EROS-DOCK: Protein-Protein Docking Using Exhaustive Branch-and-Bound Rotational Search

Maria Ruiz Echartea, Isaure Chauvot de Beauchêne, David Ritchie

EROS-DOCK: Protein-Protein Docking Using Exhaustive Branch-and-Bound Rotational Search

Maria Elisa Ruiz Echartea, Isaure Chauvot de Beauchêne, David Ritchie

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Protein Docking by Branch-and-Bound Search

EROS-DOCK: Protein-Protein Docking Using Exhaustive Branch-and-Bound Rotational Search

Maria Elisa Echartea Ruiz,[1] Isaure Chauvot de Beauchêne,[1] and David W. Ritchie [1,*]

[1] University of Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein-protein docking algorithms aim to predict the 3D structure of a binary complex using the structures of the individual proteins. This typically involves searching and scoring in a six-dimensional space. Many docking algorithms use FFT techniques to exhaustively cover the search space and to accelerate the scoring calculation. However, FFT docking results often depend on the initial protein orientations with respect to the Fourier sampling grid. Furthermore, Fourier-transforming a physics-base force field can involve a serious loss of precision.

Results: Here, we present EROS-DOCK, an algorithm to rigidly dock two proteins using a series of exhaustive 3D rotational searches in which non-clashing orientations are scored using the ATTRACT coarse-grained force field model. The rotational space is represented as a quaternion "π-ball", which is systematically sub-divided in a "branch-and-bound" manner, allowing efficient pruning of rotations that will give steric clashes. The algorithm was tested on 173 Docking Benchmark complexes, and results were compared with those of ATTRACT and ZDOCK. According to the CAPRI quality criteria, EROS-DOCK typcially gives more acceptable or medium quality solutions than ATTRACT and ZDOCK.

Availability: The EROS-DOCK program is available for download at http://erosdock.loria.fr.

Contact: dave.ritchie@inria.fr.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein-protein docking algorithms aim to predict how two proteins interact to form a complex. Docking algorithms usually involve two main tasks: (1) sampling the possible relative orientations of the two proteins, and (2) calculating an interaction energy or docking score at each position. Although the protein docking problem has been studied for over 25 years, developing accurate and efficient protein docking algorithms remains a challenging problem due to the size of the search space, the approximate nature of the scoring functions used, and often the inherent flexibility of the protein structures to be docked (for reviews, see e.g. Halperin *et al.* (2002); Bonvin (2006); Ritchie (2008); Huang (2014)).

Under the simplest rigid-body assumption, the "ligand" protein is moved in a six-dimensional (6D) rotational and translational space with respect to a fixed "receptor" protein. Currently, many docking algorithms represent each protein in a three-dimensional (3D) Cartesian grid, and use fast Fourier transform (FFT) techniques to accelerate the calculation of a scoring function based on the degree of overlap or correlation between the two grids in different relative orientations. For example, GRAMM uses a 3D grid to represent the shapes of two proteins (Tovchigrechko and Vakser, 2005). DOT calculates sum of the intermolecular electrostatic energies (Roberts *et al.*, 2013). The Hex (Ritchie and Kemp, 2000) and FRODOCK (Garzon *et al.*, 2009) algorithms apply similar principles using spherical polar representations in order to accelerate the docking search in rotational coordinates. Other algorithms such as ZDOCK (Chen *et al.*, 2003) and PIPER (Kozakov *et al.*, 2006) use 3D FFT translational searches over scoring functions derived from knowledge of known protein-protein interfaces. On the other hand, docking algorithms such as HADDOCK (Dominguez *et al.*, 2003) and ATTRACT (Zacharias, 2003) explicitly move and score atomistic or "coarse-grained" (CG) representations of the ligand and receptor using Monte-Carlo or gradient-based techniques to guide the search towards local energy minima. Thus, such approaches require multiple initial starting orientations in order to cover the 6D search

space. At each trial orientation, the receptor-ligand interaction energy is calculated using pair-wise distances between the relevant atoms or CG "beads", respectively. Thus, the scoring functions in such approaches may be considered as physics-based, rather than knowledge-based. However, since pair-wise atom or CG bead distances must be calculated explicitly in physics-based scoring functions, FFT-based acceleration cannot be used and the computational cost scales as $O(N * M)$ per trial orientation for N ligand and M receptor atoms or beads, respectively. However, in ATTRACT, this cost can be greatly reduced by pre-calculating the receptor potential energies on a 3D grid (de Vries and Zacharias, 2017).

Here, we present a novel docking algorithm which retains the exhaustive nature of FFT-based search algorithms while still using a sensitive physics-based CG scoring function. However, rather than calculating an $O(N * M)$ interaction energy explicitly at every grid point, we use a quaternion "π-ball" to represent the space of all possible 3D Euler angle rotations, and we recursively sub-divide the π-ball in order to cover the rotational space in a systematic way. It has been shown previously that there is a mapping between points in the π-ball space and Euler angle rotations, and that distances calculated between pairs of points in the π-ball are always greater or equal to the angular distances between the corresponding pairs of Euclidean space rotation matrices (Hartley and Kahl, 2009). In other words, coordinate distances in the quaternion π-ball representation provide upper bounds for the corresponding rotational distances in Euler angle rotation space. This important property has been exploited previously to develop efficient branch-and-bound based search algorithms for the problem of finding the optimal registration of two 3D point clouds (Chin *et al.*, 2014; Bustos *et al.*, 2014) which is a common problem in computer vision. In this paper, we apply for the first time a similar branch-and-bound based rotational search to the 6D rigid-body protein docking problem. However, instead of aiming to optimize the 3D registration of two objects represented by point clouds, our aim here is to find the global maximum of all possible pair-wise CG bead docking energies while simultaneously avoiding regions of the search space that lead to forbidden steric clashes. Since rigid body docking is essentially a 6D search problem, we divide the search space into multiple 3D rotational sub-problems, each of which can be treated in parallel using a separate π-ball search tree. The π-ball allows potentially very large regions of a 3D rotational search space to be pruned as soon as it can be established that any rotation within a well-defined sub-region of the search space will cause more than a given number of steric clashes.

In the current implementation of our approach, which we call "EROS-DOCK" (for Exhaustive Rotational Search based Docking), we use CG beads from the standard ATTRACT (attractive plus repulsive) CG force field model (Fiorucci and Zacharias, 2010), where the atoms of each amino acid are represented using from 2 to 4 beads. This reduces considerably the $O(N * M)$ cost of each energy calculation compared to an all-atom representation. We use all attractive pairs of receptor and ligand surface beads to define the initial starting orientations for a full 6D docking search. Additionally, in order to detect steric clashes more efficiently, we pre-calculate "super-beads" from clusters of buried (i.e. non-surface) receptor and ligand beads.

EROS-DOCK has several advantages over existing FFT-based and real-space docking algorithms. In particular, (i) the chosen force field model is calculated exactly, since no potential grids (real or complex) are used; (ii) there are no limits on the sizes of the proteins to be docked, (iii) the protein starting orientations are determined automatically, and (iv) all docking runs are completely deterministic (i.e. two separate runs give identical results), since no random sampling is involved.

We tested EROS-DOCK on the structures of 173 protein-protein complexes taken from the Protein Docking Benchmark (v4) (Hwang *et al.*, 2010) in order to compare our search algorithm with that of ATTRACT, since both EROS-DOCK and ATTRACT use exactly the same physics-based CG scoring function. Additionally, we also compared EROS-DOCK with ZDOCK, as an example of a widely used FFT-based docking algorithm. These tests were performed using the 3D structures of the unbound components of each target complex, and were assessed using the standard CAPRI quality classes (i.e. acceptable, medium, and high quality) (Méndez *et al.*, 2003) with respect to the known crystal structures of the bound complexes.

2 Methods

2.1 Defining Initial Docking Contact Poses

It is reasonable to suppose that the interface in many protein complexes will have several pairs of ligand and receptor beads whose distances are close to the optimal distance for the corresponding bead types. Therefore, we first studied the distribution of ATTRACT CG bead distances in existing protein complexes in the Protein Docking Benchmark (v5) (Vreven *et al.*, 2015). To do this, we used FATCAT (Godzik and Ye, 2004) to superposed each unbound structure onto its complex, and we calculated its intermolecular bead-bead distances. We found that each benchmark complex has at least one pair of surface beads that is within just 0.2 Å of the minimum energy bead distance (here called R_{min}, see Figure 1) of the corresponding ATTRACT interaction energy curve. Because a deviation of only 0.2 Å between a trial orientation and the optimal bead distance may be considered to be negligible, and because it is almost certain that every protein complex will have at least one pair of such beads, it follows that all possible pairs of receptor and ligand attractive surface beads may be used to define a set of initial docking contact poses.

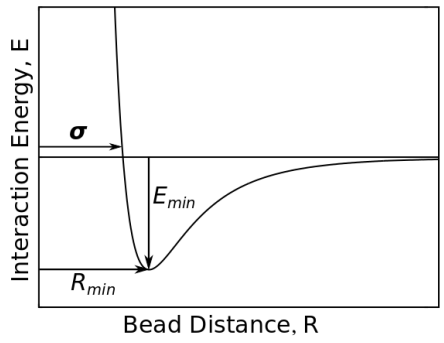

Fig. 1. Plot of an ATTRACT CG bead interaction energy E as a function of the pair-wise bead distance, R. Any distance R less than σ is considered to be a steric clash.

More specifically, for each such pair of receptor and ligand surface beads, (i, j), the receptor bead i is placed at the coordinate origin and the receptor's centre of mass is placed on the negative z axis. Similarly, ligand bead j is placed on the positive z axis at a distance R_{min} from the origin, and the ligand's centre of mass is placed on the positive z axis. This is illustrated in Figure 2. Since the action of making the receptor and ligand centres of mass co-linear with the z-axis is purely for convenience, it can be seen that each placement of one pair of beads absorbs three degrees of freedom, thus leaving a purely 3D rotational search problem. Clearly, when starting a docking search from such an initial configuration, any rotation of the ligand about the coordinate origin will keep ligand bead j in perfect contact with the receptor bead i.

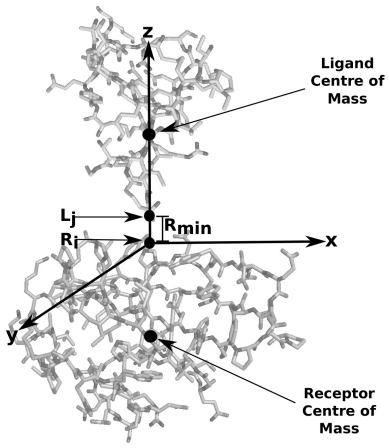

Fig. 2. Illustration of an initial docking pose in which a pair of surface beads R_i and L_j are co-located on the z-axis at their optimal distance R_{min}, thus leaving a purely 3D rotational search of a moving ligand with respect to a fixed receptor.

2.2 3D Rotation Searches using a Quaternion π-Ball

Figure 3 illustrates the notion of a quaternion π-ball. In order to subdivide this 3D rotational space, it is convenient to consider the π-ball as being inscribed in a cube of side 2π, in which any point within the π-ball may be mapped to an Euler rotation defined by the three Euler rotation angles, (α, β, γ). Points within the π-ball may be represented as a unit quaternion, $\underline{Q} = (\cos(\theta/2), \sin(\theta/2)\underline{u})$, where $\underline{u}$ is a unit vector from the coordinate origin, and θ represents the radial distance from the origin. A mapping from the π-ball coordinate system to Euler rotation angles (α, β, γ) in conventional 3D space (using the "z-y-z" convention for Euler angle rotations) may be achieved by setting $\alpha = \theta$ and $\underline{u} = (\sin\beta\cos\gamma, \sin\beta\sin\gamma, \cos\beta)$.

Conceptually, a series of sample rotations is generated by dividing the initial π-ball into 8 cubes, and by then recursively sub-dividing each such cube into smaller cubes until a given angular threshold is reached. From 3D geometry, the distance Δ_s from the centre of cube s to any one of its vertices is given by

$$\Delta_s = \frac{\sqrt{3}}{2} D_s, \tag{1}$$

where D_s is the length of the side of cube s (initially $D_0 = 2\pi$). Thus, Δ_s may be considered as the bounding radius of cube s. At each iteration of an angular search, the centre of the s^{th} cube, $\underline{Q}_s(\theta, \underline{u})$, may be used to define a 3D sample rotation, $\underline{R}_s(\alpha, \beta, \gamma)$, that may be used to rotate the ligand beads into a new trial orientation with respect to the fixed receptor beads. The set of all possible sample rotations from the π-ball cube centres are collected as a set of nodes in a 3D "search tree" data structure. In practice, however, in order to define a specific bounding radius, α, for the leaf nodes, the top level cube size is calculated by successively doubling the leaf node cube size until $D_0 \geq 2\pi$. Any cube centre $\underline{Q}_s(\theta, \underline{u})$ having $|\underline{u}| > \pi$ represents an invalid rotation and is ignored. However, as described below, it is often not necessary to evaluate a docking energy for every node in the search tree.

2.3 Distance Measures in 3D Rotation Spaces

It is intuitively obvious that two similar quaternions or rotation matrices will rotate a given object into similar orientations. However, unlike ordinary Cartesian space, many different angular distance metrics may be defined for rotational spaces. Here, we take as our starting point the angular

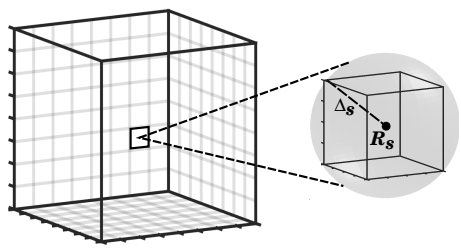

Fig. 3. Representing 3D rotation space as a quaternion π-ball. The π-ball may be subdivided by inscribing it in a cube, and by then subdividing the cube into 8 equal sub-cubes. The subdivision may then be repeated recursively to obtain progressively smaller regions of rotational space. The centre of each sub-cube is used to define a rotational sample, $\underline{R}_s$, for docking, and the radius of the cube, Δ_s, provides an upper bound on the angular distance between $\underline{R}_s$ and any other point within the sub-cube's volume.

distance relations (see Lemmas 1 and 2 of Hartley and Kahl (2009))

$$\theta(\underline{R}_s\underline{v}, \underline{R}_t\underline{v}) \leq d_\theta(\underline{R}_s, \underline{R}_t) \leq d_Q(\underline{Q}_s, \underline{Q}_t), \tag{2}$$

where $d_Q(\underline{Q}_s, \underline{Q}_t)$ represents the Euclidean distance between a pair of quaternion points, $\underline{Q}_s$ and $\underline{Q}_t$, in the π-ball space, $\underline{R}_s$ and $\underline{R}_t$ represent the 3D rotation matrices that correspond to $\underline{Q}_s$ and $\underline{Q}_t$, respectively, and $\theta(\underline{v}, \underline{v}')$ represents the angle between two Cartesian space vectors. The functions $\theta(\underline{v}, \underline{v}')$ and $d_Q(\underline{Q}, \underline{Q}')$ may be calculated as the inverse cosine of the dot product of the corresponding vector components, whereas the angular distance, $d_\theta(\underline{R}, \underline{R}')$, may be calculated by extracting the angular part of the matrix $\underline{M} = \underline{R}^{-1}.\underline{R}'$ in the axis-angle representation of $\underline{M}$ (Hartley and Kahl, 2009).

Since we are mainly concerned with how a sample 3D rotation matrix might move a ligand bead compared to some reference rotation, Equation 2 says that an upper bound for the angular difference in the bead's positions after having applied the two rotations may be obtained from their quaternion representations in the π-ball. Conversely, if the angular distance $d_\theta(\cdot, \cdot)$, between two rotation matrices is greater than the radius Δ_s of a π-ball sub-cube, then it follows that the corresponding quaternion rotation coordinates must fall within different π-ball sub-cubes. This property is used to define a branch-and-bound search in rotational space.

2.4 Branch and Bounds Search using Bead Cone Angles

In order to prune the rotational search efficiently, we begin each 3D rotational docking search by building a list of all possible receptor and ligand attractive surface bead pairs, (a, b), and for each pair we use the corresponding ATTRACT potential energy curve to define a minimum allowed contact distance σ_{ab}, such that a pair-wise bead distance less than σ_{ab} is considered as a steric clash (see Figure 1). Letting $\underline{R}_a$ and $\underline{L}_b$ represent the position vectors of beads a and b, and letting $R_a = |\underline{R}_a|$ and $L_b = |\underline{L}_b|$ denote the corresponding vector lengths, then clearly beads a and b will never give a steric clash under any ligand rotation if $|R_a - L_b| > \sigma_{ab}$. Otherwise, it will be necessary to calculate explicitly whether a particular rotation might cause a steric clash.

While steric clashes are commonly calculated according to a Euclidean distance threshold, here it is more convenient to work with angular distances. More specifically, we first use $\underline{R}_a$ and $\underline{L}_b$ to calculate the rotation $\underline{R}_c^{ab}$ that will place the ligand bead centre $\underline{L}_b$ as closely as possible to the centre of the receptor bead, $\underline{R}_a$. We call $\underline{R}_c^{ab}$ a "clash rotation", because it will cause a steric clash if $|\underline{R}_a - \underline{R}_c^{ab}.\underline{L}_b| < \sigma_{ab}$ (see Figure 4(A)). A list of clash rotations may be calculated just once for each starting pose. Now, if $\underline{R}_c^{ab}$ causes a steric clash between beads a and b then there must exist an infinite number of sample rotations, $\underline{R}_s^{ab}$, which are "near" to $\underline{R}_c^{ab}$ and which will cause the ligand bead to sweep

out a cone in 3D space while remaining in contact with the receptor bead (Figure 4(B)). Hence we use the cosine rule to define a "cone angle", β_{ab}, as

$$\cos\beta_{ab} = (R_a^2 + R_b^2 - \sigma^2)/(2R_aR_b). \quad (3)$$

Then, letting ω represent the angular difference in the ligand position when rotated by a sample rotation $\underline{R}_s$ and its position when rotated by the clash rotation $\underline{R}_c^{ab}$, we have

$$\omega = \theta(\underline{R}_s.\underline{L}_b, \underline{R}_c^{ab}.\underline{L}_b). \quad (4)$$

In this way, we may compare the angles ω and β_{ab} to determine whether the rotation $\underline{R}_s$ causes beads a and b to clash.

More importantly, since Δ_s represents an upper bound on the angular difference between $\underline{R}_s$ and any other point in sampling cube s, then if $\omega > \Delta_s$ we infer that the rotation $\underline{R}_c^{ab}$ must belong outside cube s. In a similar manner, if $\omega > \beta + \Delta_s$, we can infer that *no* rotation within cube s can cause a steric clash between beads a and b (see Figure 5(A)). Conversely, if $\omega > \beta - \Delta_s$, we can infer that *any* rotation from within cube s will cause a steric clash between beads a and b. (Figure 5(B)). Finally, as noted above, if $\omega < \beta$, we infer that the rotation $\underline{R}_s$ causes a steric clash between a and b. However, in the context of a systematic search, sub-dividing cube s could yield further rotational samples that might not cause clashes.

In a similar manner, we note here that some sampling cubes may intersect the boundary of the π-ball. In such cases, if the centre of a cube lies outside the π-ball, then its rotational sample, $\underline{R}_s$, is not meaningful and is discarded. However, the cube remains a candidate for sub-division because the centres of some of its children may still correspond to meaningful rotations.

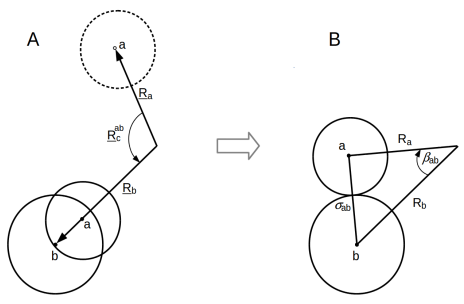

Fig. 4. (A) Illustration of the clash rotation, $\underline{R}_c^{ab}$, between ligand bead a and receptor bead b. $\underline{R}_a$ and $\underline{R}_b$ represent the position vectors of beads a and b, respectively. (B) Illustration of the clash cone angle, β, calculated from the ligand and receptor vector lengths, R_a and R_b, and the contact distance, σ, from the ATTRACT potential for the pair (a, b).

2.5 Coloring the 3D Rotation Search Tree

As indicated above, each node in the rotation search tree is visited recursively for each bead pair in order to color it according to whether it gives a steric clash or not. In order to eliminate sample rotations that lead to steric clashes as early as possible, we first use a simple clustering algorithm to assign any overlapping non-surface beads to a small number of buried "super-beads" (details not shown). These super-beads are then added to the list of potential clash pairs, and the list is sorted in order of decreasing cone angle because bead pairs having large clash cone angles are more likely to allow a node that always clashes to be detected and colored early in the search. Then, in a first pass, each pair of beads from the clash list is used to color the nodes in the tree according to whether a node always gives a steric clash or whether only its central sample rotation

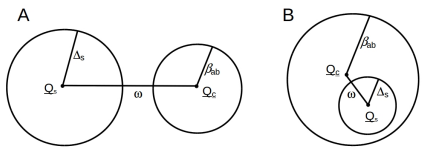

Fig. 5. Schmatic illustration of two important angular relationships in the branch-and-bound search. (A) The case of $\omega > \beta_{ab} + \Delta_s$ in sub-cube s of the π-ball. In this case, the clash rotation, $\underline{R}_c^{ab}$, cannot fall within the rotation volume of sub-cube s, and hence no rotation from within this sub-cube can cause a steric clash between beads a and b. (B) The case of $\omega < \beta_{ab} - \Delta_s$. In this case, the clash rotation, $\underline{R}_c^{ab}$, lies entirely within the rotation volume of sub-cube s, and hence any rotation from within this sub-cube must cause a steric clash between beads a and b.

gives a clash. As soon as a node has been colored as *"Always Clashing"*, it and all of its children may be ignored by subsequent bead pairs, and a counter in the parent node is incremented. Thus, whenever all of the children of a given node are colored as *Always Clashing*, then the parent node is assigned *Always Clashing* as well.

2.6 Calculating Non-Clashing Docking Energies

After the clash status of each π-ball node has been determined, the tree is traversed once more to calculate exact ATTRACT energies for only the non-clashing nodes. The list of non-clashing orientations is then sorted by ATTRACT energy, and the top 100 solutions per π-ball are saved into a global list. Once all of the top 100 solutions per bead pair have been gathered in the global list, the global list is sorted and the top 50,000 orientations are saved as the best solutions found for that target complex.

2.7 Adjustable Parameters

EROS-DOCK has a small number of adjustable parameters. Here, we describe the three most important parameters. Firstly, as mentioned above, the angular search resolution, α, is the most important parameter, since the computational cost of the algorithm scales as $O(1/\alpha^3)$. A large value of α gives faster execution, but if α is too large, then many good solutions will be missed. We currently use a default of $\alpha = 7.5°$, which corresponds to a resolution of $7.5°$ in each of the three Euler rotation angles. Secondly, In order to calculate which beads are surface beads, the accessibility of each bead is initially tested by rolling a probe bead over the surface of each protein. From some early experiments, we determined heuristically that a good value for the rolling bead probe radius is 2.5 Å. Finally, again based on heuristic tests, we determined that up to two bead clashes may be tolerated per pair-wise orientation before calling a steric clash for the corresponding node of the π-ball. Note that clashes are calculated using the cone angle representation (Section 2.4) and that ATTRACT energies are calculated only for rotational orientations that pass the clash count threshold.

3 Results

3.1 Defining the Initial Docking Poses

Our study of protein-protein complexes from the Protein Docking Benchmark (v5) revealed that a large number of protein interfaces contain at least one pair of beads at almost the optimal distance, according to their distance-dependent interaction energy in the ATTRACT CG force field. For example, 90% of the benchmark complexes have at least one pair of receptor-ligand interface beads within 0.2 Å of their optimal separation. More details of the observed percentage of such complexes are shown in Table 1 for pairs of beads in bound complexes and also pairs from the

Table 1. Percentages of complexes from the Docking Benchmark (v5) that contain at least one pair of beads within a distance of δ_R of their optimal separation. The structures of the unbound partners are fitted onto the corresponding structures of the bound complex.

Complex	δ_R/Å	% of Complexes
Bound	< 0.10	93
	< 0.23	100
Unbound	< 0.10	86
	< 0.20	98

corresponding superposed unbound structures. Therefore, a list of initial docking orientations for each target complex was constructed by locating each pair of unbound receptor and ligand surface beads at their optimal separation at the origin (see Figure 1) on the assumption that at least one such pair will resemble a near-native pair in the target complex.

3.2 Sampling the 6D Translation-Rotation Docking Space

For each pair of initial starting poses, a 3D rotational search of the moving ligand with respect to a fixed receptor at the coordinate origin was performed. In order to prune the search before physically moving any ligand beads and calculating their ATTRACT energies, nodes in the 3D search tree were colored according to their steric clash status, as described in Methods. An angular resolution (π-ball node radius) of 7.5° was specified, which gives a tree depth of 7 levels including the root node. For any non-clashing node in the tree, the corresponding node rotation was applied to the ligand and the total interaction energy for that node was calculated as the sum of the ATTRACT pair-wise CG interaction energies. For each starting pose, the best 100 rotations were saved. These orientations were then gathered to form a global list of up to 50,000 6D orientations which was then sorted.

3.3 Efficient π-ball Angular Search

To illustrate the efficiency of the π-ball representation, we may consider as an example the 1OYV target complex. This target gives a total of 18,534 attractive surface bead pairs. Given that a default rotational resolution of α=7.5° leads to a π-ball tree of 42,961 nodes, it follows that the theoretical maximum number of pair-wise orientations for which energies should be computed for this example is 796,239,174. However, EROS-DOCK determined that in fact a total of only 54,874,405 orientations were non-clashing, meaning that 93.11% of the search space was pruned before calculating any energies. Overall, for the 173 benchmark complexes tested here, we calculate that on average 93.76% of the π-ball search space is pruned, and that interaction energies need to be calculated only for the remaining 6.24% of orientations.

It is worth noting that, since the π-ball angular inequalities used here are exact, no solutions are falsely pruned, and therefore the search is guaranteed to be exhaustive for the given angular resolution and clash threshold parameters. It also worth noting that the overall algorithm is very easily parallelized using symmetric multiprocessing techniques on contemporary multi-core processors. More specifically, we assign one π-ball data structure to each available processor core, and starting bead pairs are assigned to processor cores as soon as they become available. The following experiments were performed using 48 cores from two Intel E5-2860 2.4 GHz processors. Each docking calculation required approximately 12 Gb of memory.

Naturally, the execution time varies according to the size of the molecules. For instance, for the easy cases, the target 2OOB with 111 residues and 3,414 starting orientations gave the shortest execution time of 4.33 min. On the other hand, the target 1I9R has 863 residues and 93,442 surface bead pairs, and gave the longest execution time of 184.14 min. Table 2 shows the overall shortest, longest, and average execution times for each target category.

3.4 Comparing EROS-DOCK with ATTRACT and ZDOCK

Because EROS-DOCK uses the ATTRACT coarse-grained force field model, we first compare the results of EROS-DOCK with those of ATTRACT in order to study the effect of our new sampling strategy. However, because ATTRACT performs energy minimizations whereas EROS-DOCK does not, for a fairer comparison we apply energy minimizations using the ATTRACT toolkit to the top 50,000 solutions of each target docked by EROS-DOCK. These results are subsequently called EROS-MIN.

We also compare results with ZDOCK version 3.0.2 (Pierce *et al.*, 2011) in order to examine the difference between the use of exhaustive CG sampling and regular FFT sampling using a pairwise statistical interaction potential. The results presented for ZDOCK were obtained using default parameters and random starting orientations for both receptor and ligand. Since EROS-DOCK performs dense rotational sampling, we also ran ZDOCK using its dense (6°) sampling option. However, the results were less favorable than using ZDOCK's default 15° sampling mode. Therefore, we show here only ZDOCK results using 15° sampling.

For the ATTRACT runs, the ligand starting positions were generated by the standard ATTRACT search procedure which gave a set of points evenly distributed over the receptor surface (the actual number depends on the size of the receptor), and at a distance from the receptor surface that depends on the ligand's radius of gyration. The ligand was placed on each starting point, and 228 ligand rotations were applied to generate approximately equally distributed ligand orientations. For each receptor starting position and ligand orientation, 1,000 minimization steps were applied using the ATTRACT force-field with grid acceleration, a final sum of pairwise atom-atom energies was calculated, the structures were ranked by ATTRACT energy, and redundant structures (RMSD < 0.2 Å) were discarded.

Figure 6 summarises the number of successfully docked targets obtained by ZDOCK, ATTRACT, EROS-DOCK, and EROS-MIN for the 173 benchmark complexes, as a function of the CAPRI docking quality criteria. For example, Figure 6(A) shows the distribution of targets having at least one acceptable, medium, or high quality docking solution within the ranks 1, 10, 100, and 1,000 for the 173 benchmark complexes. At each rank threshold the number of successful docking cases is represented by a bar. In the same way, Figures 6 (B), (C), and (D) show the results according to the "easy", "medium", and "difficult" classifications, as determined by the Benchmark authors. More detailed results are provided as Supplementary Material.

It should be noted that in Figure 6, the total number of acceptable solutions includes the number of medium and high quality solutions. Hence, for example, Figure 6 (A) shows that EROS-MIN found acceptable solutions ranked within the top 100 solutions for 156 out of 173 target complexes, of which 88 are also classed as medium quality solutions and 21 as high quality solutions. Because different proteins will often have different numbers of surface beads, EROS-DOCK generally calculates a different number of initial docking poses for each target complex. However, we did not find any relationship between the quality of the docking solutions and the number of starting poses (details not shown).

In general, Figure 6 (A) shows that EROS-MIN produces more acceptable solutions than the other algorithms, except at the top 10 where the results are comparable with those of EROS-DOCK and ZDOCK. This indicates that several of the basic EROS-DOCK solutions are close

enough to a near-native local energy minimum to benefit from a subsequent minimization step. Regarding medium quality solutions, Figure 6 (A) shows that the performance of EROS-DOCK, EROS-MIN, and ZDOCK is generally comparable at each level, except that ZDOCK finds noticeably more acceptable solutions in the top 10 while ATTRACT generally finds fewer acceptable or better solutions. For high quality solutions, EROS-MIN performs better than the other methods.

Since EROS-MIN and ATTRACT use the same force field and scoring function, any difference in their performance must be due to their different sampling strategies. ATTRACT uses a heuristic sampling scheme, while EROS uses an exhaustive search. Therefore, we believe that ATTRACT is prone to miss some energy minima when the energy landscape fluctuates rapidly, but it will find the local minimum in each energy basin it explores. On the other hand, EROS-DOCK is less likely to miss basins, but will not find the minimum in each basin.

To investigate this further, we compared the energy of the top-ranked solutions found by ATTRACT and by EROS-DOCK before minimization. We found that EROS-DOCK finds solutions with lower energy than the lowest-energy solution of ATTRACT in 163 out of 173 cases (in 142/173 cases when considering only differences above 1 Kcal/Mol). These lower-energy solutions found by EROS-DOCK correspond to basins not explored by ATTRACT. This confirms that the better performance of EROS-MIN over ATTRACT is due to a more exhaustive initial sampling by EROS-DOCK, allowing to find more local minima after minimization of the low-energy basins found by EROS-DOCK.

When considering the results by target difficulty, Figure 6 (B) shows that the best solutions produced by each algorithm for the easy targets are mainly of acceptable and medium quality, and the number of successfully docked targets is comparable, especially among EROS-MIN, EROS-DOCK, and ZDOCK. On the other hand, EROS-DOCK (i.e. without minimisation) finds fewer high quality solutions than the other algorithms. For medium difficulty targets, Figure 6 (C) shows that the best solutions obtained by each algorithm are mainly of acceptable quality, and the number of successfully docked targets is again comparable. A similar profile of results is seen for the difficult targets (Figure 6 (D)). However, the total number of targets and number of high quality solutions obtained by any method for the medium and difficult target groups are generally quite small, making it difficult to make meaningful comparisons between the different algorithms. Nonetheless, it interesting to note that ATTRACT is the only algorithm to obtain high quality solutions in the top 1,000 for some difficult targets (Figure 6 (D)).

As mentioned above, energy minimizing the basic EROS-DOCK solutions (EROS-MIN) increased the number of targets with high quality models for the easy targets, and it increases the number of targets with acceptable or medium quality solutions at all the difficulty classifications. This demonstrates the utility of using EROS-DOCK as an exhaustive initial docking search engine to propose high quality trial orientations which could be refined using flexible docking procedures or short molecular dynamics simulations.

Table 2. Summary of EROS-DOCK execution times, grouped by benchmark category.

	Target	No. Residues	No. Starting Pairs	Execution Time / min
Easy Targets				
Shortest Time	2OOB	111	3,414	4.33
Longest Time	1I9R	863	93,442	1272.72
Average Time		453	29,323	184.14
Medium Targets				
Shortest Time	1SYX	191	6,971	14.78
Longest Time	1BGX	1,230	195,965	3512.2
Average Time		512	40,951	367.39
Difficult Targets				
Shortest Time	1PXV	282	10,070	20.74
Longest Time	1DE4	1,641	245,456	4700.2
Average Time		573	55,073	631.43

4 Conclusion

We have developed *an exhaustive real space CG* docking algorithm called EROS-DOCK. A novel feature of our approach is the use of a quaternion π-ball representation of 3D rotational space which allows the notion of branch-and-bound search to be applied for the first time to the protein docking problem. We have demonstrated that our branch-and-bound search using the ATTRACT CG force field model typically gives more acceptable or better solutions, especially when a final energy minimization step is applied, when compared to the well-known and highly optimised ATTRACT and ZDOCK docking programs.

While the current implementation of EROS-DOCK is slower than ATTRACT, we believe there is scope to optimize the EROS-DOCK code and search parameters. Furthermore, we expect that our approach will be particularly suitable for docking very large protein structures that will have many more local energy minima than the examples studied here. We also expect that our angular search algorithm will be particularly useful when some knowledge of the interface residues is available (as in "data-driven" docking) and when multiple interfaces must be considered in multi-component docking problems. We are currently investigating both such possibilities.

Acknowledgements

Funding

M. E. Echartea Ruiz is funded by a CORDI-S (Inria) doctoral contract.

Conflict of Interest

None declared.

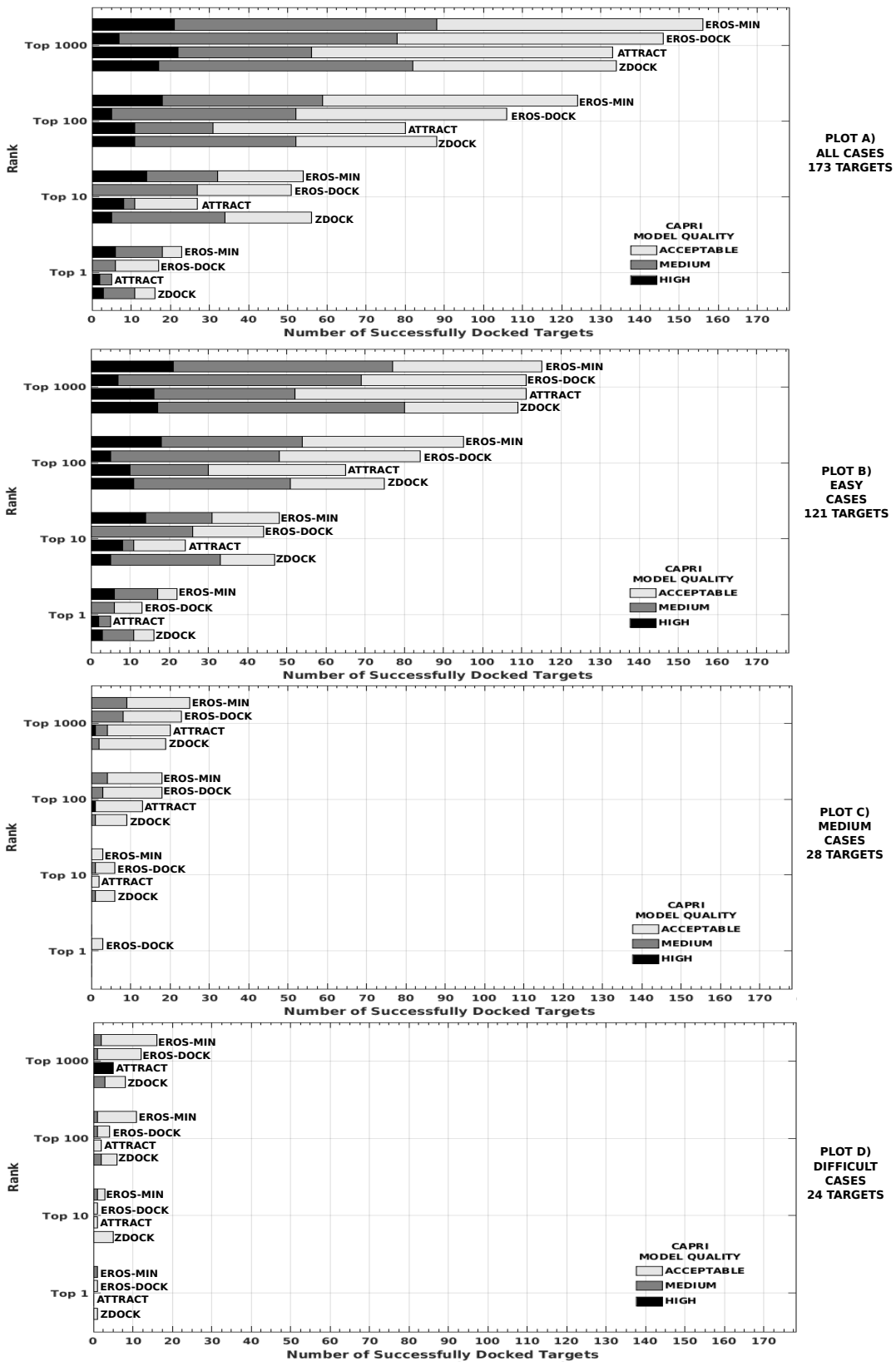

Fig. 6. Results obtained by EROS-DOCK, ATTRACT and ZDOCK for 173 unbound target complexes from the Protein Docking Benchmark (v4). The plots show the number of complexes docked with acceptable, medium, and high quality according to the CAPRI quality criteria.

References

Bonvin, A. (2006). Flexible protein-protein docking. *Current Opinion in Structural Biology*, **16**, 194–200.

Bustos, A. P., Chin, T.-J., Eriksson, A., Li, H., and Suter, D. (2014). Fast rotation search with stereographic projections for 3d registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(11), 2227–2240.

Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. *Proteins: Structure, Function, Genetics*, **52**, 80–87.

Chin, T.-J., Bustos, A. P., Brown, M. S., and Suter, D. (2014). Fast rotation search for real-time interactive point cloud registration. In *Proceedings of the 18th Meeting of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D'14, pages 55–62, New York, NY, USA. ACM.

de Vries, S. J. and Zacharias, M. (2017). Fast and accurate grid representations for atom-based docking with partner flexibility. *Journal of Computational Chemistry*, **38**(17), 1538–1546.

Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, **125**, 1731–1737.

Fiorucci, S. and Zacharias, M. (2010). Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins: Structure, Function, Bioinformatics*, **78**(15), 3131–3139.

Garzon, J. I., Lopéz-Blanco, J. R., Pons, C., Kovacs, J., Abagyan, R., Fernandez-Recio, J., and Chacon, P. (2009). FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics*, **25**(19), 2544–2551.

Godzik, A. and Ye, Y. (2004). FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Research*, **32**(suppl 2), W582–W585.

Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, Genetics*, **47**, 409–443.

Hartley, R. I. and Kahl, F. (2009). Global optimization through rotation space search. *International Journal of Computer Vision*, **82**, 64–79.

Huang, S.-Y. (2014). Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug Discovery Today*, **19**(8), 1081–1096.

Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). Protein-protein docking benchmark version 4.0. *Proteins: Structure, Function, Bioinformatics*, **78**(15), 3111–3114.

Kozakov, D., Brenke, R., Comeau, S. R., and Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, Bioinformatics*, **65**, 392–406.

Méndez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins: Structure, Function, Genetics*, **52**, 51–67.

Pierce, B. G., Hourai, Y., and Weng, Z. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS One*, **6**(9), e24657.

Ritchie, D. W. (2008). Recent progress and future directions in protein-protein docking. *Current protein and Peptide Science*, **9**(1), 1–15.

Ritchie, D. W. and Kemp, G. J. L. (2000). Protein docking using spherical polar Fourier correlations. *Proteins: Structure, Function, Genetics*, **39**(2), 178–194.

Roberts, V. A., Thompson, E. E., Pique, M. E., Perez, M. S., and Ten Eyck, L. F. (2013). Dot2: Macromolecular docking with improved biophysical models. *Journal of Computational Chemistry*, **34**(20), 1743–1758.

Tovchigrechko, A. and Vakser, I. A. (2005). GRAMM-X public web server for protein-protein docking. *Nucleic Acids Research*, **34**, W310–W314.

Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., Chaleil, R., Jiménez-García, B., Bates, P. A., Fernandez-Recio, J., Bonvin, A. M., and Weng, Z. (2015). Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *Journal of Molecular Biology*, **427**(19), 3031–3041.

Zacharias, M. (2003). Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science*, **12**(6), 1271–1282.