

# A Geometric Filter for Unbound Protein-Protein Docking\*

Peter Lloyd, Guillermo Sapiro, and David Baker<sup>†</sup>

**Keywords:** Protein-protein docking, geometric filtering, protein holes, statistical filtering, inter quartile range.

## 1 Introduction

The protein-protein docking problem [1, 3, 4], that is, the task of assembling two separate protein components into their biologically-relevant complex structure, is important for several reasons. First, it is of relevance to cellular biology, where function is accomplished by proteins interacting with themselves and with other molecular components. Second, it presents one of the fundamental tests of the understanding of molecular biophysics, requiring a sophisticated knowledge of molecular motions and free energy calculations. Finally, an important post-genomic goal is the determination of the structural details of interactions between all pairs of proteins that bind, and computational tools offer an inexpensive means to prepare large-scale studies. Most currently available algorithms employ a docking procedure based either on the chemical and energetic properties of the protein molecules or on geometric and topological properties. These techniques alone have deficiencies in computing a successful dock, and for a docking algorithm to function correctly, a combination of these two types of techniques must be utilized. In this work we extend the contribution in [2], which uses a very detailed representation of side-chain energetic and conformational freedom. Our contribution is in the form of a geometric and statistical based filter, needed to improve the docking results and to reduce the computational cost demanded by such accurate representation. This filter consists of a surface feature identification algorithm constructed on a geometric grid-based technique, along with a simple numeric comparison of proteins based on statistical measurements. The filter has been able to eliminate over 70% of the low energy decoys or false candidates produced by the accurate energetic computations, and shows great promise for the continued development of protein docking algorithms based on the combination of biochemical with geometric criteria.

## 2 The algorithm

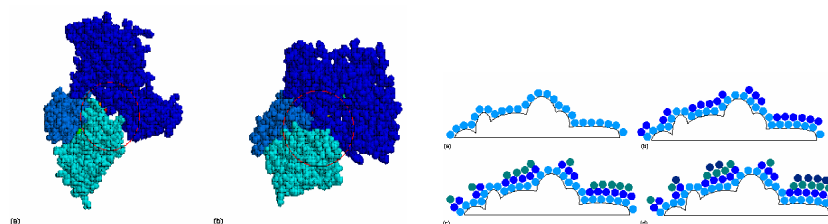
Our approach is based on an algorithm first proposed by Gray *et al.* [2], which uses a free energy surface to represent each protein. Initially, a decoy (or candidate dock) is created by placing it in a random orientation and translating one partner in the protein structure along the line of protein centers. Afterwards, a rigid-body Monte Carlo search is employed by translating and rotating one of the proteins around the surface of the other protein. This search is considered to be a low-resolution one due to a reduced representation of the amino acid residues which are based just on side-chain centroid positions. At this point in the algorithm, a score depicting the correctness of the protein can be calculated based on a low-resolution analysis of the residue-residue interactions between the protein molecules. The algorithm then adds explicit side-chains to the protein backbones by using a backbone-dependent rotamer packing algorithm. After the side-chains have been applied to the protein, the rigid-body displacement is optimized, and the local minimum of the energy function is found. A score based on the full-atom analysis is developed. This entire process is repeated enough times to create approximately  $10^5$  candidates/decoys per target. Once all of the decoys have been created the best-scoring decoys are clustered hierarchically based on a pair-wise root-mean-square distance (rmsd) calculation. The clusters that contain the most proteins are selected for a final docking prediction, and are ranked according to the sizes of the clusters. State-of-the art results are reported with this technique.

Our procedure to hybridize a geometric approach to protein-protein docking with a chemical and physical approach focuses primarily on a filter that is capable of distinguishing which candidates/decoys from those created by the technique just described should be kept for further docking refinement and which decoys should be discarded due to inherent geometric flaws in the molecule itself (see Figure 1, left). The filtration algorithm that we use is divided into two primary parts, a geometric surface reconstruction and shape feature extraction and analysis component, and an efficient statistical measurement algorithm. The geometric surface reconstruction algorithm builds on the idea that incorrect protein structures during protein-protein docking will be likely to contain voids at the binding interface (not detectable by previous global complementarity measures). It is found that many small holes on the surface of the protein are indicative of an incorrect structure. A tightly joined binding interface supports the notion of (both local and global) shape complementarity between docking proteins, and often a large degree of complementarity between proteins is

---

\*Work supported by DARPA, NSF, and NIH.

<sup>†</sup>Department of Electrical and Computer Engineering, University of Minnesota; and Department of Biochemistry, University of Washington lloyd0053@umn.edu, guille@ece.umn.edu, dabaker@u.washington.edu



**Figure 1: Left:** This figure illustrates the fact that while a docking algorithm based almost purely on protein energetics predicts a successful dock, the prediction may be inaccurate from a geometric standpoint, even if good complementarity scores are achieved. (a) A decoy/candidate from the target set IEO8, which has a very high score based on [2] scoring scheme, but does not dock very tightly due to inherent geometric flaws in the decoy's design. (b) Another decoy from target set IEO8 representing a tighter fitting dock. Pure complementarity measures will also fail in detecting such small geometric problems. **Right:** Two-dimensional cross-section of a protein molecule. The black contour represents an abstraction of the atoms composing the surface of the protein, the light blue different shades represent water molecules that have been placed to create an imaginary surface for the protein, and the other water layers are each represented by a circle with a different shade of blue. (a) The pseudo-surface, (b) one additional layer of water molecules, (c) two additional layers, (d) three additional layers. (This is a color figure.)

thought of as a positive factor to a docking score. Our geometric algorithm detects and measures the size of concave regions, holes, and handles on the surface of the protein, and uses statistics on them to filter out non-native decoys.

Our geometric surface reconstruction and surface feature measurement algorithm begins by constructing a pseudo-surface of the protein molecule. This is accomplished by mapping an imaginary layer of water molecules onto the actual surface of the protein as in Figure 1, right. This imaginary water surface is mapped by measuring the distance between the centers of the water molecules and the atoms composing the protein. Each atom in the protein is represented abstractly as a sphere with a fixed van der Waals radius, which essentially creates a space filling diagram or van der Waals surface of the molecule. Each water molecule is also represented abstractly as a sphere but with a fixed different size radius. A minimum sized three-dimensional matrix is placed over the protein molecule, and the size of each grid box in the matrix is the minimum distance that the centers of any two water molecules can be apart from one another without the molecules colliding. The radii and placement of the molecules is carefully constrained based on basic bio-chemically oriented concepts.

After the pseudo-surface has been plotted, we can start to identify the various critical features for docking which exist on the surface of the protein molecule. This is accomplished by placing additional layers of virtual water molecules in an evenly spaced formation on top of the initial layer which composes the pseudo-surface (see Figure 1). The number of layers used is discussed in our extended report. The placement of these water molecules obeys the same distance rules and constraints as the placement of the water molecules composing the surface. The only difference between the initial layer and any additional layers is that the additional layers must be placed so that they fill in the various surface features of interest instead of reflecting the shape and contour of the protein's surface. The process of filling in the surface features starts by analyzing the empty space around each water molecule of the pseudo-surface. Each of these molecules will contain at most fourteen available degrees of freedom; six degrees for movements along each of the axis, and eight degrees for the diagonal movements. A degree of freedom is considered available if that position surrounding the water molecule does not cause a collision with a pre-existing water molecule or with an atom in the protein molecule itself.

Once the layering of water molecules has been completed we must identify and measure the individual critical surface features which exist. This is accomplished through a novel hierarchical distance-based clustering technique together with a peaks detection process. The clusters volumes are then measured, statistically studied, and used for the geometric filter.

To this geometric filter we added a simple measure of compactness, the inter quartile range, and used the combination of both to filter out decoys. The exact use of the holes volumes and the inter quartile range in the filter was based on statistical analysis of native and decoy protein-protein docks. We tested the algorithm on data produced by [2] showing that 70% of the low energy decoys are filtered out on average by this simple criteria. Additional details on the geometric computation and results for each individual protein are presented at the meeting.

## References

- [1] C. Camacho and S. Vajda, "Protein-protein association kinetics and protein docking," *Current Opinion in Structural Biology* **12**, pp. 36-40, 2002.
- [2] J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker, "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations," *J. Mol. Biol.* **331**, pp. 281-299, 2003.
- [3] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov, "Principles of Docking: An overview of search algorithms and a guide to scoring functions," *PROTEINS: Structure, Function, and Genetics* **47**, pp. 409-443, 2002.
- [4] G. R. Smith and M. Sternberg, "Prediction of protein-protein interactions by docking methods," *Current Opinion in Structural Biology* **12**, pp. 28-35, 2002.