7	GOLD:GoldScore	69%
8	X-score1.2:HMScore	69%
9	HarmonyDOCK	65%
10	SYBYL:F-Score	65%
11	SYBYL:ChemScore	60%
12	DS:Ludi2	58%
13	SYBYL:PMF-Score	48%
14	DS:Jani	45%
15	DS:PMF	44%
16	SYBYL:G-Score	41%
17	SYBYL:D-Score	30%

5. CONCLUSION

A few hundred protein-ligand pairs have been processed using popular docking software. Scorings and rankings of docking programs accuracy were prepared and analysed using developed scripts. The LigandFit method performed best in our tests, and its score was used as a reference point for our investigations.

A set of poses scoring functions based on different kinds of mathematical means was designed and implemented to create a method which chooses a consensus pose for each complex among all poses generated by docking programs. The highest accuracy was observed for Arithmetic-harmonic (+5) mean predictions. It performed better than the best performing program: LigandFit in the RMSD-sum test (e.g., 1rdj complexes).

The computational techniques and, by extension, computing power increase very fast. The scoring accuracy and quality of docking applications have improved significantly in the last few years. Nevertheless, there are still too many limitations to base the drug-discovering process exclusively on *in silico* methods. The computational docking could play an important role in rational drug design, but it cannot be applied without *in vitro and in vivo* validations.

ACKNOWLEDGMENTS

This work was supported by EC BioSapiens (LHSG-CT-2003-503265) and OxyGreen (KBBE-2007-212281) 6FP projects, as well as the Polish Ministry of Education and Science (N301 159735, PBZ-MNiI-2/1/2005, and others). The eHiTS software was provided on educational license by SimBioSys Co.

Data: All the data has been obtained from PDBbind Database (www.pdbbind.org/). The site is an international resource with uniform data of biological macromolecules. It includes a collection of experimentally measured binding affinity data, provided exclusively for the protein-ligand complexes available in the Protein Data Bank (PDB). For simulations, the diverse set of 610 well-known complexes was performed. The native structures of the complexes as well as the structure of each molecule separately are available. The chosen proteins differ in their biological activity. The ligand files were downloaded in the MOL2 file format and the protein structures in the PDB file format.

Hardware: Molecular docking simulations were performed partially on KDM computing cluster thanks to the kindness of the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw. The data preparation, following simulations, and result analysis were carried out on an Intel(R) Core(TM)2 Duo CPU, 2.00 GHz with 2 GB RAM.

Graphics: The molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Block, P., Sotriffer, C.A., et al. 2006. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.* 34, D522–D526.

Bohm, H.J. 1992a. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Design* 6, 61–78.

Bohm, H.J. 1992b. LUDI: rule–based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided. Mol. Design* 6, 593–606.

Bohm, H.J. 1994. On the use of LUDI to search the Fine Chemicals Directory for ligands of proteins of known three–dimensional structure. *J. Comput. Aided Mol. Design* 8, 623–32.

Bohm, H.J., Boehringer, M., et al. 2000. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.* 43, 2664–2675.

Eldridge, M.D., Murray, C.W., et al. 1997. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Design* 11, 425–45.

Gohlke, H., Hendlich, M., et al. 2000. Knowledge–based scoring function to predict protein–ligand interactions. *J. Mol. Biol.* 295, 337–356.

Goss, K.U., and Schwarzenbach, R.P. 2003. Rules of thumb for assessing equilibrium partitioning of organic compounds: Successes and pitfalls. *J. Chem. Educ.* 80, 450–455.

Jorgensen, W.L., and Tirado–Rives, J. 2005. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J. Comput. Chem.* 26, 1689–1700.

Katchalski–Katzir, E., Shariv, I., et al. 1992. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA* 89, 2195–2199.

Lodish, H., Berk, A., et al. 2000. Chemical foundations. Covalent bounds and noncovalent interactions, 31–39. *In* Lodish, H., Berk, A., Zipurskyet, L., et al., eds. *Molecular Cell Biology,* 4th ed. W.H. Freeman, New York.

Meng, E.C., Shoichet, B.K., et al. 1992. Automated docking with grid–based energy evaluation. *J. Comput. Chem.* 13, 505–524.

Mitchell, T. 1997. Genetic algorithms, 249–270. *In* Mitchell, T., ed. *Machine Learning*. McGraw Hill, New York.

Ritchie, D.W., and Kemp, G.J. 2000. Protein docking using spherical polar Fourier correlations. *Proteins* 39, 178–194.

Rodwell, V.W. 1999. Water and pH, 28–37. *In* Murray, R.K., Granner, D.K., and Rodwell, V.W., eds. *Harper's Biochemistry.* Appleton & Lange, New York.

Terfloth, L. 2003. Drug design, 597–618. *In* Gasteiger, J., and Engel, T., eds. *Chemoinformatics: A Textbook.* Wiley, New York.

Wermuth, C. 2003. Strategies in the search for new lead compounds or original working hypotheses, 69–88. *In* Wermuth, C.G., ed. *The Practice of Medicinal Chemistry,* 2nd ed. Academic Press, London.

Address correspondence to:
Dr. Dariusz Plewczynski
Interdisciplinary Centre for Mathematical
and Computational Modelling
University of Warsaw
Pawinskiego 5a Street
02–106 Warsaw, Poland

E-mail: darman@icm.edu.pl