Bioinformatics Based Ligand-Docking and in-Silico Screening

Daisuke Takaya, Mayuko Takeda-Shitaka, Genki Terashi, Kazuhiko Kanou, Mitsuo Iwadate, and Hideaki Umeyama*

School of Pharmacy, Kitasato University; 5–9–1 Shirokane, Minato-ku, Tokyo 108–8641, Japan.

Received December 7, 2007; accepted February 20, 2008; published online February 29, 2008

We report a novel method, ChooseLD (CHOOse biological information Semi-Empirically on the Ligand Docking), which uses simulated annealing (SA) based on bioinformatics for protein–ligand flexible docking. The fingerprint alignment score (FPAScore) value is used to determine the docking conformation of the ligand. This method includes the matching of chemical descriptors such as fingerprints (FPs) and the root mean square deviation (rmsd) calculation of the coordinates of atoms of the chemical descriptors. Here, the FPAScore optimization for the translation and rotation of a rigid body is performed using the Metropolis Monte Carlo method. Our ChooseLD method will find wide application in the field of biochemistry and medicine to improve the search for new drugs targeting various proteins implicated in diseases.

Key words flexible docking; in-silico screening; fingerprint; Tanimoto coefficient; simulated annealing; CHOOse biological information Semi-Empirically on the Ligand Docking

Many protein targets implicated in diseases have been discovered through biochemical experiments.[1,2] As a result, the competition between pharmaceutical companies and other organizations to discover drug-like compounds, which inhibit or activate those protein targets, is fierce.[3,4] Experimental screening for drug-like compounds has usually been performed using an industrial robot to determine the interaction of the compound with a drug target protein. Since the cost of such screening is extremely high, in-silico screening of compounds for potential activity against the target protein is becoming popular. Many pharmaceutical companies are using in-silico screening programs, such as DOCK[5] AutoDock[6] and GOLD[7] created by Ewing *et al.*, Goodsel *et al.* and Jones *et al.*, respectively. DOCK 4.0 is a program used for automated molecular docking of flexible molecules, where the intermolecular interaction is described with the non-bonded terms of the AMBER[8] molecular mechanic potential of Lennard Jones, 12-6 dispersion term and 12-10 hydrogen bond term. AutoDock also uses the Lennard Jones 12-6 dispersion and 12-10 hydrogen bond terms. The screening program GOLD is based on a different 8-4 dipersion[7] force field. All three programs, DOCK, AutoDock and GOLD, use classical mechanical potentials. Here, we report a novel method, ChooseLD (CHOOse biological information Semi-Empirically on the Ligand Docking), which uses simulated annealing based on bioinformatics for protein–ligand flexible docking. Our docking method is mainly based on information such as fingerprint (FP) of a chemical descriptor, with some use of available information on the predicted protein structure and X-ray or NMR structures of protein–ligand complexes. Figure 1 shows the schematic diagram of our protein–ligand docking protocol. In the upper left corner of the diagram, we placed a target protein of interest and aim to select one or more ligands with low molecular weight. The query amino acid sequence of the target protein is aligned in a filter with a CE Z-Score of 3.7[9] of the Protein Data Bank (PDB)[10] database, which includes ligand molecules, termed the family ligand set. Alignment methods such as PSI-BLAST[11] are then applied.

A FP of a chemical descriptor is determined on the basis of covalent bonds that the molecule is composed of. FP is

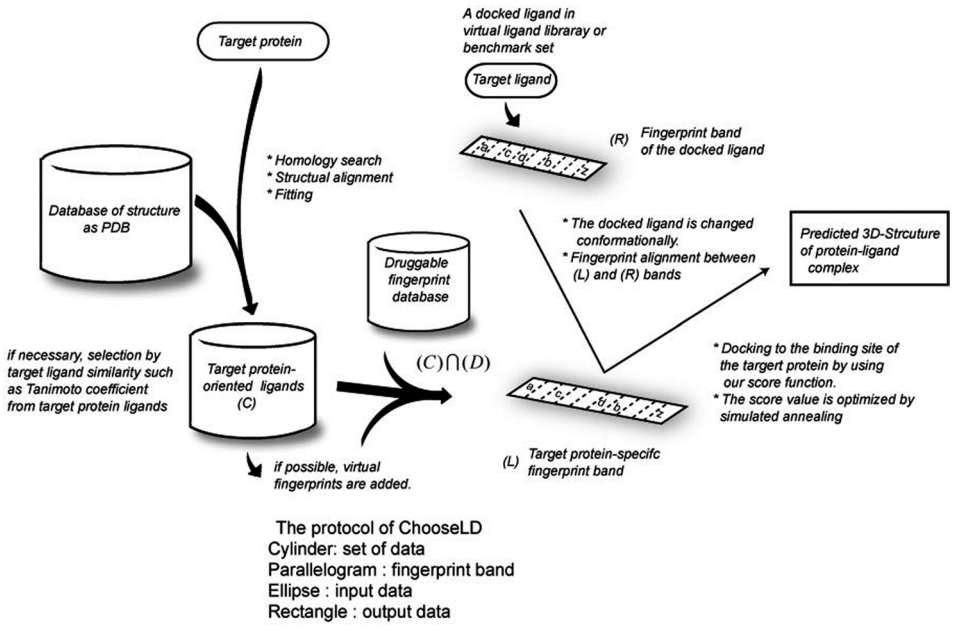

Fig. 1. Scheme Map of ChooseLD Protein–Ligand Docking

Target ligand is docked to the isolated target protein, and the predicted ligand–protein 3D-structure is produced after the fingerprint alignment and the simulated annealing procedures.

* To whom correspondence should be addressed. e-mail: umeyamah@pharm.kitasato-u.ac.jp

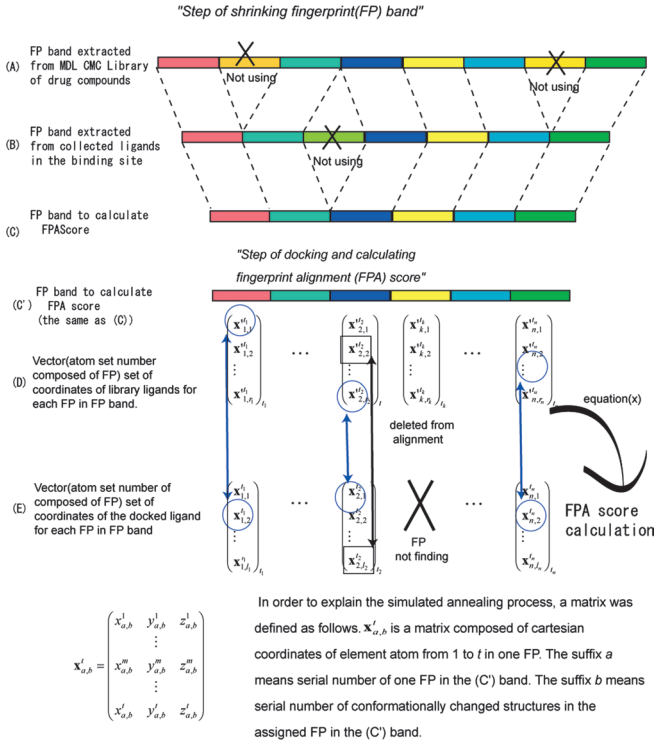

Fig. 2. Steps Used in the Shrinking of the Fingerprint Band

(A) FP band extracted from MDL CMC Library of drug compounds. (B) FP band extracted from collected ligands in the binding site. (C) FP band used to calculate FPAScore. (C′) The same as (C) band. (D) Vector of library ligand coordinates set for the same FP, and vector set for the (C′) band. (E) Vector of docked ligand coordinates set for the same FP, and vector set for the docked ligand FPs. FPs which exist in (A) and (B) bands are assigned to (C) band. FPs are able to have different Cartesian coordinates.

composed of two, three or four atoms as FP composition elements. Moreover, this FP includes information about the atom-type, such as used in SYBYL,[12] and bond-type; single, double, triple or aromatic. A long band composed of multiple FPs corresponding to all FPs extracted from the entire collection of the drug compounds contained in the MDL Comprehensive Medicinal Chemistry (CMC) Library[13] is shown as the (A) band in Fig. 2. A band composed of all nonredundant FPs extracted from the family ligand set is shown as the (B) band. All the FPs of the (A) band and the (B) band are compared, and the unmatched FPs in A and B are deleted to generate band (C). The (C) band is subsequently used to calculate the FP alignment score (FPAScore). In the (B) band, a modified FP may also be added, which can be generated from multiple FPs obtained from several ligands in the family ligand set. The (C′) band is the same as the (C) band. Each FP in the (C′) band corresponds to the assembly of several atoms, each with its own Cartesian coordinates. If the same FP is found in the (B) band in the family ligand set or in the mixing of the ligand set during the extraction process of FPs, this FP is then considered to be redundant or degenerative. However, such redundant FPs still have Cartesian coordinates different from other FPs in the (C′) band. In the (D) band, the FPs of the same type, but with different coordinates in the family library of ligands, compose a Cartesian coordinates vector of the redundant FP set. In the (E) band, the FPs of the same type, including different Cartesian coordinates of the docked ligand, compose another Cartesian coordinates vector of the redundant FP set. The selection of one Cartesian coordinates element from a Cartesian coordinates vector in the (D) band or the (E) band is performed in the simulated annealing process.

$$\begin{aligned} FPAScore &= F(aligned\_fp, fp\_rmsd, molecule) \\ &= BaseScore(aligned\_fp, fp\_rmsd) \\ &\times fp\_volume(molecule) \\ &\times fp\_contact\_surface(molecule) \end{aligned} \quad (1)$$

$$BaseScore(aligned\_fp, fp\_rmsd) = \frac{RawScore(aligned\_fp)}{1+\ln(fp\_rmsd^{k1}+1)} \quad (2)$$

$$fp\_volume(molecule) = \ln\frac{1.0+nafp^{k2}}{1.0+nap^{k3}} \quad (3)$$

$$fp\_contact\_surface(molecule) = \frac{\sum_{i=1}^{n} density\_of\_atom(atom(i))}{total\_density\_of\_atom(molecule)} \quad (4)$$

Fig. 3. The Equation Used to Calculate the FPAScore

Equations 2, 3 and 4 are substituted for three terms in Eq. 1.

Next in the process, the FPAScore is newly defined by Eq. 1 shown in Fig. 3. FPAScore is used to determine the most stable docking conformation of the ligand. The FPAScore value is optimized in the simulated annealing (SA) calculation carried out 1000 times. A single SA process is carried out by changing the molecular position of the ligand ten thousand times without changing its conformation. This method includes the matching of chemical descriptors and the root mean square deviation (rmsd) calculation of the coordinates of atoms of the chemical descriptors. The FPAScore optimization for the translation and rotation of a rigid body is performed using the Metropolis Monte Carlo method. This optimization process is performed one thousand times to obtain the maximum FPAScore, which determines the most stable ligand conformation. The value of RawScore(aliged_fp) in the Eq. 2 is maximized in the SA calculation process accompanied by variation of the value of $\ln(fp\_rmsd^{k1}+1.0)$. The fp_rmsd is the rmsd value that is the result of the least-square fitting using the FP alignment. The nafp in the Eq. 3 is the number of docking ligand atoms covering the FP region. The nap is the number of docking ligand atoms covering the target protein region. In the calculation of the benchmark sets, both k2 and k3 are equal to one. Atom(i) in the Eq. 4 is the sequential atom number of a ligand. Density_of_atom(atom(i)) shows the ligand atom number in the area of direct interaction with the FP. Total_density_of_atom(molecule) is the denominator for the ligand and is used to standardize the numerator. The meanings of each of the Eqs. 2, 3 and 4 that make up Eq. 1 may correspond to a particular physicochemical property. Equation 2 may reflect the stability of Gibbs free energy, Eq. 3—the occupancy of the docking-ligand in the binding site and Eq. 4 may describe the contact ratio of the docking-ligand to the binding site.

Results and Discussion

In order to test the protein–ligand docking method based upon bioinformatics, two benchmark tests were performed by

(a)

Tc Range	k1						failed PDBID
	1.0	2.0	3.0	4.0	5.0	6.0	
	success rate(%)						
0.56 – 0.08	46.0	54.6	57.6	58.9	56.3	58.6	1V4S, 1G9V
0.76 – 0.08	50.0	61.0	60.7	62.1	59.4	62.1	1V4S
0.96 – 0.08	55.6	62.1	64.4	65.2	65.8	64.8	1V4S
average	50.5	59.2	60.9	62.1	60.5	61.8	

(b)

Docking soft	success rate (%)		
	Corina	MINI	average
DOCK	21.6	20.6	21.1
AutoDock	26.2	27.0	26.6
GOLD ChemScoreSTD	45.5	45.3	45.4
GOLD GOLDScoreLib	44.1	44.9	44.5
GOLD GOLDScoreSTD	45.2	46.7	46.0
	success rate (%)		
ChooseLD TC 0.56 – 0.08			40.1
ChooseLD TC 0.76 – 0.08			44.8
ChooseLD TC 0.96 – 0.08			46.4

Fig. 4. Success Rates for Two Benchmark Sets for Tanimoto Coefficient (Tc) Ranges, 0.56—0.08, 0.76—0.08 and 0.96—0.08

(a) 85 benchmark set. The k1 value in Eq. 2 in Fig. 3 was varied from 1.0 to 6.0. Docking calculation of 83, 84 and 84 protein targets succeed in above three ranges, respectively. Since family ligand sets were not found, the ligand-docking results were not obtained for 1V4S and 1G9V. (b) 133 benchmark set. The success rate of ChooseLD is compared with DOCK, AutoDock and GOLD. Corina[17] and MINI[15] show the method to determine initial ligand conformation. ChooseLD uses the furthest conformation from the experimental conformation. 116 protein targets were used in the 133 PDB targets.[15]

using two database sets composed of either 85[14] or 133[15] PDB structures. Each PDB code in the 85 benchmark includes a druggable protein and a drug-like ligand.[14] On the other hand, each PDB code in the 133 benchmark set, which is completely different from the 85 benchmark set, includes a druggable protein and generic compound as a ligand. After the ligand molecule was docked to the target protein in the two benchmark sets, the rmsd value of Cartesian coordinates between the docked ligand and the X-ray analyzed ligand was calculated. If the rmsd between the predicted and experimental results is equal or less than 2.0 Å, the docked conformation is considered to be acceptable or close to the experimental value.[7] If the docking state is within 2.0 Å, docking is considered successful. Using the 85 benchmark set, the constant k1 value was optimized to be 4.0 by varying it from 1.0 to 6.0 as shown in Fig. 4a. Three regions of 0.56—0.08, 0.76—0.08 and 0.96—0.08 of Tanimoto coefficient[16] (Tc) were used for calculating the success rate of docking using this test set. Each of the FPs of the docking-ligand is sequentially compared with one of the FPs of a single ligand in the family ligand set. The ratio of FP identity determines the Tc. In this set, the FPAScore calculation for each target protein was performed ten times. When the optimized k1 value was used, the average success rate for 85 targets was 62.1%. For the k1＝4.0 value, the success rates of Tc ranges 0.56—0.08, 0.76—0.08 and 0.96—0.08 were 58.9, 62.1 and 65.2%, respectively. If the Tc values using chemical descriptors are 0.56, 0.76 and 0.96, the test ligand's interaction with each protein in comparison with the docked ligand, is regarded to be close, very close and almost identical, respectively. In the docking calculations carried out for the 133 benchmark set using the k1 value of 4.0, the success rates of Tc ranges 0.56—0.08, 0.76—0.08 and 0.96—0.08 were 40.1, 44.8 and 46.4%, respectively (Fig. 4b).

The FPAScore calculation for one target protein was also performed ten times. Compared with the success rate using programs DOCK, AutoDock and GOLD,[15] shown in the table of Fig. 4b, the success rates of our ChooseLD program are almost equivalent to those of the docking program, GOLD. Our program is mainly based upon a bioinformatics basis set called FP using new Eq. 1, while the GOLD program is based upon the classical mechanics potentials including information such as the number and type of hydrogen bonds between the target protein and the interacting ligand. Thus, direct comparison of the success rates of the two programs may be meaningless. Nevertheless, our program is comparably more powerful than other availabe programs, when the researchers want to carry out protein–ligand docking and in-silico screening of a target protein. In the future, the number of PDB codes with the interacting ligand will be increased. It is anticipated that such an increase will improve the success rate of our ChooseLD method. We believe that our ChooseLD method will find wide application in the field of biochemistry and medicine to improve the search for new drugs targeting various proteins implicated in diseases.

Experimental

Docking calculations were performed using 200 CPUs with Linux clusters. These CPUs consist of Pentium4, Core2Duo and Opteron, which have various clock frequencies. Depending on the number of atoms of a docked ligand or a library ligand, one docking cost time is about 1.5—27 min for one CPU.

References

1) Wood E. R., Truesdale A. T., McDonald O. B., Yuan D., Hassell A., Dickerson S. H., *Cancer Res.*, **64**, 6652—6659 (2004).
2) Stamos J., Sliwkowski M. X., Eigenbrot C., *J. Biol. Chem.*, **277**, 46265—46272 (2002).
3) Nakamura K., Yamamoto A., Kamishohara M., Takahashi K., Taguchi E., Miura T., Kubo K., Shibuya M., Isoe T., *Mol. Cancer Ther.*, **3**, 1639—1649 (2004).
4) Nakamura K., Taguchi E., Miura T., Yamamoto A., Takahashi K., Bichat F., Guilbaud N., Hasegawa K., Kubo K., Fujiwara Y., Suzuki R., Kubo K., Shibuya M., Isoe T., *Cancer Res.*, **66**, 9134—9142 (2006).
5) Ewing T. J. A., Makino S., Skillman A. G., Kuntz I. D., *J. Comput. Aided. Mol. Des.*, **15**, 411—428 (2001).
6) Morris G. M., Goodsell D. S., Halliday R. S., Huey R., Hart W. E., Belew R. K., Olson A. J., *J. Comput. Chem.*, **19**, 1639—1662 (1998).
7) Jones G., Willett P., Glen R. C., Leach A. R., Taylor R., *J. Mol. Biol.*, **267**, 727—748 (1997).
8) Case D. A., Cheatham T. E., III, Darden T., Gohlke H., Luo R., Merz K. M., Jr., *J. Comput. Chem.*, **26**, 1668—1688 (2005).
9) Shindyalov I. N., Bourne, P. E., *Protein Engineering*, **11**, 739—747 (1998).
10) Westbrook J., Feng Z., Chen L., Yang H., Berman H. M., *Nucleic Acids Res.*, **31**, 489—491 (2003).
11) Altsuchul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J., *Nucleic Acids Res.*, **27**, 3398—3402 (1997).
12) Tripos Inc., South Hanley Road, St. Louis, MO 63144-2913, U.S.A. (1699) (http://www.tripos.com).
13) Symyx Technologies, Inc., Corporate Address: 3100 Central Expressway, Santa Clara, CA 95051.
14) Hartshorn M. J., Verdonk M. L., Chessari G., Brewerton S. C., Mooij W. T. M., *J. Med. Chem.*, **50**, 726—741 (2007).
15) Onodera K., Satou K., Hirota H., *J. Chem. Inf. Model.*, **47**, 1609—1618 (2007).
16) Godden J. W., Xue L., Bajorath J., *J. Chem. Inf. Comput. Sci.*, **40**, 163—166 (2000).
17) Sadowski J., Gasteiger J., *J. Chem. Inf. Comput. Sci.*, **34**, 1000—1008 (1994).