J. Mol. Biol. (1998) **276**, 265–285

JMB

Rapid Refinement of Protein Interfaces Incorporating Solvation: Application to the Docking Problem

Richard M. Jackson, Henry A. Gabb and Michael J. E. Sternberg*

Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, PO Box 123 44 Lincoln's Inn Fields London WC2A 3PX, UK

A computationally tractable strategy has been developed to refine protein-protein interfaces that models the effects of side-chain conformational change, solvation and limited rigid-body movement of the subunits. The proteins are described at the atomic level by a multiple copy representation of side-chains modelled according to a rotamer library on a fixed peptide backbone. The surrounding solvent environment is described by ''soft'' sphere Langevin dipoles for water that interact with the protein *via* electrostatic, van der Waals and field-dependent hydrophobic terms. Energy refinement is based on a two-step process in which (1) a probability-based conformational matrix of the protein side-chains is refined iteratively by a mean field method. A side-chain interacts with the protein backbone and the probability-weighted average of the surrounding protein side-chains and solvent molecules. The resultant protein conformations then undergo (2) rigid-body energy minimization to relax the protein interface. Steps (1) and (2) are repeated until convergence of the interaction energy. The influence of refinement on side-chain conformation starting from unbound conformations found improvement in the RMSD of side-chains in the interface of protease-inhibitor complexes, and shows that the method leads to an improvement in interface geometry. In terms of discriminating between docked structures, the refinement was applied to two classes of protein-protein complex: five protease-protein inhibitor and four antibody-antigen complexes. A large number of putative docked complexes have already been generated for the test systems using our rigid-body docking program, FTDOCK. They include geometries that closely resemble the crystal complex, and therefore act as a test for the refinement procedure. In the protease-inhibitors, geometries that resemble the crystal complex are ranked in the top four solutions for four out of five systems when solvation is included in the energy function, against a background of between 26 and 364 complexes in the data set. The results for the antibody-antigen complexes are not as encouraging, with only two of the four systems showing discrimination. It would appear that these results reflect the somewhat different binding mechanism dominant in the two types of protein-protein complex. Binding in the protease-inhibitors appears to be ''lock and key'' in nature. The fixed backbone and mobile side-chain representation provide a good model for binding. Movements in the backbone geometry of antigens on binding represent an ''induced-fit'' and provides more of a challenge for the model. Given the limitations of the conformational sampling, the ability of the energy function to discriminate between native and non-native states is encouraging. Development of the approach to include greater conformational sampling could lead to a more general solution to the protein docking problem.

Keywords: Langevin dipole; side-chain optimisation; solvation energy; protein docking

**Corresponding author*

Present address: R. M. Jackson Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT.

Abbreviations used: LD, Langevin dipole; vdW, van der Waals; RMSd, root-mean-squared deviation; CDR, complementarity-determining region; BPTI and HPTI, bovine and human pancreatic trypsin inhibitor.

0022–2836/98/060265–21 $25.00/0/mb971519

Introduction

The structural nature of biological specificity is central to molecular biology as it underlies all types of molecular recognition process from the formation of large assemblies to enzyme-drug interactions. In particular, the prediction of protein-protein complexes, commonly known as the protein docking problem, presents a challenge to theoretical methods to model biological specificity because it imposes several requirements: (1) the adequate exploration of conformational space in order to sample productive binding geometries; (2) an ability to deal with conformational changes induced in the unbound protein structures on binding; (3) the adequate treatment of the thermodynamics of molecular association, including the large and sometimes opposing forces that favour association, including modelling solvation, interaction energies and conformational entropy. Here we address issues (2) and (3) given initial rigid-body docked complexes. We have developed a protein-protein interface refinement method that takes into account conformational change of side-chains and limited translational/rotational movement of the subunits relative to one another. A computationally tractable strategy to model the effects of solvation is included in addition to conventional molecular interaction energies.

There are many different algorithms designed to dock two protein molecules, they all initially use rigid-body approaches (e.g. see Shoichet & Kuntz *et al.*, 1991; Cherfils *et al.*, 1991; Jiang & Kim, 1991; Bacon & Moult, 1992; Walls & Sternberg, 1992; Katchalski-Katzir *et al.*, 1992; Totrov & Abagyan, 1994; Fischer *et al.*, 1995). We have recently developed such a method based on shape and electrostatic complimentarity using Fourier correlation theory (Gabb *et al.*, 1997). The algorithm FTDOCK has been used to dock six enzyme-inhibitor and four antibody-antigen complexes, starting from the unbound conformations. Here, as in previous work (Jackson & Sternberg, 1995), we concentrate on scoring predicted complexes generated by rigid-body docking.

The motivation for the work was to find a set of energy components that describe a protein in a water environment and that can be used in a rapid optimisation procedure to locate an energy minimum for that system. Here, we model the solvation energy by a ''soft'' sphere Langevin dipole for water, that interacts with the protein *via* electrostatic, van der Waals and hydrophobic components. The model is less sophisticated than all-atom statistical mechanical models for water such as Monte Carlo or molecular dynamics simulations. The large computational resources needed by these methods make them unsuitable for studying the large numbers of complexes generated by predictive docking. However, the model is more sophisticated than simple atomic solvation parameters (ASPs) commonly used to augment the vacuum potential in conformational refinement studies (e.g. see Wilson *et al.*, 1991; Cummings *et al.*, 1995; Weng *et al.*, 1996).

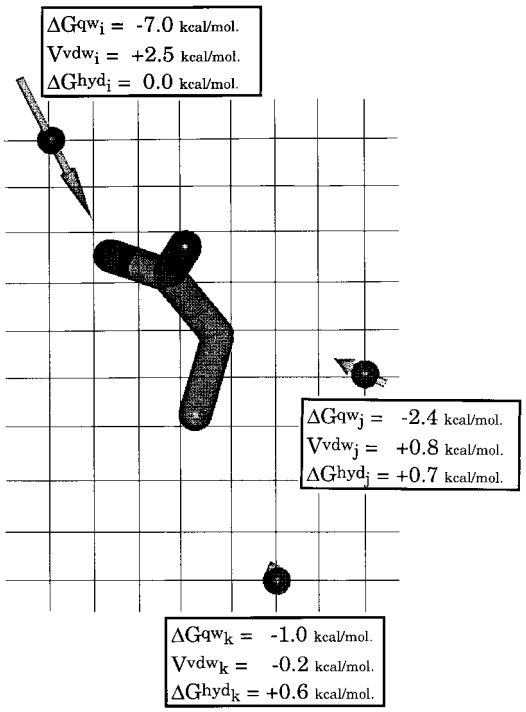

Figure 1. A representation of the soft sphere Langevin dipole model for solvation, with three representative particles in differing environments around a propionate ion. The individual electrostatic (ΔG^{qw}), van der Waals (V^{vdw}), and hydrophobic (ΔG^{hyd}) components to the interaction are shown in each case (as modelled by equations (4), (12) and (5), respectively) . The particles are part of a fine cubic grid surrounding the solute. Individual energy contributions are scaled by $(\Delta/3)^3$ (where $\Delta = 1$ Å/grid in this case) to reproduce the average density of water. The total interaction energy is calculated by summing over all sites and including the bulk contribution (modelled by equation (2)), which includes the solvent interaction outside the Langevin zone.

The soft sphere Langevin dipole (see Figure 1) treats water molecules as discrete dipoles that interact with the electric field of the protein but subject to random thermal fluctuations that reduce the effective electric field at the dipole itself. The model is a semi-empirical microscopic description as opposed to one that treats the solvent as a continuum such as the Poisson-Boltzmann equation (for a review of methods, see Harvey, 1989). In the following approach a solvent field is included in a mean field approach for predicting protein side-chain conformations. The model consists of a grid representing the centres of water molecules. These interact with the protein *via* an electrostatic energy component modelled by a Langevin function (Hill, 1956), a van der Waals component modelled by the Lennard-Jones equation and a hydrophobic component modelled by an electrostatic field-dependent hydrophobic energy. Calculation of the

solvation component is rapid and in spite of its apparent simplicity the model is able to reproduce the solvation energies of a large number of small organic molecules, including protein side-chain analogues.

The solvation treatment can be readily incorporated into an all-atom simulation of biomolecules where side-chains are modelled as discrete rotamers that interact with the backbone and the rotamers of adjacent residues. In essence, each residue feels the average of all environments weighted by their respective probabilities. Given the location and energy of interaction of the rotamers, the method iteratively refines a matrix of side-chain probabilities to give a self-consistent solution (Lee, 1994; Koehl & Delarue, 1994). The solvation potential is included in this self-consistent mean field method for predicting protein side-chain conformations. The presence of a particular water molecule at a particular site around a side-chain rotamer is modelled probabilistically and depends on the occupancy of that site by a side-chain atom of any other amino acid side-chain.

The method provides a fast means of conformational optimisation because all interactions are pre-calculated. Therefore, computation grows linearly with the number of degrees of freedom as opposed to exponentially as with conventional space searching algorithms, such as Monte Carlo and molecular dynamics simulations. The model has been combined with a rigid-body energy minimization routine to allow rapid optimisation of protein-protein interfaces. Thus, the method presented here takes into account the solvent environment and side-chain flexibility as well as a limited degree of translational and rotational movement. The speed of the method allows rapid screening of possible binding geometries and overcomes many of the limitations inherent to rigid-body docking.

Theory and Calculations

"Soft" sphere Langevin dipole model for solvation

The soft sphere Langevin dipole (LD) is an approach to model the solvation free energy of a solute with its surrounding water environment. The interacting water molecule is represented by a van der Waals particle, a Langevin dipole and a field-dependent hydrophobic energy (see Figure 1). The model described here is based on that developed by Luzhkov & Warshel (1992), but differs in detail, since it was not possible to refine iteratively the Langevin field to a self-consistent solution for all dipoles simultaneously due to the independent multiple-copy side-chain representation of the protein used here.

The model can be used to estimate solvation free energies by calibrating the solvent dipole μ_0 and including a hydrophobic energy to reproduce solvation free energies at room temperature. The total solvation energy is given by:

$$\Delta G^{\text{solv}} = \Delta G^{\text{qw}} + V^{\text{vdw}} + \Delta G^{\text{hyd}} + \Delta G^{\text{bulk}} \quad (1)$$

The protein-solvent electrostatic interaction energy (ΔG^{qw}) and the hydrophobic energy (ΔG^{hyd}) are described in detail below. V^{vdw} is the sum of van der Waals terms between all solute and solvent atoms as described by the Lennard-Jones (12-6) potential. The term ΔG^{bulk} is an additional term for the electrostatic energy of interaction between solute charges and the bulk solvent outside the LD grid representation. The interaction of a monopole or dipole moment of the solute at the designated centre of the system are modelled by the Kirkwood equation, which reduces to the Born equation for ions and the Onsager equation for dipoles:

$$\Delta G^{\text{bulk}} = -166Q_{\text{b}}^2/R_{\text{lgv}}(1 - 1/\varepsilon_{\text{w}}) - 166\mu_{\text{b}}^2/R_{\text{lgv}}^3[(2\varepsilon_{\text{w}} - 2)/(2\varepsilon_{\text{w}} + 1)] \quad (2)$$

Q_{b} and μ_{b} are the monopole and dipole moments, respectively, of the solute at the geometric centre of the system. R_{lgv} is the spherical radius for the grid of Langevin dipoles from the geometric centre. ε_{w} is the dielectric constant of bulk water ($\varepsilon_{\text{w}} = 80$). This additional term is of particular significance for charged species where the bulk interaction can be large.

The average polarisation of a solvent molecule was calculated as follows. The solute molecule is surrounded by a three-dimensional cubic grid that has two spacings. The inner region up to 4 Å distant from any protein atom has a grid spacing of 1 Å, whilst the region beyond this out to 10 Å has a coarser spacing of 3 Å. The resultant Langevin, van der Waals and hydrophobic energies are scaled by a factor $(\Delta/3)^3$, where Δ is the grid spacing in Å, in order to preserve the average density of water (as represented by a cubic grid) at the experimental value for water of 1 g/dm^3. A fine inner grid spacing was found to be important for reproducing solvation energies that were independent of the orientation of the molecule on the grid. Each grid point is occupied by a point dipole, which is polarised towards the local field as approximated by a Langevin type function (Warshel & Levitt, 1976):

$$\mu_i^{\text{L}} = e_i\mu_0(\coth \zeta_i - 1/\zeta_i)$$
$$\zeta_i = 332\ C'\mu_0\xi_i^0/k_{\text{B}}Td(r) \quad (3)$$

where μ_i^{L} is the effective Langevin dipole with an upper limit of μ_0, e_i is a unit vector in the direction of the local electric field, ξ_i^0, which arises due to the solute permanent charges. k_{B} is the Boltzmann constant and T is the temperature. C' is an adjustable parameter, set here to unity. $d(r)$ is a screening function that represents the reduction of the electric field at the ith dipole due to all the other Langevin dipoles. $d(r)$ was assigned a value of 1.0 for $r_{ij} \leqslant 4$ Å and a value of 3.4 for $r_{ij} > 4$ Å. This represents no solvent screening within what is essen-

tially the first solvation shell of the solute molecule, followed by a large increase beyond this distance (Russell & Warshel, 1985). The equation was solved non-iteratively (Warshel & Levitt, 1976; Russell & Warshel 1985) with the distance-dependent screening function chosen to give the relevant polarisation of the Langevin dipoles and hence reproduce the electrostatic solvation energy component. The Langevin contribution to solvation is then given by:

$$\Delta G^{qw} = \Delta G^{lgv} = 332/2 \sum_i \mu_i^L \xi_i^0 \qquad (4)$$

where the factor of 1/2 allows for the energy invested in orienting the solvent dipoles in the field, and the energy is summed over all Langevin dipoles.

In contrast to previous implementations of the Langevin model, we did not automatically exclude grid points within a specific van der Waals distance of the solute i.e. a hard sphere representation. Instead the dipole was treated as a soft sphere, where both the electrostatic and van der Waals (vdW) interaction were calculated. If the energy for the interaction (solvation enthalpy) was greater than a cut-off energy the particle was excluded from further consideration. The cut-off was set at 0 kcal/mol, since the unfavourable vdW repulsion term quickly overcomes any favourable electrostatic interaction. This soft sphere treatment and the fine grid spacing were both found to be critically important in reproducing the observed solvation energies and making the result independent of the orientation of the molecule on the grid.

The hydrophobic contribution ΔG^{hyd} is generally associated with a surface area (Richards, 1977) or excluded volume (Richmond, 1983) dependent term in treatments of solvation. In addition to the final model described below, simple surface area and excluded volume-dependent models were tested for their ability to reproduce observed free energies of solvation for hydrocarbons and (at the other end of the spectrum of dielectric response) polar solutes including ions. In accordance with the findings of Luzhkov & Warshel (1992), a field-dependent hydrophobic effect resulted in the closest agreement with experiment for the full range of dielectric responses in the solute data set. In this model the hydrophobic effect is dependent on the local field on the surface solvent molecules. When the field is small (e.g. for non-polar solutes) the hydrophobic effect is retained. However, as the field strength increases, the hydrophobic effect decreases, eventually to zero at high field strengths (such as exist around ions). The rationale for the field dependence is that the hydrophobic effect is primarily entropic at room temperature and relates to the ''freezing out'' of low-energy orientations of the water molecule on going from bulk solution to the surface of a non-polar solute. However, an increase in polarity of the solute will increase the number of low-energy orientations available to the surface water, and this increase in the degrees of freedom reduces the hydrophobic effect. We found the simple model described above to be almost as good as more complex models involving more extensive parameterisation.

The hydrophobic energy is given by;

$$\Delta G^{hyd} = \sum_i (n_i E^{hyd} - 332/2\mu_i^L \xi_i^0) \qquad (5)$$

$$(\Delta G_i^{hyd} \geqslant 0, \text{ and } n_i = 1, 2, 3, 4)$$

where i is summed over all Langevin dipoles within the surface volume element described below. The model involves defining a surface volume element, described as all water centres within a distance r_{min}, where r_{min} is the minimum of the Lennard-Jones well (see (equation (12), $r_{min} = (2A_i \times A_j / B_i \times B_j)^{1/6}$). Each contact of a water molecule with a solute molecule is penalised by a constant amount, $E^{hyd} = 1.6$ kcal/mol, which represents the free energy change of moving a water molecule from bulk to the first solvation shell of the solute. If there is more than one such defined contact the penalty is increased by the number of solute contacts, n_i, to a maximum value of four contacts (which represents the removal of all first solvation shell water molecules in a tetrahedral ice lattice), which represents a water molecule almost completely removed from bulk solution. If the negative of the free energy of interaction of the Langevin dipole with the solute is greater than or equal to the hydrophobic penalty, ΔG_i^{hyd}, the energy penalty is zero. Conversely, if the field is zero the full penalty is retained.

To benchmark this model, solute geometries were built and energy minimised using the INSIGHT molecular modelling package (Biosym Technologies, Inc.). Optimisation of the parameters of the Langevin and hydrophobic models discussed above was performed as follows. The only adjustable parameters for the model are μ_0 and $d(r)$ in the Langevin equation, and n_i and E^{hyd} in the field-dependent hydrophobic equation. The values of n_i and E^{hyd} can be parameterised initially by using hydrocarbon data (which have zero electrostatic field strength) and then testing their transferability to polar solutes. μ_0 must be close to the dipole moment of water and was taken here as 0.35 eÅ, which is standard in LD solvation energy calculations (Warshel & Russell, 1984). As described above, $d(r)$ was fitted to reproduce observed solvation free energies. There is no solvent screening in the first solvation shell (<4 Å) and a constant value above this distance. The only other adjustable parameters are the atomic vdW parameters and point charges, which are usually optimised for a given solvation model to reproduce the experimental data (e.g. see Lee *et al.*, 1993, Sitkoff *et al.*, 1994). We tested two existing force fields: the AMBER united atom force field (Weiner *et al.*, 1984) and the AMBER/PARSE force field with vdW parameters taken from AMBER and the

point charges taken from the PARSE force field (Sitkoff *et al.*, 1994: the PARSE force field is parameterised to reproduce solvation energies of small molecules using a polarisable cavity model using the Poisson-Boltzmann equation). Readjustment of vdW parameters or point charges was not performed (since PARSE has a united atom charge representation for aliphatic groups) with the exception that aromatic ring hydrogen atoms of groups Phe, Tyr and Trp were assigned vdW parameters and point charges to be consistent with the PARSE parameterisation in the AMBER/PARSE force field.

In the calculations a dielectric constant of $\varepsilon = 1$ and an atom-atom based cut-off of 10 Å was used. It was found that use of a cut-off that is equal to the grid radius gives results that are the least sensitive to the grid radius being used (which was also 10 Å). Solute-solvent electrostatic interaction energy beyond this point (ΔG^{bulk}) was modelled by the Kirkwood equation (equation (2)).

Self-consistent mean field approach and inclusion of solvation energy

We have implemented a self-consistent mean field approach to optimise protein side-chain conformations, given the main-chain atom co-ordinates. Here, we give only a brief description of the method as applied here, the details have been presented elsewhere (Lee, 1994; Kohl & Delarue, 1994). The method describes a protein of **N** residues whose main-chain co-ordinates (N, C^{α}, C, O and C^{β}) are defined. Each residue side-chain, i, (with the exception of residues defined by the backbone, Gly, Ala and Pro) has a discrete number, K_i, of possible conformations (side-chain rotamers). Hence one can define a conformational matrix **CM** of dimension, **N** by max(K_i), in which each rotamer, k, has a probability of $CM(i,k)$, which is bounded by the condition that the sum of the probabilities for a given residue, i, must be equal to 1.

The object of the mean field approach is to determine the most probable set of side-chain rotamers from an ensemble of rotamers. In such a closed system the potential of mean force, $E(i,k)$, on the kth rotamer of residue, i, is given by:

$$E(i,k) = V(\boldsymbol{x}_{ik}) + V(\boldsymbol{x}_{ik}, \boldsymbol{x}_{\text{mc}}) + \sum_{j=1}^{N} \sum_{j \neq i}^{K_j} \sum_{l=1} CM(j,l) V(\boldsymbol{x}_{ik}, \boldsymbol{x}_{jl}) + E_{\text{sol}}(i,k) \quad (6)$$

where V is the potential energy, $\boldsymbol{x}_{ik}$ are the co-ordinates of atoms in rotamer k of residue i and $\boldsymbol{x}_{\text{mc}}$ are the co-ordinates of atoms in the protein main-chain. The first term in (equation (6) represents the internal energy of the rotamer. The second term represents the interaction energy between the rotamer and all the main-chain atoms. These two values are constant for a given rotamer on a given main-chain. The third term represents the interaction energy between the rotamer and all the rotamers of other residues weighted by their respective probabilities. Note that rotamers of a given residue, i, do not interact with one another.

In addition to the potential generated by other protein atoms, the fourth term $E_{\text{sol}}(i,k)$ represents the potential of mean force, $E_{\text{sol}}(i,k)$, acting at rotamer, k, of residue, i, due to the surrounding solvent environment. The solvent potential acting at a rotamer is in turn dependent on the occupancies of rotamers in surrounding residues. In the present implementation the protein is surrounded by a solvent grid and all solvent sites that are within a cut-off of any rotamer ($\leqslant 4$ Å) and not in vdW conflict ($V_{\text{vdw}} \leqslant 0$) with the fixed main-chain atoms of the protein are retained. The free energy of interaction ($G(\boldsymbol{x}_{ik}+\boldsymbol{x}_{\text{mc}}, \boldsymbol{x}_{\text{LD}})$) between the main-chain plus side-chain atoms of a given rotamer and a ''soft'' sphere Langevin dipole (vdW, electrostatic and hydrophobic components) within a given cut-off ($\leqslant 4$ Å) are calculated. Hence each residue rotamer has a number, $M_{i,k}$, of precalculated interactions with all possible surrounding solvent sites. The LD of each solvent site is then checked to see if it is in vdW conflict of any rotamer of any other side-chain in the system. If the LD is in conflict with a rotamer then the probability of the site is dependent on the probability, $CM_{\text{conflict}}(j,l)$, of the rotamer in conflict. If no rotamer is in conflict the probability of the site is 1. Thus the additional solvation term is:

$$E_{\text{sol}}(i,k) = \sum_{\text{LD}=1}^{M_{i,k}} \left[1 - \sum_{j=1}^{N} \sum_{j \neq i}^{K_j} \sum_{l=1} CM_{\text{conflict}}(j,l) \right] \times G(\boldsymbol{x}_{ik} + \boldsymbol{x}_{\text{mc}}, \boldsymbol{x}_{\text{LD}})$$

$$\text{with} \quad 0 \leqslant \left[1 - \sum_{j=1}^{N} \sum_{j \neq i}^{K_j} \sum_{l=1} CM_{\text{conflict}}(j,l) \right] \leqslant 1 \quad (7)$$

where $\boldsymbol{x}_{\text{LD}}$ is the co-ordinate of the soft sphere Langevin dipole, and the probability of a particular rotamer Langevin dipole interaction is $\leqslant 1$ but $\geqslant 0$, but depends on occupancy of that site by adjacent side-chain rotamers.

Given the effective potentials acting on all K_i possible rotamers of residue, i, the probability of the rotamer can be calculated according to the Boltzmann principle as:

$$CM(i,k) = \frac{e^{-E(i,k)/RT}}{\sum_{k=1}^{K_i} e^{-E(i,k)/RT}} \quad (8)$$

where R is the Boltzmann constant and T the temperature. RT is equal to 0.592 kcal/mol at 298 K. The values of $CM(i,k)$ are substituted into (equation (6) and values of $E(i,k)$ recalculated. This process is repeated until values of $CM(i,k)$ reach convergence. In this formalism, all possible rotamer-backbone, rotamer-rotamer and rotamer-solvent interactions need to be calculated only once. Only the elements

of the probability matrix **CM** are refined. In the refinement process, all elements of the conformational matrix are initialised as $1/K_i$ or zero and a matrix ''memory'' is set to reduce oscillations in the conformational matrix during refinement (Kohl & Delarue, 1994) , such that **CM** is updated as:

$$\mathbf{CM} = \mathbf{CM}\lambda + (1-\lambda)\mathbf{CM}_{\text{old}} \quad (9)$$

where $\lambda = 0.5$. Convergence was monitored by root-mean-squared change in amplitude of the conformational matrix according to:

$$\text{rms}MAT = \sqrt{\sum_{i=1}^{N}\sum_{k=1}^{K_i}(CM(i,k) - CM_{\text{old}}(i,k))^2} \quad (10)$$

and secondly by the change in energy of the average structure after each step. Convergence was deemed to be achieved when the change in the conformational matrix rms$MAT < 10^{-4}$. The predicted structure corresponds to the highest probability rotamer for each residue.

Rotamer library and side-chain construction

The rotamer library from Tuffery *et al*. (1991) was used to define the side-chain dihedral angles χ_1, χ_2 and χ_3. The missing dihedral angles in the library χ_4 for arginine and lysine were defined as 180° and χ_5 for arginine as 0°. On taking into account rotational symmetry about χ_2 and χ_3, three rotamers of Glu and three of Tyr were removed from the library, since they appear twice. Excluding the non-rotameric glycine, alanine and proline residues, the number of rotamer conformations in the library varies between three for valine through to 16 for lysine. These rotamers are constructed in addition to the original side-chain conformation (included as an additional rotamer) taken from the PDB co-ordinate file. All main-chain atoms (N, C^{α}, C, O and C^{β}) were considered static. Alternate side-chain rotamers were constructed from the bond lengths and angles in the existing co-ordinate file after regularisation. The side-chain χ angles were used to generate new co-ordinates for alternate rotamers. Hydrogen atoms were added using existing bonds and geometries with hydrogen atoms being added according to their equilibrium bond lengths, bond angles and dihedral angles. After this stage additional rotamers were generated to specify the hydrogen positions of Ser, Thr, Cys and Tyr. For Ser, Thr and Cys(h) the *gauche*$^-$, *trans* and *gauche*$^+$ conformations of the χ_2 rotamer defined as (C^{α}-C^{β}-O(S)$^{\gamma}$-H$^{\gamma}$) were generated for each parent rotamer. For Tyr the two *cis* and *trans* conformations of the χ_6 rotamer defined as ($C^{\varepsilon 1}$-C^{ζ}-O^{η}-H^{η}) were generated for each parent rotamer. For histidine, both the $N^{\varepsilon 2}$ and $N^{\delta 1}$ protonated forms of histidine were generated for each rotamer. A decision on the connectivity of cysteine residues must be made. An automated procedure was implemented that uses distance constraints to define disulphide-bonding Cys residues ($r_{C\alpha\text{-}C\alpha} < 7$ Å and $r_{C^{\beta}\text{-}C^{\beta}} < 4.5$ Å), based on a classification procedure described by Harrison (1996).

Potential energy function

Here, we develop a fully integrated energy scheme that describes a protein in solution. In theory it should be possible to simply implement the solvation scheme described for solute solvation above using a dielectric constant of $\varepsilon = 1$ to describe both the protein-protein and protein-solvent electrostatic interaction energies. This would be appropriate if: (1) we were using a thermal averaging procedure to describe the protein-protein interactions; (2) solute polarisability effects were implicit in the model and; (3) we explicitly simulated the effects of solvent counter-ions. Furthermore, a major problem of applying a microscopic approach (as opposed to a macroscopic approach) is that we are directly calculating all of the energy terms, and this involves the sum of large contributions (namely, V^{qq} , the interaction between the charges of the rotamer of interest and the charges of the rest of the protein and ΔG^{qw}, the charge-solvent interaction). Small errors in calculation can lead to a large absolute error. The treatment of ion pairs in solution is a classic example of such a situation involving large energy contributions that essentially cancel out (see Warshel & Russell, 1984). Such a system is particularly relevant to modelling protein-protein interactions. In continuum approaches one chooses an appropriate dielectric constant to represent the solute and solvent media, and the mutual compensation of V^{qq} and ΔG^{qw} can be assumed by using a large dielectric constant.

An approach to reduce the absolute error in applying the Protein dipoles Langevin dipoles model to simulate the energetics of proteins (as opposed to solute solvation) has been described by Warshel and co-workers and is referred to as the scaled microscopic model (Lee *et al*., 1993). It assumes a uniform internal dielectric constant ($\varepsilon_{\text{in}}^{\alpha}$) for the protein and solvent LD grid, which is in turn surrounded by bulk solvent ($\varepsilon_{\text{w}} = 80$). All terms are obtained by the standard approach with the exception that solute polarisability terms are assumed to be implicit in the model. The elctrostatic energy is given as:

$$\Delta G^{\text{elec}} = V^{qq} + \Delta G^{qw} + \Delta G^{\text{hyd}} \quad (11)$$

where V^{qq} is the vacuum Coulombic interaction energy between charged groups of the protein. Each of the terms is calculated as described above, assuming $\varepsilon = 1$ and is scaled by the factor $1/\varepsilon_{\text{in}}^{\alpha}$. We used a value of $\varepsilon_{\text{in}}^{\alpha} = 4$ with the AMBER/PARSE force field, a non-bonded cut-off of 10 Å and a distance dependent dielectric for the protein-protein non-bonded interactions ($\varepsilon = r_{ij}$). The vdW interactions between any two atoms A and B were modelled using the Lennard-Jones (12-6) potential and the geometric mean approximation for the

vdW parameters:

$$V^{\mathrm{vdw}} = [A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6}] \qquad (12)$$

Hydrogen atoms were considered to have zero radius. Only residues whose C^{β} atoms (C^{α} of glycine) are within 15 Å of the C^{β} of residue i are included in the electrostatic and vdW non-bonded energy summation. The cut-off corresponds to the distance between the C^{β} atoms of two fully extended arginine residues placed end-to-end.

To ensure that all interaction energies are constrained to be within boundaries and therefore produce a smoother energy surface, the following scheme was used. Since we use discrete rotamers and V^{vdw} tends to infinity as r_{ij} tends to 0, V^{vdw} was truncated to a maximum value of 2.5 kcal/mol. This value was chosen to correspond with an optimal electrostatic interaction scheme, in which the minimum allowed distance separation between two interacting charges q_i and q_j was 3 Å for two heavy atoms, 2 Å for a heavy atom and hydrogen and 1 Å for two hydrogen atoms i.e. atom pairs that were closer than these values were re-scaled to these values for the purpose of calculating electrostatic energies.

Torsional potentials are not included in the energy function, since side-chain conformations were derived directly from observed rotamers. However, in the present implementation of the model we replace the internal energy of the rotamer (the first term in (equation (6)) with an energy relating to the observed rotameric preferences for particular side-chain conformations. The rationale for this procedure is that protein side-chains have distinct preferences for certain conformations and the molecular mechanics based energy landscape of the intra-molecular side-chain energy is either highly discontinuous or completely featureless making it unsuitable for use as an energy term. The internal energy of a particular rotamer is given by a Boltzmann weighted energy function relating to the observed frequency of occurrence, $P_{i,k}$ of the rotamer in the population distribution:

$$V(\mathrm{x}_{ik}) = -RT \ln\left[P_{i,k} \Big/ \sum_{k=1}^{K_i} P_{i,k} \right] \qquad (13)$$

The observed frequencies for rotamers were taken from the rotamer library presented by Tuffery *et al.* (1991). The side-chain conformations of the original structure are assigned to a particular rotameric class based on whether the side-chain torsion is within ±40° from the same torsion angle in the rotamer library. Non-rotameric residues, i.e. those rotamers that do not fall into any observed rotamer class, were assigned a minimum population frequency of 2%, thus limiting the maximum energy penalty for sparsely populated torsional space.

Cysteine residues involved in disulphide bridges were assigned (as discussed above) on Cys-Cys distance criteria. In order to direct their formation, a harmonic distant-dependent potential was applied to the S^{γ}-C^{β} ($K_{S\gamma\text{-}C\beta} = 50$ kcal/mol. Å^2) and S^{γ}-S^{γ} ($K_{S\gamma\text{-}S\gamma} = 50$ kcal/mol. Å^2) atom pairs, in a similar manner to the ECEPP energy function (Momany *et al.*, 1975).

Rigid-body minimization and refinement

Following each complete cycle of side-chain mean field optimisation, rigid-body minimization (Islam, 1986) was performed on the resultant coordinates. Only interface residues whose C^{β} atoms (or C^{α} in the case of glycine) are within 15 Å of any C^{β} of the other molecule are included in the minimization. The larger molecule (enzyme or antibody FAB) was kept stationary while the six degrees of freedom (three rotational and three translational) of the rigid smaller molecule (inhibitor or protein antigen) are moved according to the path determined by the derivatives to minimise the intermolecular interaction energy. Numerical derivatives were calculated using the same all-atom model and potential energy function described above (except here the solvation contribution cannot be included). These were used in a steepest-descents minimization routine. The maximum rotation step size was 1° and maximum translational step size 0.3 Å. Minimisation continues until the energy of the system decreases by less than 10^{-6} kcal/mol for any given step. The coordinates are updated according to the resultant translational/rotational vectors and a further cycle of side-chain optimisation was performed.

All calculations were carried out on a Silicon Graphics Power Challenge multiprocessor workstation. Timings are given in Table 1, for the CPU time of both the side-chain mean field optimisation and rigid-body minimization steps, on a single R10000 processor. Timings are given for only the largest and smallest of the protease-protein inhibitor and antibody-antigen systems, both with and without solvent present. Note, timings are given for an average structure, but individual run times could vary considerably depending on the number of cycles of refinement and the extent of the protein-protein interface in contact.

Results

Evaluation of procedure

Solvation energies of small molecules in solution

Solvation energies calculated using the soft sphere Langevin dipole model (described above) for amino acid side-chains, their charged analogues, small hydrocarbons and water are given in Table 2 for the AMBER/PARSE force field. Given that the soft sphere Langevin dipole model has relatively few adjustable parameters and that the charges and vdW parameters have not been adjusted, the results are in very good agreement with experiment with the least-squares fit $\Delta G^{\mathrm{solv}}_{\mathrm{calc}} = 0.98\Delta G^{\mathrm{solv}}_{\mathrm{exp}} + 0.26$ and a correlation coeffi-

Table 1. Timings

System		Total no. structures	Run time (h:min)	Average time per structure (min:s)
Subtilisin-chymotrypsin In.	*In vacuo*	26	1:16	2:55
Subtilisin-chymotrypain In.	Solvated	26	4:51	11:12
Kallikrein-BPTI	*In vacuo*	365	25:10	4:08
Kallikrein-BPTI	Solvated	365	82:26	13:33
HyHEL5-lysozyme	*In vacuo*	519	58:41	6:47
HyHEL5-lysozyme	Solvated[a]	50	38:26	46:07
HyHEL10-lysozyme	*In vacuo*	765	104:59	8:14
HyHEL10-lysozyme	Solvated[a]	50	35:09	42:11

All calculations were carried out on a Silicon Graphics Power Chellenge multiprocessor workstation. Timings are given for the CPU time of both the side-chain mean field optimisation and rigid-body minimization steps, on a single R10000 processor.
[a] Results for the top fifty structures of the in vacuo refinement.

cient of 0.93 for the uncharged species (results are more highly correlated if charged species are included). The results using AMBER charges and vdW parameters were poor in comparison (with the least-squares fit $\Delta G^{solv}_{calc} = 0.71\Delta G^{solv}_{exp} + 1.37$ and a correlation coefficient of 0.87; results not shown) with notable differences from experiment for aromatic and alcohol side-chain groups.

Discrepancies between G^{solv}_{exp} and ΔG^{solv}_{calc} using the AMBER/PARSE force field arise because the PARSE force field was parameterised using a polarisable cavity model with appropriately parameterised solute atomic radii. These do not necessarily correspond to the effective cavity produced by the soft sphere model used here. Indeed, there is clearly over polarisation of the solvent for the amide moiety (Asn, Gln and backbone) and sulphide moieties (Met and Cys) with the AMBER/PARSE charges and radii. Furthermore, although ΔG^{solv}_{calc} for charged species are in qualitative agreement with experiment:

$$\Delta G^{solv}(asp-) > \Delta G^{solv}(glu-) > \Delta G^{solv}(cys-)$$
$$> \Delta G^{solv}(lys+) > \Delta G^{solv}(his+)$$

the method overestimates the value of ΔG^{solv} for cations (by up to 10%) but underestimates it for anions (by up to 7%). This could be corrected by re-scaling the vdW or charge parameters. However, as in previous microscopic and macroscopic treatments of solvation, this arises because we have a static model that assumes a unit solute-solvent radial distribution function (as opposed to reproducing the radial distribution function

Table 2. Experimental and calculated free energies of hydration

Molecule	Residue	ΔG^{hyd}	ΔG^{qw}	ΔG^{bulk}	V^{vdw}	ΔG^{solv}_{calc}	ΔG^{solv}_{exp}
Methanol	Serine	4.0	−5.5	−0.4	−3.0	−4.9	−5.1
Ethanol	Threonine	6.2	−5.4	−0.4	−3.9	−3.5	−4.9
N-Butylamine	Lysine	8.9	−5.2	−0.4	−6.7	−3.4	−4.4
Methylthiol	Cysteine	5.1	−3.2	−0.3	−4.1	−2.4	−1.2
Methyl ethyl sulphide	Methionine	9.2	−5.1	−0.3	−6.4	−2.7	−1.5
Acetic acid	Aspartic acid	5.9	−6.4	−0.4	−5.8	−6.7	−6.7
Propanoic acid	Glutamic acid	8.0	−6.1	−0.4	−6.7	−5.2	−6.5
Acetamide	Asparagine	5.6	−10.2	−1.6	−5.4	−11.5	−9.7
Propionamide	Glutamine	6.5	−10.5	−1.5	−6.0	−11.4	−9.4
N-Propyl guanidine	Arganine	10.2	−8.6	−0.4	−8.9	−7.8	−10.9
Toluene	Phenylalanine	11.1	−2.8	0.0	−9.2	−0.8	−0.8
p-Cresole	Tyrosine	12.3	−6.8	−0.2	−10.2	−5.0	−6.1
Methylindole	Tryprophan	15.1	−7.1	−0.5	−12.0	−4.5	−5.9
N-Methyl acetamide	Main-chain	6.7	−11.2	−1.6	−6.7	−12.8	−10.1
Water		1.3	−6.4	−0.5	−1.1	−6.7	−6.3
Ethane		5.8	0.0	0.0	−4.0	1.8	1.8
Propane	Valine	7.0	0.0	0.0	−5.0	2.0	2.0
Butane	Isoleucine	8.2	0.0	0.0	−5.9	2.3	2.2
Isobutane	Leucine	8.4	0.0	0.0	−6.1	2.3	2.4
Pentane		9.8	0.0	0.0	−7.5	2.3	2.3
Hexane		11.0	0.0	0.0	−8.4	2.6	2.6
N-Butyl ammonium	$Lysine^+$	3.3	−57.5	−14.6	−1.4	−70.2	−69.2
Methylthiol ion	$Cysteine^-$	0.0	−57.0	−15.6	1.1	−71.5	−76.8
Acetate ion	$Aspartate^-$	0.0	−58.1	−15.3	−1.4	−74.8	−80.7
Propionate ion	$Glutamate^-$	1.1	−57.6	−15.2	−1.6	−73.3	−79.1
Methyl imidazolium	$Histidine^+$	2.0	−53.0	−14.8	−4.1	−69.9	−64.1
N-*p*-Guanidinium	$Arginine^+$	4.0	−52.7	−14.5	−6.3	−69.5	−66.1

All energies are in kcal/mol.

directly with the given charge and vdW parameters, such as with statistical mechanics methods) so that the effective solute-solvent separation corresponds more closely to the vdW distance than the true equilibrium distance.

The PARSE/AMBER force field reproduces the experimental solvation energies well considering the limited parameterisation of the model. Furthermore, the model reproduces the physics of solvation in the context of an all-atom model that includes vdW and electrostatic energy components as well as a field-dependent hydrophobic term. This can be readily incorporated into a molecular mechanics based approach that, when combined with the mean field approach, is fast to calculate.

Convergence characteristics and locating the global minimum

Locating the global minimum in an energy landscape is the key goal of an optimisation procedure. However, in systems with many degrees of freedom and a highly convoluted energy landscape such as exists in proteins, it is difficult to prove that a global optimum has been found. The advantage of the mean field approach is that the finite number of degrees of side-chain freedom constitute a closed (or constrained) system.

In the mean field approach the final probability matrix **CM** is influenced by initial probabilities (the initial $CM(i,j)$ values assigned to side-chain rotamers) when refinement is carried out at 298 K. If in initiating the conformational matrix one rotamer is chosen at random and given a probability of 1 whilst other rotamers are given a probability of zero for each residue and then optimisation is carried out at 298 K, the result for a series of different initial conformational matrices is a series of different final matrices with different energies and side-chain distributions (as demonstrated by Koehl & Delarue, 1994). However, temperature is the key issue, since biasing the probability matrix at relatively low temperature causes the system to get trapped in local energy minima.

We looked at several refinement procedures. Two of the protocols have been described by Koehl & Delarue (1994) in which; (i) the initial conformational matrix is set to zero. After the first step of refinement the probability of a rotamer is described by its internal energy and its interaction with the backbone, and (ii) all rotamers of a residue are initially given the same probability ($CM(i,j) = 1/K_i$).

We also started with randomly biased initial probability matrices (described above) and looked at other methods including (iii) a perturbation approach in which the system is initially perturbed with a temperature jump to 6000 K followed by refinement at 298 K in subsequent steps. This is almost identical with (ii), since the heat jump wipes clean any initial assumptions made about the rotamer probabilities. The other protocols were simulated annealing methods, in which (iv) the initial temperature was set at 6000 K and the temperature was reduced in subsequent steps according to:

$$T_i = gT_{i-1} \tag{14}$$

until 298 K was reached. g is a scaling factor that can have a value between 0 and 1. And (v) the temperature is linked to the convergence characteristics (see (equation (10)) of the conformational matrix according to:

$$T_i = T_{298\mathrm{K}} + \mathrm{rms}MAT_{i-1}T_{1000\mathrm{K}} \tag{15}$$

values of rmsMAT tend to zero as the optimisation converges.

The convergence characteristics in terms of energy during the course of refinement of the trypsin-BPTI interface are given in Figure 2. All five of the methods described above converge to the same minimum (same values for **CM**), albeit by different pathways. Furthermore, this minimum corresponds to the lowest energy conformational matrix produced when starting from a series of biased probability matrices described above. This behaviour was found for all the other systems tested (results not shown), and would appear to be a general result. This suggests that a global minimum is achieved for a given fixed backbone geometry and a finite number of side-chain rotamers using any of the protocols (i) to (v). Note, this is not the same as finding a global minimum for the (unconstrained) system as a whole.

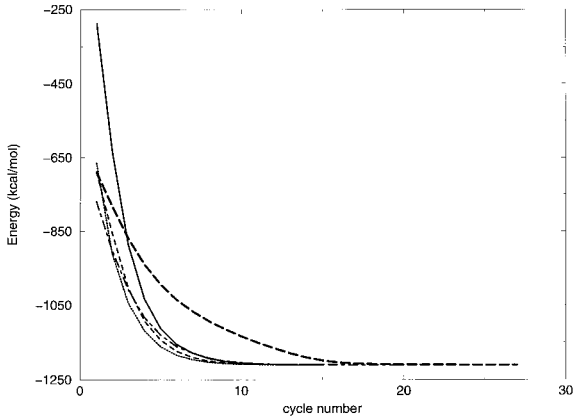

Figure 2. Convergence characteristics of diferent refinement protocol in terms of the energy of predicted strucutres in the course of refinement of the trypsin-BPTI interface. Initialise all $CM(i,j)$ as zero (continuous line). Initialise all $CM(i,j)$ as $1/K_i$ (dotted line). Perturbe initial matrix with temperature jump of 6000 K, followed by refinement at 298 K (short dashed line). Initialise at 6000 K cool according to $T_i = gT_i - 1$ (long dashed line). System cooled according to $T_i = T_{298\ \mathrm{K}} + \mathrm{rms}MAT_{i-1}T_{1000\ \mathrm{K}}$ (dot-dashed line).

Application to protein docking

Refinement of protein-protein complexes: using unbound conformations

We present a detailed analysis of three protease-protein inhibitor complexes in which refinement was carried out on the unbound structures. The unbound structures were optimally superimposed in terms of their C^{α}-RMS on the bound crystal co-ordinates and refined using alternate cycles of mean field side-chain optimisation followed by rigid-body minimisation until the interaction energy reached a minimum. The resulting conformations were compared with the crystal complexes, and the energetics of side-chain interactions were compared with those of the crystal complex after refinement. The object of the study was to establish if the refinement procedure leads to an improvement in conformation and energetics at the protein-protein interface given a suitably close docked solution to the complex but using the unbound conformations of the two proteins. The systems include trypsin-BPTI, subtilisin-chymotrypsin inhibitor and α-chymotrypsin-ovomucoid, which have been extensively studied in previous protein docking studies (Shoichet & Kuntz, 1991; Bacon & Moult, 1992; Fischer *et al*., 1995). The analysis concentrates on the inhibitor molecules because these molecules recognise the target protein and undergo almost all the conformational changes on binding. Only residues at the interface (residue C^{β} within 10 Å of any C^{β} of the enzyme. The C^{α} of glycine is used) whose interaction energy was above a threshold value (-1 kcal/mol < interaction energy > 1 kcal/mol) were included in the analysis. We present several structural and energetic measures by which the refinement procedure can be assessed, these include: (i) graphical representation of the conformations of inhibitor residues at the interface before and after refinement (see Figure 3(a) to (c)) (ii) side-chain RMS deviations from the crystal co-ordinates before and after refinement (see Figure 4(a) to (c)), (iii) energies of side-chain interactions following refinement for both the unbound and bound conformations (see Figure 5(a) to (c)). We now consider each of the systems in turn.

Trypsin-BPTI. The conformational changes brought about by refinement predict both the conformations and interaction energies of the key residues Lys15, Arg17 and Arg39 very well. In particular, Figure 4(c) shows the large reduction in side-chain RMSd for these residues in spite of the overall increase in C^{α}-RMSd following refinement. The guanidinium and amine moieties of these residues move to occupy the same locations as those seen in the crystal complex, and are predicted to have the most favourable interactions with the enzyme in agreement with the refinement of the trypsin-BPTI crystal complex.

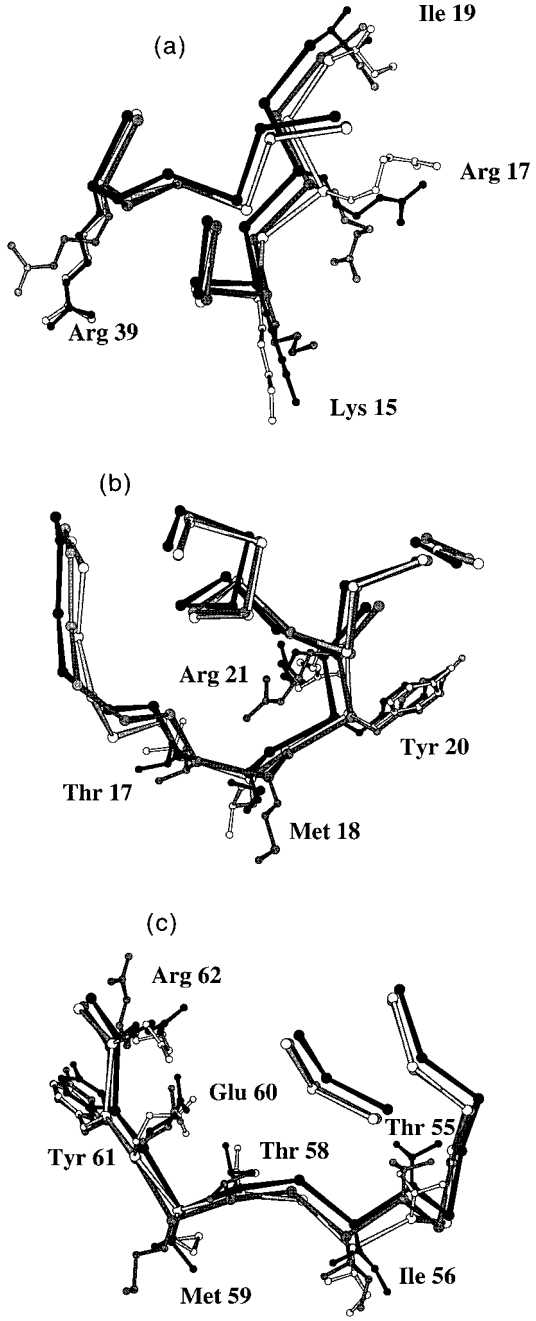

Figure 3. Comparison of the conformation of inhibitor side-chains at the interface. The experimental crystal complex (in black), before (in grey) and after (in white) refinement. (a) Trypsin-BPTI. (b) α-Chymotrypsin-ovomucoid. (c) Subtilisin-chymotrypsin inhibitor.

α-Chymotrypsin-ovomucoid. The improvement in conformation brought about by refinement is less clear than in trypsin-BPTI. This is largely because the side-chain conformations of the interface residues do not change much on binding (with the exception of Thr17, Arg21 and Asn36) and improvements in the positioning of residues is somewhat obscured by a 1 Å shift in the C^{α}-backbone relative to the crystal following refinement.

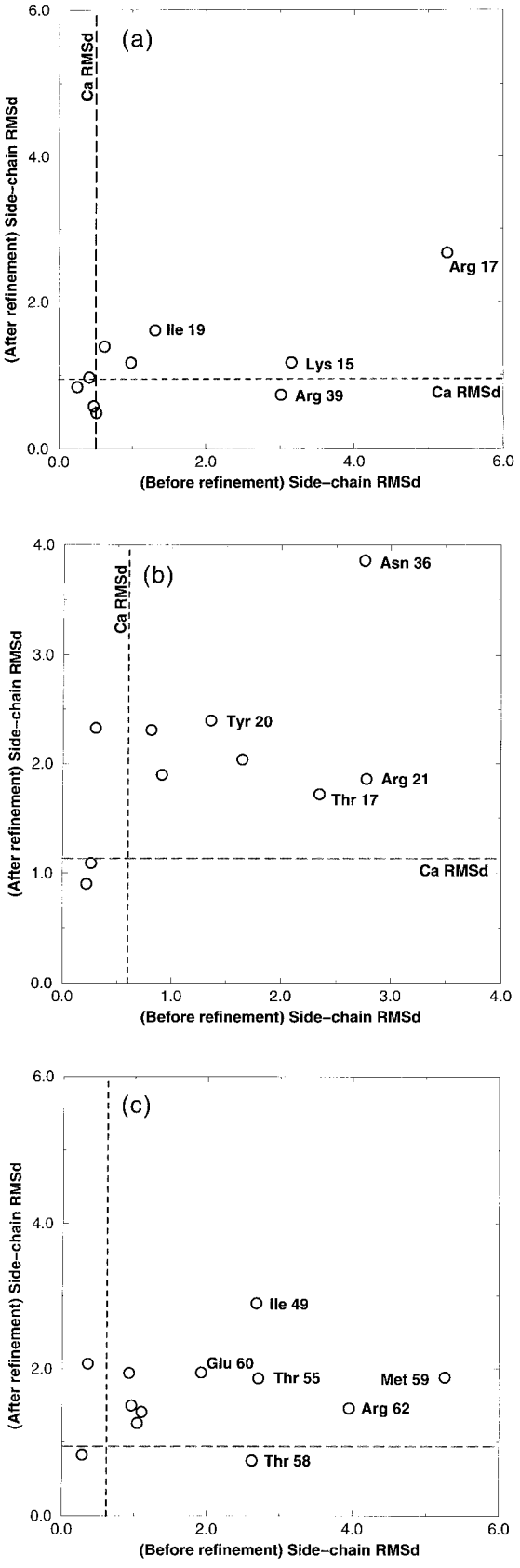

Figure 4. Inhibitor side-chain RMS deviations from the experimental crystal co-ordinates before and after refinement. The average C^{α} atom RMSd of the residues present is shown as a broken line for comparison. (a) Trypsin-BPTI. (b) α-Chymotrypsin-ovomucoid. (c) Subtilisin-chymotrypsin inhibitor.

The structural comparison of the key binding residue Leu/Met18 cannot be made because it is a different species of the protein in the bound and unbound structures. However, Met18 has the same χ1 and χ2 angles as Leu18 in the bound conformation following refinement. Rearrangement of Arg21 on refinement allows it to interact more favourably with Asp64 and Asp35 of the enzyme in agreement with the crystal structure of the complex. Following refinement there is high correlation between interaction energies for interface residues of the bound and unbound structures.

Subtilisin-chymotrypsin inhibitor. The improvement in conformation following refinement is considerable. The inhibitor binding residues Thr55, Thr58, Met59 and Arg62 all undergoing conformational changes to adopt similar conformations to those in the crystal complex, this can be seen clearly in terms of the reduction in side-chain RMSd. Following refinement there is reasonable correlation between the predicted residue interaction energies for the bound and unbound structures.

These results show that if the unbound conformations of the proteins adopt a similar orientation to the crystal complex, then an improvement in structure (as measured by side-chain RMSd) relative to the unbound starting conformations can be achieved by refinement. Also, the energies of interaction for residues following refinement of the unbound structures appear to be highly correlated with those predicted following refinement of the experimental crystal complex.

Using the energies of refined docked protein complexes to rank solutions

One of the motivations for this work was to be able to use the refinement process to optimise protein-protein interfaces of rigid-body docked complexes. Thus we are able to test the hypothesis that the solvent-mediated interaction energy can be used to rank the solutions, and successfully distinguish ''true'' from ''false'' positives. A Fourier correlation docking algorithm, FTDOCK, using shape and electrostatic complementarity to generate putative complexes has been described by us (Gabb *et al.*, 1997). The following results are based on structures reported in Table 3 of Gabb *et al.* (1997), which were generated following a global search of binding space with local refinement (1.5 Å surface thickness) and loose filtering (i.e. using biochemical constraints to exclude solutions). Results from ''loose'' filtering represent the structures that fulfil the constraint of the contact of any ligand atom with any active site residue of the enzyme or residue of the antibody CDRs (information that is commonly available in most docking studies). This involves refinement of between 26 (subtilisin-chymotrypsin inhibitor) and 762 (HyHEL10-lysozyme) structures per system. Also, we analysed structures generated by Shoichet & Kuntz (1991) using the program DOCK. In

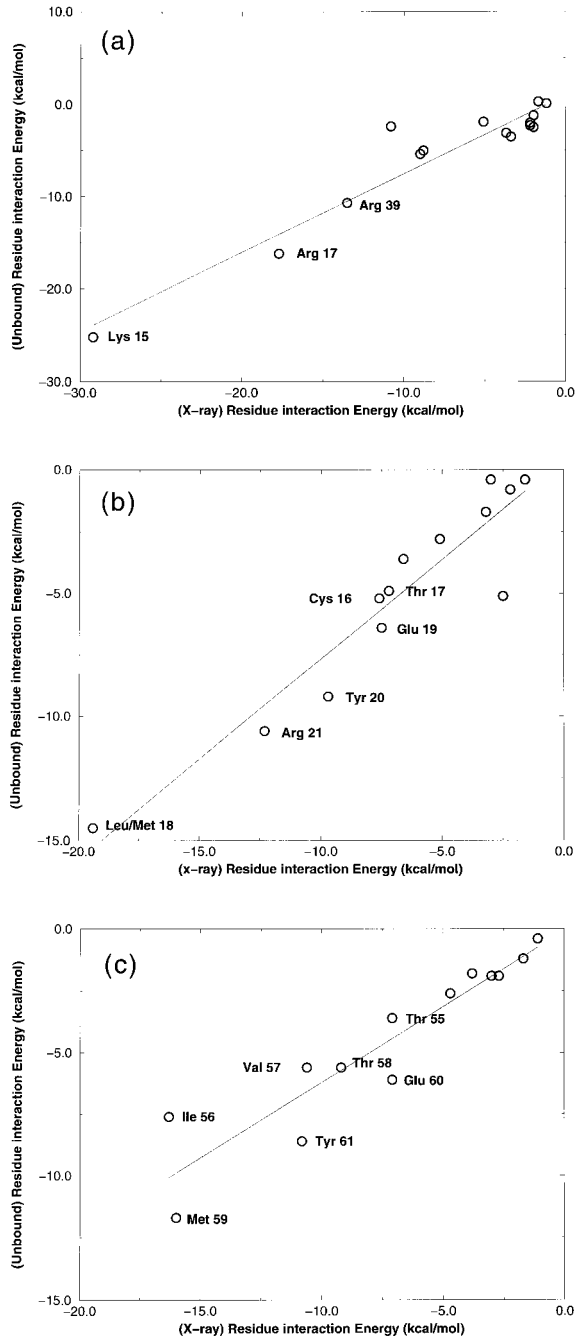

Figure 5. Energies of side-chain interactions for the inhibitor following refinement for both the unbound and bound conformations. (a) Trypsin-BPTI (correlation coefficient, R^2 0.96). (b) α-Chymotrypsin-ovomucoid (R^2 0.94). (c) Subtilisin-chymotrypsin inhibitor (R^2 0.94).

addition, for all systems, we carried out refinement on both (i) the crystal complex (with the bound conformations of the subunits) and (ii) a ''best-fit'' solution generated using the unbound structures optimally superimposed (in terms of their C^{α}-RMSd) on the bound crystal co-ordinates. The latter represents the best possible solution that could be achieved by docking of the unbound structures (in terms of the C^{α}-RMS fit).

The results of the refinement *in vacuo* (with no solvent present) are given in Table 3 for the five protease-protein inhibitor systems (for which both bound and unbound conformations of the subunits are available), and in Table 4 for the four antibody-antigen systems. For the antibody-antigen structures HyHEL5 and HyHEL10 only the unbound form of lysozyme was available, so bound forms of the antibody FABs had to be used in docking. The results of refinement with the solvent model included are given in Table 5 for the five protease-protein inhibitor systems and Table 6 for two of the antibody-antigen systems, HyHEL5 and HyHEL10 (since the *in vacuo* results with D1.3 and D44.1 were poor, these systems did not warrant further investigation with the solvent-mediated energy function). It should be noted that in the solvated systems, since we calculate ΔG^{solv} for the complexed forms only, desolvation energies cannot be compared between the crystal complex and the unbound structures or between those of two unbound structures from different systems (e.g. the FTDOCK (1sup/2ci2) and DOCK (1sbc/2ci2) forms for subtilisin-chymotrypsin inhibitor). The reason is that the solvation energy of the isolated subunits is different for different systems due to differing numbers of residues and differing conformations of side-chains. Both factors could potentially give rise to large differences in ΔG^{solv} that have nothing to do with binding. The results in Tables 3 to 6 are given in terms of (i) the C^{α} atom RMSd from the crystal complex, (ii) the energies of interaction between the enzyme and inhibitor in Tables 3 and 4 and the solvent-mediated interaction energies plus the relative desolvation energies ($\Delta\Delta G^{solv}$) of binding in Tables 5 and 6, and (iii) the ranking of the solutions in terms of energy with respect to all the other docked solutions analysed. The energies and rankings in bold are those of the FTDOCK generated structures. A close solution is defined as any with a C^{α} RMSd for interface residues from the crystal complex of <2.5 Å (Gabb *et al.*, 1997). The two close solutions of highest rank and any other close solution ranking in the top ten are shown. The rankings in italics are those of the crystal complex, best-fit or DOCK structures when that one structure is added to the FTDOCK data set (to allow comparison only) i.e. where would that structure rank were it to be a solution generated by FTDOCK? Only the top six close DOCK solutions are included for comparison.

The dependence of the solvation energy on the representation of the molecule embedded in a solvent grid can be assessed by the several different runs in which the molecule undergoes rotation about the geometric centre. Each of the three best-fit unbound conformations (analysed above) were rotated $\sim 10^{\circ}$ about the x, y, z, $x+y$, $x+z$ and $y+z$ axes. The error in the solvation energy calculations is given by the standard deviation (for seven runs). Trypsin-BPTI ($\Delta G^{solv} - 459.2(\pm 3.4)$ kcal/mol), α-chymotrypsin-

ovomucoid (ΔG^{solv} – 519.9(±1.2) kcal/mol), subtilisin-chymotrypsin inhibitor (ΔG^{solv} – 492.2(±3.9) kcal/mol). The grid dependence is comparable to that obtained with the Poisson-Boltzmann method (sd 0.5 to 2.7 kcal/mol at 1.75 grids/Å: Jackson & Sternberg, 1995) and within acceptable bounds.

One important feature of the *in vacuo* results is that in all systems except one (HyHEL10) the refined crystal complex has a more favourable interaction energy than any of the docked solutions analysed. This is a strong indication that the energy function is of predictive value in assessing the structural validity of complexes. The energy gap between the bound and unbound structures suggests that a greater level of atomic complementarity is achievable than that found by the conformational refinement of the unbound conformations.

Protease- protein inhibitor complexes. In three of the five complexes (kallikrein-BPTI, chymotrypsin-ovomucoid, chymotrypsin-HPTI) a close solution

Table 3. Calculated potential of mean force for complex formation *in vacuo* for protease- protein inhibitor systems

System	Structure[a]	C^{α}-RMSd[b]	E^{int}	Rank[c]
Trypsin-BPTI	Crystal complex (2ptc[d])	0.00	−112.8	*1/230*
	Best-fit solution (2ptn[e]/4pti[f])	0.59	−81.3	*2/230*
	(1) DOCK (2ptn/4pti)	0.62	−83.5	*2/230*
	(2) DOCK	0.96	−79.4	*3/230*
	(3) DOCK	0.63	−71.1	*3/230*
	(1) FTDOCK (2ptn/4pti)	**1.62**	**−53.9**	**47/229**
	DOCK best false-positive	6.64	−72.0	*3/230*
	FTDOCK best false-positive	**7.26**	**−85.3**	**1/229**
α-Chymotrypsin-ovomucoid	Crystal complex (1cho[g])	0.00	−96.2	*1/87*
	Best-fit solution (5cha[h]/2ovo[i])	0.65	−71.8	*3/87*
	(1) DOCK (5cha/2ovo)	0.88	−79.7	*2/87*
	(2) DOCK	0.83	−78.7	*2/87*
	(3) DOCK	0.99	−75.9	*3/87*
	(4) DOCK	0.71	−71.7	*4/87*
	(5) DOCK	0.71	−71.0	*4/87*
	(6) DOCK	0.79	−64.5	*4/87*
	(1) FTDOCK (5cha/2ovo)	**1.77**	**−80.6**	**1/86**
	(2) FTDOCK	**2.02**	**−77.1**	**2/86**
	(3) FTDOCK	**1.28**	**−71.8**	**3/86**
	(4) FTDOCK	**1.36**	**−64.4**	**4/86**
	DOCK best false-positive	3.42	−53.4	*19/87*
	FTDOCK best false-positive	**5.33**	**−63.9**	**5/86**
Subtilisin-chymotrypsin Inhib	Crystal complex (2sni[i])	0.00	−96.7	*1/27*
	Best-fit solution (1sbc[k]/2ci2[l])	0.53	−60.9	*4/27*
	Best-fit solution (1sup/2ci2)	0.30	−64.4	*2/27*
	(1) DOCK (1sbc/2ci2)	0.59	−68.5	*2/27*
	(2) DOCK	0.58	−67.4	*2/27*
	(3) DOCK	0.84	−65.4	*2/27*
	(4) DOCK	0.60	−64.2	*2/27*
	(5) DOCK	0.68	−64.1	*2/27*
	(6) DOCK	0.71	−57.4	*6/27*
	(1) FTDOCK (1sup/2ci2)	**2.59**	**−45.1**	**12/26**
	(2) FTDOCK	**1.85**	**−38.1**	**16/26**
	DOCK best false-positive	3.57	−47.9	*10/27*
	FTDOCK best false-positive	**7.47**	**−72.7**	**1/26**
Kallikrein-BPTI	Crystal complex (2kai[m])	0.00	−106.7	*1/365*
	Best-fit solution (2pka[n]/1bpi[o])	0.64	−68.3	*51/365*
	(1) FTDOCK (2pka/1bpi)	**1.28**	**−97.0**	**2/364**
	(2) FTDOCK	**2.10**	**−66.6**	**66/364**
	FTDOCK best false-positive	**6.35**	**−97.8**	**1/364**
α-Chymotrypsinogen-HPTI	Crystal complex (1cgi[p])	0.00	−113.1	*1/95*
	Best-fit solution (1chg[q]/1hpt[r])	1.28	−61.4	*12/95*
	(1) FTDOCK (1chg/1hpt)	**2.19**	**−71.9**	**2/94**
	FTDOCK best false-positive	**6.63**	**−82.8**	**1/94**

All energies are in kcal/mol.

[a] Structures include: (1) the experimental crystal complex; (2) a best-fit solution (involving optimal superposition of C^{α}-co-ordinates of the unbound subunits on the crystal complex); (3) structures generated by the program DOCK (Schoichet & Kuntz, 1991); (4) structures (results in bold) generated using the program FTDOCK (Gabb *et al.*, 1997). See the text for further details.

[b] The all C^{α} atom RMSd of the complex from the experimental crystal complex.

[c] Rank in bold is for the FTDOCK-generated data set. Rank in italics is that of any other generated geometry given it was a solution generated by FTDOCK. See the text for further details.

[d] Marquart *et al.* (1983); [e]Walter *et al.* (1982); [f]Marquart *et al.* (1983); [g]Fujinaga *et al.* (1987); [h]Blevins & Tulinsky (1985); [i]Bode *et al.* (1985); [j]McPhalen & James (1988); [k]Neidhart & Petsko (1988); [l]McPhalen & James (1987); [m]Chen & Bode (1983); [n]Bode *et al.* (1983); [o]Parkin *et al.* (1996); [p]Hecht *et al.* (1991); [q]Freer *et al.* (1970); [r]Hecht *et al.* (1992).

Table 4. Calculated potential of mean force for complex formation *in vacuo* for antibody-antigen systems

System	Structure	C^{α}-RMSd	E^{int}	Rank
D44.1- lysozyme	Crystal complex (1mlc[a])	0.00	−101.8	*1/591*
	Best-fit solution (1mlb[a]/1lza[b])	0.60	−43.8	*185/591*
	(1) FTDOCK (1mlb/1lza)	**3.2**	**−48.2**	**110/590**
	(2) FTDOCK	**3.0**	**−44.6**	**168/590**
	FTDOCK best false-positive	**13.0**	**−72.2**	**1/590**
D1.3-lysozyme	Crystal complex (1fdl[c])	0.00	−98.2	*1/708*
	Best-fit solution (1vfa[d]/1lza[b])	0.46	−50.0	*238/708*
	(1) FTDOCK (1vfa/1lza)	**2.9**	**−53.1**	**178/707**
	(2) FTDOCK	**2.1**	**−49.7**	**243/707**
	FTDOCK best false-positive	**14.7**	**−82.3**	**1/707**
HyHEL5-lysozyme	Crystal complex (2hfl[e])	0.00	−96.8	*1/521*
	Best-fit solution (2hfl/1lza[b])	0.30	−79.6	*14/521*
	(1) FTDOCK (2hfl/1lza)	**1.8**	**−69.6**	**29/520**
	(2) FTDOCK	**1.9**	**−66.0**	**43/520**
	FTDOCK best false-positive	**10.4**	**−89.3**	**1/520**
HyHEL10-lysozyme	Crystal complex (3hfm[f])	0.00	−83.3	*11/766*
	Best-fit solution (3hfm/1lza[b])	0.34	−83.9	*10/766*
	(1) FTDOCK (3hfm/1lza)	**1.1**	**−84.1**	**9/765**
	(2) FTDOCK	**2.5**	**−81.5**	**11/765**
	FTDOCK best false-positive	**17.4**	**−105.1**	**1/765**

All energies are in kcal/mol.
[a] Braden *et al.* (1994); [b]Maenaka *et al.* (1995); [c]Fischmann *et al.* (1991); [d]Bhat *et al.* (1990); [e]Sheriff *et al.* (1987); [f]Padlan *et al.* (1989).

ranks in the top four answers generated by the FTDOCK, with the *in vacuo* energy function. In four of the five complexes, a close solution generated by FTDOCK ranks in the top four answers if solvent is included in the energy function. In all the examples, a close solution ranks in the top four answers if structures generated by both the FTDOCK algorithm and DOCK (Shoichet & Kuntz,

Table 5. Calculated potential of mean force for complex formation in solution for protease-protein inhibitor systems

System	Structure	C^{α}-RMSd	$E^{int} + \Delta\Delta G^{solv}$	Rank (solvated)
Trypsin-BPTI	Best-fit solution (2ptn/4pti)	0.59	−63.5	*1/230*
	(1) DOCK (2ptn/4pti)	0.62	−67.5	*1/230*
	(2) DOCK	0.96	−60.5	*1/230*
	(3) DOCK	0.63	−60.4	*1/230*
	(1) FTDOCK (2ptn/4pti)	**1.62**	**−48.9**	**11/229**
	DOCK best false-positive	6.64	−50.6	*6/230*
	FTDOCK best false-positive	**8.61**	**−59.3**	**1/229**
α-Chymotrypsin ovomucoid	Best-fit solution (5cha/2ovo)	0.65	−56.3	*1/87*
	(1) DOCK (5cha/2ovo)	0.83	−60.2	*1/87*
	(2) DOCK	0.99	−59.4	*1/87*
	(3) DOCK	0.71	−56.3	*1/87*
	(4) DOCK	0.88	−54.9	*2/87*
	(5) DOCK	0.71	−52.5	*2/87*
	(6) DOCK	0.79	−47.4	*3/87*
	(1) FTDOCK (5cha/2ovo)	**1.28**	**−56.1**	**1/86**
	(2) FTDOCK	**1.77**	**−50.4**	**2/86**
	(3) FTDOCK	**1.36**	**−46.2**	**3/86**
	(4) FTDOCK	**2.02**	**−41.5**	**5/86**
	DOCK best false-positive	7.04	−38.8	*6/87*
	FTDOCK best false-positive	**7.59**	**−45.0**	**4/86**
Subtilisin-chymotrypsin Inhib	Best-fit solution (1sup/2ci2)	0.30	−39.0	*1/27*
	(1) FTDOCK (1sup/2ci2)	**2.59**	**−29.3**	**4/26**
	(2) FTDOCK	**1.85**	**−17.8**	**14/26**
	FTDOCK best false-positive	**7.75**	**−38.3**	**1/26**
Kallikrein-BPTI	Best-fit solution (2pka/1bpi)	0.64	−34.6	*69/365*
	(1) FTDOCK (2pka/1bpi)	**1.28**	**−57.0**	**2/364**
	(2) FTDOCK	**2.10**	**−44.3**	**31/364**
	FTDOCK best false-positive	**6.77**	**−63.8**	**1/364**
α-Chymotrypsinogen HPTI	Best-fit solution (1chg/1hpt)	1.28	−32.5	*21/95*
	(1) FTDOCK (1chg/1hpt)	**2.19**	**−42.7**	**3/94**
	FTDOCK best false-positive	**6.63**	**−51.2**	**1/94**

All energies are in kcal/mol.

Table 6. Calculated potential of mean force for complex formation in solution for antibody-antigen systems

System	Structure	C^{α}-RMSd	E^{int}	Rank
HyHEL5-lysozyme	Best-fit solution (2hfl/1lza)	0.30	−65.0	*12/51*
	(1) FTDOCK (2hfl/1lza)	**1.8**	**−46.0**	**39/50**
	(2) FTDOCK	**1.9**	**−26.4**	**49/50**
	FTDOCK best false-positive	**10.8**	**−81.2**	**1/50**
HyHEL10-lysozyme	Best-fit solution (3hfm/1lza)	0.34	−63.4	*6/51*
	(1) FTDOCK (3hfm/1lza)	**2.5**	**−66.3**	**4/50**
	(2) FTDOCK	**1.1**	**−61.0**	**6/50**
	FTDOCK best false-positive	**12.4**	**−69.9**	**1/50**

All energies are in kcal/mol.

1991) are considered. Thus, starting from unbound conformations and using the energy refinement procedure developed here it is possible to find highly ranked ''true'' positives from amongst the large number of ''false'' positives.

Success is dependent on whether the critical residue interactions can be reproduced on refinement. Failure is because some close docked structures do not allow the formation of these interactions during refinement; however, this is not a direct function of the RMS-fit of the model to the crystal (e.g. see the results of kallikrein-BPTI or chymotrypsin-HPTI below). Indeed, the rugged nature of the free energy landscape and/or the limitations of our approach are sufficient to obscure some of the close docked solutions following refinement, implying that molecular recognition occurs at a very specific level.

The inclusion of solvent in the energy function is advantageous. In the subtilisin-chymotrypsin inhibitor system, the highest ranked close solution goes from 12th *in vacuo* to fourth in the presence of solvent, and in trypsin-BPTI the top ranked close solution goes from 47th to 11th. The other three systems retain highly ranked close solutions in the presence of solvent. Furthermore, certain highly ranked false positives that have very favourable interaction energies *in vacuo* have large desolvation penalties on binding and therefore rank poorly in the presence of solvent (see the discussion of false positives below).

It is useful to analyse the success and failure of the refinement procedure as it applies to each of the protease-protein inhibitor systems in terms of structural features and the possible limitations of the side-chain rotamer representation and fixed backbone conformation.

α-Chymotrypsin-ovomucoid. There is little conformational change in the main chain on binding, or in the interface side-chains, with the exception of Met I18 and Arg I21 side-chains (see Figure 3(b)). The replacement of Leu I18 in the crystal complex (1cho) by Met I18 in the unbound forms (5cha/2ovo) presents no steric problem in the best-fit solution. Following refinement, the methionine side-chain dihedrals $\chi 1$ and $\chi 2$ adopt those of leucine, with Met having the single largest favourable contribution to binding (which like Leu represents ~20% of the favourable interaction energy). Following refinement of the FTDOCK results, the top four ranked solutions are all close solutions. This system represents the most convincing case for combining the refinement procedure with the rigid-body docking procedure.

Subtilisin-chymotrypsin inhibitor. There is only a small degree of conformational change in the main chain on binding. However, four interface side-chains of the inhibitor (Thr I55, Thr I58, Met I59 and Arg I62) that are important for binding undergo conformational change on binding (see Figure 3(c)). These changes are predicted well in the best-fit model by the side-chain refinement. However, in spite of a (theoretical) high ranking for the best-fit solution, the two close solutions generated by FTDOCK score comparatively poorly in this example (other FTDOCK runs did however generate solutions that would have ranked first; unpublished results). This shows the sensitivity of the results to relatively small changes in the interface.

Trypsin-BPTI and kallikrein-BPTI. Again in both these systems there is little main-chain conformational change on binding (see Figure 3(a)), with conformational change being largely confined to the important inhibitor binding residues Lys I15 Arg I17 and Arg I39. The ability of the Lys I15 to bind into the deep substrate-specificity pocket of trypsin or kallikrein is paramount (contributing 26% and 30% of the favourable interaction energy, respectively). However the restricted access and depth of the pocket, mean that the conformation of the side-chain is highly constrained. Indeed in terms of close FTDOCK conformations (19 for KAI and eight for PTC) only one kallikrein-BPTI solution (which has the lowest RMSd) can productively form this interaction. Translational or rotational shifts at the interface or the limited number of lysine rotamer conformations preclude the other solutions from forming an ion-pair interaction with Asp E189 at the bottom of the pocket. Disappointingly, the best-fit kallikrein-BPTI structure cannot form this interaction. This would appear to be a result of the limited rotamer choice

for lysine disallowing this interaction on steric grounds.

α-Chymotrypsinogen-HPTI. This is the only enzyme-inhibitor system in which the backbone conformation of the inhibitor binding loop (involving residues Glu I12 to Cys I16) undergoes significant conformational change between the bound and unbound states. One result of this is that the best-fit solution has a somewhat different binding interface from the crystal complex. In this best-fit solution Tyr I18 cannot bind effectively in the enzyme specificity pocket in spite of being able to make some other residue-residue interactions seen in the complex. In consequence, this solution has a fairly poor interaction energy. In the refined crystal complex Tyr I18 contributes by far the single largest favourable residue contribution (27%) to binding. The top ranked FTDOCK close solution can however bind Tyr I18 in the specificity pocket (making the single largest favourable residue contribution, 18%, to binding). However, in order to do this there are significant differences in the residue environments of the binding loop (Tyr I10-Arg I21). Only residues Gly I15-Tyr I18 can be considered to be equivalent to those of the crystal complex in spatial and conformational terms.

"False" positive solutions. We have not carried out an extensive analysis of the high-scoring ''false'' positive solutions generated by FTDOCK. However, it is interesting to focus on the trypsin-BPTI system simply due to the nature of the results and because high-scoring false positive docked solutions have been commented on before for this system (Shoichet & Kuntz, 1991; Bacon & Moult, 1991; Cherfils & Janin, 1993). As might be expected, any arginine or lysine residue that can bind in the trypsin specificity pocket will have a favourable interaction energy. Indeed, in agreement with the studies above, false positive solutions involving Lys I15 (with the molecule binding in a different orientation) or Lys I26 binding in the specificity pocket are ranked highly. However, two different solutions (the *in vacuo* DOCK and FTDOCK best false positive solutions, given in Table 3) that rank very highly, show how false positive solutions can maximise the available interaction energy. As can be seen from Figure 5(c), in addition to Lys I15 (which contributes 26% of the favourable interaction energy), Arg I17 and Arg I39 make the second and third most favourable interactions (contributing 16% and 12% of the favourable interaction energy, respectively). The two high-ranking false positives make use of these two sites in addition to the specificity pocket (see Figure 6(a) and (b)) using either long-chain arginine or lysine residues. Thus, they harness much of the interaction energy available to the native state. However, with the inclusion of solvent in the energy function both solutions (in particular the FTDOCK best false positive) are ranked much lower due to large desolvation energy penalties associated with binding.

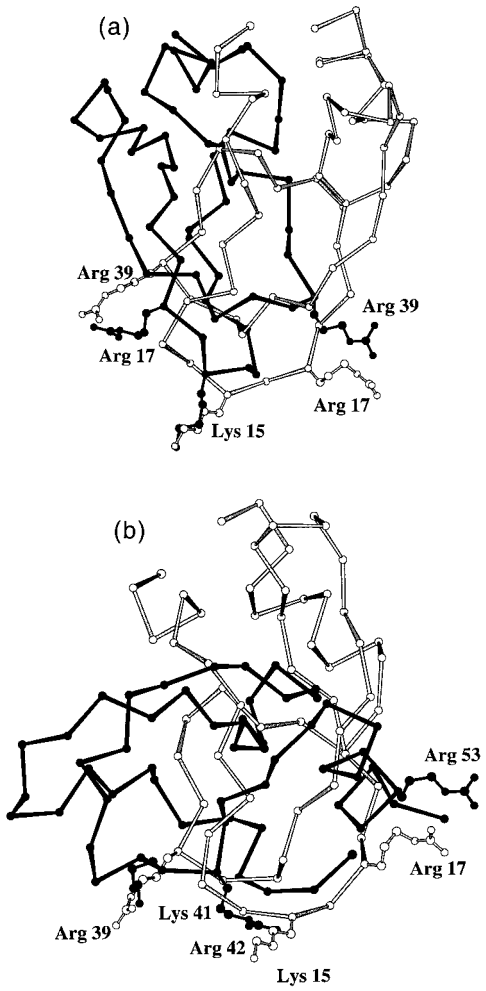

Figure 6. Comparison of the experimental crystal (in white) and the two top-ranking false positives (in black) following *in vacuo* refinement of docked structures of trypsin-BPTI generated by (a) DOCK (Schoichet & Kuntz (1991) and (b) FTDOCK (Gabb *et al*., 1997). The trypsin C^{α} co-ordinates have been optimally superimposed. Only the BPTI C^{α} trace and important Arg/Lys interface binding residues (see the text) are shown.

Antibody-antigen complexes. In only one of the four FAB-lysozyme systems does a close solution rank in the top ten answers generated by the FTDOCK algorithm. This rises to two out of four if we consider any close solution in the top 30 answers. However, it should be noted that in the two most successful systems HyHEL5 and HyHEL10 bound conformations of the antibody (and an unbound conformation of lysozyme) had to be used in the docking study. In the D44.1-lysozyme and D1.3-lysozyme systems, both the lysozyme and antibody structures used in docking were unbound crystal structures. This may be the reason why HyHEL5 and HyHEL10 perform much

better than D44.1 and D1.3. Certainly, whilst both the best-fit solutions for HyHEL5 and HyHEL10 rank highly, those of D44.1 and D1.3 perform the worse of the systems studied. It would appear that the main-chain conformational changes that occur on binding in these two systems mean that any methodology that does not treat main-chain flexibility is likely to fail. The advantages of including solvation in the energy function are more difficult to judge here than in the enzyme-inhibitor systems, since only two of the systems give close solutions in the top 50 ranked *in vacuo* solutions. There is improvement in the best-ranking close solution in HyHEL10 from ninth to fourth, however, the results for HyHEL5 get worse. Again it is useful to further analyse each system in turn.

FAB D44.1-lysozyme. Rigid-body docking is a poor approximation in this case. In particular, a conformational change on binding in the main-chain atoms of lysozyme of a loop involving Pro70 and Gly71 (of magnitude 2.1 Å and 2.4 Å, respectively, for C^{α} atoms following rigid-body superposition of the two structures) cause steric conflict when the unbound structures adopt the best-fit conformation. This clash cannot be alleviated with a fixed backbone without compromising the observed interactions at the interface. Furthermore, the side-chain of lysozyme Arg45 cannot adopt the conformation seen in the crystal whereby it forms an interaction with Glu H50 (presumably due to the fact this arginine rotamer is not present in the rotamer library). This causes the local rearrangement of the conformations of Trp L94 and Tyr H59, with a significant loss in binding energy. The best-fit unbound conformation would have ranked 185th in the *in vacuo* energy ranking for 590 structures processed. There are only three close RMSd solutions found from docking, these rank 110th, 168th and 348th.

FAB D1.3-lysozyme. A local rearrangement of the hypervariable loop H3 on binding involving Asp H100, Tyr H101 and Arg H102 (of magnitude ~1 Å for C^{α} atoms following rigid-body superposition of the two structures) places the loop in a somewhat different conformation in the bound crystal complex, which allows the dominant interactions of Asp H100 and Tyr H101 to form with the antigen. This is associated with a difference in interaction energy of ~20 kcal/mol when comparing the energy refined unbound and bound conformations. In addition, several other residue interactions are predicted to be weaker in the unbound complex than in the crystal complex, leading to the observed difference in binding energy. In particular Gly H31 and Tyr H32 (of H1) Asp H54 (of H2) Tyr L32 (of L1) Tyr49 and Tyr L50 (of L2) Trp L92 and Ser L93 (of L3).

In addition to the hypervariable loop H3, several of these residues have been identified as undergoing systematic displacement on binding (Bhat *et al.*, 1990), giving rise to an induced-fit mechanism of antigen binding. The best-fit unbound conformation would have ranked 238th *in vacuo* for 707 structures processed. There are only four close RMSd solutions found from docking, these rank 178th, 243rd, 272nd and 387th.

FAB HyHEL5-lysozyme. The inability of Arg45 and Arg68 of lysozyme to reproduce the native side-chain conformations on refinement of the unbound complex causes steric conflicts at the antibody-antigen interface that result in the loss of binding energy. Conflict between Arg45 and Trp L90 results in rearrangement of the Trp ring, and there are unresolvable vdW conflicts between Arg68 and Trp H33. These residue interactions alone account for a loss of ~15 kcal/mol in binding energy. In spite of this the Arg residues undergo considerable change in conformation on refinement and largely mimic (within the constraints imposed by the rotamer library) the interaction seen in the bound complex. The best-fit unbound conformation would have ranked 14th in the *in vacuo* energy ranking of the 520 structures processed. There are the only two close RMSd solutions found from docking and they rank 29th and 43rd. Inclusion of solvent in the energy function brings about a marginal improvement in rank of the best-fit structure following refiment of the top 50 *in vacuo* structures. However, the two close RMSd solutions rank more poorly.

FAB HyHEL10-lysozyme. This case is the only antigen-antibody complex that is not affected by the limitations imposed by a rigid backbone and rotamer library description of side-chains. In this case, three important long-chain binding residues of lysozyme, Arg21, Lys96 and Lys97 can all interact favourably with the antibody in the unbound complex. Arg21 and Lys97 do so through different rotamers from the bound complex (due to slightly differing main-chain conformations) but place the guanidinium and amine functional groups in the same locations as the bound complex. Only Arg21 undergoes a major change in side-chain conformation on binding. The best-fit unbound conformation would have ranked tenth in the *in vacuo* energy ranking of the 765 structures processed. There are only three close RMSd solutions found from docking they rank 9th, 11th and 27th. Interestingly this is the only system where false positive dockings have a better *in vacuo* binding energy than the refined crystal structure. Inclusion of solvation further enhances the ranking of all these solutions following refinement of the top 50 *in vacuo* structures.

Discussion

Here, we address the problem of whether the protein-protein interaction energy plus the differences in solvation energy can be used to dis-

Table 7. Ranking of docked structures by FTDOCK and by potential of mean force following refinement *in vacuo* and in solution

System	Total no. structures	$N \leqslant 2.5$ Å[a]	Rank FTDOCK (local refinement 1.5 Å)	Rank E^{int}	Rank $E^{int} + \Delta\Delta G^{solv}$
Trypsin-BPTI	229	8	16	47	11
α-Chymotrypsin ovomucoid	86	5	11	1	1
Subtilisin-chymotrypsin Inhib.	26	2	8	12	4
Kallikrein-BPTI	364	18	130	2	2
α-Chymotrypsinogen-HPTI	94	1	3	2	3
D44.1-lysozyme	590	4	41	110	ND
D1.3-lysozyme	707	2	176	178	ND
HyHEL5-lysozyme	519	2	228	29	39[b]
HyHEL10-lysozyme	762	6	65	9	4[b]

[a] Number of structures with C^{α} RMSd for interface residues from crystal complex $\leqslant 2.5$ Å.
[b] Results for the top 50 structures of the *in vacuo* refinement.

tinguish between true and false positives generated by rigid-body docking.

The ranking of the highest scoring close solution generated by FTDOCK are given for all the systems in Table 7 ranked according to the scoring function of (1) FTDOCK shape and electrostatic complementarity (Gabb *et al.*, 1997) (2) the *in vacuo* interaction energy (E^{int}) and (3) the solvent-mediated interaction and solvation energies ($E^{int} + \Delta\Delta G^{solv}$). Clearly, refinement of the protein-protein interface leads to a higher rank for the top close solution in all but two cases (D1.3 and D44.1). Inclusion of solvent in the refinement process preserves high-ranking close solutions or improves the ranking in all but one case (HyHEL5). In several of the systems there is an order of magnitude improvement in the ranking of a close solution over FTDOCK rankings; furthermore, where the method is successful a close solution is likely to be found in the top four structures. In summary, this indicates that where the results are not limited by conformational sampling, the energy function developed here is a good measure of protein-protein complementarity.

We carried out detailed analysis on three protease-protein inhibitor complexes, trypsin-BPTI, α-chymotrypsin-ovomucoid and subtilisin-chymotrypsin inhibitor. These three complexes have been analysed in previous studies that attempted to discriminate between docked complexes (Shoichet & Kuntz, 1991; Jackson & Sternberg, 1995; Weng *et al.*, 1996). The results show that starting from best-fit unbound docked conformations of the subunits both the structural (as defined by improvement in side-chain RMSd) and energetic criteria (as defined by the high correlation for residue contributions to binding for the unbound and bound states) show that the refinement procedure reproduces the crystallographically determined binding mode and its energetics. The best-fit solutions rank in the top four structures against a background of between 26 and 229 other structures. This strongly suggests that if the important interactions involved in the recognition process can be formed in the refinement stage then there is a high probability that a true positive structure will be ranked very highly.

As in a previous study (Jackson & Sternberg, 1995) using a continuum model (Jackson & Sternberg, 1994) for protein-protein interactions, the energy function developed here ranks true positives above the best false positive for the docked complexes generated by the study reported by Shoichet & Kuntz (1991) using the program DOCK. However, the present study represents a more stringent test of the energy function due to the size of the FTDOCK-generated data sets, and the greater number of systems analysed. Furthermore, conformational refinement at the interface can determine the important residue interactions in binding, hence, the present method represents a further improvement of modelling protein interactions given its inclusion of conformational flexibility as well as its treatment of solvation.

In terms of conformational sampling, solutions need to be generated at the docking stage that are sufficiently close to the conformation of the complex to allow the key interactions to form during refinement. Conformational change of the backbone on binding (e.g. the best-fit solution for α-chymotrypsinogen-HPTI) or the inability of the side-chain rotamer library to model the conformations of charged long-chain residues (e.g. the best-fit solution for kallikrein-BPTI) can still result in failure. However, highly ranked close solutions were found for both the above examples in spite of the poor results for the best-fit models.

It would appear that for the highly specific protease-inhibitor systems success is usually achieved, which is generally a reflection of the rigid-body nature (at least with regard to conformational changes in the backbone) of the subunit interactions. In the case of these systems, solvent would appear to further enhance the rankings of close solutions. However, inclusion of solvent does not turn close solutions with poor *in vacuo* energies into high-ranking solutions. If specific favourable interactions are not formed on refinement then a particular solution is unlikely to be highly ranked, irrespective of the presence of solvent.

We have not carried out a detailed analysis of false positive solutions. However, it was found that a few alternative solutions could form equally

favourable interactions to the native best-fit structure. In the case of the trypsin-BPTI system, the inhibitor can bind in more than one conformation in which Lys or Arg residues bind in three negatively charged surface pockets of the enzyme active site. This type of binding is similar to that employed by the inhibitor, in that it maximises the interaction with the enzyme. However, these alternative binding modes are discriminated against if solvation is included in the scoring function, emphasising the importance of solvation for excluding incorrect geometries.

In contrast to the protease-protein inhibitor structures, the antibody-lysozyme results are less encouraging. A combination of a rigid backbone representation and limited rotamer coverage are likely to be the cause of failure in these systems. Clearly, the non-rotameric tendencies of arginine, lysine and glutamic acid are of concern (Schrauber *et al*., 1993), since these residues in particular often play a key role in binding. Certainly the inability of arginine rotamer conformations to mimic the interactions seen in the crystal structures are a major cause of the weaker interaction energies seen in the best-fit unbound complexes of FAB HyHEL5 and D44.1. It would be advantageous to increase the coverage of conformational space for these non-rotameric residues in view of their importance in binding. Recently, Desmet *et al*. (1997) have described a more extensive backbone dependent rotamer library that allows greater coverage of conformational space. Alternatively, Monte Carlo refinement of interface side-chains could be performed (e.g. see Totrov & Abagyan, 1994), or a systematic conformational search (e.g. see Bruccoleri & Karplus, 1987). However, currently, such search methods are too computationally intensive for the type of study carried out here in which potentially hundreds of docked geometries are screened.

Other factors could also play a role in our inability to predict antibody-antigen interactions. Firstly, antibodies in general have binding constants in the nanomolar range as opposed to enzyme-inhibitor complexes that have binding constants in the pico and femtomolar range. Therefore the degree of specificity is likely to be less with a concomitant rise in the background noise from false positives, making discrimination somewhat harder. Secondly, specific ion binding or proton uptake/release may play an important role in binding in certain systems. Thirdly, the general role of electrostatic and shape complementarity seems somewhat less well defined in the case of antibody-antigen interactions when defined by Fourier correlation theory.

The inclusion of side-chain flexibility and interface refinement in a general rigid-body docking methodology and our study of both protein-inhibitor and antibody-antigen interactions allow us to make some generalisations about the mechanism of protein-protein recognition in the two systems. In the protease-protein inhibitors there is essentially a ''lock and key'' binding mechanism where conformational change is largely limited to side-chains, consequently the screening method works well. In the antibody-antigen systems there is an induced-fit mechanism (Wilson & Stanfield, 1993) where more extensive rearrangement of the backbone conformation on binding (e.g. D1.3 and D44.1) can cause problems for the method. The energy function developed here can discriminate between near-native structures and false dockings if the near-native structures can reproduce the residue interactions seen in the crystal complex. The limitations of the method are mainly due to conformational sampling. Development of a method that adequately models structural plasticity and side-chain flexibility may lead to a more general solution to the protein docking problem.

Acknowledgements

We thank the following: Suhail Islam for providing several subroutines used in the rigid-body minimization routine; Baldomero Oliva for valuable discussions; the Lloyd's of London Tercentenary Foundation for support of R. M. J., and the European Molecular Biology Organisation for support of H. A. G. in the form of fellowships.

References

Bacon, D. J. & Moult, J. (1992). Docking by least-squares fitting of molecular surface patterns. *J. Mol. Biol.* **225**, 849–858.

Bhat, T. N., Bentley, G. A., Fischmann, T. O., Boulot, G. & Poljak, R. J. (1990). Small rearrangements in strucutres of Fv and Fab fragments of antibody D1. 3 on antigen binding. *Nature,* **347**, 483–485.

Blevins, R. A. & Tulinsky, A. (1985). The refinement and the structure of the dimer of α-chymotrypsin at 1. 67 Å resolution. *J. Biol. Chem.* **260**, 4264–4270.

Bode, W., Chen, Z., Bartels, K., Kutzbach, C., Schmidt-Kastner, G. & Bartunik, H. (1983). Refined 2 Å X-ray crystal structure of porcine pancreatic kallikrein a and a specific trypsion-like serine proteinase. Crystallization and structure determination and crystallographic refinement and structure and its comparison with bovine trypsin. *J. Mol. Biol.* **164**, 237–283.

Bode, W., Epp, O., Huber, R. & Laskowski, M. (1985). The crystal and molecular structure of the third domain of silver pheasant ovomucoid. *Eur. J. Biochem.* **147**, 387–395.

Braden, B. C., Souchon, H., Eisele, J.-L., Bentley, G. A., Bhat, T. N., Navaza, J. & Poljak, R. J. (1994). Three-dimensional structures of the free and the antigen-complexed Fab form monoclonal anti-lysozyme antibody D44. 1. *J. Mol. Biol.* **243**, 767–781.

Bruccoleri, R. E. & Karplus, M. (1987). Prediction of the folding of short peptide segments by uniform conformational sampling. *Biopolymers,* **26**, 137–168.

Chen, Z. & Bode, W. (1983). Refined 2. 5 Å X-ray crystal structure of the complex formed by porcine kallikrein a and the the crystallization, Patterson search, structure comparison with its components and with bovine trypsin-pancreatic trypsin inhibitor complex. *J. Mol. Biol.* **164**, 283–311.

Cherfils, J. & Janin, J. (1993). Protein docking algorithms: simulating molecular recognition. *Curr. Opin. Struct. Biol.* **3**, 265–269.

Cherfils, J., Duquerroy, S. & Janin, J. (1991). Protein-protein recognition analysed by docking simulation. *Proteins: Struct. Funct. Genet.* **11**, 271–280.

Cummings, M. D., Hart, T. N. & Read, R. J. (1995). Atomic solvation parameters in the analysis of protein-protein docking results. *Protein Sci.* **4**, 2087–2099.

Desmet, J., Wilson, I. A., Joniau, M., Demaeyer, M. & Lasters, I. (1997). Computation of the binding of fully flexible peptides to proteins with flexible side-chains. *FASEB. J.* **11**, 164–172.

Fischer, D., Lin, S. L., Wolfson, H. L. & Nussinov, R. (1995). A Geometry-based suite of molecular docking processes. *J. Mol. Biol.* **248**, 459–477.

Fischmann, T. O., Bentley, G. A., Bhat, T. N., Boulot, G., Mariuzza, R. A., Phillips, S. E. V., Tello, D. & Poljak, R. J. (1991). Crystallographic refinement of the three-dimensional structure of the D1. 3–lysozyme complex at 2. 5 Å resolution. *J. Biol. Chem,* **266**, 12915–12920.

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T. & Xuong, N. H. (1970). Chymotrypsinogen, 2. 5 Å crystal structure, comparison with alpha-chymotrypsin, and implications for zymogen activation. *Biochemistry*, **9**, 1997–2009.

Fujinaga, M., Sielecki, A. F., Read, R. J., Ardelt, W., Laskowski, M. J. & James, M. N. G. (1987). Refined structure of α-lytic protease at 1. 7 Å resolution. *J. Mol. Biol.* **195**, 397–418.

Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. (1997). Modelling protein docking using shape complementarity, electrostatics, and biochemical information. *J. Mol. Biol.* **272**, 106–120.

Gilson, M. K. & Honig, B. (1988). Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins: Struct. Funct. Genet.* **4**, 7–18.

Harrison, P. (1996). Analysis and prediction of protein structure: disulphide bridges. PhD dissertation, University of London.

Harvey, S. C. (1989). Treatment of electrostatic effects in macromolecular modelling. *Proteins: Struct. Funct. Genet.* **5**, 78–92.

Hecht, H. J., Szardenings, M., Collins, J. & Schomburg, D. (1991). Three-dimensional structure of the complex between bovine chymotrypsinogen a and two recombinant variants of human pancreatic secretory trypsin inhibitor (Kazal type). *J. Mol. Biol.* **220**, 711–722.

Hecht, H. J., Szardenings, M., Collins, J. & Schomburg, D. (1992). Three-dimensional structure of a recombinant variant of human pancreatic secretory trypsin inhibitor (Kazal type). *J. Mol. Biol.* **225**, 1095–1103.

Hill, T. L. (1956). *Statistical Mechanics,* McGraw Hill, New York.

Islam, S. A. (1986). Computer aided molecular modelling of DNA-drug interactions. PhD dissertation, University of London.

Jackson, R. M. & Sternberg, M. J. E. (1994). Application of scaled particle theory to model the hydrophobic effect: implications for molecular association and protein stability. *Protein Eng.* **7**, 371–383.

Jackson, R. M. & Sternberg, M. J. E. (1995). A continuum model for protein-protein interactions: application to the docking problem. *J. Mol. Biol.* **250**, 258–275.

Jiang, F. & Kim, S.-H. (1991). ''Soft docking'': matching of molecular surface cubes. *J. Mol. Biol.* **219**, 79–102.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C. & Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between proteins and ligands by correlation techniques. *Proc. Natl Acad. Sci. USA,* **89**, 2195–2199.

Koehl, P. & Delarue, M. (1994). Application of a self-consitent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249–275.

Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* **236**, 918–939.

Lee, F. S., Chu, Z. T. & Warshel, A. (1993). Microscopic and semimicroscopic calculations of electrostatic energies in proteins by POLARIS and ENZYMIX programs. *J. Comput. Chem.* **14**, 161–185.

Luzhkov, V. & Warshel, A. (1992). Microscopic models for quantum mechanical calculations of chemical processes in solution: LD/AMPAC and SCAAS/AMPAC calculations of solvation energies. *J. Comput. Chem.* **13**, 199–213.

Maenaka, K., Matsushima, M., Song, H., Sunada, F., Watanabe, K. & Kumagai, I. (1995). Dissection of protein-carbohydrate interactions in mutant hen egg-white lysozyme complexes and their hydrolytic activity. *J. Mol. Biol.* **247**, 281–293.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallog. sect. B.* **39**, 480–490.

McPhalen, C. A. & James, M. N. G. (1987). Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry,* **26**, 261–269.

McPhalen, C. A. & James, M. N. G. (1988). Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-c-subtilisin Carlsberg and CI-2-subtilisin Novo. *Biochemistry,* **27**, 6582–6598.

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interations, hydrogen bond interactions and intrisic torsion potentials for the naturally occuring amino acids. *J. Phys. Chem.* **79**, 2361–2366.

Neidhart, J. J. & Petsko, G. A. (1988). The refined crystal structure of subtilisin Carlsberg at 2. 5 Å resolution. *Protein Eng.* **2**, 271–276.

Padlan, E. A., Silverton, E. W., Sheriff, S., Cohen, G. H., Smith-Gill, S. J. & Davies, D. R. (1989). Structure of an antibody complex. Crystal strucutre of the HyHEL-10 Fab-lysozyme complex. *Proc. Natl Acad. Sci. USA.* **86**, 5938–5942.

Parkin, S., Rubb, B. & Hope, H. (1996). Structure of bovine pancreatic trypsin-inhibitor at 125k – definition of carboxyl terminal residues gly57 and ala58. *Acta. Crystallog. sect. D,* **52**, 18–29.

Richards, F. M. (1977). Areas, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Richmond, T. J. (1984). Solvent accessible surface area and excluded volume in proteins: analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.* **178**, 63–89.

Russell, S. T. & Warshel, A. (1985). Calculation of electrostatic energies in proteins: the energetics of ionized groups in bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* **185**, 389–404.

Schrauber, H., Eisenhaber, F. & Argos, P. (1993). Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol.* **230**, 592–612.

Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J. & Davies, D. R. (1987). Three-dimensional structure of an antibody-antigen complex. *Proc. Natl Acad. Sci. USA.* **84**, 8075–8079.

Shoichet, B. K. & Kuntz, I. D. (1991). Protein docking and complementarity. *J. Mol. Biol.* **221**, 327–346.

Sitkoff, D., Sharp, K. A. & Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978–1988.

Totrov, M. M. & Abagyan, R. A. (1994). Detailed ab initio prediction of lysozyme-antibody complex with 1. 6 Å accuracy. *Nature Struct. Biol.* **1**, 259–263.

Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). A new approach to the rapid determination of protein side-chain conformations. *J. Biomol. Struct. Dynam.* **8**, 1267–1289.

Walls, P. H. & Sternberg, M. J. E. (1992). New algorithm to model protein-protein recognition based on surface complementarity. *J. Mol. Biol.* **228**, 277–297.

Walter, J., Steigemann, T. P., Singh, H., Bartunik, H., Bode, W. & Huber, R. (1982). On the disordered activation domain in trypsinogen. Chemical labelling and low temperature crystallography. *Acta Crystallog. sect. B.* **38**, 1462–1472.

Warshel, A. & Levitt, M. (1976). Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227–249.

Warshel, A. & Russell, S. T. (1984). Electrostatic interactions in biological systems and in solution. *Quart. Rev. Biophys.* **17**, 283–422.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–784.

Weng, Z., Vajda, S. & Delisi, C. (1996). Prediction of protein complexes using empirical free energy functions. *Protein Sci.* **5**, 614–626.

Wilson, C., Mace, J. E. & Agard, D. A. (1991). Computational method for the design of enzymes with altered substrate specificity. *J. Mol. Biol.* **220**, 495–506.

Wilson, I. A. & Stanfield, R. L. (1993). Antibody-antigen interactions. *Curr. Opin. Struct. Biol.* **3**, 113–118.

Edited by F. E. Cohen

(Received 11 July 1997; received in revised form 3 November 1997; accepted 3 November 1997)

Note added in proof: For availability of an executable version of the interface refinement program MULTIDOCK see http://www.icnet.uk/bmm or contact M. J. E. S. (e-mail: m.sternberg@icrf.icnet.uk.