A randomized global optimization method for protein-protein docking

B. Addis • F. Schoen

Dipartimento di Sistemi e Informatica via di Santa Marta, 3 50139 Firenze (Italy)
b.addis@ing.unifi.it • schoen@ing.unifi.it

In this paper we report results on the problem of docking two large proteins by means of a two-phase monotonic basin hopping method. Given an appropriate force field which is used to measure the interaction energy between two biomolecules which are considered as rigid bodies, we used a randomized global optimization methods based upon the repeated use of local searches. These local searches include two phases, the first of which is aimed towards obtaining a close match between the two interacting bodies by means of the inclusion of a penalty term in the energy, the reduction of the van der Waals radii of the atoms and the elimination of the electrostatic contributions. A single relatively small test example was chosen and used to tune the parameters of our method; the set of constants found this way was then consistently used to dock several other complexes obtained from the Brookhaven Protein Database. The results are very encouraging as in most cases the correct docking was eventually found and those cases in which a failure of the method was observed were later on discovered to possess characteristics which are incompatible with the hypotheses on which the experiments were made.

It is clear to the authors that rigid docking is only a rough approximation of what is really needed, i.e. flexible docking. However, in the case of protein-protein docking it has been observed by various authors that when two large molecules interact, the change in atom positions within the same molecule just occurs in the interface between the two. So it can be the case that rigid docking might prove a very valuable tool as a first phase in which two large molecules are put as close as possible. Flexible refinements can then be performed using the results of rigid docking as starting configurations. Some initial experiments within this context display quite promising results.

(*biomolecular docking, two-phase methods, basin-hopping*)

---

1. Introduction

When two biological macromolecules are close enough, they start an interaction which, in some cases, leads to the formation of a new complex. The term docking is frequently used in the scientific literature to denote the process by which two (or more) different molecules interact in such a way as to form a single complex. Being able to predict the structure of the complex resulting from a docking process is extremely relevant: knowing how (and where) a small ligand molecule docks to a protein might be useful in order to predict the effect that a new drug could produce in a living body. Hundreds of possible ligands can be docked to a large biomolecule in order to screen out those which do not appear to fit well or those whose docking site is different from the desired one, thus eliminating the highly expensive and time-consuming process of actually producing and experimenting with large sets of ligands. Most of the literature on docking methods deals with ligand-protein docking, or, in general, with the problem of docking two molecules, one of which has a very modest size. Several software packages already exist for this purpose and it can be safely assumed that the problem of docking small ligands can in many instances be solved with relatively low computational effort.

The real challenge, both for biologists and optimizers, is now protein-protein docking or, in general, the problem of docking two (or more) large biomolecules.

1.1 Difficulties of protein-protein docking

Several kinds of difficulties arise in this context. First of all there are many modelling difficulties. In fact what we know of proteins is quite different from what proteins in nature really are: there are huge databases of protein structures, but these are related to analyses mostly made on crystalized proteins or they are based on NMR observations. These structures are often re-engineered and often lack some important information; for example, crystallographic data reports only "heavy atoms", i.e., they do not report coordinates for hydrogen atoms, which account for roughly one half of the total number of atoms of a protein. So, in order to use some of the most common models for the evaluation of interactions, hydrogen atoms have to be artificially added into the structure and placed in "reasonable" positions, before attempting any docking process.

Also, and more important from the point of view of being able to check the validity of results, there is the problem of modelling the interactions. In fact, let us assume that, in

some sense, "Nature" optimizes some sort of energy, in the sense that the actual docking configuration is the one which minimizes the total energy of the complex. Of course this assumption may be disputed, but, even if it is accepted, the problem remains of finding an accurate and manageable model of energy function to be optimized. Many force fields have been defined in the literature from which an energy function can be evaluated and optimized. However, the best we can hope to obtain, is to find a docking configuration which minimizes the energy function used; the resulting docking should then be confronted with the actual docked complex, if this has been observed and inserted in the database.

The choice of a force field produces other kinds of difficulties. As an example, the Amber [AMBER] force field which we used in this paper was chosen mainly because it is a sort of standard in protein energy calculations and because its definition is of public domain. However the standard force field is capable of recognizing only molecules composed of the standard aminoacids. While it is known that different proteins are composed of sequences made only of the 20 standard aminoacids, in most cases proteins which have been observed and recorded in public databases (often as a consequence of the crystalization process) contain also so-called hetero-atoms, atoms which do not strictly belong to any specific aminoacid. The presence of these atoms rules out the possibility, for a non-biologist, to use standard force fields or, seen from another point of view, restricts the applicability of standard force fields to a very limited subset of proteins. This makes extensive numerical testing and validation quite difficult without the support of a biologist.

Given a force field and two biomolecules which we would like to dock, a global optimization method can in principle be used; it is however important to observe that in a naive approach in which all of the atoms in both molecules are free to change their relative positions, the number of variables becomes huge for most cases. Proteins usually possess a few thousands of atoms, and it is easy to obtain global optimization problems with tens of thousands of variables. Moreover, energy function evaluation is extremely costly, as it generally involves at least pair-wise contributions, so that if the number of atoms is in the thousands, the number of pairwise interactions to be evaluated at each energy function calculation is in the order of millions of operations. The situation is dramatically worse when we include three- and four-body interactions in the energy models, or when we wish to use gradient or higher order information.

Moreover, from the point of view of global optimization, docking problems are characterized by a huge number of local optima which are not global; there are no theoretical

results available on the complexity of docking, but it seems reasonable to estimate that the number of locally optimal docking configurations grows at least exponentially in the number of atoms. This conjecture obviously rules out any method which tries to obtain a certified global optimum: the best one can hope to obtain is a good local optimum, with no guarantee of global optimality.

A final difficulty in tuning global optimization methods and in assessing their capabilities or in validating their results, is the fact that quite often the putative optimum docking configuration found by a good algorithm has a total energy which is much lower than the energy, evaluated by means of the same energy function, of the actual docked units, as observed in nature. This is clearly a proof of the inadequacy of the energy model used to perform energy minimization; sometimes, as a partial criterion, an optimization algorithm is considered successful if it could detect the actual docking, even if other configurations had better rank. This is a partial remedy which can be used to partially assess the feasibility of the method, but surely it has to be complemented with other techniques to solve the real task of discovering the docking conformation of unknown complexes.

As a preliminary step, it is important to distinguish the two main classes of methods, namely rigid and flexible docking. By rigid docking we mean a process in which the two molecules, during the docking phase, are considered as rigid bodies. This way the relative positions of atoms within the same molecule do not change. This is of course a crude approximation and a strong simplification; the main advantage of assuming rigid docking is that the number of degrees of freedom is drastically reduced, from the tens of thousands in the general case to just six variables. It can in fact be assumed that one of the two molecules (usually the one with a larger number of atoms) is kept fixed (this molecule is called the *host*), while the other (*guest*) is allowed to rotate around its geometric center and to translate. The six variables are thus three rotation and three translation parameters. The problem, given any kind of force field, becomes thus a low-dimensional optimization problem. However it should be recalled that, although only six free variables exist in this case, energy evaluation again requires the computation of millions of pairwise interactions; also, even in this case, numerical experience tends to support the conjecture that the number of local optima depends on the number of atoms and not on the number of degrees of freedom. Thus even rigid docking remains an extremely challenging global optimization problem. The results from rigid docking might be used as an input to an expert biologist who can "manually" adjust the docked units to take into account flexibility. More realistically, the

results of rigid docking might be used as starting point in flexible docking methods.

As it can be easily understood, in flexible docking, as opposed to rigid one, the interacting units are not considered as rigid bodies and their shape is allowed to change in order to favour better couplings. Most flexible docking methods however do not allow every atom to move, as this would make the dimension of the optimization problem unmanageable. The main approaches within this context are those inspired by folding literature, in which the degrees of freedom are not the positions of atoms, but only a limited number of angles within the chain of aminoacids, or those based in the assumption that in flexible docking only those atoms of the two molecules which are close enough each other are allowed to move.

1.2 Brief analysis of recent relevant literature

Before introducing our approach to the docking problem, a short review of some relevant literature seems necessary. A quite detailed survey and numerical comparison of stochastic global optimization algorithms applied to the problem of docking small molecules is presented in [Diller and Verlinde1999]. The authors compare some standard global optimization methods, like random walk (in which at each step a new configuration is obtained by randomly perturbing the current one), Metropolis Monte Carlo simulated annealing, in which randomly perturbed configurations are accepted or rejected according to the standard Metropolis criterion, smoothing transformation methods, in which the objective function is smoothed, typically by means of a filter, in order to reduce or eliminate local minima. A fourth algorithm considered, called Trust (Terminal Repeller Unconstrained Subenergy Tunneling), in which after a local optimization, the objective function is modified in such a way that the current local minimum becomes a maximum from which to escape towards a lower minimum point. Similarly to classical tunneling methods (see e.g. [Levy and Montalvo1985]) the function is modified in order to avoid visiting worse local optima.

The paper by [Klepeis et al.1998] stands in a quite different position in the literature, as this is one of the few papers in which a deterministic (as opposed to randomized) global optimization method is employed. Here, the well known α-BB method [Adjiman et al.1998] is employed and tested in several cases of peptide-peptide docking; the results reported are very significative, but their applicability to large molecules seems to be quite a difficult task.

In [Totrov and Abagyan1997] a stochastic method is proposed in which random moves of a flexible ligands are performed, followed by local optimization. The model used is

one in which the (small) ligand has several degrees of flexibility, while the host (or receptor) molecule possess a limited flexibility in the neighborhood of the docking site. In [Apostolakis et al.1998], again the problem of flexibly docking a small ligand is considered. Here however while, as in the previous paper, local optimizations are performed, the energy function is gradually changed during the docking process, in order to simulate the effective interactions which occur in nature. Again, this paper considers the problem of docking small ligands, a problem which, although far from being solved, can however be considered quite mature, with good and reliable software easily available (see, for example, one of the de-facto standard packages, AutoDock [Morris et al.2001]).

In [Lenhof1995] a rigid docking method is described in which a novel scoring function is used to measure some kind of geometric complementarity between a host and a guest molecule; based upon such a score, a systematic search of possible docking conformation is performed; again the approach seems to be well suited only in the case of small ligand docking.

Recently, in [Fernandez-Recio et al.2002], a method is proposed for docking large proteins. The method consists of two phases: a rigid docking procedure, followed by a flexible docking in which only atoms in the interface between the two docked molecules are allowed to move. This approach seems to be particularly interesting, as it decomposes the problem into that of finding reasonable docking sites, which can be quite efficiently made by means of rigid docking, and that of refining the interactions between the two molecules. The good results reported in this paper encourages us in improving our methods: having a good procedure for rigid docking might prove to be a fundamental step towards a complete and computationally reasonable package for protein-protein docking.

We are perfectly aware that this literature review is far from being complete. However it seems that, while the problem of docking small ligands to large molecules, has been attacked in different ways from various authors, the problem of protein-protein docking, due to its enormous complexity, has received less attention in the literature. This paper is devoted to a procedure which, as some of our first experiments confirm, can be proposed as a first step towards automated protein docking.

2. An energy model for protein-protein docking

As it has been already pointed out, in this paper we describe a global optimization procedure based upon the use of the Amber force field. It should be again observed that the use of this force field was mainly dictated by its wide availability and its parameters, which are of public domain. Different force field should perhaps be used in order to avoid some of the pitfalls which will be described in the section devoted to numerical experiments. We trust that our method is quite robust with respect to changes in the force field and indeed we made some comparison with a different version which used the Gromos force field and obtained almost identical results.

A short description of this force field, valid for many other force fields found in the literature, is the following: the energy of a molecule can be approximated by

$$\begin{aligned} \mathrm{E} &= \sum_{\mathrm{i}\in\mathrm{L}} \frac{1}{2}\mathrm{K}_i^b(\mathrm{r}_i - r_i^0)^2 && (1)\\ &+ \sum_{\mathrm{i}\in\mathrm{A}} \frac{1}{2}K_i^\theta(\theta_i - \theta_i^0)^2 && (2)\\ &+ \sum_{\mathrm{i}\in\mathrm{T}} \frac{1}{2}K_i^\phi[1+\cos(n\phi_i - \gamma)] && (3)\\ &+ \sum_{(\mathrm{i,j})\in C}\sum \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) && (4)\\ &+ \frac{1}{2}\sum_{(\mathrm{i,j})\in C}\sum \frac{q_i q_j}{\varepsilon r_{ij}} && (5) \end{aligned}$$

Here, atoms are considered as balls, and chemical bonds as springs. The first three terms in the above expression consist of so called bonded interactions, as they refer to groups of atoms linked two by two by chemical bonds. In particular, letting L denote the set of all pairs of atoms linked by a covalent bond and r_i the distance of atoms within these pairs, term (1) represents the energy due to the oscillation of the bond length around an equilibrium value r_i^0. Symbol A denotes the set of all groups of three consecutive atoms linked by chemical bonds (i.e., atom k is bonded with atom $k+1$, and atom $k+1$ is bonded to atom $k+2$) and θ_i the angle formed by these three atoms; term (2) takes into account the energy due the oscillation of the angle around an equilibrium value θ_i^0. Term (3), as the first ones, takes in account an oscillation around some equilibrium values. The angle considered in this term is the dihedral (or torsion) one formed by the two planes identified by a group of four consecutive bonded atoms (these groups form set T).

The last two terms refer to all possible pairs of non bonded atoms (set C), and, r_{ij} is the distance between a pair of atoms (i, j) in C. Term (4) contains the attractive Van der Waals interaction and an artificial repulsive term whose aim is to avoid atom collisions. The minimum of the pairwise interaction (4) is reached when the distance of two atoms is equal to a constant which is defined as the sum of the two van der Waals radii of the atoms. A_{ij} and B_{ij} are constants which depend on the types of atoms i and j. Finally, the term in (5) is the electrostatic interaction, which depends on the electric charges of atoms, respectively denoted by q_i and q_j; ε is a constant.

While the above description can be used as a basis for molecular shape optimization (like, e.g., the well known problem of protein folding), when dealing with docking some important modifications are necessary. In particular, when rigid docking is considered, two interacting molecules become rigid bodies; thus no modification of internal structure is allowed. This has the consequence that, when evaluating the energy of a complex, all of the contributions due to bonded interactions (assuming that covalent bonds occur only inside each molecule) account for a constant term. Also, for the same reason, all of the Van der Waals and Electrostatic contribution caused by non bonded pairs *within the same molecule*, again contribute a constant term to the energy. So, in evaluating docking positions, the following interaction energy can be considered:

$$E(v) \quad = \quad \sum_{i \in \text{guest}} \sum_{j \in \text{host}} \left(\frac{A_{ij}}{r_{ij}^{12}(v)} - \frac{B_{ij}}{r_{ij}^{6}(v)} \right) + \frac{1}{2} \frac{q_i q_j}{\varepsilon r_{ij}(v)} \tag{6}$$

Here host and guest represent the sets of atoms of two different molecules, r_{ij} is the euclidean distance between two atoms (one belonging to each molecule), and v is a roto-translation vector, representing the 6 degrees of freedom (three rotational parameters and three translations) of the guest molecule with respect to the host. Although this energy function depends explicitly on six variables only, its evaluation requires the computation of all the contributions from pairs of atoms chosen in all possible ways in the two molecules.

3. Two phase monotonic basin-hopping methods

The method we used to perform rigid docking is a combination of two different approaches:

1. Monotonic Basin Hopping [Leary2000]

2. Two Phase local optimization [Locatelli and Schoen2002],[Locatelli and Schoen2003]

Monotonic basin hopping consists of repeatedly performing local optimizations starting from points randomly generated in the neighborhood of the current one: as soon as a local optimization ends up in a local minimum whose value is better than the current one, this local minimum becomes the current point.

Two-phase local searches were introduced in order to increase the region of attraction of the global minimum in Lennard-Jones cluster conformation problems. In words these local searches consist of a first phase in which, starting from a point, a local optimum of a function which includes penalty terms is found; from the local optimum found during this first phase, a standard local optimization is started. The first phase thus aims at producing good starting points for local optimization. The form and structure of the penalty terms are problem-dependent. In [Addis and Schoen2003] some preliminary experiments were performed in a simulated docking problem in which a Lennard-Jones cluster of identical atoms was cut into two parts which were then recombined by means of a docking procedure. The experience gathered in that experiments lead us to the definition of a suitable penalty which was then used in the context presented in this paper.

In some more details, the combined method consisting in monotonic basin-hopping with two phase local searches can be described as follows:

start: let $V = [V_1, \dots V_6] \in R^6$ be a starting configuration, generated by randomly displacing the guest molecule in a sufficiently large sphere centered at the geometrical barycenter of the host.
let $NoImprove := 0$;

Local Optimization: find

$$Z = \text{TwoPhaseLocOpt}(V)$$

Loop: While $NoImprove < MaxNoImprove$ do:

Perturbation: Let

$$\tilde{V} := Z + \varepsilon$$

be a perturbation of Z, where ε is a suitable random vector;

Local Optimization: find

$$\tilde{Z} = \text{TwoPhaseLocOpt}(\tilde{V})$$

Acceptance/rejection:

If

$$E(\tilde{Z}) < E(Z)$$

then let $V := \tilde{Z}$ and $NoImprove := 0$;

else let $NoImprove := NoImprove + 1$;

End loop

end: return(Z)

The two phase local optimization used in this context can be described as follows:

Procedure: TwoPhaseLocOpt(X);

Phase I: find

$$Y = [Y_1, \ldots, Y_6] \quad = \quad \arg \operatorname{local} \min_v ME(v; \alpha, \beta, \gamma)$$

using $X = [X_1, \ldots X_6] \in R^6$ as the starting point;

Phase II: find

$$Z = [Z_1, \ldots, Z_6] = \arg \operatorname{local} \min_v E(v) \tag{7}$$

using Y as a starting point;

end: return(Z).

Here the modified energy function (ME) is defined as follows:

$$ME(v; \alpha, \beta, \gamma) \quad = \quad \sum_{i \in \text{guest}} \sum_{j \in \text{host}} \left(\frac{A_{ij}}{r_{ij}^{12}(v)} - \frac{B_{ij}}{(r_{ij}/\alpha)^6(v)} \right) \tag{8}$$

$$+ \quad \beta \frac{1}{2} \frac{q_i q_j}{\varepsilon r_{ij}(v)} \tag{9}$$

$$+ \quad \gamma \sqrt{r_{ij}(v)} \tag{10}$$

In this modified energy function we use three parameters in order to obtain different penalties (however, as it will be seen in the analysis of the numerical experiments, in most of the experiments we used a single set of parameter values). In the first term of the penalized objective function (8), the van der Waals radius is changed through a rescaling of the distance term which appears in the attractive contribution (the term in which the 12-th power of the

distance appears is included in order to avoid an atom collapsing against another one). Choosing a value of α strictly greater than one has a effect which is similar to a reduction in the van der Waals radius: it is trivial to see that in a generic van der Waals term

$$\frac{A}{r^{12}} - \frac{B}{r^6}$$

the minimum occurs at

$$r^\star = \left(\frac{2A}{B}\right)^{1/6}$$

so that, as a function of α, the modified van der Waals potential has its minimum in

$$r^\star(\alpha) = \frac{1}{\alpha}\left(\frac{2A}{B}\right)^{1/6}$$

Thus, choosing $\alpha > 1$ has the effect of allowing deeper compenetration of the guest inside the host, thanks to the fact that the pairwise minimum moves towards shorter distances. The second parameter is used in order to weight the importance, during the first phase, of electric charges: much of the literature on docking reports a scarce importance of the electrostatic terms during docking, probably as a consequence of the fact that protein docking in nature occurs inside water, where the electrostatic interactions are somewhat smoothed out. Accordingly, in most of our experiments we just put $\beta = 0$. The third term in the penalized energy function is an artificial penalty used in order to strongly attract the guest molecule to the host. This term has no physical meaning (it can be thought as a kind of elastic force between the two molecules) but its use in the two phase method is extremely beneficial. The inclusion of a penalty term of this kind was already proven to be of exceptional computational value in the above cited papers on Lennard-Jones cluster conformation problems. Here it is included in order to drive the guest protein quickly towards the host one, even when starting from a very far initial point, and also in order to favor a strong complementarity between the two molecules. In some sense we could think that the guest is strongly pushed against the host; this strongly attractive term is balanced, at short distances, by the steep barrier which prevents atoms to be too close each other.

After a local minimum of the modified energy function has been found, a standard local optimization on the original energy function is started. In the case of flexible docking (which is not the subject of this paper) some authors report good results in using softer potentials (e.g., potential in which the van der Waals exponents 6-12 are lowered). It would be extremely easy to include such softer potentials either in the first or in the second phase of our method.

4. Materials and methods

The proposed procedure was coded in C++ and run on an Intel-based Linux PC at 1600Mhz. As a local optimization routine, both for the first and for the second phase, we used a C++ version of the limited memory BFGS quasi newton method described in [Liu and Nocedal1989]. The most difficult part of the numerical experiments was that of finding examples and testing the results. As it has been already observed in the introduction, the choice of a particular force field implies the difficulty (or the impossibility) of working with some sets of proteins; using Amber we were forced to consider only those complexes in which no hetero-atoms (possibly except water) were contained. We scanned the Brookhaven Protein Database in order to find proteins made only of standard aminoacids; moreover we restricted our search to complexes made of two molecules of comparable sizes. Finally, in the resulting set, we choose those complexes which were not too large. After this initial screening, we considered the following examples:

`1ciq` (proteinase inhibitor)
`1egp` (proteinase inhibitor)
`1bxp` (peptide/toxin complex)
`1l6e` (protein kinase and dimerization domain)
`1a03` (calcium-binding protein)
`1aap` (proteinase inhibitor)
`1kl0` (transferase)
`1qh2` (hydrolase inhibitor)

For each of these complexes, the following procedure was applied in order to prepare the molecules:

1. Each file was scanned for hetero-atoms; in particular, given the pre-screening made on the pdb database, no hetero-atom is present in these structure except water. If water was present, it has been deleted;

2. Each complex was scanned to check the presence of hydrogen atoms; depending on the procedure used to obtain atom coordinates, some pdb files report hydrogen positions, while some others do not. If hydrogens were absent, they were added to the pdb file by means of the `add_hydrogens` method of the BALL [Boghossian et al.1999] library. Among these examples, only `1ciq, 1aap, 1kl0` needed the insertion of hydrogen atoms.

3. the resulting PDB was locally optimized using the Amber force field and letting every atom in the complex move; this all-atom local optimization was performed in order to obtain a target protein-protein complex which is indeed a local minimum with respect to the force field used. The Amber force field was implemented without any cut-off and the local optimization used was the `minimize` method in the `ConjugateGradientMinimizer` class of the BALL library. The resulting structure was then saved as a pdb file.

4. for each molecule, composed of two aminoacid chains, one of the chain was labeled "host" and the other "guest". The criterion used has always been that of attaching the "guest" label to the smaller chain.

5. the interaction energy between the host and the guest was recorded; this information was obtained by computing the energy of the host, that of the guest and that of the complex separately and then defining

$$E_{interaction} = E_{Complex} - (E_{Host} + E_{Guest})$$

   As it was expected, only van der Waals and Electrostatic terms were significantly different from zero, while all the other components, known as stretch, bond and torsion, were null.

The following table summarizes some characteristics of the prepared complexes; in particular we report the name of the pdb file, the number of atoms in the guest and host molecules, the interaction energy between the host and the guest in the optimized complex, the root mean squared distance (RMSD) between the original and the optimized complex (for complexes for which we had to add hydrogen atoms, the RMSD was computed only with respect to heavy, i.e. non hydrogen, atoms).

pdb	# Atoms Host	# Atoms Guest	Interaction Energy (cal)	RMSD
`1ciq`	593	415	-341.092	0.487542
`1egp`	637	353	-333.396	0.412625
`1bxp`	1038	267	-408.678	1.04548
`1l6e`	763	763	-337.495	0.842386
`1a03`	1448	1448	-335.904	0.277063
`1aap`	818	818	-46.103	0.473728
`1kl0`	2011	267	-113.578	0.539769
`1qh2`	398	273	-307.143	1.21655

The numerical experiments were performed by running 150 independent instances of our two-phase monotonic basin hopping method as described in detail in section 5; in order to have more reliable and comparable results, we used the same parameters in all the tests and also used the same sequences of random numbers in generating the 150 starting configurations. Each starting configuration was obtained by translating the geometric centers of both the guest and host to the origin of the three-dimensional space; then the guest was given a uniform random rotation around its center and successively its geometric center was translated to a random position at distance 100Å from the origin. Inside the two-phase monotonic basin hopping method the perturbation vector was chosen randomly; in particular, the perturbation of the three rotation variables was chosen uniformly on $[-0.5\pi, 0.5\pi]$, while the perturbation on the three translation variables was generated uniformly in $[-5\text{Å}, 5\text{Å}]$. These parameters were chosen after some experimentation with molecule 1ciq and kept constant for all subsequent experiments.

For the definition of the modified potential energy function (8–10) we used the following parameters: $\alpha = 1.6$, $\beta = 0$ (no use of Coulomb interactions in the first phase) and $\gamma = 5$. The run was stopped as soon as for the last 30 iterations no improvement was made.

5. Discussion on Numerical results

The numerical results obtained can be more easily analyzed one by one:

1ciq : this is a relatively small example which was used as a test for the tuning of parameters. For this complex several different combination of parameters were used in order to find a single set of constants to be used in all subsequent experiments. With the set of parameters which was subsequently used for all other complexes, we obtained 2% success rate and the best docking found was indeed the correct one. Changing the γ penalty term in (10) from 5 to 1 lowered the success rate (in 150 trials)to 1%; augmenting the same parameter to 10 made the success rate go to 1.33% – it has to be remarked that we used common random numbers, so that in each experiment the same starting configurations were used. Among many other trials, here, following what many authors suggest, we tried also a first phase potential in which the van der Waals term is softer, an effect which was obtained by lowering the 12-6 exponents in (8) to 8-4, but no success was obtained. Also we obtained no success when eliminating the γ penalty term.

1egp : for this complex, which consists of two relatively small-sized molecules, we could find the correct docking in 4.7% of the runs performed (i.e. we obtained 7 successes in 150 random trials). For this complex we compared the performance of our two-phase method with a single phase one, i.e. with a standard monotonic basin hopping method, initialized from the same starting points and with all the relevant parameters kept constant. In 150 trials no success was obtained: the best configuration, which was found in 56% of the runs, had an interaction energy -225.152 cal, much worse than the correct one; the computed RMSD of the best result obtained without the use of the first phase was 10.23, an unacceptable result in this context. As a confirm of this, figures 1 and 2 display the optimal docking found and the one obtained with just the single phase basin hopping (in these as well as in the following figures, we colored in light blue the host and in orange the guest molecule).

This result shows that the compression term introduced in the penalty, as well as the reduction in the van der Waals radii, is extremely beneficial.

1bxp : in this case the optimum docking is found 16 times out of 150 trials (10.7% success rate)

1l6e : here the success rate was even higher, 13.3%, but there were false positives, that is the correct docking was not the one with minimum interaction energy. We found 5 different configurations which, according to the interaction energy, ranked best than the correct one. This is clearly not a failure of the algorithm, but of the force field: if the docking observed in nature is indeed a global minimum of the energy function, this is a clear proof that the Amber force field is not an appropriate model for the interaction energy. Finding a more appropriate model of the energy is a difficult problem which several authors are dealing with. We are however quite confident that our algorithm, given a different and more reliable force field, will certainly improve its performance on these examples.

1a03 : the correct docking was found as the best configuration in just one trial (success rate: 0.67%). It might be the case that with some parameter tuning this docking could have been discovered more easily, but in order to obtain a validation of our method, the experiments were all performed with exactly the same parameter sets.

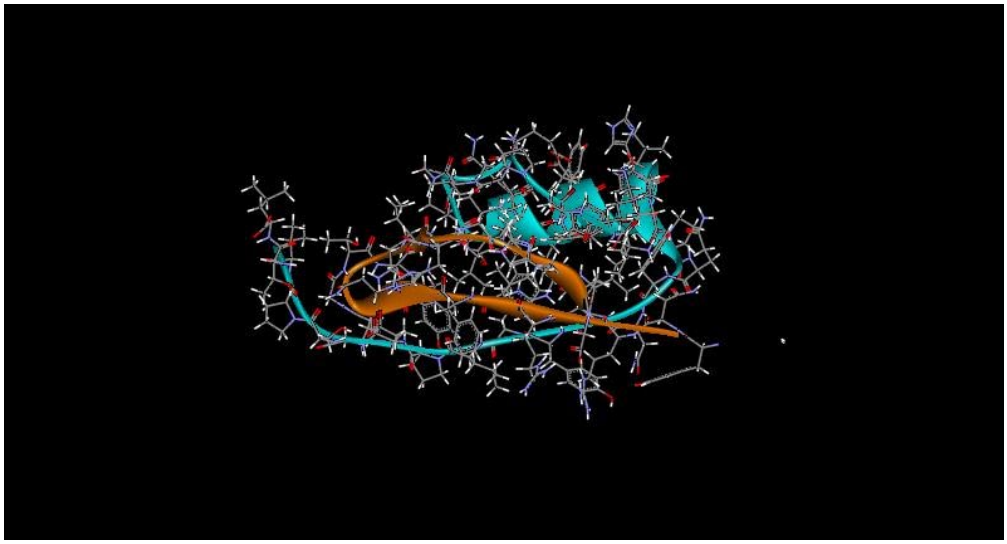

Figure 1: Optimal docking of 1egp found with our two-phase method (energy: -282.889)

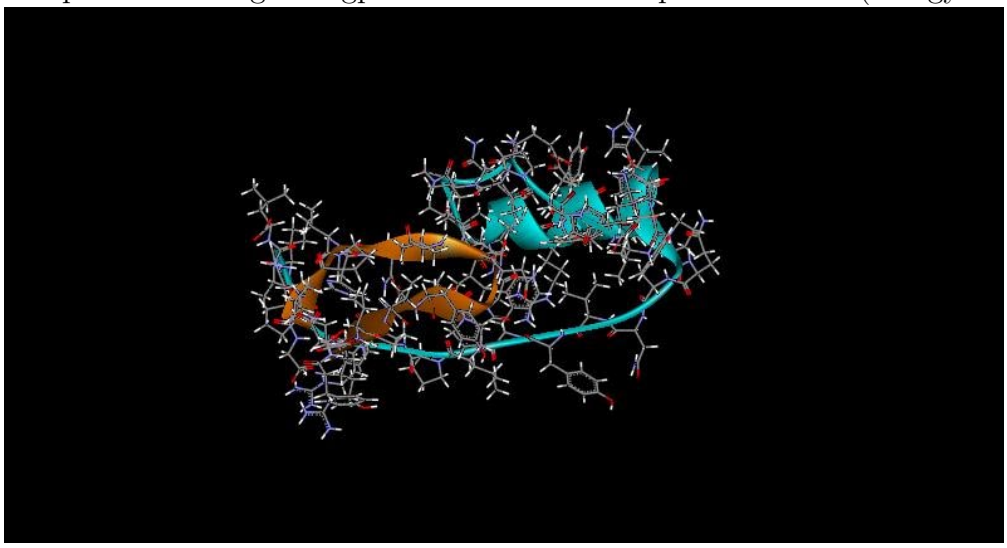

Figure 2: Optimal docking of 1egp found with a standard basin-hopping method (energy: -225.152)

1aap : here our method never found the correct docking and consistently obtained configuration with a significantly lower energy than the actual one. After an inspection of the pdb file a reasonable explanation of this failure could be found: in the actual docking reported in the pdb file, the interface between the two molecules is filled with water, which we, in our experiments, removed: this removal drastically changes the energies in the interaction and prevents the possibility of obtaining the correct docking. Also,

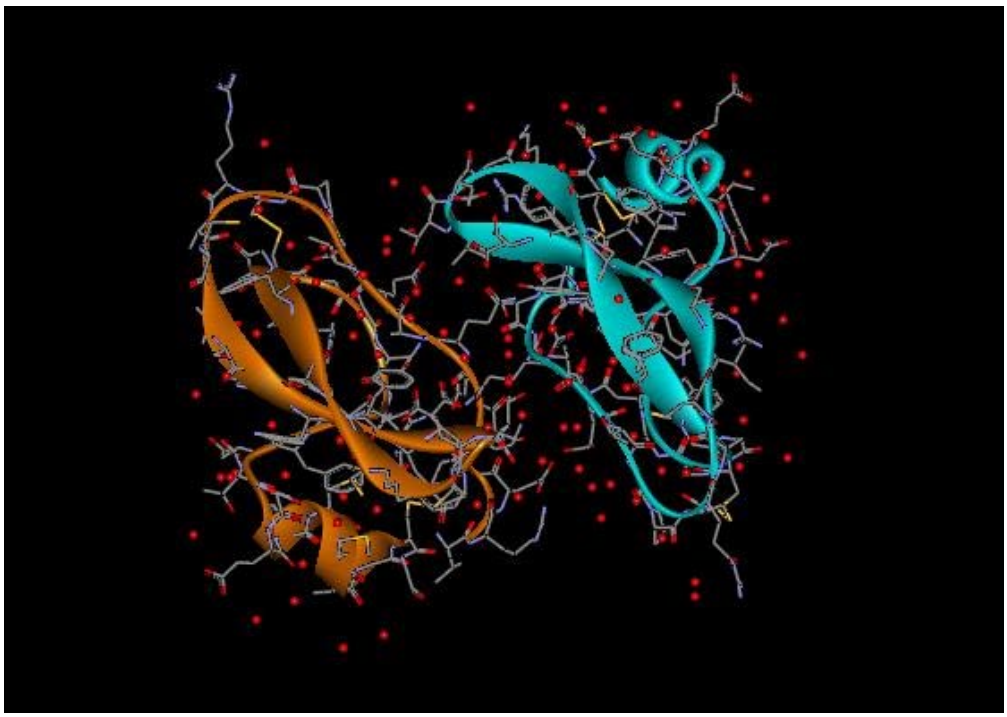

Figure 3: 1aap: red spheres represent the oxygen atoms of water molecules

it becomes clear the reason why a lower energy is found: taking away water molecules, the two proteins can be matched more closely one to the other, thus significantly lowering the van der Waals contribution. Unfortunately our method cannot cope with water molecules not only because of the lack of knowledge of the correct parameters of water in the Amber force field. What makes our method unsuitable for this problem is the fact that if water molecules were included in the model, each of them should considered as a single rigid body, independent of all the others. A better alternative would be to use a different force field which is capable of implicitly taking into account the effect of the solvent.

1kl0 : here the situation observed in the previous example goes to an extreme: we could

indeed find the correct docking (with success rate 1.3%), but most of the other configurations had a better ranking - i.e. almost all of the results were false positives. Again this is a consequence of the force field used; in this case, looking at the picture of the molecular structure in figure 4, it is easy to notice that there are many "good" docking sites.

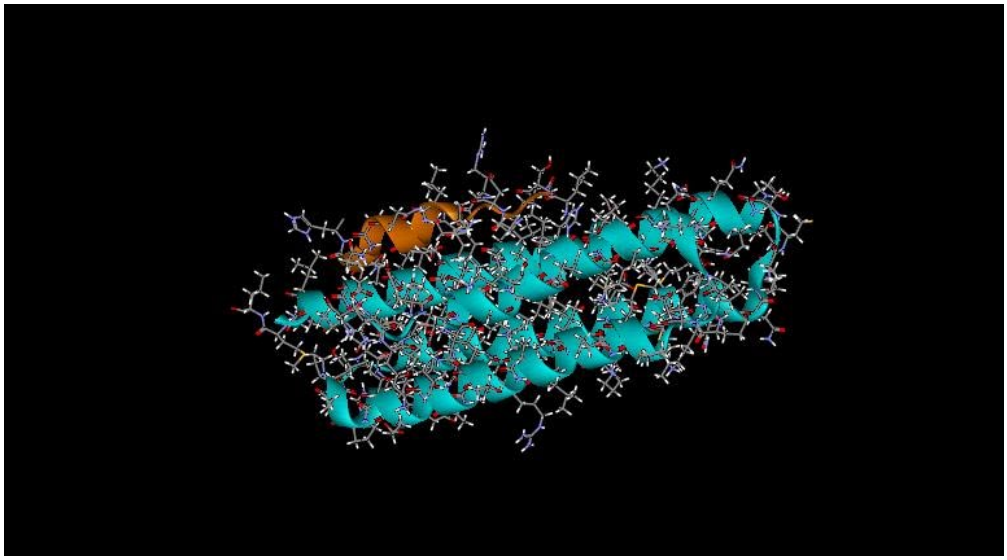

Figure 4: Optimal docking in 1kl0

1qh2 : here we got no success in 150 runs. However, after looking more closely to the structure of this molecule, and conforted by the opinion of expert biologists, we discovered that this example is not a docked complex, as between what we erroneously labelled the host and the guest some chemical bonds exists (in particular, there are some sulphur bridges): this bonds are so strong that the two parts are always complexed – i.e., this is not a docking problem as the host and guest never, in nature, exist separately. The failure of the method, in this case, is caused by the inappropriateness of the example.

We actually performed a huge number of different experiments on different structures. Here we reported some of those results which we consider as most significant. There is a great difficulty, in this field, in performing sensible numerical experiments, as data on complexes, although easily available, are very difficult to use, and no test set is available in the literature. The failure on `1qh2`, if not understood thanks to the help of biologists, might have caused our research to stop. It would be very good if an official dataset of docking problems, coupled

with an energy function, could be available for the scientific community in order to compare algorithms on a recognized and common test set.

6. Conclusions

We think that the results obtained using a two-phase monotonic basin-hopping method for rigid docking of large molecules are very encouraging; the computational times necessary for running 150 independent test on each of the previously shown complexes varied between 20 and 40 CPU hours on a Linux Intel-based PC at 1600MhZ. Of course the real challenge is to dock, allowing flexibility, two unbounded molecules. We are in an initial stage towards this objective and the first results obtained are quite encouraging. In particular, following ideas contained in [Fernandez-Recio et al.2002], we are using our rigid docking procedure on two unbounded molecules; starting from the configurations obtained by means of rigid docking, we perform local optimizations in which the atoms which are in the interface between the two molecules are allowed to move, thus giving some degrees of flexibility to the complex. After this local optimization on the flexible molecules we perform a small random rototranslation of the guest with respect to the host (like in rigid docking) and re-optimize in a flexible way the atoms in the interface between the host and the guest molecule. This part of the procedure is then repeated in a way which is analogous to the standard monotonic basin hopping method. Some preliminary results obtained within this framework are quite encouraging; the details of this part of the research will appear elsewhere.

Acknowledgments

This research has been partially supported by Progetto FIRB "Ottimizzazione Non Lineare su Larga Scala"

References

[Addis and Schoen2003] Addis, B. and F. Schoen: 2003, 'Docking of atomic clusters through nonlinear optimization'. *submitted*.

[Adjiman et al.1998] Adjiman, C., I. Andorulakis, and C. Floudas: 1998, 'A Global Optimization Method, αBB, for General Twice-Differentiable Constrained NLPs - II.

Implementation and Computational Results'. *Computers and Chemical Engineering* **22**, 1159–1179.

[AMBER] AMBER, 'Amber Homepage', `http://www.amber.ucsf.edu/amber/amber.html`.

[Apostolakis et al.1998] Apostolakis, J., A. Pluckthun, and A. Caflisch: 1998, 'Docking Small Ligands in Flexible Binding Sites'. *Journal of Computational Chemistry* **19**, 21–37.

[Boghossian et al.1999] Boghossian, N., O. Kohlbacher, and H.-P. Lenhof: 1999, 'BALL - Biochemical Algorithms Library'. Technical Report MPI-I-99-1-002, Max-Planck-Institut für Informatik.

[Diller and Verlinde1999] Diller, D. J. and C. L. Verlinde: 1999, 'A Critical Evaluation of Several Global Optimization Algorithms for the Purpose of Molecular Docking'. *Journal of Computational Chemistry* **20**, 1740–1751 .

[Fernandez-Recio et al.2002] Fernandez-Recio, J., M. Totrov, and R. Abagyan: 2002, 'Soft Protein–Protein Docking in Internal Coordinates'. *Protein Science* **11**, 280–291 .

[Klepeis et al.1998] Klepeis, J. L., M. G. Ierapetritou, and C. A. Floudas: 1998, 'Protein Folding and Peptide Docking: A Molecular Modeling and Global Optimization Approach'. *Computers and Chemical Engineering* **22 Suppl. 1**, S3–S10.

[Leary2000] Leary, R. H.: 2000, 'Global Optimization on Funneling Landscapes'. *Journal of Global Optimization* **18**(4), 367–383. .

[Lenhof1995] Lenhof, H.: 1995, 'An Algorithm for the Protein Docking Problem'. Technical report, MaxPlanckInstitut fur Informatik.

[Levy and Montalvo1985] Levy, A. and A. Montalvo: 1985, 'The Tunneling Method for Global Optimization'. *SIAM J. of Sci. and Stat. Comp.* **1**, 15–29.

[Liu and Nocedal1989] Liu, D. and J. Nocedal: 1989, 'On the Limited Memory BFGS Method for Large Scale Optimization'. *Mathematical Programming* **B 45**, 503–528.

[Locatelli and Schoen2002] Locatelli, M. and F. Schoen: 2002, 'Fast Global Optimization of Difficult Lennard-Jones Clusters', *Computational Optimization and Applications* **21**(1), 55–70.

[Locatelli and Schoen2003] Locatelli, M. and F. Schoen: 2003, 'Efficient Algorithms for Large Scale Global Optimization: Lennard-Jones Clusters', *Computational Optimization and Applications (to appear)*.

[Morris et al.2001] Morris, G., D. Goodsell, R. Huey, W. Hart, S. Halliday, R. Belew, and A. Olson: 2001, 'Autodock 3.05 - Automated Docking of Flexible Ligands to Receptors - User's Guide', The Scripps Research Institute .

[PDB] PDB, 'Brookhaven Protein Databank' `http://www.rcsb.org`.

[Totrov and Abagyan1997] Totrov, M. and R. Abagyan: 1997, 'Flexible Protein–Ligand Docking by Global Energy Optimization in Internal Coordinates', *PROTEINS: Structure, Function, and Genetics* **Suppl. 1**, 215–220. .