SURFACE SOLID ANGLE-BASED SITE POINTS FOR MOLECULAR DOCKING

Donna K. Hendrix* and Irwin D. Kuntz
*Department of Pharmaceutical Chemistry, *Graduate Group in Biophysics,*
University of California, San Francisco,
San Francisco, CA 94143-0446 USA

We are developing a new site descriptor for the DOCK molecular modeling program suite. Sphgen, the current site description program for the DOCK suite, describes the pockets of a macromolecule by filling a volume with intersecting spheres. DOCK then identifies possible ligand orientations in the pocket by overlapping the atoms of proposed ligands with the sphere centers. Sphgen limits use of the DOCK program to concave binding regions, but macromolecular binding regions can be solvent-exposed rather than buried pockets. We present a more general site descriptor, based on the surface solid angle, which generates site points by determining the solid angle of exposure for points on the surface of the molecule, then identifying patches of surface with similar solid angle values which are then built into site points. We find possible ligand orientations by matching shape-based site points on the ligand and protein and demanding complementary solid angle values. Orientations are evaluated using the DOCK's force field-based score, which evaluates the Coulombic and van der Waals energy. The surface solid angle descriptor displays the complementary characteristics of the interfaces of our test systems: trypsin/trypsin inhibitor, chymotrypsin/turkey ovomucoid third domain, and subtilisin/chymotrypsin inhibitor. The solid angle site points can be used by DOCK to generate orientations within 1.5Å r.m.s.d. of the crystal structure orientation.

1 Introduction

The interactions of proteins with other proteins and with DNA perform many of the signaling, recognition and catalytic functions within cells. The specificity of macromolecular interactions is due to a matching of complementary features in the interface of the complexed molecules. These features are both chemical in nature (e.g., salt bridges, hydrogen bonds, hydrophobic interactions) and geometric.

Solutions to the molecular docking problem have used approaches based upon the chemistry and geometry of macromolecules to reduce the solution space of the problem.[1,2,3,4] Lin et al.[5] define geometric "critical points" on the molecule, based upon the Connolly molecular surface. Each critical point also has an associated surface normal, and a defined character based upon the type of surface from which it was generated: cap, pit or belt. Several groups[6,2] describe the complementary nature of protein-protein and protein-ligand docking by describing the geometric interactions as protrusions which

fit into invaginations, or knobs-into-holes.

Macromolecular interactions do not always have a knobs-into-holes character, but can have large, smooth interfaces. In order to take advantage of these types of interactions with an existing docking algorithm, we have developed a site descriptor for docking based on the surface solid angle. This descriptor describes the local shape of the surface regardless of the features of the surface. We use these site descriptors as site points for DOCK, which generates orientations and evaluates them based upon electrostatic interactions.

2 Methods

2.1 Defining the surface and the local surface shape

The surface of proteins and ligands are described with Connolly's molecular surface (MS) program[7]. To calculate the solid angle, we require surface normals and associated areas in addition to coordinates for the surface points. The solid angle of each surface point is calculated using Connolly's solid angle algorithm[8], which places a test sphere center on a point, then determines the area of the test sphere which lies within the protein. The solid angle is then the portion of the surface area of the sphere that lay inside the protein, multiplied by 4π (see Figure 1). The solid angle is measured in steradians. The result of these calculations is a set of points in 3-space, each with an associated surface solid angle value.

The solid angle of a point lying outside of a surface is 0 steradians, while the solid angle of a point lying entirely within the surface is 4π steradians. Two complementary points have solid angles which sum to a value of 4π steradians. The radius of the test sphere used to calculate the solid angle is variable and set by the user. For these calculations, a solid angle radius of 5Å was used.

2.2 Determining site points: building regions

The purpose of calculating the solid angle is to use these data to dock two molecules together. Our docking algorithm grows geometrically with the number of site points. In order to reduce the docking time, the program, shapesite, reduces the number of site points by amalgamating them into shape regions. Shapesite examines near neighbor points and defines shape regions as clusters of adjoining points with similar solid angle values.

Neighbor lists are determined by a simple point-by-point search. The near neighbors of a point are defined as points within a distance of the square root of the density of surface points multiplied by two. For example, for these studies, a surface density of 1 dot/Å^2, and a neighbor search radius of 1.4Å is

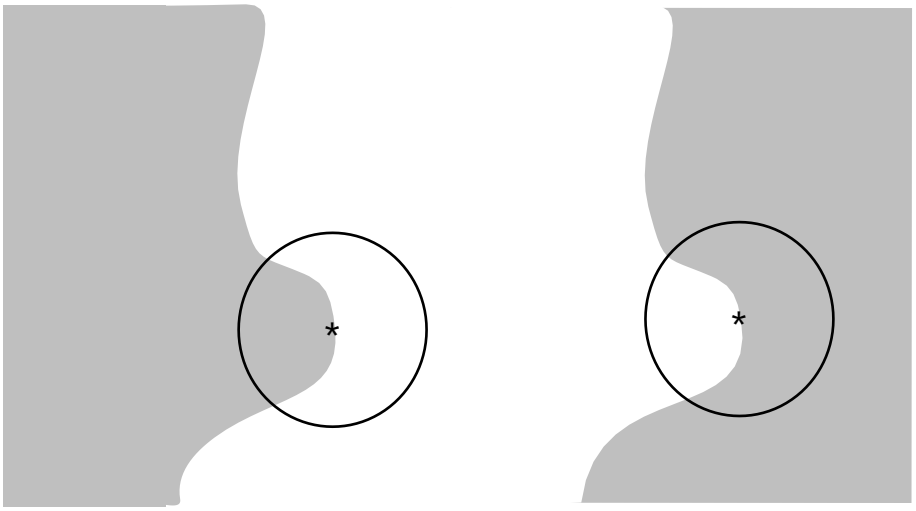

Figure 1: A description of the surface solid angle. The gray shaded area represents the interior of the protein. On the left, measuring the surface solid angle at the asterisk, approximately $\frac{1}{4}$ of the test sphere lies inside the protein, therefore its solid angle is $\frac{1}{4} \times 4\pi$ or π. On the right, measuring the surface solid angle at the asterisk, approximately $\frac{3}{4}$ of the test sphere lies inside the protein, therefore its solid angle is $\frac{3}{4} \times 4\pi$ or 3π. The surface at these two points complements, and the sum of their solid angles is 4π.

used. Up to 8 near neighbors are found. Regions begin as a seed point. The near neighbor list of the seed point is evaluated, and the neighborhood forms a region if all neighbors have a solid angle value within $\frac{\pi}{8}$ steradians. The region can grow larger by accumulating more near neighbor points if their surface solid angle is within $\frac{\pi}{8}$ steradians of the seed point. Regions have a minimum size of 5 $\AA^2$ and a maximum size of 15 $\AA^2$, with the average size of a region on a protein surface varying from 7 to 8.5 $\AA^2$. Each region has an associated solid angle value, and is represented as a site point in the DOCK algorithm by its center of mass and solid angle value. Regions may span several atoms.

The resulting site points formed from regions may vary with the order in which the surface points are searched. To determine the effect of the order of the search on the derived site points, we decoupled trypsin/trypsin inhibitor, derived site points and for the binding site first by the default order of points, which is ordered by residue in the protein from the N-terminus to the C-terminus. We then "shuffled" the residues in the data file so they were no longer in the N-to-C order (yet surface points from the same residues remain together). We re-assembled the complex and examined the complementarity of the site points from the unshuffled and shuffled data sets.

2.3 Docking with shape-based site points

Molecules are docked using version 4 of DOCK[9] and implementing the solid angle values as a shape-based filter. DOCK determines orientations by searching for distances between pairs of site points on the ligand that also exist between pairs of site points on the receptor. With shape-based site points, we additionally demand that the matched distances align such that the resulting adjacent ligand and protein site points which determine the match have complementary solid angles, with

$$solid\ angle(site\ point\ 1) + solid\ angle\ (site\ point\ 2) > 3\pi \qquad (1)$$

This shape filter is implemented as a chemical matching filter in DOCK[10].

We have selected three proteinase-inhibitor complexes for this study: chymotrypsin/turkey ovomucoid third domain (1cho)[11]; trypsin/trypsin inhibitor (2ptc)[12]; subtilisin/chymotrypsin inhibitor (2sni)[13]. These structures were selected based on their resolution, which ranges from 1.9Å to 2.1Å, and for comparison to a previous study[1]. For each structure, the complexes were decoupled and shapesite generated shape-based site points for the entire inhibitor and for an area on the proteinase which covered the binding region of the inhibitor plus an additional 5Å in all directions beyond the binding site.

Input parameters to version 4 of DOCK include the minimum distance between site points, which is the shortest distance that will be compared between pairs of site points on the protein and ligand, and a distance tolerance. Two distances whose lengths differ by the distance tolerance are considered equal distances. For these studies we selected a large minimum distance tolerance, 4Å , because of the large ligand molecules. We selected distance tolerances of 0.65Å.

After determining possible orientations, DOCK places the ligand molecule into each orientation and scores it with the force field. Once the ligand molecule is positioned, DOCK uses a rigid-body simplex minimization to find a local minimum. This minimization step is the most CPU-intensive step of the docking algorithm. In order to reduce the number of orientations minimized, we first use a "bump" filter. The bump filter evaluates the orientation and determines if there will be a significant overlap between ligand and receptor atoms.

3 Results

3.1 Regions

Shapesite quickly defines regions from solid angle data. For trypsin, a structure with 223 residues, the calculation requires less than 5 CPU seconds, and less than 2 CPU seconds for trypsin inhibitor, with 56 residues, on an SGI Octane (single processor MIPS R10000 CPU). Because the solid angle algorithm requires a comparison of all surface points to all other surface points, it requires significantly more CPU time. For example, for trypsin inhibitor calculating the solid angle on the SGI Octane requires 2,178 CPU seconds.

The formation of regions from individual points is dependent upon the order in which points are searched during the calculation; however, regardless of the order, derived site points display the complementary nature of known interfaces. For the case of trypsin-trypsin inhibitor, we compare the site points from the default ordering with site points derived from "shuffling" the residues in the input file. We examined the re-assembled trypsin-trypsin inhibitor interface for site points which lie within 2Å of one another across the interface. For the first, unshuffled run, there are 24 site points on trypsin within 2Å of a site point on trypsin inhibitor. These adjacent regions display complementarity: when the surface solid angles of the adjacent site points are summed, their average value is 3.3π steradians, and their standard deviation is 0.40π. For the shuffled data, there are 25 adjacent regions on the trypsin-trypsin interface, with an average value of 3.2π and standard deviation of 0.36π.

Table 1: Performance of DOCK runs with shapesite points. We report the number of site points on the proteinase and the inhibitor, as well as the number of orientations generated by DOCK with and without the use of shape filtering. We also report the CPU time, in minutes, to perform the DOCK runs on a Silicon Graphics Octane (R10000).

Complex	Site Points, Inhibitor	Site Points, Proteinase	Orient. without shape	CPU min.	Orient. with shape	CPU min.
2ptc	213	67	11,655,935	113.6	124,517	5.1
1cho	189	61	88,541	55.4	2,707	1.6
2sni	226	79	745,934	593.4	20,803	13.8

3.2 Macromolecular Docking

Docking studies are summarized in Tables 1 and 2. For each of the test cases, the top-scoring orientation also had the lowest r.m.s.d. from the crystal complex orientation, and it is always less than 1.5Å r.m.s.d. For comparison purposes, we report the DOCK force field score of the crystal complex, and the score of the complex minimized against the DOCK force field. For each proteinase-inhibitor complex, the DOCK structure is within 5.5 DOCK score units from the minimized crystal complex structure, and in two of the three cases, the DOCK orientation has a more favorable score.

The use of the shape-based filter vastly reduces the number of orientations searched, and therefore the computational time for docking these molecules. The reduction in both computer time and number of orientations searched approaches 100-fold, as shown in Table 1. For subtilisin/turkey ovomucoid third domain, fewer orientations were generated than for trypsin/trypsin inhibitor, however, more of those orientations passed the bump filter, so more orientations were minimized, and therefore significantly more CPU time was required to search the orientation space.

An example of the docked conformation and native crystal complex conformation of 2ptc can be found in Figure 2.

4 Discussion

Like Connolly's[2,3] and Lin et al.'s[5] molecular shape descriptors, these shape-based site points are derived from the Connolly molecular surface. Unlike Connolly's earlier attempts, we do not describe the geometric fit of proteins and ligands as strictly knobs-into-holes, but allow for a range of shape. Like Lin et al., our site descriptor is closely tied to our docking algorithm. Their

Figure 2: Trypsin inhibitor docked into trypsin (2ptc) using shapesite points. The crystal structure orientation of trypsin inhibitor is shown in black and the docked structure is in gray. The r.m.s.d. between the two structures is 0.85Å. The trypsin surface is shown in light gray.

Table 2: Results from DOCK runs with shapesite points. Reported are the DOCK force field score of the proteinase-inhibitor complex before and after minimization, the top-scoring orientation from the DOCK runs with shape-based site points and the r.m.s.d. from the unminimized complex of the top-scoring orientation.

Complex	Score, Complex	Min. Score, Complex	Score, DOCK/Shape	r.m.s.d. DOCK/Shape
2ptc	-72.26	-87.53	-85.71	0.85Å
1cho	-21.05	-75.94	-75.63	1.27Å
2sni	-50.02	-70.04	-68.65	0.37Å

algorithm makes use of critical points classified as a cap, pit or belt, and normal vector. Our method allows for a range of shape but does not use a projected normal.

Earlier efforts from this group focused on the complexes examined in this study and uncomplexed forms of the same molecules[1]. In that study, the inhibitor was partitioned into several smaller groups of 40 to 60 spheres, or site points, and the proteinase active site was represented by 40 to 90 site points. The selection and reduction of site points was a highly interactive process. Some 1-2 million orientations were generated for the complexed sites in several separate DOCK2 runs which took several days to run.

For this study, a similar number of site points were generated for the proteinases and inhibitors as were for the previous study. Our site points were generated by their shape criteria, and required no further clustering efforts. While the region-building algorithm is dependent upon the order in which the points are searched, the resulting number and complementary nature of the site points varied little with the ordering of site points.

Unlike the earlier study, we were able to simultaneously examine the entire inhibitor surface in one DOCK run. The number of orientations generated with DOCK is dependent upon the site points, a minimum distance between site points, and a distance tolerance set by the user. For these runs, the minimum distance between site points was 4Å and the distance tolerance was 0.65Å. The number of orientations generated varied from a few hundred with shape filtering to nearly 12 million without shape filtering. The differences and variations in the number of orientations generated and the significant change in time required to perform these runs (less than 10 hours) is due to the improvements in both software and hardware technologies.

With filtering based upon shape-based site points, we generated from 300 to 125,000 orientations for a protein-protein complex, depending upon the test case, which is nearly 100-fold the number of orientations generated without

shape filtering. With shape filtering, we reached the same or better orientation than without shape filtering.

Acknowledgments

This work was supported by NIH Training Grants GM08284, GM08388 and GM31497 from the Institute of Gerneal Medical Sciences, National Institutes of Health.

References

1. B.K. Shoichet and I.D. Kuntz, "Protein Docking and Complementarity", *J. Mol. Biol.* **221**, 327 (1991)
2. M.L. Connolly, "Shape Complementarity at the Hemoglobin alpha 1 beta 1 Subunit Interface", *Biopolymers* **25**, 1229-1247 (1986).
3. M.L. Connolly, "Shape Distributions of Protein Topography", *Biopolymers* **32**, 1215-1236 (1992).
4. D. Fischer, S.L. Lin, H.L. Wolfson, R. Nussinov, "A Geometry-Based Suite of Molecular Docking Processes", *J. Mol. Biol.* **248**, 459-477 (1995).
5. S.L. Lin, R. Nussinov, D. Fischer, H.J. Wolfson, "Molecular Surface Representations by Sparse Critical Points", *Proteins* **18**, 94-101 (1994).
6. I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Landgridge, T.E. Ferrin, "A Geometric Approach to Macromolecul-Ligand Interactions", *J. Mol. Biol.* **161**, 269-288 (1982).
7. M.L. Connolly, "Solvent-Accessible Surfaces of Proteins and Nucleic Acids", *Science* **221**, 709-713 (1983).
8. M.L. Connolly, "Measurement of protein surface shape by solid angles", *J. Mol. Graphics* **4**, 4-6 (1986).
9. T.A. Ewing and I.D. Kuntz, "Critical Evaluation of Search Algorithms for Automated Molecular Docking and Database Screening", *J. Comp. Chem.* **18**, 1175-1189 (1997)
10. B.K. Shoichet and I.D. Kuntz, "Matching Chemistry and Shape in Molecular Docking", *Protein Engng.* **6**, 223-232 (1993)
11. M. Fujinaga, A.R. Sielecki, R.J. Read, W. Ardelt, M. Laskowski, Jr., M.N.G. James, "Crystal and Molecular Structures of the Complex of Alpha-Chymotrypsin with its Inhibitor Turkey Ovomucoid Third Domain at 1.8 Å Resolution", *J. Mol. Biol.* **195**, 397-418 (1987)
12. M. Marquart, J. Walter, J. Deisenhofer, W. Bode, R. Huber, *Acta Crystallogr., Sect. B* **39**, 480 (1983)

13. C.A. McPhalen, M.N.G. James, "Structural Comparison of Two Serine Proteinase-Protein Inhibitor Complexes: Eglin-C-Subtilisin Carlsberg and CI-2-Subtilisin Novo", *Biochemistry* **27**, 6582-6598 (1988)