Accounting for loop flexibility during protein–protein docking

Karine Bastard, Chantal Prévost, Martin Zacharias

▶ **To cite this version:**

Karine Bastard, Chantal Prévost, Martin Zacharias. Accounting for loop flexibility during protein–protein docking. Proteins - Structure, Function and Bioinformatics, 2006, 62 (4), pp.956-969. 10.1002/prot.20770 . hal-04245445

HAL Id: hal-04245445
https://hal.science/hal-04245445v1

Submitted on 17 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

PROTEINS: Structure, Function, and Bioinformatics 00:000–000 (2005)

Accounting for Loop Flexibility During Protein–Protein Docking

Karine Bastard,[1]* Chantal Prévost,[1] and Martin Zacharias[2]
[1]*Laboratoire de Biochimie Théorique, Institut de Biologie Physico-chimique, Paris, France*
[2]*Computational Biology, School of Engineering and Science, International University Bremen, Bremen, Germany*

ABSTRACT **Although reliable docking can now be achieved for systems that do not undergo important induced conformational change upon association, the presence of flexible surface loops, which must adapt to the steric and electrostatic properties of a partner, generally presents a major obstacle. We report here the first docking method that allows large loop movements during a systematic exploration of the possible arrangements of the two partners in terms of position and rotation. Our strategy consists in taking into account an ensemble of possible loop conformations by a multi-copy representation within a reduced protein model. The docking process starts from regularly distributed positions and orientations of the ligand around the whole receptor. Each starting configuration is submitted to energy minimization during which the best-fitting loop conformation is selected based on the mean-field theory. Trials were carried out on proteins with significant differences in the main-chain conformation of the binding loop between isolated form and complexed form, which were docked to their partner considered in their bound form. The method is able to predict complexes very close to the crystal complex both in terms of relative position of the two partners and of the geometry of the flexible loop. We also show that introducing loop flexibility on the isolated protein form during systematic docking largely improves the predictions of relative position of the partners in comparison with rigid-body docking. Proteins 2005;00:000–000.**
© 2005 Wiley-Liss, Inc.*

Key words: flexible protein–protein docking; docking minimization; loop flexibility; reduced protein models; mean-field theory

INTRODUCTION

Protein-interaction maps become available for whole proteomes owing to the development of reliable methods like yeast two-hybrid analysis, mass spectroscopy, and phage display. They display protein interactions, stable or transient, strong or weak, and suggest that most proteins have interacting partners in the cell. Only a small fraction of these potential complexes are amenable for direct structural characterization by X-ray or NMR studies. However, their structure may be obtained using computational tools, such as docking procedures, if the structures of the individual components are available from precise X-ray or NMR determination, but also from low-resolution EM reconstruction or from homology modeling.[1]

Launched in 2001, the CAPRI experiment (Critical Assessment of PRediction of Interactions) provides evaluation of the docking methods on a common ground and incentives for new methodology development.[2] The motivation is to build three-dimensional structures of molecular complexes (the biologically active species) starting from their separate macromolecular components. The report of the CAPRI blind predictions outcome[3] and the publication of results from research groups that docked a large benchmark of protein pairs[4–8] provide an overview of the current state of the docking methods. Rigid-body procedures that simplify the docking problem by fitting two complementary surface characteristics provide satisfying results for systems that do not undergo important conformational change on association.[9–11] Frequently, conformational change is limited to side chains and many groups have already tackled this difficulty. Side-chain flexibility is accounted for implicitly using a "soft" interface which allows penetration of the partners[7,8] or explicitly during a refinement stage following the rigid-body search.[5,6,12] But current methods are not well adapted when large backbone motion is observed at the interface, as noticed by Chen et al.[13] (see also Vajda et al.[14]) who classified 59 protein complexes based on docking difficulty. Among them, seven cases, identified as "difficult cases," present substantial conformational changes between their unbound and bound structures that principally concern hinge/shear domain bending or large-scale loop motion. Rigid-body search generally yields a huge number of solutions which are classified using scoring functions favoring surface complementary and/or good electrostatic and desolvation properties. For the "difficult cases," back-

Abbreviations: MFT, mean-field theory; RMSD, root-mean-square deviation; PCA, principal component analysis; MD, molecular dynamics.

The work was performed at the Computational Biology, School of Engineering and Science, International University Bremen, Bremen, Germany.

*Correspondence to: Karine Bastard, Laboratoire de Biochimie Théorique, UPR 9080, Institut de Biologie Physico-chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France.
E-mail: Karine.Bastard@ibpc.fr.

Received 21 March 2005; Revised 20 May 2005; Accepted 2 August 2005

Published online 00 Month 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20770

© 2005 WILEY-LISS, INC. *This article is a US government work and, as such, is in the public domain in the United States of America.

bone movements, even limited to surface loops, can drastically modify the steric and electrostatic properties of the protein face presented to the partner. Consequently, at this step, correct solutions (i.e., with the ligand near its position in the native structure) might be very badly ranked, or even not predicted at all. The absence of a loop in a crystal structure also represents a difficulty. This was the case in the first CAPRI round for Target 1, where a surface loop was missing in the unbound structure of one partner. Predictors were expected to model possible structures for that loop to avoid the risk of missing solutions involving interactions with it. It was reported that docking without the loop induces a bias in the prediction results.[15,16] Thus, in cases where loop movements have been detected or a loop is missing, the use of rigid-body docking in the first step of the docking process seems to represent a serious limitation. New approaches are required to treat this difficulty and presently, no method is able to introduce loop flexibility during a complete protein–protein docking procedure.

Such a tool would be very useful since surface loops, consisting of roughly six to twenty amino acids, are often observed at protein–protein interfaces.1[7,18] It is generally assumed that they are more flexible than other parts of the protein.[19] Loop movements occurring during binding can reach up to 10 Å or even more.[20–22] Several types of movements could be characterized using molecular dynamics. In some cases, classical hinge motions can be well accounted for by large changes in just two torsion angles.[23] In contrast, other hinge deformations can be distributed over more torsion angles.[22,24] Sometimes, loops can undergo small rearrangements over all their length, and do not present any precise hinge region.[25] Loop motion permits specific interactions to form, to avoid steric clashes, or to enhance shape complementarity in order to allow hydrogen bonding with the interacting partner.[19,26] This is the case for zymogen where shifts of an inhibitor loop alter the pattern of hydrogen bonding and allow binding to chymotrypsinogen.[27] Loop motion can prove essential in the formation of protein–protein complexes, as it has been experimentally established in the case of the oligomerization of aerolysin protein[28] or for the binding of *Streptomyces subtilisin* inhibitors to proteases.[29] Two models have been proposed to interpret such conformational changes upon association.[30] In the induced fit model,[31] the structure of the partner and that of the receptor adapt to each other during association. This may occur because each partner modifies the chemical and steric properties of the other protein. In the preexisting equilibrium model,[32] unbound proteins exist as a population of diverse conformers, each separated by low energy barriers. The ligand binding shifts the equilibrium toward the structure observed in the complex. The two models can be illustrated by the crystal structures of a monoclonal IgE antibody, Spe7, which exits in two very different conformations resulting from large backbone alteration of two surface loops. Each conformer has been crystallized with structurally distinct antigens.[33]

As already discussed, the docking of such systems requires methods that can incorporate the loop internal flexibility from the beginning of the docking process. Soft representations, used for implicit treatment of side-chain flexibility, permit an efficient docking even in case of incorrect conformations of some interfacial side chains.[7,8] But a soft representation is clearly not adapted to loop movements since the volume scanned by such moves is very large. Thus, it seems necessary to explicitly explore the loop conformations simultaneously to the ligand position. Relaxation of the protein backbone according to precalculated collective degrees of freedom (from Principal Component Analysis—enhanced Molecular Dynamics) was successfully used for the docking of small ligands.[34,35] Nevertheless, such approach cannot be applied to consider large-scale or anharmonic loop motions.[36] Another possibility is to pre-generate the backbone deformations and to perform separate rigid-body docking of structure ensembles, sampled from NMR conformation,[37] molecular dynamics simulations,[38] or principle component restrained molecular dynamics.[39] In the last two cases, the cross rigid-body docking of structure ensembles composed of representative snapshots appeared to improve the predictions. Following the same principle, several algorithms treat loop flexibility for the docking of small ligands.[40–42] These methods generate a discrete set of receptor conformations with different loop conformations, which are then used to perform multi-rigid receptor docking, consisting in successively docking the ligand to each protein structure. These approaches are not easily adaptable to larger system, such as protein–protein complexes, which requires much more computational time. Furthermore, flexible loops can be found both on receptor and ligand proteins, which increase the combinatorial possibilities of valid complex structures. A reasonable approach consists of explicitly taking into account an ensemble of pre-generated loop conformations using a multiple copy representation and simultaneously docking the ligand to the ensemble, rather than successively docking it to single conformers of the loop. This approach was recently implemented in an algorithm that handles loop motion during association.[43] In that work, each loop copy results from ab initio construction and represents one possible conformation of the loop, with rigid backbone. The ligand position is sampled by a Monte-Carlo Simulated Annealing process. In a test case study, the method was able to construct a structure of a protein/DNA complex close to the native structure and to unambiguously predict the conformation of an interfacial loop. Based on an all-atom representation, the method is presently restricted to an exploration around the flexible loop and cannot be extended to systematic searches due to its computational cost.

In the present work, we propose to use a lower protein resolution coupled with a multiple copy representation to treat flexible loops. A reduced protein representation, with up to three pseudo atoms per amino acid, allows an extensive exploration of the possible orientations of the partners and has previously been implemented in the program ATTRACT.[44] ATTRACT considers protein sur-

face remodeling during association by introducing side-chain flexibility at an early stage of the docking process. This was shown to significantly improve the prediction of the correct interacting faces. We thus decided to apply the same approach to the loop problem. Our multiple-copy representation of the loop is treated by the mean-field theory (MFT) as described by Koehl and Delarue.[45] MFT is particularly adapted to systems where a limited subset presents many combinatorial possibilities.[46] For instance, MFT was used to determine preferential conformations of protein side chains,[47] to construct loops in protein homology modeling[48,49] and to optimize the base sequences that favor DNA deformations.[50,51] In the present approach, the flexible loop is represented by an ensemble of copies, each copy being characterized by a distinct rigid conformation for the backbone and side chains. For each ligand position, a weight is attributed to each copy following its interaction energy with the partner. This weight roughly corresponds to a probability that the loop adopts the conformation of this particular copy among the other copies. The copies contribute to the system energy in function of their own weight. Modeling the flexible loop by an ensemble of conformations is relevant to the pre-existing equilibrium model. In our approach, the oncoming of a protein partner modifies the conformation weights and thus the distribution of the conformer population. The aim of our study is to prove that our algorithm is able to deal with several possible loop conformations during a systematic docking simulation.

The method has been applied to a set of eight protein–protein complexes in which one of the two proteins presents large main-chain conformational changes between its unbound and bound forms or in which regions of the backbone are missing in the crystal structure of the unbound form. Among our benchmark set, four cases belong to the seven difficult cases of Chen et al.[13] (the three remaining cases do not present a flexible loop). The loop copy ensemble includes the loop conformation adopted in the unbound and bound forms of the protein, but also ab initio built conformations. We show that our flexible loop algorithm is able to correctly position the ligand in spite of large loop movements at the protein interface and that treating loop flexibility during docking significantly improves the results compared to rigid-body procedures. We also present the results of a "real-like" docking simulation using a protein structure determined by NMR studies. In this case, the copy ensemble is composed of several NMR loop conformations.

MATERIALS AND METHODS

Coordinates of unbound and bound conformations of both receptor and ligand proteins come from the benchmark described by Chen et al.[13] for all test cases, except *1OAZ*, and *1BRC*.* Coordinates of these test cases were obtained from the Protein Data Bank.[52]

Fn1

AQ: 4 *http://zlab.bu.edu/~rong/dock/benchmark.shtml

Protein Representation and Scoring Function

Prior to the docking process, the protein partners are translated into a reduced representation. Each residue is represented by one pseudo atom located at the Cα position and up to two pseudo atoms for each side chain (except for Gly). The side chains of Ala, Ser, Thr, Val, Leu, Ile, Asn, Asp, and Pro are represented by one pseudo atom located at the geometric center of all side-chain heavy atoms. For the side chains of Arg, Lys, Glu, Gln, His, Met, Phe, Tyr, Trp, a pseudo-atom is located exactly in the middle of Cβ and Cγ and another one is calculated as the center of geometry of all other side-chain heavy atoms. The effective interaction (called system energy and corresponding to the total energy) is described by a soft Lennard Jones-type potential between pseudo atom pair *i, j* at distance r_{ij}:

$$R(r_{ij}) = \frac{B_{ij}}{(r_{ij})^8} - \frac{C_{ij}}{(r_{ij})^6} \quad (1)$$

with B_{ij} and C_{ij} indicating repulsive and attractive LJ parameters respectively. Only a few types of repulsive and attractive LJ parameters are used, approximately representing the size and the chemical character of the amino acid.[44]

ATTRACT Protocol

The systematic docking process consists of a series of minimizations, the ligand center being placed at regular positions around the receptor surface at a distance slightly larger than its biggest radius. For each starting position, ~260 initial ligand orientations are generated. For each starting geometry, six energy minimizations (quasi-Newton minimizer) are performed using translational and orientational degrees of freedom of the ligand. The first three minimizations include a harmonic distance restraint between the two partners in order to generate an initial tight complex, followed by free minimization towards the closest minimum configuration.[44]

Mean-Field Approach

Each loop copy is characterized by a rigid backbone and rigid side chains in a specific conformation. For each starting configuration submitted to energy minimization, a weight (a probability) $C_{i,k}$ is attributed to each copy k of the loop i in function of its interaction energy $E_{i,k}$ with the ligand. Intramolecular energy of the loop is not taken into account in the calculation of $E_{i,k}$. Given the energies $E_{i,l}$ of all N_i copies of loop i, the weight of copy k ($C_{i,k}$) is calculated according the Boltzmann principle as:

$$C_{i,k} = \frac{e^{\frac{-E_{i,k}}{RT^i_{eff}}}}{\sum_{l=1}^{Ni} e^{\frac{-E_{i,l}}{RT^i_{eff}}}} \quad (2)$$

R is the Boltzmann constant and T^i_{eff} the effective temperature (see below). RT^i_{eff} is equal to 0.592 kcal/mol at 298 K. The T^i_{eff} parameter can automatically increase in response to steric clashes of the system. The term

$\sum_{l=1}^{Ni} e^{\frac{-E_{i,l}}{RT^{i}_{eff}}}$ permits to normalize the weights, each weight being comprised between 0 and 1 and the sum of the weights for the loop i being equal to 1.

Each copy k of loop i contributes to the system energy E_{sys} proportionally to its own weight:

$$E_{sys} = E_{simple} + \sum_{i=1}^{Nloop} \sum_{k=1}^{Ni} C_{i,k} \times E_{i,k} \quad (3)$$

E_{simple} corresponds to the part of the protein which is not treated in a multi-copy representation. *Nloop* is the number of flexible loops contained in the system. Note that Equation (3) is accurate if the multi-copy representation is used only on one of the two partners. Otherwise crossing terms between copies of interacting loops need to be added.

The effective energy contribution $E_{i,k}{}^{eff}$ of copy k of loop i in the system energy is:

$$E^{eff}_{i,k} = C_{i,k} \times E_{i,k} \text{ where } C_{i,k} \text{ is defined in equation } (2)$$

Thus, derivatives of the energy function with respect to the Cartesian coordinates take into account the copy weight. So the force contribution of atom A belonging to copy k of loop i for one x-component of the coordinates is:

$$\frac{\delta E^{eff}_{i,k}}{\delta x_A} = \frac{\delta}{\delta x_A}(E_{i,k} \times C_{i,k})$$

$$\times \frac{\delta E^{eff}_{i,k}}{\delta x_A} = \frac{\delta}{\delta x_A}\left(\frac{e^{\frac{-E_{i,k}}{RT^{i}_{eff}}}}{\sum}\right) \times E_{i,k} + C_{i,k} \times \frac{\delta E_{i,k}}{\delta x_A} \frac{\delta E^{eff}_{i,k}}{\delta x_A}$$

$$= C_{i,k} \times \left(\frac{E_{i,k} \times e^{\frac{-E_{i,k}}{RT^{i}_{eff}}} - E_{i,k}\sum}{RT^{i}_{eff}\sum} + 1\right) \times \frac{\delta E_{i,k}}{\delta x_A}$$

Σ corresponds to the sum of the weight of all copies l representing the loop i, $\sum = \sum_{l=1}^{Ni} e^{\frac{-E_{i,l}}{RT^{i}_{eff}}}$.

According to Equation (3), the higher the weight of a copy, the higher its importance for the system energy and the more the copy drives the minimization. After each minimization step, the copy weights are readjusted using Equation (2), which permits a continuous estimation of copy weights at each adjustment of the ligand position. Introduction of MFT in the minimizer did not impede a good convergence down to very small residual energy changes per step ($<10^{-5}$ energy units per step).

The temperature value, T^{i}_{eff} , is a key aspect of MFT since it can bias the copy weights (probabilities). Increasing T^{i}_{eff} allows to dump the interaction energies of the copy and to overcome situations when the system gets trapped due to steric clashes. This happens during minimization steps involving external distance restraint between the two partners. After a minimization step, when all copies present steric clashes with the ligand, the Boltzmann probability of every copy is equal to zero (due to the limited numerical floating point precision) at 298 K and copy weights cannot be calculated [see Equation (2)] which cause energy minimization convergence problems. In order to remedy, the effective temperature T^{i}_{eff} is gradually increased until the Boltzmann probability of one copy k of loop i differs from zero. When the protein presents several flexible loops, each loop i possesses a T^{i}_{eff} value independently from the others. Note that after each minimization step, T^{i}_{eff} is reinitialized to 298 K.

Quality Measure of the Docking Predictions

A docking simulation yields several thousand solutions, each corresponding to a ligand position and a loop conformation, which are ranked by their interaction energies. To assess the quality of the predicted complexes, two metrics are used. First, the RMSD on Cα atoms is calculated for the ligand with respect to its position in the crystal complex (called *Lrmsd*). Secondly, the fraction of native contacts (called f_{nat}) estimates the correctly reproduced residue–residue contacts. The fraction of native contacts f_{nat} is defined as the number of native residue–residue contacts in the predicted complex divided by the number of contacts in the target complex. A pair of residues on different sides of the interface is considered to be in contact if any of their atoms are within 7 Å. Note that in the assessment of CAPRI experiment, the distance required to define a native contact is 5 Å.[3] The lower resolution protein representation we use necessitated increasing that distance in order to have a good agreement with the atomic resolution. The value of 7 Å results from an adjustment aimed at reproducing f_{nat} values obtained at atomic resolution.

In order to evaluate the docking efficiency, we consider the rank of the predicted geometry closest to the native structure and the number of times geometries close to the best one are found (same minimum). Two docked complexes are counted as belonging to the same minimum if their *Lrmsd*s differ by less than 0.5 Å and if their energies differ by less than 3 RT (R, gas constant and T, room temperature). These two thresholds come from parameterization of our hierarchical algorithm over the eight test cases.

Ab Initio Generation of Loop Copies

Loop copies have been generated with the LOOP-COPY software which is an extension of the internal coordinate molecular mechanics program LIGAND.[43] LOOP-COPY builds an accurate ensemble of conformations accessible to a peptide fragment within a protein by a combinatorial approach followed by energy minimization and clustering. The free form of the receptor is used as input of the program and the flexible loop is separated from the rest of the protein. Three possible (ϕ, ψ) pair angles, which correspond to the most populated areas of the Ramachandran plots for the 20 amino acids are attributed to each amino acid of the loop, except for glycine, which has four possibilities. Viable conformations are selected on the basis of feasible loop closure and of the absence of overlap between the loop backbone and the rest of the protein. Closure of selected conformations is obtained by energy minimization with respect to the backbone and side-chain angles of the loop. Optimization of the loop/protein system is performed using a Monte-Carlo process. Monte-Carlo

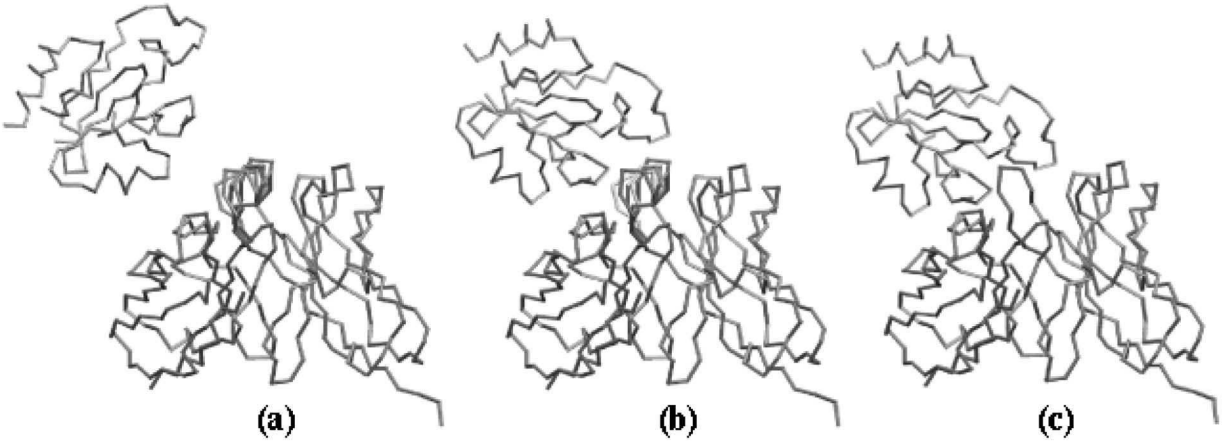


Fig. 1. Schema for the treatment of loop flexibility with the mean-field method. Green and blue lines indicate Cα traces of the receptor and ligand proteins respectively. The loop copies are represented in different colors. **a**: At each starting configuration, a weight is attributed to each copy following its interaction energy with the ligand [see equation (2) in Materials and Methods]. Weights (indicated as colors) can be equivalent if the loop copies present similar interactions with the ligand (in orange). **b**: During the ligand approach, best interacting copies (in red and in orange) present a high weight and dominate the minimization. To the contrary, yellow copies present poor interaction or steric clashes with the ligand and have thus a low weight. **c**: At the end of the minimization, the red copy presents a weight close to 1. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

random moves are applied on χ angles of both loop side chains and protein side chains situated next to the loop. Loop conformations that do not present steric clashes with the protein are clustered on the basis of pair-wise RMSD using a hierarchical clustering algorithm.[53] Conformations within a 1.5 Å clustering threshold are designated as a group, and the lowest-energy conformation within the set is chosen to represent the group. Among all the cluster representative conformations, the eight or nine lowest-energy structures are used as ab initio copies in the present study.

RESULTS

Flexible Loop Docking Algorithm

Our proposed treatment of loop flexibility was introduced in the ATTRACT software. ATTRACT performs systematic docking without using any experimental data concerning the native complex. The search process starts from regularly distributed positions and orientations of the ligand protein around the whole protein receptor. Each starting configuration is submitted to energy minimization. The effective interaction (called system energy) between the two proteins, represented by pseudo-atoms, is described by a soft Lennard-Jones potential based on amino acid size and physico-chemical character (the docking protocol is described in Materials and Methods). The originality of ATTRACT in its early version is to introduce side-chain flexibility during the exhaustive docking search by a switching process between several rotamer conformations. The minimization is performed with respect to a unique rotamer (the one with the most favorable energy) which can change during the minimization procedure. In the present work, flexible loops are represented by at least two conformations (called copies), with rigid backbone and rigid side chains, and loop flexibility is treated by the mean-field approach (MFT) instead of a switching process. In that way, minimization is performed with respect to the whole ensemble of copies. All loop copies are always present, but their weights—their probability of occupation—vary during the minimization. This is a major advantage because even the copies presenting a poor weight at a given moment participate in the system energy and can influence the direction of the minimization (principle illustrated in Fig. 1). For instance, a copy with unfavorable interactions with the ligand (low weight) at the beginning of the minimization can have a very high weight at the end of the minimization (Fig. 2). Generally, at the end of the minimization, the best copy presents a weight close to 1. Nevertheless, we observed cases where two loops had similar interactions and therefore equivalent weights at the end of the minimization.

Introduction of loop flexibility in ATTRACT does not imply a heavy additional time cost with respect to rigid-body procedures. For instance, a complete process of rigid-body docking for the system CheA/CheY (PDB code *1A0O*, 283 amino acids), corresponding to a series of energy minimizations of ~35000 starting configurations, required 12 h of CPU time on a 2.2 GHz Athlon PC. Equivalent procedures for flexible docking using two or 10 loop copies, respectively, took 13 h and 31 h. The number of copies representing the loop increases the number of step necessary to achieve the convergence of minimization.

In the present work, side-chain flexibility outside the flexible loop is not treated. However, MFT should permit to simultaneously treat loop and side-chain flexibility during docking and this approach will be the subject of future work.

In order to appreciate the influence of loop conformational changes on the docking efficiency, we performed rigid-body docking runs subsequently with the bound and

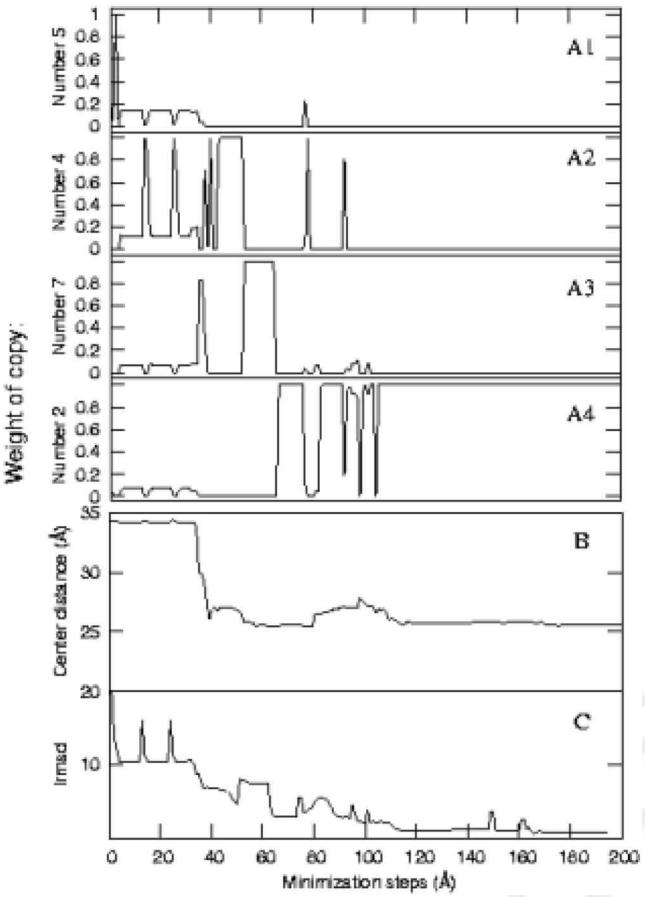


Fig. 2. Variations of the copy weight (**A** 1–4), the distance between geometric centers of the receptor and the ligand (**B**) and the *Lrmsd* value (RMSD of the ligand with respect to its position in the crystal complex) (**C**) versus minimization step during a typical docking minimization (for one starting configuration). The copy ensemble is composed of ten loop copies, but for clarity only four copies are represented in the graphs A1, A2, A3, and A4 (the six others copies present low weight during the major part of the minimization). During the 73 first steps of minimization, a force constant is applied between the two partners in order to generate a tight complex. At the beginning of the simulation, copy weights are almost equivalent because the ligand is far away from the receptor. However, the ligand approach favors the copy number 7. When the ligand adjusts to the receptor, the copy number 2 is finally selected (weight equal to 1).

unbound forms of the protein containing the flexible loop, the partner being in its bound form. The results of these simulations are then compared with flexible loop docking. For direct and accurate comparisons of different docking simulations relative to a same test case, identical starting configurations around the whole receptor were used.

Benchmark Set

Docking simulations were performed on a set of eight protein–protein complexes (Table I), composed of seven targets from the Protein–Protein Docking Benchmark developed by Chen et al.[13] and the Ige Fv Spe7 protein complexed with a recombinant thioredoxin.[33] These test cases were selected because they present significant backbone conformational change at the binding interface for one of the two partners. The number of interface Cα atoms with RMSD larger than 2 Å between unbound and bound structures after superposition is superior or equal to five. The receptor is chosen as the protein that contains the flexible loop (in most cases, except for *3HHR*, the protein with the flexible loop is the biggest between the two partners). But it is important to note that ATTRACT can also deal with flexible loops on the ligand protein and provides similar results compared to using flexible loops on the receptor. The receptors of our benchmark set present several types of flexible loops and different ways to imply the flexible loop in the interaction (Fig. 3). A β-hairpin motion is detected at the interface in *1OAZ* and *1TGS* test cases. But, in contrast to the *1OAZ* case where the eight-residue-loop is the only contact with the ligand, the β-hairpin of *1TGS* (twelve residues) only forms a small part of the interface. Contrary to the previous cases, backbone rearrangements observed at the interface of *3HHR* and *1A0O*, respectively over twenty-six and thirteen residues, do not present precise hinge regions. The deformations extend over the extremity of a helix and an adjacent loop, and we consider this whole segment as the *flexible loop*. *1CGI* and *1BTH* present several flexible loops at the interface. In the case of *1CGI*, a thirteen-residue-loop is missing in the free form and another four-residue-loop undergoes a large deformation. Two other interfacial loops present small deformations and therefore are not treated in this study. For *1BTH*, a β-hairpin (eight residues) moves as a rigid-body while a second β-hairpin (ten residues) undergoes large internal rearrangements. Another loop is missing in the free structure and is not treated here because the corresponding loop in the crystal structure has very few contacts with the ligand. Cases of *1GOT* and *1FIN* are more complicated because changes between the unbound and bound receptor are not located only on a loop. For *1FIN*, a domain motion is observed in addition to the hinge movement of a twenty-residue-loop situated between the two domains. For *1GOT*, the interface comprises an eighteen-residue-loop and a N-terminal helix of twenty-one amino acids, which is missing in the unbound form of the receptor.

Rigid-Body Docking Simulations

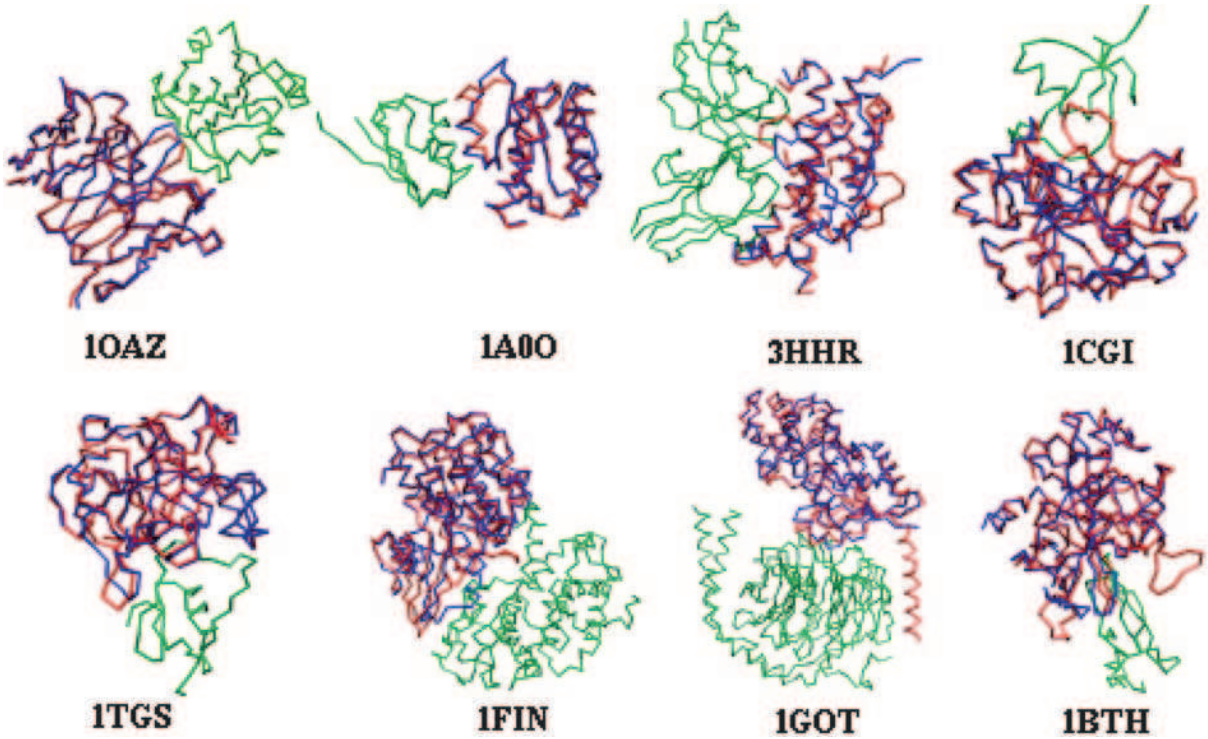
In these simulations, proteins were considered as rigid objects and no flexibility was allowed. The bound form of the ligand was successively docked to the bound and unbound forms of the receptor.

For all complexes, the docking runs using the protein partners in their bound conformations yielded predicted complexes very close to the experiment with *Lrmsd* lower than 1.2 Å and f_{nat} superior to 0.91 (Table II). For most cases, predictions closest to the native structure corresponded to the lowest energy predicted complexes (ranked first) and the same minimum was found more than 100 times. An exception is the Human growth factor/receptor complex (*3HHR*) for which the best-predicted complex (0.815 Å *Lrmsd*) is the seventh top-ranking complex and is only found seven times. The first six predictions present *Lrmsd*s greater than 2 Å.

TABLE I. Target Set†

Complex	Receptor name	Ligand name	RMSD (Å)[a]	Cα[b]
1OAZ (1)	1OAQ: Ige Fv Spe7 (100:107)	Thioredoxin	2.1/0.65	5
1A0O (1)	1CHN: CheA (89:101)	CheY	3.8/0.77	9
3HHR[c] (3)	1HGU: Human growth hormone (42:69)	Receptor	5.5/3.14	24
1CGI[d] (2)	1CGH: α-Chymo-trypsinogen (143:153)(190:193)	Pancreatic secretory trypsin inhibitor	x[d]–5.3/0.88	14
1TGS (2)	2PTN: Trypsinogen (143:154)	Pancreatic secretory trypsin inhibitor	7.0/0.86	17
1FIN[c] (3)	1HCL: CDK2 cyclin-dependent kinase 2 (146:164)	Cyclin	13.2/2.18	59
1GOT[c] (3)	1TAG: Transducin Gt-C (196:213)	GT-β-γ Gi-Cα chimera	6.1/0.99	30
1BTH[c] (2)	2HNT: Thrombin mutant (48:55)(77:86)	Pancreatic trypsin inhibitor	3.8–5.7/ 0.81	18

†The set includes antibody/antigen (1), enzyme/inhibitor (2), and other complexes (3). The receptor is the protein that contains the flexible loop(s). The extremities of the flexible part of the loop are indicated in parenthesis.
[a]RMSD on Cα atoms calculated on the flexible loop/RMSD on Cα atoms calculated on the receptor without the flexible loop, between unbound and bound forms of the receptor after superposition.
[b]Number of interface Cα atoms with RMSD larger than 2 Å between unbound and bound receptor structures after superposition.
[c]Target that belongs to the seven "difficult" cases of the benchmark cited by Chen et al.[13]
[d]The first flexible loop (143:153) is missing in the unbound receptor structure.



COLOR

Fig. 3. Set of targets composed of eight protein–protein complexes. The free structure of the receptors is superposed to the corresponding crystal structure in the complex. The unbound and bound forms of the receptor are respectively represented in blue and red. The ligand of the crystal complex is represented in green.

Docking simulations using the unbound forms yielded poor predictions for seven test cases out of eight. The geometries closest to the X-ray structures scored significantly worse than for simulations with the bound forms and their *Lrmsd* were largely increased (the average *Lrmsd* is 4.2 Å). The number of times the best geometry is found significantly decreased. For the *1CGI* test case, the *Lrmsd* of the best geometry remained correct (1.21 Å) but its ranking largely decreased (from 1 to 44). The low value of f_{nat} in spite of a good ligand placement is due to the absence of the flexible loop in the unbound receptor. Contrary to the other test cases, docking using the unbound form of *1TGS* provided results almost as good as those obtained using the bound form (with a reduced number of predictions in the same minimum) with predicted geometries close to experiment (2.12 Å *Lrmsd*) that are top-ranked. This may be explained by the fact that the flexible loop has very few contacts with the ligand in the

TABLE II. Docking to the Bound and Unbound Forms of the Receptor†

Energy (RT)	*Lrmsd*[b] (Å)	f_{nat}[b]	Rank[c]
Bound receptor			
−18.7	0.81	1.00	3 (179)
−20.8	0.74	1.00	1 (146)
−25.1	0.81	0.91	7 (7)
−34.4	1.12	0.97	1 (198)
−27.1	1.04	0.95	1 (110)
−45.2	0.54	0.97	1 (18)
−36.0	0.63	1.00	6 (124)
−32.9	0.83	0.97	1 (180)
Unbound receptor			
−12.4	4.48	0.67	44 (38)
−17.7	3.61	0.69	5 (11)
−12.1	5.23	0.19	261 (1)
−18.3	1.21	0.57	44 (418)
−22.0	2.12	0.80	1 (65)
−17.1	5.59	0.16	80 (1)
−14.7	5.53	0.16	134 (11)
−4.1	5.38	0.35	6993 (1)

[a]The PDB code of the complex is shown here for clarity, but the unbound form of the receptor was used during docking simulations reported in the right part of the table (refer to Table I).
[b]*Lrmsd* and f_{nat} are given for the predicted geometry that comes closest to the experimental structure.
[c]Indicates the rank of this predicted geometry and, in parenthesis, the number of times it was found in a systematic docking run.
[d]The flexible loop is absent in the unbound receptor in the *1CGI* test-case.

crystal structure and might not be a critical element for ligand binding.

It must be noted that energies of systems resulting from the docking of unbound or bound receptors are not comparable, since one of the two forms may lack some residues. We have checked that the number of missing residues is small and that they are not at (or near) the binding interface. To the contrary, it is possible to compare the energies of predictions involving the unbound receptor, whether they are obtained using rigid-body docking (Table II) or using flexible loop docking (see following results).

Flexible Loop Docking Simulations

In these simulations, the ligand is docked to the unbound form of the receptor and the flexible loop is represented by two copies: the loop conformations present in the unbound receptor (called unbound loop) and in the bound receptor (called bound loop). *1CGI* test case is not included in the following benchmark subset because the flexible loop is absent in the free form of the receptor.

For six test cases out of seven, the results are largely improved in comparison with rigid-body docking using the free form of the receptors (Table III). For the test cases *1OAZ*, *1A0O*, *3HHR*, *1GOT,* and *1BTH*, the *Lrmsd* values of the geometries closest to the native structure are much smaller than those obtained using rigid-body docking to the unbound receptors, even if these *Lrmsd* values do not reach those of the best predictions obtained using the rigid bound receptor (in the present simulations the best *Lrmsd* values are comprised between 1.04 and 3.19 Å). More generally, the ranking of the best geometries is largely improved in comparison with the results of rigid-body docking involving the unbound receptor. It is important to note that for predictions belonging to the same minimum as the best predicted complex, the selected copy always corresponds to the bound loop. Even for the case *1BTH* where two flexible loops are treated, the selected copy for each flexible loop corresponds to the bound loop. Predictions where the selected copy is the unbound loop show high *Lrmsd* and bad ranking. Interestingly, even for *1GOT* and *1FIN* for which the conformational changes are not only located at the interfacial loop, predictions are improved. For *1GOT*, the best *Lrmsd* value decreases from 5.53 Å to 3.19 Å and the ranking changes from 134 to 40. Note also that the number of predictions in the corresponding minimum increases (from 11 to 44). For *1FIN*, the ranking of the geometry closest to experiment (6.69 Å *Lrmsd*) is largely improved (from 80 to 5). It seems that for these two test cases, the supplementary conformational changes may impede the ligand from approaching close to its position in the crystal complex. This hypothesis was checked by performing simulations using the bound receptor and where the flexible loop was represented by the unbound and bound loops. Results are excellent since the best *Lrmsd* is 0.56 with top ranking in case of *1FIN* and the bound loop is selected in all predictions close to experiment.

The notable exception to all previous remarks again concerns the *1TGS* test case for which predictions are close to those obtained using rigid-body docking to the unbound and bound receptor. However, the number of predictions within the same minimum is smaller and the selected copy corresponding to this minimum is the unbound loop. In this test case, selection of the unbound loop copy does not prevent a good ligand position in the best predictions. A possible explanation of the algorithm failure is that the flexible loop represents only a very small part of the interface and shows very few interactions with the ligand.

Influence of the Number of Loop Copies

Flexible docking with two loop copies largely improves the prediction of the complex geometry compared to rigid-body docking. A pertinent question is how the number of copies will influence the predictions. We performed a set of simulations (called set A) using the unbound receptor, in which the flexible loop is represented by an ensemble of loop copies comprising both the unbound and bound loops and eight additional copies with different conformations. Another question arising is how the predictions will evolve if the copy ensemble does not contain the bound loop. In order to answer that question, a set of simulations (called set B) was performed using the unbound receptor with an ensemble of loop copies comprising the unbound loop and nine copies. We have chosen to build additional loop conformations ab initio rather than using geometries issued from MD simulation, NMR studies, or loop databanks. The risk of MD simulations is to observe only small rearrangement movements (large-scale loop motions occur

TABLE III. Docking to the Unbound Receptor Protein Including Two Loop Copies: the Unbound and the Bound Loops

Case[a]	Energy (RT)	*Lrmsd*[b] (Å)	f_{nat}[b]	Rank	Loop copy[c]
1OAZ	−19.2	1.04	0.94	2 (36)	B
1A0O	−20.2	2.43	0.76	3 (34)	B
3HHR	−16.6	2.1	0.78	8 (3)	B
1CGI[d]	—	—	—	—	—
1TGS	−22.4	2.04	0.81	2 (25)	U
1FIN	−30.7	6.69	0.42	5 (1)	B
1GOT	−21.5	3.19	0.66	40 (44)	B
1BTH[e]	−22.5	2.34	0.54	2 (18)	B/B

[a]and [b]same as Table II[a] and II[b].
[c]Indicates the status (U: unbound or B: bound) of the loop found within the same minimum.
[d]The flexible loop is absent in the unbound form in the *1CGI* test-case, so this case has been excluded from the present set of simulations.
[e]Two flexible loops of the receptor are treated.

TABLE IV. RMSD (Cα) Between the Ab Initio Generated Loop Conformations and the Corresponding Bound Loop†

	Number of the ab initio copy								
Test-case	1	2	3	4	5	6	7	8	9
1OAZ (from 100 to 107)	0.91	2.83	3.14	3.39	4.06	4.51	4.75	7.01	8.24
1A0O (from 89 to 101)	3.88	4.28	4.30	4.90	5.11	5.19	5.42	5.51	6.12
1CGI (from 143 to 153)	4.04	4.36	5.28	5.79	6.65	7.09	7.86	8.22	10.47
1BTH (from 48 to 55)	3.01	3.02	3.26	4.01	5.13	5.66	5.75	6.33	7.84
1BTH (from 77 to 86)	3.31	4.05	4.82	5.10	5.18	5.54	6.26	7.27	7.54

†We checked that each conformation differs from the others within each test-case by at least 2 Å.

on time scales from 10^{-3} to 10^{-1} s[54–56]). NMR structures are not always available for a given protein and loop databanks may be incomplete especially concerning interacting loops. Loop conformations built ab initio can cover a large conformational space in order to represent geometries accessible to the flexible loop.[43] Among the exhaustive conformational possibilities, best energy loop conformations were selected to compose the copy ensemble (protocol given in Materials and Methods) and are described in Table IV. Due to the predefined number of copies, these selected copies do not intend to cover the whole space accessible to the loop and actually, only in one test case an additional loop conformation with backbone close to that in the crystal is present.

Simulation sets A and B were performed on test cases *1OAZ*, *1A0O*, *1CGI,* and *1BTH* for which flexible docking with two copies yielded good results (near-native solution among the three top-ranking predictions). Table V (left) shows that the addition of extra loop copies tends to improve the docking efficiency in comparison with flexible docking with two loop copies in terms of best *Lrmsd* value and ranking of the prediction closest to experiment. The bound loop is always selected among the predictions close to experiment, except for the *1CGI* test case where the second flexible loop is represented by the bound loop 16 times over 42. Interestingly, the number of predictions close to the native complex has increased for *1A0O*, *1OAZ,* and *1BTH*. For *1A0O* and *1BTH*, the ranking is even better than obtained when we used two copies (unbound and bound loops). It might be possible that the presence of extra loop copies favor the ligand approach near the native binding site.

Simulations set B (absence of the bound loop in the copy ensemble) provides interesting results. For two test cases, the best *Lrmsd* value was lower than 3 Å and the predictions closest to the native structure were scored among the three-top ranking predictions. On the other hand for *1OAZ* and *1BTH*, the absence of the bound loop in the copy ensemble of set B impeded the approach of the ligand near a position close to that in the crystal (*Lrmsd* 4.98 Å and 4.65 Å, respectively). However for *1BTH*, the prediction closest to experiment is ranked first and belongs to a populated minimum. Note that the unbound loop is mainly selected for representing the first flexible loop. For *1OAZ*, among the copy ensemble, one copy presents a backbone conformation very similar to that of the bound loop (0.91 Å RMSD on Cα atoms) but was not selected by the program. This comes from the orientation of one critical arginine side chain completely differing from that found in the bound loop. In the crystal complex, this side chain interacts tightly with the ligand. Surprising results concern the *1A0O* and *1CGI* test cases for which the copy ensemble does not contain any copy with backbone conformation similar to that of the bound loop (the RMSD of the closest loop are 3.88 Å and 4.04 Å, respectively). Nevertheless, the docking simulations yielded ligand positions close to experiment (2.92 Å *Lrmsd* and 1.14 Å/1.31 Å *Lrmsd* respectively), moreover ranked at position 3 and 1, respectively. For *1A0O* test case, the f_{nat} value is relatively low (46% of the residue–residue native contacts are recovered), which might be explained by the important difference in backbone conformation between the selected copy and the bound loop (the flexible loop on the receptor of *1A0O* contributes to almost the whole complex interface). For

TABLE V. Docking to the Unbound Protein Receptor Where the Flexible Loop is Represented by an Ensemble of 10 Copies Containing the Unbound Loop†

Case[a]	Energy (RT)	*Lrmsd*[b] (Å)	f_{nat}[b]	Rank	Loop[b] copy
			Set A: with the bound loop		
1OAZ	−19.0	0.76	0.95	2 (47)	B
1A0O	−20.6	2.57	0.78	1 (58)	B
1CGI[d]	−23.9	1.66	0.66	1 (67)	B
1CGI[e]	−26.0	1.92	0.66	1 (42)	B/B[f]
1BTH	−21.9	2.38	0.55	1 (26)	B/B
			Set B: without the bound loop		
1OAZ	−15.3	4.98	0.18	10 (14)	6
1A0O	−16.9	2.92	0.46	3 (45)	5
1CGI[d]	−20.1	1.14	0.63	1 (38)	2
1CGI[e]	−23.8	1.31	0.64	1 (31)	2/B[g]
1BTH	−19.22	4.65	0.32	1 (34)	U/8[h]

†Contrary to simulation set A, set B does not include the bound loop.
[a]and [b]same as Table II[a] and II[b].
[c]Refers to loop number in Table IV or to unbound (U) and bound (B) loops found within the same minimum.
[d]In this case, the unbound loop is unresolved in the crystallographic structure. The unbound form was replaced by another copy.
[e]With two flexible loops: the first is represented by 10 copies and the second by two loop copies (unbound and bound loops).
[f]Only the 16 best energy complexes present the second loop in its bound form.
[g]The second loop was found 21 times in its bound form.
[h]The unbound form of the first loop was predicted 31 times. Copy 8 of the second loop was predicted 24 times.

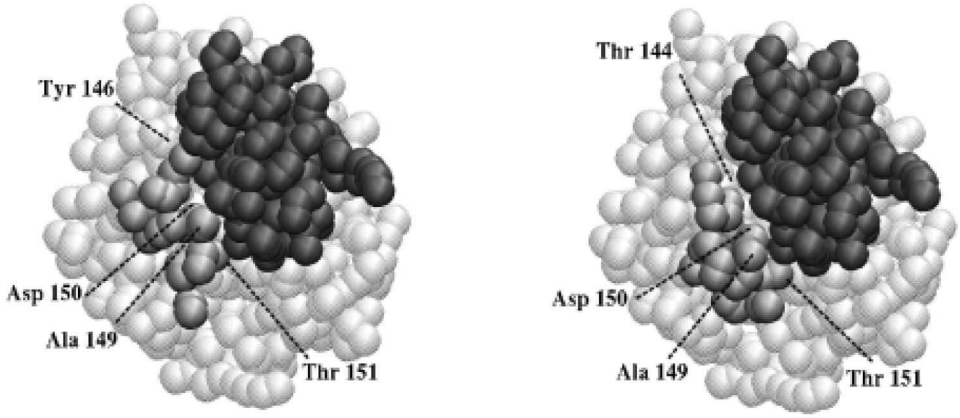


Fig. 4. Comparison between the crystal complex (left) and the predicted complex closest to native (right) for the *1CGI* test case (each sphere represents a pseudo atom in the reduced protein model). The receptor, the flexible loop and the ligand are represented, respectively, in white, in light grey and in dark grey. Alanine 149, asparagine 150, and threonine 151 of the predicted loop interact with the ligand similarly than observed in the crystal complex although threonine 144 replaces tyrosine 146 of the bound loop.

1A0O and *1CGI* test cases, in spite of the incorrect backbone conformation of the selected copy, accessible side chains point toward the same direction than the corresponding side chains in the bound loop or occupy the same space than other interacting side chains of the bound loop (Fig. 4). The *1A0O*, *1CGI*, and *1OAZ* test cases confirm the importance of the orientation of several accessible side chains on the docking efficiency, even when a reduced protein model is used. This was already demonstrated in previous work.[44]

For test case *1CGI*, the simulations always selected the same loop copy (number two, see Table IV) to represent the geometry of the first flexible loop in the best predictions, whether one or two flexible loops were accounted for. Simulations considering two flexible loops did not always select the bound form for the small second loop (Table V). However, the 12-best energy complexes correspond to predictions where the bound conformation is selected for the second flexible loop. In the crystal complex, the second loop has less contact with the ligand than the first loop, which might explain this lack of specificity. Furthermore, the second loop is not crucial for ligand binding as shown by the results of docking predictions where only the first flexible loop is accounted for (the second one being in its unbound form).

In order to establish the advantage of the multi-copy docking in comparison to the rigid-body docking of an ensemble,[37–39] we have performed two sets of separate rigid-body docking simulations for each of the loop copies on the unbound receptor of the *1A0O* test case. In the first set, the ligand was successively docked to 10 structures of the unbound receptor, each being characterized by a

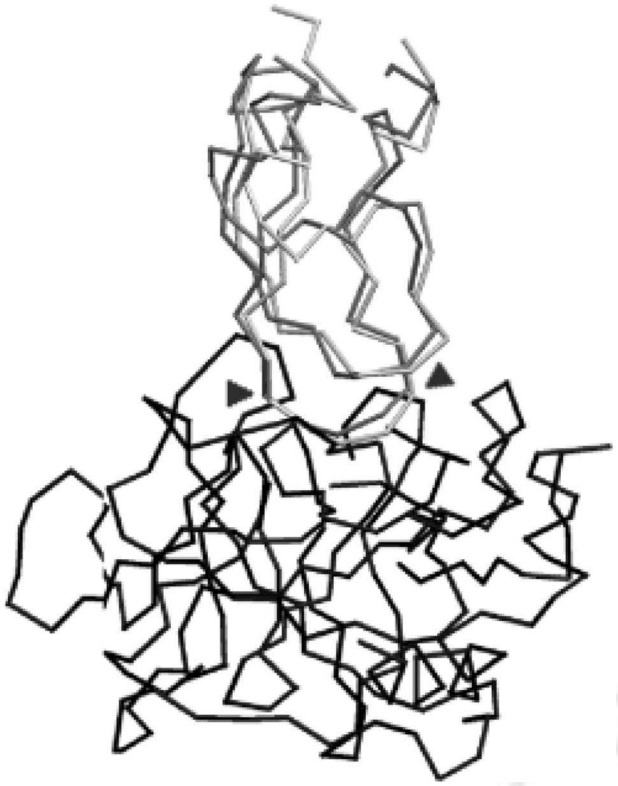


Fig. 5. Cα traces of the ligand position issue from the top-scored prediction (in light grey) of a systematic docking of BPTI to its receptor β-trypsin, compared to its position in the crystal complex (in dark grey). The receptor is represented in black. The *Lrmsd*, calculated on Cα atoms, between these ligand structures is 0.8 Å (see Results). Extremities of the flexible loop are indicated with arrows.

distinct loop conformation (eight ab initio loops, the unbound and bound loops). In the second set, the ensemble comprises 10 conformers of the unbound receptor (nine ab initio loops and the unbound loop). The first and the second sets correspond respectively to the set A and B of Table V. For the two sets, predictions are very similar in terms of ranking and *Lrmsd* in comparison with equivalent multi-copy docking. However, the number of predictions belonging to the same minimum as the best prediction has decreased (from 47 to 21 for the first set and from 45 to 15 for the second). For this example, the use of multi-copy docking clearly led to no loss of efficiency, while the benefit in terms of computer time is obvious since the rigid-body docking of an ensemble of structures took 10*12 h instead of 31 h for the multi-copy docking.

AQ: 2

Docking With Several Loop Copies From NMR Studies

In the previous simulations, the flexible loop was represented by ab initio built conformations, but it is also possible to use conformations from other origins. For instance, NMR (nuclear magnetic resonance) studies can provide several conformational states for protein flexible loops.

An NMR structure of a BPTI (bovine pancreatic trypsin inhibitor) mutant[57] (PDB code: *1JV8*) in the unbound form is available that still adopts the same structure as wild-type BPTI but shows increased conformational flexibility. The loop segment (residues 13–17) that forms the binding interface to trypsin in the trypsin-BPTI complex was represented by four conformational loop copies spanning a range of possible conformations obtained from the NMR analysis (models 1, 3, 7, and 18 of the NMR structural ensemble). We performed a systematic flexible docking of the BPTI ligand, in its unbound form with four representative loop copies, to its receptor β-trypsin, in its bound form, and we obtained predictions in which the ligand position was very close to its position in the crystal complex (PDB code: *2PTC*). The best prediction was scored first and its *Lrmsd* was 0.8 Å (Fig. 5). In this case, the *Lrmsd* was calculated between the predicted ligand and the unbound form of the ligand previously superposed to the ligand in the crystal complex, without accounting for the loop part. In that way, the *Lrmsd* only measures the difference in ligand positioning, separated from its internal conformational changes. Predictions belonging to the same minimum are numerous (188) and, for all these predictions, the algorithm was able to select the loop conformation closest to the one presents in the crystal structure of the complex (0.60 Å RMSD on Cα atoms).

DISCUSSION

Protein–protein association is often coupled to conformational changes in protein loop regions. For a realistic prediction of protein–protein complex geometries, it is essential to account for such possible conformational changes during docking simulations. Common strategies, for example, performing a systematic search using rigid protein partners followed by refinement of a number of preselected complexes, may fail completely because they disregard any favorable complex close to the realistic binding geometry during preselection.

AQ: 3

Our strategy for treating loop flexibility during docking consists of treating the selected loop regions using a mean-field approach during a multi-start docking minimization of the two partners represented by a reduced protein model. By employing a reduced representation, it is possible to position the ligand around the whole receptor surface with different orientations, and thus to systematically evaluate many thousand putative docking complexes. The computational efficiency is also due to the mean-field representation of the loop by several conformational copies because the number of additional interactions to be calculated during docking optimization is small relative to the total of protein–protein interaction. Use of a soft Lennard-Jones type interaction potential is particularly well suited for mean-field approaches that employ a discrete set of rigid conformational copies. Since the probability or relative (Boltzmann) weight of each conformation is an exponential function of its energy, softening the repulsive part of the energy decreases the weight dependency on transient steric clashes at a given moment. This allows each copy with repulsive interactions to conserve a certain degree of contribution to the minimization process. Nonetheless, this does not impede selection of a

favorable copy during the final ligand adjustment, when the attractive part of the energy dominates.

The approach was tested on several cases for which experimental X-ray and NMR studies demonstrated significant conformational changes of loop regions in at least one protein partner upon complex formation. The results show that our algorithm is able to correctly position the ligand with respect to its position in the crystal and to select the best fitting conformation among an ensemble of loop conformations proposed to represent the flexible loop. The method also shows good efficiency when dealing with two flexible loops present on the same protein. Compared with rigid-body docking, our two-loop copy docking procedure systematically predicted geometries closer to that of the native complex when applied to protein unbound forms. Furthermore, the ranking of the best geometries was largely improved and the number of predictions close to the best predicted complex increased. In the same way, our flexible docking method using nine ab initio copies showed improvements compared to unbound rigid body docking, at least for the four systems tested. The presence, among the copy ensemble, of a loop copy identical to the conformation in the protein bound form always resulted in the selection of this bound conformation in the best prediction (except for one case, *1TGS*, where the influence of the loop to association appears to be small). It is noteworthy that addition of *ab initio* copies in the copy ensemble seems to improve the results (in terms of ranking, ligand position, or number of good predictions) in comparison with docking simulations with only two loop copies (the unbound and bound loops), for three test cases. A possible explanation is that the presence of additional loop conformations can help to cross barriers of individual copy energy components and contribute to the ligand adjustment with respect to the bound loop copy. Further investigation is needed to confirm this proposition. Interestingly, in the absence of a bound loop conformation in the copy ensemble, the approach still resulted in improved docking results compared to rigid-body docking with the unbound protein form. However, in this case the loop conformation with backbone closest to that of the bound conformation was not always selected as the most favorable loop solution. Further analysis of these cases indicates that the conformations of the loop side chains also play a decisive role for loop copy selection. A loop copy with an "incorrect" backbone conformation might be selected because its side chain conformations are such that it interacts favorably with the binding cavity on the protein partner. In the present study, the method was tested for flexible loops which contain a limited number of residues. We have shown its efficiency for flexible parts up to 25 residues (*3HHR* test case). We intend to test our copy-docking method for systems which present a large part of the protein in a multi-copy representation (for instance domain motion).

This work demonstrated the feasibility of treating loop motions during a complete docking procedure. Clearly, the method must be completed in order to obtain a fully automatic docking method. First, ab initio blind predictions need to be performed on the unbound forms of both the receptor and the ligand, which requires accounting for protein side chain and loop rearrangements on both partners. A pertinent approach may consist in treating side chain flexibility by MFT instead of the previous switching process of ATTRACT since MFT is particularly adapted for multi-copy representation of elements that can interact together. Another important point is the determination of flexible parts of the protein. In the present study, limitation of the flexible loop was determined by comparing unbound and bound structures of the protein. An ab initio docking simulation will require prior evaluation of flexible regions of the free form. The flexible loops can be located using experimental data (mutations, B-factor, disordered regions in the X-ray or NMR structure), computational approaches (molecular dynamics simulations, normal mode vibrational analysis,[58] Principal Component Analysis[34,59] or covalent and noncovalent bond networks[60]) or may correspond to a protein segment with little homology to a template structure in comparative homology modeling. Finally, a specific work needs to be performed concerning the quality of the copy ensemble representing the loop, since as discussed above, it is an important parameter which influences the precision of the ligand position in our predictions. In the present study, we chose arbitrarily to represent the loop by up to 10 conformations. However, the copy number should be derived from a pertinent clustering of all possible loop geometries. Representative copies issued from this clustering should sample the available space accessible to the loop. Ab initio building allows constructing good quality conformation ensembles for short to medium loops ($\leq$ 12 amino acids)[43] but covering the conformational space available to long loops becomes a complicated task due to the explosion of combinatorial possibilities. To handle this difficulty, methods issued from robotics[61] or from sequence–structure correlations based on Bayesian statistic[62] appear very promising. However, a good sampling of the backbone conformation may not be sufficient since it is known that side-chain conformations play a major role in the binding dynamics. As suggested by the results of *1OAZ*, a rigid representation of loop side chains can represent a limitation in several cases, and accounting for loop side-chain flexibility appears necessary. This should be achieved by treating side chains flexibility by MFT using rotamer representation. However, the two levels of multi-copy representation (side-chain copies on loop copies) may burden the calculation of the energy function derivatives. An alternative solution is to replicate each possible backbone conformation and to characterize each replica with different side-chains rotamers. The reduced protein description allows representation of the conformational loop/side chain space by fewer copies than necessary at atomic resolution, and consequently it is expected that the number of copies will remain within reasonable limits. We will address this problem in future works.

Our results suggest it should be possible to access accurate predictions of partner binding faces within a prospective complex using a low-resolution approach. However, the necessary information for driving genetic and

biochemical experiments requires predicting the details of the interaction at an atomic level. In case of complexes containing interfacial flexible loops, translating the low resolution model into atomic representation not only requires adjusting the ligand position on the receptor surface and optimizing side-chain conformations: Refining the loop position and its internal backbone conformation will also be necessary as even small loop rearrangements affect the short-range protein association.[63] MD simulations should not be adapted to refine loop conformation since, as showed in this study, selected loop copy in the best predicted complex may present backbone geometry far from that of the bound loop. However, a multi-copy approach can be used coupled to atomic protein resolution.[43] The ensemble of trial loop conformations should roughly cover the same volume as occupied by the selected low-resolution loop copy together with a similar group positioning.

Protein–protein docking using protein model structures could become one main application area of the present multi-copy loop docking approach. For many proteins of known sequence it is nowadays possible to build model structures based on sequence similarity to a protein of known structure. The accuracy of such model structure depends significantly on the target–template similarity.[64] In cases of more than 50% sequence identity between target and template, the model structures can reach accuracies of 1 Å with respect to the overall protein main-chain structure. However, even in these favorable cases the structural deviation of a model from a realistic structure may not be uniformly distributed along the sequence. Often in loop regions the deviation of the structural model from a real structure might be significantly larger than 1 Å.[65,66] Errors can prevent the use of homology-built protein models in protein–protein docking simulations. A possible solution to this problem is to represent such loop regions of limited accuracy as multiple conformational copies and employ these in systematic docking simulations using the present multi-copy approach.

CONCLUSION

In this study, we bring our contribution to one important challenge of the docking field, which is introducing internal protein flexibility during systematic docking. We proposed an original strategy to account for loop adjustment during the docking search. By introducing the mean-field method in the multi-start docking protocol ATTRACT, we are able to introduce flexibility without adding heavy additional time cost. For several cases considered difficult, we showed that our flexible docking method largely improves the predictions compared to rigid protein–protein docking.

ACKNOWLEDGMENTS

We thank Drs. R. Lavery, A. Barthel, and D. Roccatano for helpful discussions. This work was specifically supported by the European Science Foundation (ESF) *Integrated Approaches for Functional Genomics* programme (2003/EG/20). This work was also supported in part by a grant from the DFG(ZA/6-1) to M.Z and a grant from the CNRS and the French Ministry of Research (DRAB 02/100) to K.B. and C.P.

REFERENCES

1. Janin J, Sepharin B. Genome-wide studies of protein-protein interaction. Curr Opin Struct Biol, 2003;13:383–388.
2. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, et al. CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 2003;52:2–9.
3. Mendez R, Leplae R, Maria LD, Wodak SJ. Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins2003;52:51–67.
4. Chen R, Weng Z. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins 2002;47:281–294.
5. Gray JJ, Moughon SE, Kortemme T, Schueler-Furman O, Misura KM, Morozov AV, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331:281–299.
6. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 2004;20:45–50.
7. Fernandez-Recio J, Totrov M, Abagyan R. Soft protein-protein docking in internal coordinates. Protein Sci 2001;11:280–291.
8. Palma PN, Kriippahl L, Wampler JE, Moura JJG. BIGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins 2000;39:372–384.
9. Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, et al. Protein docking using continuum electrostatics and geometric fit. Protein Eng. 2001;14:105–113.
10. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins 2003;52:80–87.
11. Gardiner EJ, Willett P, Artymiuk PJ. Protein docking using a genetic algorithm. Proteins 2001;44:44–56.
12. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–110.
13. Chen R, Mintseris J, Janin J, Weng Z. A protein-protein docking benchmark. Proteins 2003;52:88–91.
14. Vajda S, Camacho CJ. Protein-protein docking: is the glass half-full or half-empty? Trends Biotechnol 2004;22:110–116.
15. Camacho CJ, Gatchell DW. Successful discrimination of protein interactions. Proteins 2003;52:92–97.
16. Law DS, Eyck LFT, Katzenelson O, Tsigelny I, Roberts VA, Pique ME, Mitchell JC. Finding needles in haystacks: reranking DOT results by using shape complementarity, cluster analysis, and biological information. Proteins 2003;52:33–40.
17. Conte LL, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
18. Fetrow JS. Omega loops: nonregular secondary structures significant in protein function and stability. FASEB J 1995;9:708–717.
19. Betts MJ, Sternberg MJE. An analysis of conformational changes on protein-protein association: implications for predictive docking. Protein Eng 1999;12:271–283.
20. Joseph D, Petsko GA, Karplus M. Anatomy of a conformational change: hinged "lid" motion of the triosephosphate isomerase loop. Science 1990;249:1425–1428.
21. Sun YJ, Rose J, Wang BC, Hsiao CD. The structure of glutamine-binding protein complexed with glutamine at 1.94 Å resolution: comparisons with other amino acid binding proteins. J Mol Biol 1998;278:219–229.
22. Gerstein M, Chothia C. Analysis of protein loop closure. two types of hinges produce one motion in lactate dehydrogenase. J Mol Biol 1991;220:133–149.
23. Derreumaux P, Schlick T. The loop opening/closing motion of the enzyme triosephosphate isomerase. Biophys J 1999;74:72–81.
24. Fitzgerald MM, Musah RA, McRee DE, Goodin DB. A ligand-gated, hinged loop rearrangement opens a channel to a buried artificial protein cavity. Nat Struct Biol 1996;3:626–631.
25. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massague J, Pavletich NP. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. Nature 1995;376:313–320.
26. Kempner ES. Movable lobes and flexible loops in proteins. Struc-

tural deformations that control biochemical activity. FEBS Lett 1993;326:4–10.
27. Hecht HJ, Szardenings M, Collins J, Schomburg D. Three-dimensional structure of the complexes between bovine chymotrypsinogen A and two recombinant variants of human pancreatic secretory trypsin inhibitor (Kazal-type). J Mol Biol 1991;220:711–722.
28. Rossjohn J, Raja SM, Nelson KL, Feil SC, vander Goot FG, Parker MW, Buckley JT. Movement of a loop in domain 3 of aerolysin is required for channel formation. Biochemistry 1998;37:741–746.
29. Kojima S, Furukubo S, Kumagai I, Miura K. Effects of deletion in the flexible loop of the protease inhibitor SSI (Streptomyces subtilisin inhibitor) on interactions with proteases. Protein Eng 1993;6:297–303.
30. Goh CS, Milburn D, Gerstein M. Conformational changes associated with protein-protein interactions. Curr Opin Struct Biol 2004;14:104–109.
31. Koshland DE. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA 1958;44:98–106.
32. Tsai CJ, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. Protein Sci 1999;8:1181–1190.
33. James LC, Roversi P, Tawfik DS. Antibody multispecificity mediated by conformational diversity. Science 2003;299:1362–1367.
34. Zacharias M. Rapid protein-ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: binding of FK506 to FKBP. Proteins 2004;54:759–767.
35. Tatsumi R, Fukunishi Y, Nakamura H. A hybrid method of molecular dynamics and harmonic dynamics for docking of flexible ligand to flexible receptor. J Comput Chem 2004;25:1995–2005.
36. Balsera MA, Wriggers W, Oono Y, Schulten K. Principal component analysis and long time protein dynamics. J Phys Chem 1996;100:2567–2572.
37. Dominguez C, Bonvin AM, Winkler GS, van Schaik FM, Timmers HT, Boelens R. Structural model of the UbcH5B/CNOT4 complex revealed by combining NMR, mutagenesis, and docking approaches. Structure (Camb) 2004;12:633–644.
38. Smith GR, Sternberg MJ, Bates PA. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol 2005;347:1077–1101.
39. Grunberg R, Leckner J, Nilges M. Complementarity of structure ensembles in protein-protein binding. Structure (Camb) 2004;12:2125–2136.
40. Claussen H, Buning C, Rarey M, Lengauer T. FlexE: efficient molecular docking considering protein structure variations. J Mol Biol 2001;308:377–395.
41. Zavodszky MI, Lei M, Thorpe MF, Day AR, Kuhn LA. Modeling correlated main-chain motions in proteins for flexible molecular recognition. Proteins 2004;57:243–261.
42. Cavasotto CN,Abagyan RA. Protein flexibility in ligand docking and virtual screening to protein kinases. J Mol Biol 2004;337:209–225.
43. Bastard K, Thureau A, Lavery R, Prévost C. Docking macromolecules with flexible segments. J Comput Chem 2003;24:1910–1920.
44. Zacharias M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci 2003;12:1271–1282.
45. Koehl P, Delarue M. Mean-field minimization methods for biological macromolecules. Curr Opin Struct Biol 1996;6:222–226.
46. Elber R, Karplus M. Enhanced sampling in molecular dynamics: use of the time-dependant hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. J Am Chem Soc 1990;112:9161–9175.
47. Koehl P, Delarue M. Application of a self-consitent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J Mol Biol 1994;239:249–275.
48. Koehl P, Delarue M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nat Struct Biol 1995;2:163–170.
49. Zheng Q, Rosenfeld R, DeLisi C, Kyle DJ. Multiple copy sampling in protein loop modeling: computational efficiency and sensitivity to dihedral angle perturbations. Protein Sci 1994;3:493–506.
50. Lafontaine I, Lavery R. Optimisation of nucleic acid sequence. Biophys J 2000;79:680–685.
51. Paillard G, Lavery R. Analyzing protein-DNA recognition mechanisms. Structure (Camb) 2004;12:113–122.
52. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
53. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 1958;38:1409–1438.
54. Wade RC, Davis ME, Luty BA, Madura JD, McCammon JA. Gating of the active site of triose phosphate isomerase: Brownian dynamics simulations of flexible peptide loops in the enzyme. Biophys J 1993;64:9–15.
55. Williams JC, McDermott AE. Dynamics of the flexible loop of triosephosphate isomerase: the loop motion is not ligand gated. Biochemistry 1995;34:8309–8319.
56. Falzone CJ, Wright PE, Benkovic SJ. Dynamics of a flexible loop in dihydrofolate reductase from *Escherichia coli* and its implication for catalysis. Biochemistry 1994;33:439–442.
57. Battiste JL, Li R, Woodward C. A highly destabilizing mutation, g37a, of the bovine pancreatic trypsin inhibitor retains the average native conformation but greatly increases local flexibility. Biochemistry 2002;41:2237–2245.
58. Kovacs JA, Chacon P, Abagyan R. Predictions of protein flexibility: first order measures. Proteins 2004;56:661–668.
59. Zacharias M, Sklenar H. Harmonic modes as variables to approximately account for receptor flexibility ligand-receptor docking simulations: application to a DNA minor groove ligand complex. J Comput Chem 1999;20:287–300.
60. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. Proteins 2001;44:150–165.
61. Cortes J, Simeon T, Remaud-Simeon M, Tran V. Geometric algorithms for the conformational analysis of long protein loops. J Comput Chem 2004;25:956–967.
62. Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief WR, Wedemeyer WJ, et al. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins 2003; Suppl 6:457–468.
63. Ehrlich LP, Nilges M, Wade RC. The impact of protein flexibility on protein-protein docking. Proteins 2005;58:126–133.
64. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
65. D'Alfonso G, Tramontano A, Lahm A. Structural conservation in single-domain proteins: implications for homology modeling. J Struct Biol 2001;134:246–256.
66. Rodriguez R, Chinea G, Lopez N, Pons T, Vriend G. Homology modeling, model and software evaluation: three related resources. Bioinformatics 1998;14:523–528.