J|A|C|S
ARTICLES
Published on Web 02/16/2002

Identification of Ligand Binding Sites on Proteins Using a Multi-Scale Approach

Meir Glick, Daniel D. Robinson, Guy H. Grant, and W. Graham Richards*

Contribution from the Department of Chemistry, Central Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, OX1 3QH, U.K.

Received June 25, 2001. Revised Manuscript Received October 24, 2001

Abstract: Identification of a ligand binding site on a protein is pivotal to drug discovery. To date, no reliable and computationally feasible general approach to this problem has been published. Here we present an automated efficient method for determining binding sites on proteins for potential ligands without any a priori knowledge. Our method is based upon the multiscale concept where we deal with a hierarchy of models generated using a *k*-means clustering algorithm for the potential ligand. This is done in a simple approach whereby a potential ligand is represented by a growing number of feature points. At each increasing level of detail, a pruning of potential binding site is performed. A nonbonding energy function is used to score the interactions between molecules at each step. The technique was successfully employed to seven protein−ligand complexes. In the current paper we show that the algorithm considerably reduces the computational effort required to solve this problem. This approach offers real opportunities for exploiting the large number of structures that will evolve from structural genomics.

Introduction

Genomics, proteomics and bioinformatics are yielding novel therapeutic targets for drug discovery efforts at a rapid rate.[1,2] The genome projects reveal a plethora of new sequences. Their 3D structures will be solved by various techniques such as X-ray crystallography or nuclear magnetic resonance (NMR). Even once detailed structures are known, the design of candidate drugs that may interact with these targets is a difficult task. Novel theoretical approaches could become a major avenue for drug discovery efforts and improve our ability to deal with the abundance of information at the post-genomic era.[3] In a much publicised screen-saver project (http://www.ud.com) we have initiated a massively distributed search among a virtual library of billions of small molecules for compounds that can bind to known protein binding sites. In such circumstances, a matching algorithm such as DOCK, which employs spheres to model the binding site,[4−6] or THINK, which orients the ligand toward chemically favorable interactions zones,[7] can rapidly dock the potential ligand. Here we provide a methodology that addresses the converse question: given a drug candidate molecule and a protein's 3D structure, where will the drug candidate bind?

Let us assume the ligand and host molecules are rigid. Regrettably, a brute force search, where the ligand−host interaction energy is evaluated at all possible docking configurations, cannot be completed in a reasonable amount of computing time, particularly when employing a large protein or ligand. To gain some idea of the runtime requirements we might consider that a typical protein host might occupy a volume of some 60 Å^3. Even with a moderate translational resolution of 1 Å, this leaves 216 000 translations to search. Given a reasonable rotational resolution of 20° in each axis, and given that a potential ligand and protein might contain 35 and 3500 atoms, respectively, 1.5×10^{14} pairwise nonbonding energy evaluations will be needed to scan the complete range of possible docking configurations. Even if one can evict 99% of the points by employing various assumptions, we will still require 1.5×10^{12} evaluations. As a result, brute force approaches such as GRID[8,9] are limited to small probe groups and cannot handle a detailed ligand in a reasonable computing time. Clearly, one of the major challenges of ligand−host docking is of reducing the number of energy evaluations that need to be performed in order to locate the optimum binding position.

A different approach is based on a docking simulation in a given potential force field via a global optimization algorithm that minimizes the ligand protein interaction energy: methods such as Monte Carlo simulated annealing,[10−12] genetic algorithms,[13−16] or the multiple copy simultaneous search (MCSS)

* Corresponding author: graham.richards@chem.ox.ac.uk.

(1) Drews, J. *Science* **2000**, *287*, 1960−1964.
(2) Service, R. F. *Science* **2000**, *287*, 1954−1956.
(3) Brenner, S. E.; Levitt M. *Protein Sci.* **2000**, *9*, 197−200.
(4) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol.* **1982**, *161*, 269−288.
(5) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 505−524.
(6) Shoichet, B. K.; Kuntz, I. D. *Prot. Eng.* **1993,** *6*, 723−732.
(7) Murray, M.; Cato, S. J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 46−50.
(8) Wade, R. C.; Goodford P. J. *J. Med. Chem.* **1993**, *36*, 140−147.
(9) Wade, R. C, Goodford P. J. *J. Med. Chem.* **1993**, *36*, 148−156.
(10) Goodsell, D. S.; Olson, A. J. *Proteins: Struct. Funct. Genet.* **1990**, *8*, 195−202.
(11) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(12) Kirkpatrick, S.; Gelatt, C. D., Jr; Vecchi, M. P. *Science* **1983**, *220*, 671−680.
(13) Jones, G.; Willet, P.; Glen, R. C. *J. Mol. Biol.* **1995**, *254*, 43−53.
(14) Jones, G.; Willet, P.; Glen, R. C.; Leach, A. R. *J. Mol. Biol.* **1997**, *267*, 727−748.
(15) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293−304.

10.1021/ja016490s CCC: $22.00

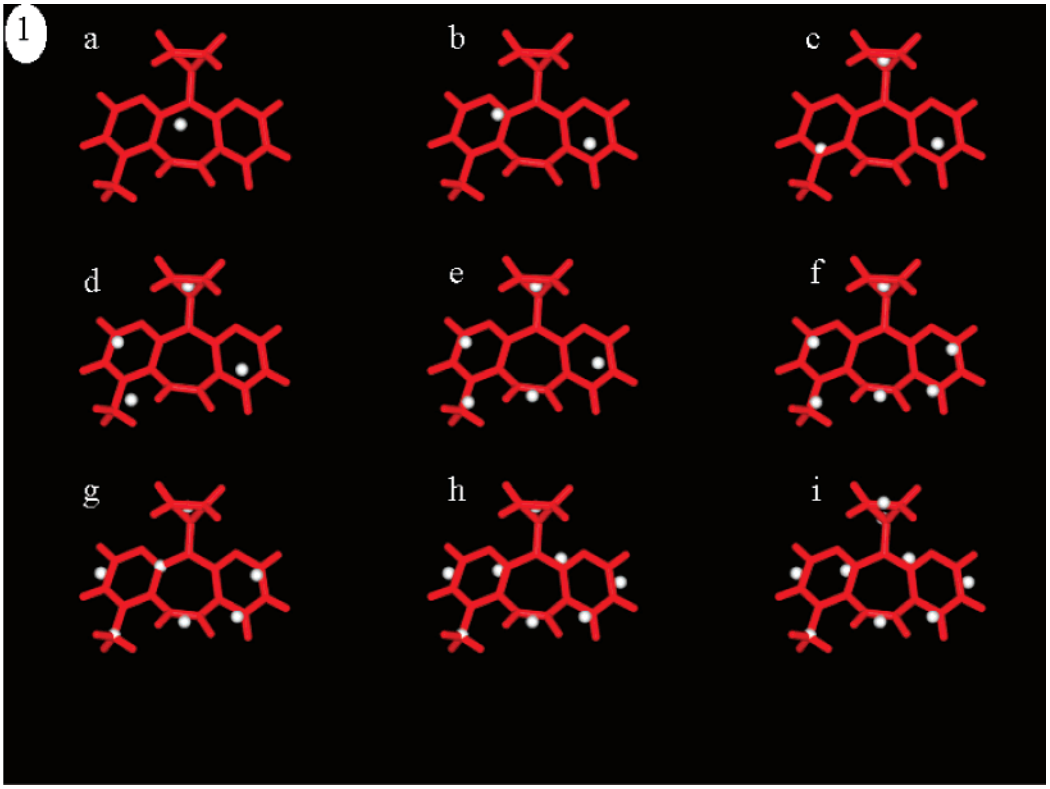

Figure 1. Example of models generated for the HIV-reverse transcriptase inhibitor nevirapine.

method[17] have proved successful in this respect. In general, these methodologies give reasonable results provided the initial position of the ligand is within, or proximate to, the binding site of the host molecule. By starting the simulation in a good position, we can restrict the number of docking configurations that need to be considered and avoid many of the local entrapments. Failure to assist the optimizer in this manner will result in an extremely long calculation that might fail to detect the global minimum.

In the face of such optimization difficulties, we devised a different approach to solve the binding site location problem. In a brute force approach, numerous operations are redundant since in many configurations the ligand atoms will be either too distant or close to the protein atoms. This work relies on the hypothesis that we can effectively evict these configurations at an early stage by employing a filtering operation thus saving a substantial amount of computing time. This strategy is adapted from methods developed in the fields of signal and image processing: the so-called *multiscale approach*.[18,19] Applying this approach enables us to return to a simpler, rapid brute force style search, removing the need for an elaborate optimization protocol.

Methods

The *multiscale approach* relies on a construct known as a *scale−space decomposition*.[18,19] This involves the application of a *scaling operator* to the original data. The effect of the scaling operator is to remove the information in our data corresponding to the highest level of detail. With the fine detail removed, the larger scale features in the data are emphasized. Repeated application of the scaling operator returns larger and larger structures in our data until either the required scale has been achieved or there is no more information left in the resulting decomposition. The selection of the scaling operator is of crucial importance if multiscale analysis via *scale−space decomposition* is to be completed successfully.

We model the ligand at various scales by employing a simple yet powerful method: the *k-means-clustering algorithm*.[20,21] By using a series of *k-means* generated clusters, we obtain series of ligand models, each containing one more feature point than the previous model. These feature points are well distributed to ensure that each model yields the best possible description of the ligand for the number of points generated. In our *k-means* procedure implementation, n atoms x_1, x_2, ..., x_n fall into k clusters, $k < n$. Let m_i be the mean position of the atomic coordinates in cluster i. If the clusters are well separated, we can use a minimum-distance classifier to separate them: atom x is in cluster i if the distance between x and m_i is the minimum of all the k distances. Since there is no definite way to initialize the mean values, it is common to make initial guesses for the means m_1, m_2, ..., m_k. At the next stage, until there are no changes in any mean, we use the estimated means to classify the atoms into clusters: for all clusters, replace m_i with the mean position of all of the atoms for cluster i. Figure 1 depicts *k means* for an increasing number of feature points (from 1 to 9) on the HIV-reverse transcriptase inhibitor, nevirapine.[22] The initial cluster is at the mean position of the ligand: all of the atoms in the ligand belong to this initial cluster (Figure 1a). To initialize the means for the second cluster, we search for the atom that is furthest away from the initial cluster. This atom then becomes the temporary center of the new cluster. All atoms that are closer to this cluster center than

(16) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(17) Miranker, A.; Karplus, M., *Proteins: Strucy. Funct. Genet.* **1991**, *11*, 29−34.
(18) Forgy, E. *Biometrics* **1965**, *21*, 768−769.
(19) MacQueen, J. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*; Le Cam, L. E., Neyman, J., Eds.; University of California Press: Berkeley, CA, 1967; Volume I, Statistics.
(20) Hartigan, J. *Clustering Algorithms*; Wiley: New York, 1975**,** 84−112.
(21) Rollet, R.; Benie, G. B.; Li, W.; Wang, S.; Boucher, J. M. *Int. J. Remote Sens.* **1998**, *19*, 3003−3009.
(22) Ren, J.; Esnouf, R.; Garman, E.; Somers, D.; Kirby, C. R. I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D. I.; Stammers, D. *Nature Struct. Biol.* **1995**, *2*, 293−302.

their currently assigned cluster center change identities and are marked as belonging to the new center. The position of the cluster centers are then iterated upon to self-consistency so that each cluster center is positioned at the average position of the atoms that belong to the cluster (Figure 1b). This process may be repeated as many times as there are atoms in the ligand, with each iteration generating the next model (Figure 1c−i).

A rapid, grid-based method is employed for energy evaluation[8,9] by precalculating ligand−protein pairwise interaction energies to form a lookup table. Energies are linearly interpolated from the grid. The energy is computed by eq 1 with the Consistent Valence Force Field (CVFF)[23] all-atom model nonbonding 12−6 Lennard-Jones and electrostatic terms. $A_{i,j}$ is the repulsion parameter for the two (i,j) atoms, $B_{i,j}$ is their attractive polarizability parameter, q_i is the partial charge, and ϵ is the dielectric constant. Atom i belongs to a feature point k, and j is a protein grid point. It should be noted that the atoms' sum in all k feature points is identical to the number of atoms in the ligand. The energy is calculated between all i atoms in feature point k and each protein atom (j). The process is repeated for all k feature points. The atom types in feature point k are different but their distances to protein's atoms j are identical since they are calculated from feature point k.

$$E_{\text{pot}} = \sum_k \sum_{i,j} \left(\frac{A_{i,j}}{r_{k,j}^{12}} - \frac{B_{i,j}}{r_{k,j}^{6}} + \frac{q_i q_j}{\epsilon r_{k,j}} \right) \quad (1)$$

An array of ligand models is generated in the manner described above. The first model, a single point, is then tested at all configurations in the host molecular field. Because a single point cannot be rotated, the test is extremely simple. All configurations whose energy is lower than a given threshold are marked to be kept for the next round of calculation. In the current implementation, configurations that do not demonstrate any avidity toward their binding site, i.e., their interaction energy is not negative, are evicted from subsequent iterations. Among the positions that are rejected from the next round are those that are either too far away or too close to the host molecule to be considered as likely docking configurations. Further, this simple test distinguishes between potentially good (negative interaction binding energy) or poor ligands. As more feature points are added the more implicit this classification becomes. At each configuration that survived the first round we test the second model. The second model is formed from two points separated by a certain distance that is related to the dimensions of the "main axis" of the ligand. The orientation of the model is now important, and consequently each configuration has to consider all possible rotations of the model. However, because the model only contains only two points, the rotation is much faster than attempting to rotate the complete ligand. Once more, all configurations whose nonbonding energy is lower than a given threshold of the two models are kept for the next round. It is clear that this second model begins to filter out any configurations where the potential binding site is not large enough to hold a molecule as large as the ligand. Repeating these steps for the subsequent models removes more of the unsuitable configurations. Eventually, we will end up with only a few of the configurations surviving, usually long before we run out of models to test. These surviving configurations show the regions where it is possible for the ligand to dock successfully and the translation of the docked ligand.

Results

To test the methodology a number of host molecules complexed with ligands (Figure 2) were downloaded from the Protein Data Bank (PDB).[24] The ligands were deleted, and an attempt was then made to find the correct docking site of the ligands. In all test cases, we placed the host protein in a box of dimensions 3 Å greater in each direction than the extent of the protein. We employed a molecular grid with a 0.7 Å resolution, and a rotation angle of 5°. A distance dependent dielectric constant of $\epsilon = 4r$ was used.

Streptavidin/Biotin. We utilized the streptavidin complex with biotin[25] (PDB entry 1stp; resolution 2.6 Å). The search results are shown in Table 1 and Figure 3a. The distance between the centroid of biotin in the crystal structure and the centroid of its predicted conformation is 1.15 Å. The multiple hydrogen bonds to N23, S45, N49, and D128 that are among the factors that allow a tight binding to the protein[16] (shown by a dashed line) are persevered in our predicted conformation. The quality of the *population* of results was evaluated by calculating the distance between the centroid of the five lowest energy conformations and the centroid of the ligand in the crystal structure, which is 0.93 Å in this case. Conjugate gradients local optimization was then employed on the complex with biotin in its predicted position. The protein's atoms were held fixed, and the ligand was allowed to move. The calculation converged after 356 iterations (12 s on R10000 single processor) when a convergence criterion of 0.01 kcal/Å had been achieved. The distance between the centroid of biotin in the crystal structure and the centroid of its predicted conformation after the minimization was reduced to 0.37 Å.

McPC-603/Phosphocholine. The immunoglobulin McPC603 Fab−phosphocholine complex[26] was retrieved as PDB entry 2mcp. Since the force field was unable to assign reliable partial charges to the phosphocholine ligand, its partial charges were calculated with the Gaussian 98 program[27] (Revision A.7) using a Hartree−Fock calculation with the STO-3G basis set. The distance between the centroid of phosphocholine in the crystal structure and the centroid of its predicted conformation is 2.02 Å. The phosphocholine recognition by McPC-603 is predominantly electrostatic in character, primarily due to the influence of Arg H52.[28] As can be seen from Figure 3b, the distance between the Nη1 in the positively charged side chain of Arg H52 and the P atom in the negatively charged phosphocholine moiety is 3.12 Å in the crystal structure, while in our predicted conformation this distance is 4.66 Å. The distance between the centroid of the 5 lowest energy conformations and the centroid of the ligand in the crystal structure is 1.70 Å. Employing the same minimization protocol as in the streptavidin/biotin test case, convergence was achieved after 159 iterations (17 s on R10000 single processor). The distance between the centroid

(23) MSI, San Diego, CA.

(24) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(25) Weber, P. C.; Ohlendorf, D. H.; Wendolski, J. J.; Salemme, F. R., *Science* **1989**, *243*, 85−88.
(26) Padlan, E. A.; Cohen, G. H.; Davies, D. R. *Ann. Immunol. Paris Sect. C* **1985**, *136*, 271−276.
(27) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian, Inc., Pittsburgh, PA, 1998.
(28) Novotny, J.; Bruccoleri, R. E.; Saul, F. A. *Biochemistry* **1989**, *28*, 4735−4749.

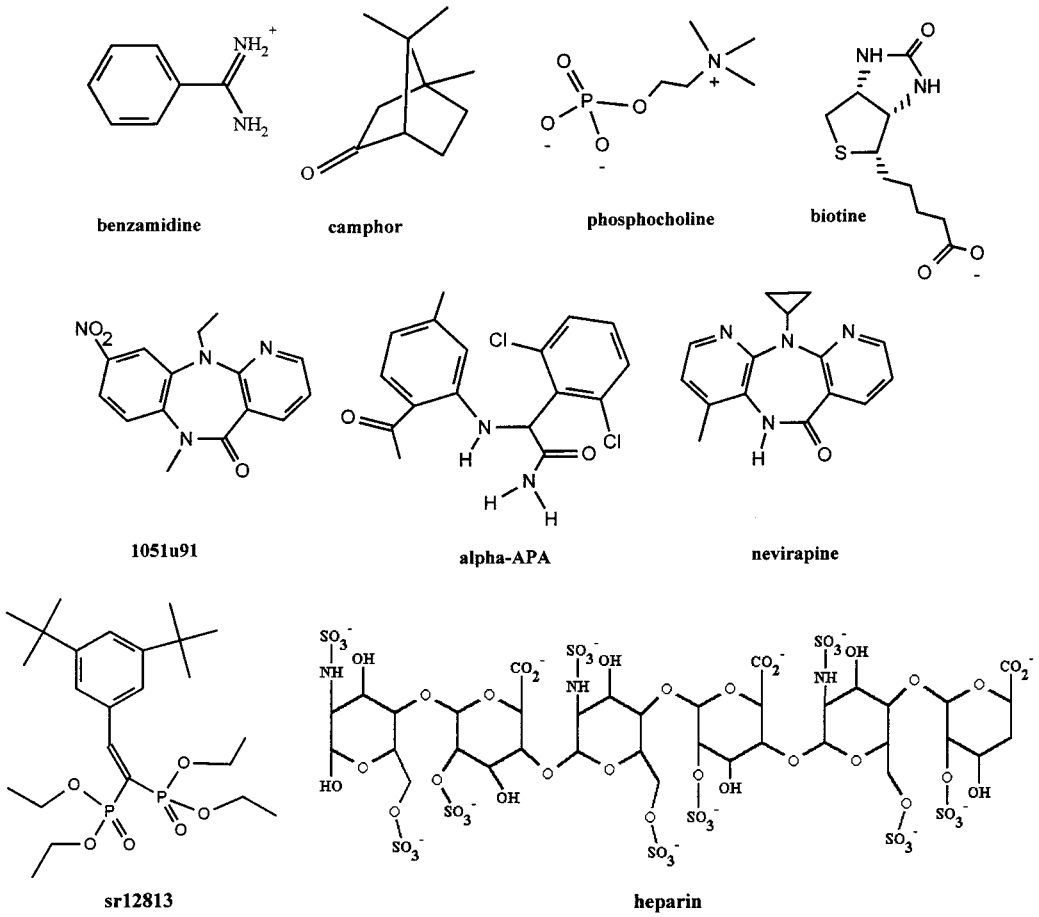

Figure 2. Chemical formulas for the nine ligands selected as test cases.

Table 1. Results for the Protein−Ligand Complexes

protein	PDB code	resolution (Å)	ligand	no. of atoms in ligand[a]	initial no. of translations	distance[b] between ligand centroid in the crystal and: ligand centroid in the lowest energy cofnormation	distance[b] between ligand centroid in the crystal and: centroid of the five low energy conformations	distance[b] between ligand centroid in the crystal and: ligand centroid in the lowest energy conformation followed by a local minimization[c]
streptavidin	1stp	2.6	biotin	31	257 816	1.15	0.93	0.37
McPC-603	2mcp	3.1	phosphocholine	24	984 528	2.02	1.70	1.90
β-trypsin	3ptb	1.7	benzamidine	18	353 920	1.68	1.59	0.38
cytochrome P-450$_{cam}$	2cpp	1.63	camphor	27	772 500	0.62	0.85	0.54
HIV-reverse transcriptase	1vrt	2.2	nevirapine	34	2 666 664	1.15	0.95	0.32
	1rt3	3.0	1051U91	36	2 666 664	1.62	0.90	
	1vru	2.4	α-APA	30	2 666 664	1.20	1.49	
Human nuclear pregnane X receptor	1ilg	2.5						
	1ilh	2.75	SR12813[e]	75	709 152	1.20		0.65

[a] Including hydrogens. [b] Distances are given in Å. [c] Conjugate gradients minimization until convergence criteria of 0.01 kcal/Å is achieved where the protein's atoms are held fixed. [d] Ligand was docked to the apo-structure (1ilg) and compared to the complexed one (1ilh). [e] SR12813 can bind in three distinct orientations. Results are shown for the first one.

of phosphocholine in the crystal structure and the centroid of its predicted conformation after the minimization was 1.90 Å.

***β*-Trypsin/Benzamidine.** The benzamidine-inhibited β-trypsin[29] was downloaded as PDB entry 3ptb. The distance between the centroid of biotin in the crystal structure and the centroid of its predicted conformation is 1.68 Å. As can be seen from Figure 3c, the aromatic ring clearly fills the hydrophobic binding pocket. The distance between the centroid of the five lowest energy conformations and the centroid of the ligand in the crystal structure is 1.59 Å. Again, the ligand was minimized and converged after 894 iterations (25 s on R10000 single processor) The distance between the centroid of benzamidine in the crystal structure and the centroid of its predicted conformation after the minimization was reduced to 0.38 Å.

Cytochrome P-450$_{cam}$/Camphor. The PDB entry 2cpp[30] contains a protoporphyrin group with Fe^{3+} and camphor.

(29) Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. *Acta Crystallogr., Sect. B* **1983**, *39*, 480−490.

(30) Poulos, T. L.; Finzel, B. C.; Howard, A. J. *J. Mol. Biol.* **1987**, *195*, 687−700.

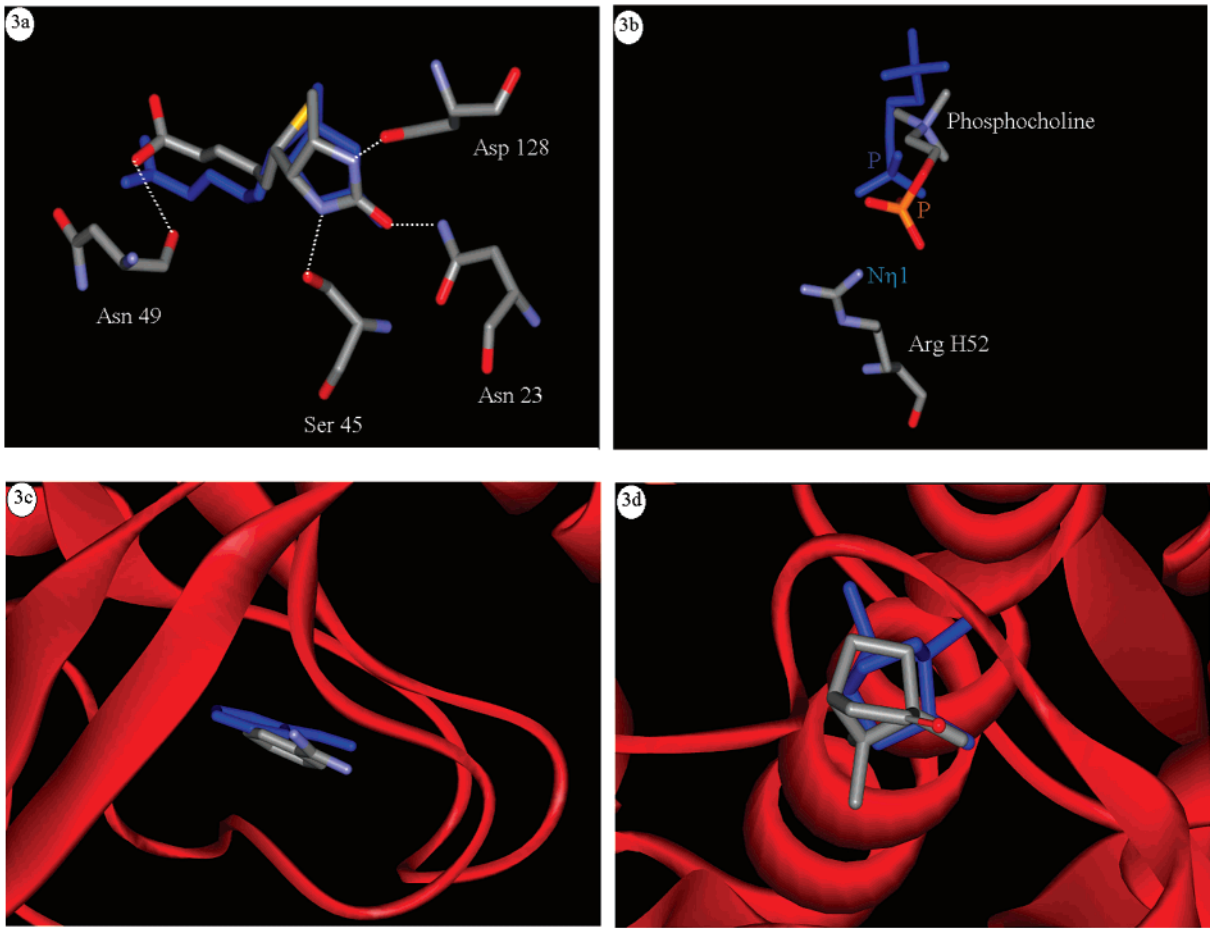

Figure 3. Comparison between the predicted (shown in blue) and the crystal structure conformations (color codes: O = red, C = gray, N = blue, S = yellow, P = orange) of the ligand. Protein shown as a red ribbon. Key: (a) streptavidin/biotin, with H-bonds shown by dashed lines; (b) McPC-603/phosphocholine results; (c) β-trypsin/benzamidine; (d) cytochrome P-450$_{cam}$/camphor.

Because the force field did not contain proper partial charges for the protoporphyrin group with Fe^{3+}, the partial charges for this moiety were calculated as before. The distance between the centroid of camphor in the crystal structure and the centroid of its predicted conformation is 0.62 Å. The distance between the centroid of the five lowest energy grid points and the centroid of the ligand in the crystal structure is 0.85 Å. Figure 3d compares the low-energy solution created by the algorithm to camphor in the crystal structure. The algorithm successfully detected the translation of the camphor, and the correct orientation of camphor was included in the low-energy population but not ranked as the one with the lowest energy. We employed the same minimization protocol as in the streptavidin/biotin test case. The minimization converged after 677 iterations (55 s on R10000 single processor). The distance between the centroid of benzamidine in the crystal structure and the centroid of its predicted conformation after the minimization was reduced to 0.54 Å.

HIV-Reverse Transcriptase/Nonnucleoside Inhibitors (NNIs). We utilized the PDB file 1vrt[22] (resolution 2.2 Å) of HIV-reverse transcriptase as a host molecule for this test case and attempted to dock three non nucleoside inhibitors (NNIs). The first inhibitor was nevirapine, which is complexed with the protein in this PDB entry. The second inhibitor was 1051U91 taken from the AZT resistant HIV-1 reverse transcriptase complex[31] (PDB entry 1rt3), which included the mutations D67→N, K70→R, T215→F, and K219→Q. The root-mean-square (rms) deviation of the active sites (P95, L100, K101, V106, E138, V179, Y181, Y188, G190, F227, W229, L234, H235, and Y318, a total number of 130 "heavy" atoms) in these two structures is 1.69 Å. The high rms value is due to Y181, which adopts a different rotamer in these structures. The third inhibitor was α-anilino phenyl acetamide (α-APA) taken from the PDB entry 1vru[22] (resolution 2.4 Å). The rms of this active site's residues to the active site of 1vrt was 0.72 Å.

Three docking stages for nevirapine are shown in Figure 4. The 20 000 lowest energy positions at the end of the first iteration are shown in Figure 4a. It can be seen that this population shows points either not too distant or too close to the protein. Figure 4b shows the points remaining after the third iteration. It can be seen that the remaining points converge to various pockets in the protein. Figure 4c shows the remaining points at the last iteration. Clearly, all points are focused in the actual binding site. The distances between the centroid of the ligand in the 1vrt crystal structure and the predicted conformation's centroid with the lowest energy are 1.15, 1.62, and 1.20 Å for nevirapine, 1051U91, and α-APA, respectively. For the five lowest energy grid points' centroids, this distance is 0.95, 0.90, and 1.49 Å for nevirapine, 1051U91, and α-APA,

(31) Ren, J.; Esnouf, R. M.; Hopkins, A. L.; Jones, E. Y.; Kirby, I.; Keeling, J.; Ross, C. K.; Larder, B. A.; Stuart, D. I.; Stammers, D. K. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9518−9523.

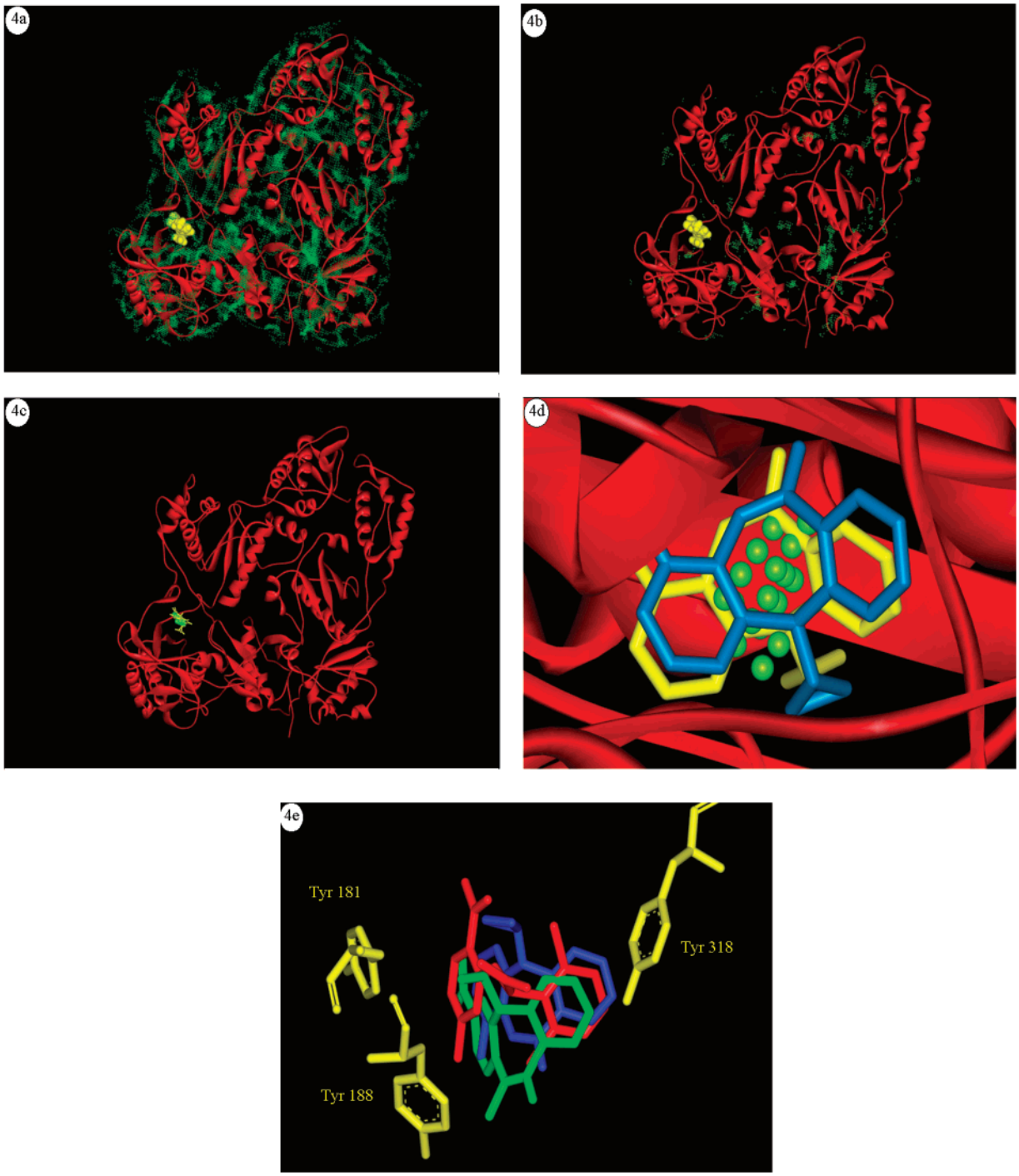

Figure 4. Search results on HIV-reverse transcriptase. Lowest energy grid points (shown in green) (up to 20 000 points). Protein shown as a red ribbon. Key: (a) first iteration for nevirapine; (b) third iteration for nevirapine; (c) last iteration for nevirapine; (d) conformation of the predicted nevirapine conformation shown in blue and conformation of the ligand in the crystal structure shown in yellow. (e) The interactions between various HIV-reverse transcriptase inhibitors in their predicted position and the active site's residues. Nevirapine is shown in blue, 1051U91 is shown in green, α-APA is shown in red, and active site residues are shown in yellow. Hydrogens are not shown.

respectively. Analysis of the interactions between the active site residues and the predicted positions of the inhibitors agrees with the results published by Ren et al.[22] showing that all three inhibitors make extensive contacts (distance between "heavy atoms" <4.0 Å) with Y181, 188 and Y318 as can be seen in Figure 4e (6 atoms for nevirapine, 10 atoms for 1051U91, 8 atoms for α-APA). We employed the same minimization protocol as in the streptavidin/biotin test case. The minimization converged after 448 iterations (157 s on R10000 single processor). The distance between the centroid of nevirapine in the crystal structure and the centroid of its predicted conformation after the minimization was reduced to 0.32 Å.

Human Nuclear Pregnane X Receptor (hPXR)−SR12813. The human nuclear pregnane X receptor (hPXR) plays a critical role in mediating dangerous drug−drug interactions.[32] Watkins

(32) Watkins, R. E.; Wisely, G. B.; Moore, L. B.; Collins, J. L.; Lambert, M. H.; Williams, S. P.; Willson, T. M.; Kliewer, S. A.; Redinbo, M. R. *Science* **2001**, *292*, 2329−2333.

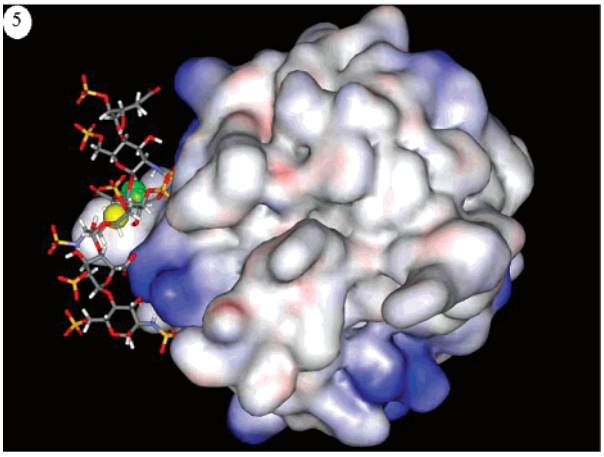

Figure 5. Search results on heparin (shown as a stick model)-basic fibroblast growth factor (bFGF). The surface of bFGF is shaded from red (negative electrostatic potential) to blue (positive electrostatic potential). Centroids are shown as spheres (yellow = ligand in crystal structure, green = five low-energy conformations).

et al.[32] have recently solved the apo structure of the ligand-binding domain of hPXR and the complex with the cholesterol-lowering drug SR12813 at resolutions of 2.5 Å (pdb file 1ilg) and 2.75 Å (1ilh), respectively. Strikingly, SR12813 can bind in three distinct orientations.[32] The distance between the centroids of the orientations ranges between 1.35 and 2.44 Å. The RMS deviation between the binding site (L206, S208, L209, V211, L240, M243, M246, S247, F251, F281, C284, Q285, F288, W299, M323, L324, H407, F410, F420, a total number of 338 "heavy" atoms) in the apo and ligand-bound forms is 1.00 Å. To test if the algorithm can handle protein plasticity, or a situation were the ligand is not well adapted, we employed our program on the apo structure and compared the binding results to the complexed one. The distances between the centroids of the three SR12813 orientations in the complex crystal structure and the centroid of its predicted conformation were 1.20, 2.05, and 2.91 Å respectively. Subsequent 701 iterations using the above minimization protocol (147 s on R10000 single processor) reduced the distances to 0.65, 1.11, and 1.82 Å, respectively.

Heparin-Basic Fibroblast Growth Factor (bFGF). This differs from the above test cases that involved small ligands, since heparin is a biomolecule (extents of 24.05 Å × 13.68 Å × 13.08 Å, 152 atoms) and we were interested in a qualitative result. We utilized the pdb file 1bfc (2.2 Å resolution) containing heparin hexamer fragment−basic fibroblast growth factor (bFGF) complex.[33] Unlike most complexes that fit the standard "lock and key" paradigm, with the ligand and host exhibiting a high degree of surface complementarity, the heparin-binding site on bFGF is sterically ill defined.[34] Instead it appears as if the ligand is held in place due to electrostatic interactions between the highly negatively charged heparin and its binding site. The calculation started with an initial number of 219 356 translations. The binding site was clearly identified (Figure 5).

Discussion

We have devised a tool that is able to locate reliably the binding site for a specified host−ligand pair. As demonstrated by the results, we have been extremely successful in this aim. In all quantitative test cases, except McPC-603/phosphocholine, the distance between the centroid of the ligand in the crystal structure and the centroid of its predicted conformation ranged from 0.62 to 1.68 Å. Bearing in mind that the resolution of the McPC-603/phosphocholine complex (2mcp) is only 3.1 Å, this may hint that a distance of 2.02 Å is a sensible result. In the case of hPXR/SR12813 the algorithm showed the best result for the first orientation (1.20 Å) of the ligand. In addition it showed reasonable results for the second (2.05 Å) and third (2.91 Å) orientations as well.

Further we have shown that not only the lowest energy solution but the *population* of the low-energy solutions is very accurate. In all test cases the distance between the centroid of the ligand in the crystal structure and the centroid of its predicted conformation *population* ranged from 0.85 to 1.70 Å. On average, the *population* results were better than the lowest energy results. This fact suggests that the algorithm is consistent and converges to one site−the binding site.

The McPC-603/phosphocholine and HIV-reverse transcriptase/NNIs experiments represent a particularly harsh test of the technique. The phosphocholine ligand is extremely small (11 "heavy" atoms) relative to the size of the McPC-603 (442 residues, 3401 "heavy" atoms). As can be seen from the crystal structure data (PDB file 1vrt) and Figure 4, the HIV-reverse transcriptase host molecule is an extremely large molecule (926 amino acids, 7625 "heavy" atoms) compared to the three NNIs that we attempted to dock: 20, 22, and 21 "heavy" atoms for nevirapine, 1051U91, and α-APA, respectively. Because of this disparity in size it is likely that the host will present many pockets into which the inhibitor could dock, in addition to the true binding site. α-APA and 1051U91 are difficult test cases since they are taken from a different conformation of the active site and their bioactive conformation, despite their relative rigidity, might vary to a certain degree among complexes. Further, unlike other test cases such as McPC-603/phosphocholine, the interactions between the host molecules and the ligand are hydrophobic and not electrostatic. Although we are comparing three different ligands, the small distances between the centroids in the predicted and the crystal structure demonstrate that the active site was successfully detected.

In addition to the translation, the orientation of the ligand in the binding site was found in all test cases. In six out of eight, it has been ranked as the one with the lowest energy. The first exception was hPXR/SR12813, which is an ambiguous test case, since there are three "correct" orientations. In camphor, the correct rotation was included in the low-energy set of solutions but not ranked as the one with the lowest energy. These findings raise the idea that employing more sophisticated cost functions than CVFF nonbonding energy terms, such as an empirical free energy function that estimates the free energy change upon binding[16] or a finite difference Poisson−Boltzmann method to represent the electrostatic properties of the molecules may yield more accurate energies. In the current implementation we employ a linear interpolation approach to obtain the energies from the grid. Using a more sophisticated interpolation strategy may further improve the results.

The current implementation of the *multiscale approach* is conducted on a discrete search space. As a result, the accuracy is limited both by the resolution of the grid and the size of the

(33) Faham, S.; Hileman, R. E.; Fromm, J. R.; Linhardt, R. J.; Rees, D. C. *Science* **1996**, *271*, 1116−1120.
(34) Bitomsky, W.; Wade, R. C. *J. Am. Chem. Soc.* **1999**, *121*, 3004−3013.

rotation angle. We have shown that our method produces an excellent starting point for rapid local optimization employed on a continues search space using the same energy function. As demonstrated by the results, in all test cases, except McPC-603/phosphocholine, the distance between the centroid of the ligand in the crystal structure and the centroid of its predicted conformation after a short optimization ranged from 0.32 to 0.65 Å. Again, since the resolution of the McPC-603/phosphocholine complex (2mcp) is only 3.1 Å, we find a distance of 1.90 Å as a reasonable result. We have shown that the minimization converges rapidly. Such results hint that the discrete search method we employ locates the ligand in close proximity to its "real" position and is reliable.

Most docking algorithms employ as test cases fixed structures of enzyme and/or ligand as taken from the enzyme-ligand complexes (PDB files). Here, one may raise the claim that there is a bias toward locating the true enzyme-ligand complex when scanning the interaction space. We have shown in two different test cases that the algorithm can cope with protein plasticity. In the HIV-reverse transcriptase/NNIs test case we have shown that we can successfully dock ligands taken from other crystal structures in a different bioactive conformation. In the hPXR-SR12813 system we successfully identified the binding site in the complexed protein where the input for our algorithm was the protein in its unbound conformation.

Our approach offers several advantages over algorithms for detection of pockets on the surface of proteins such as LIGSITE.[35] We have shown that instead of suggesting several binding pockets that exist in a large protein such as the HIV-reverse transcriptase, the algorithm is sensitive enough to detect the correct one. Further, it is sensitive enough to position the ligand in the correct binding orientation. The algorithm performed well in the case of heparin-bFGF compelx where there is a large ligand and no binding pocket.

Currently no attempt has been made to account for the flexibility of the ligand explicitly. There are several reasons for this: first, one must generate a series of likely conformers by selecting a few low-energy conformations from a molecular dynamics or Monte Carlo simulation of the ligand. Second, the earlier models contain only the grossest structural information of the ligand. Such coarse information is somewhat robust to subtle changes in ligand conformation. It is felt that this resilience will bestow the basic docking algorithm with a degree of invariance to the ligand's conformation, as hinted in the HIV-reverse transcriptase/NNIs and hPXR-SR12813 test cases. The extent of the model's invariance to ligand flexibility and how such flexibility could be readily incorporated into the process would require further examination. Currently very few docking procedures can take into account the flexibility of the host molecule as a whole. Some techniques consider the residues near the binding site to be flexible on the presumption that the gross structure of the host will remain unaffected by the ligand's presence. Once more, it is felt that the coarseness of the lower models may assist in providing a certain degree of invariance to the conformation of the side chains in or near the active site. Further work would be required to prove this point.

(35) Hendlich, M.; Rippmann, F.; Barnickel, G. *J. Mol. Graph. Model.* **1997**, *389*, 359−363.

Conclusions

We have shown how the use of a *multiscale approach* enables one efficiently to break a problem down into a number of small steps. Dismantling a problem in this manner enables efficient distribution of computing time so that only the most fruitful areas are considered in any detail. We believe that the approach introduced here has real value in exploiting the results emerging from studies in structural biology and may evolve to become a valuable tool to analyze the data from the structural genomics projects.

The approach seems to hold great promise for the future—in this initial model the interactions between the enzyme and the substrate are evaluated via simple nonbonding energy terms. Despite that, the results are accurate, and the calculations converge rapidly. This fact supports the ability of the *multiscale approach* to represent the ligand in a reliable manner.

Acknowledgment. This work was supported by the Wellcome Trust and partially by the National Foundation for Cancer Research.

JA016490S