Molecular Docking: A Problem With Thousands Of Degrees Of Freedom

Miguel L. Teodoro[1] mteodoro@rice.edu

George N. Phillips Jr[2] phillips@biochem.wisc.edu

Lydia E. Kavraki[3] kavraki@rice.edu

1 Department of Biochemistry and Cell Biology and Department of Computer Science, Rice University
2 Department of Biochemistry and Department of Computer Science, University of Wisconsin-Madison
3 Department of Computer Science and Department of Bioengineering, Rice University

Abstract

This paper reports on the problem of docking a highly flexible small molecule to the pocket of a highly flexible receptor macromolecule. The prediction of the intermolecular complex is of vital importance for the development of new therapeutics as docking can alter the chemical behavior of the receptor macromolecule. We first present current methods for docking which however have several limitations. Some of these methods consider only the flexibility of the ligand solving a problem with a few tens of degrees of freedom. When the receptor flexibility is taken into account several hundreds or even thousands of degrees of freedom need to be considered. Most methods take into account only a small number of these degrees of freedom by using chemical knowledge specific to the problem. We show how to use a Singular Value Decomposition of Molecular Dynamics trajectories to automatically obtain information about the global flexibility of the receptor and produce interesting conformations that can be used for docking purposes.

1 Introduction

The application of computational methods to study the formation of intermolecular complexes has been the subject of intensive research during the last decade. It is widely accepted that drug activity is obtained through the molecular binding of one molecule (the ligand) to the pocket of another, usually larger, molecule (the receptor), which is commonly a protein. A complex of a protein with a therapeutical drug is shown in Figure 1. In their binding conformations, the molecules exhibit geometric and chemical complementarity, both of which are essential for successful drug activity. The computational process of searching for a ligand that is able to fit both geometrically and energetically the binding site of a protein is called molecular docking.

The rapid generation of quality lead compounds is a major hurdle in the design of therapeutics, so that accurate automated procedures would be of tremendous value to pharmaceutical and other biotechnology companies. However, designing a drug based on the knowledge of the target receptor structure as determined by current experimental techniques is a process prone to error. The two major reasons responsible for failures are inaccuracies in the energy models used to score potential ligand/receptor complexes, and the inability of current methods to account for conformational changes that occur during the binding process not only for the ligand, but also for the receptor. Although this problem has been partially solved by incorporating ligand flexibility in search methods, predicting receptor structural rearrangements is a very complex problem which has not been solved. The docking problem is analogous to an assembly-planning problem where the parts are actuated by molecular forcefields and have thousands of degrees of freedom.

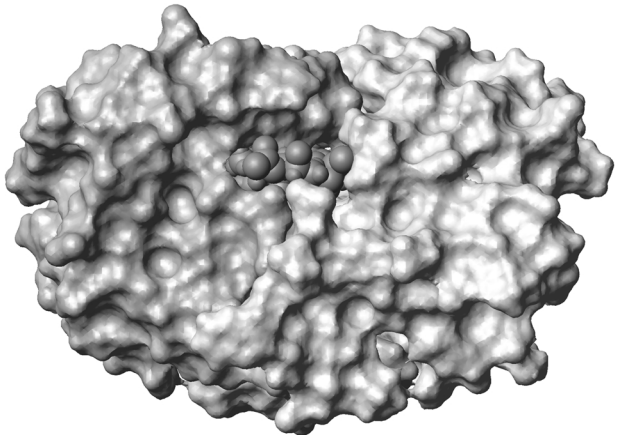

Figure 1 – Therapeutic drug molecule (small molecule towards the center of the figure) bound to protein receptor (HIV-1 protease). The drug molecule fits tightly in the binding site and blocks the normal protein function.

In this article we report on the current methods used to solve the docking problem and on some of the problems and possible solutions to incorporate protein flexibility in the docking process. Section 2 introduces some of the terminology and concepts relevant to this problem and Section 3 reports on some of the docking methods used in academia and industry. However, the models described in Section 3 follow the assumption of a rigid protein which limits their use. In Section 4 we describe some of the methods currently under development to model protein flexibility. We will also describe our own model which incorporates full protein flexibility in docking and can be easily automated.

2 Molecular Modeling

A molecule is characterized by a pair (A; B), in which A represents a collection of atoms, and B represents a collection of bonds between pairs of atoms. Information used for kinematic and energy computations is associated with each of the atoms and bonds. Each atom carries standard information, such as its van der Waals radius. Three pieces of information are associated with each bond: (i) bond length, is the distance between atom centers; (ii) bond angle, is the angle between two consecutive bonds; (iii) whether the bond is rotatable or not (for an illustration of rotatable bonds see Figure 2). Since bond lengths and angles do not change significantly, it is common practice to consider them fixed. Thus the degrees of freedom of the molecule arise from the rotatable bonds. The three dimensional embedding of a molecule defined when we assign values to its rotatable bonds is called the conformation of the molecule. Ligands typically have 3-15 rotatable bonds, while receptors have 1,000-2,000 rotatable bonds. The dimension of the combined search space makes the docking problem computationally intractable.

One key aspect of molecular modeling is calculating the energy of conformations and interactions. This energy can be calculated with a wide range of methods ranging from quantum mechanics to purely empirical energy functions. The accuracy of these functions is usually proportional to its computational expense and choosing the correct energy calculation method is highly dependent on the application. Computation times for different methods can range from a few milliseconds on a workstation to several days on a supercomputer.

In the context of docking, energy evaluations are usually carried out with the help of a scoring function and developing these is a major challenge facing structure based drug design[1]. No matter how efficient and accurate the geometric modeling of the binding process is, without good scoring functions it is impossible to obtain correct solutions. The two main characteristics of a good scoring function are selectivity and efficiency. Selectivity enables the function to distinguish between correctly and incorrectly docked structures and efficiency enables the docking program to run in a reasonable amount of time.

A large number of current scoring functions are based on forcefields that were initially designed to simulate the function of proteins[2,3]. A forcefield is an empirical fit to the potential energy surface in which the protein exists and is obtained by establishing a model with a combination of bonded terms (bond distances, bond angles, torsional angles, etc.) and non-bonded terms (van der Waals and electrostatic). The relative contributions of these terms are adjusted for the different types of atoms in the simulated molecule by adjusting a series of empirical parameters. Some scoring functions used in molecular docking have been adapted to include terms such as solvation and entropy[4]. A separate approach is to use statistical scoring functions that are derived using experimental data[5].

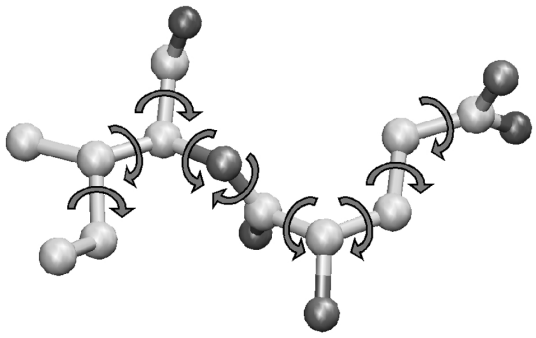

Figure 2 - A drug molecule. Spheres represent atoms and bonds connecting them are represented by sticks. Curved arrows represent the rotatable degrees of freedom around bonds.

3 Rigid Protein Docking

Most of the docking methods used at the present moment in academic and industrial research assume a rigid protein. To illustrate the methodology used by these methods we will briefly discuss three of the most common programs used for docking: Autodock[4], Dock[6] and FlexX[7].

Autodock uses a kinematic model for the ligand similar to the one illustrated in Figure 2. The ligand begins the search process randomly outside the binding site and by exploring the values for translations, rotations and its internal degrees of freedom, it will eventually reach the bound conformation. Distinction between good and bad docked conformations is carried out by the scoring function. Autodock is able to use Monte Carlo methods or simulated annealing (SA) in the search process and in its last version introduced the ability to use genetic algorithms (GA). The routine implemented in the recent release is a Lamarkian genetic algorithm (LGA), in which a traditional GA is used for global search and is combined with a Solis and Wets local search procedure. Morris *et al*[4] show that the new LGA is able to handle ligands with a larger number of degrees of freedom than SA or traditional GA.

FlexX and Dock both use an incremental construction algorithm which attempts to reconstruct the bound ligand by first placing a rigid anchor in the binding site and later using a greedy algorithm to add fragments and complete the ligand structure. Although these programs are more efficient than Autodock in the sense that they require fewer energy evaluations there exist some tradeoffs. One of main problems is that it is not trivial to choose the anchor fragment and its choice will determine what solutions can be obtained. Also the greedy algorithm propagates errors resulting from initial bad choices that lead to missing final conformations of lower energy.

In order to solve the docking problem conformation methods using standard robotics techniques such as probabilistic roadmap planning have been recently described[8,9]. In addition to being successful in finding the correct docking conformation these methods are

useful in identifying possible binding sites and in providing a computational efficient description of the dynamics of ligand binding.

4 Modeling Protein Flexibility

One of the greatest challenges facing current structure-based rational drug design is the integration of protein flexibility in docking methods used to screen databases of possible therapeutic compounds. In this section we present a survey of the approaches under research to model protein flexibility and at the end of this section we present our own solution to this problem.

Current docking methods follow the assumption that protein structures are rigid entities and that it is the ligand that during the binding process changes its three-dimensional structure to find the best spatial and energetic fit to the protein's binding site. This assumption follows the model of lock-and-key binding first proposed by Emil Fischer in 1890. However, a better description of the mechanism of interaction between a protein and its ligand was given by Koshland in 1958 with the induced-fit model. In this model both the protein and the ligand are flexible and when they interact to form a complex both structures change their conformation to form a minimum energy perfect-fit. Unfortunately doing an exact modeling of the flexibility available to the protein during the binding process is still far beyond our present computational capability. Whereas conventional ligand modeling techniques are able to handle up to approximately 30 degrees of freedom when searching for a docked conformation, modeling the full flexibility of the protein requires more than 1000 degrees of freedom, even for a small size protein.

Although the methods described in Section 3 have shown reasonable success in screening for candidate drugs, several studies have exposed their problems and limitations[10,11]. These problems are especially important when non-negligible changes in conformation are present during the binding process. This leads to final docking results that entirely fail to identify potential drug candidates or otherwise assign them very poor binding scores. To overcome limitations from the rigid protein assumption, several approximations have been used to model protein flexibility. These approximations can be divided in two groups: models which try to account for the flexibility of the protein in the binding region and models which simulate the flexibility as a whole.

4.1 Partial Protein Flexibility

The first approximation used in modeling partial protein flexibility was the soft-docking method first described by Jiang and Kim[12]. The principle underlying this method consists of decreasing the van der Waals repulsion energy term between the atoms in the binding site and those in the ligand. This method could result in final solutions that include physically impossible atom collisions. Nevertheless, due to the mobility available to the protein atoms in the binding site, it is possible that there is a low energy rearrangement of these that would eliminate collisions while maintaining the conformation of the ligand returned during its conformational search. This method has the advantage of being computationally efficient as it still describes the protein using fixed coordinates. The method is also easy to implement since it does not require changes to the energy evaluation function besides changing van der Waals parameters.

The most common approximation used to incorporate partial protein flexibility in modeling the binding process is to select a few degrees of freedom in the protein binding site and do a simultaneous search of the combined ligand/protein conformational space. Incorporating select degrees of freedom from the binding site in the conformational search process is based on the assumption that these degrees of freedom are the ones playing a major role in determining the conformational changes during the binding process. This choice requires deep chemical understanding of the system under study and is therefore difficult to automate. Furthermore, even for proteins which are considered relatively rigid, Murray *et al*[11] show that protein backbone changes often play a critical role and these are difficult to model using only a few degrees of freedom. The optimization techniques used for this approach are the same as for the rigid protein but are now required to handle a larger number of degrees of freedom resulting in overall less efficiency.

One of the earliest reports of using select degrees of freedom from the protein was described by Jones *et al*[13] and was implemented in the program GOLD (Genetic Optimization for Ligand Docking). This program improves on the rigid protein model by performing a conformational search on the binding site with the aim of improving the hydrogen bonding network between the protein and the ligand. Hydrogen bonds are local electrostatic interactions between pairs of atoms which play an important energetic role in ligand recognition and binding. GOLD selects the degrees of freedom in the binding site that correspond to reorientations of hydrogen bond donor and acceptor groups. These degrees of freedom represent only a very small fraction of the total conformational space that is available but should account for a significant difference in binding energy values. More recent studies have been reported[14-16] in which other degrees of freedom from aminoacid sidechains are also used in the conformational search. These are searched using stochastic methods with arbitrary step sizes or using rotamer libraries[17]. Rotamer libraries consist of discrete sidechain conformations of low energy which are usually determined from statistical analysis of structural data derived experimentally.

4.2 Full Protein Flexibility

Ideally ligand docking to a protein could be simulated using Molecular Dynamics (MD). This has the advantage that not only it takes into account all the degrees of freedom available to the protein but also enables an explicit modeling of the solvent. Furthermore,

accurate energy calculations can also be carried out using the free energy perturbation method. Unfortunately, modeling proteins using MD is computationally expensive, and the computational power necessary to simulate the full process of diffusion and ligand binding without any approximations will be out of our reach for many years to come. Recently Mangoni *et al*[18] reported a modification to the standard MD protocol which reduces the computational time required for the docking simulation. The protocol consists of separating the center of mass motion of the ligand from its internal and rotational motions by coupling the different degrees of freedom to separate thermal baths. This optimization allows the ligand to sample the space surrounding the binding site faster while maintaining correct interactions with both protein and solvent.

An alternative approach to model full protein flexibility is to generate an ensemble of rigid protein conformations that together represent the conformational diversity available to the protein. These conformations can later be docked to a database of ligands using traditional rigid-protein/flexible-ligand methods. There are several possible methods to generate the ensembles, but unfortunately their accuracy is proportional to the difficulty in obtaining them. The most accurate ensemble is the one determined exclusively from experimental data. An example is the case where several structures of protein/ligand complexes are determined using X-ray crystallography bound to different candidate drugs. Under these circumstances it is usually possible to observe alternative binding modes directly[19]. Another less accurate option is to use the ensemble of structures that results from an experimental protein structure determination using the NMR (Nuclear Magnetic Resonance) technique. This docking methodology was first reported by Knegtel *et al*[20]. Finally, one can generate an ensemble using computational methods such as Monte Carlo (MC) or MD sampling. The accuracy of these alternatives is closely related to the accuracy of the force field used and is limited by the ability of these computational techniques to effectively sample the conformational space[21]. Docking to an ensemble of structures generated using MD was first reported by Pang and Kozikowski[22].

A different representation for full protein flexibility is to divide the protein in tightly coupled domains whose constituent atoms move collectively as one. Hinges connect the domains and the motion of the protein is simulated similarly to an articulated robot. Required conformational changes inside domains can be handled using minimization. An application of this model to the docking problem was reported by Sandak *et al*[23].

4.2.1 Modeling Protein Flexibility With Collective Modes Of Motion

The approach we are presentely investigating to account for full protein flexibility while reducing the computational complexity of the problem is to use the concept of essential dynamics[24]. This formulation divides the conformational space accessible to the protein into two subspaces: (1) an essential subspace containing only a few degrees of freedom which correspond to major modes of anharmonic motion and describe most of the positional fluctuations; and (2) a nonessential subspace consisting of constrained harmonic motions. By using only the major modes of motion in the essential subspace of the protein it is possible to simulate an approximation to the interaction between a protein and its ligand in a conformational space of much lower dimensionality.

The mathematical formulation we use to determine the collective major modes of motion is the Singular Value Decomposition (SVD) of the displacement matrix derived from a molecular dynamics simulation[25,26]. As an alternative to MD data it is also possible to use ensembles of structures determined experimentally either by X-ray crystallography or by NMR.

The SVD of a matrix, A, is defined as:

$$A = U \Sigma V^T, \quad (1)$$

where U and V are orthonormal matrices and Σ is a nonnegative diagonal matrix whose diagonal elements, σ_i, are the singular values of A. The columns of matrices U and V are called the left and right singular vectors, respectively. Matrix A is constructed by the column-wise concatenation of atomic displacement vectors, for each time sample during a molecular dynamics run. The left singular vectors of the SVD of A will span the space sampled by the protein during the entire simulation. The left singular vectors corresponding to the largest singular values reflect the major modes of motion in the protein and span the essential subspace. The right singular vectors are projections of the dynamics trajectory along the left singular vectors. The advantage of this mathematical transformation is that it changes the basis of representation of our problem. Whereas initially all our degrees of freedom were identically important, using this method we are able to rank our collective degrees of freedom by the order of their eigenvalues. Moreover, this method does not require an intimate knowledge of the system in order to select a few degrees of freedom. The choice is determined by the eigenvalue rank.

The data used in the SVD computation was taken from a 500ps simulation of HIV-1 protease in a box of water molecules using periodical boundary conditions and full electrostatic computation. In Figure 3 we show the rank order eigenvalue spectrum of the SVD analysis of the coordinate data for all 3120 atoms in our protein system. In this plot only the first 30 out of 9360 eigenvalues are shown. The largest eigenvalue accounts for 18% of the cumulative eigenvalue sum and the first 20 account for 63%. It is clear from these values that on the new basis only a few degrees of freedom account for most of the conformational variation. Given this result we are able to approximate the most significant part of the motion in a space with significantly fewer dimensions. One problem with this approximation is that some of the motions in this new conformational space lead to high-

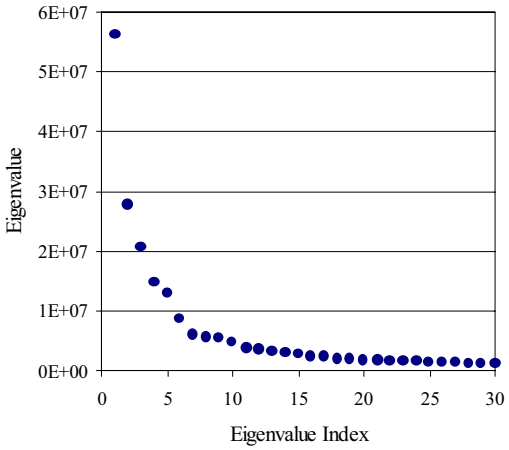

Figure 3 - Rank order eigenvalue spectrum of a SVD analysis of a 500ps MD trajectory.

energy conformations due to distortions in the internal structure of the ligand. We are currently dealing with this problem by performing standard energy minimization methods. We are also developing fast geometry correction methods for the internal structure of the protein which could help eliminate this problem.

In Figure 4 (center representation) we show the backbone representation for HIV-1 protease as determined by X-ray crystallography. The arrows show the mapping of the high dimensional first left singular vector motion into several Cartesian vectors on the backbone of the protein. The directions of the arrows indicate the direction of the motion at that position of the protein and the size of the arrows indicate the relative magnitudes of motion from one region to another. This mapping is in accordance to what would be expected for HIV-1 protease with most of the motion concentrated in flexible loops and on the two flexible flaps that cover the binding site. Using this experimental structure and the first left singular vector we can "actuate" the protein along this degree of freedom in one direction or the other (left and right representations). It is important to note that although the protein is moving as a whole, we are using only one degree of freedom of the new basis to describe that motion. This method can also be used to generate ensembles of structures for rigid docking as described in Section 4.2. In comparison with the ensemble extracted directly from the MD trajectory our method generates a more representative ensemble since the sampling is being done over the most significant degrees of freedom.

5 Discussion

Most of the docking programs presently being used simulate the binding of a flexible ligand to a rigid biological receptor. This model does not reflect the actual physical process of binding and limits or in some cases even prevents the correct identification of potential drug candidates. In this paper we reviewed some of the approaches under research to incorporate protein flexibility in the docking simulation. Some of these approaches have drawbacks such as high computational cost, limited sampling of the receptor conformational space, or require a deep understanding of the biological system making automation difficult. Here we described an alternative method to model protein flexibility based on the SVD of a molecular dynamics trajectory. This procedure is of general applicability, requires a practical amount of computational power and is easily automated.

Our discussion reveals the challenging representational and computational problems that need to be addressed to arrive to efficient molecular docking techniques. We believe that the work done in robotics on kinematics can help in the accurate simulation of protein flexibility and reduce the need of expensive energy minimizations. We also believe that the development of probabilistic path planners that can deal with many degrees of freedom robots will lead to the developmentof planners that lead to the docking of the flexible ligand in a flexible protein [9]

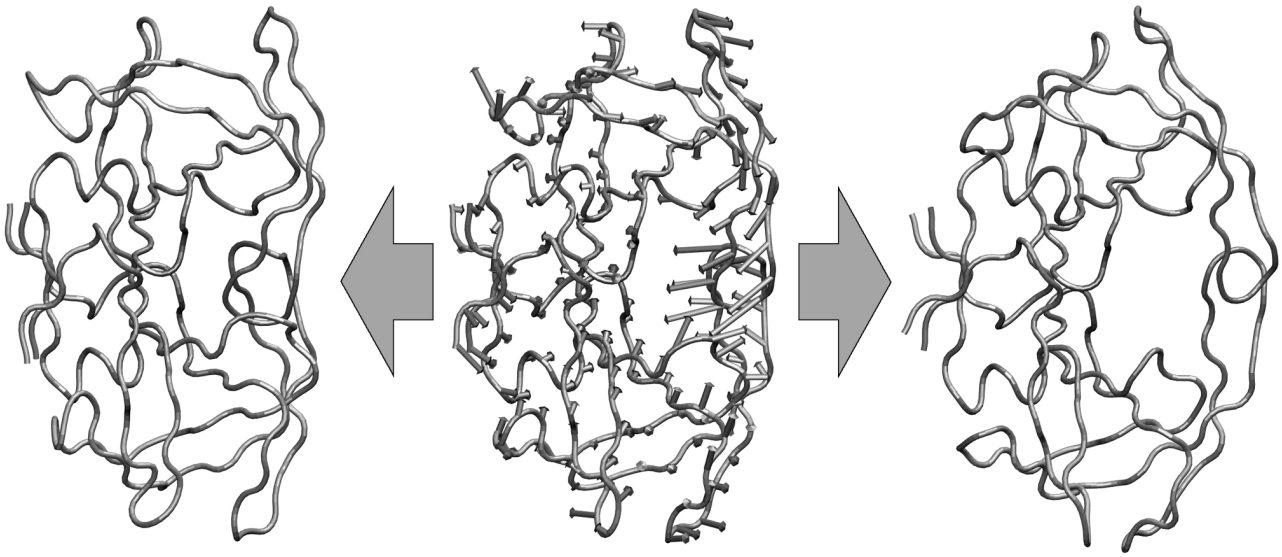

Figure 4 - First collective mode of motion for HIV-1 protease. The backbone structure of the protein as determined experimentally is shown in the center. Arrows indicate the mapping of the first left singular vector in the most significant collective motion. The structure can be "actuated" using this single degree of freedom (left and right structures).

Acknowledgements: Miguel Teodoro is supported by a PRAXIS XXI Pre-doctoral fellowship from the Portuguese Ministry of Science. George Phillips and Miguel Teodoro are partially supported by ATP 003604-0120-1999. Work on this paper by Lydia Kavraki is supported in part by NSF IRI-970228, NSF CISE SA1728-21122N, ATP 003604-0120-1999 and a Sloan Fellowship. All authors are affiliated with the W.M. Keck Center for Computational Biology.

References

1. Vieth, M., Hirst, J.D., Kolinski, A. & Brooks, C.L.I. Assessing energy functions for flexible docking. *J Comp Chem* **19**, 1612-1622 (1998).
2. Cornell, W.D. et al. A second generation force field for the simulation of proteins and nucleic acids. *J Am Chem Soc* **117**, 5179-5197 (1995).
3. MacKerell, A.D., Bashford, D., Bellot, M., Karplus, M. & al, e. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102**, 3586-3616 (1998).
4. Morris, G.M. et al. Automated docking using a Lamarkian genetic algorithm and an empirical binding free energy function. *J Comp Chem* **19**, 1639-1662 (1998).
5. Muegge, I. & Martin, Y.C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* **42**, 791-804 (1999).
6. Ewing, T.J.A. & Kuntz, I.D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J Comp Chem* **18**, 1175-1189 (1997).
7. Kramer, B., Metz, G., Rarey, M. & Lengauer, T. Ligand docking and screening with FlexX. *Med Chem Res* **9**, 463-478 (1999).
8. Bayazit, O.B., Song, G. & Amato, N.M. Ligand Binding with OBPRM and Haptic User Input. in *2001 IEEE International Conference on Robotics and Automation (ICRA'01)* (Seoul, Korea, 2001).
9. Singh, A.P., Latombe, J.C. & Brutlag, D.L. A motion planning approach to flexible ligand binding. in *International Conference on Computational Biology, ISMB* (1999).
10. Brem, R. & Dill, K.A. The effect of multiple binding modes on empirical modeling of ligand docking to proteins. *Prot Sci* **8**, 1134-43. (1999).
11. Murray, C.W., Baxter, C.A. & Frenkel, A.D. The sensitivity of the results of molecular docking to induced fit effects: Application to thrombin, thermolysin and neuraminidase. *J Comput Aided Mol Des* **13**, 547-562 (1999).
12. Jiang, F. & Kim, S.H. "Soft docking": matching of molecular surface cubes. *J Mol Biol* **219**, 79-102. (1991).
13. Jones, G., Willett, P., Glen, R.C., Leach, A.R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**, 727-48 (1997).
14. Apostolakis, J., Pluckthun, A. & Caflisch, A. Docking small ligands in flexible binding sites. *J Comp Chem* **19**, 21-37 (1998).
15. Schnecke, V., Swanson, C.A., Getzoff, E.D., Tainer, J.A. & Kuhn, L.A. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins* **33**, 74-87. (1998).
16. Totrov, M. & Abagyan, R. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins* **Suppl**, 215-20 (1997).
17. Leach, A.R. Ligand docking to proteins with discrete side-chain flexibility. *J Mol Biol* **235**, 345-56 (1994).
18. Mangoni, M., Roccatano, D. & Di Nola, A. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins* **35**, 153-62 (1999).
19. Munshi, S. et al. An alternate binding site for the P1-P3 group of a class of potent HIV- 1 protease inhibitors as a result of concerted structural change in the 80s loop of the protease. *Acta Crystallogr D Biol Crystallogr* **56**, 381-8. (2000).
20. Knegtel, R.M., Kuntz, I.D. & Oshiro, C.M. Molecular docking to ensembles of protein structures. *J Mol Biol* **266**, 424-40. (1997).
21. Clarage, J.B., Romo, T., Andrews, B.K., Pettitt, B.M. & Phillips, G.N., Jr. A sampling problem in molecular dynamics simulations of macromolecules. *Proc Natl Acad Sci U S A* **92**, 3288-92 (1995).
22. Pang, Y.P. & Kozikowski, A.P. Prediction of the binding sites of huperzine A in acetylcholinesterase by docking studies. *J Comput Aided Mol Des* **8**, 669-81. (1994).
23. Sandak, B., Wolfson, H.J. & Nussinov, R. Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. *Proteins* **32**, 159-74 (1998).
24. Amadei, A., Linssen, A.B. & Berendsen, H.J. Essential dynamics of proteins. *Proteins* **17**, 412-25 (1993).
25. Romo, T.D., Clarage, J.B., Sorensen, D.C. & Phillips, G.N., Jr. Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. *Proteins* **22**, 311-21 (1995).
26. Teodoro, M.L., Phillips, G.N. & Kavraki, L.E. Singular Value Decomposition of Protein Conformational Motions. in *Currents in Computational Molecular Biology* (ed. Satoru, M., Shamir, R. & Tagaki, T.) 198-199 (Universal Academy Press, Inc., Tokyo, 2000).