Short Papers

Evaluation of Geometric Complementarity between Molecular Surfaces Using Compactly Supported Radial Basis Functions

Luis A. Diago and Ernesto Moreno

Abstract—One of the challenges faced by all molecular docking algorithms is that of being able to discriminate between correct results and false positives obtained in the simulations. The scoring or energetic function is the one that must fulfill this task. Several scoring functions have been developed and new methodologies are still under development. In this paper, we have employed the Compactly Supported Radial Basis Functions (CSRBF) to create analytical representations of molecular surfaces, which are then included as key components of a new scoring function for molecular docking. The method proposed here achieves a better ranking of the solutions produced by the program DOCK, as compared with the ranking done by its native contact scoring function. Our new analytical scoring function based on CSRBF can be easily included in different available docking programs as a reliable and quick filter in large-scale docking simulations.

Index Terms—Molecular docking, compactly supported radial basis functions.

1 INTRODUCTION

As shape plays a crucial role in biomolecular recognition and function, the techniques for description of molecular shape and shape complementarity [1] are of prime importance in relevant applications such as drug design and protein engineering [2], [3], [4], [5]. The complementarity between two molecular surfaces can be analyzed from the existence of a similarity between the functions that describe the geometry of the surfaces and a correspondence between their physicochemical characteristics. Recent reviews [6], [7] of docking and scoring methods report that over 60 docking programs and more than 30 scoring functions have been disclosed to date. Several studies comparing different approaches have been published in the past decade (see [7] for details). As the development of universal, fast, and highly accurate docking/scoring methods is still a long way to go [7], this paper proposes a new way for the evaluation of the geometric complementarity that is useful to increase the speed and accuracy of the current molecular docking algorithms employed in the screening of large databases of small-molecule structures.

In earlier studies, Max and Getzoff [8], Duncan and Olson [9], and Leicester et al. [10], [11] established that protein surface shapes can be represented quite accurately using spherical harmonics and that many useful surface properties can be derived from this representation [9]. However, solving the harmonic expansion coefficients involves a considerable amount of computation. In a more recent study [12], real spherical harmonic expansion coefficients were used as 3D shape descriptors for protein binding pockets and small ligand molecules, and shape similarity was computed as the L2-distance in the coefficient space. Although this approach can be employed to compare shapes, it loses information, since the original shape cannot be reconstructed from its descriptors, especially for nonstar-like surfaces of proteins and ligands at high resolution [12]. In another recent report, Cai et al. [13] described an efficient virtual screening algorithm using spherical harmonic molecular surfaces. The authors in [15] extended a previous method described by Ritchie and Kemp [14] using a shape difference function to speed up the computations. However, as mentioned by Cai et al. [13], the implementation of a global docking search with six degrees of freedom remains difficult because the surface integration for calculating the expansion coefficients is computationally expensive. Other programs (for example, PatchDOCK [15] and FRED [16]) dock the small molecules using shape complementarity or interaction site matching algorithms to optimize the contact surface between the ligand and the protein, thereby allowing extremely fast rigid docking procedures. However, neither of these approaches is well suited for rescoring the solutions predicted by other programs.

In this study, we present our implementation of the Compactly Supported Radial Basis Functions (CSRBF) [17], which are commonly used to interpolate scattered data for 3D reconstruction, to create analytical representations of molecular surfaces. To our knowledge, this mathematical formalism has not been used before with this purpose. The main goal of our work was to develop a new scoring function, based on CSRBFs, to evaluate the geometric complementarity between two molecular surfaces, and implement this function within a molecular docking algorithm. The performance of our new scoring function was extensively evaluated by comparing it with the contact type of scoring [1] implemented in the program DOCK [18], [19], [20], in docking simulations that were run for hundreds of protein-ligand complexes.

1.1 Compactly Supported Radial Basis Functions

The theory of radial basis functions for interpolating scattered data and reconstruct 3D objects was first introduced in [21] and [22]. Later on, Wendland proposed the use of CSRBF [17] to reduce the computational cost of solving the system of linear algebraic equations that arises in this type of problem. A CSRBF representation of an n-dimensional object is an implicit function of the form:

\[
s(\mathbf{x}) = p(\mathbf{x}) + \sum_{i=1}^{N} \lambda_i \phi(||\mathbf{x} - \mathbf{x}_i||),
\]

where \(p\) is a polynomial of degree one, the \(\lambda_i\) are the CSRBF coefficients, and \(\phi\) is a real-valued function called CSRBF, centered at each point \(\mathbf{x}_i\). The solution for the set of \(\lambda_i\) and the coefficients of \(p(\mathbf{x})\) that will satisfy the set of interpolating conditions \(s(\mathbf{x}_i) = f_i\) for a given set of \(N\) points \(\mathbf{x}_i\), is obtained solving the linear system:

\[
\begin{pmatrix}
\Phi & P \\
P^T & 0
\end{pmatrix}
\begin{pmatrix}
\lambda \\
c
\end{pmatrix}
=
\begin{pmatrix}
f \\
0
\end{pmatrix},
\]

where \(\Phi_{ij} = \phi_{ij}(||\mathbf{x}_i - \mathbf{x}_j||)\) and \(P_{ik} = p_k(\mathbf{x}_i)\), with \(i, j = 1, \ldots, N\) and \(k = 1, \ldots, l\). Note that we use the notation \(\mathbf{x} = (x, y, z)\) for points \(\mathbf{x} \in \mathbb{R}^3\). For interpolating in \(\mathbb{R}^3\), \(l = 4\) and the monomial basis \(\{1, x, y, z\}\) is used for polynomials of degree one with \(c = (e_1, \ldots, e_6)\) as the coefficients that give \(p(\mathbf{x})\) in terms of this basis. As the ill-conditioning of the system in (2) increases with the number of CSRBF centers \(\mathbf{x}_i\), in practice, the coordinates of \(\mathbf{x}_i\) are normalized between 0 and 1 by dividing each original coordinate \(\mathbf{x'} = (x'_1, y'_1, z'_1)\) by a maximum value \(M = \max(x'_{\text{max}} - x'_{\text{min}}, y'_{\text{max}} - y'_{\text{min}}, z'_{\text{max}} - z'_{\text{min}})\) (i.e., \(x_i = \frac{x_i - x_{\text{min}}}{M}\)). Subindexes \(\text{min}\) and \(\text{max}\) indicate minimum and maximum values for each coordinate.

After normalization, the support size for the function \(\phi_{ij}\) or, in
other words, the range of influence of a point \( x_i \) is given by \( r_0 \) 
\((0 < r_0 \leq 1)\). Here, we use the CSRBF function proposed by Wendland to represent 3D surfaces: 
\[ \phi_i = (1 - r)^3 (4r + 1) \]
where \( \phi(r) = \phi_0(r) = \phi(r/r_0) \).

### 1.2 Evaluating the Complementarity between Two Analytical Surfaces

Given a set of points \( P_1 \) (e.g., the molecular surface of a protein binding site), the approach of polynomial-interpolated measurement (PIM) [23] can be used to determine which of the \( L \) sets of points \( P_{1}, P_{2}, \ldots, P_{L} \), that describe different surfaces (e.g., different orientations of one ligand in the binding site of a protein), is closer (best similarity) to the set \( P_1 \). The \( L \) sets have \( N_l \) points, whereas \( P_1 \) has \( N \) points. Here, the distance between two sets of points \( P_1 \) and \( P_2 \) is defined by the following expression (3) [23].

The lower the distance, the higher the similarity of the two surfaces.

\[
\text{Dist}(P_1, P_2) = \|s_1 - s_2\|_{P_1/P_2} = \frac{1}{N_1} \sum_{i=1}^{N_1} \|s_1(x) - s_2(x)\|^2 + \frac{1}{N_2} \sum_{i=1}^{N_2} \|s_1(x) - s_2(x)\|^2.
\]

### 1.3 The Program DOCK and Its Contact Scoring Function (FC)

A molecular docking program explores ways in which two molecules, such as a drug and an enzyme, might fit together. One of the two molecules is generally a protein (the “receptor”), whereas its counterpart is usually a small molecule (the “ligand”). DOCK, a docking algorithm first reported in 1982 [18] and since then in constant development [19], [20], implements a geometric matching algorithm to superimpose ligand atoms onto points that are the centers of spheres constructed on the binding site surface. The program DOCK includes a contact scoring function (FC) [1], which is a simple distance-based weighted sum of the number of contacts between ligand and protein atoms, given by the expression:

\[
FC = \sum_{i=1}^{N_l} \sum_{j=1}^{N_r} F_{ij},
\]

where \( N_l \) and \( N_r \) are the number of ligand and receptor atoms, respectively,

\[
F_{ij} = \begin{cases} 
1.0 & \text{if } 2.3 \text{ Å} \leq d_{ij} \leq 3.5 \text{ Å} \\
e^{-d_{ij}^2} & \text{if } 3.5 \text{ Å} < d_{ij} \leq 5.0 \text{ Å} \\
0.0 & \text{if } 5.0 \text{ Å} < d_{ij}
\end{cases}
\]

and \( d_{ij} \) is the distance between ligand atom \( i \) and receptor atom \( j \). If two atoms are too close (\( d_{ij} < 2.3 \text{ Å} \)), the evaluated complex is penalized with a very large value in the function FC. If an optimization of FC is performed, then a few number of bumps should be allowed, since the rigid-body minimization [24] can recover ligand poses from such clashes. Atomic van der Waals (VDW) overlaps are penalized by checking if two atoms approach closer than a fraction (contact clash overlap) of the sum of their VDW radii. In this way, orientations are evaluated, ranked, and optimized using the function FC and the minimizer [24].

In order to evaluate a large number of ligand poses in a reasonable amount of time, approximate scoring functions must be used. Once again, numerous solutions to this problem have been proposed, including a variety of empirical and physics-based terms [6], [7]. Besides the contact function, DOCK also uses an energy scoring function based on the AMBER molecular mechanics force field [18], [19], [20]. However, the simplest, contact-based scoring function is usually used for the screening of large databases of small-molecule structures. As has been reported [20], the orienting method samples near-crystal ligand orientations well, but none of the current scoring functions can discriminate well among the top-ranked orientations.

Here, we propose a method to improve the ranking of the docking solutions based on the analytical representation of molecular surfaces. The DOCK’s FC was selected as reference for comparison with our proposed scoring function, because both functions evaluate the molecular interactions without taking into account the chemical nature of the atomic contacts, and also because our function was devised for a fast evaluation of shape complementarity as an initial screen, e.g., in the search for drug candidates, as is the case for the FC function.

### 2 Proposed Method

#### 2.1 Representing Molecular Surfaces with CSRBF

The CSRBF surface representation needs a set of points as a support. Here, we used the Connolly dot surface [25] as source of points, both for the protein receptors and the ligands. In this algorithm, a sphere is rolled over the van der Waals atomic spheres to generate a smooth outer surface contour represented by a uniformly distributed cloud of dots. We used a sphere radius of 1.4 Å and dot densities of 1 and 4 Å\(^{-2}\) for the protein and the ligand, respectively. Additional, off-surface points were included together with the surface dots to define the normal directions. Then, the analytical representations are obtained by solving the corresponding system of linear equations in (2) [17], [22]. The off-surface points were created only at the inside or outside of every surface point \( x_i \), as suggested in [22], in order to avoid the null solution for the interpolating conditions \( s(x_i) = 0 \), which is obtained if only the points on the molecular surface are used. Thus, new interpolating conditions \( s(x_i) = d_i \neq 0 \) were added for the off-surface points. The off-surface function values of \( d_i \) are assumed to be positive outside the object and negative in the inside.

As an example of this representation, Fig. 1 shows the polygonal mesh of the CSRBF obtained for one of the protein-ligand complexes included in this study.

#### 2.2 Defining a New Analytical Scoring Function Based on CSRBF

We took advantage of the analytical representation based on CSRBF to assess the geometric complementarity of two molecular surfaces. Using the polynomial-interpolated measurement (PIM) approach [23], we developed a new scoring function for docking simulations, which is based on the calculation of the geometric complementarity between the surfaces described by \( s_1(x) \) and \( s_2(x) \), according to (1).

The new scoring function was defined as:

\[
FS = \sum_{i=1}^{N_l} D_{i, s_1},
\]

where

\[
D_{i, s_1} = \begin{cases} 
-1.0 & \text{if } 0 < d_{i, s_1} \leq 0.5 \text{ Å} \\
-0.5 & \text{if } 0.5 < d_{i, s_1} \leq 1.0 \text{ Å} \\
-0.1 & \text{if } 1.0 < d_{i, s_1} \leq 1.5 \text{ Å} \\
0.0 & \text{if } 1.5 < d_{i, s_1}
\end{cases}
\]

and \( d_{i, s_1} \) is the distance from the point \( x_i \) to the surface described by \( s_1(x) \). The scoring function FS sums the contributions of every point \( i \) of the binding site surface \( s_1(x) \) according to the distance \( d_{i, s_1} \) from the surface of the ligand \( s_1(x) \). It is important to remark that in comparing the molecular surfaces described by the CSRBF, we should take special care of the problem associated with the changes of scale. The representation by CSRBF is invariant to rotations and translations, but it is not invariant to changes in scale. Due to the normalization carried out to obtain the analytical representations \( s_1(x) \) and \( s_2(x) \), both the ligand
and the binding site change their scales according to their size. This problem was solved simply by reversing the transformation performed in the normalization, changing the coefficients $c = (c_1, \ldots, c_4)^T$ computed from the normalized coordinates $x_i$ by $c'_i = (c'_{1}, \ldots, c'_{4})^T$, computed as follows: $c'_1 = c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4$, $c'_2 = \alpha c_2$, $c'_3 = \alpha c_3$, and $c'_4 = \alpha c_4$, with $\beta_2 = x'_2 / M$, $\beta_3 = y'_3 / M$, $\beta_4 = z'_4 / M$, and $\alpha = 1 / M$. Note that the set of $\lambda_i$ does not change. After reverse normalization, a first-order approximation was used to calculate the distance ($d_{s_1}$) in (5). The main advantage of this function based on analytical representations of the surfaces is that the evaluation can be performed quickly and easily.

3 RESULTS AND DISCUSSION
3.1 Running DOCKing Simulations for Hundreds of Protein-Ligand Complexes

In order to evaluate the performance of our new analytical scoring function (FS), as compared with the contact scoring function of the program DOCK, we used two data sets of 315 and 488 protein-ligand complexes, respectively, extracted from a database previously compiled by us [26]. In these selected complexes, the ligands have 20-40 heavy atoms and more than 70 percent of their molecular surface embedded in the protein binding site.

Since our scoring function FS is not incorporated into the program DOCK, we used the strategy of selecting from each docking run a relatively large number of possible solutions, as ranked by DOCK's native scoring function FC, then rescore these ligand orientations with our function FS, and compare the results of the two different rankings.

First, a docking simulation was run for each of the 315 complexes of the first data set, keeping rigid the ligand and using the chemical filter denoted as R2 in [26]. The binding site points to be matched with ligand atoms were constructed using our own method, as described in [27]. In order to ensure an exhaustive exploration of the conformational space, we set the maximum

![Fig. 1. Visualization of the CSRBF representation for a protein binding site (PDB code 1SLY). A software was developed in Visual C++.NET using the VTK class library to visualize the CSRBF surfaces. The displayed surfaces were constructed using values of $r_0 = 0.04$ (panel a) and $r_0 = 0.13$ (panel b), the time for solving the system was 0.09 and 6 seconds, respectively. The same support radius $r_0 = 0.13$, which offered a good compromise between the quality of the representation and the calculation speed, was used for docking simulations with hundreds of protein-ligand complexes. (a) $r_0 = 0.04$ and (b) $r_0 = 0.13$.](image)

![Fig. 2. Comparison between the FC and FS according to their "fitness" and "ranking capability," for a test case (PDB structure 1GSU). Calculations were done for the best $N_0 = 10$ solutions obtained from the docking run. In the four panels, an alphabetical order was assigned to the solutions to indicate an increasing order of the RMSD value. Ordinal numbers indicate the order provided by FC or FS to each solution. (a) $D_{FC} = 0.564$, (b) $D_{FS} = 0.234$, (c) $R_{FC} = 0.216$, and (d) $R_{FS} = 0.867$.](image)
number of ligand orientations the program can generate to a high value (10^5). No optimization of the orientations was performed in these runs for the first data set.

For each complex, we kept the best 10 ligand orientations (solutions) given by the program. These orientations were then classified into “good” (correct) and “bad” (incorrect) solutions, depending on the root-mean-square deviation (RMSD) from the experimental geometry. “Good” solutions are those with an RMSD value less than 2 Å. For each of the 10 best ligand orientations, the rotation-translation matrix calculated by the DOCK program was stored to make it possible to reproduce each orientation later on, starting from the initial ligand geometry.

In order to rescore the obtained solutions using our scoring function FS, we computed the analytical representations of the interacting protein and ligand surfaces. The stored rotation-translation matrixes were then used for positioning the analytical surfaces of the ligand in correspondence with each of the orientations produced in the docking run. For each generated orientation, the value of FS was computed using (5).

Similar calculations were performed for the second data set (488 complexes), but with two differences in the docking procedure. First, to demonstrate that our results do not depend on the number of solutions (N_0) used for comparison, we collected the best N_0 = 50 solutions given by FC, which were then rescored with our FS function. Second, since the function FC

---

The figure illustrates the comparison of the performance of the FS (bars in black) and the FC (bars in gray) according to the “fitness” value D_F (a), (c), and (e) and “ranking capability” R_F (b), (d), and (f), for the best N_0 = 10 solutions obtained for the set of 315 PDB complexes (a) and (b), and for the best N_0 = 50 solutions obtained for the set of 488 complexes (c), (d), (e), and (f). (e) and (f) correspond to docking calculations using the minimizer.
is very sensitive to small differences in ligand pose, and therefore, the use of the minimizer [24] might significantly affect the results, we ran two different types of docking calculations: with and without the minimizer.

3.2 Comparing the Performance of FS and FC

The scoring functions FC and FS were compared according to their “fitness” and “ranking capability” of the solutions: two parameters defined in this work are explained below.

3.2.1 Fitness of a Scoring Function

This parameter was defined as a way to measure the correlation between the scores given by the scoring function and the deviations (RMSDs) of the corresponding ligand geometries from the experimental or reference orientation.

First, the values of FC, FS, and RMSD for each of the best orientations were normalized using the expression $f(x) = \frac{f(x) - f_{\min}}{f_{\max} - f_{\min}}$. The values of $f_{\min}$ and $f_{\max}$ correspond to the minimum and maximum values of each function for the best $N_0$ orientations. Then, the fitness $D_F$ for a particular scoring function “F” was defined as the mean quadratic difference between the normalized function and the normalized RMSD:

$$D_F = \sqrt{\frac{1}{N_0} \sum_{i=1}^{N_0} (f(i) - \text{rmsd}(i))^2},$$

where, the normalized functions have been denoted with small letters.

To illustrate our procedure, we show in Fig. 2 the results from these calculations for one protein-ligand complex (PDB code: 1GSU). Comparison of the graphs in Figs. 2a and 2b shows that the analytical function (FS) performs much better in ranking the solutions in correlation with their RMSD values, which is reflected by the calculated fitness ($D_{FS} = 0.234$ vs. $D_{FC} = 0.564$).

3.2.2 Ranking Capability of a Scoring Function

The “ranking capability” ($R_F$) evaluates the ability of a scoring function to place “good” solutions at the top of the list of the $N_0$ best solutions:

$$R_F = \frac{1}{N_0} \sum_{i=1}^{N_0} \delta_i,$$

where $\delta_i = 1$, if the orientation at position $i$ of the ranking list (according to the scoring function F) is a good orientation (RMSD < 2 Å), otherwise $\delta_i = 0$. Thus, the scoring function whose “good” solutions are located at the top will have a higher value of $R_F$.

Figs. 2c and 2d show that, for the selected complex, the FS produces a better grouping of the good solutions toward the top of the list, as reflected by the calculated ranking capabilities $R_{FS}$ and $R_{FC}$.

3.3 Comparing the Performance of FS versus FC for Hundreds of Protein-Ligand Complexes

The same procedure described in the previous section was conducted for the selected sets of 315 and 488 PDB complexes. The histograms in Figs. 3a and 3b show the performance (according to $D_F$ and $R_F$) of FC and FS for the best $N_0 = 10$ solutions obtained for the 315 complexes of the first data set.

Fig. 3a shows the distribution of the fitness values ($D_F$). The histogram for FS is shifted toward zero as compared to the histogram for FC ($D_{FS} = 0.37 \pm 0.15$ versus $D_{FC} = 0.52 \pm 0.11$), demonstrating the better performance of the analytical function.

The histogram in Fig. 3b shows the distribution of the ranking capabilities ($R_F$) for the two scoring functions. In this case, the number of complexes used to compute the values of $R_F$ was smaller than in Fig. 3a, because we considered only those complexes where the ordering given by FC and FS was different. As average, the value of $R_{FS}$ was higher than the value of $R_{FC}$ (0.66 and 0.52, respectively). Furthermore, the number of complexes whose $R_F$ value is close to 1 (0.8 < $R_F$ < 1) is higher when using our CSRBF-based scoring function FS.

Similar results were obtained for the distribution of the fitness values ($D_F$) when using the data set of 488 complexes, as shown in Fig. 3c. For this set, the resoring was performed for the best $N_0 = 10$ solutions of the contact function. The histograms obtained for the distribution of the ranking capability (Fig. 3d) show a similar behavior for both scoring functions.

3.4 Comparing the Performance of FS versus Optimized FC

When the program DOCK was run for the set of 488 complexes using the score minimizer, the average number of good solutions per complex, among the $N_0 = 50$ top-ranked solutions, increased from 22 to 33. The average running time also increased from 10 to 705 seconds per complex. After rescoring with the FS function, the distributions of the fitness values $D_F$ were very similar to those obtained without using the minimizer (Fig. 3e), showing a better performance of the FS function. The distributions of the $R_F$ values remained as before, without significant differences between the two scoring functions. We should take into account that the value of $R_F$ is determined mostly by the single top-ranked good solution, with only a minor influence of the positions of the rest of the good solutions.

A more detailed comparison of the ranking performance of the contact scoring function FC and our proposed analytical scoring function FS, for the best $N_0 = 50$ solutions collected for the set of 488 complexes. FCO corresponds to the optimized solutions obtained by applying the minimizer in the docking runs, while FSO corresponds to the rescored optimized solutions, using the scoring function FS.
4 Conclusion

In this paper, we have used the description of 3D objects by means of radial basis functions with compact support to evaluate the geometric complementarity between molecular surfaces. Employing Connolly dot surfaces as sources of points for the analytical representations allowed the use of a single value of the support radius for all complexes. Based on these CSRBF representations, we have defined a new scoring function for docking simulations, which is very simple and fast to calculate. Calculations run for hundreds of protein-ligand complexes showed that our scoring function performs better than the widely used contact scoring function of the program DOCK in evaluating the geometric complementarity between molecular surfaces, and therefore, provides a better ranking of the docking solutions. Implementing this new function in the context of DOCK or other docking programs may improve the results in large-scale screenings aimed at identifying new drug candidates.

Acknowledgments

This work was developed when Luis A. Diago worked in the Department of Bioengineering, Faculty of Electrical Engineering, Havana Institute of Technology, Cuba.

References


For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.