# Accurate Protein Docking by Shape Complementarity Alone

Sergei Bespamyatnikh[1], Vicky Choi[2] Herbert Edelsbrunner[3] and Johannes Rudolph[4]

Co-corresponding authors:

Herbert Edelsbrunner, Department of Computer Science, Duke University, Durham, North Carolina 27708, tel.: 919-660-6545, fax: 919-660-6519, email: edels@cs.duke.edu.

Johannes Rudolph, Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710, tel.: 919-668-6188, fax: 919-613-8642, email: rudolph@biochem.duke.edu.

Running title
Accurate Protein Docking by Shape Complementarity Alone

Number of pages: 25
Number of tables: 2
Number of figures: 5

Submitted as pdf-file.

No supplementary material.

[1] Department of Computer Science, University of Texas, Dallas, Texas.

[2] Department of Computer Science, Duke University, Durham, North Carolina.

[3] Department of Computer Science, Duke University, Durham, and Raindrop Geomagic, Research Triangle Park, North Carolina.

[4] Department of Biochemistry, Duke University Medical Center, Durham, North Carolina.

# Abstract

Elucidating the molecular details of protein-protein interactions is essential to understanding cellular processes. Given the recent increase in protein structural information, mostly of monomeric proteins, we now have the data necessary to address protein-protein interactions by computational approaches. Previous attempts at in silico protein docking generate a multitude of answers that have similarly high scores for both correctly and incorrectly docked proteins. We have developed the first algorithm that successfully predicts the re-docking of known protein-protein complexes without any false positives. Because our algorithm is based on complementarity alone, this implies that shape matching suffices for recognizing correct docking configurations. The essential features needed to achieve accurate protein docking are a fine covering of the space of rigid motions and a score function that counts the number of atoms at close distance. Our results provide a proof-of-principle for the development of faster docking methods based on shape complementarity alone that incorporate protein flexibility.

**Keywords.** Protein-protein interactions, protein docking, shape complementarity, rigid motions, quaternions.

# Abbreviations and Notation Used

| | |
|---|---|
| RMSD | root mean square distance |
| PDB | protein data bank |

| | |
|---|---|
| $A; a_i, \alpha_i$ | protein; center, radius of $i$-th sphere |
| $B; b_j, \beta_j$ | protein; center, radius of $j$-th sphere |
| $\delta, 2\delta$ | packing radius, step-size in $\mathbb{R}^3$ |
| $\varepsilon, 2\varepsilon$ | packing radius, step-size in $\mathbb{S}^3$ |
| $\lambda, 2\lambda$ | (half) the distance threshold |
| $\chi$ | number of collisions or bumps |

| | |
|---|---|
| $\mathbb{R}^3, \mathbb{S}^3$ | three-dim. Euclidean space, sphere |
| $\mu, \tau, \varrho$ | rigid motion, translation, rotation |
| $\mathbf{p}, \mathbf{q}$ | unit quaternions |
| $\mathbf{x}, \mathbf{y}$ | pure imaginary quaternions |
| $x, y$ | points in $\mathbb{R}^3$ |

Highly organized assemblies of proteins control most cellular events. In the past few years, incredible progress has been made in cataloging all the potential players in these assemblies through multiple genome sequencing projects. Comparative genomics and proteomics facilitated by these projects are answering many questions regarding the function and importance of specific genes. In the next few years, a three-dimensional view will be added to this catalog of sequences by the determination of large numbers of high-resolution protein structures (Montelione and Anderson 1999). These structure determinations will be supplemented by increasingly accurate homology models of proteins of unknown structure. Thus, we will soon have a relatively complete picture catalog of most of the critical components of the cell along with a map showing their network connections. This catalog and map will provide the groundwork to frame the true question of interest to the biochemist. Namely, how do the individual proteins form complexes and dynamically function together to generate the cell circuitry and its detailed time-dependent responses to various stimuli? This is a three-dimensional puzzle of enormous scale and, at the same time, an intricate parallel algorithm.

**Towards protein assemblies.** Protein-protein interactions appear to be one key to understanding complex assemblies and their elucidation has been a major research goal for many years. High-resolution crystal structure determination has provided one of the most powerful methods, revealing the molecular details of many protein-protein interfaces. Protein-protein interface maps have defined crucial residues and interactions, mostly for stable isolable complexes (Jones and Thornton 1996). Yet, there exist limitations to the amount of information we can glean through crystallography. First, protein structure determinations are usually performed using individual proteins and the elucidation of the structures of complexes becomes progressively more difficult with the number of components due to protein expression and purification problems, the difficulty of obtaining crystals of diffraction quality, and the solution of these large structures. Second, structure determinations of complexes reveal only the static three-dimensional view generated by co-crystallization and do not address the dynamic processes involved in the formation of these complexes or the artifacts caused by the crystallization conditions (e.g. partially truncated proteins, complexes lacking one or more subunits, lattice-forming residues). Thus, at this point and for the near future, we have a wealth of structural information about individual proteins, yet a severe limit on the structural tools with which to study their interactions.

In addition to crystallography, mutagenesis has been utilized to study protein-protein interactions. Multiple mutagenesis schemes directed by crystal structures or by random searches and subsequent biochemical analyses have revealed not only the sites of interaction, but also the associated energetics. Alanine scanning mutagenesis has been one of the most successful methods, allowing for an unbiased search and elucidation of molecular interfaces for binding domains (Wells 1991). The major limitation of mutagenesis studies is the amount of labor required to perform a thorough study of a given system. Not only must two proteins of interest be cloned, purified, and quantitatively assayed for their binding interaction, but this must be performed for many, if not hundreds, of individual mutants of each of the two proteins. Although more high-throughput methods have recently attempted to address some of these limitations (Weiss et al. 2000), exhaustive mutagenic analysis of protein-protein interactions will continue to be a time-consuming process.

Electron microscopy, fluorescence spectroscopy, nuclear magnetic resonance spectroscopy, and immunofluorescence serve as additional biochemical methods used to study protein-protein interactions. While all of these techniques provide important answers in their unique way, in the long run they will be insufficient to understanding the multitude of interactions we must decipher to understand cellular regulation. Furthermore, it is speculated that many of the most important protein-protein interactions are transient and/or weak, features that do not lend themselves well to any form of currently available analysis. New methods must therefore be developed to predict and then automate the analysis of potential protein-protein interactions using the tools of modern mathematics and the power of computer processing.

**Prior work on docking prediction.** It is believed that one of the essential components of any protein-protein interaction is the shape recognition that must occur between the two surfaces: two proteins that form a complex must have good surface complementarity. Physically, the van der Waals spheres cannot overlap in space and it has been noticed that protein-protein interfaces generally do not contain large empty or water-filled holes (Hubbard and Argos 1994). Theoretically, the protein-protein docking problem could be addressed by shape matching algorithms (Connolly 1986), although this has not been successfully demonstrated until this paper. Over the past years, a number of research groups have developed protein-protein docking algorithms that incorporate shape complementarity in a variety of different ways. (For comprehensive reviews, see (Elcock et al. 2001; Halperin et al. 2002). To reduce the complexity to six degrees of freedom, the two proteins are treated as rigid bodies. To overcome the typical dilemma that there are still too many possible rotations and translations that must be explored, a protocol that detects potential binding patches or pronounced geometric shapes is implemented. Finally, the scoring function evaluates geometrical complementarity with or without explicitly calculating free energies of binding.

It appears that the major limitation of all the previously described techniques is that they generate multiple possible docking positions with no way of distinguishing the correct one. Specifically, the docking software generates the correct solution as well as other high scoring solutions that do not correspond to a properly docked complex. Distinguishing true docking from the sometimes large number of false positives can be a daunting problem, especially if it needs to be addressed experimentally in the laboratory. Incorrect docking solutions have been reduced by a number of research groups in a variety of ways, most often by including binding determinants other than simple surface complementarity. For example, hydrogen bonding, electrostatic energy, solvation or hydrophobicity terms have been included. Alternatively, conformational adjustments to binding have been taken into account using molecular flexibility and limited energy minimization protocols

Despite the incorporation of these additional parameters, some of which are computationally expensive, these previously described docking procedures still generate a significant number of false positives. In addition, they appear to work best with protein-protein complexes that have deep binding sites (e.g. protease-inhibitor complexes), reminiscent of small molecule ligands interacting in the binding pockets of their proteins. These types of interactions are relatively well addressed by software such as DOCK (Ewing et al. 2001) and are not necessarily applicable to the larger and shallower surface interactions often found in protein-protein complexes. As demonstrated by the poor success rate in a recent blind docking contest (CAPRI), protein-protein docking methods still have a long way to go before they become useful to the biological community (Janin et al. 2003).

Believing that local shape complementarity is of greatest importance, we have developed an accurate unbiased exhaustive search algorithm based solely on geometric shape matching that successfully generates a re-docked pair starting from a random relative positioning of the two proteins. The advantage of our method is that it apparently does not yield any false positives.

## Results

We implemented an algorithm to predict docking configurations, fine-tuned it using a model system, and applied it to a suite of twenty-five re-docking problems.

**Overall strategy.** Examination of the interaction surface between two proteins reveals a highly complex interface characterized by extensive shape complementarity. In our approach to the protein-protein docking problem, we reason that it should be possible to take two protein surfaces and use this shape complementarity to match these proteins by exhaustively searching a large number of potential docking orientations. To preserve sufficient surface detail, we use a scoring function that counts the number of non-overlapping pairs of spheres on each protein at most at distance $2\lambda = 1.5\text{Å}$ from each other. Additionally, we allow a limited number of collisions between spheres, or bumps, $\chi \leq 5$. One of the reasons that we use this simple scoring function is the difficulty of searching in the space of rigid motions. This space has six dimensions, three for rotation and three for translation that we sample independently. Each combination of translation and rotation is a rigid motion for which we compute the score function. Dense sampling of the six-dimensional space requires a huge number of samples. For example, a step-size of $0.4\text{Å}$ over $16\text{Å}$ in length gives $64,000$ translations. Doing this for $12,036$ different rotations amounts to roughly $7.7 \times 10^8$ sample points in the space of rigid motions for which a scoring function that matches the atoms in one protein to another needs to be calculated, a total of almost $10^{15}$ calculations for even a small protein complex. As pointed out by Connolly (Connolly 1986), this is a computationally expensive approach, even now, fifteen years later with significantly faster processors.

For our test system, we have relied primarily on the barnase/barstar complex (the PDB file is 1BRS, consisting of chain A with 864 atoms and chain D with 691 atoms (Buckle et al. 1994). Barnase is a bacterial ribonuclease and barstar is its intracellular protein inhibitor. There exists a high degree of shape and charge complementarity between the two proteins and the 1BRS complex has sufficiently few atoms to allow extensive testing of software and input parameters. Because of the extent of biochemical information available on this complex, it is ideally suited for future exploration of the importance of specific residues in computational docking. We use global and local RMSD to evaluate the success of our docking experiments.

**Sampling density.** In our algorithm, we need to determine the translational and rotational sampling density required to successfully re-dock a protein-protein complex. Too sparse a sampling would be expected to miss the correct solution whereas too dense a sampling would be prohibitively expensive computationally. The translation density is varied by changing the step-size $2\delta$ of the translational grid. Because the distance between any two grid points is at least $2\delta$, we can draw non-overlapping spheres of radius $\delta$ around the points. Similarly, the rotation density is varied by changing the step-size $2\varepsilon$ or, equivalently, the maximum radius for which the spheres around the sampled points do not overlap. We use a density screening experiment to determine feasible values of $\delta$ and $\varepsilon$. For efficiency reasons, this screening is done using a local search in translational and rotational space. Specifically, we choose a small neighborhood of twenty different small random perturbations of the initial position that is big enough to contain some small but not too small number of rigid motions. We evaluated the highest scoring solution after varying $\delta$ and $\varepsilon$ in terms of both average score function and average RMSD over the twenty random starting perturbations. As seen in Figure 1, there exists the expected trade-off between the quality of prediction and the density of sampling. The best score of 319 was obtained for $\delta = 0.1$ and $\varepsilon = 0.0376$, with an average RMSD of $1.47$Å. We note that the score function appears to fall off gently with progressively sparser sampling in either translational or rotational space. In contrast, the RMSD drops off more dramatically at $\varepsilon = 0.0676$, indicating that many of these docking solutions do not correspond to the correct solution. Visual inspection of individual docking attempts confirms that the top-scoring solution, and all those with RMSD $\leq 3.5$Å, have re-docked the 1BRS complex in the correct location; see Figure 2. Thus, not only does the highest scoring solution generate the correct answer, but all other top scoring solutions in each individual docking experiment, also generate the correct answer. This demonstrates the robustness of our protein-protein docking algorithm based on shape alone.

**Initial position and false positives.** Any valid protein docking algorithm must be insensitive to the initial orientation of the proteins to be docked and should not generate high scoring solutions that are incorrect (false positives). We investigated the outcome of re-docking barnase (chain A) onto barstar (chain D) of the 1BRS complex following fifty different random rotations of barstar using a full exhaustive search. Based on the initial density screen, we chose the parameters $\delta = 0.4$ and $\varepsilon = 0.0476$ for this experiment and each docking took about one hour on a cluster of eighty computers (600 MHz to 2 GHz). As seen in Figure 3 (A), the range of top scores varies from 195 to 329 in a fairly evenly distributed manner. Of the fifty random starting rotations, forty-six gave the correct docking solution as indicated by RMSDs below $3\text{Å}$. The most noteworthy observation we have made is indicated by the empty space in the upper right hand plot of RMSD vs. score in Figure 3 (B). Although there does not exist a direct linear correlation between score function and RMSD, we significantly do not obtain any false positives. That is, there is no high scoring docked complex that yields a high RMSD. This is in contrast to previously described docking programs, where the correct solution is among the top ten, hundred, or more scores (Norel et al. 1995; Lenhof 1996; Ackerman et al. 1998; Fernández-Recio et al. 2002). We do not achieve such favorable results using sparser sampling of rotational or translational space. Thus, our protocol consistently generates believable results without the need to weed out false positives, an important characteristic if data from docking programs are to be trusted and utilized by the experimentalist.

**Testing other data sets.** We have also tested our program with a diverse data set of twenty-five other protein complexes, some of which have been used in previous docking studies (Norel et al. 1994; Fischer et al. 1995; Norel et al. 1995; Lenhof 1996; Ackerman et al. 1998). To improve our chances of finding the correct solution, we increase the translational sampling density from $0.4\text{Å}$ to $0.2\text{Å}$. This increases the computation time to between eleven and thirty-eight hours per complex on the cluster of eighty microprocessors. Additionally, we evaluate the docking solutions not only by RMSD, but also by RMSD*, a value that reflects the deviation of all atoms near the region of the docking interaction. Essentially, correct docking solutions that only have a slightly wrong angle can lead to high and potentially misleading overall RMSD values because of the contribution of residues far from the docking site. Therefore, it has been previously suggested that an alternative calculation such as RMSD* is more reflective for evaluating docking experiments (Fernádez-Recio et al. 2002). Our results show that the top scoring solution with $\chi \leq 5$ always generates the correct docking solution with an RMSD (RMSD*) below $3.7\text{Å}$ ($1.9\text{Å}$), and is not accompanied by a single false positive (Table 1).

A comparison of the top score to the score of the naive complex, the docked structure as read from the database, shows no apparent correlation. The top score can exceed the naive score (e.g. 4SGB and 1BUH), and this is always, except for 3SGB, accompanied by an increase in the number of allowed bumps. We believe this occurs because our re-docking, in the absence of explicit water molecules, brings the two proteins closer together than in the naive complex, where water forms a non-scoring part of the interface. At other times, the top docking score lies significantly below the naive score (e.g. 1A22 and 3HLA). The bump parameter $\chi$ is mostly ($60\%$) at its maximum allowed value of 5 and use of $\chi > 5$ begins to generate false positives (data not shown). Importantly, we have successfully re-docked all twenty-five complexes with no changes to the algorithm and its parameters, and there was no highest scoring solution that yielded an incorrectly docked conformation. This is in contrast to the results observed for other docking software to date. As one examines the second and third highest docking scores, we detect docking solutions that are incorrect for eighteen of the twenty-five complexes tested. For most of these false positives, the docking score is significantly below the top scoring solution (e.g. 1F47, 1TGS, and 1MCT) whereas 1JAT, 3HLA, and 1STF have scores that are not much below the top score. The occurrence of these false positives does not appear to correlate with relative buried surface area but may be indicative of the number of highly buried water molecules (data not shown).

## Discussion

Ideally, an algorithm for protein-protein docking would take the information contained in the individual three-dimensional structure files of two proteins to generate a new data file containing the coordinates of the docked complex. This is a lofty goal given the potential for conformational rearrangements of both side-chains and backbones in protein docking. Thus, we, as others before us, are developing and validating algorithms using known protein-protein complexes as test data. Although re-docking of such naive complexes is an artificial system compared to the desired goal of docking proteins based on their structures free in solution, it is a necessary step in the development and validation of an accurate docking algorithm that does not generate false positives. Believing that local shape complementarity is crucial to protein-protein recognition, we have developed our algorithm using geometric criteria alone. This is not a novel approach. However, it has proven difficult to implement and rigorously test given the difficulties of defining a shape-matching algorithm with sufficient sensitivity and attention to detail without becoming prohibitively expensive in computational time. Thus, most previous methods reduce the complexity of the protein surface, and supplement their geometry-based approach with electrostatics, H-bonding, hydrophobicity, etc. These techniques have allowed some successes in arriving at the correct docking solution, or at least in sorting through large numbers of incorrectly docked solutions with scores higher than the correct solution (false positives). Given that our aim is to develop fast algorithms for protein-protein docking based on geometry alone, including ones that allow for local conformational flexibility in de novo docking, we needed to generate proof-of-principle data supporting such a geometry-only approach. We have thus implemented a simplistic but informative and successful algorithm for protein-protein docking.

Below we further discuss important insights of broader significance to protein-protein docking.

**Covering and sampling transformation space.** Recall that the score of a relative placement of proteins $A$ and $B$ is the number of pairs of non-overlapping van der Waals spheres at distance at most $2\lambda$ from each other. The score is therefore a function from the space of rigid motions to the integers. As one may easily imagine, the score of two arbitrarily close rigid motions can be arbitrarily different. In other words, it is possible that the solution to a docking problem corresponds to a small island of rigid motions with high score in the middle of a vast sea of motions with low score and/or high bump number. Loosely speaking, our score function does not "sense" the solution even if the rigid motion is nearby. We have indeed observed that attention to detail in terms of evenly distributed high density sampling of translational and rotational space is essential if we want to find the correct solution without predicting false positives. The mere observation that we can detect the correct solution for all re-docking problems listed in Table 1 is strong evidence that our score function is well-suited to distinguish good from mediocre matches, another detail that contributes to our success. Similar to the score function used in ZDOCK (Chen et al. 2003), we count pairs of close spheres as opposed to spheres involved in such pairs. Indeed the only but possibly important difference is that our score is computed without resolution dependent error. The success of this score function is perhaps based on the similarity to the van der Waals force that attracts two proteins.

The running time of our algorithm is dominated by the sheer number of rigid motions required to explore in order to cover the space of rigid motions sufficiently fine. We have a three-dimensional space of translations, which we cover with a grid of step-size $2\delta$, and a three-dimensional space of rotations, which we cover with a point set of step-size $2\varepsilon$. The total number of rigid motions we generate is thus proportional to $1/\delta^3\varepsilon^3$. In other words, refining the covering to half the step-size, both for translations and for rotations, increases the running time of our algorithm by a factor of 64. We literally find ourselves between a rock and a hard-place: increasing the density beyond what we used to get the reported results is not practical, and decreasing the density would severely affect the quality of the results.

Any valid protein docking algorithm must be insensitive to the initial orientation of the proteins to be docked. A sensitivity to the initial orientation would bias all results and be problematic when we apply the algorithm to two separate protein structures and attempt to dock them de novo without knowledge of the correct answer. It is also of importance in our model system of re-docking known protein-protein complexes. Thus, in order to avoid cheating by re-docking back to the original position, we have implemented a random rigid motion of one of the two proteins in every experiment. The details of how this random rigid motion is computed can be found in the Methods section. It is important that this initial perturbation is independent of the way we cover the space of rigid motions, else there would be a possibility that the two steps interact with each other.

**Diversity of test set.** An important criterion for any computational docking algorithm is that it will be generally applicable to a wide variety of protein-protein complexes. Traditionally, the majority of test cases in protein docking algorithms have been the highly similar protease-inhibitor complexes (e.g. 1TGS, 2PTC, 3SGB, 4SGB). The interactions in these complexes are dominated by the tight binding of a peptide strand of the inhibitor into a groove on the enzyme; see Figure 4. As reflected by the low RMSD* values, docking of these protease-inhibitor complexes is predicted extremely well by our algorithm, as has also been found using other methods (Table 1). The core of the interaction motif shows a good overlap of the original and the docked structures for both the side chains and the backbone (Figure 4). Any observed deviations tend to arise from slight twists within the binding pocket that do not affect the quality of the docking prediction. Given the high degree of similarity among these complexes, an algorithm that correctly docks one of these complexes should be able to dock all the others. The protein-protein complexes that have a broader surface-to-surface interaction are more interesting from a biochemical point of view and more difficult from a computational point of view (e.g. 1A22, 1BUH, 1FIN, 1TX4, 3YGS). In the literature, these types of complexes have proven more difficult to dock than the protease-inhibitor complexes. Therefore we were pleased to find that our algorithm correctly docked all of these types of complexes (Table 1). Even, 3YGS, despite its very small and flat interaction area, shows a correct docking conformation, see Figure 5. Given that all of the primary amino acid contacts at the interface are preserved, the quality of these docking results are sufficient for the biochemist to proceed with further investigations (e.g. site-directed mutagenesis).

In conclusion, we are able to re-dock a wide variety of protein-protein complexes using geometric shape-matching criteria alone without the occurrence of false positives. This is an encouraging and significant result for the field of computational docking of proteins, given the complex and lengthy calculations required to dock proteins using energetic considerations. Our success lies both in the scoring algorithm that has sufficient attention to detail and in a sufficiently dense search-grid. We thus provide a proof-of-principle result for further development of faster docking algorithms based on shape complementarity alone. In the future, we will enhance our approach to incorporate the conformational flexibility required for docking of unbound proteins.

## Materials and Methods

Our docking algorithm is based on a purely geometric approach, in which we treat the two proteins as solid objects and search for the best fit between them. To that end, we exhaustively explore the space of rigid motions and evaluate each motion using a score function. Hence the two topics of this section: rigid motions and scoring.

**Representing rigid motions.** A *rigid motion* is a transformation that preserves distances and orientations. There are two types of rigid motions in $\mathbb{R}^3$: *translations* that preserve difference vectors and *rotations* that preserve the origin. Indeed, every rigid motion can be expressed as a rotation followed by a translation: $\mu = \tau \circ \varrho$. The translation is obtained by adding a three-dimensional vector to every point $x$ in $\mathbb{R}^3$: $\tau(x) = x + t$. Sampling the space of translations is thus as easy as sampling vectors in $\mathbb{R}^3$, but rotations are more subtle. It is customary to express a general rotation as the composition of three rotations about the coordinate axes, using *Euler angles* to parametrize the three special rotations, see eg. (Leach 1996). A disadvantage of that composition is the difficulty to sample the space of rotations uniformly, which is the reason we use quaternions instead of Euler angles. Specifically, we present a point $x = (x_1, x_2, x_3)$ by the 4-vector $\mathbf{x} = (0, x_1, x_2, x_3)$ and perform a rotation by multiplying $\mathbf{x}$ with a unit quaternion $\mathbf{q} = (q_0, q_1, q_2, q_3)$ from the left and its conjugate $\mathbf{q}^* = (q_0, -q_1, -q_2, -q_3)$ from the right. This operation yields a 4-vector $\mathbf{y} = (y_0, y_1, y_2, y_3)$ that corresponds to a point $y = (y_1, y_2, y_3)$ in $\mathbb{R}^3$. The map that moves $x$ to $y$ has a simple geometric interpretation: it is the rotation by the angle $\varphi = 2 \arccos q_0$ about the line spanned by the vector $(q_1, q_2, q_3)$ in $\mathbb{R}^3$. We see that two unit quaternions $\mathbf{p}$ and $\mathbf{q}$ represent different rotations unless $\mathbf{p} = \pm\mathbf{q}$. this implies that $\mathbb{S}^3$ is a double covering of the space of rotations.

The following result from differential geometry is useful for sampling rotations. Let $\mathbb{S}^2$ be the unit 2-sphere centered at the origin in $\mathbb{R}^3$, let $x$ be an arbitrary but fixed point on $\mathbb{S}^2$, and write $y = y(\mathbf{q})$ for the image of $x$ under the rotation that corresponds to $\mathbf{q}$, as before.

UNIFORMITY LEMMA If $\mathbf{q}$ is sampled uniformly at random in $\mathbb{S}^3$ then $y$ is uniformly distributed on $\mathbb{S}^2$.

We note that this is a desirable but far from trivial property. For example, the common sampling of Euler angles that translates into picking a line uniformly at random from the space of directions and an angle uniformly at random from $[0, 2\pi)$ does not enjoy this property. The Uniformity Lemma justifies the use of the geometry of $\mathbb{S}^3$ to measure the distance between rotations, which is the angle between the two unit quaternions defining the rotations: $\arccos \langle \mathbf{p}, \mathbf{q} \rangle$.

**Covering rigid motions.** Our algorithm explores the space of rigid motions by selecting a finite collection of points in this space that leaves no large empty gap. In other words, for every possible rigid motion there is a nearby selected rigid motion. Equivalently, we cover the space of rigid motions with small spherical neighborhoods. We describe this separately for translations and for rotations.

Let protein $A$ consist of $m$ atoms with centers denoted as $a_1, a_2, \ldots, a_m$ and let protein $B$ consist of $n$ atoms with centers denoted as $b_1, b_2, \ldots, b_n$. The *centroids* of the two sets are the average vectors: $\overline{a} = \frac{1}{m} \sum a_i$ and $\overline{b} = \frac{1}{n} \sum b_j$. It is convenient to first translate the two proteins so that both centroids lie at the origin of $\mathbb{R}^3$. After that initial translation, let $D_A$ be the maximum absolute coordinate of any $a_i$, and similarly let $D_B$ be the maximum absolute coordinate of any $b_j$. We define cubes $C_A = [-R_A, R_A]^3$ and $C_B = [-R_B, R_B]^3$ that contain $A$ and $B$ with the property that any translation that leads to non-overlapping cubes leads to protein placement with zero score and thus does not have to be considered by our algorithm. It is sufficient to set $R_A = D_A + r_{\max} + \lambda$ and $R_B = D_B + r_{\max} + \lambda$, where $r_{\max}$ is the largest radius of any van der Waals sphere in the data set and $\lambda$ is the separation threshold used in the definition of the score function. (Smaller non-cubic boxes with this property are possible and have been implemented but for the purpose of simplifying the discussion are not described here.) A translation $\tau(x) = x + t$ leads to overlapping cubes if and only if $t$ is a point in the cube $C = [-R, R]^3$, where $R = R_A + R_B$. To cover the space of translations represented by this cube, we choose a *step-size* $2\delta > 0$ and let $T_\delta$ be the portion of the regular cubic grid with step-size $\delta$ that lies within $C$:

$$T_\delta \quad = \quad \{(2i\delta, 2j\delta, 2k\delta) \mid -R \leq 2i\delta, 2j\delta, 2k\delta \leq R\},$$

where $i$, $j$ and $k$ are integers. The points in the cube that maximize the distance to the grid are the centers of the grid cells. These points are at distance $\sqrt{3}\delta$ from the grid points, which implies that the closed balls with centers in $T_\delta$ and radii $\sqrt{3}\delta$ cover $C$.

We treat rotations in a similar manner, by covering $\mathbb{S}^3$ with balls of small radius. There are many ways to cover $\mathbb{S}^3$, but none is as straightforward as taking the regular grid in $\mathbb{R}^3$. We choose a parameter $\varepsilon > 0$, which we refer to as the *packing radius,* and select a collection of points $S_\varepsilon$ in $\mathbb{S}^3$ such that no two are closer than $2\varepsilon$ and no point can be added to $S_\varepsilon$ without violating that property. The maximality of the collection implies that no point in $\mathbb{S}^3$ is at distance $2\varepsilon$ or more from a point in $S_\varepsilon$. This implies that the closed balls with centers in $S_\varepsilon$ and radii $2\varepsilon$ cover $\mathbb{S}^3$. The algorithm we use to select the points proceeds in layers of 2-spheres sweeping out $\mathbb{S}^3$. We begin with the equator 2-sphere, given by $x_0 = 0$ and $x_1^2 + x_2^2 + x_3^2 = 1$, and proceed in steps of distance $2\varepsilon$. For each 2-sphere, we proceed in layers of circles sweeping out the 2-sphere. Again we begin with the equator and proceed in steps of distance $2\varepsilon$. Finally, for each circle, we distribute as many points as possible at equal distances at least $2\varepsilon$ apart. The resulting set is maximal almost everywhere, except possibly near the poles of the 2-spheres and near the poles of the 3-sphere. We could make the set maximal by adding a few more points, but we did not in our implementation because the gaps are far between and not very large.

The total volume of $\mathbb{S}^3$ is $2\pi^2$. Each ball of radius $\varepsilon$ covers roughly $4\pi\varepsilon^3/3$ of that volume, which implies we cannot have more than roughly $3\pi/2\varepsilon^3$ points. Since the packing is not perfect, we expect only about half that number. Table 2 shows the numbers we get for a few values of $\varepsilon$. Recall that $\mathbb{S}^3$ covers the space of rotations twice. We thus use only half the unit quaternions generated by the algorithm, namely those with positive first non-zero coordinates.

**Random rotation.** In a *re-docking problem* we are given two proteins in complexed position, and the problem is to re-discover this position from separate descriptions of the proteins. To avoid using any knowledge of the given position, we apply a random rigid motion to one of the proteins and use this configuration as input to our docking algorithm. We really only worry about rotations since the initial step of our algorithm moves the centroids to the origin and therefore erases any knowledge of the correct translation in the first step. To pick a random rotation, we choose a point in $\mathbb{S}^3$ uniformly at random. By the Uniformity Lemma, this procedure is not only justified but indeed mandatory for else the algorithm could take unfair advantage of the statistical bias in which the two proteins are presented. There are various methods we could use to choose a point in $\mathbb{S}^3$ uniformly at random, one being described in (Marsaglia 1972):

1. Pick numbers $x, y, z, w$ uniformly at random in $[-1, 1]$.
2. If $x^2 + y^2 > 1$ or $z^2 + w^2 > 1$ then reject the selection and repeat Step 1. Else return $\mathbf{q} = (x, y, uz, uw)$, with

$$u = \sqrt{\frac{1 - x^2 - y^2}{z^2 + w^2}}.$$

Note that $x^2 + y^2 + u^2 z^2 + u^2 w^2 = 1$, so $\mathbf{q}$ is indeed a unit quaternion, as required.

**Computing the score.** Given a rotation $\varrho$ and a translation $\tau$, the number of *collisions* or *bumps* is the number of pairs of van der Waals spheres, one from $A$ and one from $B$, that have a non-empty intersection. The *score* is the number of non-intersecting pairs at distance at most some threshold $2\lambda$ from each other:

$$\begin{aligned} \text{bump}(\varrho, \tau) &= \text{card}\,\{(i, j) \mid \|a_i - b_j\| < \alpha_i + \beta_j\}, \\ \text{score}(\varrho, \tau) &= \text{card}\,\{(i, j) \mid \alpha_i + \beta_j \le \|a_i - b_j\| \le \alpha_i + \beta_j + 2\lambda\}, \end{aligned}$$

where $\alpha_i$ and $\beta_j$ are the radii of the van der Waals spheres with centers $a_i$ and $b_j$. The straightforward way of computing the two numbers compares all atoms of $A$ with all atoms of $B$, which takes time proportional to $mn$, and this for each rigid motion. We speed up the process by computing the bump and score values for all translations corresponding to a single rotation at once. To do this, we use two arrays that have an entry for each translation $\tau$ defined by a vector in $T_\delta$. Initially, all entries are zero.

```
for each rotation $\varrho$ do
  for each pair $(i, j)$ do
    for each translation $\tau$ do
      if $\|a_i - \tau(b_j)\| \leq \alpha_i + \beta_j + 2\lambda$ then
        if $\|a_i - \tau(b_j)\| < \alpha_i + \beta_j$ then bump$[\tau]$++
                                          else score$[\tau]$++
        endif
      endif
    endfor
  endfor
endfor.
```

So far, we still use the same amount of time, namely $mn$ for each combination of rotation and translation. We save time by restricting the innermost `for`-loop to only a small subset of all translations, namely the ones for which the cubes of side length $2\alpha_i + 2\lambda$ centered at $a_i$ and of side length $2\beta_i + 2\lambda$ centered at $b_i$ have a non-empty intersection. The amount of time this modification saves depends on the sizes of the two proteins and is usually about three orders of magnitude.

**Computing the root mean square distance.** After identifying promising docking positions using the score and bump functions, we evaluate their accuracy using two different standard root mean square procedures. The RMSD procedure calculates the root mean square distance between the initial points $b_j$ and the corresponding computed points obtained by applying, in this sequence, the translation $\tau_0$ of the centroid to the origin, the random rotation $\varrho_r$, and the computed rigid motion $\mu$:

$$\mathrm{RMSD}(\mu) \quad = \quad \sqrt{\frac{1}{n} \sum_{j=1}^{n} \|b_j - b'_j\|^2},$$

where $b'_j = \mu \circ \varrho_r \circ \tau_0(b_i)$. The RMSD$^*$ procedure does the same but restricts the sum to those residues of protein $B$ whose van der Waals spheres in the naive complex are within distance $2\lambda$ of the spheres of protein $A$.

## Acknowledgments

HE is partially supported by grants CCR-00-86013 and EIA-99-72879 from NSF, JR is partially supported by grant R01 GM61822-01 from NIH, and both are supported by a BGT Postdoc Program from Duke University. We acknowledge the use of the Duke Internet Systems and Storage (ISSG) Group “Devil” cluster funded by NSF CISE-RI grant EIA-9870724.

# References

ACKERMAN, F., HERRMANN, G., POSCH, S. AND SAGERER, G., 1998. Estimation and filtering of potential protein-protein docking positions. *Bioinformatics* **14**, 196–205.

BUCKLE, A. M., SCHREIBER, G. AND FERSHT, A. R., 1994. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. *Biochem.* **33**, 8879–8889.

CHEN, R., LI, L. AND WENG, Z., 2003. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80–87.

CONNOLLY, M. L., 1986. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interfaces. *Biopolymers* **25**, 1229–1247.

ELCOCK, A. H., SEPT, D. AND MCCAMMON, J. A., 2001. Computer simulation of protein-protein interactions. *J. Phys. Chem. B* **105**, 1504–1518.

EWING, T. J., MAKINO, S., SKILLMAN, A. G. AND KUNTZ, I. D., 2001. DOCK 4.0: search strategies for automated molecular docking of flexible molecular databases. *J. Comput. Aided Mol. Des.* **15**, 411–428.

FERNÁNDEZ-RECIO, J., TOTROV, M. AND ABAGYAN, R., 2002. Soft protein-protein docking in internal coordinates. *Prot. Science* **11**, 280–291.

FISCHER, D., LIN, S. L., WOLFSON, H. J. AND NUSSINOV, R., 1995. A geometry-based suite of molecular docking processes. *J. Molec. Biol.* **248**, 459–477.

HALPERIN, I., MA, B., WOLFSON, H. J. AND NUSSINOV, R., 2002. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409–443.

HUBBARD, S. J. AND ARGOS, P., 1994. Cavities and packing at protein interfaces. *Protein Sci.* **3**, 2194–2206.

JANIN, J., HENRICK, K., MOULT, J., EYCK, L. T., STERNBERG, M. J. E., VAJDA, S., VAKSER, I. AND WODAK, S. J., 2003. CAPRI: a critical assessment of predicted interactions. *Proteins* **52**, 2–9.

JONES, S. AND THORNTON, J. M., 1996. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**, 13–20.

LEACH, A. R., 1996. *Molecular Modeling. Principles and Applications.* Longman, Harlow, England.

LENHOF, H.-P., 1996. Parallel protein puzzle: a new suite of protein docking tools. In *Forschung und wissenschaftliches Rechnen: Beiträge zum Heinz-Billing-Preis 1996,* GWDG-Bericht Nr. 44, T. Plesser and P. Wittenberg, 31–48.

MARSAGLIA, G., 1972. Choosing a point from the surface of a sphere. *Ann. Math. Stat.* **43**, 645–646.

MONTELIONE, G. T. AND ANDERSON, S., 1999. Structural genomics: keystone for a Human Proteome Project. *Nat. Struct. Biol.* **6**, 11–12.

NOREL, R., FISCHER, D., WOLFSON, H. J. AND NUSSINOV, R., 1994. Molecular surface recognition by a computer vision-based technique. *Protein Engin.* **7**, 39–46.

NOREL, R., LIN, S. L., WOLFSON, H. J. AND NUSSINOV, R., 1995. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. *J. Molec. Biol.* **252**, 263–273.

WEISS, G. A., WATANABE, C. K., ZHONG, Z., GODDARD, A. AND SIDHU, S., 2000. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci. USA* **97**, 8950–8954.

WELLS, J., 1991. Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol.* **202**, 390–411.

| PDB | chains | sc | b | RMSD | RMSD* | | PDB | chains | sc | b | RMSD | RMSD* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1A22** | A: 1466 | 415 | 4 | – | – | | **1A4Y** | A: 3410 | 393 | 5 | – | – | |
| | B: 1578 | 270 | 5 | 2.03 | 1.44 | | | B: 993 | 278 | 5 | 1.93 | 1.31 | |
| | | 249 | 3 | 3.16 | 2.11 | | | | 235 | 5 | 25.87 | 28.46 | FP |
| | | 228 | 4 | 49.81 | 27.96 | FP | | | 222 | 4 | 47.49 | 44.90 | FP |
| **1BI8** | A: 2113 | 290 | 7 | – | – | | **1BUH** | A: 2311 | 185 | 0 | – | – | |
| | B: 1378 | 227 | 5 | 2.88 | 2.16 | | | B: 605 | 236 | 5 | 0.75 | 0.53 | |
| | | 219 | 5 | 61.47 | 57.93 | FP | | | 214 | 4 | 40.36 | 28.53 | FP |
| | | 218 | 4 | 69.47 | 73.89 | FP | | | 211 | 5 | 65.18 | 65.65 | FP |
| **1BXI** | A: 646 | 302 | 4 | – | – | | **1CHO** | E: 1750 | 263 | 1 | – | – | |
| | B: 1023 | 241 | 4 | 2.58 | 0.99 | | | I: 400 | 272 | 4 | 1.79 | 0.78 | |
| | | 235 | 5 | 3.33 | 1.52 | | | | 214 | 4 | 2.55 | 1.52 | |
| | | 223 | 4 | 1.57 | 1.01 | | | | 209 | 6 | 13.98 | 19.19 | FP |
| **1CSE** | E: 1920 | 269 | 3 | – | – | | **1DFJ** | E: 951 | 277 | 4 | – | – | |
| | I: 522 | 265 | 4 | 1.88 | 0.63 | | | I: 341 | 270 | 5 | 2.40 | 1.03 | |
| | | 265 | 4 | 1.88 | 0.63 | | | | 248 | 5 | 1.55 | 0.72 | |
| | | 222 | 5 | 40.75 | 35.41 | FP | | | 214 | 4 | 46.58 | 32.06 | FP |
| **1F47** | A: 135 | 170 | 3 | – | – | | **1FC2** | C: 354 | 194 | 0 | – | – | |
| | B: 1129 | 217 | 5 | 1.83 | 0.97 | | | D: 1656 | 233 | 5 | 2.85 | 1.15 | |
| | | 169 | 5 | 33.91 | 32.96 | FP | | | 223 | 4 | 3.56 | 1.07 | |
| | | 164 | 5 | 19.87 | 11.44 | FP | | | 205 | 5 | 62.06 | 42.81 | FP |
| **1FIN** | A: 2398 | 464 | 13 | – | – | | **1FS1** | A: 333 | 239 | 1 | – | – | |
| | B: 2101 | 272 | 5 | 4.01 | 2.40 | | | B: 909 | 252 | 5 | 1.35 | 0.64 | |
| | | 251 | 5 | 56.66 | 53.87 | FP | | | 213 | 5 | 37.43 | 30.05 | FP |
| | | 246 | 5 | 78.59 | 73.79 | FP | | | 201 | 5 | 2.90 | 1.37 | |
| **3HLA** | A: 2189 | 440 | 7 | – | – | | **1JAT** | A: 1193 | 232 | 1 | – | – | |
| | B: 829 | 254 | 4 | 1.76 | 1.85 | | | B: 1025 | 230 | 5 | 0.96 | 0.59 | |
| | | 235 | 5 | 7.06 | 7.53 | FP | | | 224 | 5 | 25.40 | 38.28 | FP |
| | | 231 | 5 | 1.42 | 1.30 | | | | 219 | 5 | 51.09 | 45.04 | FP |
| **1JLT** | A: 948 | 486 | 2 | – | – | | **1MCT** | A: 2029 | 483 | 19 | – | – | |
| | B: 963 | 355 | 5 | 1.40 | 1.21 | | | I: 265 | 348 | 4 | 3.18 | 1.68 | |
| | | 286 | 5 | 3.37 | 2.32 | | | | 275 | 3 | 3.90 | 2.74 | |
| | | 284 | 3 | 2.06 | 1.81 | | | | 259 | 4 | 36.02 | 26.64 | FP |
| **1MEE** | A: 1948 | 343 | 1 | – | – | | **2PTC** | E: 1629 | 287 | 2 | – | – | |
| | I: 530 | 287 | 5 | 1.35 | 0.82 | | | I: 454 | 321 | 4 | 0.96 | 0.53 | |
| | | 234 | 5 | 2.59 | 1.52 | | | | 277 | 4 | 3.85 | 1.51 | |
| | | 209 | 5 | 49.45 | 47.50 | FP | | | 238 | 3 | 3.36 | 1.16 | |
| **3SGB** | E: 1310 | 246 | 2 | – | – | | **4SGB** | E: 1310 | 245 | 2 | – | – | |
| | I: 380 | 280 | 2 | 0.58 | 0.54 | | | I: 380 | 280 | 5 | 1.57 | 0.59 | |
| | | 235 | 4 | 1.83 | 0.98 | | | | 249 | 3 | 1.03 | 0.66 | |
| | | 229 | 5 | 2.89 | 1.37 | | | | 226 | 3 | 2.35 | 1.41 | |
| **1STF** | E: 1655 | 279 | 1 | – | – | | **1TEC** | E: 2004 | 291 | 0 | – | – | |
| | I: 789 | 276 | 3 | 3.72 | 1.17 | | | I: 522 | 268 | 3 | 1.66 | 0.82 | |
| | | 264 | 3 | 1.79 | 0.76 | | | | 243 | 5 | 3.94 | 1.48 | |
| | | 246 | 5 | 35.22 | 36.51 | FP | | | 242 | 5 | 1.98 | 0.72 | |
| **1TGS** | Z: 1646 | 336 | 2 | – | – | | **1TX4** | A: 1579 | 330 | 2 | – | – | |
| | I: 416 | 316 | 3 | 1.28 | 0.59 | | | B: 1378 | 293 | 5 | 1.54 | 1.05 | |
| | | 285 | 5 | 1.82 | 1.05 | | | | 220 | 5 | 37.93 | 33.65 | FP |
| | | 206 | 5 | 30.55 | 27.41 | FP | | | 219 | 5 | 52.87 | 52.77 | FP |
| **3YGS** | C: 763 | 159 | 2 | – | – | | | | | | | | |
| | P: 790 | 223 | 5 | 1.09 | 0.71 | | | | | | | | |
| | | 213 | 5 | 32.98 | 20.53 | FP | | | | | | | |
| | | 204 | 5 | 48.73 | 49.31 | FP | | | | | | | |

Table 1: The respective first column gives the number of atoms/spheres of the two chains forming the complex. The respective last column indicates the false docking predictions. We note that 1CHO only has two scores with at most five bumps.

| $\varepsilon$ | 0.0676 | 0.0576 | 0.0476 | 0.0376 |
|---|---|---|---|---|
| # points | 4,852 | 6,552 | 12,036 | 25,432 |

Table 2: The number of balls of radius $\varepsilon$ our algorithm packs into the 3-sphere.

# Figure Legends

Figure 1: Searching with finer grid points, although computationally more expensive, yields higher scoring values in re-docking experiments of the two chains of the 1BRS complex. The two horizontal axes represent the translational step-size $2\delta$ and the rotational step-size $2\varepsilon$. The vertical axis represents the average score (A) or average RMSD (B) from twenty different random starting positions.

Figure 2: Overlay of the individual docking solutions from ten of twenty different starting positions of 1BRS using $\delta = 0.2$ and $\varepsilon = 0.0576$. Chain D is represented as a protein surface and serves as the fixed chain. The ten different chains A have been truncated to the interaction area (residues 24 to 47) and are shown in ribbons of different colors (original in black). The average score of 262 (with range from 167 to 364) and the average RMSD of $1.98\text{Å}$ (with range from $0.87$ to $3.45\text{Å}$) are represented in Figure 1. Note that despite the broad range of scores and RMSDs, all the docking solutions are correct at the site of interaction.

Figure 3: (A) Score (circles) and RMSD (triangles) vs. test numbers, as sorted by score, for the top scoring solution for each of the fifty different starting rotations using the parameters $\delta = 0.2$ and $\varepsilon = 0.0376$. (B) RMSD vs. top scoring solution for each of fifty different starting rotations.

Figure 4: Overlay of chain I of docked (yellow ribbon) and naive (red ribbon) with chain E (blue surface) of 3SGB complex ($\text{RMSD} = 0.58\text{Å}$). Side chains of residues 15 to 19 from chain I are shown in licorice to emphasize the predominance of the single peptide strand of chain I binding into the deep groove of chain E that is typical of a protease-protease inhibitor complex.

Figure 5: Overlay of chain P of docked (yellow ribbon) and naive (red ribbon) with chain E (blue surface) of 3YGS complex ($\text{RMSD} = 1.09\text{Å}$). Side chains of residues 11 to 15, 53 to 54, 57, 61, and 64 from chain P are shown in licorice to emphasize the predominance of the discontinuous and broad interaction surface between the two chains that is typical of most protein-protein complexes.