A Method for Biomolecular Structural Recognition and Docking Allowing Conformational Flexibility

Bilha Sandak[1] *, Ruth Nussinov[3,4] and Haim J. Wolfson[2]

[1] *Department of Applied Mathematics and Computer Science,*
Weizmann Institute of Science, Rehovot 76100, Israel
E-mail: billie@wisdom.weizmann.ac.il

[2] *Computer Science Department, Sackler Faculty of Exact Sciences,*
Tel Aviv University, Tel Aviv 69978, Israel
E-mail: wolfson@math.tau.ac.il

[3] *Laboratory of Experimental and Computational Biology,*
SAIC, NCI-FCRDC, Bldg 469, rm 151,
Fredrick, MD 21702, USA
E-mail: ruthn@fcrfv1.ncifcrf.gov

[4] *Sackler Institute of Molecular Medicine, Faculty of Medicine,*
Tel Aviv University, Tel Aviv 69978, Israel
E-mail: ruthnu@ccsg.tau.ac.il

Running title: A Method for Structural Recognition Allowing Flexibility

Keywords: *flexible docking, molecular docking, induced fit, structure-based drug design, object recognition*

*Correspondence should be addressed to Bilha Sandak, Department of Applied Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel; Fax: +972-8-9342945, +972-8-9344122; Tel: 972-8-9344312; E-mail: billie@wisdom.weizmann.ac.il

Abstract

In this work we present an algorithm developed to handle biomolecular structural recognition problems, as part of an interdisciplinary research endeavor of the Computer Vision and Molecular Biology fields. A key problem is rational drug design and in biomolecular structural recognition is the generation of binding modes between two molecules, also known as molecular docking. Geometrical fitness is a necessary condition for molecular interaction. Hence, docking a ligand (e.g., a drug molecule or a protein molecule), to a protein receptor (e.g., enzyme), involves recognition of molecular surfaces. Conformational transitions by 'hinge-bending' involves rotational movements of relatively rigid parts with respect to each other. The generation of docked binding modes between two associating molecules depends on their three dimensional structures (3-D) and their conformational flexibility. In comparison to the particular case of rigid-body docking, the computational difficulty grows considerably when taking into account the additional degrees of freedom intrinsic to the *flexible molecular docking* problem. Previous docking techniques have enabled hinge movements only within small ligands. They have not addressed hinge-bending motions of receptor protein domains, although these type of transitions are significant. Our approach allows hinge induced motions to exist in either the receptor or the ligand molecules of diverse sizes. We achieve this by adapting a technique developed in Computer Vision & Robotics for the efficient recognition of partially occluded articulated objects. These types of objects consist of rigid parts which are connected by rotary joints (hinges). Our method is based on an extension and generalization of the Hough transform and the Geometric Hashing paradigms for rigid object recognition. We show experimental results obtained by the successful application of the algorithm to cases of bound and unbound molecular complexes, yielding fast matching times. While the 'correct' molecular conformations of the known complexes are obtained with small RMS distances, additional, predictive good-fitting binding modes are generated as well. We conclude by discussing the algorithm's implications and extensions, as well as its application to investigations of protein structures in Molecular Biology and recognition problems in Computer Vision.

Introduction

Here we present and apply a Computer Vision & Robotics based algorithm to structural recognition problems in Molecular Biology. A key problem in rational drug design and biomolecular pattern matching is the generation of docked configurations between a pair of molecules. Molecular docking is an essential step in many biochemical activities such as regulatory mechanisms, chemotherapy and toxicity. The ability to automatically predict molecular binding modes is important for example, in structure based drug design. Such an automatic tool can aid in database searches for drugs fitting target enzymes, inhibiting or altering the enzymes' undesired activity.

As structural complementarity is a necessary condition for a successful docking, the molecular docking problem is first approached by determining geometrically acceptable solutions. Subsequently, the proposed bound complexes can be checked for their biochemical feasibility. During the process of molecular association, either of the participating molecules, i.e., the ligand or the protein receptor, may undergo conformational changes, enabling their access and binding. The induced conformational transitions involve flexible movements of molecular parts, in the form of rotational movements of relatively rigid subparts about hinges, namely *hinge-bending* movements (see Fig. 1). Hinge induced movements in receptors are important for example in enabling access and better fitting of substrates or inhibitors to the active site of an enzyme receptor. Movements of enzyme domains may thus induce catalytic or inhibition processes upon binding substrates or inhibitors (see Fig. 1C,D). Molecular docking depends on the spatial structures of the associating molecules and on their conformational flexibility. This problem is referred to as the *flexible molecular docking* problem.

Rigid docking is a particular case of the general flexible docking problem, where a rigid ligand ('key'), is docked onto a rigid protein receptor ('lock'). Heuristic rigid body methods have been developed (see for example [Katchalski-Katzir et al., 1992, Jiang and Kim, 1991, Wang, 1991, Kuntz et al., 1982, Fischer et al., 1995]). Yet, these methods do not take into account the conformational transitions that molecules may undergo. This shortcoming is exemplified by a database search for inhibitors to the HIV-1 protease [DesJarlais et al., 1990], which enables the prediction of only rigid docked conformations. By allowing flexibility in either the ligand or receptor molecules, additional candidate inhibitors may be obtained. These are most likely to be ignored during the limited and restrictive rigid-body docking search, resulting in missing potential therapeutic agents. It is therefore important and beneficial to account for alterations in ligands and receptors molecular conformations when addressing the docking problem.

Previous flexible docking techniques account for induced hinge flexibility only within small ligands, and do not enable subparts (domain) motions in receptors [DesJarlais et al., 1986, Leach and Kuntz, 1992, Mizutani et al., 1994, Leach, 1994, Smellie et al., 1991, Goodsell and Olson, 1990, Ghose and Crippen, 1985, Jones et al., 1995, Clark and Ajay, 1995, Jones et al., 1997, Rarey et al., 1996, Welch et al., 1996]. Partial receptor flexibility is allowed by Leach [Leach, 1994] which enables side-chain [1] flexibility, and Jones et al. [Jones et al., 1995] which allows partial flexibility in hydrogen-bonding groups. None of these techniques enable hinge induced domain within receptors, although theoretically it seems feasible. Both DesJarlais *et al.* and Leach & Kuntz [DesJarlais et al., 1986, Leach and Kuntz, 1992] use the DOCK based approach [Kuntz et al., 1982] for docking a flexible ligand onto a rigid receptor. The ligand is represented as a composition of its rigid parts. Each part is matched separately within the docking site. DesJarlais *et al.* [DesJarlais et al., 1986] do not assume dependency on the order in which the ligand's parts are being recognized. However, Leach & Kuntz [Leach and Kuntz, 1992] have the 'most rigid' informative part of the ligand docked in the initial step of the algorithm. The remaining parts are docked using systematic exploration of the ligand space. The method accounts for both shape complementarity and hydrogen-bonding complementarity. Leach [Leach, 1994] employs a combination of two search methods used for determining the optimal arrangement of amino-acid side chains and ligand conformations. The ligand's conformational space is systematically searched. Ghose & Crippen [Ghose and Crippen, 1985] present a method for determining feasible binding modes of a flexible ligand in the receptor's site using 'distance geometry'. The method superimposes the two bodies, without explicitly calculating the corresponding transformation, according to the intra-atomic and inter-atomic distances constraints. A combinatorial search of the conformational space is being done. Smellie *et al.* [Smellie et al., 1991] consider atoms capable of hydrogen bonding. The method is based on clique finding which is a NP-complete problem. Combinatorial matching is conducted by Mizutani *et al.* [Mizutani et al., 1994] for hydrogen bonds between the groups of a pre-chosen ligand part and of those of the receptor's. The remaining ligand subparts are examined systematically, reconstructing minimal

[1] A protein molecule is built up from a repertoire of 20 amino acids. The *side chain* (attached to the central backbone carbon atom) distinguishes between the different amino acids.

energy conformations. Goodsell & Olson [Goodsell and Olson, 1990] employ the Metropolis (simulated annealing) for conformation searching combined with energy evaluations. Since reproducibility and convergence are not guaranteed, a global solution cannot always be obtained. The same problem exists in the method of Jones *et al.* [Jones et al., 1995]. The authors adapt a 'genetic algorithm' that uses a stochastic 'evolutionary' strategy for exploring the full conformational flexibility of the ligand, with partial flexibility of hydrogen bonding groups of the protein receptor. Additional genetic algorithms based approaches are, e.g., of Clark *et al.* [Clark and Ajay, 1995] and Jones *et al.* [Jones et al., 1997]. Recently, Rarey *et al.* [Rarey et al., 1996] developed an efficient method (FLEXX) for docking flexible organic ligands into protein receptors, by combining a model of the physico-chemical attributes of the docked molecules, with the sampling of the conformational space. The incremental strategy employed is analogous to the aforementioned strategy of Leach & Kuntz. The docking tool places an interactively chosen base fragment, whereas the additional ligand fragments are incrementally placed into the active site. A similar approach is of Welch *et al.* [Welch et al., 1996] which automatically select the base fragment and the docking site.

In this work we describe and apply our general algorithm which allows conformational flexibility in either the ligand or the receptor molecules. Regardless of the existence and the extent of the a-priori unknown hinge bending movement and the docking site, our algorithm incorporates the rigid subpart matching technique and the global consistency checks as an integral part of the recognition process. We simultaneously match all parts of the molecule, and do not apply a systematic exploration of the conformational space.

By considering the molecules as 3-D structures specified by their molecular surface representation, techniques originating in Computer Vision & Robotics can be applied to discover the docked solutions (conformations). The recognition process we face in the docking problem is reminiscent of the automatic part assembly problem in Robotics and of the partially occluded 3-D object recognition task in Computer Vision. Since we are matching complementary surfaces, our recognition task can be formulated as follows : *Given a database of previously known objects (models), and a newly observed scene (target) with numerous cluttered objects, recover all the occurrences of the database objects in the scene, even if they are partially occluded.* The analogy is quite obvious. The database of objects becomes a database of ligands (e.g., drug or protein molecules), and the newly observed scene is the protein receptor. Partial occlusion and additional object clutter is analogous to the fact that only part of the ligand molecular surface binds to a part of the receptor surface with *no a-priori* knowledge of the section that will exhibit the match, i.e., the location of the binding site. The close analogy between rigid docking and object recognition extends also to the flexible case, which is usually referred to in Computer Vision & Robotics as *articulated object recognition.* Articulated objects are objects consisting of rigid parts which are connected either by rotary or sliding (prismatic) joints. The analog of a hinge in the molecule is a rotary joint in an articulated object. The conformational transitions we allow is rotational movements of the molecular substructures about hinges, in either the ligand or the receptor molecules. The extended and adapted recognition task is thus formulated as: *Given a database of known ligands (models), and a newly introduced receptor (target), recover all ligands which exhibit substantial partial surface match with the receptor surface, without colliding with the receptor. If the ligands contain hinges, solve the surface matching problem by recovering the ligand in a plausible conformation, without having the parts self-collide.* The roles of the ligands and the receptors can be interchanged, since the mathematical problem is symmetrical.

Our docking method has originated from previous work in Computer Vision by Wolfson [Wolfson, 1991], which extended the rigid body matching technique based on Geometric Hashing [Lamdan et al., 1990] and the generalized Hough transform [Ballard, 1981], to an articulated object recognition technique. An implementation of this approach for 2-D application of industrial tool recognition in photographs (such as scissors and pliers), has been reported [Beinglass and Wolfson, 1991].

We successfully have applied our adapted 3-D algorithm to interesting molecular complexes cases, in which hinge-bending flexibility is allowed in either the ligand or the receptor molecules, yielding fast matching times of their surfaces. We verify our method by docking molecular complexes of pre-determined bound structures[2]. The 'correct' docked conformations have been obtained, consistent with experimental observations. Furthermore, the transformations which generate the geometrically correct docked solutions, are among the high scoring ones. Predictive, good-fitting, alternate binding modes have been generated as well, for both bound and unbound molecular complexes.

[2] A *bound* structure is a structure where the ligand and the receptor have been determined by examining their complex. In case the ligand and receptor molecular structures have been determined separately, it is called an *unbound* case (structure).

The Algorithm

The algorithm enables protein domain motions at hinge regions of the molecule, and hinge rotational movements of ligand subparts. This is achieved by the general approach of our algorithm allowing 3-D rotation at the hinge(s). The geometrical acceptable docking solutions are optionally filtered to yield solutions restricted to hinge movements of one degree of freedom around known axes, namely rotational bond movements. Thus, one can choose the general geometric model of a full 3-D rotation which allows a reasonable approximation of hinge regions of large molecular domains and of few consecutive bonds (due to the short bond length). Alternatively, the restricted geometrical model can be chosen, where the hinge movements are about rotational bonds.

Overview

As molecules interact at their surfaces, the ligand and the receptor molecules are described as sets of 'interest points' which represent their molecular surfaces. The determination of the molecular surface representation employed, i.e., the choice of interest point sets, is a non-trivial task. Two major issues are considered for a discrete point description of the molecular surface. On the one hand we require the point representation to be as precise as possible, so as to have a high resolution of the surface features. On the other hand, too dense a representation can lead to intractable execution times and high memory consumption, when processing huge point sets. So far we have used two types of interest point sets, which meet the requirements of being both accurate and sparse. The surface representations we employ have been devised by Kuntz *et al.* [Kuntz et al., 1982] by and Lin *et al.* [Lin et al., 1994]. The former describes the surface invaginations and cavities, which facilitates docking flexible ligands onto the active site of the receptor. The latter representation describes key features of the surface by computing critical points for convex, concave and saddle regions of the surface. This description facilitates ligand docking onto hinge bent domain parts of the receptor molecule. It can also describe molecules having 'flat' surfaces.

Our algorithm is divided to two phases, the *preprocessing phase* and the *recognition phase*. An overview of these phases is given below, followed by a detailed description of each phase and its stages. A general outline of the algorithm is depicted in Fig. 2. A summary of the application parameters is displayed in Table 1. In order to simplify the exposition, we describe our algorithm for the case where we position the hinge(s) in the ligands, although one can interchange the the roles of the ligands and the receptors.

Preprocessing - The ligands information is coded into a look-up (hash) table which serves as the database of ligands. The stored information is invariant to rotations and translations, since the ligands may undergo this type of transformation in order to dock to the given target receptor. The hinge locations for every stored ligand are also determined in this phase, by employing chemical considerations. As this phase is independent of the receptor molecules, it can be executed off-line.

Recognition - The receptor structure is introduced to the system. Ligands surfaces having partial fit to the receptor surfaces are recovered, yielding the required transformations for the docking of the ligands' parts onto the receptor. This is achieved by casting a vote for every match between a ligand's surface patch and a receptor's surface patch (*matching stage*). The surface patch is a geometric configuration of the 'interest points'. The hinge position within the docked site is derived from the transformation between the corresponding ligand subpart and the receptor surface patches. The vote is cast, therefore, for the ligand together with the 'new' location of the hinge. The high scoring transformations (hinge locations) are verified to ensure obtaining geometrically acceptable solutions (*verification stage*). This is done by rejecting the transformations which cause the ligands parts to collide with the receptor structure (*collision check*) and with each other (*self collision check*).

We next detail both phases.

Preprocessing

(a) The ligand is represented as a set of 'interest points'.

(b) The (known) hinge positions are picked as the origin of 3-D Cartesian coordinate frames, which will be called the 'ligand frames'. The orientation of these frames is set arbitrarily. The angle spanned by the bonds connected to the hinge is computed, for the use of bond rotation constraint, optionally employed in the verification stage.

(c) For each non-ordered, non-collinear triplet of interest points, in each ligand part, we define three unique triplet based Cartesian frames, one for each triplet point. When considering a triplet of points, an internal cyclic order is defined, so that the frames are positioned as follows: The origin of each frame is defined at the respective triplet point, the x-axis as the line from the point to the neighboring point, the z-axis as the normal to the triangle plane, and the y-axis as the cross product of the x and z axes. These are the 'triplet frames'. The distinctive geometric 'shape signature' of the triplet, namely the non-ordered triplet of the triangle side lengths serves as an address to a look-up (hash) table. The information stored at this entry is the ligand identification, part number, and the transformations between the 'triplet frames' and the 'ligand frame'. Thus, e.g., on a part with two hinges, two sets of transformations will be stored for each triplet point, while on a part with one hinge only, one set of transformations will be stored.

A triplet of points defines the shape signature, or the *invariant*. A minimal and maximal distance constraint (*trig_max_len* and *trig_min_len*, respectively), are introduced in order to reduce the number of triplets stored and matched. A triplet is considered for matching, only if the distance between the composing interest points are within given constraints. This proximity heuristic is based on an assumption that a match between two objects contain relatively dense regions of matching points, and is not composed only of isolated points distant in space. The values of the distance constraints should be large enough to provide for reliable matching. Fewer look-up table records are allocated, when posing the distance constraints, due to the reduction in the number of triplets (invariants) considered. Hence, less table records are processed in the sequel.

Recognition

(a) The molecular structure of the receptor, is represented by its set of 'interest points'.

(b) All ordered non-collinear triplets of the 'interest points' are considered for the matching stage. For each such triplet, the triplet based Cartesian frames are calculated as explained above. The lengths of the triangle sides are computed. Note that this calculation is invariant under rotation and translation, thus (almost) congruent ligand triangles should have similar values. The look-up table is addressed according to the computed triplet of distances. For each ligand-record present at that entry in the table, the 'candidate ligand frames' are computed by applying the pre-recorded transformations at that entry to the receptor 'triplet frame'. The origins of the candidate 'ligand frames' are the candidate *hinge locations*. Votes are cast for the identity of the ligand molecule together with the locations (and orientations) of the 'candidate (hinge-centered) ligand frames'.

There is an option to have all triplets of receptor interest points considered (*run_type*='regular run'), or apply pruning, such that the number of triplets are drastically reduced (*run_type*='rapid run'). This is achieved by dividing the receptor's interest point set to eight segments (octants), and an adjoining ninth segment, whose 'corners' are positioned at the geometrical center of each octant. The matching is then conducted for the triplets within each segment. The adjoining ninth segment, which partially overlaps all eight octants, comprises triplets from two or more octants. When constructing the triplets of the ninth segment, triplets not shared by more than one octant are discarded to avoid ambiguity with the already processed triplets of the considered octant. As in the preprocessing phase, the triplets distance constraints are introduced as well.

(c) Finally, the accumulator of votes is searched for high scoring pairs of *(ligand, hinge location)*. This hinge location actually defines the 3-D translation that the ligand subpart has to undergo in this 'candidate' docking. Hinges receiving a large number of votes are picked. Subsequently, their translations are filtered in the next step (d), for geometrical compatibility.

Since we are interested only in high scoring transformations, we pick hinge locations receiving a large number of votes from both parts connected by it. The hinge location is actually the 3-D translation from the original hinge position of the ligand to its new candidate location. The high scoring hinge locations are determined according to the *voting_threshold* which is a minimal percentage value of the number of votes received by the highest scoring hinge location. Clustering the hinge locations is achieved by the discretization of the continuous 3-D parameter space of the respective translations. The size of the 3-D bins of the parameter space is determined according to the *prm_space_sedo* parameter.

(d) A potential match implies existence of complementarity between receptor-ligand surface patches. However, other regions of the two molecules may collide. In the verification stage, the respective transformations of each of the parts, between the initial (preprocessing phase) 'ligand frame' and the computed (recognition phase) 'candidate ligand frame' are applied to the atoms in each of the parts of the ligand. Transformations which result in the penetration of a ligand part into the receptor (collision check) or yielding collisions between the parts of the ligand (self collision check), are discarded.

In order to speed up the collision check, the ligand part transformations are optionally clustered according to their rotations for every candidate translation. We quantify the degree of rotation using the *angular distance* value of the rotation matrix. This value corresponds to the rotation angle around an equivalent axis, computed as $arccos\frac{trR-1}{2}$, where R is the rotation matrix of the transformation, and tr is its trace. The rotations having the lowest angular distance in every cluster, are chosen as the representative rotations of the corresponding clusters. These rotations and their respective translations are applied to the ligands parts checked for collision with the target receptor. The parameter which specifies the angular distance cluster size is *verf_cluster_size*. It is the user's choice whether to have the collision check carried out for all transformations (*verification_type*='no_prune'), or to apply pruning of the rotations via the clustering mechanism (*verification_type*='prune').

The receptor and the ligand molecules are assumed to collide, if the distance between a ligand atom and a receptor atom is smaller than the sum of their respective van der Waals radii minus a proximity threshold (*collision_distance*). The same criterion is applied to the ligand parts. To speed up the collision check we reduce the size of the space checked. The receptor molecule is divided into eight segments (octants) sharing the geometric center of the molecule. The collision check between a ligand atom and a receptor atom is conducted only in the appropriate receptor's segment (octant), which consists of the respective receptor's atoms. The 'goodness' of a solution is evaluated by employing a score which is based on the number of the ligand's van der Waals spheres which are in contact with the receptor spheres. We refer to this score as the *contact_percentage*. A ligand sphere is assumed to be in contact with a receptor sphere if the distance between the ligand atom and a receptor atom is smaller than the sum of their respective van der Walls radii plus a proximity threshold (*contact_distance*).

Only binding modes receiving a *contact_percentage* which is higher than the *contact_threshold*, are considered for the self penetration check. These high contact binding modes are optionally passed to a 'chemical check' verifying whether the allowed hinge bending motions are restricted to rotational bond movements. This is done by computing the angle(s) spanned by the bonds connected to the hinge(s) in the current binding mode handled. If the angle(s) are similar to the angle(s) computed for the stored ligand in the look-up table (see Preprocessing stage), the binding mode is passed to the self collision check. Else, it is discarded.

When testing the acceptable binding modes for self penetration, there is an option to reduce the number of atoms checked by using the *atom_jump* constraint. It specifies the relative number of the next atom to be checked. For example, if *atom_jump* equal 3, it means that every third atom in the molecule's part is checked. This parameter is important when conducting the *self collision check* for very large molecules, such as enzyme receptors. Otherwise for small molecule, the parameter has no effect (i.e., its value is set to one). As mentioned above, the self collision check employs the same criterion for rejecting self penetration causing transformation, as being done by the collision check.

Complexity Analysis

Since the roles of the ligands and the receptors can be interchanged, the molecular structure stored in the database (look-up table), are referred to as 'models'. The molecular structure introduced in the recognition phase, onto which the models are docked, is referred to as the 'target'.

Our research concentrates on the case where the look-up table contains only one model. The complexity described in this section is analyzed accordingly. The notations used throughout this description are summarized in the nomenclature of Table 2.

Preprocessing Phase Complexity

The shape signature of the object, or the object invariant, is defined as a triplet of interest points, allowing a unique definition of a 3-D rotation and translation, which superimposes a model structure, onto the target structure.

Assuming a model consisting of N equal parts, I_m, the number of invariants obtained is

$$I_m = N \times \frac{m/N(m/N-1)(m/N-2)}{6} , \tag{1}$$

where m is the number of interest points in the model. The number of invariants is reduced by $1/N^2$ in comparison to the rigid model. The complexity of the term I_m, can be written as

$$I_m = O(\frac{1}{6N^2} \times m^3) . \tag{2}$$

The complexity of the number of invariants imply that the execution times of the program are reduced as the model under study is divided to more and more parts. Since less model invariants are stored in the look-up table in the preprocessing phase, less model invariant records are handled in the subsequent recognition phase, thus less candidate transformations are considered in the process.

Since $1/6N^2$ is constant, we obtain that

$$I_m = O(m^3) . \tag{3}$$

The insertion time of a model invariant (a record) to the look-up table is assumed $O(1)$. In general, for a database consisting of L models, the complexity of the preprocessing phase is of order $L \times m^3$. For a database consisting of a single model, we obtain that complexity of the preprocessing phase, C_p, is

$$C_p = O(m^3) . \tag{4}$$

Look-up (Hash) Table Manipulation Complexity

The models' information is stored in the look-up (hash) table, so that the models description is memorized for the recognition phase. The data structure considered for the look-up table is such, that a short access time to the table and its entries is enabled. Although the preprocessing phase can be executed offline, we require the insertion time of a model invariant record, in the appropriate entry, to be as rapid as possible. Moreover, in the recognition phase, we require each target (scene) invariant 'query' to find its corresponding model invariants in the shortest time possible, since each entry in the table may contain more than one record within it.

Since we have defined three inter-point distances as an address to the look-up table, the table is a three dimensional structure, quantized (or discretized) into bins having a resolution denoted as q. The resolution value is determined by the *hash_tab_reso* parameter. Each bin covers the 3-D interval of q^3. The number of bins, B, in the table, for a given distance constraints set *trig_max_len* and *trig_min_len* is

$$B = \left(\frac{D}{q}\right)^3 , \tag{5}$$

where D is the difference between the maximum distance constraint (*trig_max_len*) and the minimum distance constraint (*trig_min_len*).

The average bin occupancy, i.e., the average number of records stored in each table entry, or the average number of invariants stored in a bin, assuming a homogeneous distribution is

$$R = \frac{I_m}{B} , \tag{6}$$

where I_m is the number of model invariants.

The insertion time of a record to an entry in the table is of order 1. A record is entered as the first record, even if many records already reside in the entry. The average number of operations carried out for each target (scene) query, i.e., the number of operations required for accessing the records stored in an entry, is R, assuming a homogeneous distribution.

Recognition Phase Complexity

Since the recognition phase is composed of the matching stage, which is the core of our algorithm, and the subsequent verification stage, the complexities of the stages are analyzed separately.

Matching Stage Complexity The matching stage considers all interest points of the target if *run_type* of regular run is chosen. That is, all triplets of interest points are considered. Hence, the number of target invariants is

$$I_{t_F} = n(n-1)(n-2) , \tag{7}$$

where n is the number of interest points of the target. The following complexity approximation is thus written to I_{t_F}

$$I_{t_F} = O(n^3) . \tag{8}$$

Another option to process the target structure is to choose a *run_type* of a rapid run, i.e., the target interest point space is divided to eight segments (octants) and an adjoining ninth one. Assuming equal number of interest points per octant, the number of target invariants is

$$I_{t_R} = 9 \times n/8(n/8-1)(n/8-2) \; , \tag{9}$$

giving the complexity approximation of

$$I_{t_R} = O(\frac{1}{57} \times n^3) \; . \tag{10}$$

In general, we can write that the approximation of the complexity of the number of target invariants in the recognition is

$$I_t = O(n^3) \; . \tag{11}$$

Based on Equation 11 we can write that the complexity of the matching stage is

$$C_M = O(n^3 \times R) \; , \tag{12}$$

where R is the average access time to a record in a look-up table entry. The actual complexity thus depends on the size of the bins of this table. Large bins are obviously undesirable, since they do not contribute much to the discrimination process as well. In this case, we 'vote' for many candidates simultaneously. If one avoids extremely large bins, this access time will be constant, thus achieving practical matching stage complexity of

$$C_M = O(n^3) \; . \tag{13}$$

However, it should be taken into consideration that there may be a trade-off when defining small bins and consequently reducing the value of R. As the resolution of the look-up table has a dependency on the accuracy of the data stored, small bins may cause the program to loose the correct solution, since in this case, small inaccuracies in the input data may not be tolerated.

Verification Stage Complexity The complexity of the verification stage is composed of the complexities of the collision and self-collision checks.

The complexity of the *collision check* is

$$C_{cc} = O(m_A \times n_A \times f(voting\_threshold, I_t, R)) \; , \tag{14}$$

where m_A is the number of docked model atoms, and n_A is the number of target atoms. The term $f(voting\_threshold, I_t, R)$ represents the number of candidate transformations passing the voting threshold criterion. That is, only for these transformations the model part is transformed, and checked for collision within the target structure it is docked into. The transformations are deduced from the target invariants and look-up table occupancy, hence the number of transformations depends on the number of target invariants, I_t, on the average look-up table occupancy per entry, R, and on the voting threshold. Since the molecular space of the target is divided to eight segments, and the model may be divided to N parts (segments and parts assumed equal in this discussion), the complexity is

$$C_{cc} = O(\frac{m_A}{N} \times \frac{n_A}{8} \times f) \ , \tag{15}$$

thus, speeding up the check.

The number of transformations passing the voting threshold f, is composed of the transformations new hinge locations (translations) and their rotations. By clustering the rotations the number of transformations checked is significantly reduced as only one representative is chosen from each cluster of the size *verf_cluster_size*. Thus, the order of the number of operations carried in the *collision check* is further reduced.

In the *self collision check*, the docked model parts are checked for self penetration. Assuming the model is equally divided to N parts, the average number of operations carried out

$$C_{scc} = \begin{pmatrix} N \\ 2 \end{pmatrix} \times \frac{{m_A}^2}{N^2} \times g(contact\_threshold, f') \ , \tag{16}$$

where the first term in the equation corresponds to the number of times required to apply the self collision check to every two parts of the molecule. Since in our current application we are handling molecules with two and three parts, this number is very small. The term $g(contact\_threshold, f')$ represents the number of high contact candidate binding modes that have passed the collision check. As such, this number depends on the number of transformations passing the collision check, f', and on the contact threshold criterion, which defines a contact percentage value for binding modes to be considered. If one considers only the binding modes which have undergone rotational bond movements, then the number of candidate binding modes checked for self penetration is reduced. Introduction of the *atom_jump* parameter can drastically reduce the number of operations executed, since by definition every $atom\_jump - 1$ atoms are skipped during the check to deliberately cause acceptable sparseness. Thus the average number of operations is reduced to

$$C_{scc} = \binom{N}{2} \times \frac{1}{atom\_jump^2} \times \frac{{m_A}^2}{N^2} \times g \ . \tag{17}$$

The *atom_jump* parameter is not used in the collision check, since in this process, the contact percentage of the model part is calculated, and it is necessary to consider all atoms in the calculation.

We have tested a grid-based approach for conducting the collision check. The target structure is 'placed' on a grid, and the transformed model atoms are checked for collision with the target atoms. This is carried out by checking whether the bins corresponding to the van der Waals radii of the query model atoms intersect with the already occupied bins of the target van der Waals spheres. This approach proved inefficient both in memory and time. It is inappropriate for the self collision check, as the molecular subparts are transformed for every configuration checked. Recently, Halperin & Overmars [Halperin and Overmars, 1994] have developed an algorithm for conducting intersection queries in molecules. The complexity of their approach depends on the data structures used. The perfect hashing data structure, which describes the target molecule composed of n balls, is constructed using $O(n)$ space and randomized preprocessing time. The intersection queries are answered in time $O(1)$ for balls whose radii are not greater than the radius of the largest ball. This algorithm should be investigated for its efficiency, robustness and memory consumption when employed on our molecular cases, as the molecules we dock have a highly diverse range of sizes.

Experimental Results

The atom coordinates considered as input to our algorithm, have been determined by the X-ray crystallography and NMR techniques, and stored in the Brookhaven Protein Data Bank (PDB) as 3-D molecular structures [Bernstein et al., 1977] (see ref. [Branden and Tooze, 1991] for a description of the structure determination techniques). In order to verify our algorithm, we have investigated *bound* molecular structures. This type of complexes comprise the 3-D structures (atom coordinates) of ligand(s) and receptor molecules bound together. By applying our method to molecules extracted from bound configurations, we reproduce the binding mode which is in agreement with the empirical observation. Since the ligand and the receptor have been extracted from bound complexes, the 'correct' geometrical solutions are those with rotations and translations close to zero. The results we have obtained for the bound cases have small root mean squared distances (*RMSD*), as compared with their native crystal/NMR structures. We define the term 'best solution', as the solution having the lowest RMSD values of the molecule-parts. This definition serves as a measurement for the performance of our algorithm when verifying it. It has no meaning in solving 'real' docking problems, where the input is the ligand and the receptor molecular structures are determined separately. A general score is used, having the solutions of the bound and unbound cases ranked according to the size of the contact area(s) of the docked ligand and receptor molecules. In addition to the correct solutions obtained by our method, good-fitting predictive binding modes have been generated as well. Fast matching (recognition) times have been obtained for all cases studied. The executions of the program have been conducted on the SGI-Challenge R8000 machine with 512MB of memory.

Previously, we have docked hinge-bent ligands of various molecular complexes such as the NADPH ligand and the dihydrofolate reductase (DHFR) receptor, using a preliminary implementation of our algorithm [Sandak et al., 1995]. In [Sandak et al., 1996b] we have investigated docked configurations of the U75875 inhibitor and the HIV-1 protease by positioning the hinge at different locations on the inhibitor molecule. In [Sandak et al., 1996a] we allow hinge induced conformational flexibility in either a peptide ligand or the calmodulin receptor molecule, for both complexed (bound) and unbound receptor structures. As we have extended our method to account for multiple (double) hinges, we also explore this case. Here, we apply our extended method to diverse molecular cases. We summarize the results obtained in Table 3 which appears in 'Performance Summary' section (see also [Sandak, 1997]. Following this section, is a detailed description of docking investigations and the biological functions of three cases, which correspond to the MTX/DHFR, NAD-/LDH and MBP/maltose columns of Table 3. The first two cases (the methotrexate ligand and the DHFR receptor; the NAD-lactate ligand and the lactate dehydrogenase receptor), account for docking a flexible ligand onto to a rigid receptor. An unbound docking investigation is conducted for the maltose binding protein receptor and the maltose ligand, where receptor conformational flexibility is permitted.

Performance Summary

A summary of the performance measurements appears in Table 3. Following is a description of the table's entries.

Input size The first four rows of the table correspond to the input size of the different molecular cases explored. The entries comprise of: m, the number of model interest points; n, the number of target interest points; m_A, the number of model atoms and n_A, the number of target atoms.

Molecular cases under study The first six columns of both parts of the table comprise the cases where the hinge is positioned in the ligand. The seventh column summarizes the case where two hinges are introduced in the ligand. The remaining three columns correspond to the cases where the hinge is positioned in the receptor molecule. Bound (complexed) cases are dealt with in the first eight columns, whereas unbound cases are handled in the last two. The data of bound ligand and receptor are extracted from PDB files describing the structure of the complexed molecules. The data of an unbound pair of molecules are extracted from two different PDB files.

Execution time measurements The next group of runtime measurements consist of the execution time measured for the different stages of the algorithm. These are the preprocessing phase, the matching stage, the collision check, and the self collision check. The three latter cases comprise the recognition phase of the algorithm. The verification stage in this phase is composed of the collision check and the self collision check. Although the preprocessing stage can be executed offline, here, the 'total' time measurements are the summation of the execution time of the recognition and preprocessing phases.

Best solution The final group of rows accounts for the information relating to the best solution obtained by our algorithm. For the unbound structures it is defined as the binding mode yielding the highest contact percentage. For the bound (complexed) structures, it is defined as a high contact binding mode having the smallest RMSD measurement for the parts of the model from its original determined PDB coordinates. The 'RMSD' rows depict the RMSD values obtained for the parts of the model. The RMSDs of the domain parts of the receptor were not computed for the unbound cases. There is no meaning for this type of measurement as defined here, when dealing with open and closed conformations of the receptor molecule. The hinge location of the best solution is ranked according to the number of votes it has received in the matching stage, prior to the verification stage. The 'rank percentage' is computed as the ratio (in percents) between the rank of the hinge location of the best solution, and the total number of hinge locations. For example, in the U75875 and HIV-1 protease case, the best solution received a voting score which was ranked as the third highest hinge location. Since 20775 hinge locations were voted for, we obtain the 'rank percentage' of 0.01% ($3/20775 \times 100$). The 'final position' row displays the position of the best solution obtained at the end of the run, ranked according to the contact percentage. This ranking is carried out for the final group of binding modes, which survive the verification stages. These binding modes are the geometrically acceptable solutions, optionally restricted to rotational bond flexibility. Their number appears in the 'binding modes' row.

Below we give a detailed description of the docking investigations conducted for the cases appearing in the third, fourth and eight columns of Table 3.

Docking Methotrexate and Dihydrofolate Reductase

Nucleotides are the building blocks of nucleic acids. The biosynthesis of nucleotides is a vital process, since these compounds are indispensable precursors for the synthesis of both RNA and DNA. Without RNA synthesis, protein synthesis is halted, and unless cells can synthesize DNA, they cannot divide. Nucleotides are also necessary for continuous repair of DNA, which is critical for cell survival. They further play an important role in all major aspects of the metabolism. The inhibitors to nucleotide biosynthesis, are very toxic to cells. The toxicity of the inhibitors has been used to combat cancer, as well as in the treatment of certain diseases resulting from infection by viruses, bacteria or protozoans. *Methotrexate (MTX)* and a number of related compounds inhibit the reduction of dihydrofolate to tetrahydrofolate, a reaction catalyzed by the *dihydrofolate reductase (DHFR)* enzyme. The inhibitor prevents the synthesis of thymidylate in replicating cells. MTX is used as an anticancer drug, preventing the replication of cancerous cells. However, it is highly toxic for all dividing, normal body cells (for more detailed information see ref. [Zubay, 1993]).

We have carried out our docking investigations positioning the hinge at the C9 atom of the MTX ligand (see Fig. 3). The DHFR receptor is assumed rigid. The atom coordinates of the MTX/DHFR molecular complex have been taken from the 3DFR data entry file of the PDB [Bolin et al., 1982]. We have obtained the results depicted by Fig. 4. Prior to the verification stage, the total number of candidate hinge locations receiving votes are 5162. These are presented as translation distances in the plot. The translation distance is computed as the l_2-norm of the transformation's translation vector ($\sqrt{x^2 + y^2 + z^2}$). As the figure shows (Fig. 4a), the 'correct' translations have been obtained among the high scoring ones. This is depicted in Fig. 4c,d, where the translation distances are plotted against their score. The translation distance of the transformation which yields the lowest RMSD of the MTX parts (i.e., the 'best' solution), received the rank of 33 (Fig. 4c) with the score of 268 (Fig. 4d). This hinge location is thus among the 0.64% highest scoring 5162 candidate hinge locations.

The solutions having the highest contact percentages values (see Fig. 4b) correspond to the 'correct' solutions, and to the additional good fitting binding modes. The 'best' solution has been obtained with small RMSD from the original ligand crystal structure as depicted in Fig. 5. 1.48Å for the first part of the MTX ligand and 2.64Å for the second. The translation distance of the first part is 0.49Å, and that of the second is 0.36Å. The respective angular distances are 45.4^0 and 63.1^0. The x, y and z coordinate translations of the first part are -0.45, -0.17 and 0.05 angstroms. The rotation angles are 29.25, 33.33 and -13.41 degrees. As for the second part, the translation values of the three coordinates are 0.33, -0.14 and 0.01 angstroms. The three angles of rotation are 43.14, 9.93 and -50.93 degrees. At the end of the run, the best solution is the ranked as the second top scoring solution. Some of the alternate good-fitting binding modes are depicted in Fig. 6.

The execution times of the program are: 0.32 minutes for the matching stage. 0.58

minutes for the collision check, and 0.32 minutes for the self collision check, giving a total of 1.22 minutes. The number of interest points of the DHFR receptor is 427. The combined number of interest points of the two parts of the MTX ligand is 33. There are 13 points in the first part and 21 points in the second. The two parts are joined by the hinge point. The number of atoms in the DHFR receptor is 1294, and in the MTX ligand is 33 (13 in the first part, and 21 in the second part).

Docking NAD-lactate and Lactate Dehydrogenase

Numerous enzymes catalyze the transfer of electrons from one substrate to another. In many of these oxidation-reduction reactions, one or more protons are also removed from the substrate that becomes oxidized and are added to the substrate that is reduced. Enzymes that catalyze such oxidation reduction reactions, are generally called dehydrogenases. *Lactate dehydrogenase (LDH)* plays an important role in carbohydrate (energy) metabolism, such as in the heart muscle tissue. It catalyzes the interconversion of lactate and pyruvate using NAD as a coenzyme [3] (for a review see ref. [Holbrook et al., 1975]). The NAD is a common electron acceptor in the biological oxidation. LDH catalyzes the reversible transfer of a hydride ion (H^-) to NAD (*NAD-lactate* or *S_lac_NAD*) from lactate. The enzyme binds its substrate in an obligatory order, i.e., LDH binds NAD in the absence of lactate, but will not bind lactate in the absence of NAD. Thus, the LDH is NAD-dependent.

For carrying out our docking investigations, we have positioned the hinge at the O3 atom of the NAD-lactate ligand molecule (see Fig. 7). The LDH receptor is assumed rigid. The atom coordinates of the NAD-lactate/LDH complex were taken from the 5LDH data entry file of the PDB [Grau et al., 1991]. An acceptable cluster of 'correct' solutions has been obtained with a relatively high contact percentages as compared with the original crystal structure. The total number of candidate hinge locations receiving votes is 35053. The 'correct' translations are among the highest scoring ones. The lowest RMSD solution, i.e., the 'best' solution, has a translation distance which is ranked at position 75, having a score of 572. Since the best hinge location is ranked out of a total of 35053 candidate hinge locations, it is in the first 0.21% high scoring hinge locations. This result is obtained after the matching

[3] A *coenzyme* is an organic molecule, smaller than the enzyme, which possess structural properties, or special chemical reactivities, that are useful for catalyzing biochemical reactions in concert with the enzyme.

stage, prior to the verification stage.

The best transformation has been obtained with RMSD of 1.55Å for the first part of the NAD-lactate ligand, and 1.84Å for the second part. The translation distance of the first part is 1.2Å, and the angular distance is 17.04^0. For the second part we have obtained a translation distance of 0.77Å, and angular distance of 22.89^0. The translation values of the x, y, z coordinates of the first part are -0.28, 1.16 and 0.05 angstroms. The rotation angles are 12.85, 13.07 and 1.74 degrees. The coordinate translations of the second part is 0.38, 0.64 and -0.23 angstroms. The three angles of rotations are 20.51, 19.73 and -8.5 degrees. The best solution is ranked as the 13th top scoring solution at the end of the run. Additional, good-fitting predictive binding modes have been generated as well (results not shown here).

The execution times of the program are 1.65 minutes for the matching stage. 2.33 minutes for the collision check, and 0.32 minutes for the self collision check, giving the total of 4.3 minutes. The number of interest points of the LDH receptor is 621. The combined number of interest points of the two parts of the NAD-lactate ligand is 51. 23 points are in the first part and 29 points are in the second. The hinge point joins the two parts. The number of atoms in the LDH receptor is 2560, and in the NAD-lactate ligand is 51 (23 atoms in the first part and 29 points in the second).

Docking Maltose and Maltose-Binding Protein

There are a variety of substrates which are transported through the periplasm. The periplasm is the space between the inner and outer membrane of a bacterium. The substrates transported include peptides, amino-acids, vitamins, monosaccharides, and ions. The bacterial periplasmic system, which transport these substrates, consists of initial receptors, the periplasmic substrate-binding proteins, and protein components that translocate the substrate from the periplasm to the cytoplasm. The substrate first binds to the binding protein, and these interact with the membrane-bound complex, which translocates the substrate to the cytoplasm with concomitant ATP hydrolysis [Oh et al., 1993]. There are about two dozen periplasmic binding proteins, all of which are monomeric, having two distinct globular domains separated by a deep grove or cleft. Some examples are the LAO (lysine/isoleucine/valine)-binding protein [Oh et al., 1993] and the GBP (galactose/glucose)-binding protein.

Here we study the binding conformations of the *maltose/maltodextrin binding protein (MBP)*. Without the bound maltose, the two domains of the MBP receptor, are farther apart and the cleft is wide open ('open form') (see Fig 8a). Upon binding the maltose, the ligand is engulfed in the cleft between the domains, such that the ligand-induced conformational change results in a 'close form' of the protein receptor. The bound maltose is buried in the groove and is almost completely shielded from the bulk solvent. The two domains, the N-domain and the C-domain, are joined by three linkages, two at the base of the cleft, and the third at its side. The N-domain consists of residues [4] 1-109 and 264-309. The C-domain comprises residues 114-258 and 316-370. These lobes are connected by the three interdomain segments, which consist of residues 110-113, 259-263 and 310-315, respectively. The main structural distortion upon maltose binding, are mediated entirely by hinge bending of the main chain of the segments. There is a hinge opening about an axis through the first and second segments (residues 111 and 261) [Sharff et al., 1992].

We have studied the binding modes of the unbound MBP receptor to the maltose ligand. The atom coordinates of the unbound MBP have been taken from the 1OMP data entry file of the PDB [Sharff et al., 1992]. The atom coordinates of the maltodextrin have been extracted from the bound complex of the 2MBP file [Spurlino et al., 1991]. We have divided the MBP molecule to two parts according to its domain partitioning. The hinge has been positioned at the C_α atom of the Glu^{111} residue, as it is assumed that this residue is significant in the ligand-induced domain motion, located at the base of the binding cleft [Sharff et al., 1992].

In Fig. 8 we depict a potential binding mode we have obtained by our algorithm, of the unbound MBP with the maltose ligand. It is the geometrical solution yielding the highest contact percentage. The maltose molecule is composed of 23 atoms which are engulfed by the MBP receptor. Although the ligand is a very small molecule, the binding mode we portray in Fig. 8b, yields a very good fit between the ligand and the receptor (the highest scoring solution). All of the maltose ligand atoms are in contact with the MBP receptor. This result manifests the ability of our method to successfully dock very small molecules having small matching surface, as well as large ones. The unbound MBP has undergone a significant translation and rotation of its two domains, enabling it to transform to a 'closed conformation' from its native 'open form' (20 degrees difference). It is docked in a reversed

[4] During protein synthesis, the amino acids constructing the molecule are linked to each other as *residues*, creating polypeptide chains.

fashion as compared with the bound MBP, as the maltose ligand has a somewhat round and symmetrical structure. The running times recorded are 0.05 minutes for the matching stage, 1.08 minutes for the collision check and 0.35 minutes for the self collision check, giving the total of 1.48 minutes (excluding the offline preprocessing stage). The number of interest points describing the maltose ligand is 40. The combined number of interest points composing the two parts of the MBP receptor is 804. The first part consists of 330 points and the second part of 474 points. The number of atoms in the maltose ligand is 23, and in the MBP receptor is 2862 (1244 atoms in the first parts and 1619 atoms in the second part).

Discussion

In this work we present a general approach, to docking a ligand onto a receptor, allowing hinge-bending motions of relatively rigid parts in either the ligand or receptor molecules. Our main algorithmic tool is the generalization and the extension of the generalized Hough transform technique, which was originally developed for partially occluded, articulated object recognition in Computer Vision & Robotics. A major problem in Computer Vision is object recognition in cluttered scenes, where the objects may be partially occluded. The recognition problem becomes especially challenging when the objects are, so called, articulated objects, namely objects consisting of rigid parts connected by either rotary or sliding joints. Efficient techniques to handle these types of problems have been suggested [Wolfson, 1991]. From a geometric standpoint, there exists an analogy between the problems of object matching and assembly in Computer Vision & Robotics, and of molecular structural comparison and docking in Molecular Biology. In both cases one seeks to discover subparts of the objects (molecules) under study, which have a similar geometric structure (molecular surfaces patches). In general, in both cases, one does not have an a-priori knowledge of subpart (binding site) which will exhibit the match. The analogy between these types of problems brought about this interdisciplinary research endeavor (see [Nussinov and Wolfson, 1991] for previous related research).

Considering the implications of our algorithm to the recognition problems in Computer Vision our technique has the following attributes:

1. *A technique for addressing 3-D to 3-D matching problems.* We have extended the 2-D generalized Hough transform based technique to 3-D, thus potentially enabling to handle occluded articulated object recognition problems originating from diverse research fields.

2. *A full 3-D joint rotation of the object during matching.* This general geometrical model enables all rotary trajectories possible by point rotation in space. It is not restricted to one (two) degree(s) of freedom rotation around an axis (axes). The latter restriction is optionally applied during the subsequent verification stage.

3. *Efficient handling of a large number of interest features.* Usually algorithms in Computer Vision handle small number of interest features specifying the objects (typically

10-20 interest features). Our matching algorithm efficiently handles objects consisting of hundreds of interest features, as well as objects composed of tens of interest points (we have matched objects consisting of 33 to 804 feature points). Typical matching run-times are a few seconds to around a minute, depending on the number of interest features of the matched objects (run-time measurements have been conducted on a SGI R8000 machine).

4. *Handling of noisy sampled input within cluttered scenes.* Due to the complex nature of the molecular structure determination process, the input data to our algorithm can have a range of resolutions and accuracies. Our method yields reasonable results, manifesting robustness and tolerance in tackling this difficulty. Also, our method handles occluded objects in highly cluttered scenes.

5. *Complexity of the matching is* $O(n^3 \times R)$, where n is the number of the target interest features. The power of 3 represents the object's shape signature (surface patches) complexity. The shape signature is defined as a triplet of points, allowing a unique definition of a 3-D rotation and translation, which superimposes a model structure to the target structure. The model objects are indexed into a look-up table (hash-table), allowing the target scene to be simultaneously matched with all previously stored model(s). R is the average number of records within the associative look-up table entries, i.e., the average 'access time' to the table. Constant value of R may result in a practical complexity of order n^3.

6. *Incorporation of the information from all parts of the objects, in the recognition process, regardless of the part's size.* We exploit the fact that the different parts belong to the same object and share common joints. There is no order in the recognition of the parts, and no additional degrees of freedom are incorporated in the matching process. Thus, the parts are not restricted to being highly informative, and the global consistency checks are conducted within the recognition process itself.

The capabilities of our algorithm, manifested in the Molecular Biology applications presented here, have several attractive features.

1. *The hinge is introduced either into a large receptor molecule allowing domain motions,*

or into the ligand. Todate, other methods allow hinges to be positioned strictly in small ligands, and only partial (e.g., side-chain) flexibility is permitted in the receptors.

2. *Docking of diverse sized molecules.* Our general method allows hinge induced motions to exist in either variable size receptors, or in diverse size ligands. The sizes of the molecules range from tens of atoms to thousands of atoms (we have docked molecules of 23 to 2862 atoms).

3. *Good quality predictive results.* The 'correct' conformations are obtained with small RMS deviations from the experimental (crystal/NMR determined) configurations. Additional predictive, 'high-scoring' (good-fitting) binding modes, are generated as well.

4. *A substantial conformational change between the bound and unbound structures is handled,* yielding compatible and acceptable docked configurations.

5. *Short execution times for matching the molecular surfaces.* Seconds to around a minute on a SGI R8000 machine, mostly depending on the target molecule size.

6. *Decrease in running times* for the incorporation of additional hinge(s).

7. *Entire molecular surfaces are considered,* assuming no knowledge of the binding site.

8. *Unbound and bound NMR and crystal molecular structures are successfully docked.* In general, docking NMR structures presents a higher degree of difficulty, as they are determined from an ensemble of relatively flexible conformations present in solution.

It is the combination of the above that enables the application of this tool to molecular motions which are predominantly hinge. However, there are some limitations. If the molecular parts are relatively small, e.g., a part comprising a single bond, they may not be efficiently matched by the method, as small subparts are described by a small number of interest features. A plausible remedy for this type of cases, is to explore the conformational space. In addition, when dealing with molecular databases, the data structures employed here require a significant memory space. Thus, when extending the method to handle database screening, an alternative storage areas should be considered, e.g., external, direct access storage.

Our ability to predict good fitting binding modes of the docked conformations, provides a useful research tool in biomolecular structural studies, and can aid in rational drug design. The association between the receptor and the ligand is an essential step in many biological processes, such as chemotherapy, regulatory mechanisms and toxicity. Knowing the 3-D structure of the associating molecules, our ability to predict their molecular interactions, can potentially lead to the discovery and to the design of a chemical or a compound specifically fitting a target molecule. Structure-based drug design is an increasingly promising approach for rational drug development. Many therapeutic agents have been recently spawned by this approach, for treating diseases such as cancer and AIDS (under human trials). In comparison with the traditional approaches, automation can provide the speed-up required for large-scale screening efforts of drug databases. For example, the Fine Chemical Directory (FCD) database [Gunner et al., 1991] and the Chemical Structural Database (CSD) [Allen et al., 1979]. In the classical methods, several thousands of compounds may have to be synthesized and tested, before a suitable drug is found [Branden and Tooze, 1991]. The structure-based methodology, can yield drugs which are 'tailored' for their target, more quickly and less expensively than other ways. Owing to these advantages, the compounds can potentially be more specific, more potent, and less toxic than those discovered using other methodologies [Bugg et al., 1993, Navia and Murcko, 1992, Martin, 1991, Kuntz, 1992]. An efficient method for molecular docking, as presented here, geared toward screening molecular databases, can contribute to achieving these goals. Many compounds of pharmacological interest can be simultaneously screened for docking to a target molecule. The flexibility we allow, in either the ligand or receptor molecules, can yield additional therapeutic agents that otherwise are most likely to be ignored and missed when using the more restrictive rigid-body docking techniques.

We have obtained the 'correct' binding modes which are consistent with experimental results, by successfully reproducing known docked configurations of the associating molecules. The alternate docked configurations that we predict, i.e., the additional, good-fitting binding modes between the two molecules, provide predictions for plausible molecular interactions and conformations. This capability is important to biomolecular research and to structure-based drug design, in the case where the 3-D molecular structures have been elucidated. Also, the predictive binding modes are important for investigations of molecular interactions.

Following are some additional features that may be incorporated into our algorithm for improving its efficiency and execution times. The algorithm presented here may be extended to account for chemical filtering during the matching process. This may be done, for example, by labeling the 'interest points' according to chemical properties, and accepting only geometrically and chemically consistent matches (e.g., see [Rarey et al., 1996]). Not only will this procedure obliviate the need for carrying out chemical verification of all geometrically acceptable docked conformations, but it would also filter out many biologically unacceptable solutions, speeding up the verification stage. Furthermore, since rotational bond movement is more restrictive than a 3-D hinge rotation, this restriction may be incorporated in the algorithm (matching stage) and improve its performance for the cases where a full 3-D point rotation is not required (e.g., see [Rigoutsos and Califano, 1996]). Parallelization of the algorithm may also speed up the calculation.

The generality of our approach allows its application toward additional structural/pattern matching problems in molecular biology, such as protein structure comparisons. These can be exemplified by a database search for hinge-bent motifs, which can aid in investigations of protein folding. Moreover, our algorithm can be applied to recognition problems in Computer Vision, such as problems arising from diverse fields of medical imaging, CAD/CAM, robotics assembly, mobility and manipulation.

Acknowledgments

We thank Drs. Shuo L. Lin, Robert Jernigan and Jacob Maizel, for helpful discussions, encouragement and interest. The research of R. Nussinov has been sponsored by the National Cancer Institute, DHHS, under Contract No. 1-CO-74102 with SAIC. The contents of this publication do not necessarily reflect the views or policies of the DHHS, nor does mention of trade names, commercial products, or organization imply of endorsement by the U.S. Government. The research of H. Wolfson and R. Nussinov in Israel has been supported in part by grant No. 95-0028 from the U.S.-Israel Binational Science Foundation (BSF), Jerusalem, Israel, by the Rekanati fund and by a basic research fund from Tel Aviv University. This work forms part of the Ph.D. Thesis of B. Sandak, Tel Aviv University and is supported in part by the Carl F. Gause Minvera Center for Scientific Computation.

References

[Allen et al., 1979] Allen, F., Bellard, S., Brice, M., Cartwright, B., Doubledaz, A., Higgs, H., Hummelink, T., Hummelink-Peters, B., Kenward, O., Motherwell, W., Rodgers, J., and Watson, D. (1979). The Cambridge Crystallographic Data Center: computer-based search, retrieval, analyses and display information. *Acta Crystallogr. B*, 35:2331–2339.

[Ballard, 1981] Ballard, D. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122.

[Beinglass and Wolfson, 1991] Beinglass, A. and Wolfson, H. (1991). Articulated object recognition, or, how to generalize the generalized Hough transform. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conf.*, pages 461–466. Maui, Hawaii.

[Bernstein et al., 1977] Bernstein, F., Koetzle, T., Williams, G., Meyer, E., M.D., B., Rodgers, J., Kennard, O., Shimanonchi, T., and Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542.

[Bolin et al., 1982] Bolin, J., Filman, D., Matthews, D., Hamlin, R., and Kraut, J. (1982). Crystal structures of eschorichia coli and lactobacillus casei dihydrofolate reductase refined at 1.7Å resolution. *J.Biol.Chem*, 257:13650–13662.

[Branden and Tooze, 1991] Branden, C. and Tooze, J. (1991). *Introduction to Protein Structure*. Garland Publishing, Inc., New York and London.

[Bugg et al., 1993] Bugg, C., Carson, W., and Montgomery, J. (1993). Drugs by design. *Sci. Am.*, Dec:92–98.

[Clark and Ajay, 1995] Clark, K. and Ajay (1995). Flexible ligand docking without parameter adjustment across four ligand receptor complexes. *J. Comp. Chem.*, 16:1210–1226.

[DesJarlais et al., 1990] DesJarlais, R., Kuntz, I., Furth, P., Alvarez, J., Ortiz de Montellano, P., DeCamp, D., Babe, L., and Craik, C. (1990). Structured-based design of nonpeptide inhibitors specific for the human immunodeficiency virus 1 protease. *Proc. Natl. Acad. Sci. USA*, 87:6644–6648.

[DesJarlais et al., 1986] DesJarlais, R., Sheridan, R., Dixon, J., Kuntz, I., and Venkataraghavan, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.*, 29:2149–2153.

[Fischer et al., 1995] Fischer, D., Lin, S. L., Wolfson, H. J., and Nussinov, R. (1995). A geometry-based suite of molecular docking processes. *J. Mol. Biol.*, 248:459–477.

[Ghose and Crippen, 1985] Ghose, A. and Crippen, G. (1985). Geometrically feasible binding modes of a flexible ligand molecule at the receptor site. *J. Comp. Chem.*, 6:350–359.

[Goodsell and Olson, 1990] Goodsell, D. and Olson, A. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct., Funct., Genet.*, 8:195.

[Grau et al., 1991] Grau, U., Trommer, W., and Rossmann, M. (1991). Structure of the active ternary complex of pig heart lactate dehydrogenase with S-lac-NAD at 2.7Å resolution. *J. Mol. Biol.*, 151:289.

[Gunner et al., 1991] Gunner, O., Hughes, D., and Dumont, L. (1991). *Chem. Inf. Comput. Sci.*, 31:408.

[Halperin and Overmars, 1994] Halperin, D. and Overmars, M. (1994). Spheres, molecules, and hidden surface removal. In *Proc. 10th ACM Symposium on Computational Geometry*, pages 113–122. Stony Brook.

[Holbrook et al., 1975] Holbrook, J., Liljar, A., Steindel, S., and Rossman, M. (1975). In *The Enzymes*, pages 190–292, New-York. Academic Press. (Boyer P.D. ed.) 3rd edit. vol. 11.

[Jiang and Kim, 1991] Jiang, F. and Kim, S. (1991). Soft docking: Matching of molecular surface cubes. *J. Mol. Biol.*, 219:79–102.

[Jones et al., 1995] Jones, G., Willet, P., and Glen, R. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.*, 245:43–53.

[Jones et al., 1997] Jones, G., Willett, P., Glen, R., Leach, A., and Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267:727–748.

[Katchalski-Katzir et al., 1992] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A., Aflalo, C., and Vakser, I. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA*, 89:2195–2199.

[Kuntz, 1992] Kuntz, I. (1992). Structure-based strategies for drug design and discovery. *Science*, 257:1078–1082.

[Kuntz et al., 1982] Kuntz, I., Blaney, J., Oatley, S., Langridge, R., and Ferrin, T. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, 161:269–288.

[Lamdan et al., 1990] Lamdan, Y., Schwartz, J., and Wolfson, H. (1990). Affine Invariant Model-Based Object Recognition. *IEEE Trans. on Robotics and Automation*, 6(5):578–589.

[Leach, 1994] Leach, A. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.*, 235:345–356.

[Leach and Kuntz, 1992] Leach, A. and Kuntz, I. (1992). Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comp. Chem.*, 13(6):730–748.

[Lin et al., 1994] Lin, S. L., Nussinov, R., Fischer, D., and Wolfson, H. J. (1994). Molecular surface representation by sparse critical points. *Proteins: Struct., Funct., Genet.*, 18:94–101.

[Martin, 1991] Martin, Y. (1991). Computer-assisted rational drug design. *Methods in Enzymology*, 203:587–613.

[Mizutani et al., 1994] Mizutani, M., Tomioka, N., and Itai, A. (1994). Rational automatic search method for stable docking models of protein and ligand. *J. Mol. Biol.*, 243:310–326.

[Navia and Murcko, 1992] Navia, M. and Murcko, M. (1992). Use of structural information in drug design. *Current Opinion in Structural Biology*, 2:202–210.

[Nussinov and Wolfson, 1991] Nussinov, R. and Wolfson, H. J. (1991). Efficient detection of three-dimensional motifs in biological macromolecules by computer vision techniques. *Proc. Natl. Acad. Sci. USA*, 88:10495–10499.

[Oh et al., 1993] Oh, B., Pandit, J., Kang, C., Nikaido, K., Gokcen, S., Ames, G., and Kim, S. (1993). Three-dimensioanl structures of the periplasmic lysine/arginine/ornithine-binding proteins with and without a ligand. *J. Bio. Chem.*, 268:11348–11355.

[Rarey et al., 1996] Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996). A fast flexible docking method using incremental construction algorithm. *J. Mol. Biol.*, 261:470–489.

[Rigoutsos and Califano, 1996] Rigoutsos, I. Platt, D. and Califano, A. (1996). Flexible 3D-substructure matching & novel conformer derivation in very large databases of 3D-molecular information. *Research Report RC 20497 (90854), IBM Research Division.* Submitted.

[Sandak, 1997] Sandak, B. (1997). *A Method for Biomolecular Structural Recognition Allowing Conformational Flexibility.* PhD thesis, Computer Science Dept., Tel Aviv University.

[Sandak et al., 1995] Sandak, B., Nussinov, R., and Wolfson, H. J. (1995). An automated Computer-Vision & Robotics based technique for 3-D flexible biomolecular docking and matching. *Computer Applications in the Biosciences*, 11:87–99.

[Sandak et al., 1996a] Sandak, B., Nussinov, R., and Wolfson, H. J. (1996a). Docking of Conformationally Flexible Proteins. In *Seventh Symposium on Combinatorial Pattern Matching*, pages 271–287, Laguna Beach, California, USA. Springer Verlag. Lecture Notes in Computer Science Vol. 1075.

[Sandak et al., 1996b] Sandak, B., Wolfson, H. J., and Nussinov, R. (1996b). Hinge-bending at molecular interfaces: Automated docking of a dihydroxyethylene-containing inhibitor of the HIV-1 protease. In *Journal of Biomolecular Structure & Dynamics, Vol 1,Proceedings of the Ninth Conversation, Eds., R.H. Sarma and M.H. Sarma*, pages 233–252, New York. Adenine Press.

[Sharff et al., 1992] Sharff, A., Rodseth, L., Spurlino, J., and Quiocho, F. (1992). Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry*, 31:10657–10663.

[Smellie et al., 1991] Smellie, A., Crippen, G., and Richards, W. (1991). Fast drug-receptor mapping by site directed distances: a novel method of prdiciting new pharmacological leads. *J. Chem. Inform. Comput. Sci.*, 31:386–392.

[Spurlino et al., 1991] Spurlino, J., Lu, G., and Quiocho, F. (1991). The 2.3Å resolution structure of the maltose or maltodextrin binding protein, a primary receptor of bacterial transport and chemotaxis. *J. Bio. Chem.*, 266:5202–5219.

[Wang, 1991] Wang, H. (1991). Grid-search molecular accessible algorithm for solving the protein docking problem. *J. Comp. Chem.*, 12:746–750.

[Welch et al., 1996] Welch, W., Ruppert, J., and Jain, A. (1996). Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.*, 3:449–462.

[Wolfson, 1991] Wolfson, H. J. (1991). Generalizing the generalized Hough transform. *Pattern Recognition Letters*, 12(9):565 – 573.

[Zubay, 1993] Zubay, Z. (1993). *Biochemistry*. Wm. C. Brown Publishers.

Figure Legend

Figure 1

A schematic illustration of hinge-bending. (**A**) A flexible ligand consisting of three rigid parts connected by two hinges. Two consecutive parts share a common hinge. (**B**) The flexible ligand displayed in (A), is docked to a receptor (bold-face line type). (**C**) A flexible receptor consisting of two rigid domain parts connected by a hinge - 'open form'. (**D**) The flexible receptor displayed in (C), binds a ligand (light line type) - 'closed form'.

Figure 2

A general outline of the algorithm. The 'model' entity represents the molecular structure of a ligand (or a receptor). The 'target' entity represents a receptor (or a ligand) onto which the model(s) are docked. The preprocessing phase can be carried out offline for each new model stored.

Figure 3

The spatial structure of methotrexate (MTX), showing the two parts of the ligand joined by the hinge positioned at the C9 atom.

Figure 4

The MTX/DHFR flexible docking results. (**a**) The number of votes as a function of the translation distance received by the two parts of the MTX ligand. The results displayed here are obtained from the matching stage of the algorithm, i.e., prior to verifying them in the collision and self-collision checks. (**b**) The 'goodness' of the generated binding modes. The percentage of the MTX ligand's van der Waals spheres which are in contact with the DHFR receptor spheres (ordinate), are plotted as a function of translation distances for the different binding modes. (**c**) The candidate hinge locations represented by their translation distances (ordinate), plotted as a function of their scoring rank (abscissa). (**d**) The score, i.e., the number of votes received by the corresponding translation distances in (c) versus the rank.

Figure 5

The best 'correct' docked conformation of the MTX ligand, predicted by our algorithm, is represented as a yellow stick. It is plotted against the pink stick and van der Waals sphere representation of the crystal structure. The DHFR molecule of the crystal complex is drawn as a ribbon. The RMSD of the first part of the docked ligand is 1.48Å from the respective part of the original crystal structure. The RMSD of the second part is 2.64Å. The best solution almost coincides with the crystal structure.

Figure 6

Examples of the MTX/DHFR geometrically acceptable binding conformations. The ligand molecule is drawn in its van der Waals solid sphere representation. The DHFR molecule is represented as a white ribbon. The location and orientation of the DHFR is the same in all figures. (**a**) The original crystal bound complex. (**b**) A predicted binding mode of the MTX ligand. (**c**) An additional predicted binding mode of the MTX ligand.

Figure 7

The spatial structure of NAD-lactate, showing the two parts of the ligand joined by the hinge positioned at the O3 atom.

Figure 8

The *unbound* case where the hinge is in the unbound maltose-binding protein (MBP) receptor, and the maltose ligand is rigid. We position the hinge at the C_α atom of the Glu^{111}, pointed at by the arrow. The maltose ligand is represented as van der Waals solid spheres. The MBP receptor is plotted as a ribbon. (**a**) The unbound MBP receptor and the maltose ligand extracted from two different PDB files. The MBP is in its 'open conformation'. (**b**) The unbound MBP receptor is docked with the maltose ligand yielding a 'close conformation' of the receptor which engulfs the ligand.