FDS: Flexible Ligand and Receptor Docking with a Continuum Solvent Model and Soft-Core Energy Function

RICHARD D. TAYLOR,[1,*] PHILIP J. JEWSBURY,[2] JONATHAN W. ESSEX[1]
[1]*Department of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, UK*
[2]*AstraZeneca, Mereside, Alderley Park, Macclesfield, Cheshire, SK10 4TG, UK*

Received 13 August 2002; Accepted 5 March 2003

Abstract: The docking of flexible small molecule ligands to large flexible protein targets is addressed in this article using a two-stage simulation-based method. The methodology presented is a hybrid approach where the first component is a dock of the ligand to the protein binding site, based on deriving sets of simultaneously satisfied intermolecular hydrogen bonds using graph theory and a recursive distance geometry algorithm. The output structures are reduced in number by cluster analysis based on distance similarities. These structures are submitted to a modified Monte Carlo algorithm using the AMBER-AA molecular mechanics force field with the Generalized Born/Surface Area (GB/SA) continuum model. This solvent model is not only less expensive than an explicit representation, but also yields increased sampling. Sampling is also increased using a rotamer library to direct some of the protein side-chain movements along with large dihedral moves. Finally, a softening function for the nonbonded force field terms is used, enabling the potential energy function to be slowly turned on throughout the course of the simulation. The docking procedure is optimized, and the results are presented for a single complex of the arabinose binding protein. It was found that for a rigid receptor model, the X-ray binding geometry was reproduced and uniquely identified based on the associated potential energy. However, when side-chain flexibility was included, although the X-ray structure was identified, it was one of three possible binding geometries that were energetically indistinguishable. These results suggest that on relaxing the constraint on receptor flexibility, the docking energy hypersurface changes from being funnel-like to rugged. A further 14 complexes were then examined using the optimized protocol. For each complex the docking methodology was tested for a fully flexible ligand, both with and without protein side-chain flexibility. For the rigid protein docking, 13 out of the 15 test cases were able to find the experimental binding mode; this number was reduced to 11 for the flexible protein docking. However, of these 11, in the majority of cases the experimental binding mode was not uniquely identified, but was present in a cluster of low energy structures that were energetically indistinguishable. These results not only support the presence of a rugged docking energy hypersurface, but also suggest that it may be necessary to consider the possibility of more than one binding conformation during ligand optimization.

 J Comput Chem 24: 1637–1656, 2003

Key words: flexible ligand docking; flexible protein; Monte Carlo; GB/SA; rotamer library

Introduction

The prediction of small molecule binding modes to macromolecules of known three-dimensional structure is often referred to as the "docking" problem.[1] The results from such studies can be used not only to direct structure based drug design, but also to analyze key interactions between a ligand and receptor. Moreover, the application of docking methods in structure based virtual screening using small molecule databases is now becoming routine.[2] It has been shown with numerous algorithms that for an idealized system comprising a flexible ligand, a well-defined rigid receptor, and a known active site center, the ligand can be docked to the target in a short time frame[3] with reasonable success. It is difficult to compare objectively the success of the current algorithms because most programs are validated on data sets of protein–ligand complexes that vary considerably in both size and diversity. However, given the rigid receptor model, success rates for elucidating the crystallographic binding mode to within 2 Å RMSD are typically

Correspondence to: J. W. Essex; e-mail: J.W.Essex@soton.ac.uk

*Present address: Astex Technology Ltd., 436 Cambridge Science Park, Cambridge, CB4 0QA, UK

Contract/grant sponsor: AstraZeneca (to R.D.T.)

quoted between 50% and 80%,[4–6] using data sets containing approximately 100 complexes.

The notion of a rigid conformation for the receptor is not necessarily valid, particularly if the apo-protein structure undergoes conformational change upon complexation. This can be manifested as an extreme backbone movement as seen in the HIV-1 protease[7] complex, where the protein undergoes a "hinge" type conformational change upon binding. Other proteins such as Neuraminidase[8] exhibit side-chain movement coupled with explicit water expulsion when different substrates are bound to the receptor. Thus, incorporating conformational flexibility of both the ligand and the protein in a docking procedure is arguably important to analyze a novel substrate or modifications of a known ligand binder. Existing methods typically use multiple rigid protein conformations and sequentially dock against each receptor conformation.[9] Alternatively, an average description of several protein conformations is used.[10] We have previously summarized a large number of docking methods along with the techniques used to incorporate protein flexibility, either implicitly or explicitly, during the docking of small molecules.[11]

A further issue in the docking problem is the treatment of solvation. Continuum solvent models have been widely used to provide computationally inexpensive solvent descriptions.[12] To date, these models have primarily been used in docking studies to score structures generated in the gas phase, rather than directing the course of a simulation in the solution phase.[13,14] This work has focused on implementing a continuum solvent model that not only provides an inexpensive solvent description but also permits increased sampling.

The aim of this work was to develop a complete suite of docking algorithms that address the docking paradigm and to assess the approximations associated with existing techniques. The complete set of algorithms were designed to fulfill the following criteria:

1. Provide an accurate thermodynamic description of the complex that is incorporated into a docking strategy, with a solvent description that is both realistic and computationally tractable.
2. Incorporate flexible behavior of both receptor and ligand during the docking procedure.
3. Compare and contrast the energy hypersurface for the rigid and flexible receptor models.
4. Demonstrate the searching function's ability to overcome high energy barriers and adequately sample the energy hypersurface.
5. Provide an extendable platform for future development.

Overview of the Method

The simulation methodology reported here is a hybrid approach where the first stage is a dock based on deriving sets of simultaneously satisfied hydrogen bonds using graph theory and a recursive distance geometry algorithm. For this stage the protein is treated as a rigid receptor. The output structures are reduced in number by cluster analysis based on distance similarities. These structures are then submitted to a modified version of the Monte Carlo program MCPRO 1.4[15] using the AMBER All Atom[16] molecular mechanics force field. Sampling is increased using a rotamer library to direct some of the protein side-chain movements along with large dihedral moves, and a softening function for the nonbonded force field terms. The Generalized Born/Surface Area (GB/SA)[12] continuum solvent model is also used. This solvent model is not only less expensive than an explicit representation but also yields increased sampling. A new parametrization is also developed for the GB/SA model that is consistent with the AMBER-AA force field.

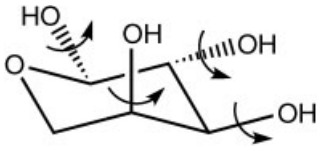

Figure 1. α-L-Arabinose from 1ABE complex with L-Arabinose-binding protein.

A single test system, the α-L-Arabinose ligand (Fig. 1) bound to L-Arabinose-Binding Protein from Escherichia coli (1ABE)[17] was taken from the test set of GOLD[5] (an existing docking program) and was used to parametrize the soft-core function, and refine the simulation protocol. An additional 14 systems from the same test set were subsequently used to further validate the suite of algorithms and the simulation protocol.

In what follows, each stage of the algorithm will be considered in turn, followed by a discussion of the results for the 1ABE complex. The results for the additional 14 systems are summarized, and will be discussed in more detail elsewhere.[18]

Geometry-Based Docking

The first stage of this method uses a flexible ligand and rigid receptor approximation in a hydrogen-bond directed docking. This geometry based docking is used to determine diverse multiple starting points that are submitted to the Monte Carlo algorithm. Although this first stage is computationally inexpensive compared to the Monte Carlo component, the scoring function, which is based solely on satisfying hydrogen bond geometries, is obviously less rigorous than a full molecular mechanics force field. However, as a first approximation this is a method that generates diverse structures sampling potentially important regions of the active site.

This preliminary rigid receptor dock is a three-stage process: first, the hydrogen bond motifs or cliques are determined; second, embedding is used to generate structures that satisfy the cliques; third, the cliques are clustered to produce approximately five starting conformations for the Monte Carlo dock.

The limiting step for this preliminary dock is the embedding algorithm using the program DGEOM95,[19] which takes approximately 3 min per clique, using a MIPS R12000 processor. In our hands, an average of 100 cliques can therefore be submitted to the embedding algorithm, DGEOM95, based on hardware constraints. The complexity of the calculation is simplified by constraining the protein to the crystallographic coordinates and keeping all ring systems, both aliphatic and aromatic, rigid.

Clique Detection

The hydrogen bond docking was implemented using a distance geometry protocol that describes the protein–ligand complex in distance space. All intramolecular atom pair distances in the crystallographic structures are calculated, along with the upper and lower bounds on these distances that effectively include intramolecular flexibility. For the initial distance geometry docking phase, the active site is assumed rigid, and hence the protein upper and lower bounds are equal and identical to the interatomic distances calculated from the crystallographic starting structure. However, dihedral flexibility is permitted for the ligand; thus, the upper and lower bounds for 1,4 (and greater) ligand intramolecular distances are assigned with DGEOM95[19] using the program's default rules.

The distance geometry method can be explained by considering two intermolecular hydrogen bonds between the protein (p) and ligand (l). For the two hydrogen bonds to simultaneously occur the following equations must be satisfied,

$$D_{p1,p2} \leq U_{l1,l2} + \delta \quad (1)$$

$$D_{p1,p2} \geq L_{l1,l2} - \delta \quad (2)$$

where $D_{p1,p2}$ is the distance between protein atoms $p1$ and $p2$ and $U_{l1,l2}$ and $L_{l1,l2}$ are the respective upper and lower bounds for the distances between ligand atoms $l1$ and $l2$. The parameter δ is an approximate measure of the hydrogen bond distance. In graph theory terminology a hydrogen bonded atom pair represents a node and an edge occurs between two nodes when two hydrogen bonds can be simultaneously satisfied. A clique is defined as a list of hydrogen bonded atom pairs (one from the protein and one from the ligand) that can be simultaneously satisfied given the geometrical constraints of eqs. (1) and (2). Thus, finding all cliques of the docking graph finds all maximal hydrogen bond matchings of the ligand to the protein active site, where each match must be between donors and acceptors, and not mismatches of donors with donors or acceptors with acceptors.

The clique finding technique is a modified version of the recursive algorithm by Bron and Kerbosch,[20] which has been applied to hydrogen bond searches by Smellie et al.[21] A recent analysis showed this to be the most effective approach for finding all maximal cliques.[22] The method uses a backtracking branch-and-bound technique whereby branches are eliminated that cannot lead to a clique, thereby ensuring an efficient search of the minimum number of branches in the graph tree.

Clique Filtering

The clique analysis calculates multiple sets of hydrogen bond constraints that are subsequently used to produce three-dimensional docked structures in an embedding process. Although the size of the docking tree is undetermined at the start of the algorithm because it is an NP complete problem, the number of cliques can be of the order of 10^3. Several filters or constraints were applied during the clique detection and to post-process the cliques, to reduce both the number and the size of the cliques. Based on hardware considerations, approximately 100 cliques are required for the embedding process. As such the number of hydrogen bonds per clique (or cardinality) was restricted. The clique finding algorithm was repeated with the cardinality increasing from an initial value of two until the required number of cliques were generated.

Increasing the number of constraints, particularly for a single atom, in the embedding process drastically reduces the probability of a successful embed, as such only one-to-one mappings are therefore permitted in a clique, that is, each atom may be involved in only one hydrogen bond. The rationale for this decision is that a single hydrogen bond between a protein atom and ligand atom will dictate a similar region of the binding site as multiple hydrogen bonds to those atoms, recalling that the purpose of this initial docking phase is to determine diverse starting points for the Monte Carlo method. These structures should therefore sample a diverse range of potentially important regions of the active site. A similar technique was used by Smellie et al.,[21] but this was not a post-process filter and rejected a clique if the number of edges per node exceeded a user defined upper limit. This methodology was not adopted, as possible solutions that may sample important regions of the active can be rejected. By permitting only one-to-one mappings, although the clique size was reduced, identical cliques were sometimes found. Thus, once the clique filter was applied all identical cliques were removed and the embedding process was started.

Clique Embedding

The output from the clique detection and filter is a list of hydrogen bonded atom pairs that can simultaneously satisfy the bounds matrix for the ligand and the distance matrix for the protein. Cartesian coordinates are generated from these constraints using the embedding algorithm in DGEOM95,[19] which attempts to satisfy simultaneously three sets of distance bounds, namely, the intermolecular hydrogen bonds, the ligand intramolecular distance bounds, and the bounded distance from the active site center to the ligand, while eliminating atomic overlap.

Clique Clustering

Hardware considerations dictated that approximately five structures are submitted to the Monte Carlo algorithm. However, the number of embedded structures is of the order 10^2. Therefore, the structures determined from the clique analysis and embedding process are analyzed using clustering techniques to identify similar binding motifs. For simplicity, molecular similarity was determined using a basic root-mean-square distance (RMSD) score to produce diverse structures representative of cluster sets that sampled important regions of the active site.

In this clustering algorithm, seeds are required from each cluster. Each cluster is simply a set of near-neighbors, where a near-neighbor is defined as a structure with an RMSD within a user defined range, typically 3 Å. From these lists the structure with the largest number of near-neighbors is always chosen and will be referred to as a seed. The motivation for this choice of seed is the larger the number of near-neighbors the more statistically important that region. The seed and all of the members of the cluster are then removed from all other lists and the next seed (i.e., the structure with the largest number of near-neighbors) is chosen, and the procedure repeated.

Table 1. Clique Calculation Results for Two Values of δ, Giving the Total Number of Cliques Found (Cliques), the Largest Number of Crystallographic Hydrogen Bonds (H-Bonds) in a Single Clique and the Total Embed Time for All Cliques.

		Nonfiltered		Filtered		
Cardinality	δ/Å	Cliques	H-bonds	Cliques	H-bonds	Embed time/hours
2	3.5	742	4	128	2	6.4
2	5.0	15262	8	698	5	34.9[a]

[a]Estimated embedding time.

It should also be noted that the RMSD cutoff for the near-neighbors was automatically increased or reduced to give approximately five seeds. An additional consideration in the clustering algorithm is the analysis of singletons, where a singleton is defined as a structure with no near-neighbors. These were included in the seed count when the RMSD cutoff was systematically altered, although only true singletons were considered; false singletons were discarded. A false singleton is defined as a singleton which previously had a near-neighbor set, which was eliminated when another group of the clustered structures were removed from all other near-neighbor sets.

Geometry-Based Docking Optimization

In this first stage of geometry-based docking the crucial parameters for clique generation are the cardinality and delta (δ) value. The cardinality describes the minimum number of hydrogen bonds in a clique, which is automatically incremented by one in a step-wise fashion from an initial value of 2, until such time as between 50 and 500 cliques are generated. A δ value of 3.5 Å is used, which corresponds to an approximate hydrogen bond distance between two nonhydrogen atoms. Having generated cliques using the modified Bron and Kerbosch algorithm[20] the postprocess filter algorithm was applied to each set, which ensures only one-to-one mappings, and that all degenerate sets are removed. The choice of δ can be justified by considering the prefilter and postfilter clique sets for the 1ABE complex. Table 1 summarizes the difference in the maximum number of crystallographic hydrogen bonds generated in a single clique, before and after the application of the clique filter, for the 1ABE test system using a δ value of 3.5 and 5.0 Å and a constant cardinality of 2.

It is clear from Table 1 that the modified Bron and Kerbosch algorithm with $\delta = 5.0$ Å produced a single nonfiltered clique containing all eight hydrogen bonds observed in the crystallographic structure. This result demonstrates the success of the algorithm, in that the complete hydrogen bond network can be found for a complex hydrogen bond system. This was the lowest value of δ required to generate a clique with the full complement of hydrogen bonds observed in the crystal structure. However, utilizing this protocol produces an impracticable number of cliques even after the postprocess filter is applied. Although the embed for a δ value of 5.0 Å was not attempted, the estimated time is 34.9 h based on each embed requiring approximately 3 min on a MIPS R12000 processor.

A corollary of relaxing this δ constraint is the production of a large number of cliques, where not only is the average clique size increased, but many cliques are produced using unrealistically wide bounds matrices. The probability of a successful embed is subsequently reduced due to both the unfeasible bound set and the increase in the average clique size, which results in an increase in the number of constraints per clique. It is worth noting that the main purpose of the clique generation is to provide diverse sets of starting structures for the Monte Carlo algorithm that contain some, but not necessarily all, hydrogen bonds observed in the crystallographic structures. Furthermore, the application of the filter to the clique sets was crucial to a successful embed, as a large number of constraints greatly reduces the probability of embedding the ligand into the active site. For these reasons a small value of δ is considered acceptable, even though the probability of finding the X-ray structure in the first phase is reduced.

Embedding the cliques using the program DGEOM95 is the next stage in the clique analysis followed by the clustering protocol. This process involves the systematic modification of the RMSD cutoff such that approximately five seeds are generated. As previously stated, this number is a consequence of hardware restrictions. The clustered seeds, which include all true singletons, are a diverse set of structures sampling potentially important regions of the active site.

Energy Based Docking

The diverse seeds generated from the geometry based docking are used as multiple starting points to the Monte Carlo energy based docking. The algorithm is based around the MCPRO 1.4 program[15] with a continuum solvent model, soft-core annealing and explicit protein side chain movement.

Continuum Solvation Model

To determine the most probable binding modes of a ligand–protein complex using a molecular mechanics force field it is important to consider how the solvent affects the behavior of the system. A solution to this problem would be to model the solvent molecules explicitly as an integral part of the system. However, for large biological simulations this is computationally very expensive. In

this work the solvent was modeled as a continuous medium surrounding the solute, providing the solvation effects for comparatively little computational effort. An additional benefit provided by this model is the increased sampling due to the removal of steric clash with explicit waters.

The treatment of the solvent as a statistical continuum was achieved using the generalized Born/solvent-accessible surface area (GB/SA)[12] algorithm. Continuum methods have shown some success in ligand–protein docking,[14] although the solvation term is often used in a snapshot fashion, i.e., to rank structures that have been generated in vacuum. In this work, the solvation term is calculated and included "on-the-fly," such that the change in solvation free energy arising from each stochastic move is then used in the Metropolis acceptance test for the Monte Carlo simulation.

The Surface Area (SA) term in the GB/SA model is defined as the area over which the center of a water molecule of radius 1.4 Å can move while maintaining unobstructed contact with the molecule in question, where the solute is modelled as interconnected spheres based on van der Waals radii. SA is calculated using an exact surface area algorithm for interconnected spheres,[23] based on the routines in TINKER[24] (a molecular mechanics package). Each atomic surface area was multiplied by 7.2 kcal mol^{-1} $Å^{-2}$ for the purposes of calculating the associated free energy term.[12]

For the Generalized Born component, the calculation of the Born radii is crucial. For this work, the Pairwise Descreening Approximation (PDA) of Hawkins et al.[25] was adopted. This yields a dependency of the Born radii on the summation of pairwise solvent-accessible terms rather than the solvent-accessible area of the whole molecule. Still and coworkers[26] have recently developed a fast analytical method for the calculation of approximate Born radii. Their algorithm was compared with that of Hawkins et al. on many small organic molecules, with the conclusion that both gave results of similar average unsigned error when compared to experimental data and were of comparable computational speed. The method by Hawkins has been chosen because this implementation was more consistent with the MCPRO data structure.

The GB/SA algorithm was fully integrated into the Monte Carlo program MCPRO version 1.4. Although the algorithm was based on the routines in TINKER, significant modifications were required, particularly for the Born radii calculations, to be used with the MCPRO program, to calculate the solvation free energy.

Parametrization of the Continuum Model

The PDA approach proposed by Hawkins et al.[25] decomposes the eclipsed surface area into pairwise terms, which is a fast and simple analytical calculation but tends to over estimate the eclipsed surface area if two spheres surrounding the central sphere intersect. Following the precedent set by Hawkins et al., this is compensated for by reducing the van der Waals radii with a single scaling factor. A separate scaling factor, based on atom types, is then applied to the Born radii to compensate for the reverse problem where an exposed surface area is unable to contribute to the solvation term due to it being buried in a narrow gap between atoms. This is often referred to as the descreening effect.

Table 2. Calculated Hydration Free Energies (kcal mol^{-1}) for the Authors' Generalized Born Parametrization (A-GB) and Jayaram et al.[27] Parametrization (J-GB).

Molecule	Exp	A-GB	J-GB
Methanol	−5.08	−3.25	−4.79
Ethanol	−4.90	−2.20	−3.46
Ammonia	−4.31	−5.92	−6.14
Methylamine	−4.57	−2.70	−2.57
Ethylamine	−4.50	−1.23	−1.17
Methylthiol	−1.24	−4.85	−2.04
Acetone	−3.85	−5.14	−6.59
2-Butanone	−3.64	−3.82	−5.79
Acetaldehyde	−3.50	−4.28	−5.19
Propionaldehyde	−3.44	−3.20	−3.86
Acetic acid	−6.70	−7.62	−9.32
Propionic acid	−6.47	−6.42	−7.99
Acetamide	−9.72	−9.78	−9.23
Propionamide	−9.42	−8.94	−8.17
Benzene	−0.87	−0.51	−0.84
Toluene	−0.76	+0.29	+0.16
Pyridine	−4.70	−2.01	−2.98
Phenol	−6.62	−4.91	−4.82
N-butyl-Ammonium	−69.24	−69.86	−69.05
Acetate ion	−80.65	−80.97	−80.75
Mean unsigned error		1.2	1.1

Hawkins et al. used a single van der Waals scaling factor, as suggested by Still et al., and optimized the Born radii scaling factors to best reproduce known hydration free energies of more than 100 organic molecules using SM2 atomic radii and AM1 derived partial charges. By optimizing these parameters Hawkins et al. achieved an average unsigned error of 0.38 kcal mol^{-1}. Using this methodology a further parametrization by Jayaram et al.[27] gave the van der Waals and Born radii scaling factors associated with the PDA for the AMBER-AA force field.[16] It should be noted that this parametrization used different van der Waals scaling factors for different atom types, whereas Hawkins et al. used a single scaling factor. A separate parametrization was undertaken here consistent with the AMBER-AA force field using 20 of the molecules from the 32 molecule test set of Jayaram et al., using five Born radii scaling factors and a single van der Waals radii scaling factor for uncharged species. The Jayaram et al. parameterization used a total of 12 parameters for the uncharged species.

The van der Waals scaling factor was 0.88 for all atoms in neutral molecules following the original implementation by Still et al. To generate the scaling factors for the Born radii, first the hydrogen, carbon, and nitrogen parameters were fitted using a simplex algorithm. These values were then fixed and the oxygen and sulphur parameters were determined again using the simplex algorithm. It was found that the charged atoms O2 (sp^2 oxygen in anionic acids) and N3 (sp^3 nitrogen) required different van der Waals scaling factors of 0.80 and 0.89, respectively. These factors were obtained through a systematic search. The 20 molecules used for parametrization are given in Table 2 where the geometries were based on standard AMBER conformations and the charges were derived using the RESP[28] methodology.

Table 3. Simplex Optimized Born Radii Scaling Factors for the PDA.

Atom	Scale factor
H	0.825
C	0.692
N	0.942
O	0.882
S	0.925

The parametrization gave an an averaged unsigned error of 1.2 kcal mol^{-1} compared with Jayaram et al. of 1.1 kcal mol^{-1}. This difference may be attributed to the use of fewer parameters in our implementation. Optimization of the Born radii scaling factors using a simplex algorithm gave the parameters shown in Table 3.

To calculate rapidly the GB/SA term, the algorithm has been structured so that the initial computation includes a complete free energy of hydration calculation for the starting configuration. Subsequent moves in the MCPRO program are based on a random choice of protein residue (or inhibitor) followed by random displacements of atoms for this moving fragment or molecule. Therefore, the solvent-accessible surface area calculation only requires updating for the moving residue and all atoms close enough to be affected by the move. Furthermore, the Born radii calculation can be optimized in a similar way where the summation term only requires updating for the moving residue and the interaction of this residue with all other atoms. This is a similar approach to the GB/SA frozen atom approximation by Guvench et al.[29]

Dihedral Sampling

To increase the sampling of conformational space not only has the GB/SA algorithm been implemented but the increased sampling of dihedrals has also been addressed. The basic Monte Carlo sampling scheme for proteins involves picking a residue at random and changing the side-chain dihedral angles by a random value within a user defined range. To develop a more efficient sampling scheme, instead of simply applying random dihedral moves, the conformation of the side chain can also be chosen at random from a rotamer library of stable conformations[30] identified from protein crystal structures. This modification was coded as an alternative to the standard random dihedral move so that for a user determined number of moves the dihedral angles are assigned based on the rotamer states picked at random from the library. This method is using prior knowledge of low energy rotamer states to assist the sampling such that the acceptance ratio is approximately 40%, even though the moves can potentially be very large. Some flexibility within the rotamer states was allowed, with the dihedral angles being to within $\pm 10°$ of the stable rotamer state.

To complement the rotamer library moves, large dihedral moves in the range $\pm 180°$ were periodically applied to both the protein side chain dihedrals and the ligand dihedrals. This range is considerably more than standard dihedral moves, which are typically in the range ± 5–$15°$. These moves were implemented as a combined strategy with the assignment of random rotamer states to enhance movement between local minimum energy conformations.

Soft Core Interaction Function

Adequate sampling is pivotal to achieving an efficient Monte Carlo dock. Although the GB/SA methodology coupled with novel dihedral moves can improve sampling, large rigid body moves of the ligand within the active site are still very inefficient. It is generally accepted that approximately 40% acceptance of moves yields optimum efficiency. However, for this to be the case the rigid body rotations and translations of the ligand can only take approximate maximum values of 0.1° and 0.03 Å respectively. This is caused by the short range repulsive interactions that tend to infinity at low interatomic separation leading to rough energy surfaces with high energy barriers separating local minima. To overcome this problem, methodology for softening the repulsive intermolecular potential first introduced in free energy calculations[31] was adopted. An alternative approach to achieve extensive sampling of a rugged potential energy hypersurface is simulated annealing. In the early stages of this method, an elevated temperature is used which is slowly reduced, enabling more extensive sampling. However, the selective annealing using a soft-core function focuses specifically on the barriers to adequate sampling, in this case the repulsive nature of the Lennard–Jones potential at low interatomic separation. Thus, the underlying potential is modified in the early stages of the anneal. This is in contrast to the unselective nature of simulated annealing, which affects the total potential energy, enabling sampling of high energy states, but the underlying potential is not altered.

Soft-Core Optimization

Given the original soft-core interaction function[31] of the form

$$V_{LJ}(\mathbf{r}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{(\alpha\sigma_{ij}^{6} + r_{ij}^{6})^2}\right) - \left(\frac{\sigma_{ij}^{6}}{(\alpha\sigma_{ij}^{6} + r_{ij}^{6})}\right)\right] \qquad (3)$$

where α is the annealing parameter, it was envisaged that α would be slowly reduced from 1 to 0 throughout the course of the simulation, returning the full Lennard–Jones potential in the final simulation stages.

The original implementation of the softening function described in eq. (3) proved to be inadequate as short range attractive electrostatic terms tended to dominate in the early stages of the annealing process. From initial vacuum studies of the 1ABE complex it was found that there is an intrinsic balance between the nonbonded terms. Over-softening of a single term can lead to the unsoftened terms dominating the sampling, and in extreme cases molecular locking can occur, where two nonbonded atoms occupy the same space; the molecules effectively fuse together. Unphysical "lock-up" would occur at the initial high values of α where a finite value of the Lennard–Jones function occurs at short interatomic distances, rather than the asymptotic behavior in the unsoftened form. Favorable electrostatic interactions could then dominate, causing the two atoms on different molecules to become

Table 4. Summary of AMBER-AA Parameters Used to Parametrize the Soft-Core Lennard–Jones and Coulombic Energies.

Atom	Residue	Charge	σ/Å	ε/kcal mol^{-1}
HE2	Gln	0.42510	1.0691	0.0157
OG1	Thr	−0.67610	3.0665	0.2104

superimposed. This highlights the importance of providing an effective balance between the softened Lennard–Jones term and other nonbonded interactions.

To ensure a balance between the nonbonded terms a weighting power, m, was introduced to the soft-core function [eq. (4)].

$$V_{LJ}(\mathbf{r}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{(\alpha^m\sigma_{ij}^6 + r_{ij}^6)^2}\right) - \left(\frac{\sigma_{ij}^6}{(\alpha^m\sigma_{ij}^6 + r_{ij}^6)}\right)\right] \quad (4)$$

The parameter m is optimized in conjunction with the softening potential for the remaining nonbonded interactions, to avoid anomalies such as "lock-up."

For the optimization process a typical nonbonded hydrogen atom (HE2 from a Glutamine) and oxygen atom (OG1 from threonine) were considered using AMBER-AA force field parameters given in Table 4.

The Lennard–Jones pair potential is shown in Figure 3(a) with $m = 1$, which illustrates the finite value of the softened Lennard–Jones energy at high α. As α is linearly decreased from 0.9 down to 0.0 in 0.1 decrements the softened potential reduces to the original Lennard–Jones curve.

A similar softening strategy was adopted for the coulombic term where a functional form was chosen to mimic the softened Lennard–Jones term, i.e., a finite value for the potential at low interatomic separation [eq. (5)].

$$V_{\text{Coul}}(\mathbf{r}) = \frac{(1-\alpha)^n q_i q_j}{4\pi\varepsilon_0(\alpha + r_{ij}^2)^{0.5}} \quad (5)$$

The modified shape of the electrostatic function is given by the introduction of the α weight in the denominator and the numerator

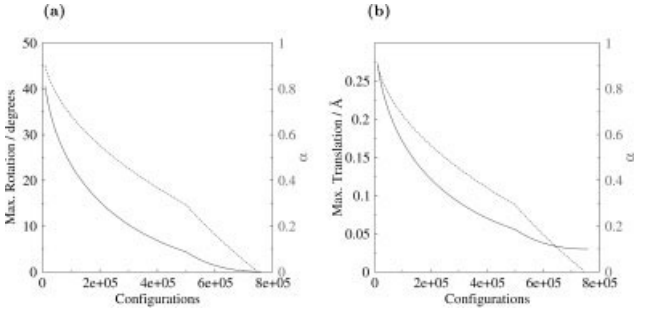

Figure 2. Variation of annealing parameter α (dashed) throughout the simulation with the corresponding maximum rotation ranges (**a**) and maximum translation range (**b**).

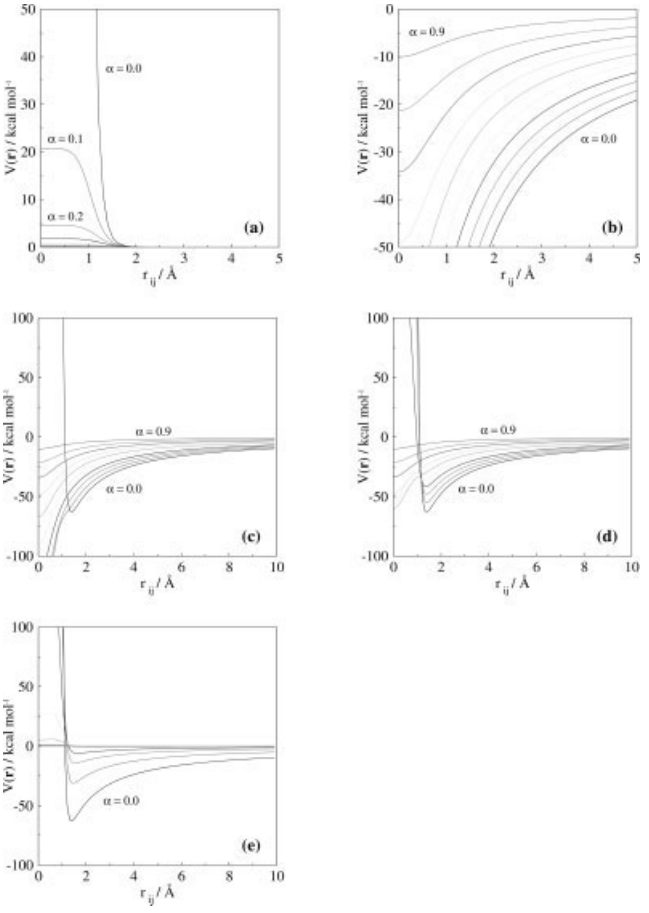

Figure 3. Softened nonbonded energy components with decreasing α in 0.1 decrements, for an oxygen/hydrogen AMBER atom pair. (**a**) LJ function ($m = 1$). (**b**) Coulombic function ($n = 1$). (**c**) Combined LJ and Coulombic ($m = 1$, $n = 1$). (**d**) Combined LJ and Coulombic ($m = 3$, $n = 1$). (**e**) Combined LJ and Coulombic ($m = 3$, $n = 6$).

weight $(1 - \alpha)^n$; n is to be optimized to balance the Lennard–Jones soft-core potential. The coulombic potential with decreasing α is shown in Figure 3b for $n = 1$.

The importance of the weighting powers m and n is shown in Figure 3c for the combined Lennard–Jones and coulombic potential, where it is clear that a 1:1 ratio for m and n results in the electrostatic terms dominating. Systematic increase of the Lennard–Jones weight to $m = 3$ is shown as an intermediate stage in the optimization process in Figure 3d, where the electrostatic terms are still dominating giving a negative energy at low atomic separation. The weighting powers were finalized as $m = 3$ and $n = 6$ (Fig. 3e), which gives a small but finite repulsive term at low atomic separation when α is close to unity corresponding to the early stages of the simulation. This permits a close contact distance between nonbonded atoms and in extreme cases they are effectively able to pass through each other.

Complications associated with combining the full GB/SA with softened nonbonded energies were illustrated by modeling the 1ABE test case. A short simulation protocol was adopted using 10

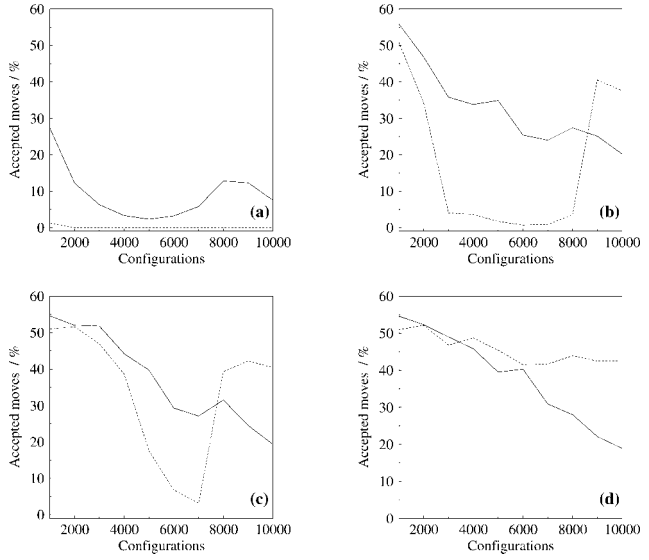

Figure 4. Accepted moves for protein (solid) and ligand (dashed) with (**a**) no GB/SA softening, (**b**) softened with $k = 1$, $l = 1$, (**c**) softened with $k = 1$, $l = 2$, (**d**) softened with $k = 1$, $l = 3$.

batches of 2000 configurations with a 1:1 ratio for the ligand and protein side-chain random moves. The annealing parameter α, was turned off through the course of the simulation, using a linear reduction from $\alpha = 0.9$ (high softening) down to $\alpha = 0.0$ (no softening, hardened potential) in 0.1 decrements, with the previously described Lennard–Jones and coulombic softening function. During this time the complete GB/SA solvation model was calculated for each configuration and added to the vacuum potential energy with softened nonbonded terms.

The percentage of accepted moves for both the ligand and protein is shown in Figure 4a using the full GB/SA solvation potential. The softened Lennard–Jones and coulombic terms are slowly hardened throughout this simulation using the linear annealing protocol. In this instance, the GB/SA solvent model dominates the intermolecular movement resulting in unphysical states, such that all subsequent ligand moves are rejected. This is clearly shown in Figure 4a; the percentage of accepted ligand moves drops to zero almost instantaneously, whereas common acceptance ratios in standard MC methods are often up to 50%.

Independent weighting powers k and l were therefore applied to the surface area and coulombic parts of the GB/SA algorithm using the following equation

$$V_{\mathrm{GBSA}}(\mathbf{r}) = (1 - \alpha)^k \gamma \mathrm{SA} + (1 - \alpha)^l V_{\mathrm{GB}} \qquad (6)$$

where γ is an empirically derived coefficient[12] with a value of 7.2 kcal mol^{-1} Å^{-2}, SA is the surface area, and V_{GB} is the electrostatic contributions to the solvation free energy. Through systematic variation of k and l, for low values of l, intermediate lock-up states were observed as indicated by the temporary drop in the acceptance percentages in Figure 4b and c. These lock-ups were not observed for $l = 3$ (Fig. 4d), which was chosen as the final value of l. Although the acceptance ratio seemed insensitive to the variation of k, the value of $k = 1$ was used, so that the GB/SA energy was not dominated by the surface area terms.

Through observation of the simulation using the softened intermolecular potentials for the 1ABE test case, it was apparent that although molecular lock-up was not seen using the previously described parametrization, during the early stages of the anneal, the remaining unsoftened nonbonded 1,4 interactions were dominating the potential energy. Therefore, to reduce the magnitude of these interactions a $(1 - \alpha)^2$ was applied to the 1,4 interactions to bring the magnitude in line with the other softened nonbonded terms.

Annealing Protocol

It was found that the van der Waals, Lennard–Jones, GB/SA and nonbonded 1–4 interactions all required softening to avoid erroneous structure "lock-up" while providing an appropriate acceptance ratio. Furthermore, each softening function was optimized such that a single energy component did not dominate in a manner inconsistent with the complete force field. The initial tests for this protocol made two assumptions: a constant move size, and linear scaling of α. This methodology uses a maximum rigid body rotation of 0.1° and translation of 0.03 Å for each MC random move, during which α is reduced linearly from 0.9 to 0.0.

By monitoring the RMSD of the ligand between batches of 10,000 configurations throughout the simulations, using a linear decrement of α and a fixed move size, it was seen that most movement occurred in the early stages of the simulation with the largest RMSD between batches of approximately 0.5 Å. It was decided that an increase in the coverage of conformational space was required, which was achieved by modifying the maximum rigid body translation and rotational ranges and spending more simulation time using the softened potential.

A three-stage annealing protocol was developed. This required a new variable, x, which was increased linearly from 0.01 to 1 to yield α defined in eq. (7). The power in eq. (7) was chosen to give a smooth variation of α (see Fig. 2a and b). Alternative scaling methods were investigated, however this protocol produced a smooth potential energy profile avoiding anomalies such as lock-up.

$$\alpha = 1 - x^{1/2} \qquad (7)$$

The three-stage variation of the counter x is given in Table 5. Each batch is performed with a fixed value of x, with x being incremented by δx between batches. It should be noted that the

Table 5. Three Phase Annealing Protocol.

Stage	Configurations	No. batches	δx	Initial x
1	10,000	50	0.01	0.01
2	10,000	25	0.02	0.50
3	5,000	4	0.00	1.00

final stage of the anneal, corresponding to $\alpha = 0$, uses the full potential, i.e., no softening is applied.

Thus, approximately 50% of the simulation time is spent in the first phase for which α varies from 0.9 to 0.3, during which time the most extensive sampling occurs. Having decided upon the variation of α, the size of the attempted ligand move was then linked to α. This was achieved throughout the simulation by gradually reducing the rigid body translation and rotation move sizes, between batches. Obviously, a soft potential increases the probability of a large move size being accepted but, as previously stated, there is a fine balance between large conformational coverage and avoiding anomalies such as lock-up. The maximum translation and rotation ranges are given by eqs. (8) and (9).

$$\text{max.translation} = \texttt{RDELS2}(1+10\alpha^2) \tag{8}$$

$$\text{max.rotation} = \texttt{ADELS2}(1+500\alpha^2) \tag{9}$$

`RDELS2` and `ADELS2` are MCPRO parameters that correspond to the maximum translation and rotation ranges and are assigned the values 0.03 Å and 0.1°. These ranges were chosen based on the smooth variation of the function, shown in Figure 2. In the early stages of the simulation, the scaling functions give a maximum translation and rotation of approximately 0.25 Å and 40°, falling to 0.03 Å and 0.1° at the end. This was shown to give a reasonable acceptance ratio of approximately (40 ± 10)% throughout the simulation of the 1ABE test case; this is discussed further in the Results section.

An example of softening the Lennard–Jones and electrostatic terms to yield increased sampling was implemented by Bouzida et al.[32] in a MC ligand docking study. A similar soft-core function was used for the Lennard–Jones term, combined with a softened coulombic calculation with the AMBER force field, and a volume based description of desolvation. Their motivation for selective softening was the inadequacy of the standard AMBER force field in exploring the binding energy landscapes which are extremely rugged with multiple high energy barriers. They also note the problems associated with simulated annealing whereby docking simulations can become trapped in metastable local minima. A significant difference between the Bouzida et al. MC docking study and that reported here, is that Bouzida et al. used a softened potential with fixed variables for the soft-core function, and the anneal was achieved by a standard simulated annealing of the temperature. As a result, the complex does not experience the full potential, even in the later stages of the simulation. Furthermore, no attempt to quantify or address the problems associated with disturbing the balance between the nonbonded interactions by introducing softening functions was presented. Using this methodology two ligand complexes were tested in a rigid protein and flexible ligand dock, of which only one test case was able to determine the experimental binding mode.

Ligand Preparation

For both the protein and ligand, the AMBER[16] all atom force field was adopted. The ligand cartesian coordinates obtained from the embedding and clustering processes were used to generate a Z-matrix with the program AUTOZMAT supplied with MCPRO 1.6.[15] The original crystal structure conformation is used for the ligand charge derivations. For consistency with the AMBER-AA force field,[16] the RESP methodology[28] was used in conjunction with electrostatic potentials calculated at the 6-31G* level of theory using the GAMESS program.[33]

Structural Considerations

All explicit waters were removed for all stages of the docking protocol. Furthermore, the all-atom force field requires the addition of hydrogen atoms because the pdb structures do not contain this information; this addition is achieved using PEPZ supplied with MCPRO 1.6.[15] The hydrogen positions are based on standard conformations, and therefore atoms, from different residues, could be placed on top of one another. In this case the hydrogens are moved manually to obtain a sensible starting structure. The histidine protonation states were assigned identically to that in the GOLD test set.[5] The GOLD test set assumes the majority of histidine residues to be uncharged in the HID or HIE form, which is reasonable, because most of the experiments for these structures were carried out at pH 6.5. These states were verified visually by assessing which protonation position would enhance the hydrogen bonding the most. In each instance the protonation states, assigned by the authors of GOLD, were reasonable. A further point is that hydrogen bond donating residues such as threonine, serine, and tyrosine could form alternative hydrogen bonding networks if different hydrogen atom conformations are adopted. Again, the GOLD assignments were assessed visually to determine whether appropriate changes would enhance the system's ability to form hydrogen bonds. There was no evidence that changing the conformations assigned in the GOLD test would significantly enhance hydrogen bonding, and hence, the GOLD conformations were adopted.

Combining periodic boundary conditions with an Ewald sum,[34] which models long-range electrostatic interactions, is a common protocol to model systems within a molecular mechanics framework. An alternative method that is computationally less expensive truncates the protein system as a sphere, typically of radius 10–20 Å, centered on the active site where all protein residues outside this sphere are discarded; such a method was used in this work. Within this sphere two regions are defined, a flexible inner reaction zone near the center of the active site and a constrained rigid outer region. Although this spherical scheme has been shown to affect free energies of hydration in simple systems,[35] it has been successfully applied to a wide range of ligand–protein complexes[8,36–38] and hence, is the method of choice for this study.

Based on the embedded and clustered structures, the protein is cropped such that only residues within a certain distance (referred to as a cutoff distance) of the ligand atoms for the seed structures are included. In this case, if any atom in a residue is within the cutoff distance the complete residue is included in the simulation of every seed. This ensures that only important regions, as defined by the clustered seeds of the protein, are included. The process is repeated for calculating the fixed and moving regions, where the cutoff for each region was 12 and 6 Å, respectively, giving an

average of 30 moving residues and 60 fixed residues. This truncation of the protein system can convert nonterminating backbone atoms into terminating atoms; these are retained as uncharged species.

Monte Carlo Parameters

The previous section has demonstrated the benefits of the soft-core function, and has shown how barriers to sampling are reduced by modifying the potential energy functional form, and that a judicious choice of annealing protocol can yield increased sampling. Furthermore, the rigid body rotation and translation ranges for the ligand, that are linked to the annealing parameter α, have been discussed. However, the MC method requires additional protocol parameters to be defined.

As previously stated, the anneal used a three-stage protocol; the first 50 batches of 10,000 configurations use a slow anneal, the next 25 batches of 10,000 configurations use a faster anneal and the final four batches of 5000 configurations use the unsoftened potential energy function. In the MCPRO implementation of the MC method, each random move is either applied to the ligand or a protein residue. For this docking strategy the number of protein moves to inhibitor moves was a 1:1 ratio. A simulation temperature of 37°C was maintained throughout the potential annealing.

If the decision is made to perform a protein residue move, a residue is picked at random. This residue is then moved in one of three ways; either a standard protein residue move, a large dihedral move, or a rotamer library move.[30] A standard protein residue move applies a random change to the dihedrals with a maximum displacement in the range $\pm$2–15° along with random sampling of the bond angles. A large dihedral move performs a random dihedral change for all moving dihedrals in the chosen protein residue or ligand of up to $\pm$180°. The rotamer library move assigns the residue dihedrals at random within 10° of the predetermined rotamer states. Twenty percent of the protein dihedral moves were rotamer library moves, a further 5% were large dihedral moves, and all other moves were standard dihedral moves. It should be noted that although bond angle and dihedral motion for protein side chains were permitted, all rings systems were fixed along with the protein backbone.

A similar procedure to the standard protein residue move was adopted for intramolecular sampling of the ligand. However, the ligand cannot only undergo internal movement, through the sampling of bond angles and dihedrals, but also rigid body rotations and translations. As previously described, the maximum rigid body translation and rotation ranges are linked to the annealing parameter, α, and are subsequently reduced throughout the simulation.

During the early stages of the simulation, when the nonbonded terms are softened, to restrict molecular drift of the ligand from the active site, a half-harmonic restraining potential of 25 kcal mol^{-1} Å^{-2} was applied. This potential was applied to the ligand if the closest atom to the geometric center of the ligand moved further than 5 Å from the active site center, as defined in the GOLD test set.

Two additional points worth noting are the use of a large nonbonded residue-based cutoff, which ensures all residues are included in the energy evaluations throughout the simulation. Also, in the early stages of the simulation, for computational efficiency, the solvent accessible area calculations were updated every 10,000 configurations, although the Born calculation was updated for each configuration because this polarization term was found to be more sensitive to changes in geometry.

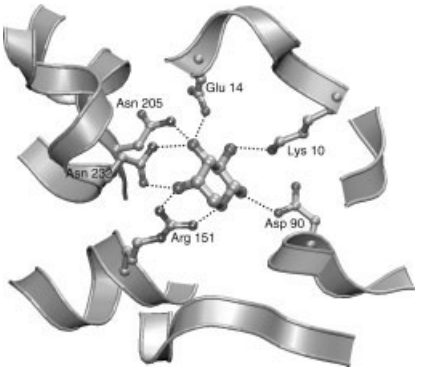

Figure 5. 1ABE complex: crystal structure with hydrogen bonds.

Geometry-Based Docking Results

The geometry-based docking protocol consisting of clique analysis, embedding, and clustering of structures was tested on the 1ABE complex.

A total of 128 cliques were generated, of which 67 were successfully embedded and subsequently clustered using an RMSD cutoff of 3.3 Å to yield five seeds. Approximately 50% of attempted embeds are therefore successful using a maximum of five attempts each. Increasing from 5 to 10 attempts did not drastically increase the success of an embed.

The intermolecular hydrogen bonds present in the crystal structure and the five seeds produced from the geometric docking were analyzed using HBPLUS 3.06,[39] which uses a maximum heavy atom distance between donor and acceptor of 3.9 Å and a minimum angle between donor, acceptor, and acceptor antecedent of 90°. Two seeds (Seed 3 and Seed 4) have found single intermolecular hydrogen bonds that are present in the X-ray structure. A further two seeds (Seed 1 and Seed 5) have each found two hydrogen bonds that are also present in the X-ray structure.

The complete hydrogen bond network for the crystal structure for the 1ABE complex is shown in Figure 5, with all eight hydrogen bonds marked. Figure 6 shows two seeds (Seed 3 and Seed 5) that highlight the diversity of the structures obtained from the geometric dock. Both structures contain hydrogen bonds that are present in the crystal structure, and the RMSD between the structures is 5.5 Å.

Geometric docking for this test complex has therefore satisfied the original requirement in that the required number of structures have been generated, which are not only diverse but also reproduce some (but not all) of the hydrogen bonds found in the crystal structure.

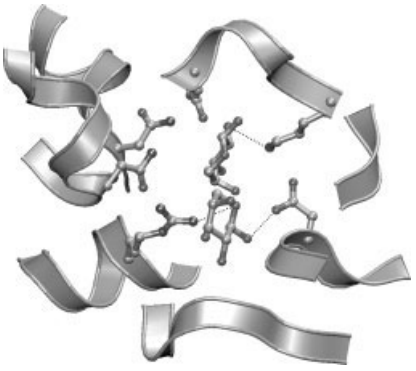

Figure 6. 1ABE complex: two seeds (Seed 3 and Seed 5) showing diversity of structures between different seeds.

Energy-Based Docking Results

An analysis of the final docking phase using the novel MC method requires the following questions to be answered:

1. Has the X-ray structure been generated?
2. If the X-ray structure is found, is it the lowest energy structure, i.e., without prior knowledge would this structure be identified as the best dock?
3. If the X-ray structure is not found, are there parts of the ligand (or protein) that resemble the X-ray conformations, and are key interactions still present?
4. Can failures in generating the X-ray structure be rationalized, i.e., can they be attributed to factors such as poor X-ray structure resolution or the omission of explicit waters?
5. How do the energy hypersurfaces for docking compare, depending on whether protein side-chain flexibility is included.

To help answer these questions the MC docking method was first tested using a rigid protein system and a flexible ligand, which will be referred to as a rigid protein dock (RPD). This is a similar methodology to that employed by the GOLD program. The protocol was then repeated using full side-chain flexibility of the protein, which will be referred to as a flexible protein dock (FPD).

For both the RPD and FPD the seeds derived from the geometric dock were used as starting configurations for the MC annealing protocol. It should be noted that annealing refers to the annealing of the potential energy function; the temperature is kept constant at 37°C throughout the simulation. Furthermore, the X-ray structure for the ligand was used as an additional starting point for both the RPD and FPD. This simulation is defined as Seed 0. The anneal of the crystal structure determines whether the algorithm can return to the experimental binding mode; if the MC sampling is insufficient, this should in theory be one of the most successful docks. Obviously, this structure should not be included in the seed ranking because it uses the X-ray structure as the starting conformation and therefore prior knowledge of the binding geometry.

A standard MC simulation was also tested (referred to as MC optimization), which uses the unsoftened potential energy function at constant temperature (Seed 0a). The standard MC optimization of the crystal structure should produce the lowest potential energy to which the annealed structures may be compared. This optimization also gives an indication of the energy fluctuation associated with sampling a small region of phase space (which may correspond to the global minimum) assuming that the standard MC optimization only samples, at best, a small number of local minima separated by small energy barriers. If the annealed structures are significantly lower in energy than the MC optimization of the crystal structure, this indicates the potential energy function is inadequate.

To determine whether the X-ray structure has been found a commonly used calculation is the RMSD (for nonhydrogen atoms) of the docked structure with the X-ray structure. This calculation takes into account symmetry effects. In this work an acceptable RMSD is considered to be less than approximately 2.0 Å. Structures with a larger RMSD must be treated with caution, and should be inspected visually to assess the degree of success in obtaining the crystallographic binding mode. It is not sufficient to rely solely on the RMSD, because a small translation may give a large RMSD shift, but the primary interactions may still be present. Each docking result has therefore been assessed visually and the RMSD with the crystal structure has been calculated. These results are presented along with the potential energies and intermolecular hydrogen bonds.

The RPD was a success for the 1ABE complex. Two of the seeds (Seed 4 and Seed 5) generated complexes that found the experimental binding motif to within 0.36 Å RMSD of the crystal structure. These two conformations gave the lowest average potential energies over the final full-potential batches, within 2 kcal mol^{-1} of the crystal structure. A summary of the energies and RMSD with the crystal structure is given in Table 6 for the five annealed seeds (Seed 1–Seed 5), the annealed crystal structure (Seed 0), and the MC optimized crystal structure with no softening (Seed 0a). Twice the standard error is given in parentheses; this is derived by dividing the final stages of the simulation, using the full unsoftened potential, into two equal batches. An increase in batch size for the final stages of the simulation was not deemed to be necessary because no significant improvement in the RMSD value was obtained in the final stages of the docking protocol.

Thus, Seed 4, Seed 5, and Seed 0a are considered to be in a low energy cluster of conformations, all within an RMSD of 0.36 Å of the crystal structure. A further point is that Seed 0a is a standard MC simulation of the crystal structure with the full potential energy function throughout. It is therefore unlikely that the annealed seeds will be as low in energy, because Seed 0a is effectively just optimization of a structure that should already be close to the global minimum.

Seed 0 (the annealed crystal structure) is also ranked sixth out of the RPD simulations even though the starting structure was the closest to the global minimum. This suggests that starting from the crystal structure will not necessarily assist the docking because the sampling is extensive. It is therefore important for the first docking stage to produce diverse starting structures that sample potentially important regions of the active site, because not all of the seeds were able to reproduce the experimental binding motif.

The potential energy for all seeds in the last stages of the RPD simulation is shown in Figure 7a. It should be noted that for all simulation graphs the solid black trace monitors the crystal struc-

Table 6. Summary of 1ABE Rigid Protein and Flexible Ligand Dock, for Five Diverse Starting Structures (Seed 1–Seed 5), Annealed Crystal Structure (Seed 0) and MC Optimized Crystal Structure (Seed 0a).

Seed	RMSD/Å	Potential energy/ kcal mol^{-1}	GB/SA energy/ kcal mol^{-1}	Coulombic and Lennard–Jones/ kcal mol^{-1}	Number of hydrogen bonds
0a	0.29 (0.00)	−1498 (1)	−1265 (0)	−106 (1)	8 (1)
4	0.36 (0.11)	−1496 (0)	−1263 (0)	−106 (0)	6 (0)
5	0.24 (0.02)	−1496 (1)	−1267 (2)	−103 (3)	7 (1)
1	1.64 (0.11)	−1488 (0)	−1287 (2)	−80 (2)	6 (1)
2	2.63 (0.05)	−1487 (0)	−1263 (0)	−108 (0)	8 (0)
0	2.89 (0.03)	−1479 (1)	−1288 (0)	−58 (1)	7 (0)
3	2.38 (0.07)	−1474 (1)	−1272 (0)	−74 (0)	5 (2)

ture docked with the annealing protocol (Seed 0), while the dashed black trace shows the standard MC optimization of the crystal structure (Seed 0a). The colored traces correspond to the different seeds, and therefore, diverse starting points for the simulation. The corresponding RMSD with the crystal structure, throughout the entire simulation, is shown in Figure 7b.

It is clear from Figure 7b that the largest fluctuations in the RMSD occurs in the initial 250,000 configurations. This corresponds to the stage in the simulation with the softest potential energy, and α (the annealing parameter) is between 0.9 and 0.5. It should be noted that the large fluctuation of the RMSD (up to 4 Å) is an indication of the extensive sampling achieved using the annealing protocol. After approximately 300,000 configurations, the yellow seed and brown seed (Seed 4 and Seed 5, respectively) find the crystallographic binding mode and do not undergo any significant geometry changes after this point. The red seed (Seed 1) seems to fluctuate between two binding geometries, one of which is within an RMSD of 1.64 Å with the crystal structure. The remaining seeds do not find the crystallographic binding mode.

A second set of docking simulations were then performed that included side-chain flexibility. The potential energies and RMSD with the crystal structure from the FPD are given in Figure 7c and d, and are summarized in Table 7.

Using the FPD, the blue seed (Seed 3) has found the X-ray structure to within 1.24 Å RMSD. However, this is not the lowest energy structure, the green seed (Seed 2) is at least 10 kcal mol^{-1} lower in energy. Furthermore, the yellow seed (Seed 4) is indistinguishable from Seed 3 based on potential energy. Seed 3 would therefore not be chosen as the most successful dock using only the potential energy score, even though the RMSD with the crystal structure is the lowest.

Although Seed 0 for the FPD has achieved a low RMSD (see Table 7) and is one of the lowest energy structures, obviously this cannot be considered a success. This structure used the crystal conformation as the starting structure, and is therefore using prior knowledge to dock the ligand into the active site. As previously stated, Seed 0 is only a guide to monitor the behavior of the annealing protocol using a starting structure, which is very close to the global minimum. If Seed 0 is not always the most successful dock, this indicates that the method is able to sample extensively regions of configuration space over a short time frame. This issue of adequate sampling is a limitation of the standard MC method.

These results can be analyzed by considering the effect of incorporating flexibility into the protein side chains in the FPD. From Figure 7c it is clear a much broader overlap of energies are produced, which tend to fluctuate more than the RPD energies. This is not surprising, because an extra degree of complexity has been introduced into the problem. By comparing the standard errors in the potential energies for the MC optimized crystal structure using the FPD and RPD, the FPD has a larger average error of at least ± 5 kcal mol^{-1} compared with a negligible error for the RPD. This value is not only shown in the standard error for the MC optimized crystal structure but it is also verified by visual analysis of Figure 7c. If this is an indication of the energy fluctuations that can be produced for what is likely to be the global minimum, then seeds within this error can be considered isoenergetic based on potential energy alone. Remembering that the standard error is only a guide to the energy fluctuations observed in the final stages of the simulation, using the full potential energy function, Seed 2 has therefore found a binding mode that is energetically indistinguishable from the crystal structure.

A further point concerns the RMSD of the seeds, with the crystal structure, using the FPD (Fig. 7d). It is clear that even greater fluctuations in the RMSD are observed in the FPD compared with the RPD (Fig. 7b). Furthermore, a stable binding mode is found after approximately 500,000 configurations, which is approximately twice the simulation time required for a stable binding mode to be found in the RPD. This is not surprising because the dimensionality of the problem is increased using the FPD, compared with the RPD method.

One possible reason for the production of a low RMSD structure (Seed 3) with an energy that is greater than another binding mode (Seed 2) that has a higher RMSD (see Table 7) is that the simulation has not been run for sufficient time. Hence, the FPD protocol was applied to the system twice more, using the output structures from the previous anneals as the starting points for the new anneals. The third anneal is probably the most interesting, as the green seed (Seed 2) achieves the lowest RMSD with the crystal structure (0.78 Å), the results for which are summarized in Table 8.

Thus, by applying the annealing protocol three times, the X-ray structure is found by one of the annealed seeds (Seeds 1–5), to within an acceptable RMSD. The next question is, can this structure be identified as the lowest energy structure from the other

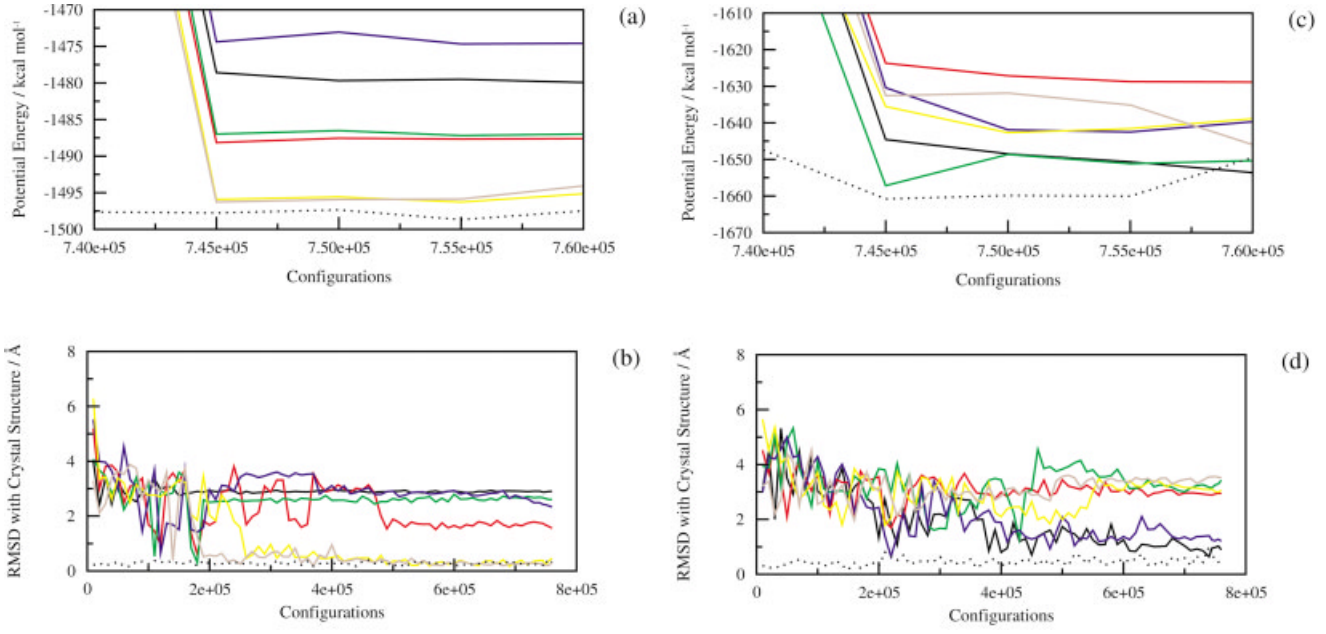

Figure 7. 1ABE complex: Rigid protein dock, (**a**) energy and (**b**) RMSD. Flexible protein dock, (**c**) energy and (**d**) RMSD.

seeds? Unfortunately, the answer to this question is no; there are two other binding modes found by Seed 3 (blue Seed) and Seed 1 (red Seed) which are iso-energetic within the standard error. Although the RMSD of Seed 1 and Seed 3 are similar these are two different binding modes. Seed 1, Seed 2, and Seed 3 are therefore in a cluster of low-energy binding modes that cannot be distinguished using the potential energy, and Seed 2 has an RMSD of 0.78 Å with the crystal structure.

To analyze the results from the third application of the annealing protocol, the final conformation of the seed with a low RMSD (Seed 2) is shown in Figure 8, along with the residues that have the largest side-chain movements. The corresponding crystal structure is shown in the transparent ball-and-stick representation. It is clear that the single largest side-chain movement is for Asp 90, where the χ_1 dihedral angle (N—CA—CB—CG) moves from $-64.6°$ in the crystal structure, to $-160°$ after the annealing protocol; these conformations are very similar to two values observed in the rotamer libraries. As a result of this movement, only 6 ± 1 intermolecular hydrogen bonds are present in the annealed structure rather than the full complement of eight hydrogen bonds present in the original crystal structure. However, the binding geometry is still consistent with the crystal structure. A rationalization for this side-chain movement is difficult to determine, owing to the complex coupled nature of the molecular mechanics force field. However, the movement of Asp 90 to this position does not produce any new hydrogen bonds; in fact, a hydrogen bond with Lys 10 is no longer present. The movement could be attributed to the residue being in a position on the extremity of the protein, and by adopting this conformation the polar residue becomes more solvent exposed.

Figure 9 shows the blue seed (Seed 3), which is indistinguishable from Seed 2 based on the potential energy. Seed 3 also shows a similar movement of Asp 90, due to the reasons previously described for Seed 2. Interestingly, the primary difference between the protein structures for Seed 2 and Seed 3 is the movement of the Asp 89, which forms a bridging hydrogen bond between the ligand, protein, and an explicit water molecule in the crystal structure. The χ_1 dihedral angle (N—CA—CB—CG) is approximately $-170°$ in the crystal structure (and Seed 2) but is $-56.9°$ in the binding mode of Seed 3. An obvious reason for this movement in Seed 3 is the inability of an implicit solvent model to reproduce explicit and very specific hydrogen bonded water interactions. In fact, owing to the movement of the ligand in Seed 3, the conformation permits a hydrogen bond between the Asp 89 and O4

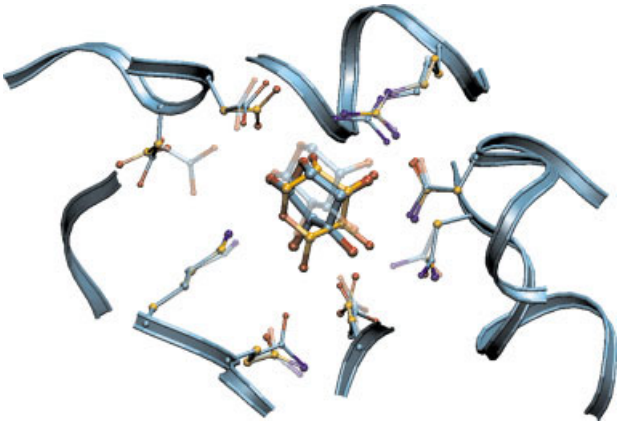

Figure 9. Most successful flexible dock (third batch) with crystal structure (transparent) and iso-energetic binding mode.

Table 7. Summary of 1ABE Flexible Protein and Flexible Ligand Dock, for Five Diverse Starting Structures (Seed 1–Seed 5), Annealed Crystal Structure (Seed 0) and MC Optimized Crystal Structure (Seed 0a).

Seed	RMSD/Å	Potential energy/ kcal mol^{-1}	GB/SA energy/ kcal mol^{-1}	Coulombic and Lennard–Jones/ kcal mol^{-1}	Number of hydrogen bonds
0a	0.53 (0.18)	−1658 (6)	−1298 (2)	−91 (1)	6 (1)
2	3.28 (0.15)	−1652 (2)	−1272 (15)	−100 (2)	6 (0)
0	0.96 (0.01)	−1649 (6)	−1254 (15)	−98 (0)	6 (0)
4	3.04 (0.03)	−1640 (1)	−1310 (24)	−75 (6)	4 (2)
3	1.24 (0.01)	−1639 (5)	−1247 (17)	−107 (1)	6 (0)
5	3.54 (0.03)	−1636 (8)	−1252 (3)	−81 (2)	6 (0)
1	2.93 (0.04)	−1627 (3)	−1261 (5)	−93 (1)	7 (0)

of the ligand. Only minor changes in the other protein side chains are observed.

A further encouraging result is the lack of false positives, i.e., no structure is able to produce a binding mode that is lower in energy than the MC optimized crystal structure, within the computational error. Obviously, this is an important consideration.

To examine whether the experimental temperature factors (or B factors) in the original pdb file give an indication of the side-chain movement observed in the FPD anneal, the B factors for Asp 89 and Asp 90 were compared with the remaining residues. However, the B factors for these residues were not noticeably higher than those for the other residues.

An additional consideration is that if iso-energetic binding modes are found that are indistinguishable based on potential energy alone, the result could be an artefact of truncating the protein system. This question was examined by performing simulations on the full 1ABE complex using a 40 Å cutoff with the same moving residues. In this instance two iso-energetic binding modes were found that were within 4 kcal mol^{-1} of the crystal structure but the RMSD with the crystal structure was greater than 4 Å. This indicates that the observation of multiple binding modes that are iso-energetic is not an artefact of truncating the potential. These annealing simulations required approximately 400 h to run, which is an order of magnitude more than that for the truncated system. The average simulation time for the truncated protein system is between 30–40 h on an AMD 750 MHz Athlon(tm) processor.

Although it is important to reproduce the experimental binding modes, it is also important to assess the success of the searching function in achieving a large coverage of conformational space. This is analyzed by monitoring the difference in the ligand RMSD between the final structures of the current and previous batches of 10,000 configurations, to determine the degree of movement between each batch for both the rigid and flexible dock (Fig. 10a and b). The RMSD with respect to the last configuration of the previous batch can be as much as 5 Å, which highlights the extensive sampling.

The limitations of a standard MC optimization procedure without an annealing protocol was investigated by repeating the flexible docking simulations with the five diverse seeds and crystal structure, but with the full potential energy function at all times. The RMSD with the previous batch and the RMSD with the crystal structure are shown in Figure 10c and d. It is clear that the sampling for a standard MC optimization is not as extensive as the annealing protocol, because the maximum RMSD with the previous batch is 0.5 Å, compared with up to 5 Å for the annealing protocol. Analysis of the RMSD with the crystal structure supports this observation, as only the red seed (Seed 1) seems to sample more than one binding mode throughout the simulation.

To further validate the protocol the average acceptance ratio for RPD and FPD was recorded for each batch (Fig. 11a and b). An average acceptance ratio of (40 ± 10)% was achieved. As previously stated, this is a reasonably efficient ratio, which is particularly important, because the rigid body translations and rotations of

Table 8. Summary of 1ABE Flexible Protein and Flexible Ligand Dock (Third Batch), for Five Diverse Starting Structures (Seed 1–Seed 5), Annealed Crystal Structure (Seed 0), and MC Optimized Crystal Structure (Seed 0a).

Seed	RMSD/Å	Potential energy/ kcal mol^{-1}	GB/SA energy/ kcal mol^{-1}	Coulombic and Lennard–Jones/ kcal mol^{-1}	Number of hydrogen bonds
0a	0.53 (0.18)	−1658 (6)	−1298 (2)	−91 (1)	6 (1)
3	2.32 (0.01)	−1648 (11)	−1264 (22)	−117 (4)	7 (0)
1	2.61 (0.06)	−1647 (2)	−1256 (9)	−117 (3)	6 (2)
2	0.78 (0.07)	−1645 (10)	−1290 (22)	−89 (4)	6 (1)
0	3.16 (0.03)	−1635 (1)	−1288 (8)	−70 (2)	5 (2)
4	3.81 (0.00)	−1629 (3)	−1316 (16)	−78 (1)	4 (1)
5	1.91 (0.01)	−1619 (10)	−1249 (22)	−86 (5)	5 (2)

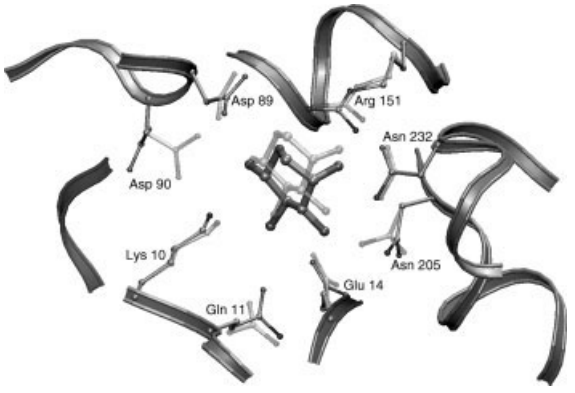

Figure 8. Most successful flexible protein dock (third batch) with crystal structure (transparent).

the ligand in the early stages of the simulation are at least an order of magnitude larger than those used to achieve a similar acceptance ratio in a standard MC simulation.

Summary of the Method

A description of the implementation and execution procedure to operate the suite of docking algorithms has been given, along with the optimization protocol tested on the 1ABE complex. The algorithm involves the generation of diverse docked geometries by satisfying intermolecular hydrogen bonding through distance geometry. These structures are then used in a more time-consuming MC simulation based approach, in which solvation and large-scale side-chain motion of the receptor are allowed. Furthermore, to facilitate sampling of the ligand in the binding site, the potential energy function is carefully softened. The importance of the annealing protocol and the optimization of the soft-core function to traverse efficiently the rugged potential energy hypersurface has been demonstrated. An annealing protocol has been established that yields a softened potential energy that is balanced in a manner consistent with the force field. The potential energy can be turned on in a smooth and controlled fashion, and the specific barriers to sampling have been reduced. Consequently, extensive sampling is observed throughout the annealing procedure.

The geometric dock based on hydrogen bonds was able to generate a diverse set of starting structures that included some hydrogen bonds observed in the X-ray structure.

The RPD was a success where the two lowest energy structures corresponded to the crystallographic binding mode. An RMSD of 0.36 Å was achieved. After three applications of the FPD protocol, three iso-energetic binding modes, one of which was within 0.78 Å of the crystallographic geometry, were generated.

The protein side chains were extensively sampled in the FPD, and for all but two of the residues the approximate crystallographic conformations were observed. These movements were attributed to an increase in solvent accessibility and the omission of an explicit water molecule.

Summary of Results for 14 Complexes

To validate further the method and to compare the energy hypersurface for the rigid and flexible receptor approximations, the algorithm has been applied to a further 14 complexes chosen from the GOLD data set.[5] These results will be summarized here and described in detail elsewhere.[18]

A summary of the RPD results for the 14 complexes is given in Table 9. The crystallographic binding mode was observed in 12 out of the 14 docked complexes, with a RMSD of 1.46 Å or less, using a rigid receptor and flexible ligand. Of the 12 binding modes that were similar to the crystallographic binding mode, 8 were the single, lowest potential energy modes. It should be noted that although the 1DBB complex produced a structure with the lowest RMSD that was not ranked the most favorable in energy, the lowest energy structure had a RMSD with the crystal structure of 2.01 Å. This is therefore considered a successful dock. Of the remaining four complexes, two found an alternative binding mode, which was energetically indistinguishable from the structure that was closest to the crystallographic binding mode. A further two complexes (1STP and 2ACK) produced alternative binding modes that were lower in energy than the seed with the lowest RMSD.

As a further point, none of the docked seeds produced a potential energy lower than the energy obtained though standard MC simulation of the crystal structure (within the error limit). This is a strength of the potential energy function with a continuum solvent model; no false positives were produced.

Only 2 of the 14 complexes were unable to reproduce the crystallographic binding mode. The failure to dock successfully the 1FKG complex is attributed to the large number of flexible dihedrals. However, it was unclear why 1MCR was a failure, although a standard MC simulation of the crystal structure showed significant movement of the ligand. This could imply that the crystal structure binding mode is unstable in our model. Because no explicit waters were included in the pdb, it was impossible to identify whether specific interactions between the protein, ligand, and explicit waters stabilized the binding.

Having applied the RPD protocol to the data set, the protocol was then tested using a flexible receptor. This is considered to be the first stage in validating a completely flexible docking algorithm. Further extensions to this work should apply the method using the apo-protein structure, to evaluate the procedure in determining the bound conformation of the protein, given only the free protein structure.

A summary of the FPD results are given in Table 10. When side-chain flexibility of the protein was included in the docking protocol, 10 out of the 14 complexes produced the crystallographic binding mode. Eight of the 10 successful complexes contain the crystallographic binding mode in clusters of low-energy structures that are indistinguishable based on potential energy. Of these eight complexes, three were the single lowest energy conformations and the remaining five contained at least one other binding mode of similar energy. In the case of complex 6ABP, visual inspection of the potential energy graphs showed fluctuations of at least ± 5 kcal mol^{-1}, a value that is underestimated by the standard errors reported in Table 10. For this reason, the lowest energy and lowest RMSD structures are considered to be energetically indistinguishable. Through inspection of the X-ray structures for the five

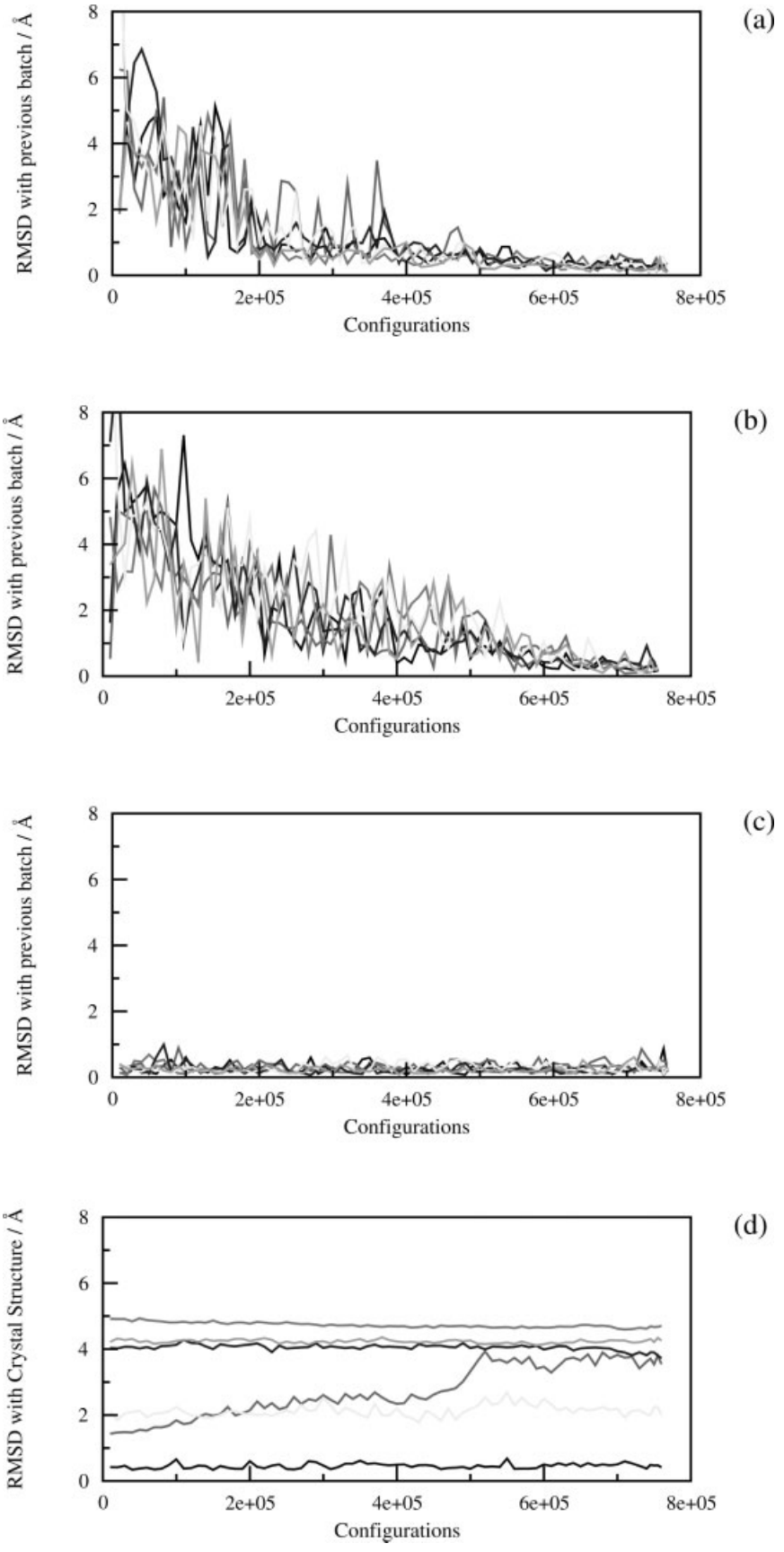

Figure 10. 1ABE complex: (**a**) Rigid protein dock, RMSD with previous batch. (**b**) Flexible protein dock, RMSD with previous batch. (**c**) Flexible protein dock (no annealing) RMSD with previous batch. (**d**) Flexible protein dock (no annealing) RMSD with crystal structure.

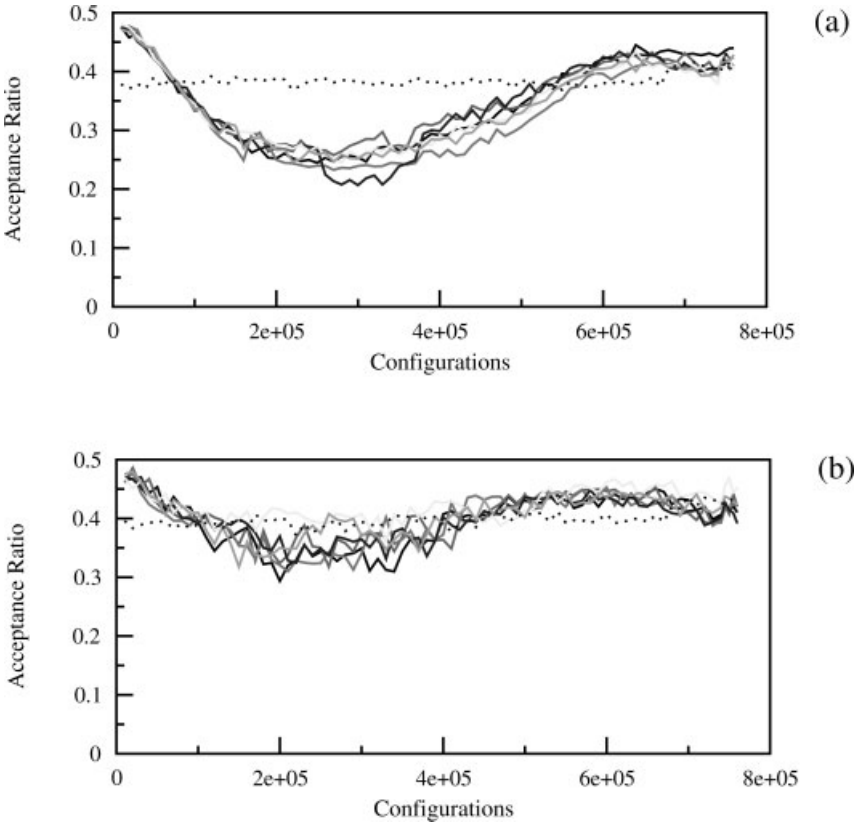

Figure 11. 1ABE complex: (**a**) Flexible protein dock acceptance ratio. (**b**) Rigid protein dock acceptance ratio.

complexes containing multiple binding modes, four contained explicit waters in the binding site. These explicit waters were not included in the docking protocol, and this could explain the observed multiple binding modes. The inclusion of several explicit water molecules may therefore enable the crystallographic binding mode to be isolated from other low-energy binding modes. A combined explicit/implicit solvent model has already been developed by the author using the OPLS-UA force field, and has been applied to free energy calculations for the Neuraminidase system by Wall et al.[40] By including these explicit waters the stability of the binding mode was enhanced for some of the Neuraminidase complexes.

As expected, fewer complexes were able to reproduce the experimental binding mode when protein side-chain flexibility was incorporated into the docking method, because a further degree of complexity has been introduced to the method. By increasing the dimensionality of the problem, the average error in the potential energy increased from ± 1 kcal mol^{-1} in the RPD to ± 5 kcal mol^{-1} in the FPD. The distinction between different binding modes is therefore more difficult. Furthermore, by permitting side-chain flexibility alternative binding modes were found. In the case of the 2CGR complex, the movement of a tryptophan resulted in a new binding mode that included a new hydrophobic interaction between the ligand and protein. This binding mode was not observed in the RPD.

The three failures in the FPD, that were successes in the RPD were 2CHT, 1ACJ, and 4CTS. The 2CHT binding site contains a large number of highly flexible residues such as arginine. From inspection of the side-chain conformations it is clear that many of the arginines adopt significantly different conformations compared with the crystal structure, thus making alternative binding modes for the ligand accessible. Similarly, in 4CTS flexible arginine and asparagine residues adopt different conformations. This combination of a complex hydrogen bonding network coupled with the flexible nature of the protein residues and the small ligand size with respect to the active site, results in a difficult test for the docking method. In contrast, 1ACJ is predominantly a hydrophobic binding pocket, and the movement of a single tyrosine residue enables alternative binding modes to be accessed.

Conclusions

In this article, a docking procedure that explicitly incorporates ligand flexibility, protein side-chain flexibility, and solvation has been described. The methods have been optimized on the 1ABE complex, and then tested on a further 14 structures. On adopting a rigid receptor approximation, the binding mode observed in the X-ray crystal structure was found for 87% of the systems examined. In 73% of the systems, the X-ray binding mode was in the

Table 9. Summary of Docking Results with a Flexible Ligand and Rigid Receptor.

	Lowest RMSD			Lowest energy		Optimized X-ray		
PDB	RMSD	E	Rank	RMSD	E	RMSD	E	Anneal no.
1DBB	0.63	−404 (2)	5/5	2.01	−438 (1)	1.00	−445 (0)	1
1STP[a]	1.25	−1276 (0)	=2/5	2.37	−1318 (2)	0.46	−1339 (0)	2
2SIM[a]	1.44	−1412 (1)	=1/5	4.14	−1413 (0)	0.65	−1426 (1)	2
1FKG	2.73	−1263 (2)	4/5	6.13	−1306 (0)	0.57	−1336 (0)	1
1ACJ[a]	0.46	−2002 (0)	1/5	0.46	−2002 (0)	0.45	−2002 (0)	1
2ACK[a]	0.49	−2138 (0)	2/5	3.64	−2146 (0)	0.49	−2182 (0)	3
1MCR	2.62	−145 (3)	=1/5	2.62	−145 (3)	2.87	−150 (0)	1
4CTS	0.39	−843 (0)	1/5	0.39	−843 (0)	0.45	−843 (0)	1
6ABP[a]	1.46	−1533 (4)	=1/4	1.46	−1533 (4)	0.29	−1544 (0)	3
1DBJ	0.61	−1039 (1)	1/5	0.61	−1039 (1)	0.53	−1041 (1)	1
1LST	0.45	−2667 (1)	1/4	0.45	−2667 (1)	0.34	−2676 (1)	1
1MRK[a]	1.25	−2323 (0)	1/3	1.25	−2323 (0)	1.09	−2328 (0)	2
2CGR[a]	0.79	−1602 (1)	1/4	0.79	−1602 (1)	0.94	−1601 (3)	1
2CHT	0.82	−834 (0)	1/5	0.82	−834 (0)	1.07	−840 (1)	3

RMSDs are given in Å, energies in kcal mol^{-1}, and the Anneal no. corresponds to the batch of the annealing protocol that yielded the structure with the lowest RMSD. The figures in parentheses correspond to twice the standard errors on these energies.

[a]Indicates explicit waters were present in the X-ray structure of the binding site; = indicates alternative binding modes of similar energy were found.

group of lowest energy structures, and in 60% of the cases it was the unique, lowest energy structure. This docking performance is comparable to that reported for existing algorithms. On adopting a flexible protein receptor model, the binding mode observed in the X-ray crystal structure was found for 73% of the systems, was part of the group of lowest energy structures in 60% of the cases, but was the unique, lowest energy structure in only 20% of the systems.

The results presented suggest the presence of a funnelled energy hypersurface using the rigid receptor approximation. However, by including protein flexibility the energy surface is seen to

Table 10. Summary of Results for Flexible Ligand and Flexible Receptor.

	Lowest RMSD			Lowest energy		Optimized X-ray		
PDB	RMSD	E	Rank	RMSD	E	RMSD	E	Anneal no.
1DBB	0.78	−1274 (4)	=1/5	0.78	−1274 (4)	0.92	−1284 (7)	3
1STP[a]	1.28	−1525 (0)	4/5	2.32	−1611 (6)	0.79	−1632 (4)	2
2SIM[a]	1.34	−2281 (5)	=1/5	6.04	−2287 (7)	0.80	−2283 (7)	3
1FKG	4.57	−1461 (2)	1/5	4.57	−1461 (2)	1.10	−1495 (7)	1
1ACJ[a]	3.01	−2161 (2)	5/5	4.87	−2175 (1)	1.43	−2172 (5)	2
2ACK[a]	1.73	−2180 (1)	1/5	1.73	−2180 (1)	0.47	−2218 (8)	2
1MCR	3.59	−1813 (2)	2/5	4.50	−1819 (1)	4.18	−1822 (9)	2
4CTS	3.57	−2186 (4)	3/5	3.64	−2206 (0)	1.32	−2229 (4)	3
6ABP[a]	0.93	−1836 (3)	=1/4	3.70	−1847 (3)	0.90	−1842 (4)	3
1DBJ	0.92	−1626 (3)	1/5	0.92	−1626 (3)	0.66	−1637 (7)	2
1LST	0.79	−2783 (6)	1/4	0.79	−2783 (6)	0.25	−2821 (1)	2
1MRK[a]	1.30	−2532 (11)	=1/3	1.30	−2532 (11)	1.67	−2580 (1)	2
2CGR[a]	0.67	−1830 (7)	=1/4	5.02	−1838 (2)	0.85	−1862 (9)	2
2CHT	1.83	−1953 (11)	3/5	3.82	−1997 (4)	1.18	−1987 (6)	3

RMSDs are given in Å, energies in kcal mol^{-1}, and the Anneal no. corresponds to the batch of the annealing protocol that yielded the structure with the lowest RMSD. The figures in parentheses correspond to twice the standard errors on these energies.

[a]Indicates explicit waters were present in the X-ray structure of the binding site; = indicates alternative binding modes of similar energy were found.

become more rugged with multiple minima that are indistinguishable using the potential energy function described.

The notion of an ensemble of docked substates that are indistinguishable, based on a molecular mechanics model of the potential energy, is a feature that is observed in both the RPD and FPD. It is unclear whether multiple binding modes of similar energies is a real effect or an artefact of the potential energy model. However, there is significant literature evidence to support the view that this may be a true reflection of reality. A similar observation was noted by Verkhivker et al.,[41] where the importance of conformational substates of ligands interacting with a rigid receptor were highlighted. This problem is accentuated when flexibility is included into the protein model, which can lead to not only a rugged potential energy hypersurface but also a diverse cluster of binding modes of similar potential energies. Other docking programs have also produced multiple binding modes that are similar in energy. The authors of the docking program Darwin,[42] which uses a GA search strategy with the CHARMM force field, also note the dynamic nature of molecules, and that multiple docking modes may be a reasonable reflection of reality. Wade and coworkers[43] observed three binding modes of 1S-camphor using multiple copy molecular dynamics to cytochrome P450cam. However, only one of these modes was observed in the crystal structure. The dynamic nature of proteins is also discussed in depth by Rejto and Freer.[44] They suggest that protein energy landscapes are frustrated and are characterized by multiple minima separated by large energy barriers, where many near degenerate binding modes are accessible. If the notion of multiple binding modes with similar energies is a true reflection of reality then as a consequence more than one structure should be considered during ligand optimization. Hilpert et al.[45] have shown the importance of alternative binding modes in the design of potent and highly selective thrombin inhibitors. During a search for novel inhibitors, a new and unexpected binding mode to thrombin was discovered. Subsequent modifications of these new inhibitors reproduced the expected binding mode. By using the method presented in this article, it is possible that the two different binding modes would be observed in a docking calculation.

Multiple binding modes are not usually seen with the rigid receptor approximation used in the majority of current docking algorithms. Alternative binding modes were mainly seen when extensive movement of the protein side chain was permitted, which was achieved using rotamer libraries and the soft-core function. By using standard minimization techniques such as conjugate gradient minimization or standard MC simulations, the conformations of the side chains and the ligand will not be sampled extensively. Although several docking methods have addressed the problem of obtaining adequate sampling,[14,32] the degree of coverage of search space that is achieved using these methods is unclear.

It has been shown that extensive sampling has been achieved in both the RPD and FPD. This is not only indicated by the large difference in RMSD between anneal batches (up to 5 Å) but also by the fact that the anneal of the crystal structure was not always the most successful dock. This has implications for the starting structures generated by the geometric dock. It is not essential to use the crystal structure as the starting conformation, although it seems the diversity of the structures is important; rarely will more than one seed achieve a successful dock. An alternative approach to generating starting structures for the MC docking protocol may be random conformations. However, the geometric docking method produces diverse starting structures that sample potentially important regions of the active site.

Including side-chain flexibility into the simulation using the bound conformation of the protein is only the first stage in the difficult problem of docking a flexible ligand into a flexible binding site. The next stage to be addressed in future work, is the use of the apo-protein structure and the bound form for an alternative ligand, as starting structures for the FPD. Prior to attempting this, however, it was necessary to demonstrate that many of the crystallographic conformations could be reproduced after extensive side-chain sampling, starting from the bound protein conformation.

Many complexes also have their binding mediated by specific bridging waters that can be crucial in determining the most likely binding motif. It is notable that multiple iso-energetic binding modes were generally identified in systems where explicit water molecules had been removed from the structure for the docking procedure. Although this methodology is not presented here, a combined explicit/implicit solvent model was successfully developed with Wall[40] and applied to the study of Neuraminidase. Thus, the inclusion of a few explicit water molecules, in conjunction with a continuum solvation model, will also be investigated.

A significant disadvantage of this docking methodology is the cost taken to perform a single anneal (approximately 30 h on a 750 MHz AMD Athlon processor). For the rigid receptor approximation, a grid representation for the stationary parts of the system will significantly reduce the computational time by an order of magnitude. However, because a primary aim of this work was to address the issue of protein conformational flexibility, the benefits from a grid representation may be limited. Furthermore, the use of an accurate solvation model is a very expensive procedure, and although considered important for this study it is often omitted in other docking methods. An additional consideration is the use of explicit electrostatics; in most current docking methods the electrostatics are included using simple hydrogen bond approximations to model short range attractive electrostatics. However, to model protein flexibility the omission of important terms such as electrostatics may give erroneous results. It should be remembered that none of the docked structures gave a potential energy lower than that obtained through standard Monte Carlo simulation of the crystal structure (within the error limits), thereby validating the potential energy function used.

Finally, the docked structures generated here have been scored using a potential energy function. It is possible that the isoenergetic binding modes may be distinguishable if their relative free energies of binding could be determined. This work is currently in progress.

Acknowledgments

We should thank Prof W. L. Jorgensen for making the MCPRO software available to us. J.W.E. is a Royal Society University Research Fellow.

References

1. Blaney, J. M.; Dixon, J. S. Perspect Drug Discov 1993, 1, 301.
2. Walters, W. P.; Stahl, M. T.; Murcko, M. A. Drug Discov Today 1998, 3, 160.
3. Bissantz, C.; Folkers, G.; Rognan, D. J Med Chem 2000, 43, 4759.
4. Stahl, M.; Rarey, M. J Med Chem 2001, 44, 1035.
5. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. J Mol Biol 1997, 267, 727.
6. Baxter, C. A.; Murray, C. W.; Waszkowycz, B.; Li, J.; Sykes, R. A.; Bone, R. G. A.; Perkins, T. D. J.; Wylie, W. J Chem Inform Comput Sci 2000, 40, 254.
7. Miller, M.; Schneider, J.; Sathyanarayana, B. K.; Toth, M. V.; Marshall, G. R.; Clawson, L.; Selk, L.; Kent, S. B. H.; Wlodawer, A. Science 1989, 246, 1149.
8. Wall, I. D.; Leach, A. R.; Salt, D. W.; Ford, M. G.; Essex, J. W. J Med Chem 1999, 42, 5142.
9. Murray, C. W.; Baxter, C. A.; Frenkel, A. D. J Comput Aid Mol Des 1999, 13, 547.
10. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. J Mol Biol 2001, 308, 377.
11. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. J Comput Aid Mol Des 2002, 16, 151.
12. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. J Am Chem Soc 1990, 112, 6127.
13. Zou, X. Q.; Sun, Y. X.; Kuntz, I. D. J Am Chem Soc 1999, 121, 8033.
14. Apostolakis, J.; Pluckthun, A.; Caflisch, A. J Comput Chem 1998, 19, 21.
15. Jorgensen, W. L. MCPRO; Yale University: New Haven, CT, 1996.
16. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179.
17. Quiocho, F. A.; Vyas, N. K. Nature 1984, 310, 381.
18. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W., in preparation.
19. Havel, T. F.; Kuntz, I. D.; Crippen, G. M. Bull Math Biol 1983, 45, 665.
20. Bron, C.; Kerbosch, J. Commun ACM 1973, 16, 575.
21. Smellie, A. S.; Crippen, G. M.; Richards, W. G. J Chem Inform Comput Sci 1991, 31, 386.
22. Eleanor, J. G.; Artymiuk, P. J.; Willett, P. J Mol Graphics Mod 1997, 15, 245.
23. Richmond, T. J. J Mol Biol 1984, 178, 63.
24. Ponder, J. W. TINKER 3.6; Washington University: St. Louis, MO, 1998.
25. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Chem Phys Lett 1995, 246, 122.
26. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. J Phys Chem A 1997, 101, 3005.
27. Jayaram, B.; Sprous, D.; Beveridge, D. L. J Phys Chem B 1998, 102, 9571.
28. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. J Phys Chem 1993, 97, 10269.
29. Guvench, O.; Weiser, J.; Kolossvary, I.; Still, W. C. J Comput Chem 2002, 23, 214.
30. Desmet, J.; Demaeyer, M.; Hazes, B.; Lasters, I. Nature 1992, 356, 539.
31. Liu, H. Y.; Mark, A. E.; van Gunsteren, W. F. J Phys Chem 1996, 100, 9485.
32. Bouzida, D.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Gehlhaar, D. K.; Larson, V.; Luty, B. A.; Rose, P. W.; Verkhivker, G. M. Int J Quantum Chem 1999, 72, 73.
33. Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. J Comput Chem 1993, 14, 1347.
34. Ewald, P. Ann Phys 1921, 64, 253.
35. Essex, J. W.; Jorgensen, W. L. J Comput Chem 1995, 16, 951.
36. Essex, J. W.; Severance, D. L.; Tirado–Rives, J.; Jorgensen, W. L. J Phys Chem B 1997, 101, 9663.
37. Pierce, A. C.; Jorgensen, W. L. J Med Chem 2001, 44, 1043.
38. Rizzo, R. C.; Tirado–Rives, J.; Jorgensen, W. L. J Med Chem 2001, 44, 145.
39. McDonald, I. K.; Thornton, J. M. J Mol Biol 1994, 238, 777.
40. Wall, I. D.; Taylor, R. D.; Leach, A. R.; Ford, M. G.; Jewsbury, P. J.; Essex, J. W., in preparation.
41. Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. J Comput Aid Mol Des 2000, 14, 731.
42. Taylor, J. S.; Burnett, R. M. Proteins 2000, 41, 173.
43. Das, B.; Helms, V.; Lounnas, V.; Wade, R. C. J Inorg Biochem 2000, 81, 121.
44. Rejto, P. A.; Freer, S. T. Prog Biophys Mol Biol 1996, 66, 167.
45. Hilpert, K.; Ackermann, J.; Banner, D. W.; Gast, A.; Gubernator, K.; Hadvary, P.; Labler, L.; Muller, K.; Schmid, G.; Tschopp, T. B.; Vandewaterbeemd, H. J Med Chem 1994, 37, 3889.