Principles of flexible protein–protein docking

Nelly Andrusier,1‡ Efrat Mashiach,1‡ Ruth Nussinov,2,3 and Haim J. Wolfson1*

1 School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel
2 Basic Research Program, SAIC-Frederick, Inc., Center for Cancer Research Nanobiology Program NCI – Frederick, Frederick, Maryland 21702
3 Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

ABSTRACT

Treating flexibility in molecular docking is a major challenge in cell biology research. Here we describe the background and the principles of existing flexible protein–protein docking methods, focusing on the algorithms and their rational. We describe how protein flexibility is treated in different stages of the docking process: in the pre-processing stage, rigid and flexible parts are identified and their possible conformations are modeled. This pre-processing provides information for the subsequent docking and refinement stages. In the docking stage, an ensemble of pre-generated conformations or the identified rigid domains may be docked separately. In the refinement stage, small-scale movements of the backbone and side-chains are modeled and the binding orientation is improved by rigid-body adjustments. For clarity of presentation, we divide the different methods into categories. This should allow the reader to focus on the most suitable method for a particular docking problem.

INTRODUCTION

Most cellular processes are carried out by protein–protein interactions. Predicting the 3D structures of protein–protein complexes (docking) can shed light on their functional mechanisms and roles in the cell. Understanding and modeling the bound configuration are major scientific challenges. The structures of the complexes provide information regarding the interfaces of the proteins and assist in drug design. Docking can assist in predicting protein–protein interactions, in understanding signaling pathways and in evaluating the affinity of complexes.

Upon binding, proteins undergo conformational changes that include both backbone and side-chain movements. Backbone flexibility can be divided into two major types: large-scale domain motions, such as “shear” and “hinge-bending” motion [Fig. 1(a–d)], and disordered regions such as flexible loops [Fig. 1(e)]. The first docking methods treated proteins as rigid bodies in order to reduce the search space for optimal structures of the complexes. However, ignoring flexibility
Figure 1
Protein flexibility types. (a, b) Shear motion, demonstrated in two conformations of S100 Calcium sensor (PDB-id: 1K9P, 1K9K). The blue helix "slides" on the rest of the protein. (c, d) Hinge motion, demonstrated in two conformations of LAO binding protein (PDB-id: 1LAO, 1LAF). The hinge location is shown as a green sphere. (e) Flexible loop in the ribosomal protein L1 (PDB-id: 1FOX). The different conformations of the loop were determined experimentally by NMR.
could prevent docking algorithms from recovering native associations. Accounting for flexibility is also essential for the accuracy of the solutions. In addition, flexibility must also be taken into account if the docked structures were determined by homology modeling or if loop conformations were modeled.\(^5\)

Incorporating flexibility in a docking algorithm is much more difficult than performing rigid-body docking. The high number of degrees of freedom not only significantly increases the running time, but also results in a higher rate of false-positive solutions. These must be scored correctly in order to identify near-native results.\(^6\) Consequently, existing docking methods limit the flexibility to certain types of motions. In addition, many of these methods allow only one of the proteins in the complex to be flexible.

The general scheme of flexible docking can be divided into four major stages as depicted in Figure 2. The first is a preprocessing stage. In this stage the proteins are analyzed in order to define their conformational space. An ensemble of discrete conformations can be generated from this space and used in further cross-docking, where each protein conformation is docked separately. This process simulates the conformational selection model.\(^7,8\) The analysis can also identify possible hinge locations. In this case the proteins can be divided into their rigid parts and be docked separately. The second is a rigid-docking stage. The docking procedure aims to generate a set of solution candidates with at least one near-native structure. The rigid docking should allow some steric clashes because proteins in their unbound conformation can collide when placed in their native interacting position. The next stage, called refinement, models an induced fit.\(^9\) In this stage, each candidate is optimized by small backbone and side-chain movements and by rigid-body adjustments. It is difficult to simultaneously optimize the side-chain conformations, the backbone structure and the rigid-body orientation. Therefore, the three can be optimized in three separately repeated successive steps. The resulting refined structures have better binding energy and hardly include steric clashes. The final stage is scoring. In this stage the candidate solutions are scored and ranked according to different parameters such as binding energy, agreement with known binding sites, deformation energy of the flexible proteins, and existence of energy funnels.\(^10,11\) The goal of this important stage is to identify the near-native solutions among the candidates. In this review we do not address the scoring function problem. A detailed review of scoring schemes was published by Halperin et al.\(^3\)

Two flexible docking reviews were published recently.\(^12,13\) These articles present a variety of docking methods that incorporate protein flexibility. However, we believe that a more detailed and comprehensive review would be useful for the docking community. Our review provides a description of the computational algorithms behind the docking methods. The division into categories used in this article can help the reader choose the most suitable method for a particular docking problem.

Our review includes three major parts, corresponding to different flexible docking procedures. The first part describes protein flexibility analysis methods. The second part discusses the treatment of backbone flexibility in current docking algorithms. The third, side-chain refinement part reviews methods for prediction of bound side-chain conformations. Finally, existing methods that handle both backbone and side-chain flexibility are described.

![Figure 2](image_url)

**Figure 2**

A general scheme of flexible docking procedure. (a) Methods for protein flexibility analysis are described in the “Protein Flexibility Analysis” section. (b) Rigid docking with soft interface, ensemble docking of different conformations and backbone refinement methods are described in the “Handling Backbone Flexibility in Docking Methods” section. (c) Side-chain refinement methods are described in the “Handling Side-Chain Flexibility in Docking Methods” section. (d) Rigid body optimization methods are mentioned in the Discussion.
**Protein Flexibility Analysis**

Protein flexibility analysis methods, reviewed below, can be classified into three major categories:

1. Methods for generating an ensemble of discrete conformations. Ensembles of conformations are widely used in cross-docking and in the refinement stage of the docking procedure. The different conformations can be generated by analyzing different experimentally solved protein structures or by using Molecular Dynamics (MD) simulation snapshots.

2. Methods for determining a continuous protein conformational space. The conformational space can be used as a continuous search space for refinement algorithms. In addition, many flexible docking methods sample this pre-calculated conformational space in order to generate a set of discrete conformations. This group of methods includes Normal Modes Analysis (NMA) and Essential Dynamics.

3. Methods for identifying rigid and flexible regions in the protein. These methods include the rigidity theory and hinge detection algorithms.

**Conformational ensemble analysis**

Using different solved 3D structures (by X-ray and NMR) of diverse conformations of the same protein, or of homologous proteins, is probably the most convenient way to obtain information relating to protein flexibility. Using such conformers, one can generate new viable conformations which might exist during the transition between one given conformation to another. These new conformations can be generated by ‘morphing’ techniques which implement linear interpolations, but have limited biological relevance.

Known structures of homologs or of different conformations of the same protein can also be useful in detecting rigid domains and hinge locations. Boutonnet et al. developed one of the first methods for an automated detection of hinge and shear motions in proteins. The method uses two conformations of the same protein. It identifies structurally similar segments and aligns them. Then, the local alignments are hierarchically clustered to generate a global alignment and a clustering tree. Finally the tree is analyzed to identify the hinge and shear motions. The DynDom method uses a similar clustering approach for identifying hinge points using two protein conformations. Given the set of atom displacement vectors, the rotation vectors are calculated for each short backbone segment. A rotation vector can be represented as a rotation point in a 3D space. A domain that moves as a rigid-body will produce a cluster of rotation points. The method uses the K-means clustering algorithm to determine the clusters and detect the domains. Finally, the hinge axis is calculated and the residues involved in the inter-domain bending are identified.

The HingeFind method can also analyze structures of homolog proteins in different conformations and detects rigid domains, whose superimposition achieves RMSD of less than a given threshold. It requires sequence alignment of two given protein structures. The procedure starts with each pair of aligned Ca atoms and iteratively tries to extend them by adding adjacent Ca atoms as long as the RMSD criterion holds. After all the rigid domains are identified, the rotation axes between them are calculated. Verbitsky et al. used the geometric hashing approach to align two molecules, and detect hinge motifs. The method can match the motifs independently of the order of the amino acids in the chain. The more advanced FlexProt method searches for 3D congruent rigid fragment pairs in structures of homolog proteins, by aligning every Ca pair and trying to extend the 3D alignment, in a way similar to HingeFind. Next, an efficient graph-theory method is used for the connection of the rigid parts and the construction of the full solution with the largest correspondence list, which is sequence-order dependent. The construction simultaneously detects the locations of the hinges.

**Molecular dynamics**

Depending on the time scales and the energy barrier heights, molecular dynamics simulations can provide insight into protein flexibility. Molecular dynamics (MD) simulations are based on a force field that describes the forces created by chemical interactions. Throughout the simulation, the motions of all atoms are modeled by repeatedly calculating the forces on each atom, solving Newton’s equation and moving the atoms accordingly. Di Nola et al. were first to incorporate explicit solvent molecules into MD simulations while docking two flexible molecules. Pak et al. applied MD using Tsallis effective potential for the flexible docking of few complexes.

Molecular dynamics simulations require long computational time scales and therefore are limited in the motion amplitudes. For this reason they can be used for modeling only relatively small-scale movements, which take place in nanosecond time scales, while conformational changes of proteins often occur over a relatively long period of time (~1 ms). One way to speed up the simulations is by restricting the degrees of freedom to the torsional space, which allows larger integration time steps. Another difficulty is that the existence of energy barriers may trap the MD simulation in certain conformations of a protein. This problem can be overcome by using simulated annealing and scaling methods during the simulation. For example, simulated annealing MD is used in the refinement stage of HADDOCK in order to refine the conformations of both the side-chains and the backbone. Riemann et al. applied poten-
tial scaling during MD simulations to predict side chain conformations.4

In order to sample a wide conformational space and search for conformations at local minima in the energy landscape, biased methods, which were previously reviewed,35 can be used. The flooding technique,36 which is used in the GROMACS method,37 fills the “well” of the initial conformation in the energy landscape with a Gaussian shape “flooding” potential. Another similar method, called puddle-jumping,38 fills this well up to a flat energy level. These methods accelerate the transition across energy barriers and permit scanning other stable conformations.

Normal modes

Normal Modes Analysis (NMA) is a method for calculating a set of basis vectors (normal modes) which describes the flexibility of the analyzed protein.39–41 The length of each vector is $3N$, where $N$ is the number of atoms or amino acids in the protein, depending on the resolution of the analysis. Each vector represents a certain movement of the protein such that any conformational change can be expressed as a linear combination of the normal modes. The coefficient of a normal mode represents its amplitude.

A common model used for normal modes calculation is the Anisotropic Network Model (ANM), which was previously described in detail.40,42 This is a simplified spring model which relies primarily on the geometry and mass distribution of a protein (Fig. 3). Every two atoms (or residues) within a distance below a threshold are connected by a spring (usually all springs have a single force constant). The model treats the initial conformation as the equilibrium conformation.

The normal modes describe continuous motions of the flexible protein around a single equilibrium conformation. Theoretically, this model does not apply to proteins which have several conformational states with local free-energy minima. However, in practice, normal modes suit very well conformational changes observed between bound and unbound protein structures.43 Another advantage of the normal modes analysis is that it can discriminate between low and high frequency modes. The low frequency modes usually describe the large scale motions of the protein. It has been shown43,44 that the first few normal modes, with the lowest frequencies, can describe much of a conformational change. This allows reducing the degrees of freedom considerably while preserving the information about the main characteristics of the protein motion. Therefore, many studies45–47 use a subset of the lowest frequency modes for analyzing the flexibility of proteins. The normal modes can further be used for predicting hinge-bending movements,48 for generating an ensemble of discrete conformations49 and for estimating the protein’s deformation energy resulting from a conformational change.50,51

Tama and Sanejouand44 showed that normal modes obtained from the open form of a protein correlate better with its known conformational changes, than the ones obtained from its closed form. In a recent work, Petrone and Pande43 showed that the first 20 modes can improve the RMSD to the bound conformation by only up to 50%. The suggested reason was that while the unbound conformation moves mostly according to low frequency modes, the binding process activates movements related to modes with higher frequencies.

The binding site of proteins often contains loops which undergo relatively small conformational changes triggered by an interaction. This phenomenon is common in protein kinase binding pockets. Loop movements can only be characterized by high-frequency normal modes. Therefore, we would like to identify the modes which influence these loops the most, in order to focus...
on these in the docking process. For this reason, Cava- 
sotto et al.\textsuperscript{52} have introduced a method for measuring 
the relevance of a mode to a certain loop. This measure 
of relevance favors modes which bend the loop at its 
edges, and significantly moves the center of the loop. It 
excludes modes which distort the loop or move the loop 
together with its surroundings. This measure was used to 
identify modes which are relevant to loops 
within the binding sites of two cAMP-dependent protein 
kinesins (cAPKs). For each loop less than 10 normal 
ones were recomputed after each small displacement.\textsuperscript{53–55} 
This is an accurate but time-consuming method. Kirillova et al.\textsuperscript{56} have recently developed a NMA-guided 
method for exploring the conformational space 
spanned by 10–30 low frequency normal modes. This ef-
cient method requires selecting only a relatively small number of norm-
al modes to compute large conformational changes.

Since NMA is based on an approximation of a potential 
edge in a specific starting conformation, its accuracy 
deteriorates when modeling large conformational changes. Therefore in some studies, the normal modes 
were recomputed after each small displacement.\textsuperscript{53–55} 
This is an accurate but time-consuming method. The analysis uses the topology of the spring network for cal-
culating the amplitudes of the normal modes and the 
correlations between the fluctuations of each pair of resi-
dues. However, the direction of each fluctuation cannot 
be found by GNM. This analysis is more efficient both in 
CPU time and in memory than the ANM analysis and 
therefore it can be applied on larger systems. The draw-
back is that the GNM cannot be used for generating alter-
native conformational states for these proteins, which 
were later used for docking. The method succeeded in 
docking two small ligands which could not be docked to 
the unbound conformations of the cAPKs due to steric clashes.

Since NMA is based on an approximation of a potential 
edge in a specific starting conformation, its accuracy 
deteriorates when modeling large conformational changes. Therefore in some studies, the normal modes 
were recomputed after each small displacement.\textsuperscript{53–55} 
This is an accurate but time-consuming method. The analysis uses the topology of the spring network for cal-
culating the amplitudes of the normal modes and the 
correlations between the fluctuations of each pair of resi-
dues. However, the direction of each fluctuation cannot 
be found by GNM. This analysis is more efficient both in 
CPU time and in memory than the ANM analysis and 
therefore it can be applied on larger systems. The draw-
back is that the GNM cannot be used for generating alter-
native conformational states for these proteins, which 
were later used for docking. The method succeeded in 
docking two small ligands which could not be docked to 
the unbound conformations of the cAPKs due to steric clashes.

The Gaussian Network Model (GNM) is another sim-
pified version of normal modes analysis.\textsuperscript{57,58} The GNM 
analysis uses the topology of the spring network for cal-
culating the amplitudes of the normal modes and the 
correlations between the fluctuations of each pair of resi-
dues. However, the direction of each fluctuation cannot 
be found by GNM. This analysis is more efficient both in 
CPU time and in memory than the ANM analysis and 
therefore it can be applied on larger systems. The draw-
back is that the GNM cannot be used for generating alter-
native conformational states for these proteins, which 
were later used for docking. The method succeeded in 
docking two small ligands which could not be docked to 
the unbound conformations of the cAPKs due to steric clashes.

The Essential dynamics approach aims at capturing the 
main flexible degrees of freedom of a protein, given a set 
of its feasible conformations.\textsuperscript{59} These degrees of freedom 
are described by vectors which are often called essential 
ones, or principal components (PC). The conformation 
set is used to construct a square covariance matrix (3N 
× 3N, where N is the number of atoms) of the deviation 
of each atom coordinates from its unbound position or, 
alternatively, average position. This matrix is then dia-
ognalized and its eigenvectors and eigenvalues are found. 
These eigenvectors represent the principal components of the protein flexibility. The bigger the eigenvalues, the 
larger the amplitude of the fluctuation described by its 
eigenvector.

Mustard and Ritchie\textsuperscript{60} used this essential dynamics 
approach to generate realistic starting structures for 
docking, which are called eigenstructures. The covariance 
matrix was created according to a large number of con-
formations, generated using the CONCOORD pro-
gram,\textsuperscript{61} which randomly generates 3D protein conforma-
tions that fulfill distance constraints. The eigenvectors 
were calculated and it has been shown\textsuperscript{60} that the first 
2 or more (with the largest eigenvalues) can account 
for many of the backbone conformational changes that 
occur upon binding in seven different CAPRI targets 
from rounds 3–5.\textsuperscript{52} Linear combinations of the first 
eight eigenvectors were later used to generate eigenstruc-
tures from each original unbound structure of these 
CAPRI targets. An experiment that used these eigenstruc-
tures in rigid docking showed improvements in the 
results compared to using the unbound structure or a 
model-built structure.\textsuperscript{60} An ensemble of conformations 
can be generated in a similar way by the Dynamite soft-
ware,\textsuperscript{63} which also applies the essential dynamics 
approach to a set of conformations generated by CON-
COORD.

Principle component analysis (PCA) can also be based 
on molecular dynamics simulations. Unlike normal mode 
analysis, this PCA includes the effect of the surrounding 
water on the flexibility. However, the results of the analysis 
strongly depend on the simulation’s length and con-
vergence. It has been shown that most of the conforma-
tional fluctuations observed by MD simulations\textsuperscript{59} and 
some known conformational changes between unbound 
and bound forms\textsuperscript{64} can be described with only few 
PCs.

Rigidity theory

Jacobs et al.\textsuperscript{65} developed a graph-theory method 
which analyzes protein flexibility and identifies rigid and 
flexible substructures. In this method a network is con-
structed according to distance and angle constraints, 
which are derived from covalent bonds, hydrogen bonds 
and salt bridges within a single conformation of a pro-
tein. The vertices of the network represent the atoms 
and the edges represent the constraints. The analysis of the 
network resembles a pebble game. At the beginning of 
the algorithm, each atom (vertex) receives three pebbles 
which represent three degrees of freedom (translation in 
3D). The edges are added one by one and each edge con-
sumes one pebble from one of its vertices, if possible. It is possible to rearrange the pebbles on the graph as long as the following rules hold: (1) Each vertex is always associated with exactly three pebbles which can be consumed by some of its adjacent edges. (2) Once an edge consumes a pebble it must continue holding a pebble from one of its vertices throughout the rest of the algorithm. At the end of the algorithm the remaining pebbles can still be rearranged but the specified rules divide the protein into areas in such a way that the pebbles can not move from one area to another.

The number of remaining degrees of freedom in a certain area of the protein quantifies its flexibility. For example, a rigid area will not possess more than 6 degrees of freedom (which represent translation and rotation in 3D). The algorithm can also identify hinge points, and rigid domains which are stable upon removal of constraints like hydrogen-bonds and salt bridges. This pebbles-game algorithm is implemented in a software called FIRST (Floppy Inclusion and Rigid Substructure Topography) which analyzes protein flexibility in only a few seconds of CPU time. The algorithm was tested on HIV protease, dihydrofolate reductase and adenylate kinase and was able to predict most of their functionally important flexible regions, which were known beforehand by X-ray and NMR experiments.

**HANDLING BACKBONE FLEXIBILITY IN DOCKING METHODS**

Treating backbone flexibility in docking methods is still a major challenge. The backbone flexibility adds a huge number of degrees of freedom to the search space and therefore makes the docking problem much more difficult. The docking methods can be divided into four groups according to their treatment of backbone flexibility. The first group uses soft interface during the docking and allows some steric clashes in the resulting complex models. The second performs an ensemble docking, which uses feasible conformations of the proteins, generated beforehand. The third group deals with hinge bending motions, and the last group heuristically searches for energetically favored conformations in a wide conformational search space.

**Soft interface**

Docking methods that use soft interface actually perform relatively fast rigid-body docking which allows a certain amount of steric clashes (penetration). These methods can be divided into three major groups: (i) brute force techniques that can be significantly speeded up by FFT, (ii) randomized methods and (iii) shape complementarity methods. This approach can only deal with side chain flexibility and small scale backbone movements. It is assumed that the proteins are capable of performing the required conformational changes which avoid the penetrations, although the actual changes are not modeled explicitly. Since the results of this soft docking usually contain steric clashes, a further refinement stage must be used in order to resolve them.

**Ensemble docking**

In order to avoid the search through the entire flexible conformational space of two proteins during the docking or refinement process, the ensemble docking approach samples an ensemble of different feasible conformations prior to docking. Next, docking of the whole ensemble is performed. The different conformations can be docked one by one (cross-docking), which significantly increases the computational time, or all together using different algorithms such as the mean-field approach presented below.

The ensemble may include different crystal structures and NMR conformers of the protein. Other structures can be calculated using computational sampling methods which are derived from the protein flexibility analysis (molecular dynamics, normal modes, essential dynamics, loop modeling, etc). Feasible structures can also be sampled using random-search methods such as Monte Carlo and genetic algorithms.

The search for an optimal loop conformation can be performed during the docking procedure by the mean-field approach. In this method, a set of loop conformations is sampled in advance and each conformation is initialized by an equal weight. Throughout the docking, in each iteration, the weights of the conformations (copies) change according to the Boltzmann criterion, in a way that a conformation receives a higher weight if it achieves a lower free energy. The partner and the rest of the protein which interact with the loop “feel” the weighted average of the energies of their interactions with each conformation in the set. The algorithm usually converges to a single conformation for each loop, with a high weight.

Bastard et al. used the mean-field approach in their MC2 method which is based on multiple copy representation of loops and Monte Carlo (MC) conformational search. Viable loop conformations were created using a combinatorial approach, which randomly selected common torsional angles for the loop backbone. In each MC step the side-chains dihedral angles and the rotation and translation variables are randomly chosen. Then, the weight of each loop is adjusted according to its Boltzmann probability. The performance of the MC2 algorithm was evaluated on the solved protein-DNA complex of a Drosophila Prd-paired-domain protein, which interacts with its target DNA segment by a loop of seven residues. 23% of the MC2 simulations produced results.
in which the RMSD was lower than 1.5 Å, and included the selection of a loop conformation which was extremely similar to the native one. Furthermore, these results got much better energy scores than the other 77%, therefore they could be easily identified.

In a later work, the mean-field approach was introduced in the ATTRACT software and was tested on a set of eight protein–protein complexes in which the receptor undergoes a large conformational change upon binding or its solved unbound structure has a missing loop at its interaction site. The results showed that the algorithm improved the docking significantly compared to rigid docking methods.

**Modeling hinge motion**

Hinge-bending motions are common during complex creation. Hinges are flexible segments which separate relatively rigid parts of the proteins, such as domains or subdomains.

Sandak et al. introduced a method which deals with this type of flexibility. The algorithm allows multiple hinge locations, which are given by the user. Hinges can be given for only one of the interacting proteins (e.g. the ligand). The algorithm docks all the rigid parts of the flexible ligand simultaneously, using the geometric hashing approach.

The FlexDock algorithm is a more advanced method for docking with hinge-bending flexibility in one of the interacting proteins. The locations of the hinges are automatically detected by the HingeProt algorithm. The number of hinges is not limited and does not affect the running time complexity. However, the hinges must impose a chain-type topology, that is the subdomains separated by the hinges must form a linear chain. The algorithm divides the flexible protein into subdomains at its hinge points. These subdomains are docked separately to the second protein by the PatchDock algorithm. Then, an assembly graph is constructed. Each node in the graph represents a result of a subdomain rigid docking (a transformation), and the node is assigned a weight according to the docking score. Edges are added between nodes which represent consistent solutions of consecutive rigid subdomains. An edge weight corresponds to the shape complementarity score between the two subdomains represented by its two nodes. Finally, the docking results of the different subdomains are assembled to create full consistent results for the complex using an efficient dynamic programming algorithm that finds high scoring paths in the graph. This approach can cope with very large conformational changes. Among its achievements, it has predicted the bound conformation of calmodulin to a target peptide, the complex of Replication Protein A with a single stranded DNA as shown in Figure 4, and has created the only acceptable solution for the LicT dimer at the CAPRI challenge (Target 9).

Ben-Zeev et al. have coped with the CAPRI challenges which included domain movements (Target 9, 11, and 13) by a rigid body multi-stage docking procedure. Each of the proteins was partitioned into its domains. Then, the domains of the two proteins were docked to each other in all possible order of steps. In each step, the current domain was docked to the best results from the previous docking step.

This multistage method requires that the native position of a subdomain will be ranked high enough in each step. This restriction is avoided in the FlexDock algorithm which in the assembling stage uses a large number of docking results for each subdomain. Therefore, a full docking result can be found and be highly ranked even if its partial subdomain docking results were poorly ranked.

**Refinement and minimization methods for treating backbone flexibility**

Fitzjohn and Bates used a guided docking method, which includes a fully flexible refinement stage. In the refinement stage CHARMM22 all-atom force field was used to move the individual atoms of the receptor and the ligand. In addition, the forces acting on each atom were summed and converted into a force on the center-of-mass of each molecule.

Lindahl and Delarue introduced a new refinement method for docking solutions which minimizes the interaction energy in a complex along 5–10 of the lowest frequency normal modes' directions. The degrees of freedom in the search space are the amplitudes of the normal modes, and a quasi-Newtonian algorithm is used for the energy minimization. The method was tested on protein-ligand and protein-DNA complexes and was able to reduce the RMSD between the docking model and the true complex by 0.3–3.6 Å.

In a recent work, May and Zacharias accounted for global conformational changes during a systematic docking procedure. The docking starts by generating many thousands of rigid starting positions of the ligand around the receptor. Then, a minimization procedure is performed on the six rigid degrees of freedom and on five additional degrees of freedom which account for the coefficients of the five, pre-calculated, slowest frequency normal modes. The energy function includes a penalty term that prevents large scale deformations. Applying the method to several test cases showed that it can significantly improve the accuracy and the ranking of the results. However the side-chain conformations must be refined as well. The method was recently incorporated into the ATTRACT docking software.

A new data structure called Flexibility Tree (FT) was recently presented by Zhao et al. The FT is a hierarchi-
cal data structure which represents conformational subspaces of proteins and full flexibility of small ligands. The hierarchical structure of this data structure enables focusing solely on the motions which are relevant to a protein binding site. The representation of protein flexibility by FT combines a variety of motions such as hinge bending, flexible side-chain conformations and loop deformations which can be represented by normal modes or essential dynamics. The FT parameterizes the flexibility subspace by a relatively small number of variables. The values of these variables can be searched in order to find the minimal energy solution. The FLIPDock\textsuperscript{92} method uses two FT data structures, representing the flexibility of both the ligand and the receptor. The right conformations are then searched using a genetic algorithm and a divide and conquer approach, during the docking process.

Many docking methods use Monte Carlo methods in the final minimization step. For example, Monte Carlo minimization (MCM) is used in the refinement stage of RosettaDock.\textsuperscript{93,94} Each MCM iteration consists of three steps: (1) random rigid-body movements and backbone perturbation, in certain peptide segments which were chosen to be flexible according to a flexibility analysis performed beforehand; (2) rotamer-based side-chain refinement; (3) quasi-Newton energy minimization for relatively small changes in the backbone and side-chain torsional angles, and for minor rigid-body optimization.

Some docking methods\textsuperscript{93} simply ignore flexible loops during the docking and rebuild them afterwards in a loop modeling step.

**Handling Side-Chain Flexibility in Docking Methods**

The majority of the methods handle side-chain flexibility in the refinement stage. Each docking candidate is optimized by side-chain movements. Figure 5 shows a non-optimized conformation of a ligand residue, which clashes with the receptor's interface, and a correct prediction of its bound conformation by a side-chain optimization algorithm. Most conformational changes occur in the interface between the two binding proteins. Therefore, many methods try to predict side-chain conformational changes for a given backbone structure in the interaction area. The problem has been widely studied in the more general context of side-chain assignment on a fixed backbone in the fields of protein design and homology modeling. Therefore, all the algorithms reviewed in this section apply to side-chain refinement in both folding and docking methods.

To reduce the search space, most of the methods use rotamer discretization. Rotamer libraries are derived from statistical analysis of side-chain conformations in known high-resolution protein structures. Backbone-dependent rotamer libraries contain information on side-chain dihedral angles and rotamer populations dependent on the backbone conformation.\textsuperscript{96} Usually, unbound conformations of side-chains are added to the set of conformers for each residue. In this way a side-chain can remain in its original state if the unbound conformer is chosen by the optimization algorithm.
Global optimization algorithms for side-chain refinement

The side-chain prediction problem can be treated as a combinatorial optimization problem. The goal is to find the combination of rotamer assignments for each residue with the global minimal energy, denoted as GMEC (Global Minimal Energy Conformation). The energy value of GMEC is calculated as follows:

$$E_{GMEC} = \min_{r,s} \left( \sum_i E(i) + \sum_{ij} E(i,r) + E(j,s) \right)$$

(1)

where $E(i)$ is the self energy of the assignment of rotamer $r$ for residue $i$. It includes the interaction energy of the rotamer with a fixed environment. $E(i, r)$ is the pair-wise energy between rotamer $r$ of residue $i$ and rotamer $s$ of residue $j$. For each residue one rotamer should be chosen, and the overall energy should be minimal. This combinatorial optimization problem was proved to be NP-hard and inapproximable. In practice, topological restraints of residues can facilitate the problem solution.

In branch-and-bound algorithms all possible conformations are represented by a tree. Each level of the tree represents a different residue and the order of the nodes at this level is the number of possible residue rotamers. Scanning down the tree and adding self and pairwise energies at each level will sum up to the global energy values at the leaves. A branch-and-bound algorithm can be performed by using a bound function. A proposed bound function is defined for a certain level, and yields a lower bound of energy, obtainable from any branch below this level. This level bound function is added to the cumulative energy in the current scanned node and the branch can be eliminated if the value is greater than a previously calculated leaf energy.

The dead-end elimination (DEE) method is based on pruning the rotamers, which are certain not to participate in GMEC, because better alternatives can be chosen. The Goldstein DEE criterion removes a rotamer from further consideration if another rotamer of the same residue has a lower energy for every possible rotamer assignment for the rest of the residues. A more powerful criterion for dead-end elimination is proposed in the “split DEE” method (Fig. 6). Many methods use DEE as a first stage in order to reduce a conformational search space.

In addition to the rotamer reduction method by DEE, many methods also use a residue reduction procedure, which eliminates residues with a single rotamer or with up to two interacting residues (neighbors). A residue with a single rotamer can be eliminated from further consideration by incorporating its pairwise energies into the self energies of its neighbors. A residue with one neighbor can be reduced by adding its rotamer energies to the self-energies of its neighbor’s rotamers. A residue with two neighbors can be eliminated by updating the pair-wise energies of the neighbors. The Residue-
Rotamer-Reduction (R3) method\textsuperscript{107} repeatedly performs residue and rotamer reduction. When a reduction is not possible in a certain iteration, the R3 method performs residue unification.\textsuperscript{103,108} In this procedure, two residues are unified and a set of all their possible rotamer pairs is generated. The method finds the GMEC in a finite number of elimination iterations, because at least one residue is reduced in each iteration.\textsuperscript{107}

Bahadur \textit{et al.}\textsuperscript{109} have defined a weighted graph of non-colliding rotamers. In this graph the vertices are rotamers and two rotamers are connected by an edge if they represent different residues that do not have steric clashes. The weights on the edges correspond to the strength of the interaction between two rotamers. The algorithm searches for the maximum edge-weight clique in the induced graph. If the size of the obtained clique equals the number of residues, then each residue is assigned with exactly one rotamer. Since each two nodes in the clique are connected, none of the chosen rotamers collide. Thus, the obtained clique defines a feasible conformation and the maximum edge-weight clique corresponds to the GMEC.

The SCWRL\textsuperscript{101} algorithm uses a residue interaction graph in which residues with clashing rotamers are connected. The resulting graph is decomposed to biconnected components \textbf{[see Fig. 7(a,b)]} and a dynamic programming technique is applied to find a GMEC. Any two components include at most one common residue -- the \textit{articulation point}. It starts by optimizing the leaf components, which have only one articulation point. The component’s GMEC is calculated for each rotamer of the articulation point and is stored as the energy of the compatible rotamer for further GMEC calculations of adjacent components. Figure 7(a,b) demonstrate the decomposition of a residue interaction graph into components. The drawback of the method is that it might include large components, which increase dramatically the CPU time. SCATD\textsuperscript{106} proposes an improvement of the SCWRL methodology by using a tree decomposition of the residue interaction graph. This method results in more balanced decomposition and prevents creation of huge components, as opposed to biconnected decomposition. After this decomposition, any two components can share more than one common residue \textbf{[Fig. 7(c)]}. Therefore, the component GMEC is calculated for every possible combination of these common residues and stored for further calculations.

Recent methods use the mixed-integer linear programming (MILP) framework\textsuperscript{110–113} to find a GMEC. In general, a decision variable is defined for each rotamer and rotamer-rotamer interaction. If a rotamer participates in GMEC, its corresponding decision variables will be equal to 1. Each decision variable is weighted by its score (self and pair-wise energies) and summed in a global linear expression for minimization. Constraints are set in order to guarantee one rotamer choice for each residue, and that only pair-wise energies between the selected rotamers are included in the global minimal energy. Although the MILP algorithm is NP-hard, by relaxation of the integrality condition on the decision variables, the polynomial-complexity linear programming algorithm can be applied to find the minimum. If the solution happens to be integral, the GMEC is found in polynomial time. Otherwise, an integer linear programming algorithm, with significantly longer running time, is applied. The MILP framework allows obtaining successive near-optimal solutions by addition of constraints that exclude the previously found optimal set of rotamers.\textsuperscript{112} The FireDock\textsuperscript{113} method for refinement and scoring of docking candidates uses the MILP technique for side-chain optimization. An example of a successful rotamer assignment by FireDock is shown in Figure 5.

In general, for all methods which use pair-wise energy calculations, a prefix tree data structure (trie) can be used for saving CPU time.\textsuperscript{114} In a trie data structure, the inter-atomic energies of rotamers’ parts, which share the same torsion angles, are computed once.

Many of the described methods efficiently find a GMEC due to the use of a simplified energy function, which usually includes only the repulsive van der Waals and rotamer probability terms. The energy function can be extended by additional terms, like the attractive van der Waals, solvation and electrostatics. However, this complicates the problem. The SCWRL/SCATD graph decomposition results in larger components, the number of decision variables in the MILP technique increases, and so forth. For example, Kingsford \textit{et al.}\textsuperscript{112} use only van der Waals and rotamer probability terms and almost always succeed in finding the optimal solution by polynomial LP. However, when adding electrostatic term, non-polynomial ILP is often required.
A performance comparison of \( R^3 \), \( SCWRL \), and \( MILP \) methods was performed. The first test set included 25 proteins. The differences in the prediction ability of the methods were minor, since they all find a GMEC and use a similar energy function. The time efficiency of the \( R^3 \) and \( SCWRL \) methods was better than \( MILP \) for these cases. The second test set of 5 proteins was harder and the \( R^3 \) method performed significantly faster than \( SCWRL \) and \( MILP \). In addition, \( Xu \) demonstrated that the \( SCATD \) method shows a significant improvement in CPU time compared to \( SCWRL \) for the second test set.

Heuristic methods for treating side-chain flexibility

Heuristic algorithms are widely used in side-chain refinement methods because of the following reasons. First, a continuous conformational space can be used during the minimization, as opposed to global optimization algorithms, where the conformational space has to be reduced to a pre-defined discrete set of conformers. Second, different energy functions can be easily incorporated into heuristic algorithms, while global optimization methods usually require a simplified energy function. A third advantage is that heuristic algorithms can provide many low-energy solutions, while most of the global algorithms provide a single one. However, the main drawback of the heuristic methods is that they cannot guarantee finding the GMEC.

Monte Carlo (MC) is an iterative method. At each step it randomly picks a residue and switches its current rotamer by another. The new overall energy is calculated and the conformational change is accepted or rejected by the Metropolis criterion. In simulated annealing MC, the Boltzman temperature is high at the beginning to overcome local minima. Then, it is gradually lowered in order to converge to the global minimum. Finally, a quench step can be performed. The quench step cycles through the residues in a random order, and for each residue, the best rotomer for the overall energy is chosen. RosettaDock uses this rotamer-based MC approach and, in addition, performs gradient-based minimization in torsion space of dihedral angles.

The self-consistent mean-field (SCMF) optimization method uses a matrix which contains the probability of each rotamer to be included in the optimal solution. Each rotamer probability is calculated by the sum of its interaction energies with the surrounding rotamers, weighted by their respective probabilities. The method iteratively refines this matrix and converges in a few cycles. The 3D-DOCK package uses the mean-field approach for side-chain optimization with surrounding solvent molecules.

Other optimization techniques like genetic algorithms and neural networks are also applied to predict optimal conformations of side-chains. Several methods do not restrict the conformational search to rotamers. \( Abagyan et al. \) apply the biased probability MC method for minimization in the torsion angles space. Molecular dynamics simulations (described in Section Molecular dynamics) are also used to model flexibility of side-chains. SmoothDock uses short MD simulations to predict conformations of anchor side-chains at the pre-docking phase. HADDOCK uses restricted MD simulations for final refinement with explicit solvent.

Obviously, an energy function has great influence on side-chain prediction performance. Yanover showed that finding a GMEC does not significantly improve side-chain prediction results compared to the heuristic RosettaDock side-chain optimization. They showed that using an optimized energy function has much greater influence on the performance than using an improved search strategy.

Recent studies have shown that most of the interface residues do not undergo significant changes during binding. Therefore, changing unbound conformations should be performed carefully during the optimization process. In addition, when analyzing the performance of side-chain optimization methods, unbound conformations of side-chains should be used as a reference.

Handling both backbone and side-chain flexibility in recent CAPRI challenges

In recent CAPRI (Critical Assessment of PRediction of Interactions) challenges, some of the participating groups attempted to handle both backbone and side-chain flexibility. Many groups treated conformational deformations by generating ensembles of conformations, which were later used for cross-docking. Additionally, some methods, specified below, handle protein flexibility during the docking process or in a refinement stage.

The group of Bates used MD for generating ensembles of different conformations for the receptor and the ligand. Then, rigid body cross-docking was performed by the FTdock method. The best results were minimized by CHARMM and clustered. Finally, a refinement by MD was performed. It has been shown that the cross-docking produced more near native results compared to unbound docking only in cases where the proteins undergo large conformational changes upon binding.

Similar conclusions were obtained by Smith et al. The ATTRACT docking program uses a reduced protein model, which represents each amino acid by up to three pseudo atoms. For each starting orientation, energy minimization is performed on six rigid-body degrees of freedom and on additional five degrees of freedom derived from the five lowest frequency normal modes.
modes. Finally, the side-chain conformations at the interface of each docking solution are adjusted using the Swiss-PdbViewer\textsuperscript{137} and the Sander program from the Amber8 package\textsuperscript{138} is used for a final minimization.

The RosettaDock method\textsuperscript{93,94} performs an initial low-resolution global docking, which includes a Monte Carlo (MC) search with random backbone and rigid-body perturbations. The low energy docking candidates are further refined by Monte Carlo minimization (MCM). Each MCM cycle consists of: (i) backbone and/or rigid-body perturbation, (ii) rotamer-based side-chain optimization and (iii) quasi-Newton minimization on the degrees of freedom.
freedom of the backbone and/or side-chains and/or rigid-body orientation.

The HADDOCK protocol\textsuperscript{33} consists of rigid-body docking followed by a semi-flexible refinement of the interface in torsion angles’ space (of both backbone and side-chains). As a final stage, a Cartesian dynamics refinement in explicit solvent is performed.

To conclude, the treatment of internal flexibility can be performed in different stages of the docking process and in different combinations. In many cases, backbone flexibility is treated before the side-chain flexibility. For example, an ensemble of backbone conformations is often created before the docking procedure. In addition, some methods, like ATTRACT and RosettaDock, perform backbone minimization prior to further refinement. There are two reasons for this order of handling flexibility: (i) the backbone deformations have greater influence on the protein structure than the side-chain movements; (ii) side-chain conformations often depend on the backbone torsion angles. On the other hand, in the final refinement stage, leading docking groups attempt to parallelize the treatment of all the degrees of freedom, including full internal flexibility and rigid-body orientation. CAPRI challenges still show unsatisfying results for cases with significant conformational changes. Therefore the optimal way to combine side-chain and backbone optimization methods is still to be found, and further work in this direction is required.

**DISCUSSION**

Protein flexibility presents a great challenge in predicting the structure of complexes. This flexibility includes both backbone and side-chain conformational changes, which increase the size of the search space considerably. In this paper we reviewed docking methods that handle various flexibility types which are used in different stages of the docking process. These are summarized in the flowchart in Figure 8. Tables I–III, briefly specify the algorithmic approaches of these methods.

The flexible docking process is divided into three major stages. In the first stage the flexibility of the proteins is analyzed. Hinge points can be detected by Ensemble Analysis, GNM or Rigidity Theory. Flexible loops can be identified by MD or Rigidity Theory. Additionally, general conformational space can be defined by NMA, MD or Essential Dynamics. In the second stage the actual

---

**Table I**

<table>
<thead>
<tr>
<th>Method</th>
<th>Flexibility type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DynDom\textsuperscript{17}</td>
<td>Hinge bending</td>
<td>Given two conformations, clusters rotation vectors of short backbone segments and detects the rigid domains.</td>
</tr>
<tr>
<td>HingeFind\textsuperscript{18}</td>
<td>Hinge bending</td>
<td>Compares given conformational states using sequence alignment and detects hinge locations.</td>
</tr>
<tr>
<td>FlexProt\textsuperscript{20,21}</td>
<td>Hinge bending</td>
<td>Compares given conformational states, preforms structural alignment and detects hinge locations.</td>
</tr>
<tr>
<td>HingeProt\textsuperscript{48}</td>
<td>Hinge bending</td>
<td>Detects hinge locations using GNM.</td>
</tr>
<tr>
<td>CONCOORD\textsuperscript{61}</td>
<td>General flexibility</td>
<td>Generates conformations that fulfill distance constraints.</td>
</tr>
<tr>
<td>Dynamics\textsuperscript{63}</td>
<td>General flexibility</td>
<td>Generates conformations using the essential dynamics approach.</td>
</tr>
<tr>
<td>FIRST\textsuperscript{65}</td>
<td>General flexibility</td>
<td>Identifies rigid and flexible substructures using Rigidity Theory.</td>
</tr>
</tbody>
</table>

---

**Table II**

<table>
<thead>
<tr>
<th>Method</th>
<th>Flexibility type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC\textsuperscript{28}</td>
<td>Flexible loops</td>
<td>Chooses the best loop conformations from an ensemble using the Mean-Field approach.</td>
</tr>
<tr>
<td>ATTRACT\textsuperscript{51,83}</td>
<td>Flexible loops</td>
<td>Chooses the best loop conformations from an ensemble using the Mean-Field approach.</td>
</tr>
<tr>
<td>FlexDock\textsuperscript{86}</td>
<td>General flexibility</td>
<td>Energy minimization on degrees of freedom derived from the lowest frequency normal modes.</td>
</tr>
<tr>
<td>FLIPDock\textsuperscript{92}</td>
<td>Hinge bending</td>
<td>Allows hinge bending in the docking. The rigid subdomains are docked separately and consistent results are assembled.</td>
</tr>
<tr>
<td>HADDOCK\textsuperscript{32,33}</td>
<td>General flexibility</td>
<td>Searches favored conformations by a genetic algorithm and a divide and conquer approach. Uses FT data structure.</td>
</tr>
<tr>
<td>RosettaDock\textsuperscript{10,93,118}</td>
<td>General flexibility</td>
<td>Handles backbone flexibility in the refinement stage, by simulated annealing MD.</td>
</tr>
</tbody>
</table>

---
docking is performed. If hinges were identified, the subdomains can be docked separately. Furthermore, an ensemble of conformations can be generated, according to the results of the flexible analysis, and docked using cross-docking or the Mean Field approach. The docking candidates generated in this stage are refined in the third stage. This stage refines the backbone, side chains and rigid-body orientation. These three can be refined separately in an iterative manner or simultaneously. Backbone refinement can be performed by normal modes minimization. Side-chain optimization can be achieved by methods like iterative elimination, graph theory algorithms, MILP, and the Mean Field approach. The refinement of the orientation can be done by a variety of minimization methods such as Steepest Descent, Conjugate Gradient, Newton-Raphson, quasi-Newton and Simplex. Simultaneous refinement can be performed by methods like MD, MC, and genetic algorithms. The final refined docking candidates are scored and ranked.

In spite of the variety of methods developed for handling protein flexibility during docking, the challenge is yet far from resolved. This can be observed from the CAPRI results, where in cases with significant conformational changes the predictions were dissatisfying. Modeling backbone flexibility is currently the main challenge in the docking field and is addressed by only a few methods. In contrast, side-chain flexibility is easier to model and encouraging results have been achieved. The rigid-body optimization stage plays an important role in flexible docking refinement, and contributes considerably to docking prediction success. However, we believe that in order to achieve the best flexible refinement results, the refinement of the backbone, side-chains and rigid-body orientation need to be parallelized. Parallel refinement will best model the induced fit process that proteins undergo during their interaction.

Another major obstacle in the flexible docking field is the poor ranking ability of the current scoring functions. Adding degrees of freedom of protein flexibility to the search space increases the number of false-positive solutions. Therefore, a reliable energy function is critical for the correct model discrimination. The near-native solutions can be identified not only by their energies, but also by the existence of energy binding funnels. Since the ranking ability of the current methods is dissatisfying, further work in this field is required.

Finally, we would like to emphasize that although modeling internal flexibility is essential for general docking predictions, rigid docking is also extremely important. In many known cases the structural changes that occur upon binding are minimal, and rigid-docking is sufficient. The benefits of the rigid procedure are its simplicity and relatively low computational time. In addition, a reliable rigid docking algorithm is essential for generating good docking candidates for further flexible refinement.

ACKNOWLEDGMENTS

The authors thank Dina Schneidman-Duhovny and Michael Farkash for the helpful comments. E.M. was supported in part by a fellowship from the Edmond J. Safra Bioinformatics Program at Tel-Aviv University. The research of HJW has been supported in part by the Israel Science Foundation (grant no. 281/05). The research of HJW and RN has also been supported in part by the NIAID, NIH (grant No. 1UC1AI067231) and by the Binational US-Israel Science Foundation (BSF).
ject has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

REFERENCES


43. Petrone P, Pande VS. Can conformational change be described by only a few normal modes? Biophy J 2006;90:1583–1593.

Flexible Protein-Protein Docking


