Efficient Computational Algorithms for Docking, and for Generating and Matching a Library of Functional Epitopes I. Rigid and Flexible Hinge-Bending Docking Algorithms

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

[1] Intramural Research Support Program - SAIC
National Cancer Institute
Frederick Cancer Research and Development Center
Laboratory of Experimental and Computational Biology
Bldg 469, Rm 151
Frederick, MD 21702

[2] Sackler Inst. of Molecular Medicine
Sackler School of Medicine
Tel Aviv University, Tel Aviv 69978, Israel

[3] Computer Science Department
School of Mathematical Sciences
Tel Aviv University, Tel Aviv 69978, Israel

November 22, 1999

A running title: Efficient docking algorithms

*correspondence should be addressed to R. Nussinov at NCI-FCRF Bldg 469, rm 151, Frederick, MD 21702. TEL: 301-846-5579, FAX: 301-846-5598, email: ruthn@ncifcrf.gov. The publisher or recipient acknowledges right of the U.S. Government to retain a nonexclusive, royalty-free license in and to any copyright covering the article.

Abstract

In this, and the next (1), two review series, we present highly efficient, computer-vision and robotics based algorithms for docking and for the generation and matching of epitopes on molecular surfaces. We start with descriptions of molecular surfaces, and proceed to utilize these in both rigid-body and flexible matching routines. These algorithms originate in the computer vision and robotics disciplines. Frequently used approaches, both in searches for molecular similarity and for docking, i.e., molecular complementarity, strive to obtain highly accurate correspondence of respective molecular surfaces. However, owing to molecular surface variability in solution, to mutational events, and to the need to use modeled structures in addition to high resolution ones, utilization of epitopes might prove to be a judicious approach to follow. Furthermore, through the deployment of libraries of epitopes which represent recurring features, or motifs in a given family of receptors or of enzymes, in principle we *a priori* focus on the more critical groups of atoms, or amino acids, essential for the binding of the two molecules. Utilization of recurring motifs may prove more robust than single molecule matchings. In addition, *via* utilization of epitopes one can make use of information derived from evolutionary related molecules. All of the above combine to represent an approach which may be highly advantageous. Combinatorial approaches have proven their immense utility in the wet laboratory. The combination of efficient computational approaches, and utilization of such libraries may well be particularly profitable. Our highly efficient techniques are amenable to such a task.

In this, first review we focus on rigid and flexible docking algorithms. In the second review (1) we address the generation of epitopes in families of molecules. These can be used by the docking algorithms to obtain the more likely bound interfaces.

Keywords: rigid docking; flexible docking; hinge-bending; ranking; matching; library of epitopes

Introduction

In principle, to efficiently, and successfully, sift through a library of inhibitors, or drugs, seeking to optimally dock them onto a given target receptor, five elements are needed: First, a proper molecular surface representation; Second, efficient matching (docking) algorithms; Third, a reasonable way of accounting for molecular surface variability; And fourth, a practical way of handling molecular flexibility. These four components provide the first candidate molecules. The fifth critical element entails a fast, empirical way of scoring the large number of obtained solutions, and of ranking them. The next stage, which is not addressed in our two review series, is the detailed energetical evaluation of these. Docking is an alternate, complementary tool to that of drug design. In drug design, the molecule is built, normally in the binding site of the receptor, by ligating groups of atoms, in a step by step trial-and-error calculation. A comprehensive, critical review of different methods for carrying out such a design task, their merits and drawbacks, has appeared recently in *Combinatorial Chemistry & High Throughput Screening* (2). Here we focus on docking, or matching, algorithms. In these algorithms the structure of the molecule is taken as a single entity from the database.

The problem of predicting the docked conformation of a receptor–ligand molecule pair without any additional biochemical knowledge as to their binding sites is extremely complex (e.g., ref. 3-23). This problem can be defined as follows: Given the atomic coordinates of the two molecules, predict their native binding association. In principle, any patch of surface of one molecule should be matched with any patch of the other, in all rotations and translations. Clearly, the number of possibilities is immense. Furthermore, even if the active site can be predicted (*via*, for example, its being the largest cavity, e.g., ref. 24, 25), there is no assurance that the corresponding candidate trial ligand will not bind at an alternate, second site. Hence, it is preferable to *a priori* consider entire molecular surfaces, without a predefinition of the interaction site. A further complication is the molecular surface variability. In solution, the surfaces of the molecules are in constant motion. Even without taking into account the slower modes of conformational movements, such as motions of domains, subdomains, or molecular

parts, already the surface, fast mode motions implicitly forces taking account of inter-molecular penetrations of the molecule-pair. In "real life" such penetrations will be alleviated by the movements of the surface atoms (26, 27). In particular, as is intuitively obvious, allowing full fledged molecular flexibility, i.e., allowing every two atoms to move with respect to each other, while searching through databases of molecules is simply entirely infeasible. Thus, we need to devise some practical approaches in which some degree of flexibility will still be permitted.

Are there ways which would allow us to quickly sift through potential docked solutions obtained *via* such docking algorithms, and rank the native conformations at the top of the list? Even a quick glance at the results illustrates the awesome task which is involved (27). Analyses of protein–protein interfaces have already indicated that the binding interfaces are not necessarily those having the largest extent of buried surface areas. Furthermore, native–like bound conformations do not possess the largest non-polar buried surface areas of all other potentially feasible docked configurations (27). Additionally, they do not contain the largest number of hydrogen bonds, or the smallest number of unsatisfied buried polar groups. The latter no doubt can be handled by surface motions, to eliminate such potentially unfavorable energy contributions. Thus, the problem is how, despite these enormous difficulties, to still conceivably be able to detect candidate lead compounds/protein inhibitors.

Here we examine efficient current computational approaches. We point to the advantages, and disadvantages of each. We focus on two computational techniques: the first entails rigid body matching (14, 17, 26-28), and the second is a method which allows conformational flexibility of molecular parts, *via* hinge-bending motions (29-33). We describe the geometrical representation of the molecular surface they currently employ (6, 14, 34-37), and the pattern matching algorithms they apply to detect geometric surface complementarity (14, 16, 30). These techniques are rooted in computer vision and robotics (47, 48, 41). They are highly efficient, yielding candidate solutions in very short CPU matching times. They are straightforward to run, using an SGI workstation, or a PC. Nevertheless, while possessing all of these encouraging attributes, we are still faced with very serious difficulties. These are in the scoring and ranking

of the obtained, predicted docked configurations. We proceed to describe these difficulties, and outline a potential way of handling them, namely, through utilization of a *library of functional epitopes*. We describe an efficient way of generating such a library using techniques based on the same principles, and consider the pro's and con's of their utilization.

This review is divided into two parts. In the first, we describe the docking algorithms, and the generation of potential solutions. In the second, to be published in a later issue (1), we focus on techniques for the generation of functional epitopes, and their utilization.

Generation of the molecular surfaces

In the first stage, we describe the molecular surfaces of the molecules. This representation serves as an input to the matching (docking) routines. We start by putting dots on the molecular surface, utilizing the MS program of Connolly (34, 35). We proceed to select a set of "critical points" from these. As we have already described how these are computed (14, 17), here we shall only touch on the calculation very briefly. For the docking of two protein molecules onto each other, the following steps are carried out: An atom is labeled as an exterior atom if an MS dot is generated for it. Otherwise it is an interior atom. The molecule is mapped onto a 3D grid; First the interior atoms, followed by the exterior ones. The shape function is calculated for each dot. The shape function measures the local convexity of the surface. Here it is defined as the intersection between the molecule and a sphere of a given radius centered at the dot. The critical points are a subset of the local extrema of this shape function (4): Points with a high shape function represent a *hole*; those with a low shape function represent a *knob*. Our goal is to search for deep holes or high knobs. Thus, only dots with shape function $> \frac{2}{3}V$ or $< \frac{1}{3}V$ are candidates for a critical point. A candidate is chosen if its shape function is the local maxima for holes, or local minima for knobs. The surface normal orientation is computed as the segment from the center of mass of the intersection volume, and the dot at which the ball is centered. Further details are given in Norel *et al.* (14, 17). For the large protein–protein docking, both knobs and holes are calculated for each molecule. Matched knob–hole pairs are sought, making no reference as to which molecule contributes the hole, and which the knob (27). For the drug

docking (28), we compute holes in the larger receptor, and knobs in the smaller ligands. In the computation of the shape function of the ligands, the radius of the intersecting ball has been reduced, to enable the small ligand to enter and bind into a pocket of the receptor. For the smaller ligands, the van der Waals radii are further augmented (by 30%) as compared to protein–protein docking. By inflating the atoms of small molecules, we distance the molecules which otherwise might be too close. Further, the contact area between small (or medium sized) ligands and their receptors is often limited as compared to the cases of protein–protein docking. Hence, by inflating the radii, we achieve a more snug fit between the two molecules.

Docking

There are two rigid-body computer-vision based docking techniques which we have devised, implemented and utilized (14, 16, 17). In both the representation and the matching schemes are 3-D rotation and translation invariant. The principles of both are quite similar. To get the flavor of such approaches, here we describe briefly one of these. For additional details on this technique, and for a description of the second, we refer the reader to the literature (16).

To complementary align the surfaces of two molecules we need to compute a rigid transformation which superimposes the surfaces without allowing one molecule to penetrate into the other. To get a hypothesis for such a transformation it is enough to align a triplet of ordered non-collinear points (congruent triangles) from both molecules. However, it may happen that there are no 3 independent matching critical point-pairs between the receptor and the ligand. The strategy exploits only pairs of matching points with the additional geometric information on their associated normals. In order to compute a (hypothesis of a) rigid transformation we need to detect a pair of points in both molecules which share the same internal distance, and, if superimposed, have opposing surface normals. This enables both to find matched surfaces with very few matching critical points, and to reduce the run-time complexity of the program. This is unlike the original work of Connolly (6) which has inspired our approach. The failure of Connolly was simply the consequence of his too-strict a requirement, namely, 4 (independent) critical point pairs which would match between the receptor and ligand at their interface.

However, 4 independent point-pairs are not always found in the intermolecular binding region.

For each pair of points from *each* molecule (any two critical points combination for the protein–protein docking; two holes for the receptor and two knobs for the ligand in the protein–drug cases) we compute a transformation invariant "signature". The signature includes the distance between the two critical points (d); the two angles formed by the line between these critical points and their respective normals (α_1, α_2), and the torsion angle defined by the two normals and the line segment between them (ω). The best rigid transformation (38) between the two pairs of critical points is computed *only* if the signatures of the ligand and of the receptor are compatible. To be compatible implies that a knob in one molecule matches a hole in the other; that the difference between the two distances (d) in the two molecules is less than a given threshold; and that the difference between the corresponding $\alpha's$ and the two $\omega's$ is also within the prespecified limits. We further impose stronger cumulative constraints on the alignment of the normals: The sums of the differences of the two $\alpha's$, and of the three angles should not exceed certain limits. Under these conditions, in order to qualify as a potential match, if one of the normals is not well aligned, the other must compensate, and have a reasonable alignment. There are several advantages in using the surface normals in the signature, in addition to the critical points: First, we need only two critical points in the interface area. Second, the combinatorics of finding the correct match is lower (we use pairs of points, rather than triplets or quartets). And third, the normal orientation is used for fast pruning of many wrong solutions.

Geometric Scoring and Overlap Test

The matching is computed for local patches of the surface. Therefore, it is essential to verify that the solution is acceptable for the entire protein. A problem may conceivably arise when the entire ligand molecule is brought to dock onto the receptor. While there is a good complementarity at the matching interface, there could be an overlap between the two molecules elsewhere. To this end, we compute a scoring function, based on geometric features: surface contact is awarded, overlaps where ligand atom centers invade the outer shell of the molecular representation of the receptor are retained, though penalized. However, potential solutions with

serious overlaps, where ligand atoms fall into the "core" of the receptor are rejected. Allowing a certain extent of intermolecular penetrations implicitly takes into account molecular surface variability. The details of the scoring and ranking routines based on these principles, along with a simple hydrophobicity function utilized in the docking of protein–protein pairs, are described elsewhere (27).

Inspection of the obtained complexed conformations immediately reveals, however, that many molecular associations do not vary appreciably. These represent virtually the same docked solution (16). Clustering similar solutions has the advantage of both reducing the number of the solutions, and allowing us to focus on different and possibly alternate configurations. A good clustering scheme should group similar solutions. However, at the same time it should properly distinguish between alternate binding modes. The relative rotation between two conformations can be computed from their individual rotations against the initial (zero-transformation) conformation (27). If the rotations are R_a and R_b for conformations A and B, then B is rotated with respect to A by $R = R_b {R_a}^{-1}$. When the rotation matrix R is known between two conformers, one can compute from this matrix an axis of rotation (equivalent axis) so that the rotational angle between the two molecules is $acos(\frac{tr[R]-1}{2})$. The relative translation between two conformations is calculated as the distance between the mass centers of their interfaces.

This computer-vision motivated docking algorithm is highly efficient. Its fast CPU matching times, on the order of minutes on a PC, even for large protein–protein cases, allows large scale docking experiments of large and small molecules. Specifically, we have carried out such experiments on about 230 receptor–ligand molecule pairs. The docked molecules vary substantially both in size, and in the hydrophobic/polar chemical nature of their surfaces. We have docked 26 protein–protein "bound", complexed cases, and 19 "unbound" protein–protein cases, where the structures have been determined separately. These have variable, imperfectly matching surfaces. We have further docked about 160 cases of protein–drugs, 13 cases of protein–DNA and 14 cases of DNA–drugs. In none of these the location of the active site, or the identity of particular residues/atoms in either the receptor or the ligand, which participate in the binding, has been

taken into account. Obviously, if the location of the active site is known, this information can be easily incorporated in the described technique to improve speed and quality of results.

For the "bound" protein–protein cases, we obtain solutions with an RMSD $< 2.5\mathring{A}$ from the native conformation. For the "unbound" protein–protein pairs, the solutions range from $1.1\mathring{A}$ to $3.5\mathring{A}$, with 2 cases with larger RMSDs (4.7 and $5.2\mathring{A}$) (27). For the protein–drugs (28), our success rate is over 92%, defining success as obtaining a solution with an RMSD $< 2\mathring{A}$ from the native conformation. Here, in most cases the ranking of the top scoring docked conformations, as compared to their "correct" crystal complex counterparts, is reasonable, however for others it is low. Based on this examination of the docking and the results it has obtained, we have concluded that shape complementarity practically always enables attaining correct docked configurations. Nevertheless, it is insufficient for ranking.

Above, we have outlined a computer-vision based algorithm for *rigid* docking. There are several advantages to adopting such docking schemes. First, on the technical side, rigid docking algorithms are faster than those allowing molecular flexibility. Second, rigid docking computations result in fewer potential solutions than those allowing molecular flexibility. While, certainly, molecules are flexible, and hence rigid body docking may miss out on the "correct" binding configuration, nevertheless, if the flexibility is limited, treating the molecules as rigid bodies will still detect the native bound conformations. Such cases are handled simply by the so-called "error ranges", or thresholds. Up to a certain degree, and within reasonable limits, such techniques have proven to be useful. Below we extend the docking repertoire, by describing efficient methods for handling a special kind of flexibility, namely, those that allow *hinge bending* motions.

Hinge-bending flexible matching

Here we dock a ligand onto a receptor surface, allowing hinge-bending movements of domains, subdomains, or substructural parts. We allow all angular rotations, while still avoiding a conformational space search (29-33). By picking a hinge point, we divide the molecule into two parts. However, rather than dock each of the molecular parts separately, with subsequent

reconstruction of the consistently docked molecules, all parts are docked simultaneously. Furthermore, we utilize the position of the hinge from the start. Like pliers closing on a screw, the receptor automatically closes on its ligand. We allow movements either in the ligand or in the larger receptor, hence mimicking the so-called "induced" molecular fit. In principle, more than one hinge can be allowed in the docking. Contrary to intuitive expectation, there is no increase in the surface *matching* times as the number of hinges increases. However, the scoring (ranking) times increase, owing to the need to check for intra-molecular part penetration, in addition to the inter-molecular penetration check in rigid-body docking. The hinge-bending docking is illustrated in the schematic diagram of Figure 1.

Hinge bending movements are frequently observed when molecules associate (39). The movements can involve entire domains, subdomains, loops, (other) secondary structure elements, or between any groups of atoms connected by flexible joints. We have implemented the hinges at points and at bonds. Since we are dealing with chemistry, the only movements which are possible are rotations around a covalent bond. Hence, we could have limited the matching to the simpler, one degree of freedom rotation. However, by allowing full 3D rotations around a point, rather than around a bond, several rotations about (consecutive, or nearby) bonds are implicitly taken into account. Nevertheless, if required, the point rotation can be restricted to bond rotation. We may further allow several, simultaneous hinges. By allowing several hinge motions to occur at the same time, we simulate the cumulative effect of multiple mutations, each introducing a limited motion. Additionally, several hinge points are examined in nearby (non-adjacent) residues, still obtaining similar docked configurations. In practice, so far the algorithm has been implemented to enable two simultaneous hinges (29-33), although initial prototypes for multiple hinges already exist.

We have successfully applied our algorithm to a number of *bound* and *unbound* molecular configurations, achieving fast matching (recognition) times of their surfaces, for both rigid and flexible docking. As in the rigid-body case, the atomic coordinates are taken from the PDB (40). The location of the hinge has been determined by a comparison of similar structures

in different, i.e., open and closed conformations. Applying our method to molecules extracted from bound configurations, we have reproduced the binding mode which is in agreement with the experimental observation. Since the ligand and the receptor have been extracted from these types of complexes, near native geometrical solutions should be those with near zero rotations and translations. The binding modes we have obtained have small RMSD's as compared with the native crystal structures. The average RMSD of a correct solution is 1.4Å with average run-time for each complex around 1 min (on a SGI-Challenge R8000 machine). For these bound cases, correct bound configurations typically rank high. In addition, for all the cases we have tested, geometrically well-fitting, alternate binding modes have been generated as well (29-33).

Hinge-bending flexible matching: The algorithm

To simplify the description of this robotics-based method (41), we outline it here for the single hinge case, where the flexibility is allowed in the ligand molecule. Although we describe the method for flexible *ligands*, the roles of the ligands and the receptors can be reversed, with the hinges positioned in the receptors. We further describe below the enhancements we have implemented for the multiple hinges case.

The approach is based on a voting scheme for finding the most suitable ligands (out of a library of ligands), and for the respective transformations for docking each of their parts onto a target receptor. Both the ligand and the receptor molecular surfaces are represented as 3-D sets of 'interest points'. Below we first outline this molecular representation which is analogous to the one described above for the rigid body docking. We follow with a brief outline of the interest point matching method. We proceed to detail its two phases, the *preprocessing* and the *recognition*.

The surface representation The method is general, and is not restricted to a particular molecular surface representation. The surface representation of Lin *et al.* (36, 37) which has been employed here, describes the surface in terms of accurately and sparsely spaced points, placed at key locations of the molecule. The representation is independent of the initial density of dot MS points used to describe the surface by the Connolly program (34, 35). The density

of the computed key points is modest. Nevertheless, these points still describe the molecular shape adequately, highlighting its important characteristics. Additionally, as previously, the surface normals are also computed. A further useful feature is calculated here, namely, the surface area which is associated with each point. These critical points are nick-named 'cap', 'pit' and 'belt', and are computed for every convex, concave, and saddle region of the molecular surface, respectively. The computation involves projection of the center of gravity of points in the region, onto the surface. One should note that although the points obtained are sparse relative to standard Connolly surface sampling, they are usually much more dense than the knob-hole points discussed in the previous section.

An overview of the algorithm The ligands may undergo translations and rotations of their parts in order to optimally dock onto the receptor surface. The information of the ligands is stored in a database (a look-up table), which is invariant to this type of transformations. The procedure is carried out in the *preprocessing* phase of the hinge-bending algorithm. The location of the hinge in each of the ligands is pre-determined by flexibility considerations. However, as the algorithm is fast, many, alternate, hinge locations may be tested. The structure of the receptor is presented to the system in the second, *recognition* phase of the algorithm. If a ligand has a matching interest point configuration (i.e., a patch of the molecular surface) to the receptor interest point configuration, the algorithm scores a match. We do this by casting a vote for this ligand, along with the computed location of its hinge. This location is determined from the transformation between the corresponding receptor and ligand interest point surface patches. We seek the highest scoring (voted for) hinge locations of candidate ligands. We assume no knowledge of the binding site, or of the hinge locations relative to the receptor. For a detailed description, see Sandak *et al.*, ref. 29-33.

Description of the phases of the algorithm Below, we describe briefly the two phases in the method. For clarification, these are divided into steps.

Preprocessing

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

(b) The (known) hinge position is selected to be the origin of a 3-D Cartesian coordinate frame. This frame is called the *ligand frame*. The orientation of this frame is set arbitrarily.

(c) For each non-collinear triplet of interest points, in each ligand part, we define a unique triplet based Cartesian frame, called the *triplet frame*. One way to do this is by defining the origin at the first triplet point, the x-axis in the direction of the vector from the first to the second point, the z-axis, which is the normal to the triangle plane, in the direction of the cross-product of the two vectors emanating from the first triangle point, and the y-axis in the direction of the cross product of the x and z unit vectors. In practice, in order not to miss out any potential solution, three such frames are defined, one for each triplet point. The *shape signature* of each triplet is the ordered triplet of its triangle side lengths. This distinctive geometric *shape signature* of the triplet, serves as an address to a look-up (hash) table. The information stored at this entry, with this address, is the ligand identification, the part number, and the transformations between the *triplet frames* and the *ligand frame*.

Recognition

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

(b) All non-collinear triplets of the interest points of the receptor are considered in the matching stage. For each of these triplets, the triplet based Cartesian frames are calculated. Each is nick-named the *receptor triplet frame*. The mode of calculation is as described above. The lengths of the triangle sides of each of the triplets are computed. This calculation is invariant under rotation and translation, thus (almost) congruent ligand triangles will result in similar values. The look-up table formed in the previous, *preprocessing* phase, is entered using the currently computed ordered triplet of triangle side lengths of the receptor as the address. For each ligand-record present at that entry in the table, a *candidate ligand frame* is computed. The *candidate ligand frame* is computed by applying the pre-recorded

ligand transformation at that table entry, to the current *receptor triplet frame*. The origin of the *candidate ligand frame* is the candidate *hinge location*. We cast a vote for the identity of the ligand molecule, along with the location and orientation of the *candidate, hinge-centered, ligand frame*.

(c) Finally, we search for high scoring pairs of *(ligand, hinge location)*. This hinge location defines the 3-D translation that the ligand would need to undergo in this *candidate* docking. The appropriate rotations are computed separately for each part only at a later stage, i.e., in the *verification* stage. In this stage, we check the intra- and inter-molecular penetrations, and score the solutions. In the current step, hinges receiving a large number of votes are selected. High scoring hinge locations are selected according to the minimal percentage value of the number of votes received by the highest scoring hinge location.

(d) A potential match implies the existence of complementarity between patches of the surface of the receptor and of the ligand. However, as we have seen already in the rigid body docking, at the same time other regions of the two molecules may inter-penetrate (collide with) each other. The two molecules are assumed to inter-penetrate (collide) if the distance between a ligand atom and a receptor atom is less than the sum of their respective van der Waals radii minus a small proximity threshold. The same criterion is applied to the intra-molecular penetration check of the two ligand parts. Initially, the transformations for each part are clustered by their rotations. The respective transformations of each of the parts, are next applied to all atoms in each of the two parts of the ligand. Transformations yielding intra-molecular penetration (self-collision) between the two parts of the ligand, or which result in inter-molecular penetration of a ligand part into the receptor are rejected. The quality of a solution is estimated by awarding a score which is based on the number of the van der Waals spheres of the ligand which are in contact with the receptor atoms. This score is the *contact percentage*. A ligand atom is assumed to be in contact with a receptor atom if the distance between the two atoms is smaller than the sum of their respective van der Walls radii and a proximity threshold. Solutions passing these filters are ranked

by the contact percentage.

This algorithm exploits the fact that both parts of the molecule share the same hinge. This has been done by locating the origin of the reference frame of the ligand at the hinge. In this way, both parts contribute votes to a reference frame at the same location, although the orientations differ. Most importantly, by picking up votes from both of its molecular parts, a ligand, which might otherwise have only a small matching surface area with the receptor structure in each of the parts, might still score high. Hence, each of the individual parts of the ligand can obtain an insignificant score. However, by combining the votes from both of the ligand's parts, an overall acceptable match can be automatically detected.

This algorithm is general. It can handle the rigid docking as a particular case. For the rigid-body docking, the ligand reference frame is located arbitrarily.

Here we have described the algorithm for a case with a single hinge. However, it may be extended to multiple hinges. Rather than define a single ligand frame we define multiple ligand frames. Each of these is centered at a different hinge. Then, in the *preprocessing* stage, for each ligand triplet in a single part, we encode the transformations of its triplet frame to all ligand frames in that part. Hence, if a part has two hinges, we will store two transformations for each triplet. On the other hand, if a part has only one hinge, we store only one transformation. The *recognition* phase is unchanged, with the exception that each receptor triplet will participate in the voting for as many frames as the number of different transformations stored in its table entry. Details on the various stages of the algorithm as well as analysis of the run times on typical examples are given in (33).

Some examples of the hinge-bending docking

We have docked several receptor-ligand associations: methotrexate with dihydrofolate reductase; NADPH with dihydrofolate reductase; NAD-lactate with lactate dehydrogenase; maltose with maltose binding protein; HIV-1 protease with an inhibitor; and M13 with calmodulin. For all of these cases, with the exception of the M13-calmodulin, only one hinge has been used, with the hinge positioned in the ligand. To give a feeling for the actual raw numbers in these calcu-

lations, the number of interest points for each of the parts is as follows: for the HIV-1 inhibitor (U75875) the first part has 36 interest points, and the second part contains 24; the first part of the NADPH, when docked to the dihydrofolate reductase, has 27 interest points, and the second 22; the methotrexates 13 in the first part, 21 in the second; the NAD, when docked to the lactate dehydrogenase has 33 in the first part, and 29 in the second. The M13 has 103 in its first part, and 108 in the second. When the M13 is divided into three parts, by the positioning of two hinges, the number of interest points are 87, 37 and 92. When the hinges were put in the larger receptors, i.e., in the calmodulin, the first part has 265 interest points, the second 261; or the maltose binding protein has 330 in the first part, 474 in the second. The number of interest points in the receptors are (the number of the atoms in each of the molecules is in parenthesis): the HIV protease has 351 interest points (1517 atoms); the dihydrofolate reductase has 427 interest points (1294 atoms); lactate dehydrogenase 621 (2560); calmodulin 440 (1164). Maltose has 40 interest points (23 atoms). All examples were run on an SGI-Challenge R8000 machine with 512MB of memory. On this machine, the *preprocessing* times were all a few seconds when the hinges are put in the smaller molecule. If, however, the hinges are positioned in the receptor, the preprocessing times were between 1 min to 5.5 minutes. The *recognition* (matching) times were of the same order of magnitude: between 30 seconds to 1.45 minutes. The RMSD's of the predicted *versus* the crystal complex which have been obtained are also good: these range between $0.6\mathring{A}$ to $3.14\mathring{A}$ for either of the parts. (Two RMSD's are obtained for each match: one for each of the parts. If there are two hinges, 3 RMSD's are obtained, for each of the 3 parts.)

The collision detection times are considerably longer: between 1.3 minutes to 6.6 minutes when the hinge is in the ligand. However, when the hinge is in the receptor, as in the case of the calmodulin, the total collision check time (i.e., the intra-molecular check and inter-molecular collision check) is 96 minutes. We are currently optimizing this stage of the algorithm.

This very fast method is adapted from computer vision and robotics. It allows utilization of the entire molecular surfaces of the molecules when no *a priori* knowledge on the active site location is available. We do not treat conformational flexibility of side-chains. Addressing local

conformational changes is essential in order to obtain correct docked solutions, which would be of practical value. On the other hand, current techniques sampling the configurations locally, treating side-chain rotations, or allowing flexibility only in the smaller ligands, do not allow full scale domain movements. Hence, currently available methods addressing these critical issues on the one hand, and the method presented here on the other hand, are complementary, expanding the repertoire of computational docking tools. In particular, integrating the efficient hinge-bending, "induced fit" docking method described here, with a subsequent finer sampling of the local surface changes, may prove to be an efficient, practical, tool-kit for the researcher.

Previous docking techniques have been able to allow either hinge flexibility within relatively small ligands (e.g., drug molecules), or partial flexibility in protein receptors (20, 21, 42-45). Our approach allows hinge motions of domain/sub-domain/group of atoms to exist in either variable sized ligands, or in receptors. Furthermore, we are further able to incorporate several molecular hinges in the same molecule *simultaneously*. We allow full 3-D rotation at the hinge(s) where the hinge can be at any point in space. Such capabilities stem from the fact that the method is derived from computer vision and robotics (41).

The ranking problem

Here we have presented two methods for docking ligands onto the surfaces of their respective receptors. Both methods are highly efficient, producing high quality results for cases where the fit between the surfaces of the two molecules is quite good. Despite the advantages of these techniques, they face considerable hurdles as the fit deteriorates. This problem is uniformally encountered by all current methodologies. Namely, if we initiate the docking process from entire molecular surfaces, and where the structures of the molecules have been determined separately, that is, when they are not in a complexed conformation, we obtain a range of solutions. Despite repeated attempts by us and by others, the problem of how to rank the solutions, with the near-native solution ranking at, or near, the top of the list has still not been surmounted. As the number of candidate solutions may be large, detailed chemical calculations are unrealistic. On the other hand, fast empirical approaches have not yielded satisfactory results. Among the

approaches which have been utilized over the last few years we find derivation and utilization of pair-wise, knowledge-based potentials, either residue-based or atom-based; calculations of the non-polar buried surface area, or of the total buried surface area; inspection and counts of hydrogen bonds and salt bridges; minimization of the number of buried polar groups; and minimization of the energies. Any of these may work for some cases, but fail for the others. Hence, even if the docking programs are good, and are able to scan quickly large databases of drugs or inhibitors, the ranking problem is still there, constituting a huge stumbling block.

In the next review in this two-review series (1), we provide a potential alternative to these approaches. We describe an efficient algorithm to generate a library of functional epitopes. While such a combinatorial library also has its problems, it may nevertheless still provide a potentially useful evolutionary handle.

Searches of databases of drugs, allowing hinge-bending motions

Above we have illustrated searches for drugs, by directly docking them onto their potential, respective protein receptors. An alternative is to search databases of drugs by seeking similarity to ligands which are known to bind to these receptors. One can either take a given drug and search for molecules which are similar to it, or, alternatively, a superior, more robust approach is to take several ligands known to bind to the same receptor at the same site, construct a consensus structure, and utilize the resulting structural motif in the large scale comparisons. While rigid body comparisons are often carried out to this end, owing to the considerable flexibility of the small molecules, allowing the parts of the ligands to rotate with respect to each other is likely to result in more hits. The potential difficulty however, is the time that computation time that such flexible comparisons would entail.

Nevertheless, the hinge-bending algorithm described above is straightforwardly amenable to such a task. We have already described its implementation to flexible protein structural comparison, allowing swiveling motions (47). The method described by Verbitsky *et al.* extends and generalizes current approaches to searches for structural similarity between molecules by viewing proteins as objects consisting of rigid parts connected by rotary joints. During the matching, we

allow the rigid subparts to be rotated with respect to each other around swiveling points in one of the molecules. This is just as we have described in the hinge-bending docking method. For the protein structural comparisons, this technique straightforwardly detects structural motifs having hinge(s) between their domains. Other existing methods detect hinge-bent motifs by initially finding the matching rigid parts and subsequently merging these together. On the other hand, our method automatically detects recurring substructures, allowing full 3 dimensional rotations about their swiveling points. We have already shown that even for the large proteins, the method is extremely fast. It avoids the time consuming full conformational space search. Comparison of two protein structures, *without a predefinition of the motif*, takes only seconds to one minute on a workstation per hinge. Consequently, the molecule can be scanned for many potential hinge sites, allowing practically all C_α atoms to be tried as swiveling points. Hence, this algorithm provides a highly efficient, fully automated tool. Its complexity is only $O(n^2)$, where n is the number of C_α atoms in the compared proteins. As in our previous methodologies, the matching is independent of the order of the amino acids in the polypeptide chain. We have illustrated the performance of this highly powerful tool on a large number of proteins with hinge-bending domain movements. Despite the motions, known hinge-bent domains/motifs which have been assembled and classified, have been correctly identified. We have further detected additional matches as well. This approach has also been motivated by the model based recognition of articulated objects originating in computer vision and robotics. Figure 2 illustrates a schematic diagram. Details of the method are described in ref. 47.

While previously we have implemented it for protein structural comparison, we are currently extending the implementation to structural comparisons of drugs. For the drug comparisons, we allow rotations around the covalent bonds. There, rather than utilizing the C_α atoms, all atoms are used. More flexible joints are picked as potential hinge sites. Initial prototypes are already being tested (M. Sadov, R. Nussinov and H. J. Wolfson, unpublished).

Acknowledgment

We thank Dr. J. V. Maizel for discussions, and for encouragement. The research of R. Nussinov has been sponsored by the National Cancer Institute, DHHS, under Contract No. 1-CO-74102 with SAIC. The research of R. Nussinov and H. J. Wolfson in Israel has been supported in part by grant No. 95-00208 from the BSF, Israel, by a grant from the *Israel Science Foundation* administered by the *Israel Academy of Sciences*, by the Magnet fund, administered by the Israel Ministry of Industry, and by a basic science grant administered by Tel Aviv University. 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 endorsement by the U.S. Government. The research of H. J. Wolfson is partially supported by the Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University.

References

1. Nussinov, R.; Wolfson, H.; functional epitopes. *Combinatorial Chemistry & High Throughput Screening*, **1999**, To be published.

2. Apostolakis, J.; Caflisch, A.; *Combinatorial Chemistry & High Throughput Screening*, **1999**, 2, 91–104.

3. Goodsell, D.; Olson, A.; *PROTEINS: Structure, Function and Genetics*, **1990**, 8, 195–202.

4. Bacon, D.; Moult, J.; *J. Mol. Biol.*, **1992**, 225, 849–858.

5. Kuntz, I.; Blaney, J.; Oatley, S.; Langridge, R.; Ferrin, T.; *J. Mol. Biol.*, **1982**, 161, 269–288.

6. Connolly, M.; *Biopolymers*, **1986**, 25, 1229–1247.

7. Cherfils, J.; Duquerroy, S.; Janin, J.; *PROTEINS: Structure, Function and Genetics*, **1991**, 11, 271–280.

8. Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A.; Aflalo, C.; Vakser, I.; *Proc. Natl. Acad. Sci. USA*, **1992**, 89, 2195–2199.

9. Jiang, F.; Kim, S.; *J. Mol. Biol.*, **1991**, 219, 79–102.

10. Shoichet, B.; Kuntz, I.; *J. Mol. Biol.*, **1991**, 221, 327–346.

11. Wang, H.; *J. Comp. Chem.*, **1991**, 12, 746–750.

12. Cherfils, J.; Janin, J.; *Curr. Opin. Struct. Biol.*, **1993**, 3, 265–269.

13. Norel, R.; Fischer, D.; Wolfson, H.; Nussinov, R.; *Protein Engineering*, **1994**, 7, 39–46.

14. Norel, R.; Lin, S. L.; Wolfson, H.; Nussinov, R.; *Biopolymers*, **1994**, 34, 933–940.

15. Helmer-Citterich, M.; Tramonato, A.; *J. Mol. Biol.*, **1994**, 235, 1021–1031.

16. Fischer, D.; Lin, S. L.; Wolfson, H.; Nussinov, R.; *J. Mol. Biol.*, **1995**, 248, 459–477.

17. Norel, R.; Lin, S. L.; Wolfson, H.; Nussinov, R.; *J. Mol. Biol.*, **1995**, 252, 263–273.

18. Lengauer, T.; Rarey, M.; *Curr. Opin. Struct. Biol.*, **1996**, 6, 402–406.

19. Wallqvist, A.; Covell, D.; *PROTEINS: Structure, Function and Genetics*, **1996**, 25, 403–419.

20. Rarey, M.; Wefing, S.; Lengauer, T.; *J. Comp.-Aided Mol. Design*, **1996**, 10, 41–54.

21. Jones, G.; Willet, P.; Glen, R.; Leach, A.; Taylor, R.; *J. Mol. Biol.*, **1997**, 267, 727–748.

22. Kasinos, N.; Lilley, G.; Subbarao, N.; Haneef, I.; *Prot Engng*, **1992**, 5, 69–75.

23. Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. E.; *J. Mol. Biol.*, **1997**, 272, 106–120.

24. Peters, K. P.; Fauck, J.; Frommel, C.; *J. Mol. Biol.*, **1996**, 256, 201–213.

25. Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M.; *Prot. Sci.*, **1996**, 5, 2438–2452.

26. Norel, R.; Lin, S. L.; Xu, D.; Wolfson, H.; Nussinov, R.; *Structure, Motion, Interaction and Expression of Biological Macromolecules*, Ed. Sarma, R.H. & Sarma, M.H., Eds. Adenine Press, Albany, N.Y., **1998**, 33–51.

27. Norel, R.; Petrey, D.; Wolfson, H.; Nussinov, R.; *PROTEINS: Structure, Function and Genetics*, **1999**, *in press.*

28. Norel, R.; Wolfson, H.; Nussinov, R.; *Combinatorial Chemistry & High Throughput Screening*, **1999**, Submitted.

29. Sandak, B.; Wolfson, H. J.; Nussinov, R.; *J. Biomol. Struct.& Dynamics*, Proceedings of the Ninth Conversation, Eds., R.H. Sarma and M.H. Sarma, Adenine Press (New York), **1996**, 1, 233–252.

30. Sandak, B.; Nussinov, R.; Wolfson, H. J.; *Comp. Appl. BioSci*, **1995**, 11, 87–99.

31. Sandak, B.; Nussinov, R.; Wolfson, H. J.; Seventh Symposium on Combinatorial Pattern Matching, Laguna Beach, California, USA. *Lecture Notes in Computer Science,* Springer Verlag. **1996**, 1075, 271–287.

32. Sandak, B.; Wolfson, H. J.; Nussinov, R.; *PROTEINS: Structure, Function and Genetics*, **1998**, 32, 159–174.

33. Sandak, B.; Nussinov, R.; Wolfson, H. J.; *J. Comput. Biol*, **1999**, 5, 631–654.

34. Connolly, M.; *J. Appl. Cryst.*, **1983**, 16, 548–558.

35. Connolly, M.; *Science*, **1983**, 221, 709–713.

36. Lin, S.L.; Nussinov, R.; Fischer, D.; Wolfson, H.J.; *Proteins: Struct. Funct. Genet.* **1994**, 18, 94–101.

37. Lin, S.L; Nussinov, R.; *J. Mol. Graphics.* **1996**, 14, 78–90.

38. Schwartz, J.T.; Sharir, M *Int. J. Robotics Res.* **1987**, 6(2), 29–44.

39. Gerstein, M.; Lesk, A. M.; Chothia, C.; *Biochemistry*, **1994** , 6739–6749.

40. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M.; *J. Mol. Biol.*, **1977**, 112, 535–542.

41. Wolfson, H. J.; *Pattern Recog. Lett.*, **1991**, 12, 565–573.

42. Mangoni, M.; Roccatano, D.; Di Nola, A. D.; *Proteins: Struct. Funct. Genet.* **1999**, 35, 153–162.

43. Wang, J.; Kollman, P. A.; Kuntz, I. D.; *Proteins: Struct. Funct. Genet.*, **1999**, 36, 1–19.

44. Leach, A.R.; Kuntz, I.D.; *J. Comp. Chem.*, **1992**, 13, 730–748.

45. Leach, A.R.; *J. Mol. Biol.*, **1994**, 235, 345–356.

46. Verbitsky, G.; Nussinov, R.; Wolfson, H.; *Proteins: Struct. Funct. Genet.*, **1999**, 34, 232–254.

47. Lamdan, Y.; Wolfson, H.J; *Proc. 2'nd Int. Conf. on Computer Vision*, **1988**, IEEE Comp. Soc. Press, 238–249.

48. Lamdan, Y.; Schwartz, J.T.; Wolfson, H.J; *Trans. Robotics and Autom.*, **1990**, 6(5), 578–589.

Figure Legends

Figure 1. A schematic illustration of the molecular hinges allowed in ligands and receptors in the hinge-bending docking. (A) The hinges in the ligands. (B) The ligand is hinge-bent to optimally fit in the receptor active site. (C) The hinge is in the receptor. (D) The receptor is hinge-bent, closing on its ligand.

Figure 2. The hinge-bending algorithm: The origin of the reference frame is at the hinge (top). During the matching phase of the algorithm we obtain two different reference frames with relatively high number of "votes". However, both frames have the same origin.