Mark R. McGann[1]
Harold R. Almond[2]
Anthony Nicholls[1]
J. Andrew Grant[3]
Frank K. Brown[4]

[1] *Open Eye Scientific Software, Santa Fe, NM 87501, USA*

[2] *Johnson & Johnson Pharmaceutical Research and Development LLC, Springhouse, PA 19477, USA*

[3] *Astra Zeneca Pharmaceuticals, EST(Chem) 26F17, Mereside, Macclesfield, Cheshire, SK10 4TG UK*

[4] *Johnson & Johnson Pharmaceutical Research and Development, LCC, 1000 Route 202, Raritan, NJ 08869*

Gaussian Docking Functions

Received 30 January 2002; accepted 15 April 2002

Abstract: *A shape-based Gaussian docking function is constructed which uses Gaussian functions to represent the shapes of individual atoms. A set of 20 trypsin ligand–protein complexes are drawn from the Protein Data Bank (PDB), the ligands are separated from the proteins, and then are docked back into the active sites using numerical optimization of this function. It is found that by employing this docking function, quasi-Newton optimization is capable of moving ligands great distances [on average 7 Å root mean square distance (RMSD)] to locate the correctly docked structure. It is also found that a ligand drawn from one PDB file can be docked into a trypsin structure drawn from any of the trypsin PDB files. This implies that this scoring function is not limited to more accurate x-ray structures, as is the case for many of the conventional docking methods, but could be extended to homology models.* © 2002 Wiley Periodicals, Inc. Biopolymers 68: 76–90, 2003

Keywords: *docking; structure based drug design; Gaussian based docking function; electronic screening of ligands*

INTRODUCTION

The goal of any docking method is to find the best candidate ligands, i.e., those with the highest binding affinity, in the smallest possible list to be assayed.[1] With molecular dynamics or Monte Carlo simulations, it is theoretically possible to rigorously evaluate the ΔG of binding, and hence the binding affinity of a ligand.[2,3] However, these methods are far too time-consuming to be used for the screening of large cor-

Correspondence to: Frank K. Brown; email: fbrown@PRDUS.JNJ.COM

Biopolymers, Vol. 68, 76–90 (2003)

porate databases. To improve docking throughput, scoring functions are employed that involve simplified approximations of the ΔG of binding dependent on the position of the protein and ligand atoms.[4–8] Docking methods attempt to find the optimum value of the scoring function for each ligand interacting with the protein as quickly and accurately as possible. The time required to find the optimal value of a scoring function depends strongly on the number of degrees of freedom as well as the complexity of the scoring function hypersurface. Due to large number of atoms involved and the number of factors often accounted for (e.g., steric fitting, hydrogen bonding, electrostatics, torsional strain, etc.), the complexity of the hypersurface is generally very high with many locally optimal values.[9–11] A typical approximation is that of a rigid protein and a torsionally flexible ligand.[12–22] The degrees of freedom are generally 3 rotation and 3 translations for the rigid ligand with respect to the protein plus N_{tor} torsional degrees of freedom for the ligands rotatable bonds. While it is desirable to include additional degrees of freedom to account for motion of some protein side chains, the additional complexity is often prohibitively high and the protein is usually treated as completely rigid. Even with this simplification, searching for the optimal value of the scoring function is time-consuming unless the search is initially close to the solution.[12–22]

Docking methods initially attempt the optimization of two components, shape and chemical function complementarity.[8,23–26] What is described here is a Gaussian Scoring Function (GSF) that has a simpler hypersurface than traditional scoring functions because it accounts for shape alone and uses smooth, analytical functions. This function can provide docking methods with a more tractable hypersurface to search for likely docked positions of the ligand. An ensemble of the most reasonable positions can then be refined using more detailed scoring functions. Thus, the purpose of the GSF is not to replace traditional scoring function, but rather to act as a prefilter that greatly reduces the search space docking methods must explore with more complicated scoring functions. Aside from being far more tractable to optimization, an advantage of a simplified hypersurface is that the atomic positions are more tolerant to errors: long-range interactions rather than short-range interactions tend to dominate. This makes the assumption of a rigid protein more tenable. The smoother surface created by the Gaussian representation will be less sensitive to small changes in the sidechain location induced by the ligand. As a result, we are now able to correctly dock inhibitors to crystal structures of proteins solved without ligands (apo structures) or conservative homology models.

In this article we demonstrate how a GSF with only two free parameters (a hardness and an exclusion factor) can successfully distinguish crystal structure poses, dramatically reduce the search space, and allow minimization over large distances. Furthermore, we show it can be used to dock ligands from closely related proteins, supporting a role for docking as a genomics tool.

THEORY

The GSF presented here is a modified version of that proposed by Grant and Pickup.[27,28] They observed that replacing atomic hard-sphere functions with Gaussians allowed analytically tractable calculation of volumes and areas of small molecules. They extended this observation to a simple two-atom Gaussian-based function that mimics van der Waals interaction. The "hardness" of this function, the degree of interpenetrating allowed, could be adjusted by a single parameter.

Gaussian Shape

Gaussian functions are of the form

$$g(r) = C \exp(-\alpha r^2) \tag{1}$$

where r is the distance from the Gaussian center, and C and α are parameters controlling the magnitude and distribution of the function, respectively. Gaussian functions are fundamentally smooth because they are infinitely differentiable. Numerical optimization methods either implicitly or explicitly assume continuous smooth functions, and while they are often robust enough to deal with small discontinuities, their performance improves when used with truly smooth functions. In general, the greater the number of higher-order derivatives available the more efficient the optimization (i.e., Quasi-Newton over Conjugate Gradient over Simplex).

The present scoring function describes each atom with a Gaussian function centered at the atom center. The value of the Gaussian is a measure of how much the atom occupies any particular position in space. This description of shape is not binary; instead, any given point in space is both partly inside and partly outside the atom, and the degree of each varies smoothly. Although there is no discrete surface that

delineates the inside and outside of the atom, an equivalent hard-sphere radius for the Gaussian atom can be obtained by requiring the Gaussian atom's volume to be equal to that of a hard-sphere atom. The shape of a hard sphere of radius R is given by a binary function that has a value of 1 inside the sphere and 0 outside (r is the distance from the sphere center).

$$f(r) = \begin{bmatrix} 0, & r > R \\ 1, & r \leq R \end{bmatrix} \tag{2}$$

The volume of the hard-sphere atom given by Eq. (2) is

$$V_{\text{hs}} = \frac{4}{3}\pi R^3 \tag{3}$$

while the volume of the Gaussian given in Eq. (1) is

$$V_{\text{g}} = C\left(\frac{\pi}{\alpha}\right)^{3/2} \tag{4}$$

Equating the hard-sphere and Gaussian volumes from Eqs. (3) and (4) respectively gives

$$C = \frac{4}{3}\frac{(\alpha R^2)^{3/2}}{\sqrt{\pi}} \tag{5}$$

It is convenient to define a dimensionless parameter:

$$\kappa = \alpha R^2 \tag{6}$$

Substituting Eqs. (5) and (6) into Eq. (1) yields the final Gaussian expression for atom shape:

$$g(r) = \frac{4}{3}\left(\frac{\kappa^3}{\pi}\right)^{1/2} \exp\left(-\frac{\kappa}{R^2}r^2\right) \tag{7}$$

Here R is the radius of the equivalent hard-sphere atom we wish the Gaussian function to represent, and is hence determined by the properties of the atom. The remaining parameter, κ, is a freely adjustable parameter.

The parameter κ controls the distribution of the Gaussian. A κ of 1.5 maximizes the value of the Gaussian function at $r = R$ and gives the most hard-sphere-like behavior (Figure 1). Increasing the value of κ further draws more and more of the atom toward the center, eventually becoming a Δ function at $\kappa = \infty$. Decreasing the value of κ diffuses the atom (Figure 1).

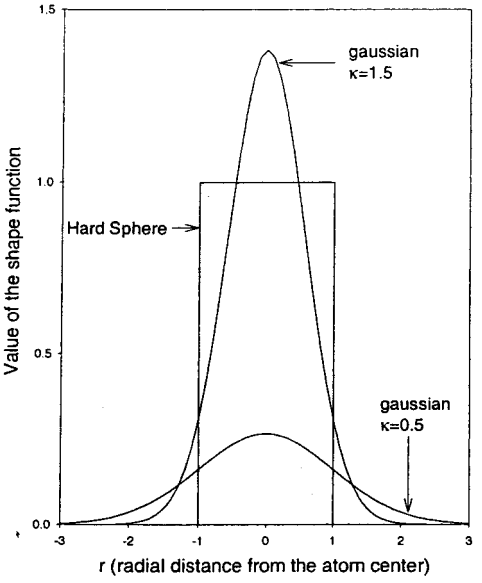

FIGURE 1 Plots of shape functions for an atom of radius one. The Gaussian shape function with a higher κ value (κ = 1.5) has a narrower, more hard-sphere-like, distribution with more of the atoms volume drawn toward the atom center, relative to the Gaussian shape function with a lower κ value (κ = 0.5).

The two critical aspects of our shape-based approach are error tolerance and robustness. By error tolerance we mean the ability of the GSF to be tolerant of errors in the protein structure, whether intrinsic, i.e., protein motion, or extrinsic, i.e., poor refinement, inadequate homology modeling. By robustness, we mean the ability of the function to accurately reproduce the effective shape of the pocket. By adjusting κ, the error tolerance vs robustness can be directly influenced. Lowering κ spreads the distribution, and makes the GSF less sensitive to atom coordinates and therefore more tolerant of error. Conversely, increasing κ increases the function's sensitivity to atomic position, making it more robust. For example, the shape functions of two identical atoms separated by a given distance are more similar (i.e., the dot product of the two functions are closer to one) at lower κ values (Figure 2). Therefore, an atom that is placed incorrectly, as compared to the crystal structure, has a shape that is more similar to the correct shape when the κ value is low. Thus, error tolerance comes at a cost of robustness. Typically, however, atomic coordinates are not known exactly and a highly robust docking function may be of little use.

Lowering κ should also reduce the number of GSF critical points (maxima or minima). This is significant because many docking methods become trapped in

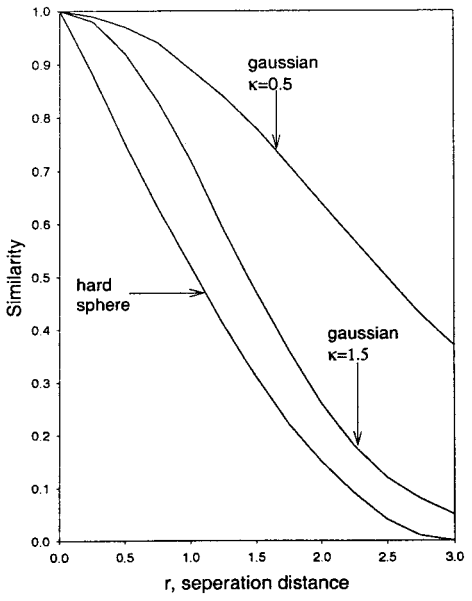

FIGURE 2 Similarity of pairs of shape functions with radius 1.5 Å separated by distance *r*, but otherwise identical. Similarity = ⇐⇐⇐f1*f2/(⇐⇐⇐f1* ⇐⇐⇐f2). A similarity value of one indicates that the functions are identical while a value of zero indicates the functions are orthogonal. At a separation distance of 1.75 Å the similarity of hard spheres, Gaussians with $\kappa = 1.5$ and Gaussians with $\kappa = 0.5$, are 0.22, 0.36. and 0.71 respectively.

local optima of the docking function while searching for the global optimum. Therefore, provided the location of the global optima is not significantly changed, reducing the number of local maxima in a docking function can greatly improve performance.

Docking Function

The Gaussian docking function originally suggested by Grant and Pickup is a summation of pairwise interactions between the ligand and protein atoms. Each pairwise interaction, $F_{i,j}$, has a volume overlap, $V_{i,j}$, and an area intersection component, $I_{i,j}$, which are functions of the distance between atoms, $d_{i,j}$:

$$F_{i,j}(d_{i,j}) = I_{i,j}(d_{i,j}) - \lambda V_{i,j}(d_{i,j}) \tag{8}$$

The λ is a constant determined as discussed later.

The volume overlap of two Gaussian atoms is found by analogy with the hard-sphere model by integrating the product of the shape functions. For two Gaussian atoms in the form of Eq. (7), this integration yields the following equation as a function of distance between the atom centers:

$$V_{i,j}(d_{i,j}) = \left(\frac{16\kappa^3}{9\pi}\right)\left(\frac{\pi R_i^2 R_j^2}{\kappa R_i^2 + \kappa R_j^2}\right)^{3/2} \exp\left(\frac{\kappa}{R_i^2 + R_j^2} d_{i,j}^2\right) \tag{9}$$

where radii R_i and R_j are the radii of atoms i and j, respectively, and $d_{i,j}$ is the distance between the two atom centers.

The area ($A_{i,j}$) of the intersection volume of two atoms i and j with radii R_i and R_j and intersection volume $V_{i,j}$ is given by the following equation[27]:

$$A_{i,j} = \frac{\delta V_{i,j}}{\delta R_i} + \frac{\delta V_{i,j}}{\delta R_j} \tag{10}$$

The component of the docking function, Eq. (8), related to the intersection area, is

$$I_{i,j}(d_{i,j}) = R_i \frac{\delta V_{i,j}(d_{i,j})}{\delta R_i} + R_j \frac{\delta V_{i,j}(d_{i,j})}{\delta R_j} \tag{11}$$

where the variables and function are the same as Eq. (9). This function is similar to Eq. (10), but has the desired dimensions of volume.

The λ is selected to maximize the pairwise interaction at a particular distance $D_{i,j}$. Grant and Pickup proposed this function could serve both for internal docking, i.e., partial shape matching and external docking, i.e., complementarity discovery depending on the choice of $D_{i,j}$.[5,6] This work focuses on external docking, i.e., fitting the ligand shape to the shape of a negative image of the protein, and $D_{i,j}$ is therefore selected as the sum of the hard-sphere radii.

$$D_{i,j} = R_i + R_j \tag{12}$$

The pairwise interaction then becomes

$$F_{i,j}(d_{i,j}) = \left(\frac{32\kappa^3}{9\pi}\right)\left(\frac{\pi R_i^2 R_j^2}{\kappa R_i^2 + \kappa R_j^2}\right)^{3/2} \times \left[1 + \kappa \frac{d_{i,j}^2 - D_{i,j}^2}{R_i^2 + R_j^2}\right] \exp\left(-\frac{\kappa}{R_1^2 + R_2^2} d_{i,j}^2\right) \tag{13}$$

Finally, a normalized docking function is defined such that

$$N_{i,j}(d_{i,j}) = \frac{F_{i,j}(d_{i,j})}{F_{i,j}^{\max}}$$

This reduces to

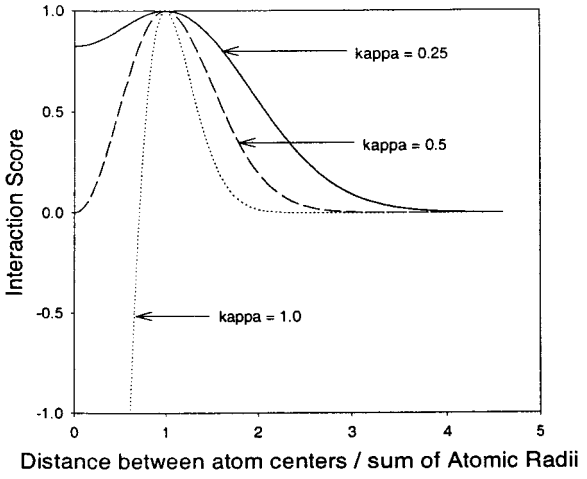

FIGURE 3 Pairwise interaction score between two Gaussian atoms [Eq. (13)] at three κ values. High scores are favorable interactions. The interaction is always most favorable at the sum of the hard-sphere distances. At a high κ values ($\kappa = 1.0$) the interaction of 2 atoms is less than that of noninteracting atoms (zero score) when the atoms are very close. However, at low kappa values ($\kappa = 0.25$) the interaction is never less than the noninteraction case (zero score).

$$N_{i,j}(d_{i,j}) = [1 + \xi(\mathrm{d}_{i,j}^2 - D_{i,j}^2)]\exp(-\xi(d_{i,j}^2 - D_{i,j}^2)) \tag{14}$$

The ξ is defined as

$$\xi = \frac{\kappa}{R_i^2 + R_j^2} \tag{15}$$

The total interaction, N, between a set of ligand atoms A and a set of protein atoms B is then simply

$$N = \sum_{i \in A} \sum_{j \in B} N_{i,j} \tag{16}$$

This is the form of the GSF initially proposed by Grant and Pickup.[3] It has reasonable behavior for $\kappa \geq 1.0$, favoring ligand positions where ligand atoms are near parts of the protein surface with high curvature. As the κ value is decreased, however, the most favorable position for the ligand atoms moves inside the protein. This is obviously not desirable. The reason is that at low κ the pairwise interaction, Eq. (13) does not sufficiently penalize atom clashes and the most favorable position for the ligand atom moves inside the protein. This effect is illustrated in (Figure 3), which shows the pairwise interaction of two atoms at several κ values. At a high κ value, around 1.0 or greater, the interaction becomes unfavorable once the atoms significantly overlap, but at lower κ values even the most severe clashes have favorable scores relative to the noninteracting spheres (zero score). The interior of the protein becomes a more favorable region for ligand atoms because there are more protein atoms making favorable surface contacts and no significant penalty for atom clashes, and the interior has more protein atoms with which to make favorable interactions. To address this problem, Eq. (16) was modified. An exclusion function, similar to that proposed by Grant, Nicholls, and Pickup for dielectric screening, was multiplied to the original GSF to give the new to give a new expression:

$$GS = \sum_{i \in A}[(\sum_{j \in B} N_{i,j})(\exp(-\gamma \sum_{j \in B} V_{i,j}))] \tag{17}$$

This is the form of the GSF reported in this article. The functions $N_{i,j}$ and $V_{i,j}$ are given by Eqs. (14) and (9), respectively, and γ is a new parameter. The new exponential term in Eq. (17) applies a clash penalty to ligand atoms that are placed interior to the protein and the parameter γ controls the strength of the penalty. An exponential form was chosen because it smoothly decreases as the ligand moves away from the protein interior. At $\gamma = 0$, it becomes the original docking function, Eq. (16).

Method

A simple docking method was used to explore the properties of our GSF, i.e., a Quasi-Newton solid-body optimization of the ligand position from random starting positions near the receptor site. For each run the center of mass of the ligand is randomly positioned within a box defined around the receptor site, the ligand is randomly rotated and then optimized. No attempt is made to avoid initial clashes between the ligand and protein.

Two sets of ligand–protein complexes drawn from the Protein Data Bank[29] were used to examine the properties of the GSF. The first is a set of 20 complexes of ligands bound to trypsin and the second is a set of 49 complexes of different ligands and proteins. The first set was used to see if the method can overcome a common criticism of rigid docking methods, i.e., that "redocking" ligand and protein from the same crystal structure is a form of "postprediction." If having the exact protein/ligand crystal structure is not necessary, apo and homology structures can be used

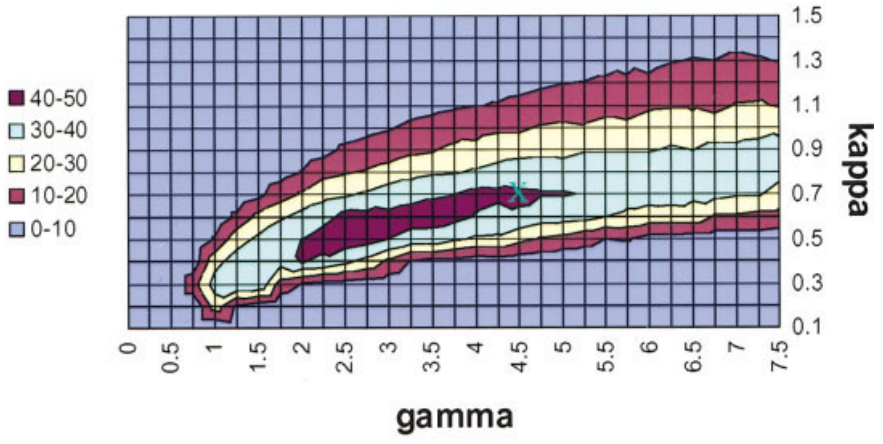

FIGURE 4 The number of docking runs, out of 1000, which resulted in a structure within 2.0 RMSD of the x-ray structure. These results are averaged over all 20 trypsin ligands. X marks the parameters selected for the dock run. The parameters were selected from the upper edge of the optimal region because this gave the best results for both large and small ligands.

with more confidence. The second test set was used to determine how the GSF performs as a general scoring function.

RESULTS

Optimum Parameters

The GSF optimal parameters (κ the Gaussian distribution parameter and γ the exclusion parameter) were determined by an iterative process using the trypsin complexes. For each κ and γ pair 1000 docking attempts were made for each ligand–receptor complex. Ligands were only docked to their own trypsin structure, not the trypsin structure of other ligand–trypsin complexes. If the ligand optimized to within 2.0 RMSD of the x-ray structure, the run was considered a success, and by summing the successes over all 1000 runs an average chance of a success was obtained. The results are shown in (Figure 4). An optimal value appear at $\kappa = 0.7$ and $\gamma = 4.5$. These are the values used for the remainder of the docking runs.

The values selected ($\kappa = 0.7$, $\gamma = 4.5$) are on the high-κ edge of the optimal region because it was found that the lower κ, the more the optimal γ varied with ligand size (Figure 4).

Docking Ligands to Non-Native Protein Structures

To test the sensitivity of the GSF to the protein structure, the 20 trypsin complexes were aligned by minimizing the RMSD between the residues near the receptor site and those same residues of a reference protein conformation. The reference trypsin structure was chosen to be the 1AQ7 complex trypsin structure and the residues used in the RMSD minimization were selected by hand. The GLN 192 residue, near the entrance to S1, was the most flexible side chain that interacted with the ligands and adopted several completely different conformations with the atomic positions varying up to 4.5 Å. The remaining relevant side-chain atoms aligned to within approximately 1.0 Å of the reference.

Each ligand was then docked into all 20 aligned trypsin proteins 1000 times. A docking run was considered a success if the optimized ligand structure was within 2.0 or 1.5 RMSD of the x-ray structures. The numbers of successful runs are shown in Tables I and II.

The average chance of docking any of the 20 ligands into any of the 20 trypsin x-ray structures and obtaining a result within either 2.0 or 1.5 RMSD of the x-ray structure is 4 and 3%, respectively. These optimizations move the ligand on average 7 Å. This is an extremely long distance for numerical optimizations that rarely result in motions greater than 2 or 3 Å when more traditional docking functions are used. The average chance of optimizing to the correct x-ray structure as a function of how far from the x-ray structure the docking run started is shown in Figure 5. As seen even starting from a distance of over 6 Å RMSD, there is still greater than a 10% chance that simple optimization will locate the x-ray structure, based on a simple shape-matching scoring function.

Table I Number of Times, Out of 1000 Trials, a Minimization from a Random Position Resulted in an Orientation Within 2.0 RMSD of the X-Ray Structure for Each Trypsin Ligand–Protein Pair

	Protein																			
Ligand	1AQ7	1AVW	1CE5	1JRS	1JRT	1MTS	1MTU	1MTV	1MTW	1PPC	1PPH	1TNG	1TNH	1TNI	1TNJ	1TNK	1XUI	1XUK	2BZA	2TBS	Average
1AQ7		2	3	9	6	1	2	0	4	3	3	1	0	1	1	4	2	2	1	0	4
1AVW	24		34	33	30	38	35	40	28	39	31	41	40	46	44	40	37	28	35	29	35
1CES	57	66		47	50	75	74	70	77	67	71	66	68	80	67	75	71	67	65	91	67
1JRS	44	29	13		22	43	31	18	32	7	17	19	18	14	13	16	25	40	21	20	25
1JRT	38	12	19	15		17	17	14	18	17	12	14	12	10	13	21	25	28	20	0	17
1MTS	0	22	39	13	0		32	7	42	19	18	18	18	23	18	19	13	36	19	12	20
1MTU	18	27	37	15	15	34		32	37	22	35	24	25	25	21	27	44	45	27	32	29
1MTV	24	2	51	10	15	22	24		38	13	1	0	2	2	0	4	21	30	19	13	16
1MTW	8	10	6	1	0	9	14	6		7	7	2	1	6	1	0	7	16	12	18	7
1PPC	56	48	70	51	44	63	59	53	65		53	46	48	26	41	45	48	76	76	40	53
1PPH	63	51	66	48	37	78	72	71	59	42		51	49	42	57	64	59	81	69	57	59
1TNG	87	94	92	77	58	108	106	106	102	97	81		105	91	104	106	75	79	90	68	91
1TNH	107	133	84	97	78	129	114	117	118	123	125	134		127	119	121	70	108	99	90	110
1TNI	0	12	2	0	0	0	0	2	0	0	0	25	18		18	15	0	0	0	12	6
1TNJ	49	31	57	53	52	43	36	63	17	30	19	56	53	44		64	23	32	57	23	43
1TNK	91	59	23	54	57	69	68	80	76	69	38	91	81	85	102		35	45	79	25	66
1XUI	22	26	18	27	17	31	39	36	29	21	25	28	30	35	28	29		32	28	32	28
1XUK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		0	0	0
2BZA	97	87	90	99	78	104	102	98	111	102	86	101	109	106	108	113	74	100		101	98
2TBS	50	66	105	112	125	76	85	62	78	63	74	66	52	72	60	61	56	69	62		75
Average	43	41	42	40	35	49	47	45	47	40	38	44	42	43	43	46	36	46	44	38	

Table II Number of Times, out of 1000 Trials, a Minimization from a Random Position Resulted in an Orientation Within 1.5 RMSD of the X-Ray Structure for Each Trypsin Ligand–Protein Pair

	Protein																			
Ligand	1AQ7	1AVW	1CE5	1JRS	1JRT	1MTS	1MTU	1MTV	1MTW	1PPC	1PPH	1TNG	1TNH	1TNI	1TNJ	1TNK	1XUI	1XUK	2BZA	2TBS	Average
1AQ7		2	2	6	4	1	1	0	4	2	3	1	0	1	1	4	0	1	1	0	3
1AVW	1		34	32	30	38	35	40	28	39	31	41	40	46	44	40	37	5	35	29	33
1CE5	56	66		47	49	75	74	69	76	67	71	65	68	80	67	74	71	67	64	90	67
1JRS	44	29	13		22	43	30	16	31	7	17	19	18	14	13	16	22	37	21	0	23
1JRT	38	0	0	15		0	1	0	1	1	0	1	2	1	12	18	0	0	0	0	5
1MTS	0	3	7	11	0		14	5	22	4	2	1	3	2	2	4	1	10	6	0	6
1MTU	15	26	37	15	14	34		22	37	20	32	21	18	21	17	23	17	40	26	0	24
1MTV	22	1	41	2	10	9	11		13	0	0	0	2	0	0	1	0	0	4	0	7
1MTW	8	2	0	1	0	5	5	0		3	2	2	1	3	1	0	1	5	3	8	3
1PPC	56	48	69	51	44	61	58	53	65		53	46	47	26	41	45	48	76	76	38	53
1PPH	63	51	65	48	37	78	70	71	59	41		49	49	42	57	64	58	80	69	49	58
1TNG	54	46	48	32	36	39	41	40	54	51	41		42	37	39	42	34	55	34	67	44
1TNH	53	52	44	43	37	53	49	50	43	48	52	57		60	46	52	44	55	45	77	51
1TNI	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0
1TNJ	14	21	47	53	52	20	17	52	10	23	12	54	53	38		64	20	30	56	19	35
1TNK	40	25	19	35	42	31	34	48	40	34	14	38	43	35	51		20	22	40	20	34
1XUI	21	26	18	27	17	30	39	35	29	21	22	5	30	35	28	27		32	28	32	26
1XUK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0
2BZA	61	34	46	36	25	43	39	37	47	44	39	43	39	43	37	43	41	59		70	43
2TBS	50	64	57	58	63	70	73	56	78	61	71	64	51	67	60	61	56	57	54		64
Average	31	26	29	28	25	33	31	31	33	26	26	28	28	28	28	31	25	32	30	30	

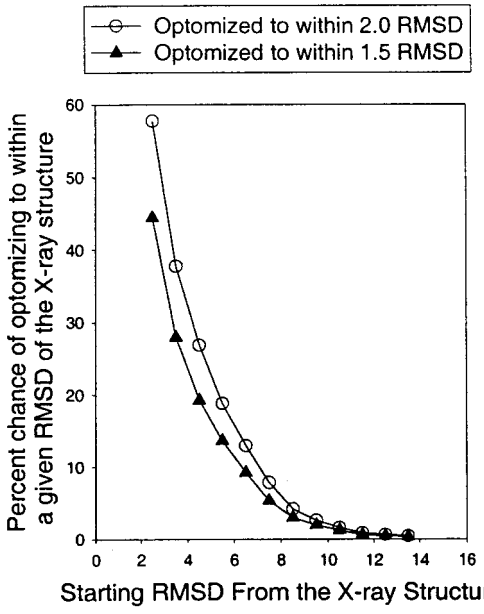

FIGURE 5 Average chance for any trypsin ligand docking successfully into any one of the trypsin structures as a function of how far it starts from the x-ray structure. Success is defined as docking to within with 1.5 or 2.0 RMSD of the x-ray structure.

The chance of the GSF optimizing near the correct structure is almost completely invariant to the particular trypsin structure used, as seen in Tables I and II. This indicates a good tolerance for errors in the protein structure. There is, however, significant variation across the different ligands. Ligands from 1AQ7 and 1XUK fail to dock near the crystal structure at all. The reason for this is seen in the x-ray structures; the ligands lie on top of the protein, not in any pocket. Thus, since the GSF is a shape-based function, it fails when the shape complementarity of ligands and protein is low. The x-ray structure and the highest scoring docked structure of the 1XUK ligand docked into the 1XUK protein are shown in (Figure 6a). The high scoring structure is somewhat similar to the x-ray structure in the S1 pocket; however, the other end of the ligand has been shifted to fit into a nearby pocket with good shape complementarity. This is due to the lack of solvent consideration. As seen in Figure 6b, crystallographic waters and a sulfate group block the pocket, preventing the ligand from fitting down into it. This could be corrected by using a model that includes the waters and phosphate group.

The method was then extended by averaging the GSF around all the trypsin structures. The 20 ligands were then docked using this averaged scoring field as

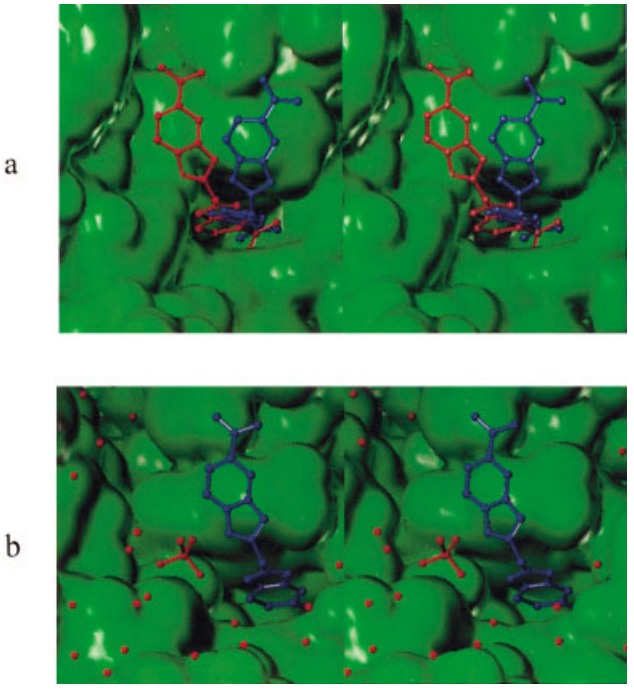

FIGURE 6 (a) The highest scoring (red) docked structure and the x-ray structure (blue) of the complex 1XUK. (b) The original PDB complex 1XUK with all crystallographic waters and the sulfate ion.

Table III Using an Average Interaction Score Over All 20 Trypsin Structures, the Chance That a Randomly Placed Ligand Will Optimize to Within a Given RMSD of the X-Ray Structure

Ligand	Within 2.0 RMSD (%)	Within 1.5 RMSD (%)
1AQ7	0	0
1AVW	5	5
1CE5	7	7
1JRS	5	5
1JRT	4	1
1MTS	2	1
1MTU	5	4
1MTV	2	1
1MTW	2	0
1PPC	6	6
1PPH	8	8
1TNG	13	6
1TNH	14	6
1TNI	0	0
1TNJ	8	7
1TNK	11	6
1XUI	3	2
1XUK	0	0
2BZA	12	5
2TBS	7	7
Overall average	6	4

a severe test of the robustness of a GSF, i.e. can such an averaged field still contain enough information to reproduce crystal-like poses. This is in the spirit of the MCSS where atoms may move in the field of many copies of the protein structure.

The outcome was nearly the same as docking to individual trypsin structures (see Table III). This remarkable result suggests a means to dock to clusters of related structures, for instance, to the kinase family. In addition, docking to inherently noisy structures,

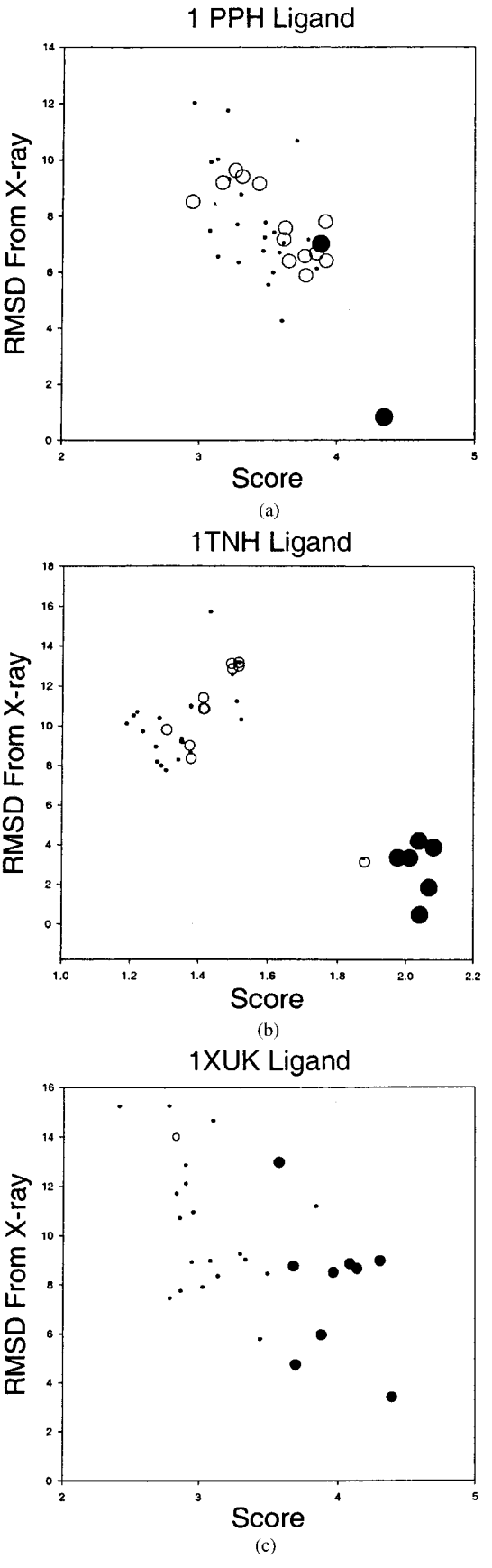

FIGURE 7 Score vs RMSD of significant local minima (over 50% of random optimizations locate one of these minima). Results are for ligands 1PPH, 1TNH, and 1XUK docked into there own enzyme x-ray structure. Dark circles are optima that contain at least 2.5% of the docking runs, open circles are optima that contain between 1.3 and 2.5% of the docking runs, and dots are optima that contain between 0.6 and 1.3% of the docking runs. Optima that contain less than 0.5% of runs are not shown. (a) Results for the ligand of complex 1PPH. (b) Results for the ligand of complex 1TNH. (c) Results for the ligand of complex 1XUK.

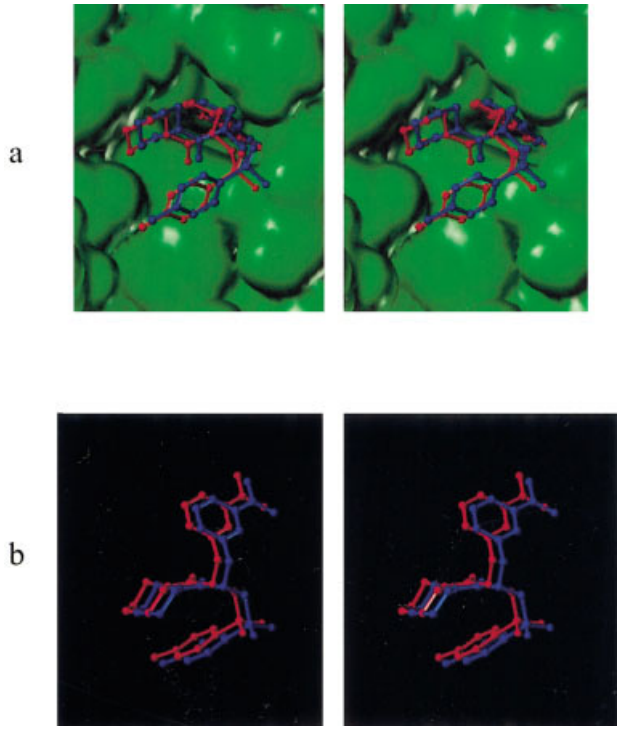

FIGURE 8 The highest scoring (red) docked structure and the x-ray structure (blue) of the complex 1PPH. Panels (a) and (b) are identical other than the presence or absence of the protein surface in the figure and the angle of the view.

such as produced by homology models, may be adequate. Research is ongoing in this area.

The GSF Hypersurface

One of the assumptions of the GSF is that the scoring hypersurface will be much simpler the more error tolerant, i.e., the lower κ. To test this hypothesis, we performed a high-resolution exhaustive scan of the GSF of the 1TNI trypsin complex. The center of mass of the ligand was placed on each point of a grid of 0.75 Å resolution and every 3° rotation about each principle axis was sampled. Local optima were determined when any translation or rotation about a particular pose led to a uniformly less favorable scores. At ($\kappa = 1.0$, $\gamma = 7.5$) there were 45,676 local optima (Figure 7a). By lowering the κ, and adjusting γ to ($\kappa = 0.2$, $\gamma = 1.5$), the number of optima was reduced to 1463 (Figure 7b). Although these exhaustive samplings were time-consuming, the observation that even a low κ, with a very smooth GSF, minima were still found near the x-ray structure has led to the development of an exhaustive search protocol competitive with stochastic methods.[8]

Although the number of optima found might seem large, the number of practical minima, i.e., that might typically be found by minimizing from random starting points, is much smaller. We investigated this by performing 5000 random optimizations we obtained results on the number of "significant" local optima in the GSF's hypersurface vs the κ and γ parameters. A "significant" optima is one that has at least a 1% chance of being located by a docking run. These results averaged over all 20 trypsin complexes and are reported in Figure 7a. The chance of locating any significant local optima is shown in Figure 7b. At high κ there are large number of different local optima; thus there are few local optima that are frequently located (Figure 7a) since most runs fall into optima that are less than 1% likely (Figure 7b). As κ is decreased many small local optima near to each other tend to coalesce into a significant optima. The number of significant local optima increases (Figure 7a) as well as the probability that any significant optima will be located (Figure 7b). Thus lowering the value of κ does reduce the number of critical points on the GSF's hypersurface.

The number of significant local optima (Figure 7a) and the probability of locating any significant local

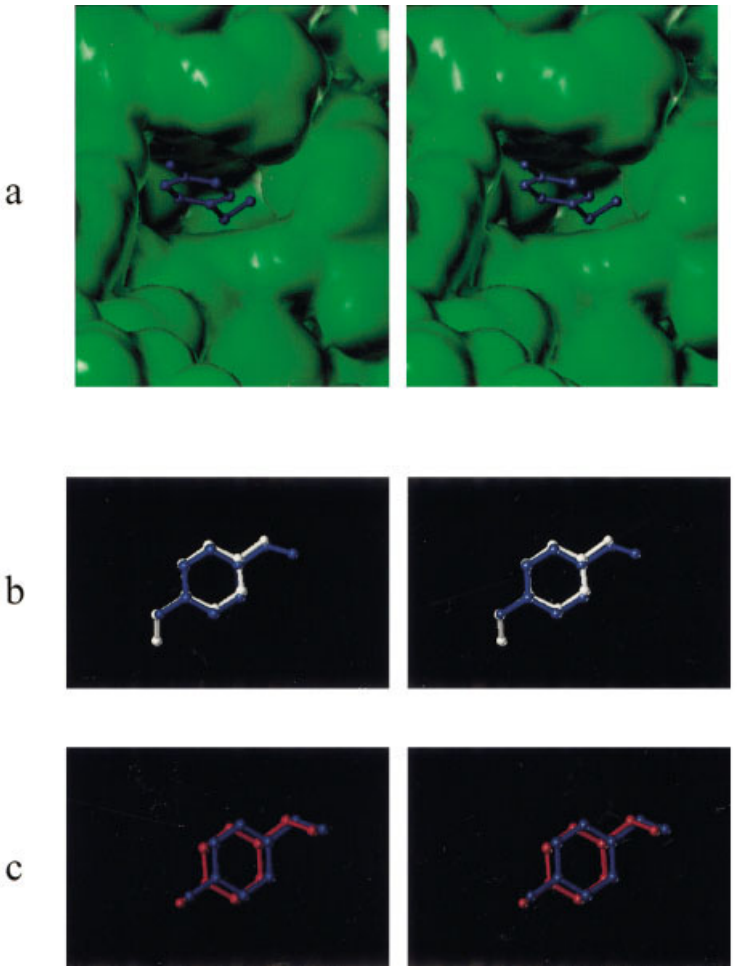

FIGURE 9 (a) The x-ray structure of 1TNH. (b) The x-ray structure of the 1TNH ligand (blue) and the highest scoring (best shape fit) docked structure (white). (c) The x-ray structure of the 1TNH ligand (blue) and docked structure (red) closest to the x-ray (i.e., the structure that should have the best shape and function fit).

optima (Figure 7b) do not depend strongly on the γ parameter. However, changing γ does alter the position of local optima and only particular values of γ will allow the GSF to locate positions near the x-ray structure (Figure 4). At very low γ and κ values many local optima move outside of the scoring grid used. This leads to the number (Figure 7a) and probability (Figure7b) of significant optima changing with γ at $\gamma < 1.0$, which is an artifact of the limits of our scoring grid not a property of the GSF.

We examined the results of 10,000 random dockings to 3 of the aligned trypsin complexes (1PPH, 1TNH, and 1XUK) more closely to determine the probability of finding the crystal pose. When an optima was defined as an RMSD cluster of less than 0.1 RMSD, approximately 1000 unique optima were typically located. However, in all 9 cases over 50% of the docked orientations fell into less than 50 local optima. For the ligands of 1PPH, 1TNH and 1XUK docked into their respective enzyme structures; the scores vs RMSD of any optimum populated more than 0.5% are shown in Figure 8. While the best binding position is usually found, it is often not the optimal GSF. The function is too simple to also discriminate between reasonable orientations. A more comprehensive scoring function should obtain better a RMS correlation.

This said, in the case of the ligand of 1PPH for each of the 3 trypsin structures the highest scoring optimum is also the closest to the x-ray structure (Figure 8a). This optimum for the 1PPH ligand docked into the 1PPH-trypsin and the x-ray structure are shown in Figure 9. In the case of the ligand of 1TNH (Figure 8b), the optimum closest to the

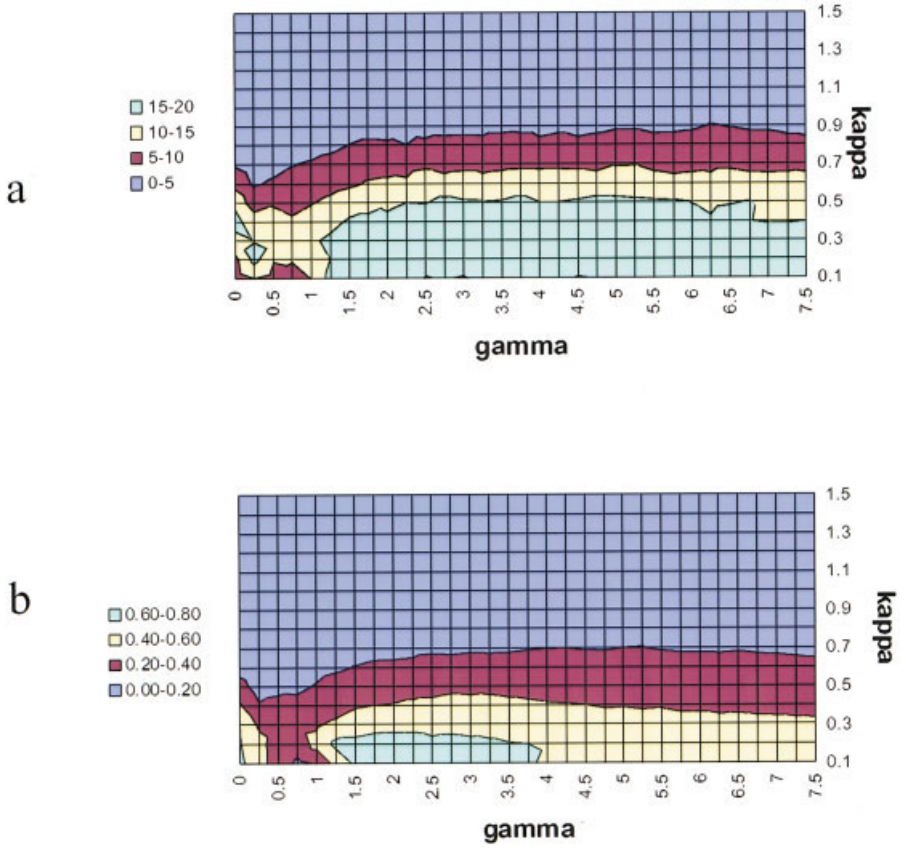

FIGURE 10 (a) Number of significant optima vs the κ and γ parameters. A significant optimum is one that has at least a 1% chance of being located by a docking run. (b) The probability that a docking run will locate a significant optima.

x-ray structure is not the highest scoring. The reason for this is that the ligand is symmetric with the exception of a single tail atom and the highest scoring structure is a 180° rotation that is the same, due to symmetry, as the x-ray structure with the exception of the single tail atom. This is shown in Figure 10. This effect tends to occur more often in smaller ligands, which overall have more positions that complement the receptor site than larger ligands. The ligand from 1XUK interestingly has high-scoring optima that are also the optima closest to the x-ray structure (Figure 8c). However these optima are still poor fits to the x-ray because, as explained earlier, the ligand and receptor shape do not compliment each other well.

Docking of Other Ligand Receptor Complexes

To test targets other than trypsin, a set of 49 ligand–receptor complexes were docked. Each ligand was separated from its protein and docked back 1000 times using (κ = 0.7, γ = 4.5) and the method described at the beginning of this section. The success rates are listed in Table IV and have the same overall success rate (4 and 3% for docking within 2.0 RMSD and 1.5 RMSD of the x-ray structure, respectively) as the trypsin complexes.

Only 6 of the 49 complexes (1cil, 1glp, 1gst, 1hdc, 1icn, and 1tbd) tested have a success rate less than 1%. The x-ray structures of these complexes do score well. However, while these positions score well they are not optimal with respect to the GSF. A physical interpretation is that the x-ray structure fits well, but a nearby position is a better fit, and due to the smoothness of the GSF there is no "energy barrier" between the two positions. Optimization therefore does not locate the x-ray structure. A different type of docking mechanism that looks for good-scoring positions, rather than optima, would be more likely to locate the x-ray structure in these 6 cases.

Table IV Chance for Each Ligand–Receptor That a Solid Body Optimization of a Ligand from a Random Starting Position Will Result in the Ligand Being Within a Given RMSD from the Experimentally Measured Orientation

Ligand	Within 2.0 RMSD	Within 1.5 RMSD
1abe	7	3
1ack	3	3
1aha	1	1
1apt	3	3
1azm	8	7
1bbp	3	3
1bma	1	1
1byb	6	6
1cbs	2	2
1cbx	8	4
1cil	0	0
1cps	2	1
1eap	4	4
1fkg	3	3
1glp	0	0
1gst	0	0
1hdc	0	0
1hfc	2	2
1hri	3	3
1icn	0	0
1ida	5	5
1lah	3	1
1ldm	3	3
1mrk	9	0
1nco	2	2
1nis	7	6
1poc	4	3
1rob	3	3
1sit	1	1
1snc	2	2
1stp	10	9
1tdb	0	0
1tng	6	3
1xid	2	0
2ak3	3	2
2gbp	9	9
2lgs	4	1
2phh	3	2
2r07	2	2
2sim	3	3
3aah	8	8
4mbn	2	1
4phv	7	7
6rsa	4	4
Overall average	4	3

CONCLUSION

A Gaussian docking function has been developed that simplifies the simulation of protein–ligand interactions. Simple docking tests indicated that this function is tolerant of deviations in the crystal structure while still being able to locate ligand positions that closely approximate the crystal structure. Specifically, 20 ligand–receptor complexes of trypsin were separated and then redocked to each of the 20 structures of the trypsin protein. The results of these docking runs showed no significant dependence on which structure of the trypsin protein was used when the ligand was docked. These runs successfully reproduced the x-ray structure for 18 out of 20 ligands. The method also shows utility across the spectrum of protein–ligand complexes.

While the docking function is too simple for reliable determination of the most likely pose of a ligand, it shows remarkable promise in guaranteed sampling of such poses, the hope being that other, more sophisticated, scoring functions may then provide this level of discrimination. Even so, this, and subsequent work, has shown that when an active site has a very specific shape the GSF can actually find the crystal structure as the rank one pose.

Finally, the computational times for docking are not presented here as no attempt was made to optimize the methods. In a subsequent paper we will describe using our GSF in a highly optimized implementation that exhaustively samples search space and yet is still faster than any stochastic method previously reported.

The authors wish to thank Geoff Skillman for useful comments and help with this paper.

REFERENCES

1. Kuntz, I. D. Science 1992, 257, 1078.
2. Åqvist, J.; Medina, C.; Samuelsson, J.-E. Protein Eng 1994, 7, 385.
3. Kollman, P. Chem Rev 1993, 93, 2395.
4. Böhm, H. J. J Comput-Aided Mol Design 1994, 8, 243.
5. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Chem Biol 1995, 2, 317.
6. Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. J Med Chem 1999, 42, 5100.
7. Stahl, M.; Rarey, M. J Med Chem 2001, 44, 1035.
8. Tame, J. F. H. J Comput-Aided Mol Design 1999, 13, 99.

9. Rejto, P. A.; Verkhivker, G. M. Proc Nat Acad Sci USA 1996, 93, 8945.
10. Rejto, P. A.; Bouzida, D.; Verkhivker, G. M. Theor Chem Acc 1999, 101, 138.
11. Diller, D. J.; Verlinde, C. J Comp Chem 1999, 20, 1740.
12. Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. J Comput-Aided Mol Design 2001, 15, 411.
13. Goodsell, D. S.; Olson, A. J. Proteins Struct Funct Genet 1990, 8, 195.
14. Gschwend, D. A.; Kuntz, I. D. J Comput-Aided Mol Design 1996, 10, 123.
15. Jones, G., Willett, P., Glen, R. C. J Mol Biol 1995, 245, 43.
16. Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M. Sandia National Laboriatories 1993, SAND93-8688.
17. Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. J Comput-Aided Mol Design 1995, 9, 113.
18. Payne, A. W. R.; Glen, R. C. J Mol Graph 1993, 11, 74.
19. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. J Mol Biol 1996, 261, 470.
20. Leach, A. R.; Kuntz, I. D. J Comp Chem 1992, 13, 730.
21. Welch, W.; Ruppert, J.; Jain, A. N. Chem Biol 1996, 3, 449.
22. Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. J Comput-Aided Mol Design 1994, 8, 153.
23. Jiang, F.; Kim, S.-H. J Mol Biol 1991, 219, 79.
24. Verkhivker, G. M.; Rejto, P. A.; Gehlhaar, D. K.; Freer, S. T. Proteins Struct Funct Genet 1996, 25, 342.
25. Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. Acc Chem Res 1994, 27, 117.
26. Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. J Comput Chem 1992, 13, 380.
27. Grant, J. A.; Pickup, B. T. J Phys Chem 1995, 99, 3503.
28. Grant, J. A.; Pickup, B. T. Comput Simul Biomol Syst 1997, 7, 150.
29. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F. J.; Brice, M. D.; Rodgers, M. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. J Mol Biol 1977, 112, 535.