J Mol Model (2010) 16:903–913
DOI 10.1007/s00894-009-0598-7

ORIGINAL PAPER

A fast protein-ligand docking algorithm based on hydrogen bond matching and surface shape complementarity

Wenjia Luo · Jianfeng Pei · Yushan Zhu

Received: 23 July 2009 / Accepted: 19 September 2009 / Published online: 13 October 2009

Abstract With the rapid development of structural determination of target proteins for human diseases, high throughout virtual screening based drug discovery is gaining popularity gradually. In this paper, a fast docking algorithm (H-DOCK) based on hydrogen bond matching and surface shape complementarity was developed. In H-DOCK, firstly a divide-and-conquer strategy based enumeration approach is applied to rank the intermolecular modes between protein and ligand by maximizing their hydrogen bonds matching, then each docked conformation of the ligand is calculated according to the matched hydrogen bonding geometry, finally a simple but effective scoring function reflecting mainly the van der Waals interaction is used to evaluate the docked conformations of the ligand. H-DOCK is tested for rigid ligand docking and flexible one, the latter is implemented by repeating rigid docking for multiple conformations of a small molecule and ranking all together. For rigid ligands, H-DOCK was tested on a set of 271 complexes where there is at least one intermolecular hydrogen bond, and H-DOCK achieved success rate (RMSD<2.0 Å) of 91.1%. For flexible ligands, H-DOCK was tested on another set of 93 complexes, where each case was a conformation ensemble containing native ligand conformation as well as 100 decoy ones generated by AutoDock [1], and the success rate reached 81.7%. The high success rate of H-DOCK indicates that the hydrogen bonding and steric hindrance can grasp the key interaction between protein and ligand. H-DOCK is quite efficient compared with the conventional docking algorithms, and it takes only about 0.14 seconds for a rigid ligand docking and about 8.25 seconds for a flexible one on average. According to the preliminary docking results, it implies that H-DOCK can be potentially used for large scale virtual screening as a pre-filter for a more accurate but less efficient docking algorithm.

Keywords Combinatorial optimization · Docking · Hydrogen bond matching · Virtual screening

W. Luo · Y. Zhu (✉)
Department of Chemical Engineering, Tsinghua University, Beijing 100084, People's Republic of China
e-mail: yszhu@tsinghua.edu.cn

J. Pei
State Key Laboratory for Structural Chemistry of Stable and Unstable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing, People's Republic of China

J. Pei
Center for Theoretical Biology, Peking University, Beijing, People's Republic of China

Introduction

The docking problem of protein and ligand, which refers to the prediction of the interaction between a macromolecule mainly protein and a small molecular target, arises in many molecular recognition based applications such as drug discovery, receptor and enzyme design. Given the structures of both the protein and the ligand, a reliable evaluation rule for a docking algorithm is its ability to find the experimental binding conformation of the ligand, usually 2.0 Å of root-mean-square deviation (RMSD) between the docked conformation of the ligand and its experimental one is used as the threshold to test different docking techniques, based on a large and carefully constructed set of protein-ligand complexes. At present, docking has been widely researched and is currently in a stage of rapid development [2, 3]. Many docking methods and commercial software programs

are now available and are in wide use. In order to predict the true protein-ligand binding mode, a typical docking protocol often uses a searching algorithm to sample a sufficiently large ensemble of binding modes, followed by a scoring function to guide the searching procedure to pick the true mode from the ensemble. The searching algorithms must take into consideration the degrees of freedom of translation and rotation of the ligand; furthermore, modern docking routines generally treat the ligand as a flexible molecule.

Existing searching algorithms can be loosely categorized into three basic types: random or stochastic methods, systematic methods, and simulation methods. Random methods sample the conformational spaces by performing changes to the ligand at each step. The changes are then accepted or rejected based on a predefined probability function. Based on the random algorithms, this method can be further classified into three types: Genetic algorithm methods, examples of which include AutoDock [1], GOLD [4] and DARWIN [5]; Monte Carlo methods, examples of which include Prodock [6], ICM [7], MCDOCK [8], DockVision [9], and QXP [10]; and Tabu search methods, which promote efficiency by preventing revisiting of already explored conformational space. PRO_LEADS [11] is an example that uses a Tabu search algorithm. Systematic methods explore the conformational space in a combinatorial way. All rotatable bonds in the ligand are rotated through 360° using a given increment to generate all possible combinations. To prevent the computation from becoming intractable due to a combinatorial explosion, one commonly used strategy is to adopt fragmentation methods, which dock parts instead of the whole ligand into the active site, and then join those parts together at a later time to get a feasible docking pose. Another strategy is the use of database methods, which explore a library of pregenerated ligand conformations. Each conformation in the library can be treated as a rigid body during the docking process. Examples of systematic methods include DOCK [12], LUDI [13], FlexX [14], ADAM [15], HammerHead [16] and FLOG [17]. Simulation methods usually include molecular dynamics and energy minimization. These methods have the drawback of being trapped in local energy minima, and are typically not used as stand-alone search techniques in an actual docking exercise. Instead, several other docking algorithms will use a simulation method as a complement. One example is DOCK [12], which performs an energy minimization calculation after each fragment addition.

Along with the searching methods comes the need for scoring functions, which are used to rank the docking poses generated in the searching process. A successful scoring function must be rigorous enough to rank the native docking pose in the front of the generated list, while it cannot be too computationally expensive, for the sake of efficiency. Current scoring functions in wide use include those that are force-field based, like CHARMM [18], AMBER [19], G-Score [20], and GoldScore [21], and those which are empirical or knowledge-based, like F-Score [14], SCORE [22], and X-Score [23], to name just a few examples of each kind.

For high throughput virtual screening, a large library of molecules needs to be screened for the discovery of new drugs, where the efficiency of the docking algorithm determines the capability of screening. Although the aforementioned successful docking algorithms can achieve high accuracy in predicting binding conformation, it predictably will take 10–20 minutes for a typical flexible docking and multiple runs will be required if higher accuracy is needed. LigandFit [24] and LibDock [25] are two examples of docking programs whose efficiency is tailored for virtual screening in large libraries. There is always a trade-off between efficiency and accuracy for any docking algorithm, as the algorithm that takes a larger conformational sampling set and more elaborate scoring functions will tend to be more accurate, but also less efficient.

In this paper, a fast docking algorithm (H-DOCK) was developed in hopes of achieving high efficiency as well as comparable accuracy by considering two dominant interactions between protein and ligand, i.e., hydrogen bonding and van der Waals interactions. The principle of the docking procedure in H-DOCK is to maximize the intermolecular hydrogen bonding and to avoid large steric hindrance between protein and ligand, the former is implemented by using a divide-and-conquer based combinatorial search, while the latter is implemented by using a simple and effective scoring function comprising mainly the van der Waals interaction between protein and ligand. It should be noted that Meyer et al. [26] first proposed hydrogen bond formation for protein-protein docking, and we have applied a similar geometrical hydrogen bond matching for protein-ligand docking in this paper. The searching algorithm in H-DOCK is deterministic, so multiple runs are not needed. Although the ligand in H-DOCK is assumed to be rigid, it can be extended into cases where the ligand is flexible by taking an ensemble of multiple conformations of ligand, and for each one of these conformations, the docking computation is carried out just as the ligand is rigid. As shown by the docking results in the latter sections, H-DOCK can achieve fast docking efficiency with comparable accuracy either for rigid docking or flexible one, these results imply that H-DOCK can be used as a stand-alone docking program or mainly as a companion filter for a conventional docking program, such as the PSI-DOCK [27], while for large database based virtual screening for drug discovery.

Materials and methods

The validation testing set

One of the most straightforward ways to validate a docking algorithm is to reproduce the structure of experimentally determined protein-ligand complexes. In 2002, Nissink et al. collected a large set of 305 protein-ligand complexes, i.e., the CCDC/Astex set [28], for the specific purpose of validating algorithms that rely on the prediction of protein-ligand interactions. These 305 complexes are distributed among different protein families and have diverse ligand structures, which makes it a good test set for docking programs. Because H-DOCK is based on hydrogen bonding, an initial structural survey of the 305 complexes was carried out to determine the number of intermolecular hydrogen bonds between the protein and ligand, using the criteria similar to that of Baker and Hubbard [29], i.e., the distances H…A and D…A must be less than 2.5 Å and 3.9 Å, respectively, and the angle ∠D-H...A must exceed 90°, no other restrictions are imposed. 271 out of the 305 complexes have at least one hydrogen bond between the protein and the ligand, and they have been selected as the test set for H-DOCK. The Protein Data Bank (PDB) names of these 271 testing cases and the number of intermolecular hydrogen bonds are shown in Table 1. All the data of the CCDC/Astex set were downloaded from http://www.ccdc.cam.ac.uk. Coordinates of hydrogen atoms on both protein and the ligand were also provided in these complexes, which were used in H-DOCK. According to Nissink et al. [28], coordinates of hydrogen atoms on both the protein and ligand are added using the SYBYL [30] software, but the orientations of rotatable OH and NH3 groups are not optimized. It indicates that the hydrogen coordinates are calculated only from atom types and hybrid states, but do not contain information regarding the intermolecular hydrogen bond network. Optimized ligands and all water molecules provided by the CCDC/Astex complexes set were not used in H-DOCK.

The complexes shown in Table 1 were used to evaluate the rigid docking efficiency of H-DOCK. For flexible docking, H-DOCK was tested on a set where each ligand had multiple conformations, which is collected by Wang et al. [31]. The test set of Wang et al. [31] contains 100 protein-ligand complexes, which come from 43 different types of proteins. Coordinates of hydrogen atoms are also calculated by using SYBYL [30]. Similarly, a structural survey of intermolecular hydrogen bonds was carried out and a subset of 93 out of the 100 complexes was selected as the final testing set since each testing case has at least one hydrogen bond between protein and ligand. The PDB names of these 93 complexes and the number of intermolecular hydrogen bonds are presented in Table 2. For each one of these complexes, an ensemble of docked conformations for the ligand molecule was generated by AutoDock [1] program. In our testing computations, the ensemble of conformations also contained the native one. It is worth noting that the decoy conformations are the top ranked docked results according to the scoring functions used by AutoDock [1], instead of randomly generated conformations [24] which usually cannot achieve high affinity with the protein. Thus it brings a great challenge for H-DOCK to identify the native docked conformation from the decoy ones.

Identification of potential hydrogen bonding sites

On the protein, the region which is within 4 Å from the ligand in the experimentally determined docked complex is defined as active region. Only the potential sites on the ligand and those in the active region of the protein will be taken into account. An identification rule similar to that of Meyer et al. [26] is used in this paper. For protein, donors include main-chain N-H, His NE2, His ND1, Lys NZ, Asn ND2, Gln NE2, Arg NE, Arg NH1, Arg NH2, Ser OG, Thr OG1, Tyr OH, Trp NE1; acceptors include main-chain C=O, Asp OD1, Asp OD2, Glu OE1, Glu OE2, Asn OD1, Gln OE1, Ser OG, Thr OG1. Alpha carbon atoms, aromatic ring acceptors, sulfur atoms and nitrogen atom acceptors are relatively weak and are not included. For the ligand, the identification is based on the atom′s hybrid state and chemical environment. Nitrogen atoms bonded with hydrogen and sp3 hybridized oxygen atoms are treated as donors, all forms of oxygen atoms are treated as acceptors, therefore it is possible for an oxygen atom to be a donor and an acceptor simultaneously, e. g., a hydroxyl oxygen is considered to be both a donor and an acceptor.

In H-DOCK, the coordinates of hydrogen atoms are required for all donor sites, and fortunately those coordinates are provided in the data of the testing set. Note that every hydrogen atom bonded with those heavy donor atoms corresponds to one potential donor site, e. g., the two hydrogen atoms bonded with ND2 of Gln each corresponds to a potential donor site.

The combinatorial method for hydrogen bond matching

Considering the geometry of a hydrogen bond D-H…A, it is required that when a hydrogen bond is formed, both the D…A distance and the ∠D-H…A angle keep in a certain range. Typically the D…A distance ranges from 2.5–3.5 Å and the D-H-A ranges from 90°to 180°. In the most ideal case, the ∠D-H…A angle will be exactly 180°and the D…A distance will be the energetically optimal bond length. In H-DOCK, the position of the acceptor atom in the ideal case is defined as the “ideal acceptor position”. In practice

Table 1 The PDB names of test set for rigid docking, and the number in the parenthesis implies the count of identified hydrogen bonds between protein and ligand

1a07(2)	1a0q(5)	1a1b(4)	1a1e(4)	1a28(1)	1a42(3)	1a4g(7)	1a4k(2)	1a4q(5)	1a6w(3)
1aaq(10)	1abe(5)	1abf(5)	1acj(1)	1acm(9)	1aco(6)	1aec(6)	1aha(2)	1ai5(2)	1aj7(2)
1ake(17)	1aoe(4)	1apt(9)	1apu(5)	1aqw(7)	1ase(6)	1atl(2)	1azm(1)	1b58(12)	1b59(1)
1b6n(5)	1b9v(4)	1baf(2)	1bbp(2)	1bgo(3)	1blh(5)	1bma(3)	1bmq(6)	1byb(6)	1byg(3)
1c1e(1)	1c2t(8)	1c5c(7)	1c5x(4)	1c83(5)	1cbs(1)	1cbx(4)	1cdg(3)	1cf8(3)	1cil(2)
1cin(2)	1ckp(1)	1com(5)	1coy(3)	1cps(2)	1cqp(1)	1ctt(4)	1cvu(1)	1d0l(7)	1d3h(1)
1d4p(3)	1dbb(2)	1dbj(1)	1dbm(2)	1dd7(1)	1dg5(3)	1dhf(5)	1did(3)	1die(3)	1dmp(3)
1dog(3)	1dr1(2)	1dwb(3)	1dwc(5)	1dwd(4)	1dy9(6)	1eap(4)	1ebg(5)	1eed(9)	1ei1(11)
1ejn(5)	1ela(4)	1elb(2)	1elc(1)	1eld(3)	1ele(3)	1eoc(2)	1epo(6)	1eta(1)	1etr(4)
1ets(4)	1ett(4)	1etz(2)	1f0r(2)	1f0s(2)	1f3d(2)	1fax(5)	1fbl(6)	1fgi(3)	1fkg(2)
1fki(2)	1flr(2)	1frp(10)	1ghb(4)	1glp(8)	1glq(7)	1gpy(6)	1hak(1)	1hdc(1)	1hef(7)
1hfc(6)	1hiv(8)	1hos(8)	1hpv(3)	1hsb(5)	1hsl(6)	1htf(3)	1hti(5)	1hvr(1)	1hyt(3)
1ibg(4)	1ida(5)	1imb(4)	1ivb(3)	1ivc(5)	1ivd(3)	1ive(2)	1ivq(7)	1jao(4)	1jap(3)
1kel(4)	1kno(2)	1lah(7)	1lcp(3)	1ldm(4)	1lic(1)	1lmo(5)	1lna(5)	1 lpm(2)	1lst(6)
1lyb(9)	1lyl(4)	1mcq(1)	1mcr(1)	1mdr(4)	1 ml1(6)	1mld(6)	1mmb(5)	1mmq(6)	1mnc(5)
1mrg(1)	1mrk(2)	1mts(4)	1mtw(5)	1nco(2)	1ngp(3)	1nis(6)	1nsd(4)	1okl(1)	1okm(1)
1pbd(4)	1pdz(5)	1pgp(5)	1poc(3)	1ppc(6)	1pph(7)	1ppi(11)	1ppl(7)	1pso(10)	1ptv(3)
1qbr(5)	1qbt(7)	1qbu(3)	1qcf(2)	1qh7(4)	1qpe(2)	1qpq(4)	1rds(9)	1rne(9)	1rnt(5)
1rob(5)	1rt2(1)	1sln(5)	1slt(4)	1snc(7)	1srf(2)	1srg(3)	1srh(3)	1srj(3)	1 stp(5)
1tdb(4)	1tka(5)	1tlp(5)	1tmn(6)	1tng(3)	1tnh(3)	1tni(2)	1tnl(2)	1tph(5)	1tpp(6)
1trk(6)	1tyl(3)	1ukz(6)	1ulb(3)	1uvs(3)	1uvt(2)	1vgc(3)	1vrh(2)	1wap(6)	1xid(4)
1xie(3)	1xkb(4)	1yds(2)	1ydt(2)	1yee(3)	2aad(7)	2ada(4)	2ak3(4)	2cgr(4)	2cht(4)
2cmd(5)	2ctc(2)	2dbl(3)	2er7(10)	2fox(9)	2gbp(7)	2h4n(2)	2ifb(2)	2lgs(6)	2mcp(3)
2mip(8)	2pcp(1)	2phh(4)	2pk4(4)	2plv(2)	2qwk(5)	2sim(5)	2tmn(3)	2tsc(3)	2yhx(6)
2ypi(5)	3cpa(4)	3erd(2)	3ert(1)	3gpb(9)	3mth(2)	3nos(5)	3pgh(2)	3ptb(4)	3tpi(5)
4aah(7)	4cox(1)	4cts(4)	4dfr(8)	4er2(8)	4est(5)	4fab(2)	4fbp(6)	4lbd(3)	4phv(6)
4tpi(5)	5abp(5)	5er1(6)	5p2p(3)	6abp(7)	6cpa(6)	6rnt(4)	6rsa(6)	7cpa(5)	7tim(6)
8gch(5)									

this position is defined to be collinear with the D, H atom and 2.8 Å away from the D atom. Although the acceptor atoms are rarely found to occur exactly at the ideal position, they do appear near the ideal position. Typically there are 10–30 potential hydrogen bonding sites on both the ligand and the protein. In this paper, a specific intermolecular hydrogen bonding relation is referred to as a bonding "mode".

Assume that there are *m* hydrogen bonding suites on the ligand and *n* on the protein, then in total there are *m!n!*

Table 2 The PDB names of test set for flexible docking, and the number in the parenthesis implies the count of identified hydrogen bonds between protein and ligand

1a46(7)	1a5g(8)	1abe(8)	1abf(8)	1adb(9)	1add(3)	1af2(4)	1apb(8)	1apt(8)	1apw(8)
1b5g(7)	1ba8(7)	1bap(8)	1bb0(7)	1bbz(3)	1bcu(1)	1bhf(4)	1bra(2)	1bxo(2)	1bzm(1)
1cbx(4)	1dhf(5)	1dr1(3)	1drf(5)	1e96(12)	1ela(4)	1etr(5)	1ets(5)	1fkb(3)	1fkf(3)
1fmo(3)	1hsl(6)	1hvr(2)	1inc(3)	1mnc(5)	1ppc(6)	1pph(7)	1rgk(3)	1rgl(2)	1rnt(5)
1sre(2)	1tet(3)	1tha(3)	1tlp(5)	1tmn(5)	1tng(3)	1tnh(3)	1tni(2)	1tnj(3)	1tnk(2)
1tnl(2)	1yyy(8)	1zzz(8)	2ak3(4)	2cgr(4)	2csc(4)	2ctc(2)	2gbp(7)	2pk4(4)	2qwb(7)
2qwc(5)	2qwd(7)	2qwe(7)	2qwf(6)	2qwg(5)	2sns(4)	2tmn(2)	2xim(2)	2xis(2)	3cpa(4)
3fx2(7)	3ptb(4)	3tmn(6)	4sga(5)	4tim(6)	4tln(3)	4xia(3)	5abp(8)	5cna(4)	5p21(10)
5sga(5)	5tln(4)	6abp(8)	6rnt(3)	6tim(5)	7abp(8)	7est(3)	7tim(6)	7tln(2)	8abp(8)
8xia(2)	9aat(5)	9abp(7)							

possible modes. This number grows exponentially as the number of potential bonding sites increases, so an exhaustive search is not tractable. Fortunately, there is a simple geometric criterion which can be used to exclude most of the impossible bonding modes. Figure 1 shows the situation where there are two potential donor sites on the ligand and two potential acceptor sites on the protein. The two donor sites are very close while the two acceptor sites are far apart. If it is required that when the hydrogen bond is formed, the acceptor atom should be within a certain distance from its "ideal acceptor position", the situation in Fig. 1 shows that hydrogen bond between site 1 and 1′ cannot coexist with hydrogen bond between site 2 and 2′ because it is impossible to put the two acceptor sites within these two spherical areas simultaneously. Here, the condition, i.e., | distance (1, 2) – distance(1′, 2′) |<2D, is the criterion which can be used to exclude the possibility that bond 1–1′ can coexist with bond 2–2′. Here D is the maximum allowed distance between the acceptor atom and its "ideal acceptor position" when a hydrogen bond is formed. D is a key parameter in the process of computation and balances the efficiency and accuracy of the algorithm. In H-DOCK, parameter D is set to be 1.0 Å. Note that in the process of the enumeration, the position of ligand relative to the protein is still unknown, so it is impossible to compute the distance between one site on the protein and the other on the ligand. However it is always possible to compute the distance between two sites which are both on one side, i.e., either on the ligand or on the protein, since this type of distance is determined only by the molecular structure when the molecule is treated as a rigid body. This criterion enables us to perform the exhaustive enumeration process quite efficiently when a divide-and-conquer scheme is applied.

Suppose that bonding sites on protein are indexed by 1, 2 … m, and sites on ligand indexed by 1′, 2′ … n′. At first no bond is formed. Considering the possible bonding status of site 1, it can then be bonded to site 1′, 2′… n′ or it can form no bond at all. Given that each one of these assumed bonding states of site 1 is true, the algorithm further enumerates the bonding status of site 2, then given that each one of the assumed bonding states of site 1 and 2 is true, it considers the bonding status of site 3, 4 … m. As shown in Fig. 2, this is essentially a tree-like exhaustive searching process that enumerates all bonding possibilities. However, the aforementioned geometric criterion can greatly speed up the search. For example, Fig. 2 shows a case where m=n=3. Suppose bond 1–1′ and 2–3′ cannot coexist due to geometric constraints. The entire sub branch below node "2–3" of the tree (surrounded by dashed line in Fig. 2) can then be cut off from the search.

This combinatorial approach and the cutoff criterion has made some improvement over the existing matching methods, such as that applied in LibDock [25], where hydrogen bonding sites as well as apolar sites are defined as "hot spots" and the matching of exact three hot spots between the protein and the ligand are searched. By considering only hydrogen bonding sites and using the geometrical criterion about the "ideal acceptor position", H-DOCK can carry out an exhaustive enumeration process quite efficiently, and is not restricted on the number of matching sites, such as the three adopted in LibDock [25].

When the enumeration process ends, many possible bonding modes are proposed, there can be up to 10^5 modes. These modes are sorted by the number of hypothetically formed hydrogen bonds in the mode in descending order, and only a small fraction of those proposals which are ranked on the top of the list are sent to the next step in the algorithm. The magnitude of the fraction is another parameter of the algorithm. In H-DOCK, the maximum number of hypothetically formed hydrogen bonds in all modes is recorded firstly, and the modes whose number of bonds is less than the maximum one by three or more are discarded, typically 10^4 modes will still remain. For each remaining mode, the placement of the ligand will be calculated according to the method described in the following section. And such a placement of the ligand, or

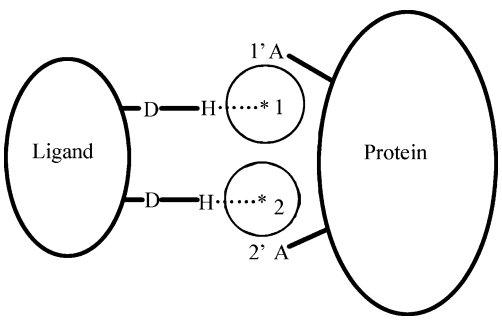

Fig. 1 A situation in which two hydrogen bonds cannot coexist

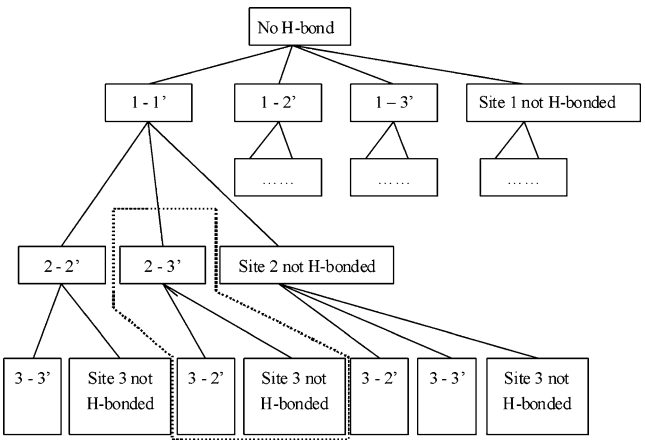

Fig. 2 Part of the searching tree of the hydrogen bond matching when m=n=3. The sub branch surrounded by the dashed line can be cut off if bond 1–1′ and 2–3′ cannot coexist due to geometric constraints

a collection of placements if only one or two hydrogen bonds are assumed to be formed in that mode, becomes one of the candidates for further filtering.

Displacement of the ligand

After the combinatorial search, the hydrogen bonding match between protein and ligand for a mode is known, then the ligand position can be located by minimizing the sum of distances between the acceptors and their ideal positions, as described in Fig. 3. Mathematically, the potential hydrogen bonding sites on the protein form one set of points, while the sites on the ligand form another. For an acceptor site, the coordinate of the point is the same as that of the acceptor atom, while for a donor site, the coordinate of the point is the "ideal acceptor position" corresponding with the donor site. The set of points on the protein is denoted as $\boldsymbol{p}=\{\boldsymbol{p}_1,\boldsymbol{p}_2,...\boldsymbol{p}_n\}$, and the set of points on the ligand is denoted as $\boldsymbol{q}=\{\boldsymbol{q}_1,\boldsymbol{q}_2,...\boldsymbol{q}_n\}$,. Assume that $\boldsymbol{P}_1$ should bonded to $\boldsymbol{q}_1$, $\boldsymbol{p}_2$ bonded to $\boldsymbol{q}_2$, and so on. The problem turns out to be the determination of the amount of rotation and translation of the set q, *i.e.*, the displacement of the ligand, here the ligand is treated as a rigid body, in order to make every pair of points as closely as possible. The optimization problem for ligand positioning can be formulated as,

$$\min_{R,T} \omega = \sum_{i=1}^{n} (p_i - Rq_i - T)^T (p_i - Rq_i - T) \qquad (1)$$

where, R is a 3 by 3 matrix representing the rotation of the ligand, and **T** is a three-dimensional vector standing for the translation. This is sometimes called "absolute orientation" problem and has been widely researched [32]. However, to find the exact solution of problem (1) is too slow no matter which numerical or analytical method is used, where typically 10^6 problems of (1) need to be solved for a flexible ligand docking case. An approximation approach is applied for the sake of efficiency. Firstly, the ligand is translated to make the gravity centers of point set $\{\mathbf{q}_i\}$ and $\{\mathbf{p}_i\}$ to be superimposed. Then the ligand is rotated around the gravity center to make the gravity center, p_1 and q_1 to be collinear and $\mathbf{p}_1$ and $\mathbf{q}_1$ are on the same side. Lastly, the ligand is rotated around the axis formed by the gravity center, $\mathbf{p}_1$ and $\mathbf{q}_1$ to make the remaining points in $\{\mathbf{q}_i\}$ to be close to their corresponding points in $\{\mathbf{p}_i\}$ as much as possible. The role of the optimization problem (1) can be illustrated more explicitly by an example in Fig. 4, where two hydrogen bonds can match simultaneously. There are two potential donor sites, i.e., N1–H1 and N2–H2, on molecule A, and two potential acceptor sites, i.e., O1 and O2, on molecule B. The "ideal acceptor positions" of these two donor sites are shown in asterisks. An extreme case is shown in Fig. 4(a), where the acceptor site O1 is placed exactly on the ideal acceptor position of donor site N1–H1, then the hydrogen bond N1–H1…O1 is perfectly formed, but hydrogen bond N2–H2…O2 is poorly formed because of the small ∠N2–H2...O2 angle. Figure 4(b) shows just the contrary to Fig. 4(a), where O2 is exactly on the ideal acceptor position of N2–H2, but O1 is far from the ideal acceptor position of N1–H1. Figure 4(c) shows the result of the optimization, which is in fact a tradeoff between schemes (a) and (b), where the sum of squares of distances between the acceptors sites and their ideal positions are

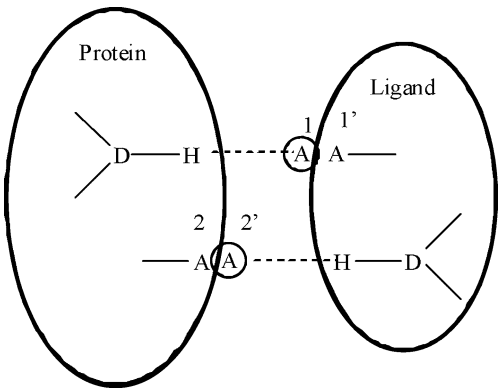

Fig. 3 The displacement of the ligand. As an example, there are one donor and one acceptor site on both the protein and the ligand. For each donor site, its "ideal acceptor position" is shown as 'A' in a circle. The 'A's without circles represent the true positions of the acceptor sites. The displacement process is to find a proper position of the ligand such that the sum of squares of 1–1′ distance and 2–2′ distance is minimized

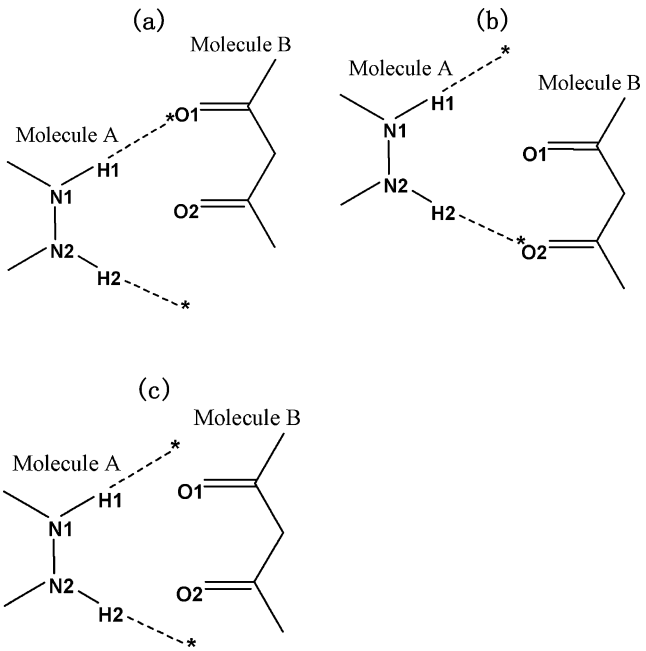

Fig. 4 The illustration of the displacement optimization

minimized. Both N1–H1…O1 and N2–H2...O2 can be hydrogen bonded in Fig. 4(c). Although only two hydrogen bonds are taken into consideration in this example, the principle of optimization problem (1) applies to any situation with more than three hydrogen bonds.

When there are less than three hydrogen bonds formed between protein and ligand, particular attention should be paid to handle these special cases. In the case where only two hydrogen bonds are formed, the ligand can rotate freely around the axis which passes through the two potential hydrogen bonding sites while keeping the objective function of formula (1) unchanged. In this situation, the algorithm will generate a collection of estimations of the position of ligand due to the possible rotation. In H-DOCK, the rotation increment is taken to be 30°. Similar approach is applied when only one hydrogen bond is formed, where the ligand can rotate around the potential bonding site freely while keeping the objective function of formula (1) unchanged. In H-DOCK, three rotational degrees of freedom are taken into account. Rotation around x, y, z axis is taken every 30° to generate all possible ligand positions. In order to eliminate unreasonable docking results, a further requirement is that the angle D-A……AA angle must be greater than 90° in the case where only one hydrogen bond is formed.

The scoring function

After the enumeration and the displacement process of the ligand, a list of candidate docking results is generated. For each of these candidates, the position of the ligand relative to the protein is known, so a simple scoring function is used to rank the results finally. In H-DOCK, the van der Waals interaction between protein and ligand is a simplified potential, as

$$VDW_{ij} = \begin{cases} 0.0 & \left(d_{ij} \geq d_{ij0}\right) \\ d_{ij0} - d_{ij} & \left(d_{ij} < d_{ij0}\right) \end{cases} \quad (2)$$

where VDW_{ij} is the potential between atom i and j, d_{ij0} is sum of van der Waals radii of atom i and j, and d_{ij} is the distance between atom i and j. The total potential between the protein and ligand is summed over all heavy atoms between the protein and the ligand, as

$$VDW = \sum_{i}^{protein} \sum_{j}^{ligand} VDW_{ij} \quad (3)$$

where, a uniform van der Waals radii, i.e., 1.7Å, is taken for all types of heavy atoms in protein or ligand. Besides the van der Waals potential, the other term of the scoring function is a penalty item for the ligand atoms if the ligand is out of the docking zone of the protein. In H-DOCK, the docking zone is defined to be the smallest rectangular zone which contains the ligand in the native docked complex, and extends 2Å outwards in every direction. The penalty item is defined as:

$$P_i = \begin{cases} 0.0 & \text{if atom i is in the docking zone} \\ d_{i0} & \text{otherwise} \end{cases} \quad (4)$$

where d_{i0} is the van der Waals radii of atom i of the ligand. Finally, the whole scoring function is described as:

$$F = \sum_{i}^{protein} \sum_{j}^{ligand} VDW_{ij} + \sum_{i}^{ligand} P_i. \quad (5)$$

In order to speed up the searching, the conventional grid based strategy [1, 8] is adopted to calculate the potential energy in H-DOCK. The 3-dimensional space of the docking zone on the protein has been sliced into small cubes with length of 0.25Å. At the center of each cube, an imaginary "probe atom" is placed, and the van der Waals potential between the probe atom and the whole protein is calculated and stored in this cube. When the scoring function needs to be evaluated for a candidate docking result, the precalculated value for each atom in the ligand is fetched from the cube in which the atom is located, and the sum of values of all atoms is the final score for that docking result.

Docking of a flexible ligand

The aforementioned approach applies to the situation where the ligand is assumed to be rigid, but in fact, there always exists some rotatable bonds in the ligand, so flexible docking is more suitable to identify more accurate complex conformations in practice. In order to adapt the H-DOCK to handle the cases where the ligand is flexible, an ensemble of conformations of the ligand is generated, and each conformation can be treated as a rigid ligand, respectively. The combinatorial method can then be applied to each conformation of ligand, as was done for rigid docking, and the same sorting procedure can be applied to all resulting docking conformations regardless whether the ligand is rigid or flexible.

The algorithmic procedure

The flowchart of the H-DOCK algorithm is stated by Fig. 5, and a detailed step-by-step description is given below, as

(1) Identify potential hydrogen bonding sites on the ligand and the protein.
(2) Combinatorial search of the potential hydrogen bonding modes between protein and ligand, and the same procedure is repeated for each conformation of the ligand when the ligand is flexible.
(3) Rank the modes according to the numbers of hypothetically formed bonds, and discard the unnecessary modes according to the algorithmic parameter.

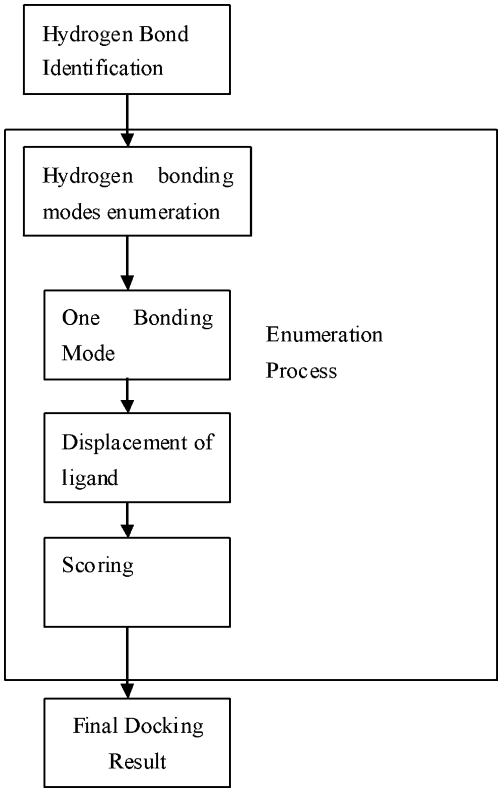

Fig. 5 The algorithmic flow chart of H-DOCK

(4) For each of the remaining modes, estimate the position of the ligand, and compute the value of the scoring function.
(5) Sort the results according to the value of the scoring function. The one with the minimum value is the final docking result, and other results whose scoring function value rank top 100 are saved as well.

Results and discussion

Docking results

H-DOCK is coded by ANSI C++ language, and its effectiveness is evaluated by two test sets, i.e., the rigid test set and the flexible test set introduced in the preceding section and presented in Tables 1 and 2. The docking results are provided in Tables 3 and 4 respectively for rigid and flexible testing, all program runs are done on a cluster core with CPU 2.0 GHz and 8G RAM. For every testing case, the program will give a ranked list of docking results. If the RMSD of the top ranked result is less than 2.0 Å, it is a successful case for the top 1 result. If at least one of the top 100 ranked results falls into the range of RMSD<2.0 Å , it is a successful case for the top 100 results. For rigid docking on test set with 271 cases described in Table 1, the success rate (RMSD < 2.0 Å) of H-DOCK reaches 56.1% and 91.1% for the top 1 and top 100 results. For flexible docking on the test set with 93 cases described in Table 2, the success rate reaches 29.0% and 81.7% for the top 1 and top 100 results, respectively. It takes just 0.14 CPU seconds on average for H-DOCK to finish a rigid test case, and 8.25 CPU seconds on average for each flexible test case with an ensemble of 101 conformations.

Table 3 The success rate of testing cases with different intermolecular hydrogen bonds for rigid test set of Table 1

Number of hydrogen bonds in the complexes	Number of cases	Success rate of top one result/ %	Success rate of top 100 results/ %
1	27	37.037	85.185
2	45	33.333	80.000
3	46	43.478	89.130
4	40	57.500	95.000
5	44	79.545	97.727
6	30	66.667	96.667
7	16	62.500	93.750
8	7	71.429	100.000
9	8	100.000	100.000
10	4	75.000	100.000
11	2	50.000	50.000
12	1	100.000	100.000
17	1	100.000	100.000
Total	271	56.089	91.144

Table 4 The success rate of testing cases with different intermolecular hydrogen bonds for flexible test set of Table 2

Number of hydrogen bonds in the complexes	Number of cases	Success rate of top one result/ %	Success rate of top 100 results/ %
1	2	0.000	100.000
2	14	7.143	64.286
3	17	23.529	70.588
4	13	15.385	69.231
5	14	42.857	92.857
6	6	33.333	100.000
7	11	45.455	100.000
8	13	30.769	84.615
9	1	100.000	100.000
10	1	100.000	100.000
12	1	100.000	100.000
Total	93	29.032	81.720

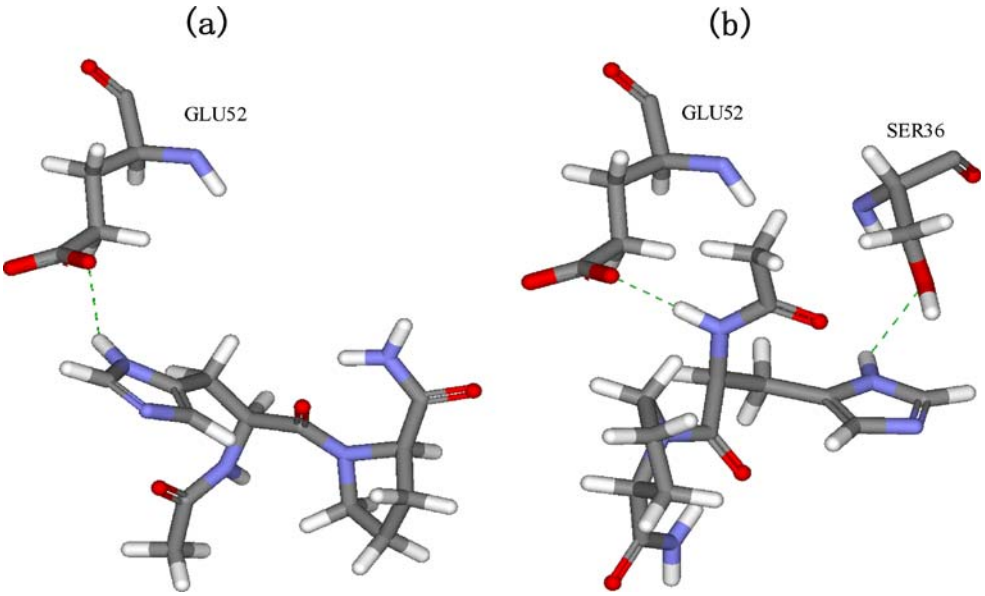

Fig. 6 The native state and docking result of complex 1mcq. Intermolecular hydrogen bonds are shown in green dashed lines. (**a**) The native complex of 1mcq. There is only one hydrogen bond between the protein and ligand. (**b**) The docking result of 1mcq. It is an erroneous result with RSMD=6.8 Å although 2 rather than 1 hydrogen bonds can be formed

Key factors of the algorithm and their impact

Tables 3 and 4 provide the success rate change with the number of hydrogen bonds in the experimental complex, the calculation results state that H-DOCK achieves higher success rate when the hydrogen bond number between protein and ligand increases. The phenomenon that the success rate is higher when more hydrogen bonds are present can be attributed to two reasons. Firstly, hydrogen bonding is not the only interaction between protein and ligand, other interactions such as solvent effect, electrostatic potentials, etc. also contribute to the binding conformation, but such effects are not considered in H-DOCK. When a strategy of maximizing the intermolecular hydrogen bonding is taken, the ligand can be placed at non-native positions where more hydrogen bonds are formed. Figure 6 shows such a situation, where Fig. 6(a) shows the native structure of complex 1mcq and Fig. 6(b) shows its docked result by using H-DOCK. There is only one hydrogen bonds in the native structure, i.e., with GLU 52, but it is possible to place the ligand in the position shown in Fig. 6(b) such that two hydrogen bonds are formed, i.e., with GLU52 and SER36. Thus the strategy of maximizing hydrogen bonds leads to an erroneous docking result with a RMSD of 6.8 Å. However, when there are more hydrogen bonds, hydrogen bonding becomes the dominant interaction between protein and ligand, and the strategy of maximizing hydrogen bonds reflects the nature of the docking process. Figure 7 compares the native state and docking result of complex 1aec. It implies that to give a correct docking result it is not necessary to reproduce all native hydrogen bonds in the docking process.

The second reason comes from the method used to place the ligand. When multiple hydrogen bonds are assumed to form, the optimization process is used to estimate the position of ligand. But, in fact the ligand still has freedom to translate and rotate while keeping all geometric constraints imposed by hydrogen bonding satisfied. The fewer the number of hydrogen bonds there are, the more freedom there is. The least squares optimization scheme does not take such freedom into account and therefore sometimes unable to reproduce the native ligand position even if the hydrogen bonding modes are correctly predicted. However with the increase of the number of hydrogen bonds, stricter geometric constraints are imposed, and the optimization turns out to be an effective way to determine the position of the ligand.

It can be observed from Tables 3 and 4 that the success rate is the lowest when there are two hydrogen bonds, even lower than that when only one hydrogen bond exists. It seems paradoxical, but such phenomenon can be extricated according to the second reason aforementioned. There still is freedom for the ligand to translate and rotate when only two hydrogen bonds are imposed, but only the rotational degree of freedom has been taken into account to generate the possible docked position. However, the case with only one hydrogen bond benefits from the three full rotational degrees taken into account to sample the ligand position, which yields higher probability of finding the native docked position.

Generally speaking, the computation time grows exponentially with the increase of the number of potential hydrogen bonding sites due to the combinatorial nature of the algorithm. For the cases of flexible docking, the computation time is proportional to the number of conformations in the ensemble, but the time is generally less than the total time of the same number of separate

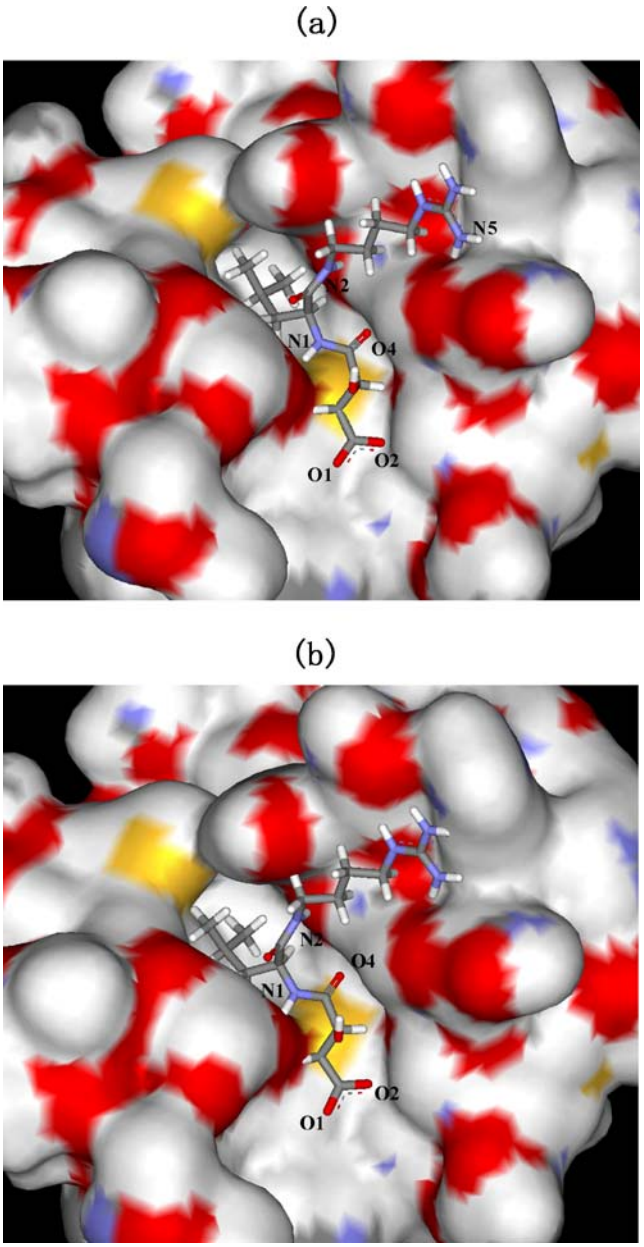

Fig. 7 The native state and docking result of complex 1aec. (**a**) The native complex of 1aec, 6 hydrogen bonds are present between atom N1, N2, N5, O1, O2, O4 and the protein. (**b**) The docking result of 1aec. The bonding mode is not exactly the same with the native state, for the N5 atom on ligand is not bonded. However it is still a successful docking result, gives a RMSD value of 0.83 Å

docking cases of rigid ligands, because some conformations are unsuitable to form intermolecular hydrogen bonds and can be easily rejected during the early stage of the enumeration process.

Comparison with other docking algorithms

According to aforementioned calculation results, H-DOCK is comparable to two commercially available docking programs especially tailored for virtual screening, i.e., LigandFit [24] and LibDock [25], no matter on efficiency or accuracy. Although completely equal-footing comparison cannot be made because of different validation sets and computation environments they were tested on. For flexible docking, both LigandFit [24] and LibDock [25] can achieve 80~90% success rates, and require <10 CPU seconds for one case. However it is worth noting that the decoy conformations used in our validation set are the docked results generated by AutoDock [1], which is different from the randomly generated conformations like those in LibDock [25] and much harder for a traditional scoring function to single out the right conformation, then it is a more rigorous validation set for practical virtual screening algorithm.

It should be noted that though only hydrogen bonding and van der Waals interactions are considered in H-DOCK, it still can achieve high accuracy comparable to those algorithms [24, 25] which consider all kinds of interactions including hydrophobic effects, electrostatic potentials, etc., besides the above ones. Only for some cases where the ligand forms no hydrogen bond with the protein, HDOCK does not apply. In the CCDC/Astex testing set [28], about 10% (34 out of 305) of complexes belong to this category and other docking principle such as hydrophobic area matching must be used. However in most situations, especially in cases when there are three or more intermolecular hydrogen bonds between protein and ligand, high success rate implies the dominate role hydrogen bonding played during the docking process, and H-DOCK has shown obvious advantages when applied to cases where multiple intermolecular hydrogen bonds are presented.

Conclusions

A fast docking algorithm (H-DOCK) based on hydrogen bond matching and surface shape complementarity between protein and ligand is developed, though H-DOCK is constructed firstly for rigid docking, it can be adapted to flexible docking by treating ligand as an ensemble of different conformations. H-DOCK has been tested on a rigid docking set with 271 cases and a flexible docking set with 93 cases, respectively, for flexible docking 101 conformations are considered for each ligand. For rigid docking test set, the success rate (RMSD<2.0 Å) of H-DOCK reached 56.1% and 91.1% for the top 1 and top 100 results. For flexible docking test set, the success rate still reached 29.0% and 81.7% for the top 1 and top 100 results. These results state that hydrogen bonding and surface shape packing reflect the dominant interaction between protein and ligand, though the success rate deteriorates when the number of intermolecular hydrogen bonds are fewer than three. It

takes just 0.14 CPU seconds on average for H-DOCK to implement a rigid test case, and 8.25 CPU seconds on average for each flexible test case, the high efficiency of H-DOCK comes from the unique combinatorial search by virtue of hydrogen bond matching and the simple but effective scoring function. Based on the accuracy and efficiency for the test results, it implies that H-DOCK is suitable to be used as a pre-filter for high throughput virtual screening for molecular recognition based molecular design as the database may contain millions of molecules.

Acknowledgments Y.Z appreciates gratefully the financial support from the National Science Foundation of China (Nos: 20506013, 20776075) and the National High Technology Research and Development (863) Program of China (Nos: 2006AA02Z337, 2008AA02Z208).

References

1. Morris GM et al (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
2. Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 32:335–373
3. Sousa S, Fernandes P, Ramos M (2006) Protein-ligand docking: current status and future challenges. Proteins: Structure, Function, and Bioinformatics 65:15–26
4. Jones G et al (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
5. Taylor J, Burnett R (2000) DARWIN: a program for docking flexible molecules. Proteins: Structure, Function, and Genetics 41:173–191
6. Trosset J, Scheraga H (1999) Prodock: software package for protein modeling and docking. J Comput Chem 20:412–427
7. Abagyan R, Totrov M, Kuznetsov D (1994) ICM-a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 15:488–506
8. Liu M, Wang S (1999) MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. J Comput Aided Mol Des 13:435–451
9. Hart T, Read R (1992) A multiple-start Monte Carlo docking method. Proteins: Structure, Function, and Genetics 13:206–222
10. McMartin C, Bohacek R (1997) QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 11:333–344
11. Baxter CA et al (1998) Flexible docking using Tabu search and an empirical estimate of binding affinity. Proteins: Structure, Function, and Genetics 33:367–382
12. Ewing TJA, Makino S, Skillman AG, Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15:411–428
13. Bohm H (1992) LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. J Comput Aided Mol Des 6:593–606
14. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489
15. Mizutani M, Tomioka N, Itai A (1994) Rational automatic search method for stable docking models of protein and ligand. J Mol Biol 243:310–326
16. Welch W, Ruppert J, Jain A (1996) Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chem Biol 3:449–462
17. Miller M, Kearsley S, Underwood D, Sheridan R (1994) FLOG: a system to select 'quasi-flexible'ligands complementary to a receptor of known three-dimensional structure. J Comput Aided Mol Des 8:153–174
18. Nilsson L, Karplus M (1986) Empirical energy functions for energy minimization and dynamics of nucleic acids Supported in part by a grant from the national institutes of health. J Comput Chem 7:591–616
19. Weiner P, Kollman P (1981) AMBER: assisted model building with energy refinement. A general program for modeling molecules and their interactions. J Comput Chem 2:287–303
20. Kramer B, Rarey M, Lengauer T (1999) Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. Proteins: Structure, Function, and Genetics 37:228–241
21. Verdonk M et al (2003) Improved protein-ligand docking using GOLD. Proteins 52:609–623
22. Tao P, Lai L (2001) Protein ligand docking based on empirical method for binding affinity estimation. J Comput Aided Mol Des 15:429–446
23. Wang RX, Lai LH, Wang SM (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aided Mol Des 16:11–26
24. Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 21:289–307
25. Diller DJ, Merz KM (2001) High throughput docking for library design and library prioritization. Proteins: Structure, Function, and Genetics 43:113–124
26. Meyer M, Wilson P, Schomburg D (1996) Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. J Mol Biol 264:199–210
27. Pei JF et al (2006) PSI-DOCK: towards highly efficient and accurate flexible ligand docking. Proteins: Structure, Function, and Bioinformatics 62:934–946
28. Nissink JWM et al (2002) A new test set for validating predictions of protein-ligand interaction. Proteins: Structure, Function, and Genetics 49:457–471
29. Baker EN, Hubbard RE (1984) Hydrogen-Bonding in globular-proteins. Prog Biophys Mol Biol 44:97–179
30. SYBYL, Tripos Inc, 1699 South Hanley Rd, St Louis, Missouri 63144, USA
31. Wang RX, Lu YP, Wang SM (2003) Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 46:2287–2303
32. Horn B, Hilden H, Negahdaripour S (1988) Closed-form solution of absolute orientation using orthonormal matrices. J Opt Soc Am A 5:1127–1135