ELSEVIER

Journal of Molecular Graphics and Modelling 21 (2003) 289–307

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites

C.M. Venkatachalam*, X. Jiang, T. Oldfield, M. Waldman

Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121, USA

Received 25 April 2002; received in revised form 11 July 2002; accepted 9 August 2002

Abstract

We present a new shape-based method, LigandFit, for accurately docking ligands into protein active sites. The method employs a cavity detection algorithm for detecting invaginations in the protein as candidate active site regions. A shape comparison filter is combined with a Monte Carlo conformational search for generating ligand poses consistent with the active site shape. Candidate poses are minimized in the context of the active site using a grid-based method for evaluating protein–ligand interaction energies. Errors arising from grid interpolation are dramatically reduced using a new non-linear interpolation scheme. Results are presented for 19 diverse protein–ligand complexes. The method appears quite promising, reproducing the X-ray structure ligand pose within an RMS of 2 Å in 14 out of the 19 complexes. A high-throughput screening study applied to the thymidine kinase receptor is also presented in which LigandFit, when combined with LigScore, an internally developed scoring function [1], yields very good hit rates for a ligand pool seeded with known actives.

Keywords: LigandFit; Protein–ligand complexes; High-throughput docking tools; Active site detection; Ligand docking

1. Introduction

The need for accurate high-throughput docking tools is receiving increasing attention as a result of the rising use of combinatorial chemistry, high-throughput screening, and the growing availability of protein targets with known three-dimensional structures. During the past several years, many methods have been presented for docking ligands into known protein active sites [2–28]. While advances in simulation methods have improved our understanding of the factors that influence binding of ligands into active sites, several aspects of the docking procedure continue to present significant computational challenges. These include adequate positional and conformational sampling of the ligand in the active site, accounting for protein flexibility and treatment of functional water molecules during ligand docking, to name just a few. On the other hand, one often has to compromise some aspects of the accuracy of the docking in order to achieve the speed needed to dock very large numbers of ligands in an acceptable amount of time. For example, using the approximation of a rigid protein, the computation of protein–ligand interaction energies can be considerably speeded up by use of grid-based energy calculations [2,9]. Even in this case, however, thorough sampling of just ligand poses can still be quite time consuming. It is therefore relevant to consider methods that can improve the efficiency of ligand pose sampling. Such methods include feature-based docking [27,29] and shape or similarity-based docking [3,4,7,8,17,23–25]. Ligand feature-based docking works by analyzing the active site of a protein to obtain a representation of the optimal locations of ligand features, such as hydrogen bond donors, acceptors and hydrophobic regions. One then considers only docking poses that overlay the ligand features onto the optimal positions in the active site. In general, shape-based docking methods involve characterizing the shape of the active site and generating ligand poses that are complementary to the shape of the receptor surface. A variety of methods have been reported differing in how the shape of the active site is represented and how the shape-compatible ligand poses are generated. In one of the earliest pioneering works in this area, Kuntz and coworkers [3,30] represented the shape of the active site using a set of overlapping spheres of various radii that touch the surface of the active site. The ligands are also similarly represented by a set of spheres. Ligands are then docked into the active site by optimally overlapping ligand sphere centers with the active site sphere centers. Fradera

* Corresponding author. Tel.: +1-858-799-5359.
E-mail address: venkat@accelrys.com (C.M. Venkatachalam).

1093-3263/02/$ – see front matter © 2002 Published by Elsevier Science Inc.
PII: S1093-3263(02)00164-X

et al. have employed a similarity measure that compares the Gaussian-based molecular fields (steric and electrostatic) of a given ligand to a reference ligand [24]. Cosgrove et al. have described an algorithm that attempts to characterize the shape of the receptor surface using a set of circular *patches* of constant curvature and generate molecular overlays using a clique-detection algorithm [23]. Hahn has employed maximum extents along principal axes of the receptor surface to represent the shape and then searches a multi-conformation database of flexible molecules to select compatible ligands [17]. Sudarsanam et al. have described a method of comparing the shape of an active site with that of a ligand based on the dimensions of the *shape ellipsoid* [7]. A database of ligands is then searched to find ligands with a shape ellipsoid of dimensions similar to that of the active site. In the present work, we describe a new approach, termed LigandFit [30], which provides a rapid accurate protocol for docking small molecule ligands into protein active sites by considering shape complimentarity between the ligand and the protein active site. It employs a shape comparison first used by Oldfield for fitting molecules to electron density distributions from X-ray studies [31,32].

2. Methods

The LigandFit docking procedure consists of two parts: (1) cavity detection to identify and select the region of the protein as the active site for docking; and (2) docking ligands to a selected site. Three-dimensional regular grids of points are employed for site detection and also for estimating the interaction energy of the ligand with the protein during docking.

The first part, cavity detection, involves the use of a flood-filling algorithm [33,34]. While the use of such a procedure for identifying voids in the protein is quite straightforward, determination of where the cavity region ends and bulk solvent begins is more problematic. Thus, an algorithm for automatic identification of the extent (boundary) of the active site is more challenging. In the algorithm described herein, we control the extent of the site by employing a parameter that is related to the size of the 'mouth' of the site.

The second part of our method, the docking procedure, employs a protocol using: (a) conformational searching of flexible ligands using a stochastic sampling to select values for variable torsion angles; (b) selection of a pose based on comparing the shape of the ligand conformation with that of the site; and (c) estimation of the goodness of docking using a grid-based energy calculation for estimating the energy of interaction of the ligand with the protein.

2.1. Cavity detection using protein shape

The cavity detection algorithm begins by first constructing a rectangular grid with a 0.5 Å spacing. The size of the grid is determined by calculating the extents of the protein along the *x*-, *y*- and *z*-axes and adding a border to these extents. Each grid point is then classified as either an occupied or free point. Occupied grid points are those that lie within contact distance of the nearest protein atom. The contact distance is set equal to the radius of the protein atom. The radius of each protein heavy atom is set at 2.5 Å, while the radius for protein hydrogen atoms is set to 2.0 Å. Grid points lying outside contact distance are free (unoccupied). Fig. 1a shows schematically the grid system around a protein. All possible active sites identified by the algorithm will be contained within the collection of free grid points.

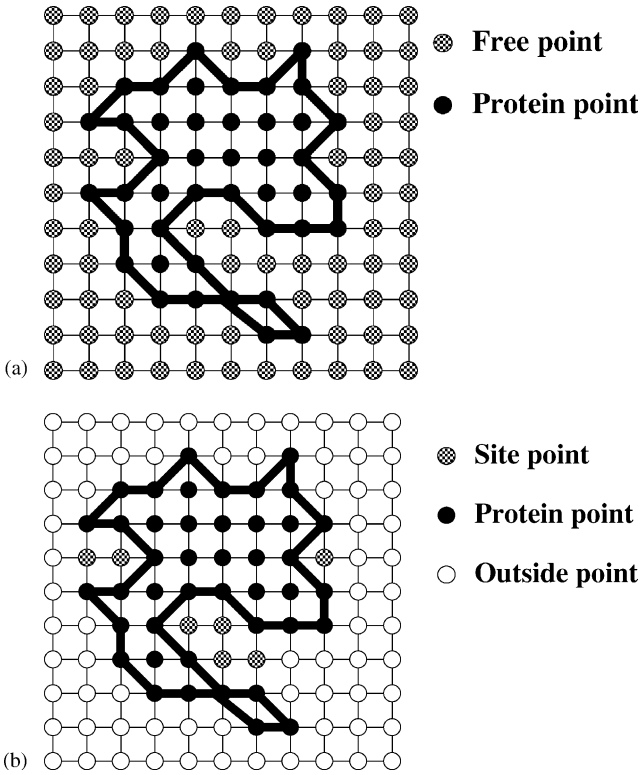

Fig. 1. A schematic representation of the grid system enclosing the protein (black boundary). In (a) protein occupied grid points are shown as black circles while the remaining free grid points are shown as shaded circles. When an eraser of a given size is employed to remove free grid points (see text for details) the free grid points become as shown in (b). Here open circles denote free grid points that have been erased leaving behind three clusters of unerased grid points identifying three possible cavities (sites).

The next step is to partition the free grid points into separate site regions. This involves removing the 'connection' between sites. These regions will generally be connected by free grid points usually lying 'outside' the protein. Individual or distinct active sites can be identified by removing these free grid points. This task is achieved by employing a cubically shaped 'eraser'. Starting from outside the grid, the role of the eraser is to remove free grid points that are encountered while sweeping the eraser across the grid, stopping whenever it comes in contact with a protein atom. Fig. 1b schematically illustrates the effect of free grid point removal by moving the eraser along the axes of the grid

system normal to the six faces of the rectangular parallelepiped. The removal of the free points lying outside the protein results in isolating sites from each other. Finally, a flood-filling procedure [33,34] is employed to collect all the grid points belonging to a given site into a single group. The algorithm is schematically illustrated in Fig. 1b in which three sites are identified.

The extent of the removal of free points depends on the size of the eraser. The smaller the size of the eraser the greater the number of points removed, resulting in smaller-sized sites. The size of the eraser roughly corresponds to the dimension of the opening of the mouth of the site.

2.2. Generating a site from a known ligand

In some cases, an experimental three-dimensional structure of a ligand docked into the active site is available. In such cases, it may be desirable to construct the site from the known ligand pose. A tool for constructing the site in this manner has also been implemented in LigandFit based on a variation of the previously discussed site finding algorithm. Instead of employing the eraser, all free grid points (i.e. grid points not occupied by the protein) that lie within the radius of any ligand atom are determined. The radius of ligand heavy atoms is set at 2.5 Å, while for ligand hydrogens, the radius is set to 2 Å. These radii values are user-adjustable. Thus, the site is obtained as the collection of all grid points occupied by the ligand and unoccupied by the protein. This ligand-based site may be used to validate the automatic eraser-based site finding algorithm as will be shown in Section 3.

2.3. Editing a site

It may be desirable to edit the sites determined from the eraser or ligand-based algorithms to help accommodate the shapes of other ligands. As such, tools are provided to edit the site by expansion or contraction as well as by deleting selected grid points. Expanding the site is achieved by adding all free grid points adjacent to a site point to the site. Contracting the site is achieved by removing all site points adjacent to a free grid point from the site. One may also expand and contract the putative site in the vicinity of user-defined regions. However, such an expansion will not include grid points occupied by the protein. Such editing facilities are useful when one has some knowledge about the binding site. In some cases, the site obtained using the automated procedure may exhibit thin protrusions that cannot accommodate any ligand atoms. It is advisable to manually delete these protrusions to obtain a more realistic representation of the shape of the site.

2.4. Docking methodology

Fig. 2 schematically illustrates the algorithm employed in the docking procedure. For a given ligand, the method is an iterative procedure in which random ligand conformations are generated a specified number of times, $N_{\mathrm{MaxTrials}}$. The procedure maintains a 'Save List' in which the best-docked structures found by the algorithm are stored. One specifies the number of structures, N_{save}, to be saved into the *Save List*. The shape of each candidate ligand conformation is compared with that of the active site. If the Save List is full (i.e. it contains N_{save} docked ligand structures), and if the shape similarity of the candidate conformation is worse than that of any saved structure (in the *Save List*), the candidate conformation is rejected. Otherwise, the candidate conformation is selected for docking. Initially when the Save List is empty, the candidate is selected for docking regardless of its shape similarity. When the Save List is full, the shape similarity comes into play. As the algorithm proceeds, the *Save List* evolves towards better-docked structures as these replace the worse ones. The following steps describe the overall algorithm.

1. Initialize count of trial conformations, N_{trials}, to 0.
2. Increment value of N_{trials} by 1. If $N_{\mathrm{trials}} > N_{\mathrm{MaxTrials}}$, go to Step 13.
3. Generate random trial conformation.
4. Compare the shape of the ligand with that of the site via the shape discrepancy, ρ.
5. If the *Save List* is not full, go to Step 7.
6. If $\rho \geq \rho_{\max}$, reject conformation and return to Step 2 (here, $\rho_{\max}$ is the shape discrepancy of that structure with the largest ρ stored in the *Save List*).
7. Accept the trial conformation for docking. The trial conformation is docked using the active site shape. As discussed below, there are four ways of generating the initial pose of the ligand into the active site. For each pose, the approximate internal strain energy of the ligand and the interaction energy of the ligand with the protein are computed. The dock energy, D, is taken as:

$$D = \text{ligand–protein interaction energy} + \text{ligand internal energy} \quad (1)$$

   The position and the orientation of the ligand is optimized by minimizing the dock energy with respect to rigid body translations and rotations of the ligand using a steepest descent method [35,36]. Further processing of the optimized ligand pose is described in the steps below.
8. If the *Save List* is full, go to Step 10.
9. Save the docked conformation to the *Save List* and go to Step 2.
10. If $D < D_{\max}$ go to Step 12 (here, $D_{\max}$ is the dock energy of that structure with the worst dock energy, i.e. highest D, stored in the *Save List*).
11. Reject conformation and return to Step 2.
12. Replace the worst structure in *Save List* with the optimized docked structure and return to Step 2.
13. Perform rigid body minimization of the ligand with respect to the dock energy for each of the ligand

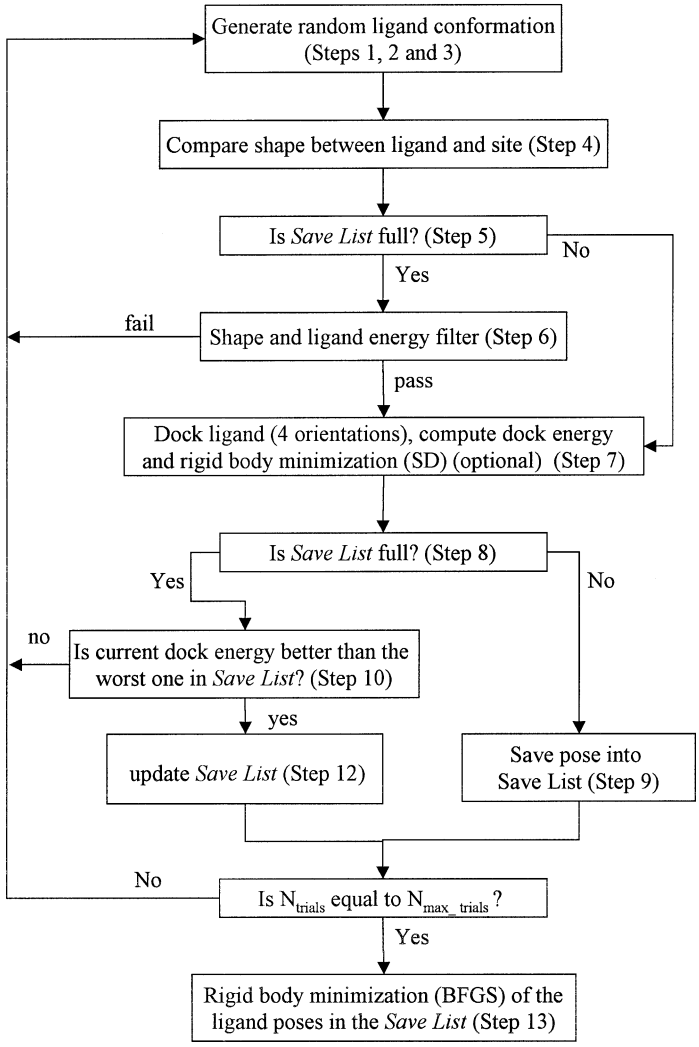

Fig. 2. Docking algorithm.

poses saved in the *Save List* using a BFGS minimizer [35,36].

In the following sections, we discuss the details of the conformation generation, the shape comparison method, docking using the shapes, and computation of the dock energy.

2.5. Ligand conformational generation

Starting from an arbitrary initial conformation, a specified number of attempts, *N*, are made to generate a random ligand conformation. In each attempt, all rotatable bonds are subjected to a quasi-random change in torsion angle. The possible values for change in torsion angle about a bond are based on the total number of rotating atoms; the idea being to allow coarser rotations for those bonds having a smaller number of rotating atoms. This is achieved by employing the following procedure [32].

For a given bond in the ligand, define the ratio σ as:

$$\sigma = \left(0.25\left[\frac{n_{\text{total}}}{n_{\text{rot}}}\right]^2\right)$$

where n_{total} is the total number of atoms in the ligand and n_{rot} the number of rotating atoms.

Next, define two arrays Step (of size 11) and Cut (of size 12) as follows:

$$\text{Step} = \{1, 2, 5, 10, 20, 30, 45, 60, 90, 120, 180\}$$

$$\text{Cut}(1) = 0$$

$$\text{Cut}(i) = 0.5 \times [\text{Step}(i-1) + \text{Step}(i)], \quad i = 2\text{–}11$$

$$\text{Cut}(12) = 9999999 \quad \text{(an arbitrary large number)}$$

Then determine integer j such that:

$$\text{Cut}(j) \leq \sigma < \text{Cut}(j+1)$$

Finally, the amount of rotation, ang, about the bond is taken as:

$$\text{ang} = \text{Rnd} \times \text{Step}(j)$$

where Rnd is an *integer* random number between 0 and 360/Step(j).

For example, using this scheme, for a molecule of 50 atoms, a torsion that rotates 25 atoms will be assigned a step size of 1°, for 10 rotating atoms, the step size will be 5°, for 5 atoms the step size will be 30° and for 3 atoms the step size will be 60°.

The rationale behind such a procedure is that while this protocol results in a greater probability of making a smaller change in torsion angle as the number of rotating atoms increases, this is not achieved at the expense of totally precluding large rotations from occurring. Thus, one reduces the chance of getting trapped in local shape minima when large rotations about bonds affecting many atoms are required for escaping such minima.

2.6. Shape comparison between the ligand and site

In the algorithm adopted herein as previously set out by Oldfield [31,32], shape comparison between the ligand conformation and active site is employed in choosing the ligand conformation for docking and also in selecting the ligand 'pose' (namely position and orientation). The site itself is a collection of grid points. The shape of a collection of points is characterized by the shape matrix:

$$M = \begin{bmatrix} \sum x^2 & \sum xy & \sum xz \\ \sum xy & \sum y^2 & \sum yz \\ \sum xz & \sum yz & \sum z^2 \end{bmatrix} \quad (2)$$

where the summations are over the coordinates of the collection of points. Let S_1, S_2, S_3 be the eigenvalues of the site shape matrix sorted such that $S_1 \geq S_2 \geq S_3$. Similarly, let L_1, L_2, and L_3 be the eigenvalues of a candidate conformation of the ligand's shape matrix (also computed as in Eq. (2) above with the sums now taken over the ligand atom coordinates of the trial conformation). The shape discrepancy, ρ is obtained as:

$$\rho = \sqrt{\left(\frac{S_1}{S_2} - \frac{L_1}{L_2}\right)^2 + \left(\frac{S_2}{S_3} - \frac{L_2}{L_3}\right)^2 + \left(\frac{S_1}{S_3} - \frac{L_1}{L_3}\right)^2} \quad (3)$$

In this expression, the ratios of the eigenvalues are compared instead of the eigenvalues themselves in order to obtain a comparison of shape without giving undue importance to volume. The shape discrepancy of the trial conformation is then compared to those of the conformations already in the Save List. If the shape discrepancy of the candidate is lower than any of the saved conformations, the candidate is accepted for docking. Otherwise, the candidate is rejected and the generation of the next trial conformation ensues.

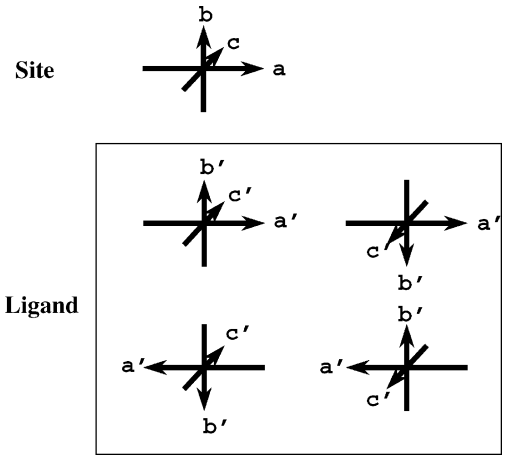

Fig. 3. Four orientations of the ligand consistent with the shape correspondence between the ligand and the site.

2.7. Docking of the ligand: calculation of dock energy

The initial docking of the ligand is obtained by alignment of the principal axes of the ligand to the principal axes of the site. As shown in Fig. 3, there are four possible orientations to be considered. As seen in Eq. (1), there are two energy terms in the expression for the dock energy, internal energy of the ligand and the interaction energy of the ligand with the protein. The interaction energy is taken as the sum of the van der Waals energy and electrostatic energy. The van der Waals energy of interaction of the ligand with the protein can be expressed as:

$$E_{\text{vdW}} = \sum_{\text{ligand–protein atom pairs } i,j} \varepsilon_{ij} \left[2\left(\frac{r^*_{ij}}{r_{ij}}\right)^9 - 3\left(\frac{r^*_{ij}}{r_{ij}}\right)^6 \right] \quad (4)$$

where

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}, \qquad r^*_{ij} = \sqrt{r^*_i r^*_j} \quad (5)$$

and r^*_i and ε_i are the van der Waals radius and energy parameters of the i-th ligand atom while r^*_j and ε_j are similarly parameters for the j-th protein atom, and r_{ij} is the distance between the i-th ligand atom and the j-th protein atom. In the calculations presented in this paper, we have employed the CFF force field [37–40] for the van der Waals parameters, r^*_i and ε_i, for the ligand and protein atoms [41].[1]

The electrostatic energy of interaction between the ligand and the protein is taken as:

$$E_{\text{ele}} = \frac{332.0716}{\varepsilon} \sum_{\text{ligand–protein atom pairs } i,j} \frac{q_i q_j}{r_{ij}} \quad (6)$$

[1] Even though we have utilized CFF force field for all the calculations here, we have also implemented the use of Dreiding force field for the docking calculations. When using the Dreiding force field, we have employed the Gasteiger charging method for determining charges on the protein and the ligand. Results obtained using the Dreiding force field will be presented elsewhere.

where q_i and q_j are the respective charges (in atomic units) on the ligand atom i and protein atom j and ε is the dielectric constant.[2] The charges are defined as part of the force field specification, and can be generated using the Cerius[2] software[3]. By default, the dielectric constant is set to unity.

The computation of the interaction energy using Eqs. (4)–(6) is quite time consuming. To improve the speed of the docking procedure, we employ a grid-based energy [2,9] estimation of the interaction energy. A rectangular grid encloses the selected site with a specified border. At each grid point, x, the van der Waals attractive, $\Phi_{\text{atr}}(x)$, and repulsive potential, $\Phi_{\text{rep}}(x)$, as well as the electrostatic potential, $\Phi_{\text{ele}}(x)$, due to the protein atoms are precomputed using Eqs. (7)–(9).

$$\Phi_{\text{rep}}(x) = \sum_{j \in \text{protein atoms}} 2\sqrt{\varepsilon_j} \left(\frac{\sqrt{r_j^*}}{R_{xj}} \right)^9 \tag{7}$$

$$\Phi_{\text{atr}}(x) = \sum_{j \in \text{protein atoms}} 3\sqrt{\varepsilon_j} \left(\frac{\sqrt{r_j^*}}{R_{xj}} \right)^6 \tag{8}$$

$$\Phi_{\text{ele}}(x) = \sum_{j \in \text{protein atoms}} \frac{332.0716}{\varepsilon} \left(\frac{q_j}{R_{xj}} \right) \tag{9}$$

where the summations are over protein atoms j, r_j^* and ε_j are the van der Waals radius and energy parameters of the j-th protein atom and R_{xj} is the distance between the j-th protein atom and grid point at x.

The computation of the potentials at the grid points is performed only once for a given protein and defined active site. The van der Waals interaction energy of a ligand is then given by:

$$E_{\text{vdW}} = \sum_{i \in \text{ligand atoms}} \left(\sqrt{\varepsilon_i}(r_i^*)^{9/2} \Phi_{\text{rep}}(x_i) - \sqrt{\varepsilon_i}(r_i^*)^3 \Phi_{\text{atr}}(x_i) \right) \tag{10}$$

where x_i is the position of the i-th ligand atom. The electrostatic interaction energy is given by:

$$E_{\text{ele}} = \sum_{i \in \text{ligand atoms}} (q_i \Phi_{\text{ele}}(x_i)) \tag{11}$$

2.8. Grid interpolation

The above Eqs. (10)–(11) for E_{vdW} and E_{ele} would be exactly equivalent to Eqs. (4) and (6) presented above if the potentials, Φ_{atr}, Φ_{rep}, and Φ_{ele} were computed at ligand atom locations. Since we calculate the values of the potentials at grid points, the values of the potentials at the ligand atom locations requires the use of an interpolation scheme. As a result of interpolation errors, the resulting estimate of the interaction energy using Eqs. (10)–(11) is consequently only an approximation to the value obtained via Eqs. (4) and (6). The most straightforward approach to interpolation on a three-dimensional grid is tri-linear interpolation [9]. However, the van der Waals potentials are highly non-linear and simple tri-linear interpolation can lead to large errors [42]. In Fig. 4 we show the errors introduced by using tri-linear interpolation to evaluate the interaction energy for a set of protein–ligand complexes from X-ray data. Here, the grid resolution is 0.5 Å and the errors represent the difference between the exact calculation using Eqs. (4) and (6) and the grid-based calculations using Eqs. (10) and (11). One can see that the errors range from about 20 to as large as 140 kcal/mol.

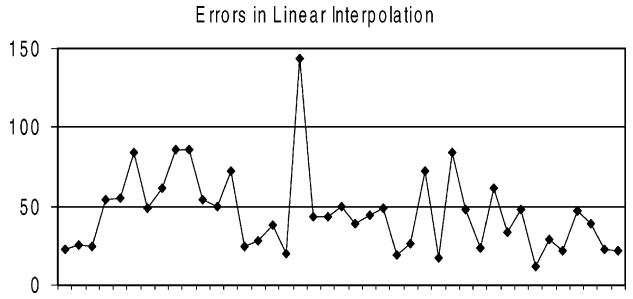

Fig. 4. Errors (kcal/mol) in linear interpolation for a set of 42 protein–ligand complexes from X-ray data, using a grid resolution of 0.5 Å and 12–6 potential.

In this study, we now describe a novel method of dramatically reducing the errors due to interpolation. The essence of the method is to transform the potential into a more smoothly varying function, perform tri-linear interpolation over the transformed function and then back-transform the interpolated value.

Consider a ligand atom i located at x_i. Estimation of the interaction energy is achieved by the following steps.

Step 1: Identify the eight corner grid points surrounding the position x_i of the ligand. For each corner grid point x, compute the values using:

$$F_{\text{atr}}(x) = (\Phi_{\text{atr}}(x))^{-1/n}, \quad F_{\text{rep}}(x) = (\Phi_{\text{rep}}(x))^{-1/m} \tag{12}$$

where m and n are positive numbers.

Step 2: Using the eight sets of values, perform tri-linear interpolation to obtain the values for $F_{\text{rep}}(x_i)$ and $F_{\text{atr}}(x_i)$ at the ligand atom position x_i.

Step 3: Back-transform the two values obtained in Step 2 using:

$$g_{\text{rep}}(x_i) = (F_{\text{rep}}(x_i))^{-m}, \quad g_{\text{atr}}(x_i) = (F_{\text{atr}}(x_i))^{-n} \tag{13}$$

Step 4: Now estimate the interaction energy of the ligand atom i with the protein using:

$$E_{\text{vdW}}(i) = \sqrt{\varepsilon_i}(r_i^*)^{9/2} g_{\text{rep}}(x_i) - \sqrt{\varepsilon_i}(r_i^*)^3 g_{\text{atr}}(x_i)$$

[2] Optionally, we have also considered a distance-dependent dielectric by taking $1/r_{ij}^2$ instead of $1/r_{ij}$ in Eq. (6).

[3] Cerius[2] LigandFit software package is available from Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121, USA.

The reduction in error depends on the choice for the values of m and n. The effect of varying m and n are analyzed in Section 3.

For electrostatic energy, we have employed simple tri-linear interpolation since the rise in energy at close distances is far less steep than in the case of van der Waals energy. Thus, the errors resulting from simple tri-linear interpolation for electrostatics are smaller and acceptable without the need for a transformed function approach.

2.9. Soft van der waals functions

The 9–6 potential employed here for estimating the van der Waals energy has a steep rise at short interatomic distances. In the context of protein–ligand docking, when treating the protein as rigid as is done in the current work and due to limited sampling of the ligand conformational space, this has undesirable consequences. For example, the 9–6 potential will unduly penalize dockings with 'mild' short contacts between ligand and protein. We have thus employed a soft potential that gradually rises to a large but finite value at zero separation between atoms. Eq. (15) shows the soft-potential modification suggested by Levitt [43] applied to the 9–6 van der Waals expression in Eq. (4).

$$E_{\mathrm{vdW}}(r_{ij}) = \frac{(2\varepsilon_{ij}r_{ij}^{*9}/r_{ij}^{9}) - (3\varepsilon_{ij}r_{ij}^{*6}/r_{ij}^{6})}{(2\varepsilon_{ij}r_{ij}^{*9}/r_{ij}^{9})\alpha(1+\beta r_{ij}^{2})+1} \tag{15}$$

where α is a parameter that controls the value of the function at $r_{ij} = 0$, and β is a parameter that controls the rate at which the function approaches the maximum value at zero separation.

To facilitate the incorporation of this softening function while using a grid-based formalism for calculating the interactions, we have adopted a modified form of Eq. (15) above to soften both the attractive and repulsive *potentials* as shown in Eqs. (16)–(17).

$$\Phi_{\mathrm{rep}}^{\mathrm{soft}}(x) = \sum_{j\in\text{protein atoms}} \frac{2\sqrt{\varepsilon_j}\left(\sqrt{r_j^*}/R_{xj}\right)^9}{2\sqrt{\varepsilon_j}\left(\sqrt{r_j^*}/R_{xj}\right)^9\alpha(1+\beta R_{xj}^2)+1} \tag{16}$$

$$\Phi_{\mathrm{atr}}^{\mathrm{soft}}(x) = \sum_{j\in\text{protein atoms}} \frac{3\sqrt{\varepsilon_j}\left(\sqrt{r_j^*}/R_{xj}\right)^6}{2\sqrt{\varepsilon_j}\left(\sqrt{r_j^*}/R_{xj}\right)^9\alpha(1+\beta R_{xj}^2)+1} \tag{17}$$

where R_{xj} is the distance between the j-th protein atom and grid point x. The expression for the softened van der Waals energy of interaction of the ligand then becomes:

$$E_{\mathrm{vdW}}^{\mathrm{soft}} = \sum_{i\in\text{ligand atoms}} \left(\sqrt{\varepsilon_i}(r_i^*)^{9/2}\Phi_{\mathrm{rep}}^{\mathrm{soft}}(x_i) - \sqrt{\varepsilon_i}(r_i^*)^3\Phi_{\mathrm{atr}}^{\mathrm{soft}}(x_i)\right) \tag{18}$$

In the original Levitt formalism [43], the limiting value of the interaction at $r = 0$ is given by the parameter α, and is independent of the r_{ij}^* and ε_{ij} values. Our modified form of the Levitt function is used to enable a factoring of the protein atom-dependent and ligand atom-dependent terms so that the function may be summed over all protein atoms independent of the ligand atom for each grid point. However, as a consequence, the limiting value at $r = 0$ for a given protein–ligand atom pair is now dependent on the values of the ligand atom parameters, ε_i and r_i. Our choice for the parameter α results in limiting values at zero that range from 100 to 1060 kcal/mol using the CFF force field parameters over the range of ligand atom types that have been parameterized.

2.10. Soft electrostatic potential

The electrostatic potential also rises to infinity as the interatomic distance approaches zero though the rise to infinity is not as steep as the van der Waals repulsive and attractive components. Nevertheless, it is desirable to soften the electrostatic energy to prevent the electrostatic energy from dominating the softened van der Waals energy at short distances.

We soften the electrostatic potential given in Eq. (9) by replacing $1/R_{xj}$ by a function $g(R_{xj})$:

$$\Phi_{\mathrm{ele}}(x) = \sum_{j\in\text{protein atoms}} \frac{332.0716}{\varepsilon} q_j g(R_{xj}) \tag{19}$$

where the smoothing function $g(R_{xj})$ is defined as:

$$g(R_{xj}) = \frac{1}{R_{xj}}, \quad \text{for } R_{xj} \geq 1,$$
$$g(R_{xj}) = k + aR_{xj}^2 + bR_{xj}^3, \quad \text{for } R_{xj} < 1 \tag{20}$$

where k is an adjustable parameter that controls the value of the electrostatic potential at zero separation. The value of the function g is unity at $R = 1$ and is k at $R = 0$. The constants[4] a and b are taken as $a = 4 - 3k$ and $b = 2k - 3$, which provides for continuity of the function and its first derivative at $R = 1$. We have taken the value of k as 2.

2.11. Ligand internal energy

The internal energy of the ligand is taken as the sum of the internal van der Waals and electrostatic energy. The van der Waals energy is computed using the 9–6 function (see Eq. (4) with the summation taken over all non-bonded atom pairs within the ligand separated by at least three bonds. Similarly, electrostatic energy is computed using Eq. (6) for the same non-bonded ligand atom pairs. Unlike intermolecular interactions, softening of the energy is not employed for the internal energy calculations, so as to avoid the generation of unreasonably high energy conformations of the ligand.

[4] The values of a and b are determined such that the function and its first derivative of $g(R)$ are continuous at $R = 1$.

2.12. Scoring docked ligands

The docking procedure described here attempts to produce ligand poses having a favorable energy of interaction with the protein. While the energy functions employed for this purpose tend to favor energetically reasonable ligand poses, they have not been derived to predict binding affinities or to prioritize ligands relative to one another. Hence, we have employed scoring functions to prioritize docked ligands. In addition to the scoring functions, such as Ludi [44], PLP [45,46] and PMF [47] developed elsewhere, we have also employed a new scoring function, LigScore, developed in-house [1]. LigScore has been developed by using the Genetic Function Approximation [48,49] to train against a set of 80 protein–ligand complexes obtained from high resolution X-ray structure determination for which experimental binding affinities (i.e. pK_i values) are available. The functional form of LigScore is simple and contains only three descriptors. Eq. (21) shows the LigScore function obtained with the CFF force field using grid-based energy calculations with the improved interpolation method described in this article.

$$\begin{aligned} \mathrm{p}K_i &= 0.517527 - 0.043650(\mathrm{vdW\_Grid\_Soft}) \\ &\quad +0.143901(C + \mathrm{\_pol}) - 0.00099039(\mathrm{Totpol\hat{}2}) \end{aligned} \tag{21}$$

where vdW_Grid_Soft is the grid-based soft van der Waals energy of interaction between the ligand and the protein. C+_pol is the total surface area of the ligand involved in attractive polar interactions with the protein and Totpol^2 is $(\mathrm{C} + \mathrm{\_pol})^2 + (\mathrm{C} - \mathrm{\_pol})^2$ where C−_pol is the total surface area of the ligand involved in repulsive polar interactions with the protein. Additional details about the LigScore function will be reported elsewhere [1].

3. Results

3.1. Performance of cavity detection: variation of site with eraser size

To assess the optimal size to be employed for the eraser, the site detection algorithm was run on a set of 75 proteins (see Fig. 5a) for various eraser sizes from 4 to 10 Å. Key questions to explore included: (a) Does a given eraser size identify the site found in a given ligand–protein X-ray structure complex (termed the X-ray site)? (b) If the X-ray site is obtained with a given eraser size, is it the largest site (as measured by the number of grid points in the site)? (c) Does the site identified encompass the ligand pose found in the X-ray structure? Fig. 5b shows the distribution of the rank (according to the size) of the site detected with respect to the site from the X-ray structure. Here, we show the distribution of the best rank obtained for the X-ray site by varying the eraser size. For 53 of the 75 proteins, the X-ray site is identified to be the largest site if a suitable eraser size is employed. Fig. 5c shows the distribution of the eraser size that detects the X-ray site as one of the sites (regardless of the rank). It can be seen that in 59 proteins, an eraser size of 5 Å identifies the X-ray site as one of the possible sites. These results are further summarized in Table 1 where the

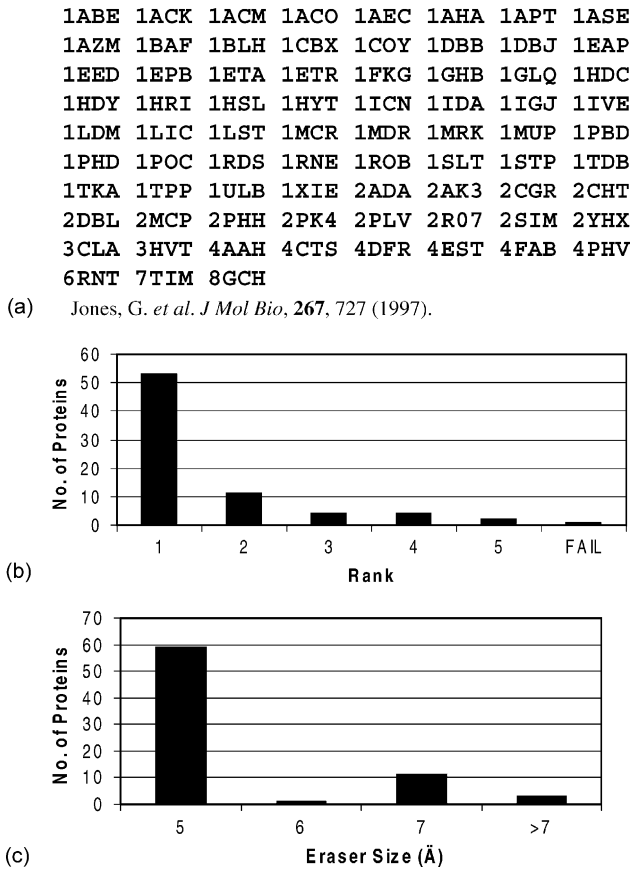

Fig. 5. (a) List of 75 unique proteins (GOLD set) from Protein Data Bank employed in the validation of site detection algorithm. (b) Distribution of the rank (according to volume) of the site detected corresponding to site in X-ray structure. The cases where the site detection failed to identify the X-ray site are also shown marked 'FAIL'. (c) Distribution of the size of the eraser that identifies the correct site.

Table 1
The effect of eraser size on rank of the X-ray site in the site finding algorithm

Eraser size	Rank	No. of proteins
5	1	45
5	2	8
5	3	2
5	4	3
5	5	1
6	2	1
7	1	5
7	2	2
7	3	2
7	4	1
7	5	1
9	1	1
10	1	2

Table 2
Performance of LigandFit in 19 selected protein–ligand complexes

PDB code	Complex	Best eraser size (Å) coverage	$L \cap E/L \cup E$ (%)	No. of ligand torsions	RMS (Å)			
					A	B	C	D
186L	Lysozyme	5	99	3	0.39	0.39	0.66	0.71
1A9U	p38 Kinase	8	29	4	0.23	5.39	1.11	8.08
1ABE-1	Arabinose-binding protein	5	100	4	0.17	0.2	0.17	0.19
1ACL	Acetylcholinesterase	7	46	11	1.24	1.26	1.23	1.95
1APT	Penicillopepsin	7	74	22	0.24	0.24	2.51	8.35
1B40	Oligo-peptide binding protein	5	80	18	0.35	4.33	1.27	1.39
1CPS	Carboxypeptidase	5	77	7	0.69	0.71	0.64	0.62
1ELA	Elastase	10	59	12	4.83	2.79	1.06	2.17
1ETR	Thrombin	7	45	10	0.32	0.34	2.29	3.48
1KIM	Thymidine kinase	5	80	4	0.22	0.21	0.35	0.32
1PHG	Cytochrome P450-CAM	5	91	3	0.27	0.35	0.31	0.57
1RBP	Retinol-binding protein	5	93	6	0.29	0.25	0.61	0.43
1STP	Streptavidin	6	89	4	0.27	0.25	0.7	0.29
2CGR	IgG2b–Fab complex	9	46	10	0.53	0.53	0.83	0.9
2QWK	Neuraminidase	7	63	6	0.4	3.76	0.54	0.68
3PTB	Trypsin	6	92	1	0.41	0.33	0.39	0.33
4DFR	Dihydrofolate reductase	6	72	8	0.2	0.21	1.11	1.24
4PHV-1	HIV protease	6	89	15	0.16	0.17	1.99	3.78
7UPJ	HIV protease	6	70	7	0.33	0.44	0.66	1.09

For A: fitting, rigid; starting ligand conformation, X-ray; site employed, X-ray ligand based. For B: fitting, rigid; starting ligand conformation, X-ray; site employed, eraser based. For C: fitting, flexible; starting ligand conformation, random; site employed, X-ray ligand based. For D: fitting, flexible; starting ligand conformation, random; site employed, eraser based. LigandFit parameters: CFF forcefield, $N_{\text{MaxTrials}} = 10{,}000$, with RBM, $N_{\text{save}} = 10$.

number of X-ray sites that are recovered with a given rank is shown for a given eraser size. For a majority of the proteins in this test set (45 out of 75 proteins), the active site is identified as the largest site correctly by an eraser size of 5 Å.

The performance of the LigandFit site detection algorithm was tested more quantitatively on 19 protein–ligand complexes listed in Table 2. This set represents 18 different proteins. For each protein, site detection was carried out using the X-ray ligand as well as using the eraser with various sizes (from 5 to 10 Å in increments of 1 Å). The site E obtained using the eraser algorithm was compared with the site L obtained using the X-ray ligand by computing the ratio:

$$\text{coverage} = \frac{L \cap E}{L \cup E} \tag{22}$$

where $L \cap E$ is the volume intersection between the two sites, and $L \cup E$ the corresponding volume union. The table shows the best eraser size for each protein and the respective coverage obtained using the eraser algorithm. More than 70% coverage is obtained with the eraser algorithm for 13 out of the 19 proteins studied here. A majority of proteins can be treated with an eraser size of 5–6 Å. Proteins requiring a very large eraser size (∼10 Å) have a very open binding site. Elastase (1ELA) is an example of such a system. Another factor that can complicate the site analysis is that, in some cases, the cavity may be larger than the region occupied by the ligand. P38 kinase complex (1A9U) is a case in point in which the coverage measure is only 29%.

3.2. Performance of the improved tri-linear interpolations

The performance improvements in the modified tri-linear interpolation discussed above were tested with various integral values for m and n in Eqs. (12) and (13). A dramatic reduction in errors was achieved with $m = n = 2$. Fig. 6a–c shows the errors obtained for a set of X-ray complexes for various values of m and n (taken as equal), using 12–6 and 9–6 potentials and using 0.5 and 0.25 Å grid resolutions. It can be seen from Fig. 6a that the unsigned maximum error is reduced from 144 to 29.4 kcal/mol when going from a 12–6 function to a 9–6 function using tri-linear interpolation with 0.5 Å grid resolution. Employing the improved interpolation with $m = n = 2$ further reduces the maximum error to 5.5 kcal/mol. Using 0.25 Å grid resolution and modified interpolation with $m = n = 2$, the maximum unsigned error further reduces to 2.8 kcal/mol as seen in Fig. 6b. The unsigned average error is 45.7 kcal/mol with tri-linear interpolation, 12–6 function and 0.5 Å grid resolution. This is reduced to 2.1 kcal/mol, with 9–6 function and improved interpolation. The error is further reduced to 0.8 kcal/mol with 0.25 Å grid resolution. Fig. 6c shows the effect of varying the value of m with $n = m$. The most significant improvement in error is achieved with $m = n = 2$ when compared to $m = n = 1$. We have employed $m =$

$n = 2$ in all the calculations presented in this paper, except otherwise noted. Three factors contribute to reducing the errors: (a) softening the repulsive potential from a 12-6 to a 9–6 form; (b) improving grid resolution from 0.5 to 0.25 Å; and (c) using improved tri-linear interpolation with $m = n = 2$.

3.3. Rigid docking to ligand-based site: the case of HIV protease and the 4PHV ligand

To validate the shape comparison and alignment procedure employed here, we consider the X-ray structure of the ligand 4PHV bound to HIV protease. Starting from the X-ray

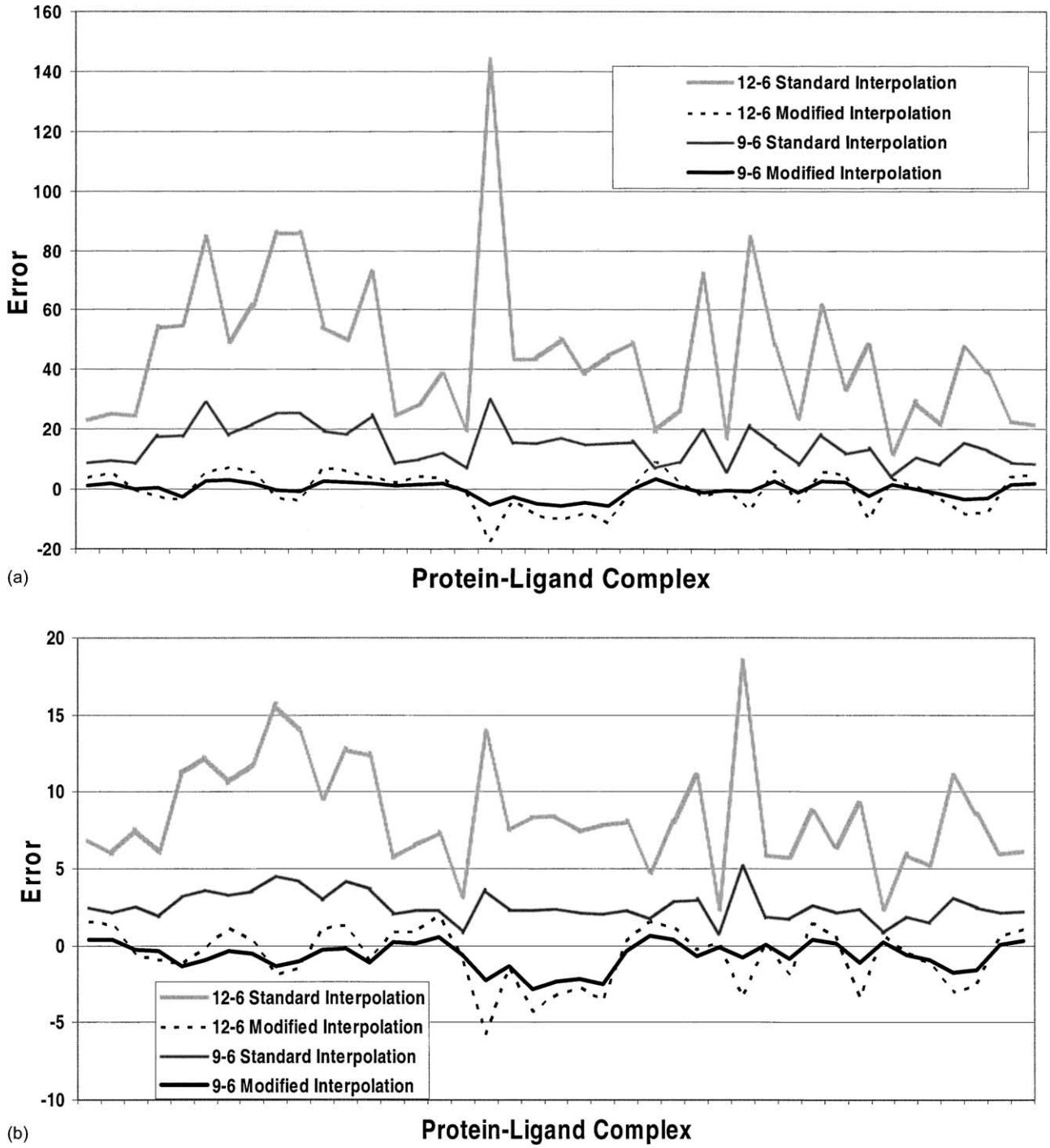

Fig. 6. Variation of error (kcal/mol) in 41 selected protein complexes, due to modified tri-linear interpolation with $m = n$: (a) with 12–6 and 9–6 potentials at grid resolution of 0.5 Å and $m = n = 2$; (b) with 12–6 and 9–6 potentials at grid resolution of 0.25 Å and $m = n = 2$; (c) with 9–6 potential and 0.25 Å for various values of $m = n$.

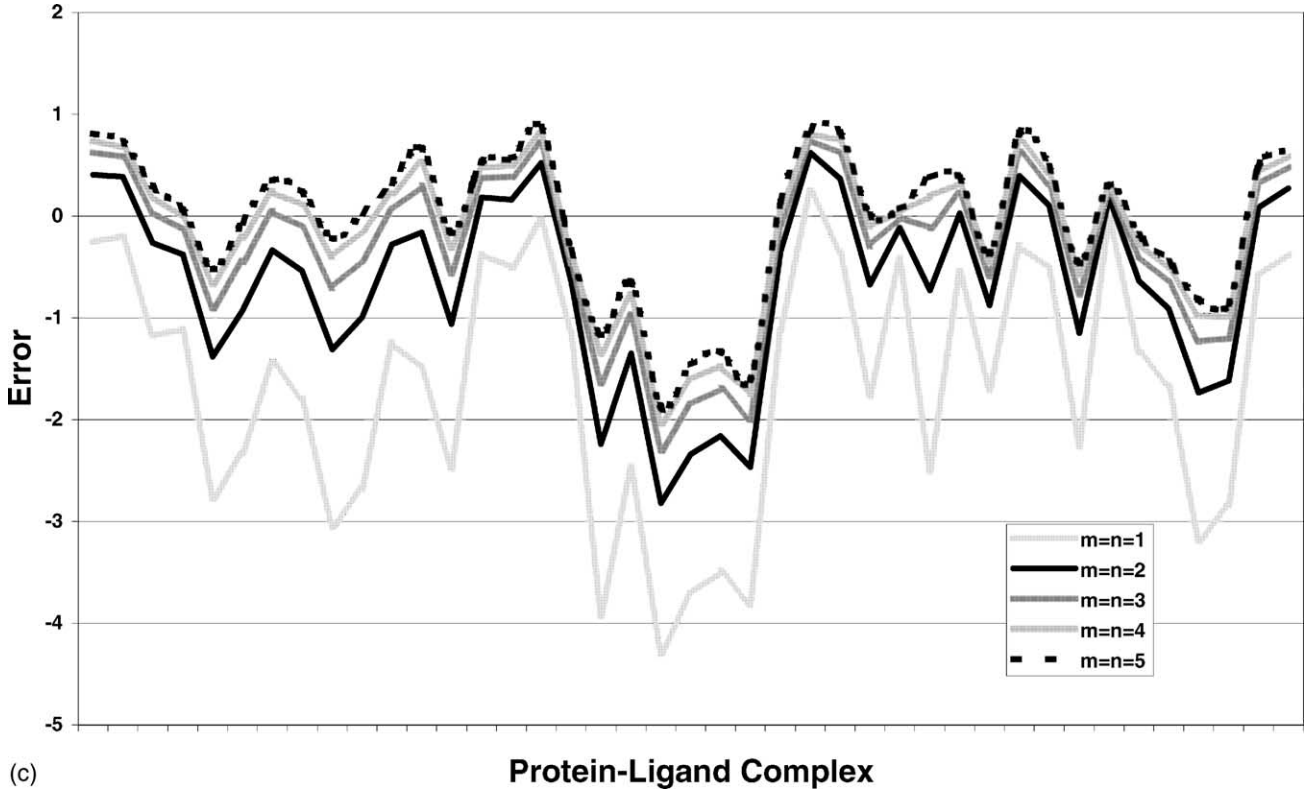

Fig. 6. (*Continued*).

conformation of the 4PHV ligand, rigid docking was performed with and without rigid body minimization using the site obtained with the eraser algorithm with various eraser sizes as well as the site derived from the X-ray ligand pose. Table 3 summarizes the dock energy and the heavy atom RMS difference from the X-ray pose obtained for all four orientations (see Fig. 3) consistent with the shape alignment. As can be seen from the last two pairs of rows in Table 3, using the site derived from the X-ray ligand pose, one obtains very good results with rigid docking of the X-ray ligand conformation. For orientations 1 and 2, one obtains an RMS of 0.70 and 0.65 Å without RBM and 0.16 and 0.28 Å with RBM, respectively. Note that the two orientations yield slightly different results, because the X-ray structure is not perfectly symmetric although the ligand itself is topologically symmetrical. Fig. 7 shows the results obtained with rigid docking to the ligand-based site with and without RBM. Fig. 8 shows similar results for rigid docking to the site obtained with a 5 Å eraser. Thus, we conclude that if a good description of the site is available, and if the ligand conformation is correct, then the rigid fitting via shape alignment yields a ligand pose very close to the X-ray pose. Note also that a near-perfect docking is achieved only with RBM. Using the eraser algorithm, the best docking with an RMS of 0.14 Å is obtained with an eraser size of 5 Å. Thus, for this protein–ligand complex the default eraser algorithm works very well. While results for all four orientations are given here for each rigid docking performed, the first two and last

Table 3
Performance of rigid fitting 4PHV Ligand in X-ray conformation to HIV protease

Eraser size (Å)	RBM[a]	Orientations (dock energy[b] RMS[c])			
		1	2	3	4
5	No	1676.7 2.9	1801.2 2.9	2716.4 6.1	2780.5 6.1
5	Yes	−143.1 0.14	−141.8 0.28	469.9 5.8	586.2 5.9
6	No	610.2 1.65	752.7 1.69	1758.4 5.7	1919.7 5.7
6	Yes	−143.1 0.17	−141.8 0.28	577.7 5.6	677.3 5.7
7	No	996.3 1.97	1139.1 2.0	2108.8 5.8	2170.0 5.8
7	Yes	−143.1 0.17	−141.8 0.28	530.8 5.9	603.7 5.6
Lig[d]	No	−84.0 0.70	−79.7 0.65	979.0 5.6	1018.8 5.6
Lig[d]	Yes	−143.1 0.16	−141.8 0.28	576.8 5.6	595.0 5.6

[a] Rigid body minimization of the ligand based on dock energy after docking using shape alignment.
[b] In kcal/mol.
[c] RMS difference (Å) between docked ligand pose and the ligand pose in X-ray structure.
[d] Site detected using the ligand pose in the X-ray structure.

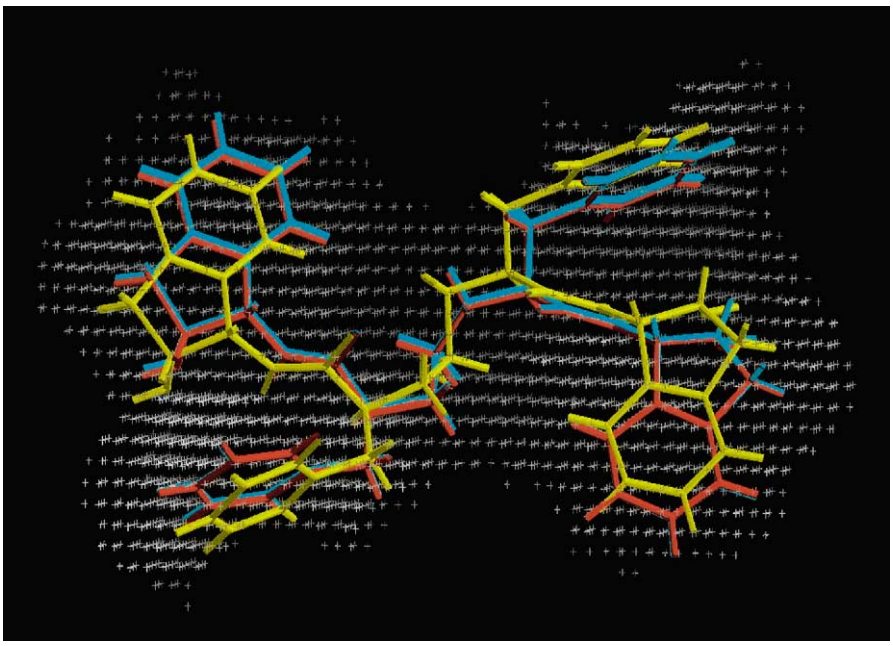

Fig. 7. Rigid docking of 4PHV ligand to site determined from X-ray ligand. Red: X-ray structure. Yellow: rigid docking without RBM. Cyan: rigid docking with RBM. The site points are shown in gray crosses.

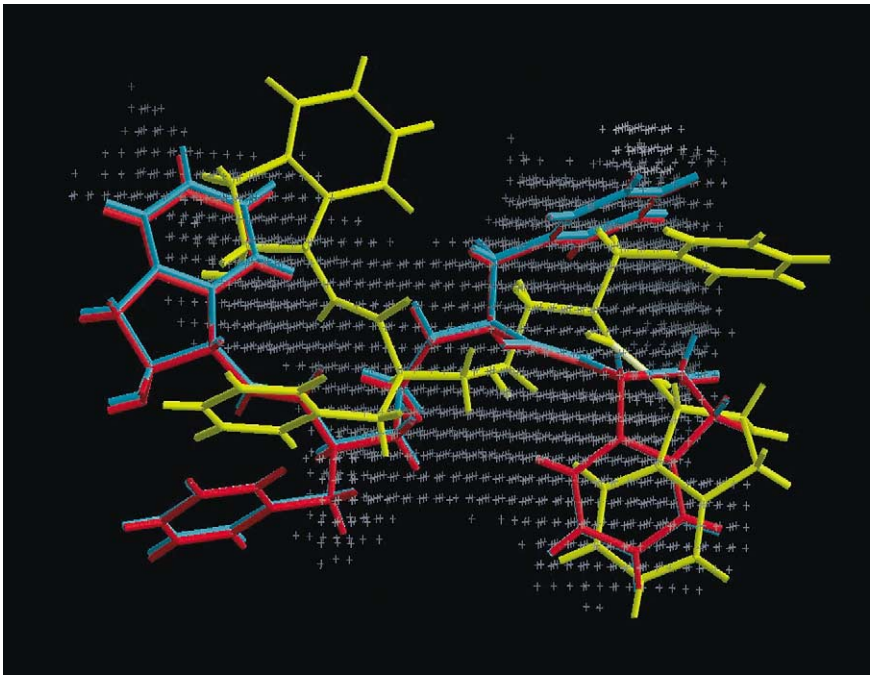

Fig. 8. Rigid docking of 4PHV ligand to site determined using a 5 Å eraser. Same color scheme as in Fig. 7.

two orientations in each case give very similar RMS values (and similar dock energies), respectively. This result is due to the topological symmetry of the 4PHV ligand.[5]

[5] The RMS calculation takes topological symmetry into account and automatically associates atom pairs between the X-ray structure and the docked structure to report the best RMS for all possible topologically equivalent pairings.

3.4. Performance of LigandFit with various ligand–protein complexes

The performance of the LigandFit algorithm was tested on the same 19 protein–ligand complexes used to analyze the performance of site detection algorithm in Table 2. For each complex, the X-ray ligand conformation was rigidly docked to the protein with the site defined using the eraser algorithm

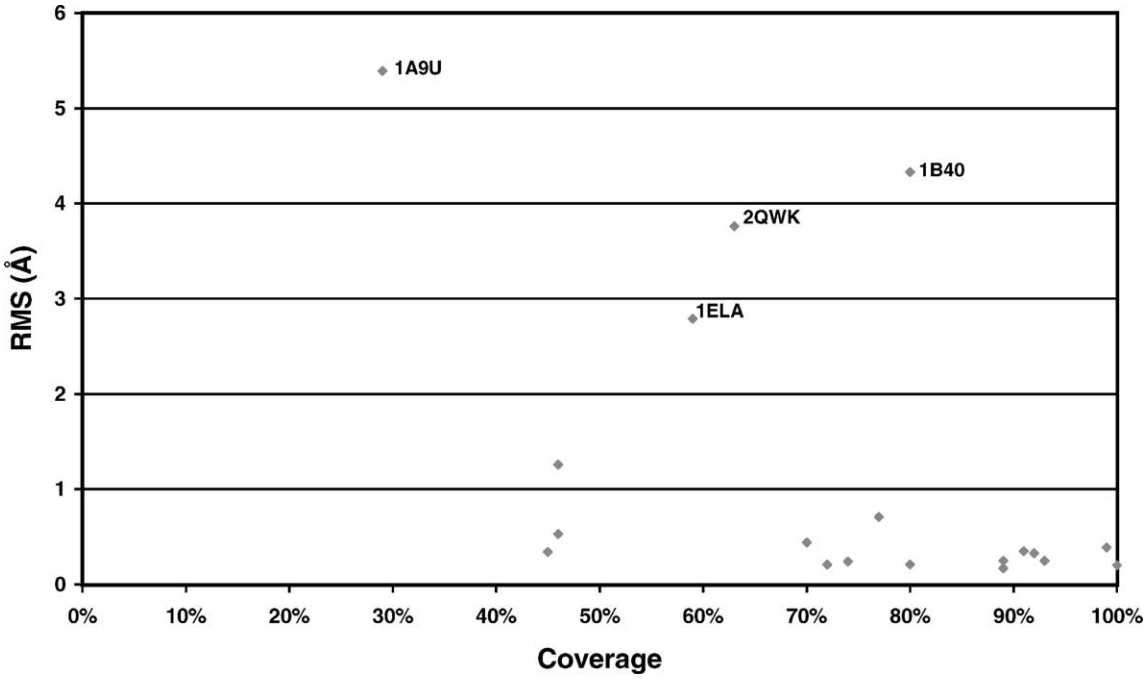

Fig. 9. Scatter plot of site coverage with best RMS with X-ray structure using rigid docking to the best eraser-based site.

as well as using the site generated from the X-ray pose of the ligand. In addition, each ligand was assigned a random conformation and was flexibly docked to the corresponding protein with $N_{\text{MaxTrials}} = 10{,}000$, $N_{\text{save}} = 10$ and employing rigid body minimization using the CFF force field. These computations were repeated using sites obtained with the eraser algorithm using various eraser sizes. The RMS difference between each saved ligand pose was determined relative to the ligand pose in the X-ray structure.

Table 2 summarizes the results. The table shows the lowest RMS difference obtained for each complex. Fig. 9 shows a scatter plot of the lowest RMS (Å) obtained by rigidly docking the X-ray ligand to the eraser-based site plotted against the site coverage. Satisfactory docking with low RMS is obtained in all the cases where the coverage is more than 70% with the exception of 1B40. In the case of 1B40, though the site coverage is high, the shape of the eraser-based site is such that the eigenvalues of the shape matrix are almost degenerate for the two largest eigenvalues. This degeneracy results in a reversal of the order of the two larger axes for the eraser-baser site with respect to the axes of the X-ray ligand-based site. Such a mismatch results in erroneous rigid docking to the eraser-based site while the matching is correct in the case of rigid docking to the X-ray-based site. Work is in progress to take corrective measures against such degeneracies by considering other orientations in the shape alignment algorithm in addition to the four orientations shown in Fig. 3. It can be seen in Table 2 that in the case of 1ELA, the rigid docking of the X-ray ligand to the X-ray ligand-based site yields a high RMS of 4.83 Å, and this is also the result of a similar degeneracy problem.

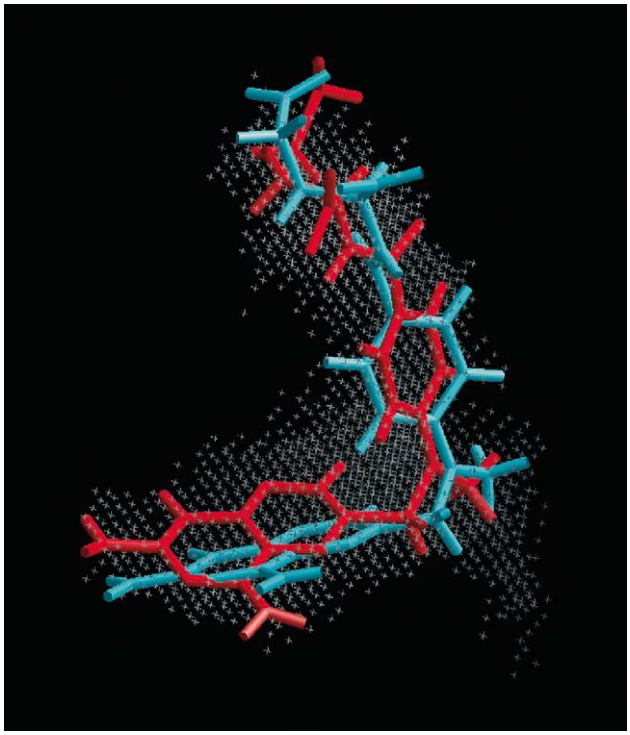

Fig. 10. Flexible docking with RBM of 4DFR ligand (dihydrofolate reductase) to site determined using a 6 Å eraser (docked structure in cyan and the site in gray). The X-ray pose of the ligand is in red. In this case, docking works quite well using the eraser algorithm for determining the site.

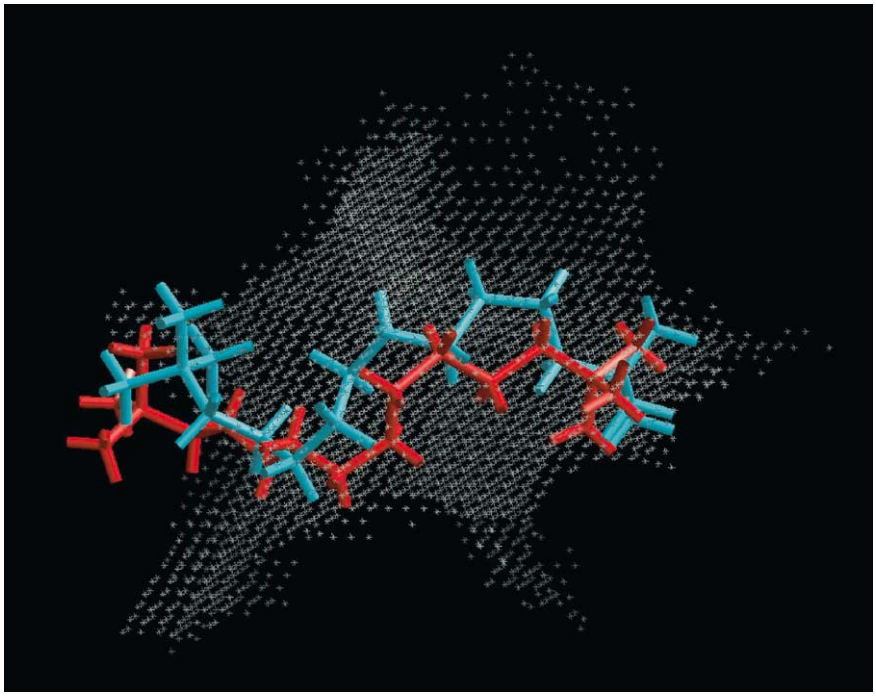

Fig. 11. Flexible docking with RBM of 1ACL ligand (acetylcholinesterase) to site determined using a 7 Å eraser (docked structure in cyan and the site in gray). The X-ray pose of the ligand is in red. Even though the site detected by the eraser algorithm does not match the region occupied by the X-ray ligand, the docking nevertheless succeeds since the shape information from the calculated site is approximately correct.

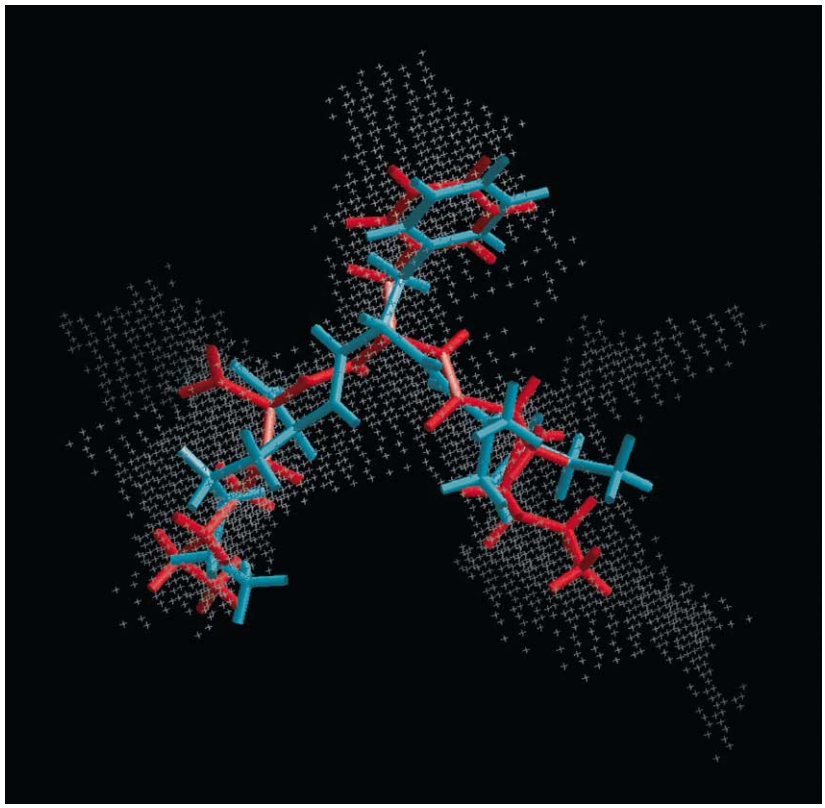

Fig. 12. Flexible docking with RBM of 1B40 ligand (oligopeptide binding protein) to site determined using a 5 Å eraser (docked structure in cyan and the site in gray). The X-ray pose of the ligand is in red. The site detected by the eraser algorithm covers significantly more region than that occupied by the X-ray ligand. The flexible docking nevertheless works quite well.

The quality of the flexible docking to the eraser-based site may be inferred from the last column of Table 2. Fourteen complexes yield an RMS of less than 2 Å while the coverage among this set varies from approximately 50–100%. Figs. 10–12 give examples of excellent flexible docking achieved with 4DFR (dihydrofolate reductase), 1ACL (acetylcholinesterase) and 1B40 (oligopeptide binding protein). On the other hand, good docking (i.e. RMS < 2.0 Å) is not obtained for the four complexes, 1A9U (p38 kinase), 1APT (penicillopepsin), 1ETR (thrombin) and 4PHV (HIV protease). Among these, the coverage of the eraser active site is quite low for p38 kinase and thrombin, namely 32 and 55%, respectively. Fig. 13 displays the sites obtained from the X-ray ligand and the site computed by the eraser-based site finding algorithm. The sites are quite dissimilar and accordingly the docked structure is different from the X-ray pose. For 1APT (22 rotatable torsions) and 4PHV (15 rotatable bonds), the conformational search procedure is unable to find the correct pose *within the specified number of iterations*. Using a very large number of iterations for 4PHV we have verified convergence to the X-ray structure, although we have been unable to verify this for the case of 1APT. However, since we have verified that the rigid docking of the X-ray ligand to the site reproduces the X-ray pose within an RMS of 0.24 Å for 1APT, we conclude that this failure to reproduce the X-ray pose arises from the inability of the Monte Carlo sampling to find the X-ray conformation within the specified number of trials.

3.5. Thymidine kinase receptor: an example of high-throughput screening using LigandFit

As an example of the feasibility of employing LigandFit for virtual high-throughput screening (vHTS), we investigated the thymidine kinase (TK) receptor studied recently by Rognan and coworkers [50]. Nine inhibitors are known for this receptor and their measured binding affinity is in the micromolar range.

The following protocol was employed for performing the vHTS for the TK receptor.

1. The three-dimensional structure of the receptor was taken from the pdb file 1kim. The ligand and all the water molecules were removed. Hydrogen atoms were added using the Cerius^2 templates for the protein residues.
2. Three-dimensional structures of nine known TK inhibitors identified by Rognan and coworkers [50] were obtained either from X-ray data (where available) or by sketching them in Cerius^2. In addition, 993 random single-fragment molecules were extracted from the WDI database with molecular weight less than 500. The structures were assigned partial charges using the Gasteiger method [51] as implemented in Cerius^2. They were then

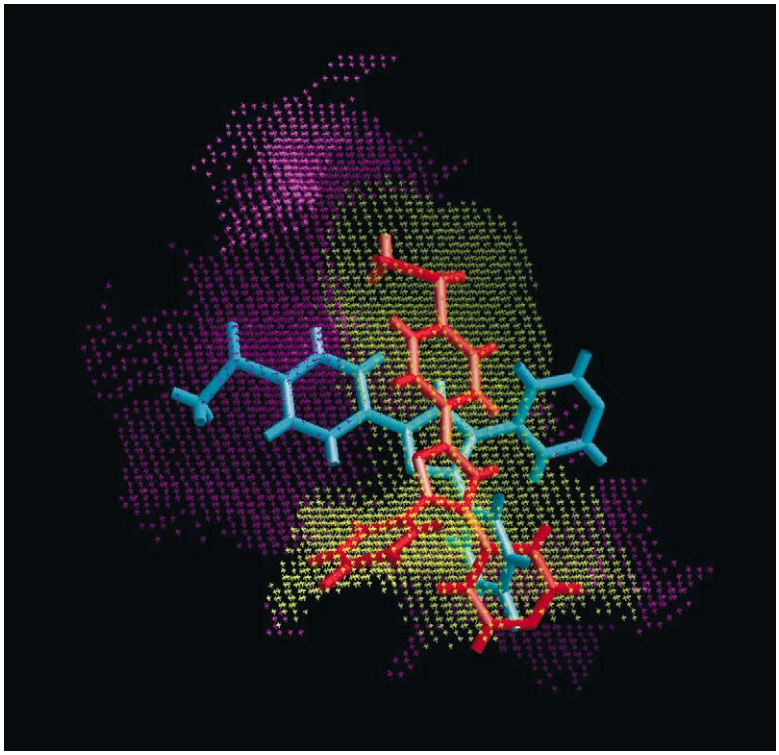

Fig. 13. Flexible docking with RBM of 1A9U ligand (p38 kinase). The X-ray ligand pose is shown in red and the site determined from the X-ray ligand is shown in yellow. The site obtained using an 8 Å eraser is shown in magenta and the flexibly docked structure using eraser-based site is shown in cyan. In this example, automatic site detection using the eraser algorithm performs poorly in identifying the binding region for this ligand and the docking based on site shape does not reproduce the binding mode seen in the X-ray structure.

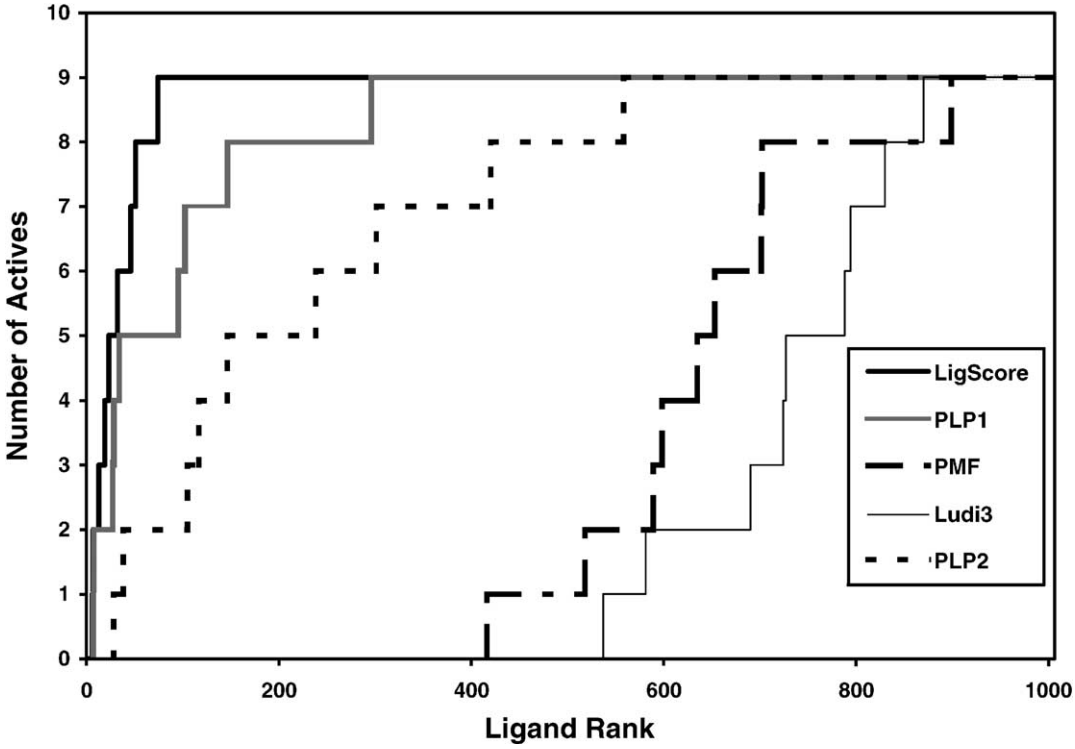

Fig. 14. Hit rate plotted against ligand rank corresponding to scoring functions LigScore, PLP1, PLP2, PMF and Ludi3.

assigned random conformations followed by energy minimization using the Dreiding force field [41]. This resulted in a dataset of 1002 molecules.

3. Using the site search algorithm of LigandFit, the largest cavity detected using an eraser size of 5 Å was selected. The site was verified with the location of the docked ligand and in 1kim from X-ray data. All 1002 ligands prepared in Step 2 were then docked to the site using the following conditions: $N_{\text{MaxTrials}} = 5000$; $N_{\text{save}} = 20$; CFF force field; distance dependent dielectric $\varepsilon = 4r$; soft potential with $\alpha = 1$ and $\beta = 0.05$; improved interpolation with $m = n = 2$; grid resolution of 0.5 Å. Docked conformations obtained for each molecule were clustered using a Leader algorithm in Cerius^2 with RMS threshold of 2.5 Å. This resulted in 2402 docked structures. The total computing time for docking all 1002 ligands was 12 h and 13 min using a single processor R10000 Silicon Graphics workstation. The average time for each molecule is 43.9 s.
4. For all the docked structures obtained in Step 3, we computed the scores using the scoring functions LigScore [1], Ludi [44], PLP version 1 [45,46], PLP version 2 [52]and PMF [54].
5. Considering each scoring function in turn, the highest scoring docked structure was selected for each molecule. This resulted in a set of 1002 docked molecules with scores corresponding to each scoring function. For each scoring function, the 1002 molecules were sorted according to their score. The ranks of the nine active ligands were determined for each scoring set. Fig. 14 shows the hit rate as a function of rank for each scoring function.

In this test, LigScore gives the best hit-rate identifying all the known actives within the first 100 compounds while PLP version 1 gives seven actives within the first 100 compounds. PLP version 2 finds seven actives within first 300 compounds. With both Ludi and PMF functions, more than 400 structures have to be screened before retrieving the first known active. One reason for the poor performance of the Ludi and PMF scoring functions is their inability to adequately penalize ligands with close contacts. Ludi score ignores close contacts and the PMF function [47] exhibits very mild repulsion at small interatomic distances. To be

Table 4
RMS values of the docked poses of thymidine kinase inhibitors to the X-ray structures

TK inhibitor (pdb code)	Rank	RMS (Å)	LigScore (CFF)
1ki7	6	0.71	5.31
1kim	7	0.64	5.26
1e2m	13	0.67	5.12
1e2k	19	0.56	4.74
2ki5	23	3.99	4.64
1ki3	32	3.71	4.6
1ki2	46	3.44	4.35
1ki6	51	0.78	4.26
1e2n	74	3.54	3.97

The LigScore values and the rank of the ligand in the screening using LigScore are also tabulated.

Table 5
Example data illustrating consensus scoring using four scoring functions on five models using $s = 40\%$

Model	Score A		Score B		Score C		Score D		Consensus score
	Score	Rank score	Score	Rank score	Score	Rank score	Score	Rank score	
1	12	**0**	55	**0**	43	**0**	241	**0**	0
2	22	**0**	46	**0**	113	**0**	283	**1**	1
3	112	**1**	92	**1**	221	**1**	299	**1**	**4**
4	78	**0**	82	**1**	182	**0**	251	**0**	1
5	98	**1**	77	**0**	192	**1**	263	**0**	2

employed with docked structures, the PMF function will require a modification that adds a stronger repulsive part at small interatomic separation. We have verified that with suitable modification of the PMF function by adding such a repulsive term at short interatomic distances, one obtains an improved hit rate [53]. The ability of both the LigScore and PLP functions in identifying most of the actives tends to validate the use of these scoring functions to analyze the results from LigandFit for this system.

3.6. Quality of the docking of the TK inhibitors

To evaluate the quality of the docking of the TK inhibitors, we have compared the docked poses to those found in the X-ray structures. In practice, however, this is somewhat complicated due to the fact that in the high-throughput screening example, we have employed the protein structure found in the 1kim complex. The nine protein–ligand complexes available from X-ray studies show slight variations in the protein three-dimensional structure even though the same protein is present in all the complexes. Some of the variations involve changes in some side-chain conformations near the active site. To compare the poses obtained from docking the known inhibitors to the 1kim protein structure with the X-ray structure of the corresponding protein–ligand complex, we have superimposed the 1kim protein structure to the other X-ray protein structures by superimposing corresponding backbone atoms in the vicinity of the active site, thereby transforming the docked ligand poses such that they are suitable for direct comparison with the poses seen in X-ray studies. Table 4 shows the RMS of the best docked pose to the X-ray pose for all the TK inhibitors. For five out of nine actives, satisfactory RMS values ranging from 0.56 to 0.78 Å are obtained. In the protein–ligand complexes 2ki5, 1ki2 and 1ki3, some protein side chains are in different conformations from the 1kim protein structure employed in the screening studies here. This may explain the large RMS found in the docking of these ligands to the 1kim protein structure. However, the case of the ligand hmtt in pdb structure 1e2n cannot be explained on this basis. While further detailed docking studies will be required to understand the reason for poor docking for this ligand to the 1kim structure, we have found that using the Dreiding force field instead of CFF yields a satisfactory pose with RMS of 1.51 Å to the X-ray pose. Further studies are also in progress to analyze the docking of TK inhibitors to their native protein structures.

3.7. Consensus scoring

Scoring functions differ in the ligand–receptor descriptors they employ to evaluate the quality of the docking. Depending on these descriptors, the importance of various factors that can influence ligand binding such as nonpolar interactions, polar interactions, solvent effects, loss of entropy of ligand upon binding, etc. may vary. The current state of development of scoring functions is such that no single function has been found to handle all protein complexes equally well. Therefore, it has been found to be beneficial to consider several scoring functions in evaluating docked structures. For example, Walters and coworkers [54] report a consensus scoring procedure that employs several scoring functions. We have employed a modified consensus scoring procedure described as follows.

Considering a set of molecules scored according to a set of scoring functions, assign a rank score of 1 to the top s% scoring molecules for a given score and assign a rank score of 0 to all the remaining molecules with lower score. Repeat this procedure for each scoring function. This results in assigning a set of binary (0 or 1) rank scores for each molecule corresponding to each scoring function. For each molecule, add all its binary rank scores to obtain its consensus score. The consensus score is an integer with maximum value equal to the number of scoring functions employed and minimum value of zero. Table 5 illustrates this simple procedure with 40% consensus scoring (i.e. $s = 40\%$). Recently, a similar consensus scoring method has been reported [55] where a rank transform is applied to individual scores before summing them to obtain a consensus score.

Performing consensus scoring using this procedure with $s = 10\%$ with LigScore and PLP version 1 and then sorting all the 1002 docked structures according to the consensus scores results in retrieving 7 actives within the first 34 molecules.[6] The use of consensus scoring improves the hit rate. Though we lose two actives, we have far fewer inactives.

[6] As seen earlier, with Ligscore, the first 100 molecules contain all 9 actives. With PLP v1, the first 100 molecules contain only 7 actives. It turns out there are only 27 non-active molecules common to both sets.

4. Summary

We have presented a shape-directed docking methodology for accurately docking ligands to protein active sites. The method uses a site detection algorithm for identifying candidate active sites within the protein. In cases where a known X-ray ligand–protein complex is available, the docked ligand structure may also be used to define the active site. A Monte Carlo conformational search procedure is used for generating candidate ligand docking conformations. A shape comparison filter is then used to evaluate each ligand conformation against the active site shape. Ligand conformations satisfying the shape comparison filter are initially docked into the active site via a shape alignment protocol based on principal axes and moments. These initial poses are further refined via a grid-based energy calculation which rapidly evaluates protein–ligand interaction energies. A novel method of dramatically reducing the errors arising from interpolation of the grid energies has been presented resulting in an accurate energy evaluation. The docking method also employs a rapid rigid body minimization of the ligand with respect to the grid-based interaction energy. We have shown that the docking procedure works very well to obtain ligand poses close to X-ray poses in 14 out of 19 diverse protein–ligand complexes for which a reasonable definition of the binding site is available. The method, in addition to being accurate, is quite fast, on average, taking under 7 s per ligand docking on a Linux PC with an AMD Athlon 1.4 GHz processor. As demonstrated with the thymidine kinase receptor, the docking methodology can be successfully employed for performing accurate high-throughput virtual screening when combined with our scoring function, LigScore, to prioritize the hits.

We have also discussed several areas where there is room for improvement. As seen in the case of p38 kinase (1A9U), when the binding site is much larger than the ligand, the fully automated method fails to identify the X-ray ligand pose. This failure can be attributed to lack of adequate ligand pose sampling in the large active site as well as deficiencies in the shape alignment algorithm for such a large active site (as compared to the ligand-based site). While for such cases, one may improve results with manual editing of the site; work is currently in progress for improving the ligand pose sampling in such cases using partitioning of the site into smaller sites in an automated manner. Furthermore, other approaches employing feature-based docking, such as the docking method of Diller and Merz [27], are available that handle such cases satisfactorily.

Acknowledgements

The authors would like to thank Paul Kirchhoff, André Krammer, Jürgen Koska, Eric Jamois, Osman Güner, Shashidhar Rao, Rob Brown, Rod Hubbard and Scott Kahn for many valuable discussions during this study.

References

[1] A. Krammer, P.D. Kirchhoff, J. Jiang, C.M. Venkatachalam, M. Waldman, LIGSCORE: a new scoring function to evaluate and prioritize ligands docked to protein active sites (in preparation patent pending).
[2] N. Pattabhiraman, M. Levitt, T.E. Ferrin, R. Langridge, Computer graphics in real-time docking with energy calculation and minimization, J. Comp. Chem. 6 (5) (1985) 432–436.
[3] R.L. DesJarlais, R.P. Sheridan, J.S. Dixon, I.D. Kuntz, R. Venkataraghavan, Docking flexible ligands to macromolecular receptors by molecular shape, J. Med. Chem. 29 (11) (1986) 2149–2153.
[4] R.L. DesJarlais, R.P. Sheridan, G.L. Seibel, J.S. Dixon, I.D. Kuntz, R. Venkataraghavan, Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure, J. Med. Chem. 31 (4) (1988) 722–729.
[5] E.C. Meng, B.K. Shoichet, I.D. Kuntz, Automated docking with grid-based energy evaluations, J. Comp. Chem. 13 (4) (1992) 505–524.
[6] D.J. Bacon, J. Moult, Docking by least-squares fitting of molecular surface patterns, J. Mol. Biol. 225 (3) (1992) 849–858.
[7] S. Sudarsanam, G.D. Virca, C.J. March, S. Srinivasan, An approach to computer-aided inhibitor design: application to cathepsin L, J. Comput. Aided Mol. Des. 6 (3) (1992) 223–233.
[8] R.L. DesJarlais, J.S. Dixon, A shape- and chemistry-based docking method and its use in the design of HIV-1 protease inhibitors, J. Comput. Aided Mol. Des. 8 (3) (1994) 231–242.
[9] B.A. Luty, Z.R. Wasserman, P.F.W. Stouten, C.N. Hodge, A molecular mechanics/grid method for evaluation of ligand–receptor interactions, J. Comp. Chem. 16 (4) (1995) 454–464.
[10] C.M. Oshiro, I.D. Kuntz, J.S. Dixon, Flexible ligand docking using a genetic algorithm, J. Comput. Aided Mol. Des. 9 (2) (1995) 113–130.
[11] D.S. Goodsell, G.M. Morris, A.J. Olson, Automated docking of flexible ligands: applications of AutoDock, J. Mol. Recog. 9 (1) (1996) 1–5.
[12] D.A. Gschwend, I.D. Kuntz, Orientational sampling and rigid-body minimization in molecular docking revisited: on-the-fly optimization and degeneracy removal, J. Comput. Aided Mol. Des. 10 (2) (1996) 123–132.
[13] G.M. Morris, D.S. Goodsell, R. Huey, A.J. Olson, Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4, J. Comput. Aided Mol. Des. 10 (4) (1996) 293–304.
[14] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, A fast flexible docking method using an incremental construction algorithm, J. Mol. Biol. 261 (1996) 470–489.
[15] W. Welch, J. Ruppert, A.N. Jain, Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites, Chem. Biol. 3 (6) (1996) 449–462.
[16] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking, J. Mol. Biol. 267 (1997) 727–748.
[17] M. Hahn, Three-dimensional shape-based searching of conformationally flexible compounds, J. Chem. Inf. Comput. Sci. 7 (1) (1997) 80–86.
[18] Y. Sun, T.J. Ewing, A.G. Skillman, I.D. Kuntz, CombiDOCK: structure-based combinatorial docking and library design, J. Comput. Aided Mol. Des. 12 (6) (1998) 597–604.
[19] J.-Y. Trosset, H.A. Scheraga, Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 8011–8015.
[20] M. Liu, S. Wang, MCDOCK: a Monte Carlo simulation approach to the molecular docking problem, J. Comput. Aided Mol. Des. 13 (5) (1999) 435–451.

[21] S. Makino, T.J. Ewing, I.D. Kuntz, DREAM++: flexible docking program for virtual combinatorial libraries, J. Comput. Aided Mol. Des. 13 (5) (1999) 513–532.
[22] J. Wang, P.A. Kollman, I.D. Kuntz, Flexible ligand docking: a multistep strategy approach, Proteins 36 (1) (1999) 1–19.
[23] D.A. Cosgrove, D.M. Bayada, A.P. Johnson, A novel method of aligning molecules by local surface shape similarity, J. Comput. Aided Mol. Des. 14 (6) (2000) 573–591.
[24] X. Fradera, R.M.A. Knegtel, J. Mestres, Similarity-driven flexible ligand docking, Proteins 40 (2000) 623–636.
[25] B.B. Goldman, W.T. Wipke, QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock), Proteins 38 (1) (2000) 79–94.
[26] L. David, R. Luo, M.K. Gilson, Ligand–receptor docking with the mining minima optimizer, J. Comput. Aided Mol. Des. 15 (2) (2001) 157–171.
[27] D.J. Diller, K.M.J. Merz, High throughput docking for library design and library prioritization, Proteins 43 (2001) 113–124.
[28] T.J. Ewing, S. Makino, A.G. Skillman, I.D. Kuntz, DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases, J. Comput. Aided Mol. Des. 15 (5) (2001) 411–428.
[29] P.D. Kirchhoff, R. Brown, S. Kahn, M. Waldman, C.M. Venkatachalam, Application of structure-based focusing to the estrogen receptor, J. Comp. Chem. 22 (10) (2001) 993–1003.
[30] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, T.E. Ferrin, A geometric approach to macromolecule–ligand interactions, J. Mol. Biol. 161 (2) (1982) 269–288.
[31] T.J. Oldfield, A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent, Acta Cryst. D57 (2001) 82–94.
[32] T.J. Oldfield, X-LIGAND: an application for the automated addition of flexible ligands into electron density, Acta Cryst. D57 (2001) 696–705.
[33] J.D. Foley, A. Van Dam, Fundamentals of Interactive Computer Graphics, Addison-Wesley, Reading, MA, 1982, p. 664.
[34] D.F. Rogers, Procedural Elements for Computer Graphics, McGraw-Hill, New York, 1985, p. 433.
[35] R. Fletcher, Practical methods of optimization, in: Unconstrained Optimization, vol. 1, 1980, Wiley, New York.
[36] L.E. Scales, Introduction to Non-Linear Optimization, Springer-Verlag, New York, 1985, p. 243.
[37] U. Dinur, A.T. Hagler, New approaches to empirical force fields, in: Reviews of Computational Chemistry, 1991 (Chapter 4).
[38] J.R. Maple, U. Dinur, A.T. Hagler, Proc. Natl. Acad. Sci. U.S.A. 85 (1988) 5350–5354.
[39] Z. Peng, C.S. Ewig, M.-J. Hwang, M. Waldman, A.T. Hagler, Derivation of class II force fields. 4. van der Waals parameters of alkali metal cations and halide anions, J. Phys. Chem. A 101 (39) (1997) 7243–7252.
[40] C.S. Ewig, R. Berry, U. Dinur, J.-R. Hill, M.-J. Hwang, H. Li, C. Liang, J. Maple, Z. Peng, T.P. Stockfisch, T.S. Thacher, L. Yan, X. Ni, A.T. Hagler, Derivation of class II force fields. VIII. derivation of a general quantum mechanical force field for organic compounds, J. Comp. Chem. 22 (15) (2001) 1782–1800.
[41] S.L. Mayo, B.D. Olafson, W.A.I. Goddard, DREIDING: a generic force field for molecular simulations, J. Phys. Chem. 94 (1990) 8897–8909.
[42] D.J. Oberlin, H.A. Scheraga, B-Spline method for energy minimization in grid-based molecular mechanics calculations, J. Comp. Chem. 19 (1) (1998) 71–85.
[43] M. Levitt, Protein folding by restrained energy minimization and molecular dynamics, J. Mol. Biol. 170 (3) (1983) 723–764.
[44] H.J. Bohm, Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or three-dimensional database search programs, J. Comput. Aided Mol. Des. 12 (4) (1998) 309–323.
[45] D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.B. Fogel, L.J. Fogel, S.T. Freer, Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming, Chem. Biol. 2 (1995) 317–324.
[46] G.M. Verkhivker, D. Bouzida, D.K. Gehlhaar, P.A. Rejto, S. Arthurs, A.B. Colson, S.T. Freer, V. Larson, B.A. Luty, T. Marrone, P.W. Rose, Deciphering common failures in molecular docking of ligand–protein complexes, J. Comput. Aided. Mol. Des. 14 (8) (2000) 731–751.
[47] I. Muegge, Y.C. Martin, A general and fast scoring function for protein–ligand interactions: a simplified potential approach, J. Med. Chem. 42 (5) (1999) 791–804.
[48] D. Rogers, A.J. Hopfinger, Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure property relationships, J. Chem. Inf. Comput. Sci. 34 (4) (1994) 854–866.
[49] D. Rogers, Evolutionary statistics: using a genetic algorithm and model reduction to isolate alternate statistical hypothses of experimental data, in: Proceedings of the 7th International Conference on Genetic Algorithm, East Lansing, MI, 1997.
[50] C. Bissantz, G. Folkers, D. Rognan, Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations, J. Med. Chem. 43 (25) (2000) 4759–4767.
[51] J. Gasteiger, M. Marsili, A new model for calculating atomic charges in molecules, Tetrahedron Lett. (1978) 3181–3184.
[52] D.K. Gehlhaar, D. Bouzida, P.A. Rejto, Reduced dimensionality in ligand–protein structure prediction: covalent inhibitors of serine proteases and design of site-directed combinatorial libraries, in: A.L. Parrill, M.R. Reddy (Eds.), Rational Drug Design: Novel Methodology and Practical Applications, vol. 719, American Chemical Society, Washington, DC, 1999, pp. 292–311.
[53] X. Jiang, unpublished work.
[54] P.S. Charifson, J.J. Corkery, M.A. Murcko, W.P. Walters, Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins, J. Med. Chem. 42 (1999) 5100–5109.
[55] R.D. Clark, A. Strizhev, J.M. Leonard, J.F. Blake, J.B. Matthew, Consensus scoring for ligand/protein interactions, J. Mol. Graphics Modell. 20 (2002) 281–295.