
Journal of Global Optimization, **30:** 1–21, 2004.

Docking of Atomic Clusters Through Nonlinear Optimization

B. ADDIS and F. SCHOEN
Dipartimento di Sistemi e Informatica, Università di Firenze, Italy, via S. Marta 3, 50139 Firenze (*e-mail:* {*b.addis, schoen*}*@ing.unifi.it*)

(Received 7 August 2002; accepted 14 August 2003)

Abstract. The problem of molecular docking is defined as that of finding a minimum energy configuration of a pair of molecular structures (usually consisting of proteins, DNA or RNA fragments). It is often assumed that the two interacting structures can be considered as rigid bodies and that it is of interest to researchers to develop methods which enable to discover the potential binding sites. Many different models have been proposed in the literature for the definition of the potential energy between two molecular structures, most of which contain at least a term (known as Van Der Waals interaction) which accounts for pairwise attraction between atoms, a repulsion term and a term which takes into account electrostatic forces (Coulomb interaction). Some well known models, and in particular those used in rigid docking, are based on the assumption that the only terms which are relevant in the process of docking are pairwise interactions between atoms belonging to the two different parts of the structure. In this paper the problem of finding the lowest energy configuration of a pair of biomolecular structures, considered as rigid bodies, is defined and formulated as a global optimization problem. In terms of dimension of the search space this formulation is not 'high-dimensional', as there are only six degrees of freedom: 3 translation and 3 rotation parameters. However the energy surface of the docking problem is characterized by a huge number of local minima; moreover each function evaluation is quite expensive (interesting structures usually possess a few thousand atoms each). So there is a strong need both of local and of global optimization procedures. In this paper a local optimization technique, based upon standard non linear programming software and a penalized objective function, is introduced and its potential usefulness in the context of global optimization is outlined.

Key words: biomolecular docking, Lennard-Jones clusters, Multistart, two-phase methods.

1. Introduction

Many important functions of the human organism are linked to the activity of one or more proteins, or some other biomolecules as nucleic acids. The activity of proteins is frequently related with their capability of linking with other molecules and change their properties. To explain this mechanism it is useful to use as an example the action of enzymes. Chemical reactions can occur only when a threshold of energy is reached. Sometimes this threshold is so high that the reaction, in normal conditions, does not take place. The enzyme acts linking to the host molecule, modifying its geometry and its binding capability; this way it allows an easier interaction with other molecules. During this process the surfaces of the molecules become very close each other, as the interaction

becomes non-negligible only at very short distances. To obtain a close match between the molecules, a strong complementarity of the two surfaces involved in the binding is required. Thanks to this distinctive characteristic, the process of approaching is called 'docking'. In this paper we will provide a tool for finding the optimal docking configuration of two molecules by means of local and global optimization applied to a function which expresses the potential energy of the system. In the next section a detailed model for the definition of the potential energy is introduced. Then, after a review of some computational approaches to the docking problem found in the literature, the general idea of two phase methods for energy minimization will be outlined. Then the application of global and local optimization to a simulated experiment consisting in splitting into two parts a molecular cluster and recovering the original conformation through an algorithm for docking will be presented. Finally some preliminary results on the problem of docking proteins of interest to biologists will be presented. The main purpose of this paper is to introduce a technique for global optimization in biomolecular docking; more refined algorithms for rigid docking and a much more extensive computational study of protein-protein docking will appear elsewhere (Addis and Schoen, 2003).

2. A Model for the Potential Energy of a Docking Configuration

A docking configuration of two molecules is assumed to correspond to the configuration in which the potential energy of the whole system is minimum. Evaluating the actual energy of a complex molecule, and, in particular, the free energy, is in general very expensive. Thus, in particular for molecular dynamics simulations, an approximate mathematical model is necessary. In the literature many different models have been proposed; most common force field expressions can be represented as follows:

$$E = \sum_{i \in L} \frac{1}{2} K_i^b (r_i - r_i^0)^2 \tag{1}$$

$$+ \sum_{i \in A} \frac{1}{2} K_i^\theta (\theta_i - \theta_i^0)^2 \tag{2}$$

$$+ \sum_{i \in T} \frac{1}{2} K_i^\phi [1 + \cos(n\phi_i - \gamma)] \tag{3}$$

$$+ \sum_{(i,j) \in C} \sum \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) \tag{4}$$

$$+ \frac{1}{2} \sum_{(i,j) \in C} \sum \frac{q_i q_j}{\varepsilon r_{ij}} \tag{5}$$

Some force fields also include an additional term which takes into account the solvation energy, which is caused by the interaction with a solvent (frequently

water). In the above model, atoms are considered as balls, and chemical bonds as springs. The first three terms in the above expression are 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 chemical 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 into 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) is the Van der Waals interaction, which is the effect of the sum of an attractive and a repulsive force. 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.

This scheme is common to a large number of different existing force fields (some of the best known of which are GROMOS[1], CHARMM[2], ECEPP[3], AMBER[4]) which mainly differ for different choices of the parameters.

In order to reduce the number of variables and the computational time a common approximation is to consider the two molecules as rigid bodies, as, in fact, it has been observed that during the docking process the deformation is relatively small. Some recent results (Fernandez-Recio et al., 2002) show that a rigid docking phase can be successfully used as a first phase followed by a refined flexible docking phase. In other words, rigid docking can be used to produce a number of potentially promising docking configurations, which are then used as starting points for an optimization phase in which at least those atoms which belong to the interface between the two molecules are allowed to move in a non-rigid way. With the assumption of rigid docking, we can significantly reduce the computational effort needed to evaluate the potential energy; this does not mean that the overall computational complexity of the docking problem is

[1]GROninger MOlecular Simulation package.
[2]Chemistry at HARvard Macromolecular Mechanics.
[3]Empirical Conformational Energy Program for Peptides.
[4]Assisted Model Building using Energy Refinement.

reduced as, in fact, it is widely believed that the Van der Waals and electrostatic contributions are responsible for the presence of a huge number of local minima; some experimental studies tend to support the opinion that the explicit addition of the other terms in problems like, e.g., protein folding or flexible docking, simplifies the optimization process.

In rigid docking we can consider one of the two molecules as fixed and the other in a position obtained through a roto-translation from a fixed position. We call the fixed molecule the *host* and the other one the *guest*. For what concerns the experiments, the guest is chosen as the smaller of the two molecules. Many contributions to the potential energy, due to the assumption of rigid docking, account for a constant and thus can be neglected during optimization. In particular all bonded interactions and all non bonded interactions within the host or the guest molecule can be neglected. It is also assumed that no chemical bond exists between the host and the guest, like e.g., sulphure or salt bridges. The only variable contributions are those due to the non bonded interaction between the two molecules. Thanks to this simplification, the problem to be solved can be formalized as follows:

$$v^* = \arg\min_{v\in\mathbb{R}^6} E(v)$$

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

Here r_{ij} is the euclidean distance between two atoms (one belonging to each molecule), and v is the roto-translation vector. The main difficulty in the solution of this problem is the extremely high number of local minima and the great variation in the shape of the objective function depending on the pair of molecules considered, so that, usually, general purpose methods are not efficient. Moreover the very high computational effort needed to evaluate the objective function and its gradient is to be taken into account: proteins have thousands of atoms and, for each function evaluation, we must compute the contribution for every pair of atoms, one in the guest and the other in the host, so the number of terms involved in each function evaluation is of the order of millions. Simplified models have been proposed in the literature, the most promising of which are those based upon pre-calculated potential grids. We plan to explore the possibility of including such approximate models in the near future.

3. A Short Review of Existing Approaches

It is quite difficult to compare different methods for solving the docking problem, as there is no unique model for the potential and, in many cases, it is not easy to obtain reliable data on the molecules used in the tests. Frequently papers report only the values of the energy and not the optimal configurations, so

results obtained with different force field cannot be compared. Other differences in the energy can be introduced by simplification of the model, for example considering rigid docking instead of flexible docking. In fact despite the fact that rigid docking can be considered as a good approximation, it is difficult, without a deep knowledge in biochemistry problems, to understand if the solution obtained with an approximate model is indeed a good solution in practice.

In the following we briefly cite some of the approaches reported in the literature for the docking problem; we recall however that, for what concerns methods based upon global optimization, a good review is found in (Diller and Verlinde, 1999). Some docking methods are based upon ligand/receptor knowledge, like e.g. (Rosenfeld et al., 1995); there it is assumed that the position of the binding site in one molecule is a priori known and an algorithm is used to dock the other molecule by fitting the position of the ligand to the binding site. This prior knowledge is usually obtained by experimental observations and crystallographic data. Other methods are based upon the geometric characteristics of the surfaces. One example of these approaches is described in (Lenhof, 1995) where, in order to reduce partially the computational complexity of the algorithm, atoms are assigned to cubes in a regular lattice in three-dimensional space. While one of the two molecules is kept fixed, the other is roto-translated, and an index is computed which counts the number of atoms of both molecules within each cube. Cubes can be internal, external or on the surface of a molecule. A fitness function is defined which takes into account the number of surface cubes of one molecule which contain atoms of the other. In this fitness function terms are subtracted when atoms of the two molecules overlap. Given this function, an exhaustive search of possible dockings is made and the best one is recorded.

In this paper we cannot go into details in the method used to perform this sort of exhaustive search: it may suffice to recall that the method prescribes to choose three atoms in both molecules and make them match as closely as possible. The algorithm runs by considering all possible matchings of triples chosen from a set of promising surface points. Most of the approaches which rely on global optimization algorithms are *ab-initio* methods: in this case no structure is generally imposed on the docking configuration, nor any prior information on the docking sites is imposed. Instead, based upon 'first principles' models for the potential and some structural knowledge, global optimization algorithms are used to drive the two molecules toward a close contact. Many authors use variations of genetic algorithms in this context, like (Wang et al., 1999), (Morris et al., 1998). In other approaches, deterministic global optimization methods, like the αBB method described in (Klepeis et al., 1998) (Floudas et al., 1999), (Androulakis et al., 1997) and (Maranas et al., 1995) are used. This method is an implementation of a classical Branch and Bound scheme for global optimization; in order to be able to prune the Branch and Bound tree, a lower bounding technique is employed based on the convexification of the objective function. Other authors, see e.g. (Shao et al., 1997) and (More and Wu, 1997), use different smoothing techniques in the context of Lennard-Jones cluster optimization.

Although the idea of smoothing has some connection with the two-phase approach we are introducing in this paper, it should be observed that classical smoothing approaches transform the objective function into another one, hopefully easier to globally optimize, and then try to track the global optimum of the smoothed function to the original one. Although the idea is interesting, published results are much worse than those obtained, on the same problems, by two-phase optimization. Smoothing techniques are based on the idea of gradually moving from a smoothed, hopefully convex, function to the original one; two phase methods, which will be introduced later, do not attempt a smoothing or a convexification, but simply transform the problem in such a way that sampling in the region of attraction of the global optimum is easier. In Totrov and Abagyan (1997), a stochastic method is proposed in which random moves of a flexible ligand 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. Also in (Apostolakis et al., 1998) 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.

It can be observed that, in order to find new docking configurations, ab-initio methods might be preferable, as they do not require much prior information. Unfortunately genetic algorithms often have the disadvantage of being quite constrained by the characteristics of the initial population which, if not chosen carefully, might prevent the global optimum to be reached. On the other hand, exact methods like αBB, might have a computational cost which is prohibitively high for protein-protein complexes composed of tens or hundreds of amino acids, unless simplifying assumptions are made which can easily lead to the discovery of only local optima. A quite unexplored field is that of problem reformulation, an example of which can be found in (Huang et al., 2002); however it is still not clear if and how such an approach could be applied to protein docking.

In closing this short review we would also like to recall that in King et al., (1996) a potential function obtained by experimental data is described, in which interactions depending on solvability and entropy changes are considered. The resultant additive term does not substantially increase the computational effort in function evaluation. In the paper some examples obtained by minimizing with the same algorithm the new potential function and a classic potential (specifically CHARMM see (MacKerell et al., 1998)) are described. Using this new model, structures are obtained whose mean square distances from the experimental data are reduced respect to those obtained with classic potential energy. So it can be affirmed that the research in global optimization algorithms for molecular docking problem is moving along at least two important and complementary directions: first that of building new and reliable algorithms which are capable, given a

potential energy function, to find the best docking configurations; second, new and more precise models for the potential energy have to be developed, which take into account experimental data, but do not add too much to the already extremely high computational cost of function and gradient evaluations.

4. An Introduction to Two-phase Methods for Global Optimization

One of the simplest methods of global optimization is Multistart, which consists of locally minimizing the objective function from different starting points obtained by random generation. In order to try to improve on Multistart, several approaches have been designed in order either to reduce the number of unnecessary local searches by means of clustering techniques, or to modify the probability distribution of the starting points. In particular a recent approach for the minimization of the potential energy of Lennard-Jones clusters of identical atoms with no charge has been recently described in (Locatelli and Schoen, 2002) and (Locatelli and Schoen, 2003). There, exploiting the special structure of the problem at hand, a two phase local optimization was introduced in which the first phase, starting from a random initial configuration, is aimed at finding a good starting point for the potential energy minimization (the second phase). The objective function in the first phase is not the original one, but a modified function M in which penalty terms are added in order to augment the probability of obtaining a good starting configuration. Here we try to follow the same idea of looking for starting points better than those obtained through random sampling. In doing so, it is important to choose properties which are general enough, otherwise we could obtain a method which is only applicable to a very specific class of molecules. We build a function which penalizes conformations without certain specific properties; although we could have used constraints in order to reduce the feasible space, we considered that this choice could produce two negative effects: first the local optimization problem becomes a constrained one, harder to solve; second, in this way a rigid cut-off of solutions is performed, at the risk of excluding the optimal one.

Our idea is to construct a general method using only a priori information valid for a large part of biomolecular docking: one of these, for example, is the strong geometric complementarity.

The general scheme of a two phase local optimization, started from a random initial configuration X, is outlined as follows:

Procedure: TwoPhaseLocOpt(X);

Phase I: find

$$Y = [Y_1, \ldots, Y_6] = \arg \text{ local } \min_{v \in \mathbb{R}^6} M(v)$$

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

Phase II: find

$$Z = [Z_1, \ldots, Z_6] = \arg \text{ local } \min_{v \in \mathbb{R}^6} E(v) \tag{7}$$

using Y as a starting point;
end: return(Z).

In order to implement a two phase method, a definition of a suitable modified potential function has to be given. In the next sections some possible approaches will be suggested for docking problems. It should in any case be stressed that the two-phase approach is a local one; a two-phase optimization can be used as a substitute for regular local optimization in any global minimization method, like Multistart or Simulated Annealing or other methods, even deterministic ones, which rely on local searches.

5. Two Phase Methods for Docking Lennard-Jones Clusters

As the problem of docking realistic biomolecules is extremely hard to solve, we chose to test our algorithms on a simplified problem and then we used some of the results on the original problem. The simplified model chosen is based on Lennard-Jones clusters. A cluster is formed by n identical atoms with no electric charge. The interaction considered is the Lennard-Jones potential, which depends only on the distance of pairs of atoms r_{ij} and has the following form:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \frac{1}{r_{ij}^{12}} - \frac{2}{r_{ij}^{6}} \tag{8}$$

The Lennard-Jones cluster problem consists of finding the optimal configuration which minimizes this energy. In this problem the variables are the $3n$ coordinates of the centers of the n atoms in three-dimensional euclidean space.

This model was chosen for two reasons: Lennard-Jones clusters have been the subject of intense study and research, so putative global optima are available; moreover, the Lennard-Jones potential can be considered as a simplified form of the Van der Waals interaction, one of the main contributions to the potential energy of biomolecules.

In order to perform experiments on docking we started from a putative optimal cluster and split it to obtain two parts; one of the two parts was randomly displaced in $\mathbb{R}^3$. Then we tried to find the optimal docking configuration, which corresponds to the known original cluster, minimizing the interaction between pairs of atoms belonging to the two different sub-clusters. As in the biomolecular problem, we called the fixed part the host and the other one the guest; v is the roto-translation vector. The following optimization problem was thus solved:

$$v^* = \arg \min_{v \in \mathbb{R}^6} E(v)$$

$$E(v) = \sum_{i \in \text{guest}} \sum_{j \in \text{host}} \left(\frac{1}{r_{ij}^{12}(v)} - \frac{2}{r_{ij}^{6}(v)} \right) \tag{9}$$

During the first phase of the algorithm we would like to obtain configurations where the contact surface is large, and, in order to do so, we tried to put the two parts as close as possible. To obtain this effect we chose to insert a penalty term of the form $\sqrt[k]{r_{ij}}$ in the modified potential in order to strongly take into account the surface interaction; this way the local optimization algorithm is driven towards configurations in which many pairs of atoms are close to each other. We also need a barrier term to avoid the overlapping of atoms; the resulting modified function chosen for this problem is represented in (10).

$$M(v) = \sum_{i \in \text{guest}} \sum_{j \in \text{host}} \frac{1}{r_{ij}^{q}(v)} + \alpha \sqrt[k]{r_{ij}(v)} \tag{10}$$

and some plots for different values of the parameters are reported in Figure 1.

In the above expression the repulsive term and the attractive one contain parameters which have to be chosen; in particular, choosing a value of q smaller than 12 has the effect of generating a barrier which is softer than that of the original Lennard-Jones potential; α is a weight which is used to tune the contribution of the attractive term and k is used in order to test different shapes of the attractive term. Although this penalized function has no direct physical interpretation, an analysis of the effects of a similar penalization term in a bio-chemical context has been performed in (Doye, 2000).

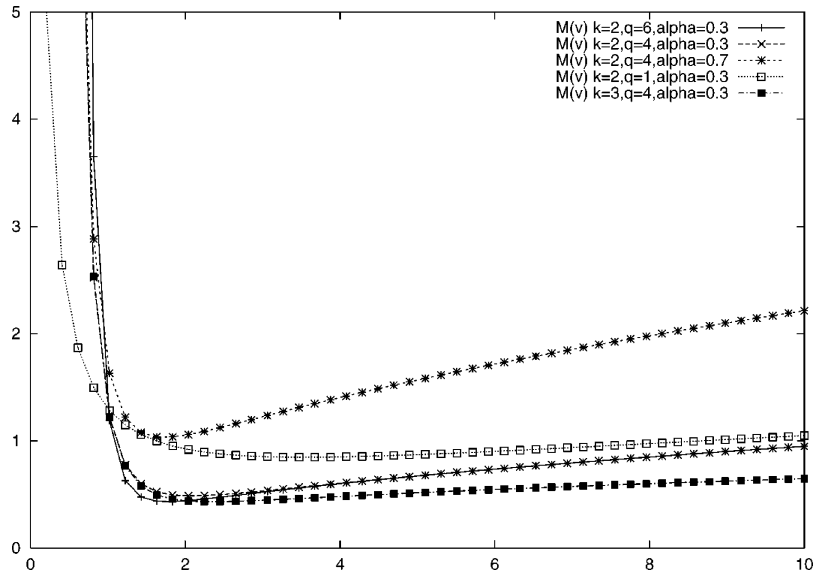

Figure 1. Modified function $M(v)$ with different parameters.

Several numerical experiments have been performed, with different choices of the parameters. We generated random initial configurations by randomly applying a rotation to the guest molecule and then displacing the guest molecule at a prefixed distance r from the host in a random direction; from each of these starting configurations, we performed local minimization both with the two phase algorithm and with a standard local method, in order to check whether the addition of a first phase is indeed useful; for this purpose, in the two phase algorithm we used the same local method as used in the single phase method, a limited memory BFGS algorithm (Liu and Nocedal, 1989). We generated 1000 random starting configurations and reported the number of successes – a run is called a success when the a priori known global optimum configuration is found. We remark that, in order to obtain more meaningful comparisons, the same starting points were used in the experiments, both with the standard and with two-phase methods. In most of the experiments starting configurations were randomly generated, as already described, with $r=6$; in some cases in which the standard algorithm could not find the correct docking, experiments were performed also with $r=3$ (see Tables 5 and 7).

We first started our testing on widely different examples and we noticed a great variation in the behavior of the method on different cases, so we chose to work on a more significant example. We decided to split the cluster to obtain two parts that could mimic biomolecular docking. In fact just an ordinary division would not be a simplified model of two proteins. It is clear that using Lennard-Jones clusters is always a rough approximation, but we only need to reproduce the characteristics that we would like to observe in the first phase of minimization and not the general behavior of the two proteins during docking. Clusters big enough to obtain a guest of a reasonable dimension were used. We chose to work in the beginning on a single example in order to tune the parameters and after to use different examples to validate the choice of parameters. Test examples were built with the aim of obtaining different kinds of docking; in fact if we divide a cluster cutting a single square and wedge-shaped part, we obtain a case which in some sense is too easy, as configurations different from the optimal one are widely different from the optimal one and possess a much higher potential energy. In our tests we used different splitting techniques to test whether the algorithm is capable of finding the unique optimal configuration among many similar configurations. In short we transformed the problem into a harder one by allowing the existence of local minima near the global one.

As an example, the optimal 68-atoms cluster was split into two parts with nearly the same dimension; one of the parts has two bulges (or protrusions) corresponding to two cavities in the other part. These two cavities are like two similar but not identical valleys, so that two close matchings are indeed possible, only one of which is the optimal one. The example was labelled by the number of atoms in the host and the guest; this test, as an example, is called 37-31. Another important test is 35-33, which is similar to the first one, but without bulges (the

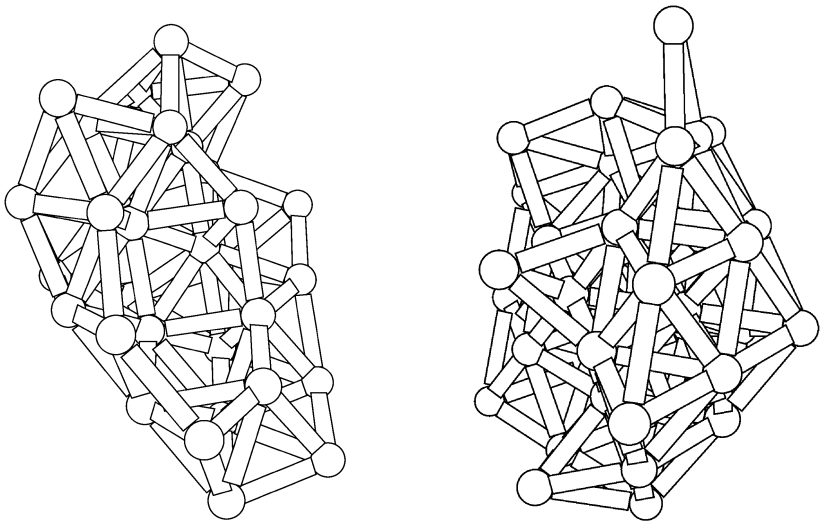

Figure 2. Example 37-31.

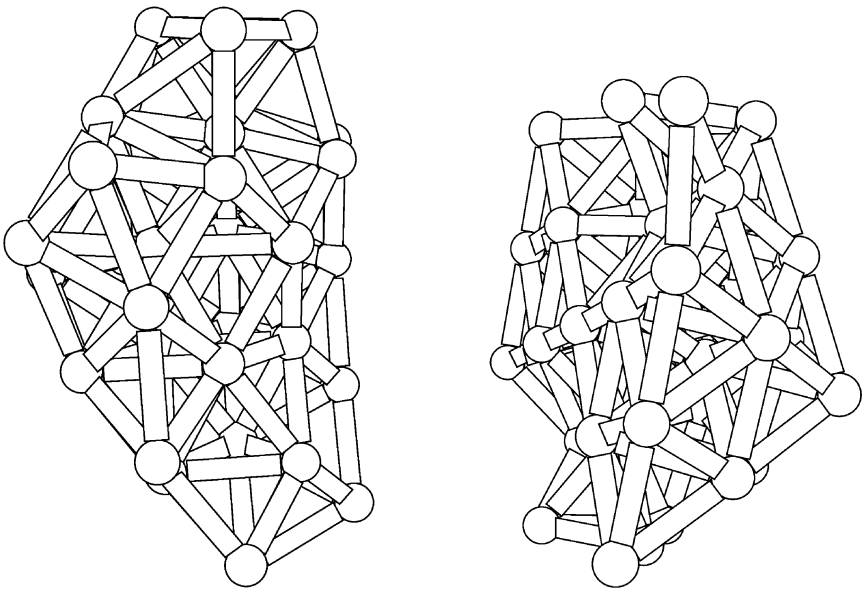

Figure 3. Example 35-33.

two atoms that form these are moved to the guest). In this example many close configurations are possible as the contact surface is quite smooth.

In the first tests (see Table 1), a barrier similar to Lennard-Jones repulsive contribution was used, so we chose high values for q, in the range 10–12; the values chosen for the root term k was 2. In some cases we observed an increase in the number of successes of one order of magnitude, but for some choices of α

Table 1. Example 37-31; $r=6$

q	α	%successes
12	0.005	2.8
12	0.01	3.0
12	0.1	2.8
12	1.0	1.2
10	0.001	2.2
10	0.01	3.1
10	0.1	2.9
7	0.0001	0.0
7	0.0005	1.7
7	0.001	2.4
7	0.003	3.5
7	0.005	3.1
7	0.01	2.6
—	—	0.0

we obtained no success or a smaller number of successes than in the single phase case.

We performed a large number of tests changing the value of the parameter q. After these tests we observed that values for q less than 6, gave the best results even for different choices of α (see Tables 2, 3). It is important to notice that while varying q, it is also necessary to change the value of α: in fact, a small variation in q can modify in a significant way the weight of the repulsive term in (10). So, in order to have a balance between the repulsive and the attractive terms, an adjustment in the value of α is needed.

A large number of systematic tests have been performed on the example 37-31, with constant values of α, changing the parameter q; Table 3 shows the best results obtained. We can notice a strong improvement obtained with q equal to 1 or 2 and α in $0.3-0.7$. With some values of the parameters used on the example 37-31 we performed other tests on different examples, some of which are reported in Tables 4–7. Even if in a few cases we noticed a strong sensitivity on the choice of parameters, in general the results obtained seem to be quite insensible to a large range of parameter choices.

From the tables it is evident that a strong improvement in the number of successes is obtained by the use of two phase methods against a very small increase of the computational effort: even if in two phase algorithms two local optimizations are performed in place of one, it can be observed that the first local optimization does not need to be carried out with high precision, and thus is usually less computationally intensive than a regular local search. Moreover, the second phase, which is performed as a normal local optimization, is started from a usually good initial configuration and it has been observed that it stops in very few iterations even if the required precision is high. Precise computational times are not reported here, but it has been observed that each two-phase local search required significantly less than two single local searches. In our experiments we

Table 2. Example 37-31, $r = 3$

q	α	k	%successes
6	0.005	4	5.4
6	0.005	2	5.4

Table 3. Example 37-31; $r = 6$

q	α	%successes
6	0.3	1.0
6	0.5	1.3
6	0.7	1.1
6	1.0	1.6
5	0.3	0.9
5	0.5	1.8
5	0.7	1.7
5	1.0	1.4
4	0.3	2.3
4	0.5	1.7
4	0.7	1.5
4	1.0	1.5
3	0.3	3.1
3	0.5	2.8
3	0.7	2.4
3	1.0	2.6
2	0.3	6.2
2	0.5	6.9
2	0.7	6.0
2	1.0	2.1
1	0.3	6.9
1	0.5	10.2
1	0.7	9.2
1	1.0	0.0

found that the success rate of two-phase local searches is often at least an order of magnitude higher than that of the standard one: even if a two-phase local search costed twice as much as a single one, the advantages would have remained outstanding.

In the following, a selection of tables of numerical results is presented. We recall that most rows in each table correspond to averages over 1000 random experiments, while for the most difficult cases 10000 runs were performed. Entries in the tables with no value for the α, q and k parameters correspond to results obtained with single phase local optimizations.

Table 4. Example 51-17

q	α	k	r	%successes
1	1.5	2	6.0	5.6
1	0.5	4	3.0	2.0
1	1.2	4	3.0	4.5
1	1.8	4	3.0	1.7
1	1.5	4	3.0	4.5
1	0.5	2	5.0	4.3
1	0.5	3	6.0	0.6
1	0.3	2	6.0	3.5
1	0.5	2	3.0	4.4
1	0.7	2	3.0	2.1
2	0.5	2	3.0	4.6
2	0.3	2	3.0	4.1
–	–	–	6.0	0.06
–	–	–	3.0	0.1

Table 5. Example 30-25.

q	α	k	r	%successes
1	0.5	2	6.0	9.5
1	0.7	2	6.0	8.9
–	–	–	3.0	0.13

Table 6. Example 35-33 $r=4.5$

q	α	%successes
6	0.005	3.3
6	0.01	2.3
6	0.1	4.7
5	0.005	2.4

Table 7. Example 37-31

q	α	k	r	%successes
1	0.5	2	6.0	0.3
2	0.5	3	6.0	6.3
–	–	–	3.0	0.18

6. Two Phase Methods for Docking Biomolecules

After the successful experiments with Lennard-Jones clusters, some preliminary results have been obtained for realistic-size problems of docking. When the molecules to be docked consist of biomolecules, several modifications are needed to the form of the first-phase modified potential function. In fact, in the expression of the potential energy, the term corresponding to Van der Waals contributions is similar to that of Lennard-Jones clusters, as the repulsive and the attractive components in both cases are proportional to r^{-12} and r^{-6} respectively; however, the coefficients are different, and, in particular, they depend on the pair of atoms considered, accounting for their different radii. So using the same barrier term for all pairs of atoms does not provide a good solution; a possibility which has been explored is thus that of replacing the barrier term in the penalty with the Van der Waals contribution. However this simple modification did not lead to a great improvement in the performance of the resulting method. This fact can be explained observing that in docking configurations which are observed in nature the two molecules form a complex which is extremely compact, with deep inclusion of a part of one molecule inside the other, like a key inside the lock. This happens in nature also thanks to the fact that the docking process is not rigid; it is frequently the case that, starting from a known docking configuration and separating the molecules, it is impossible to reconstruct the complex through rigid movements.

We can visualize this situation in a schematic way as in the docking on the left side of Figure 4. As dealing with flexible docking would lead to an enormous increase in complexity, we preferred to follow a different strategy which could enable us to obtain the correct docking with rigid roto-translations. To obtain this effect, a coefficient β greater than one was added into the expression of the modified potential to virtually reduce the Van der Waals radii; the effect of this term is that of reducing the barrier which prevents two atoms to get closer, or, equivalently, to assume that the radii of the atoms are smaller than real; this way molecules can get close enough and even they can be placed into position which,

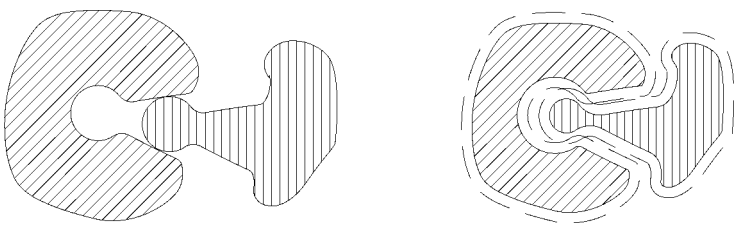

Figure 4. A difficult docking case.

Figure 5. `1auh.pdb`.

with the original radii, were unreachable through rigid movements. This new modified function is represented in (11).

$$V = \sum_{i \in \text{guest}} \sum_{j \in \text{host}} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{(r_{ij}/\beta)^6} + \alpha \sqrt[k]{r_{ij}} \right) \tag{11}$$

The examples used in this phase of algorithm test were taken from the Protein Data Bank (see (Berman et al., 2000)).

In this paper we just comment on results obtained on a theoretical docking model of the HIV virus with an RNA fragment (corresponding to the PDB file `1auh`). Figures 5–7 respectively represent the docked complex, the host (an RNA fragment) and the guest (a protein).

The docking published in the PDB databank was obtained minimizing the Van der Waals and electrostatic contributions along particular directions. In the first tests we chose to constrain the geometric center of the guest molecule to have the same sign as that of the optimal configuration reported in the PDB file. The introduction of such a constraint is justified by the fact that the host molecule in the PDB file is just a fragment of a much larger one: this fact may induce false positives which consist of docking the guest molecule to sites which are very close to where the host was truncated.

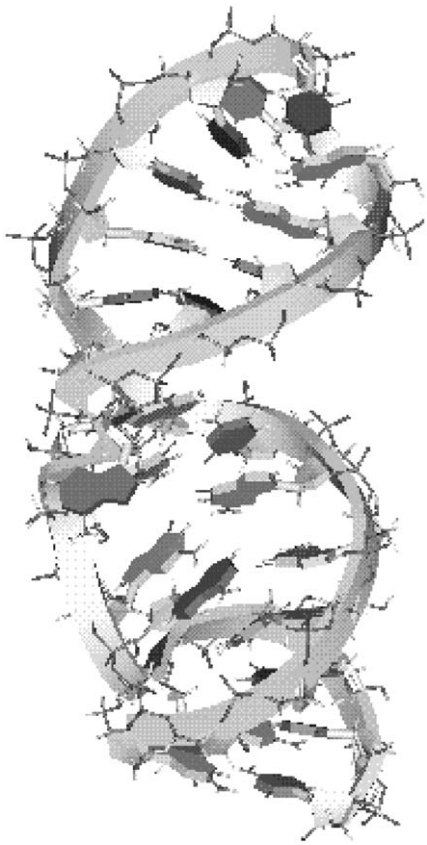

Figure 6. The host molecule.

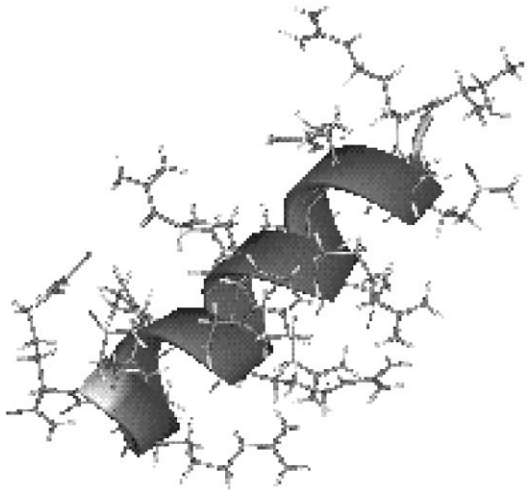

Figure 7. The guest molecule.

Table 8. `1auh` $k=2$

β	α	%successes
–	–	0.5
0.7	1.5	1.5
0.4	1.5	1.0
0.4	4.0	1.5

Table 9. `1auh` $k=2$

β	α	%successes
–	–	0.5
0.7	4.0	16.5
0.7	1.5	16.5
0.7	0.5	1.0
0.7	0.05	0.5
0.8	1.5	17.0

Using single phase optimization we could not find the optimal docking within 10000 local optimizations; so in order to be able to make comparisons with two phase methods, and considering also the fact that 10000 local optimizations required roughly 3 days of cpu time on a SUN Ultra 5 workstation, we decided to increase the probability of finding the optimal docking with the single phase algorithm by initializing the local searches in both single and two-phase methods by fixing the rotation parameters to be equal to the optimal ones, and letting only the translations vary. Of course the local optimization phases are then performed varying all six roto-translation parameters. This way we could observe 5 successes out of 1000 attempts with the single phase algorithm. The results are in Table 8; we notice the improvement obtained using two phase method; it is important to notice that the single phase local optimization fails even when starting from initial configurations which are very near to the optimal one. Other tests were performed starting with the geometric center in a fixed position and applying random rotations in the range $(-20°, 20°)$ around each axis. Considering Table 9 we notice a variation of the number of successes varying parameters, but it is important to notice that the two phase method consistently finds the global optimum with significantly higher probability than the single phase local method.

All the tests were performed on a Sun-Ultra5 workstation. A large part of the computational effort is devoted to function and gradient evaluations: for the `1auh` example 0.2 cpu seconds are needed on average for a single function and gradient evaluation. For every local minimization between one and two hundreds of function and gradient evaluations are needed and each test required 1000 local searches, thus roughly 7 cpu hours are needed for every choice of parameters.

Conclusions

In this paper a modified potential function has been proposed for the solution of biomolecular docking problem. It is important to notice that the original contribution of the paper is the definition of a suitable modified potential function which is used in order to find good starting points for local optimization. The whole procedure outlined in this paper is thus a local optimization method suited for the molecular docking problem. A first set of numerical experiments in both artificial examples and real life molecules have been performed and the results obtained embedding this local search into the most elementary global optimization method, namely Multistart, are very encouraging. For the artificial examples obtained from the splitting of Lennard-Jones clusters the improvement over traditional local searches is one or two orders of magnitude in terms of the number of local searches performed. When applied to realistic biomolecular docking problems, the improvement is even more sensible, as with the modified potential function the optimal configuration could be obtained in a reasonable number of local searches, while this could never be observed with standard local searches, even after several days of attempts.

In conclusion it seems that, despite the fact that molecular conformation problems are indeed extremely hard problems in which a huge number local optima exists, an appropriate choice of a penalty function to be applied during the first phase of any local method might be extremely beneficial. A possible explanation for this behavior is the fact that in many cases molecular conformation problems display a funnel structure; this means that the energy landscape is composed of relatively few and large valleys whose shape is however perturbed by an enormous number of small oscillations. This may explain the failure of standard Multistart methods; however the inclusion of a first phase which smooths out many uninteresting local optima might prove crucial for reaching the bottom of each energy valley.

It is also to be remarked that the results presented in this paper where obtained with extremely cheap hardware like standard personal computers or low-cost workstations. The encouraging results obtained so far provide a stimulus for investigating other research directions. In particular it is planned in the near future to explore the effect of including these two phase procedures into global optimization algorithms like, e.g., simulated annealing and to implement the resulting algorithms also on distributed architectures like clusters of personal computers. Some preliminary results on the use of Monotonic Basin Hopping coupled with the two-phase local search (Addis and Schoen, 2003) confirm the superiority of the two-phase approach.

References

Addis, B. and Schoen, F. (2003), A randomized global optimization method for protein-protein docking. Technical Report DSI 4-2003, Dip. Sistemi e Informatica – Univ. Firenze.

Androulakis, I., Floudas, C., Nayak, N., Ierapetritou, M. and Monos, D. (1997), A predictive method for the evaluation of peptide binding in pocket 1 of HLA-DRB1 via global minimization of energy interactions, *Proteins: Structure, Function and Genetics*, 29.

Apostolakis, J., Pluckthun, A. and Caflisch, A. (1998), Docking small ligands in flexible binding sites, *J. Computational Chemistry*, 19, 21–37.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P. (2000), The protein data bank, *Nucleic Acids Research*, 28, 235–242.

Diller, D.J. and Verlinde, C.L. (1999), A critical evaluation of several global optimization algorithms for the purpose of molecular docking, *J. Computational Chemistry*, 20, 1740–1751.

Doye, J.P.K. (2000), The effect of compression on the global optimization of atomic clusters, *Physical Review E*, 62, 8753–8761.

Fernandez-Recio, J., Totrov, M. and Abagyan, R. (2002), Soft Protein-Protein Docking in Internal Coordinates, *Protein Science*, 11, 280–291.

Floudas, C., Klepeis, J. and Pardalos, P. (1999), Global optimization approaches in protein folding and peptide docking. In: Farach-Colton, M., Roberts, F.S., Vingron, M. and Waterman, M. (eds.), *Mathematical Support for Molecular Biology*, Vol. 47 of *DIMACS Series*. American Mathematical Society, pp. 141–171.

Huang, H.X., Pardalos, P. and Shen, Z. (2002), Equivalent formulations and necessary optimality conditions for the Lenard-Jones problem, *J. Global Optimization*, 22, 97–118.

King, B.L., Vajda, S. and DeLisi, C. (1996), Empirical free energy as a target function in docking and design: Application to HIV-1 protease inhibitor, *Federation of European Biochemical Societies Letters*, 384, 87–91.

Klepeis, J.L., Ierapetritou, M.G. and Floudas, C.A. (1998), Protein folding and peptide docking: A molecular modeling and global optimization approach, *Computers and Chemical Engineering*, 22 Suppl. 1, S3–S10.

Lenhof, H. (1995), An Algorithm for the protein docking problem, Technical report, Max-Planck-Institut fur Informatik.

Liu, D. and Nocedal, J. (1989), On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, B45, 503–528.

Locatelli, M. and Schoen, F. (2002), Fast global optimization of difficult Lennard-Jones clusters, *Computational Optimization and Applications*, 21(1), 55–70.

Locatelli, M. and Schoen, F. (2003), Efficient algorithms for large scale global optimization: Lennard-Jones clusters to appear in *Computational Optimization and Applications*.

MacKerell, A.D.J., Brooks, B., Brooks III, C.L., Nilsson, L., Roux, B., Won, Y. and Karplus, M. (1998), *The Encyclopedia of Computational Chemistry*, Vol. 1, Chapt. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program, John Wiley & Sons, Chichester, pp. 271–277.

Maranas, C.D., Androulakis, I.P. and Floudas, C.A. (1995), A deterministic global optimization approach for protein folding problem. In: Pardalos, P.M., Shalloway, D. and Xue, G. (eds.), *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*. American Mathematical Society, pp. 133–150.

More, J.J. and Wu, Z. (1997), Issues in large scale global molecular optimization. In: Biegler, L.T., Coleman, T.F., Conn, A.R. and Santosa, F.N. (eds.), *Large Scale Optimization with Applications: Part III: Molecular Structure and Optimization*, Springer, New York, pp. 99–121.

Morris, G.M., Goodsel, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A. (1998), Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp Chem.* 19, 1639–1662.

Rosenfeld, R., Zheng, Q., Vajda, S. and DeLisi, C. (1995), Flexible docking of peptides to class I major-histocompatibility-complex receptors, *Genetic Analysis, Biomolecular Engineering*, 12, 1–21.

Shao, C.S., Byrd, R.H., Eskow, E. and Schnabel, R.B. (1997), Global optimization for molecular clusters using a new smoothing approach. In: Biegler, L.T., Coleman, T.F., Conn, A.R. and Santosa, F.N. (eds.), *Large Scale Optimization with Applications: Part III: Molecular Structure and Optimization*. Springer, New York, pp. 163–199.

Totrov, M. and Abagyan, R. (1997), Flexible protein-ligand docking by global energy optimization in internal coordinates. *PROTEINS: Structure, Function, and Genetics*, Suppl. 1, 215–220.

Wang, J., Hou, T., Chen, L. and Xu, X. (1999), Automated docking of peptides and proteins by genetic algorithm, *Chemometrics and Intelligent Laboratory Systems*, 45, 281–286.