

Automatic identification and representation of protein binding sites for molecular docking

JIM RUPPERT,* WILL WELCH, AND AJAY N. JAIN^{1,*}

Arris Pharmaceutical Corporation, 385 Oyster Point Boulevard, South San Francisco, California 94080

(RECEIVED May 16, 1996; ACCEPTED December 5, 1996)

Abstract

Molecular docking is a popular way to screen for novel drug compounds. The method involves aligning small molecules to a protein structure and estimating their binding affinity. To do this rapidly for tens of thousands of molecules requires an effective representation of the binding region of the target protein. This paper presents an algorithm for representing a protein's binding site in a way that is specifically suited to molecular docking applications. Initially, the protein's surface is coated with a collection of molecular fragments that could potentially interact with the protein. Each fragment, or *probe*, serves as a potential alignment point for atoms in a ligand, and is scored to represent that probe's affinity for the protein. Probes are then clustered by accumulating their affinities, where high affinity clusters are identified as being the "stickiest" portions of the protein surface. The stickiest cluster is used as a computational binding "pocket" for docking. This method of site identification was tested on a number of ligand–protein complexes; in each case the pocket constructed by the algorithm coincided with the known ligand binding site. Successful docking experiments demonstrated the effectiveness of the probe representation.

Keywords: binding site, molecular docking; protein–ligand interaction; protein surface representation

Computational screening has become a popular tool in the search for drug leads, and has the potential to amplify other capabilities such as high throughput screening. The approach involves "docking" potential ligands from a database of tens of thousands of small molecules against a 3D protein structure to identify those molecules that may bind to the protein, thus modulating its biological activity.

The effectiveness of a molecular docking program depends greatly on the *computational representation* of the intended binding site. This representation, referred to as a *pocket*, should reflect only those protein features implicated in the desired binding; inclusion of excess features multiplies runtimes needlessly, whereas missing features may make matching a ligand difficult or impossible. Binding pockets are typically created manually, since this only needs to be done once per screening run, and since there is often a priori knowledge of a protein's active site or favorable ligand binding motifs. However, an automatic method to identify binding sites and create pockets can eliminate human biases and oversights and offers a rigorous way to select protein features for docking. For example, one can automatically enumerate distinct variants of pockets, or one can explore known protein structures for unknown or secondary binding sites. This can be important in designing a small molecule mimetic of a large peptide or protein hormone, where it

may be necessary to modulate receptor activity at sites distinct from the native hormone binding site.

There are several requirements for an effective algorithm for representing and identifying binding sites (a *pocket-finder*). The overriding goal is to narrow the search space of possible alignments explored during docking. This demands the selection of a minimal set of protein features to which ligands will be aligned. The pocket-finder must also choose a *sticky* region of the protein, one that has significant "binding opportunity" so that a docked ligand might achieve a high affinity for the protein. The pockets produced should be *well-connected*. It should be possible to fit a single ligand molecule into the pocket, ruling out pockets that have multiple disconnected components (separated by protein) or narrow constrictions.

This paper describes a pocket-finding algorithm that satisfies these requirements. The protein surface is characterized by a set of *probes* that indicate potential hydrogen bonds (or salt bridges) and favorable hydrophobic interactions with the protein. There are only three types of probes: a *steric (hydrophobic) probe* consists of a lone hydrogen atom; a hydrogen bond *donor probe* consists of N–H; a hydrogen bond *acceptor probe* is C=O. Together, the set of probes forms a sort of "protein complement," filling much of the void in and around the protein, with donor probes surrounding hydrogen bond acceptor atoms on the protein, and vice versa. See Figure 1 for an illustration.

The probes are used directly for generating ligand alignments in docking. For efficiency, the number of probes should be mini-

*Corresponding author.

¹Current address: MetaXen, 120 Independence Drive, Menlo Park, California 94025.

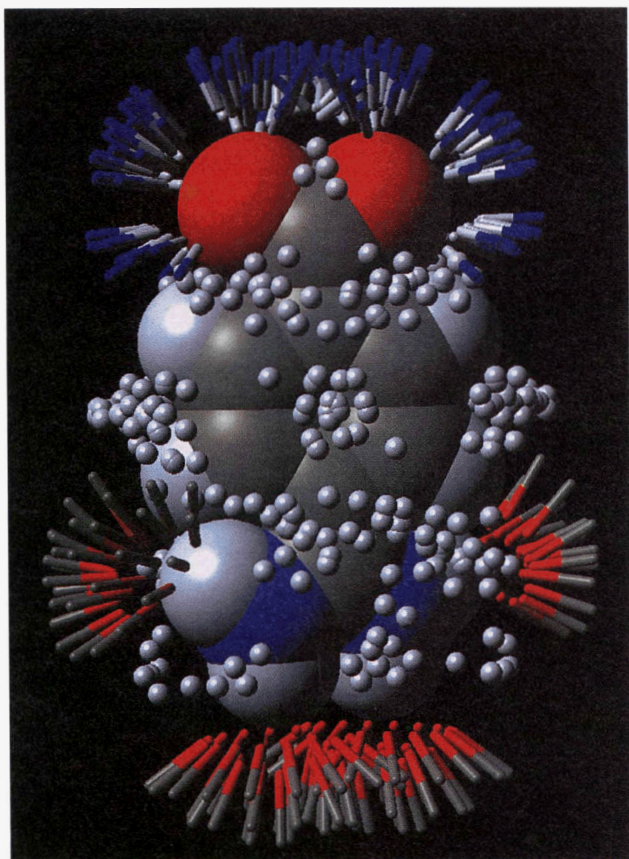


Fig. 1. Intermediate stage of probe placement around a small "protein." For clarity, a small molecule (*p*-aminobenzoic acid) is standing in for protein. Small spheres represent steric probes (i.e., an H atom), and tubes represent polar probes (i.e., N—H is a hydrogen bond donor, C=O is a hydrogen bond acceptor). The probe set has been thresholded by discarding low-scoring probes. The density will be further reduced with several filtering steps.

mized. The algorithm is quite stringent in this regard: probes are carefully positioned using a scoring function (Jain, 1996) to optimize their interaction with the protein, and then only the few best probes per protein atom are kept. When docking, conformational constraints may disallow exact ligand alignment to these probes, but the assumption is that the best ligand conformation contains near-optimal atomic interactions, and hence can be found by a robust docking program using the probe-based alignments as a starting point.

The algorithm uses the probes to compute a measure of "local protein stickiness" to identify the regions of strongest potential binding. The set of probes in each region is collected as a sticky spot. An example of a sticky spot is shown in Figure 2. Next, the geometry of the protein void is analyzed and a *pocket* is grown around each sticky spot. A pocket is a superset of the probes in the corresponding sticky spot. Figure 3 shows the pocket corresponding to the sticky spot of Figure 2. When used for docking, the sticky spot can serve as the "core" of the pocket. For instance, the Hammerhead docker (Welch et al., 1996) requires each ligand alignment to use at least one of the sticky spot probes. The remainder of the pocket allows for alternate alignments of the rest of the ligand. The user may select the desired maximum pocket size; larger pockets are useful for docking larger ligands.

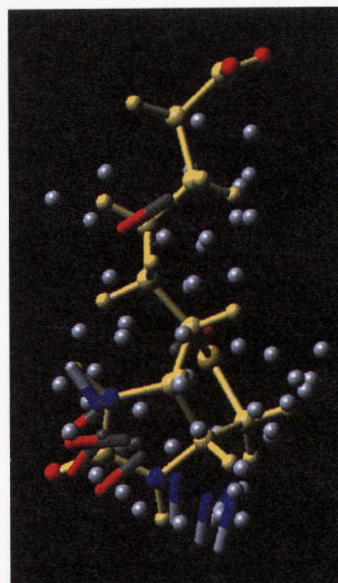


Fig. 2. Highest scoring sticky spot for streptavidin (not shown). Biotin, the natural ligand of streptavidin, is overlaid in yellow for reference. Many polar probes are closely aligned with hydrogen-bonding moieties on the ligand.

A detailed description of the algorithm is given in Methods.

The pocket-finder was tested on a number of proteins and ligand-protein complexes. After removing the ligands from the 3D structures, a pocket was automatically generated for each protein. In every case, the pocket coincided with the known ligand's binding site and had probes aligned with many of the ligand's atoms. The pockets were used successfully to re-dock the ligands to the proteins (Welch et al., 1996). The pocket-finder is a component of a docking system currently being used in ongoing screens for novel ligands of cytokine receptor targets and protease targets (Jain, 1996; Welch et al., 1996).

Results

This section evaluates the performance of the pocket-finder in several respects: the degree of similarity between the pocket probes and the ligand atoms, which is an indication of how useful the pockets will be for docking; the overlap of the sticky spots with the ligands, which indicates the pocket-finder's success at identifying the binding sites; and the extent to which the polar probes mimic the position and directionality of ligand-protein hydrogen bonds.

The performance of the pocket-finder is summarized in Table 1, which lists the protein test cases, the number of pockets identified on each protein, and the index of the pocket containing the ligand binding site. It also lists the scores of the two top-scoring pockets, to compare the algorithm's assessment of the actual binding site versus other regions of the protein. The last column lists the number of probes for the whole protein, the top-scoring sticky spot, and the resulting pocket. The pocket-finder identified from one to seven pockets on each protein, and ranked the pockets according to the score of their sticky spots. In each case, the top-scoring pocket was centered on the known ligand binding site. Figures 3, 4, 5 show pockets produced for three test cases, overlaid with the corresponding ligands for reference.

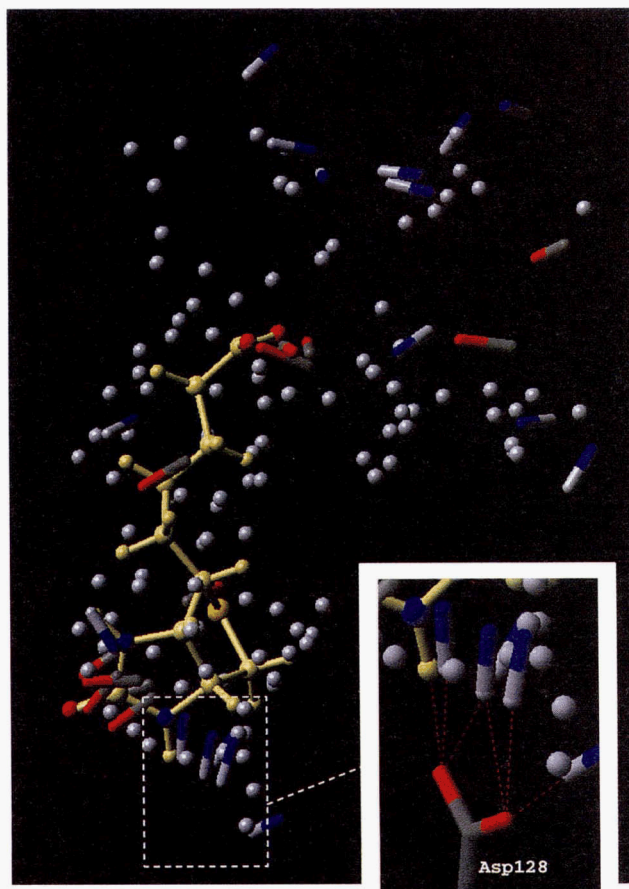


Fig. 3. Highest scoring pocket for streptavidin, derived from the sticky spot shown in Figure 2. Ligand biotin is overlaid in yellow. Inset shows closeup of some hydrogen bond donor probes and the Asp 128 they interact with on the protein.

Test cases

Eleven different X-ray crystallographic 3D protein structures were used in developing the algorithm. Table 1 lists these test cases,

which were chosen because of the wide variation in their binding affinities, binding pocket shapes, and ligand sizes. Nine were co-crystals of ligand–protein complexes, drawn from among those presented by Bohm (1994) and also used by Jain (1996). The last two test cases (chymotrypsin and streptavidin tetramer) are uncomplexed structures (no ligands in the crystal). Any water molecules present in the structures were removed.

The protein test cases had a wide variety of binding pockets, as evidenced by the variety of ligands. The streptavidin, DHFR, and trypsin cases will be discussed in some detail, as they reasonably represent the diversity of the full test set. Since the pocket identified in each case coincides with the binding site of a known ligand, the pockets will be evaluated by comparison with the bound conformations of the known ligands (biotin, methotrexate, and benzamidine, respectively). Close alignment of polar probes in a pocket with similar polar moieties on the ligand indicates that the algorithm has accurately selected favorable interactions with the protein. To the extent that the known ligand has a shape complementary to the protein, the steric probes in the pocket should be slightly inside the Van der Waals surface of the ligand. In broad terms, a good pocket for docking should “look like a ligand,” meaning it has similar probe density and probe types, and an overall shape that can accommodate a ligand. A pocket with too few features cannot effectively distinguish different potential ligands, and a pocket with too many features will generate many unproductive alignments during docking.

Probe proximity and sticky spot location

Table 2 contains the data used in evaluating the performance of the pocket-finder. It summarizes the distances between probes and bound ligand atoms of the same type (steric, hydrogen bond donor, or acceptor), for each of the nine test cases in which a bound ligand was present in the crystal structure.

For each ligand atom, the distance is computed to the nearest probe of the same type in the full pocket. The second column in Table 2 shows the minimum, average and maximum distances for each test case. The average values range between 1.00 Å and 1.91 Å, indicating that many ligand features have proximal representation in the pockets. As discussed by Jain (1996), RMS devi-

Table 1. Test cases and top-scoring pockets

Protein	Ligand	PDB reference	Number of pockets	Index of ligand-binding pocket	Scores #1, #2	Probes total, sticky, pocket
Streptavidin	Biotin	1STP	1	#1	142, —	1,968, 62, 136
DHFR	Methotrexate	4DFR	6	#1	110, 82	2,055, 52, 227
Thrombin	NAPAP	1DWD	5	#1	75, 70	3,428, 37, 205
Trypsin	Benzamidine	1PTB	1	#1	86, —	2,398, 41, 140
Carboxypeptidase	ZFV ^P (O)F	7CPA	5	#1	89, 75	3,137, 40, 108
HIV protease	L700,417	4PHV	7	#1	125, 82	2,462, 60, 263
Thermolysin	ZF ^P LA	4TMN	6	#1	116, 79	3,262, 58, 281
Cytochrome P450	Metyrapone	1PHG	2	#1	100, 59	4,441, 43, 69
Endothiapepsin	H-256	2ER6	6	#1	115, 75	3,349, 60, 381
Chymotrypsin	— ^a	4CHA	3	#1	77, 12	2,590, 36, 145
Streptavidin tetramer	— ^a	— ^b	4	#1–4	106, 84 ^c	5,397, 48–52, 85–109

^aUncomplexed (no bound ligand in crystal).

^bUnpublished structure of non-liganded streptavidin tetramer is due to Brad Katz.

^cLowest-scoring binding site pocket (#4) vs. #5.

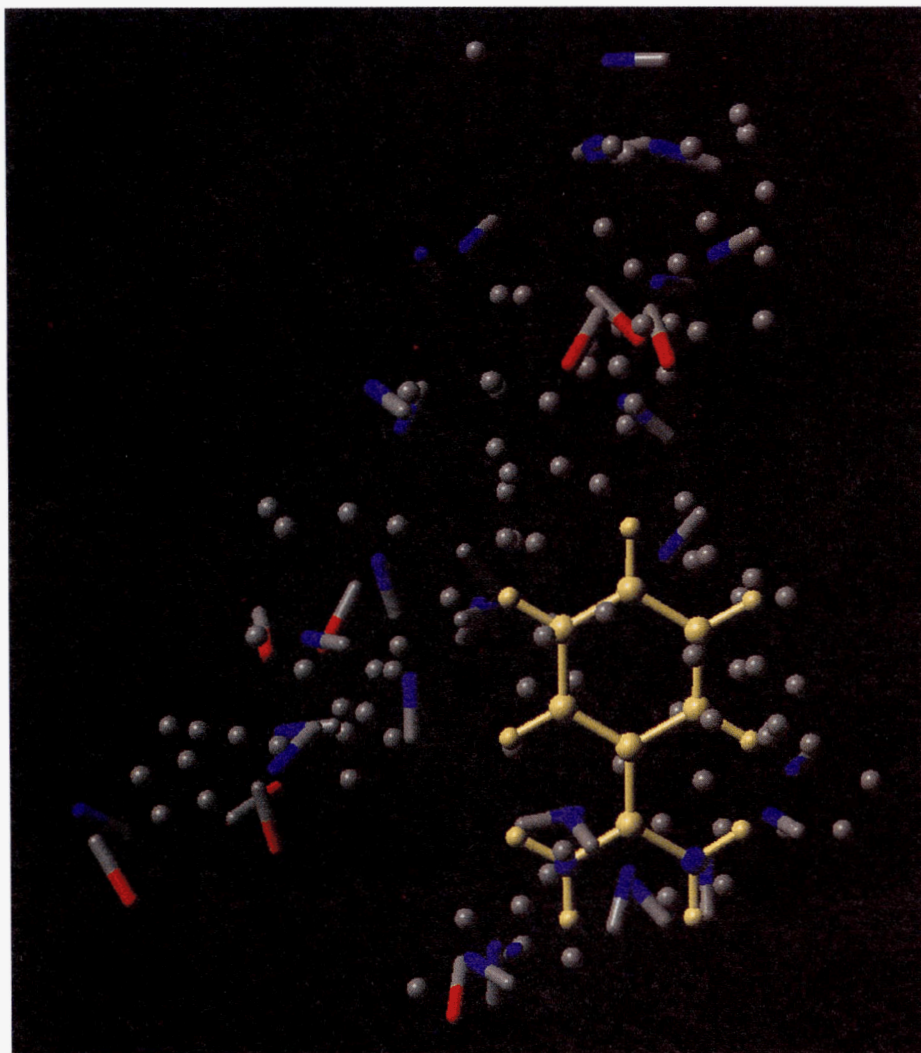


Fig. 4. Highest scoring pocket for trypsin, with ligand benzamidine overlaid for reference.

ations of up to 2.0 Å can be pulled back into a near-optimal alignment by the scoring function, so the probes in each pocket can effectively serve as ligand alignment points for molecular docking. For small ligands, the average distance is close to 1.0 Å; larger averages result in cases where the pocket does not fully cover a large ligand. The third column counts how many of the ligand atoms have a probe of the same type within 1.5 Å, and indicates that large portions of each ligand are represented with nearby probes.

The fourth column summarizes the analogous distances from the probes on each sticky spot (subset of a pocket) to atoms of a similar type on the corresponding ligand. This verifies that each sticky spot has a large overlap with the ligand, and hence has successfully identified the binding site. The fifth column counts the number of ligand atoms with nearby probes of like type, and comparison with the third column shows that although the sticky spots are generally much smaller than the entire pockets, they still represent from 9 to 27 of each ligand's atoms. For the two unliganded proteins in Table 1, the top-scoring sticky spots were visually determined to coincide with the known ligand binding site.

Hydrogen bonds in the crystal structures were also examined. The collection of test cases was determined to have 49 high-quality hydrogen bonds between ligands and proteins. Of these, 47 (94%) were found to have a corresponding probe within 1.5 Å of the participating ligand atom, whose direction was within 60 degrees of the hydrogen bond's direction. The average orientation discrepancy was 23.3 degrees.

Streptavidin test case

Figure 3 shows the pocket produced for the streptavidin test case. The protein is not shown, but for reference biotin is overlaid in its bound position from the co-crystal structure. The pocket can be compared with the sticky spot in Figure 2 from which it was derived.

The carbonyl on the lower left of biotin is known to make three hydrogen bonds with protein side chains (Weber et al., 1992). Five C=O probes represent interactions with different combinations of the protein's donor atoms. The neighboring N—Hs also make hydrogen bonds of nearly optimal geometry; note the nearly co-incident N—H probes. The inset at the bottom illustrates how the Asp 128 side chain gives rise to several donor probes representing

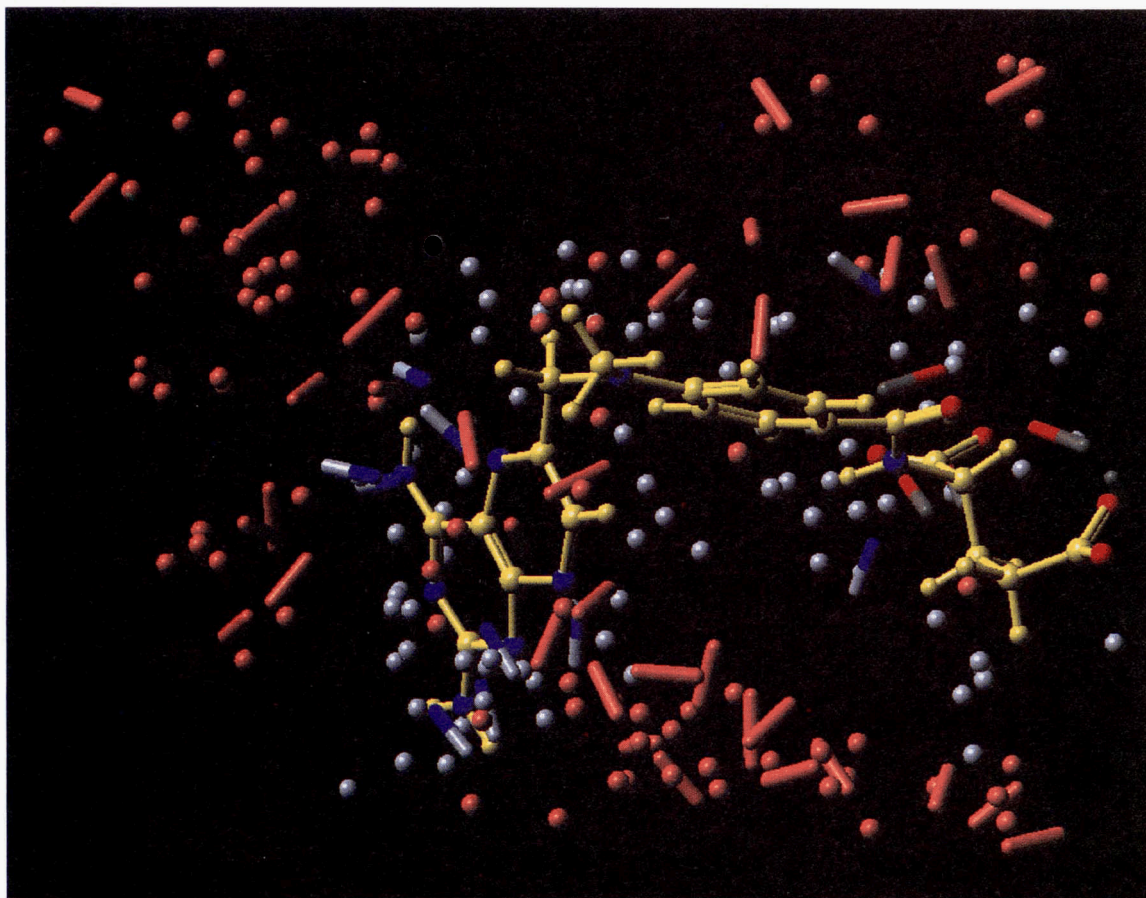


Fig. 5. Highest scoring pocket for DHFR, with ligand methotrexate overlaid for reference. Probes greater than 3 Å from the ligand are shown in pink.

potential hydrogen bonds, one of which exactly simulates the interaction biotin makes. The other polar moieties on the ligand (the thioether and the carboxylic acid) are represented by the corresponding types of polar probes. The remaining hydrophobic portion of the ligand is well-suggested by steric probes. These probes are not expected to align precisely with ligand atoms, due to the non-directional nature of Van der Waals interactions. Instead the

steric probes may be viewed as representative points on a hydrophobic surface. Most of the lipophilic hydrogens of biotin are seen to lie in this surface. A few more distant steric probes indicate small crevices in the protein not completely filled by biotin.

A comparison of Figure 2 with Figure 3 shows that the pocket extends well beyond the sticky spot at the top, near the “tail” of biotin, out into solvent on the exterior of the protein.

Table 2. Distribution of distances between probes and ligand atoms of similar types

PDB code	Ligand to pocket distances (Å) ^a		Sticky spot to ligand distances (Å)	
	Min/Avg/Max	Within 1.5 Å	Min/Avg/Max	Within 1.5 Å
1STP	0.27/1.00/1.63	26 of 31	0.27/1.28/4.05	23 of 31
4DFR	0.27/1.79/4.65	22 of 53	0.34/1.95/5.71	9 of 53
1DWD	0.25/1.57/3.47	38 of 69	0.33/1.47/3.06	14 of 69
1PTB	0.24/1.07/1.85	14 of 18	0.24/1.55/5.66	14 of 18
7CPA	0.38/1.79/4.90	30 of 74	0.38/2.22/8.81	12 of 74
4PHV	0.39/1.47/4.87	50 of 88	0.39/1.60/4.06	23 of 88
4TMN	0.13/1.39/3.04	42 of 68	0.13/1.56/5.24	23 of 68
1PHG	0.23/1.12/4.21	26 of 31	0.23/1.27/2.85 ^b	21 of 31
2ER6	0.25/1.91/7.65	52 of 127	0.25/1.49/3.43	27 of 127

^aHydrogen atoms have been added to all ligands.

^bThe maximum distance in the 1PHG case excludes 2 hydrogen bond donor probes not matching any atoms on the ligand.

Trypsin test case

The pocket produced for the trypsin test case is shown in Figure 4, overlaid with benzamidine for reference. The sticky spot (not shown) contained all the probes proximal to the ligand, and the pocket-finder added the probes above and to the left of the ligand, interacting with the exterior surface of the protein, as well as a few probes at the bottom center of the figure. The set of probes near the ligand bears a resemblance to the ligand itself: each polar N—H on the ligand has one or more corresponding N—H probes nearby, and the phenyl group is almost entirely enclosed by steric probes.

The pocket reaches approximately 10 Å further out onto the exterior of the protein. This pocket would be quite suitable for docking of larger ligands, although smaller pockets may be obtained by adjusting the pocket size parameter. The few probes in the bottom center reside in a crevice in the protein.

DHFR test case

The pocket for the DHFR test case is shown in Figure 5. Methotrexate (MTX) is overlaid for reference. The pocket was grown from a sticky spot (not shown) centered on the location of the pteridine ring system on the left side of the ligand. Probes further than 3.0 Å from the ligand are shown in pink to illustrate the overall topography of the pocket. The probes on the upper right are on the exterior of the protein. Unlike the streptavidin case, the internal cavity of DHFR includes additional branches not occupied by the ligand: a large void reaching upwards (the probes on the upper left), and a thinner tunnel reaching out towards the viewer (probes along the bottom right).

In the region of the pocket where the ligand binds, the probes correspond closely with the ligand: each polar N—H on the pteridine system has a corresponding N—H probe, and many of the polar moieties at the other end of MTX are represented by polar probes. The central hydrophobic portion is entirely represented by steric probes.

The large void above MTX is the binding site of NADPH (a co-factor of DHFR (Bolin et al., 1982)). NADPH was not present in the crystal structure used, and that portion of the protein was also identified by as a sticky spot by the pocket-finder. The two sticky spots listed for DHFR in Table 1, scoring 110 and 82, were both centered on the MTX pteridine ring system. The third best sticky spot scored 81, and covered a portion of the NADPH binding site. The portions of the pocket that extend beyond MTX allow the exploration of different binding modes during ligand docking.

Pocket-finder specificity and utility

The scores column of Table 1 indicates the difference in score between the two best pockets as identified by the algorithm. This illustrates the algorithm's ability to separate the stickiest spot on the protein from other regions. The relatively large separation in all cases (except thrombin) indicates that the algorithm is relatively insensitive to its parameter settings. Thus, non-default settings can be used (for example, to produce larger or smaller pockets), without affecting the algorithm's selectivity towards the known binding sites.

The final column in Table 1 gives the number of probes in the entire set for the protein, in the top-scoring sticky spot, and in the resulting pocket. A basic check on the efficacy of the algorithm is to note that, in each test case, the pocket-finder selected a very small subset of the input probes, correctly located on the known binding site. The full probe sets contained from 1968 to 4441

probes, (5397 for the streptavidin tetramer case), the number of probes being roughly proportional to the protein's surface area, or size. The pockets generated were derived from sticky spots that contained only 36–62 probes, e.g., about 1% to 3% of the input probes.

The pockets of streptavidin, DHFR, and trypsin that were discussed earlier all yield rapid and accurate dockings of the respective ligands using the Hammerhead program of Welch et al. (1996). The pocket-finder has also been used in successful computational screens for novel ligands of streptavidin and thymidilate synthase. The results of these experiments will be the subject of future publications.

Characterization of binding sites

The algorithm's success in identifying and characterizing known binding sites highlights two general properties shared by binding sites, at least for the studied cases involving small-molecule ligands binding largely in the interior of proteins. These binding sites are largely hydrophobic in nature relative to other parts of the protein, as evidenced by the increased weighting of hydrophobic scores required in the algorithm below. Also, at least a portion of every binding site was found to have a local score density well above the rest of the protein, as measured during the identification of sticky spots. Often, as in the cases of trypsin-like proteases and the streptavidin pockets, the local score density was associated with accessible polar moieties. However, in cases like chymotrypsin, hydrophobic score density was the dominant factor.

It is not clear whether the results of this paper will apply to protein-protein binding interactions, or to binding sites on the exterior of proteins. The sizes, shapes, and characteristics of such binding sites may differ from those of the cases studied in this paper, all of which involved internal protein cavities. Studies are ongoing to determine whether the algorithm performs well in these cases. Alternate parameter settings may be required that, for example, apply lesser weight to hydrophobicity or accept sticky spots that are less localized.

Discussion

Demarcation of a protein binding site and its representation as a pocket are critical preprocessing steps for molecular docking. The pocket-finder algorithm presented here provides a rigorous and robust way to automate these steps. It characterizes a protein with a sparse set of probes, small molecular fragments that represent potential ligand interactions with the protein. A novel technique is used to select the "stickiest" spot on the protein. The surrounding protein cavity is then analyzed geometrically to produce a reasonably shaped pocket for docking.

For rapid screening of large databases of compounds, the pocket must be a concise but accurate characterization of the protein features. To this end, only three types of probes were used: steric, hydrogen-bond donor, and hydrogen-bond acceptor. Formally charged interactions were represented by the hydrogen-bonding probes. The set of probes is as sparse as possible to reduce the docking search space by limiting the set of ligand alignments to be tested. A high degree of specificity was obtained by using an empirically tuned scoring function to optimally place the probes.

The algorithm was tested on a variety of protein and ligand-protein crystal structures. In each case, it successfully identified the known ligand binding site as the stickiest spot on the protein. Many of the probes in each pocket coincided with similar moieties

on the known ligands, which is an ideal situation for docking. The resulting pockets are accurate enough to be successful in docking known ligands to protein binding sites, and the pockets are sparse enough that rapid docking is possible.

The pocket-finder is a rigorous and effective alternative to manual preprocessing for molecular docking. Coupled with a scoring function and a docking engine, it gives a completely automated method for end-to-end computational screening. The pocket-finder is currently in use in lead discovery for both enzyme and receptor targets.

Relation to previous work

Many techniques have been previously published on analysis of protein surfaces as a preparatory step for computational docking or other applications, although automatic identification and demarcation of binding sites appears to be novel.

Most of the previous work on docking uses protein descriptors that are related to the pocket-finder's probes. The general strategy for protein characterization used in the DOCK program of Shoichet & Kuntz (1991) is to fill the protein void with a collection of spheres touching the protein surface in two places. This collection of spheres forms a negative image of the protein, and the sphere centers are used for docking alignments. The spheres can also be assigned types according to protein complementarity (Shoichet & Kuntz, 1992). The DOCK system includes a CLUSTER program for choosing clusters of features into which ligands may be docked (Shoichet et al., 1993). No automatic method is given for selecting the best cluster, but otherwise, many aspects of the cluster analysis are similar to the accretion process used in this paper. The binding site is typically selected with human intervention (Miller et al., 1994), or defaults to the location of a known ligand (Clark et al., 1995; Gehlhaar et al., 1995; Rarey et al., 1995). An alternative is to exhaustively dock against the entire protein surface (Jiang & Kim, 1991; Lin et al., 1994), which increases the computational cost.

The GRID program developed by Goodford (1985) computes the interaction energies of a user-specified probe at all points of a grid surrounding the protein to indicate the desirability of placing various ligand atom types at the grid points. Other approaches do not actually compute scores, but assign types based on proximal protein atoms: in the approach of Jiang & Kim (1991), each "surface cube" receives one of six types (positive or negatively charged, hydrogen bond donor or acceptor, polar, or hydrophobic), and in the approach of Gehlhaar et al. (1995) one of four types is assigned to each protein atom itself.

A purely geometric approach to molecular surface description is reported by Lin et al. (1994). "Critical points" are selected for faces on the Connolly surface, for instance "pits" which lie in the crotches between three atoms.

The pocket-finder's probes are most closely related to a series of approaches that utilize interaction points or molecular fragments as descriptors. The descriptor approach is typified by the work of Rarey et al. (1995), in which discrete "interaction points" of five or more types are placed around protein atoms. The descriptors used by PRO_LIGAND (Clark et al., 1995) include four types of "interaction sites": hydrogen bond donors and acceptors, which can be thought of as the two-atom molecular fragments N—H and C=O, respectively, and hydrophobic sites, which can be thought of as aromatic or aliphatic carbon atoms. PRO_LIGAND uses rule-based placement for these sites. The FLOG system (Miller

et al., 1994) combines grid and descriptor methods, computing six potential fields over a grid, then selecting grid points which are local maxima of interaction energy to be used as "match centers."

Methods

This section presents the pocket-finder algorithm in detail. There are three main steps in the pocket-finder: probe placement, sticky spot identification, and pocket accretion. First, three types of probes are placed densely around the protein to complement its surface. The set of probes is filtered, retaining only those representing the strongest interactions with the protein. Sticky spots are then located by selecting probe subsets which have the potential to make strong cumulative interactions with the protein in a small volume. Then a set of *protein-free spheres* are "inflated" in the protein void. These spheres guide an accretion process that grows the sticky spots into pockets. Finally the pockets are scored, and the top-ranked pocket is output.

Brief review of scoring function and Hammerhead molecular docking program

The pocket-finder is part of a computational screening system that also includes a scoring function and the Hammerhead program for flexible molecular docking. They are discussed in detail in Welch et al. (1996) and Jain (1996); the features of those components that pertain to the pocket-finder are reviewed here. The scoring function estimates the binding affinity between two molecules in a given alignment. It was tuned empirically on a large set of complexed ligand-protein crystal structures, and achieves an expected error of 1.0 log units. The function is continuous and differentiable with respect to the alignment, enabling optimization of the alignment via gradient descent. In the pocket-finder, the scoring function calculates each probe's affinity to the protein, and these affinities determine the strongest-interacting probes and the local "stickiness" of the protein surface. The optimization capability of the scoring function is used to optimize the alignments of the probes.

The Hammerhead docking program flexibly aligns small molecules to a known protein structure and uses the scoring function to predict their binding affinities. It screens a compound in a few seconds, a large database in a few days. Hammerhead operates directly on a set of probes produced by the pocket-finder, by aligning subsets of a ligand's atoms to the probes.

Probes for protein characterization/pocket representation

The first step in building a pocket is to locally characterize the protein surface by surrounding it with a set of probes. Each protein atom is classified as polar if it can make a hydrogen bond (or salt-bridge), or hydrophobic otherwise. Each atom is then densely surrounded by probes of the appropriate type. For instance, a protein atom that is negatively charged or is a hydrogen bond acceptor (e.g., the O from a protein's C=O) is surrounded by N—H donor probes. See Figure 1 for an illustration. The probes are placed at distances and orientations that are optimal for the interaction with the given protein atom. Long range dipoles and other interactions not considered by the scoring function (Jain, 1996) are not represented by probes.

Next, each probe's position and orientation are adjusted to maximize its interaction with all protein atoms by following the gradient of the scoring function (Jain, 1996). At this point, each probe

will have a score representing the binding contribution from a like atom on a ligand (scores are in units $-\log_{10}(K_d)$). Probes scoring below a fixed threshold are discarded. By default, the minimum score is half of the score of an ideal neutral hydrogen bond for a polar probe, and 2.5 times the score of an ideal hydrophobic interaction for a steric probe. Figure 1 shows the situation at this point, after removal of the low-scoring probes. Notice that the only steric probes remaining are those that can interact with more than one protein atom; hence the steric probes tend to crowd into the "crevices" between atoms.

Since the Hammerhead docking program operates by matching sets of ligand atoms against probes, the set of probes should be as sparse as possible (to keep docking time down) while still allowing discovery of desirable dockings. Thus, the next step reduces the probe density via a series of filters. Only a small number of probes making the best interaction with each protein atom will be kept. First, isolated probes are eliminated. Next, redundant probes are eliminated. For steric probes, any probe that is within 1.0 Å of a higher scoring steric probe is discarded. For polar probes, any probe that is within 1.5 Å and whose direction is within 60 degrees of another higher scoring polar probe of the same sign is discarded (unless otherwise specified, the position of a probe is considered to be the center of the atom at its head: the O of C=O, or the H of N-H). The filtering steps remove about 75% of the (thresholded) probes. The final set of probes will have an atom density close to that of a realistic ligand, somewhat higher in regions where several distinct forms of hydrogen bonds can be made with the protein.

Finding sticky spots

Each probe has an associated score s_i produced by the scoring function (Jain, 1996). The scores represent each probe's estimated (additive) contribution to the ligand-protein binding affinity. A region containing many high-scoring probes is thus a "sticky" part of the protein—a place where part of a hypothetical ligand might bind with high affinity.

The search for sticky spots is biased to prefer hydrophobic regions in the interior of the protein. Such pockets, when they exist, offer great potential for strong ligand binding, and many known protein-ligand binding complexes are of this type. The bias towards hydrophobicity is in general agreement with the observation that ligand-protein binding affinity is largely due to hydrophobic interactions (Jain, 1996). This bias is accomplished by multiplicatively weighting the scores of steric probes, yielding weighted scores $w_i = 4 * s_i$ for steric probes, and $w_i = s_i$ for polar probes. The value of 4 was not chosen systematically.

Sticky spots are those points with the highest local score density, computed at a point x as $\sum_i w_i$, where the summation is over all probes within 4.0 Å of x . The local score density is computed at each probe. For each probe whose local score density is at least 80% of the best, the probes within 4.0 Å are collected together into a probe group. Each of these groups is considered to be a sticky spot.

The next step is to merge redundant sticky spots, by combining groups. Given the score-sorted list of probe groups, each group is repeatedly merged with better scoring overlapping groups, unless the resulting set of probes would have a radius larger than a fixed limit r_0 . The default value is $r_0 = 5.5$ Å. This limit is chosen to bias pocket sizes towards what might be reasonably occupied by a small molecule ligand. After merging, each sticky spot's score is set to the sum of its probes' scores. Figure 2 shows the probes in

the top-scoring sticky spot computed for streptavidin using the default parameter settings. For reference, biotin is shown in its bound conformation.

One technical detail concerning the above method is that it may produce a sticky spot that is not *connected*: one that contains a pair of probes that are completely separated by protein. This can occur because the protein is not taken into account while determining sticky spots. A disconnected sticky spot leads to a disconnected pocket, which is undesirable for docking. The simple fix used for this problem was to do a preliminary *accretion* step (described in a later section), for connectivity analysis of the sticky spot. It determines which subsets of the probes in the sticky spot are connected, and the sticky spot is pruned to consist of only the largest connected subset.

Protein-free spheres

A set of spheres is used to represent the void around the protein, and allows efficient determination of the connected regions of the protein-free space suitable for ligand placement. Ideally, the set of spheres would be a minimal set that covers most of the protein-free space. However, a sufficient approximation is achieved with the following simple scheme. Spheres are placed with centers on a 1.0 Å cubical grid, and each sphere is grown until it reaches the Van der Waals surface of a protein atom. Spheres inside the protein's surface or with radii less than 0.5 Å are discarded. This set of protein-free spheres forms an approximate negative image of the protein. Figure 6 illustrates the idea of protein-free spheres in 2D.

Figure 7 shows the protein-free spheres computed for a portion of streptavidin. The cavity containing the biotin binding site is in the center (the bound position of biotin is indicated by the overlaid dotted yellow lines). The figure has been simplified by trimming large radii to 2 Å. At the top, near the carboxylic acid tail of biotin, the cavity opens out into solvent. Below biotin, there is a thin constriction in the cavity, about 2 Å wide, below which the cavity again opens into solvent. The regular pattern of spheres at the very bottom results from the edge of a bounding box containing the protein.

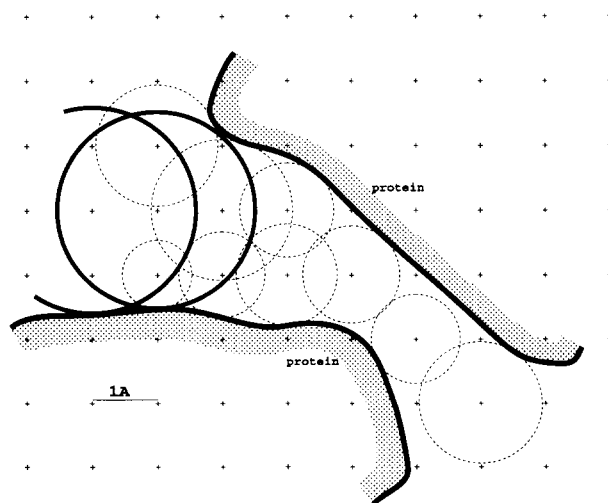


Fig. 6. 2D illustration of protein-free spheres. Only the two bold spheres are "spannable" with >2 Å interpenetration.

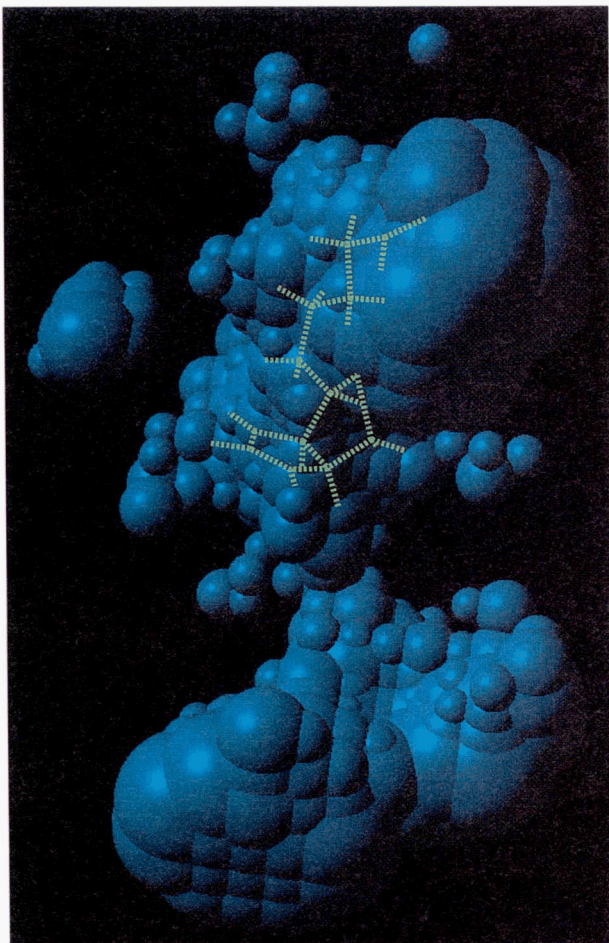


Fig. 7. Protein-free spheres for a portion of streptavidin including biotin binding site. Bound position of biotin is overlaid in dotted yellow. Note protein cavity “constriction” below biotin.

Accretion

The next step in the algorithm is to enlarge the sticky spots into pockets of the desired size by adding any nearby accessible probes. This is done via a process of *accretion* on the set of protein-free spheres. Some of the issues involved in pocket construction are illustrated in Figure 7. For instance, a large pocket that includes the biotin binding site might also extend out into solvent and the protein exterior if it were to be used for the docking of molecules larger than biotin. However, it would be undesirable for the pocket to include the isolated voids, or to extend through the thin constriction at the bottom, since no single small-molecule ligand could simultaneously dock into disconnected portions of the pocket.

A pocket's extent will be a set of overlapping spheres. Each pair of spheres is classified as *spannable*, *reachable*, or *disconnected*, depending on the amount of interpenetration. A pair of spheres is said to be *spannable* if their protein-free spheres interpenetrate at least 2.0 Å, as shown in Figure 6. An interpenetration greater than 0.7 Å is considered *reachable*, and pairs with a smaller interpenetration are disconnected. Spannable pairs of spheres tend to have large clearances to the protein, sufficient for a ligand molecule to “span” that region of space. Reachable pairs, by contrast, are closer to the protein, and indicate that a small ligand moiety may be accommodated.

More precisely, the interpenetration of two spheres having radii r_1 and r_2 , with centers at a distance d , is $r_1 + r_2 - d$. Since all spheres centers are at least 1.0 Å apart, the 2.0 Å interpenetration required of spannable pairs of spheres means that $r_1 + r_2 > 3.0$ Å, so at least one of the spheres must have a radius larger than 1.5 Å. This guarantees that the protein cavity will have a diameter of at least 3.0 Å locally. (Typically, the cavity must be slightly wider than this to be considered spannable, depending on the orientation and alignment of the spheres relative to the protein.) A sequence of spannable spheres is intended to reflect the thinnest passage in the protein void that can be spanned by a thin ligand such as a methylene chain.

For each probe in the sticky spot, the sphere with the nearest center is labeled as being in the pocket. From the labeled spheres, any spheres that are spannable are labeled, and this process is repeated up to some maximum distance d_1 from the sticky spot. This yields a skeletal pocket that spans a well-accessible portion of the protein void. This skeletal pocket is “fattened” slightly by repeatedly labeling any spheres that are reachable, out to a distance d_2 from the spannable points. Fattening helps to deal with “edge effects” from the approximation of the protein surface. The values $d_1 = 9.0$ Å and $d_2 = 1.5$ Å were found to work well over the range of proteins studied. Since r_0 is the sticky spot radius, the maximum pocket radius is $r_0 + d_1 + d_2$. Larger or smaller pockets are most effectively generated by varying the value of d_1 . To convert the set of labeled spheres into a set of probes suitable for docking, each probe whose nearest sphere center is labeled is added to the pocket. The output for each pocket is simply this set of probes. These pockets are ranked using the weighted scores of their corresponding sticky spots. More than one pocket may be produced for a protein if several sufficiently sticky spots are identified.

A sticky spot is output along with each pocket. Each sticky spot is a subset of the pocket's probes (called *anchor probes*) that can be used as a hint to the docking program, where it is required that all generated ligand matches utilize some of the pocket's anchor probes. When used in this manner, the anchor probes focus the docking program's search on ligands that match the best part of the pocket while reducing the size of the search space. This is similar to the concept of “essential points” used by Miller et al. (1994).

Acknowledgments

The authors would like to thank Mike Ross and Mike Venuti for support.

References

- Bohm H-J. 1994. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* 8:243-256.
- Bolin JT, Filman DJ, Matthews DA, Hamlin RC, Kraut J. 1982. Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Å resolution. *J Biol Chem* 257(22):13650-13662.
- Clark DE, Frenkel D, Levy SA, Li J, Murray CW, Robson B, Waszkowycz B, Westhead DR. 1995. PRO_LIGAND: An approach to de novo molecular design. 1. Application to the design of organic molecules. *J Comput Aided Mol Des* 9:13-32.
- Goodford PJ. 1985. A computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J Med Chem* 28:849-857.
- Gehlhaar DK, Verhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel LJ, Freer ST. 1995. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chemistry & Biology* 2:317-324.

- Jain AN. 1996. Scoring non-covalent ligand-protein interactions: A continuous differentiable function tuned to compute binding affinities. *J Comput Aided Mol Des* 10:427-440.
- Jiang F, Kim S-H. 1991. "Soft Docking": Matching of molecular surface cubes. *J Mol Biol* 219:79-102.
- Lin SL, Nussinov R, Fischer D, Wolfson HJ. 1994. Molecular surface representations by sparse critical points. *Proteins Struct Funct Genet* 18:94-101.
- Miller MD, Kearsley SK, Underwood DJ, Sheridan RP. 1994. FLOG: A system to select "quasi-flexible" ligands complementary to a receptor of known three-dimensional structure. *J Comput Aided Mol Des* 8:153-174.
- Rarey M, Kramer B, Lengauer T. 1995. Time-efficient docking of flexible ligands into active sites of proteins. *ISMB-95*:300-308.
- Shoichet BK, Bodian DL, Kuntz ID. 1992. Molecular docking using shape descriptors. *J Comput Chem* 13(3):380-397.
- Shoichet BK, Kuntz ID. 1991. Protein docking and complementarity. *J Mol Biol* 221:327-346.
- Shoichet BK, Kuntz ID. 1993. Matching chemistry and shape in molecular docking. *Protein Eng* 6(7):723-732.
- Weber PC, Wendoloski JJ, Pantoliano MW, Salemme FR. 1992. Crystallographic and thermodynamic comparison of natural and synthetic ligands bound to streptavidin. *J Am Chem Soc* 114:3197-3200.
- Welch W, Ruppert J, Jain A. 1996. Hammerhead: Fast, fully automated docking of flexible ligands to protein structures. *Chemistry & Biology* 3:449-462.