Q-Dock: Low-Resolution Flexible Ligand Docking with Pocket-Specific Threading Restraints

MICHAL BRYLINSKI, JEFFREY SKOLNICK
Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, 250 14th Street NW, Atlanta, Georgia 30318

Received 24 September 2007; Accepted 18 December 2007
DOI 10.1002/jcc.20917
Published online 21 February 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: The rapidly growing number of theoretically predicted protein structures requires robust methods that can utilize low-quality receptor structures as targets for ligand docking. Typically, docking accuracy falls off dramatically when apo or modeled receptors are used in docking experiments. Low-resolution ligand docking techniques have been developed to deal with structural inaccuracies in predicted receptor models. In this spirit, we describe the development and optimization of a knowledge-based potential implemented in Q-Dock, a low-resolution flexible ligand docking approach. Self-docking experiments using crystal structures reveals satisfactory accuracy, comparable with all-atom docking. All-atom models reconstructed from Q-Dock's low-resolution models can be further refined by even a simple all-atom energy minimization. In decoy-docking against distorted receptor models with a root-mean-square deviation, RMSD, from native of ~3 Å, Q-Dock recovers on average 15-20% more specific contacts and 25–35% more binding residues than all-atom methods. To further improve docking accuracy against low-quality protein models, we propose a pocket-specific protein–ligand interaction potential derived from weakly homologous threading holo-templates. The success rate of Q-Dock employing a pocket-specific potential is 6.3 times higher than that previously reported for the Dolores method, another low-resolution docking approach.

 J Comput Chem 29: 1574–1588, 2008

Key words: Q-dock; ligand docking; low-resolution docking; pocket-specific potential; protein models; threading

Introduction

Computational modeling of protein–ligand interactions is of great importance in modern structural biology and has many applications in investigating fundamental biochemical processes and in the development of new pharmaceutical compounds.[1–3] During the past years, a number of diverse algorithms for docking small molecules into receptor proteins have been developed[4–7] and evaluated in terms of docking accuracy and the ability to predict binding affinities.[8–10] In general, docking algorithms seek to identify the lowest free energy position of a ligand in the binding site of the receptor protein. These algorithms are designed to reproduce the experimentally given structure of a receptor protein complexed with a ligand and to rank all generated solutions such that the conformation closest to the experimental structure appears as the top model. There are two key elements of a docking approach: First, a scoring function is required that accurately ranks the generated set of solutions. Second, a fast and effective search algorithm is necessary to explore the conformational space of protein–ligand interactions. Search efficiency is particularly important in virtual screening experiments[11,12] that require many thousands of possible ligands to be docked into a receptor structure in an acceptable amount of time, usually no more than few minutes per ligand.

Docking programs typically utilize high-resolution receptor structures determined by experiment or theoretical modeling.[13-15] Virtual screening reveals that the success of the docking calculation typically depends on the quality of the receptor structure with the success rate decreasing from ligand-bound to ligand-free to modeled structures.[16] This drop off is correlated with the degree of protein movement in the active site; protein active site rearrangements greater than 1.5 Å lead to almost complete lack of recovery of the "true" binding mode.[17] Furthermore, decoy-docking experiments using deformed trypsin structures with a

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/0192-8651/suppmat

Correspondence to: J. Skolnick; e-mail: skolnick@gatech.edu

Contract/grant sponsor: Division of General Medical Sciences, National Institutes of Health; contract/grant numbers: GM-37408, GM-48835

Cα RMSD varying from 1 to 3 Å as targets for docking of 47 ligands experimentally known to bind to trypsin revealed that the specific native contacts between the ligands and their receptor structures are rapidly lost with the deformation of the receptor structure.[18]

On the other hand, protein models can now be routinely determined by high-throughput modeling procedures for entire proteomes. Many of the protein structures generated by structure prediction algorithms appear as attractive targets for the development of biologically active compounds.[19] As demonstrated by CASP7, the quality of theoretical methods for protein tertiary structure prediction has improved, and in many cases, predicted models are comparable to low-resolution experimental structures.[20,21] Nonetheless, these models have significant structural inaccuracies in side-chain and backbone coordinates when compared to ligand-bound, experimentally solved structures. An estimated one half of weakly homologous protein models have a RMSD from the native binding site >2 Å.[22]

A variety of different docking techniques have been developed to address this problem. Most account for receptor flexibility by docking ligands against a precalculated ensemble of receptor conformations[23] or by softening the criterion for the steric fit between the ligand and receptor.[24] To overcome the limitation of computationally expensive modeling of macromolecules, the Flexibility Tree combining a variety of efficient motion descriptors has been recently developed and implemented in the FLIP-Dock program.[25] Other docking techniques capable of dealing with significant structural inaccuracies employ a low-resolution representation of the protein. It has been shown that a ultra low (~7 Å resolution) representation of molecular structure averages all high-resolution structural details and dramatically improves the tolerance to receptor structure deformation.[26] A similar approach used to dock small molecules into low-resolution models demonstrated that even low-quality receptor structures could be efficiently utilized in docking experiments.[27] Nevertheless, most low-resolution docking approaches neglect ligand flexibility.

The desire to improve the state-of-the-art motivated us to develop Q-Dock, an approach that effectively utilizes low-quality protein structures as targets for flexible ligand docking. Q-Dock describes both the ligand and the protein in a reduced representation. Ligand flexibility is accounted for by docking an ensemble of precalculated discrete ligand conformations with Monte Carlo Replica Exchange (REMC) used to optimize the binding mode of the ligand in the binding site of the rigid receptor protein. Here, we describe the development and optimization of a coarse-grained knowledge-based potential implemented in Q-Dock. The performance of Q-Dock is compared with several popular all-atom programs for flexible ligand docking in a self-docking experiment using the crystal structures of target receptors. Next, we evaluated the efficiency of Q-Dock in a decoy-docking study against a set of distorted receptor structures whose Cα RMSD from the crystal structure ranges from 1 to 3 Å. Finally, with regards towards improving the quality of ligand–receptor pose predictions, we take full advantage of pocket-specific potentials derived from weakly homologous threading templates and apply them to the docking of ligands against modeled receptor structures.

Table 1. Predefined Chemical Groups Used to Decompose Ligands into Quasichemical Building Blocks.

Number	Description	Symbol/formula
1	Aromatic rings	mono-, heterocyclic five-, six-membered
2	Ether	−C−O−C−
3	Thioether	−C−S−C−
4	Carbonyl	>C=O
5	Thiocarbonyl	>C=S
6	Halogene	−Cl; −Br; −F; −I
7	Guanidine	$-NHC(NH_2)NH$
8	Amide	−CONH−
9	Carboxyl	−COOH
10	Amine (primary, secondary, tertiary)	$-NH_2$; >NH; >N−
11	Phosphate	$-PO_4$
12	Sulphate	$-SO_4$
13	Nitro group	$-NO_2$
14	Metals	Fe; Zn; Mg; Ca
15	Hydroxyl group	−OH
16	Thiol group	−SH
17	A fragment of aliphatic chain composed of at least 2 carbons not connected to groups 7-16	$-(C-C)_x-$; $-(C=C)_x-$; $-(C\equiv C)_x-$

Methods

Dataset

The structures of protein–ligand complexes were selected from the Protein Data Bank[28] according to the following criteria: Protein structures determined by X-ray crystallography to a resolution ≤2.5 Å and that have at least 50 residues were chosen. Organic molecules, cofactors, single nucleotides, and short peptides composed of standard or modified amino acids were considered as ligands if the number of predefined functional groups (listed in Table 1) was ≥5 and ≤25. To exclude nonspecific ligand interactions, a minimum number of five residues in contact with the ligand atoms are imposed. Interatomic contacts are calculated by LPC[29] that defines contacts based on an analysis of interatomic surfaces. Structures containing two or more ligands within 9 Å of each other were rejected. Subsequently, the complexes were subjected to a clustering procedure that uses a cut-off of 35% amino acid sequence identity between clusters. Two homologous proteins (members of one cluster) were accepted into the dataset only if the Tanimoto coefficient,[30] *TC*, calculated for their ligands was below 0.5. A high *TC* (typically 0.7–1.0) is indicative of very high chemical similarity. In this manner, a dataset of 1636 complexes was created, which can be considered as nonredundant with respect to protein–ligand interactions. This dataset was then divided into two sets: a training set of 818 complexes used to derive the statistical potential and then to optimize force field parameters and weights and a benchmark set of 818 structures used exclusively to assess the derived potentials. Training set proteins with a sequence identity ≥35% to any of the 34 targets used in the docking experiment

(described below) were exchanged with randomly selected benchmark proteins so that no proteins with $\geq$35% sequence identity to any docking target are used to derive and optimize the force field parameters.

The performance of Q-Dock was evaluated in a self-docking experiment for the set of protein–ligand complexes for which comparative assessments of several programs for all-atom flexible molecular docking were reported.[31-33] From the original dataset we removed three structures of cytochrome P-450 that contain two ligands in the binding pocket. The resulting set consists of 34 protein–ligand complexes (PDB codes: 1abe, 1abf, 1apt, 1apu, 1cbx, 1cil, 1cnx, 1etr, 1ets, 1ett, 1gsp, 1icm, 1icn, 1nnb, 1nsc, 1nsd, 1okl, 1pph, 1rhl, 1rls, 1tng, 1tni, 1tnj, 1tnk, 1tnl, 1tpp, 2ifb, 3cpa, 3ptb, 3tmn, 5abp, 5tln, 6cpa, 6tmn).

Next, we used Q-Dock in a decoy-docking study against distorted receptor structures. The decoy dataset consists of 291 models of trypsin of which 93, 101 and 97 structures have a Cα RMSD from the crystal structure of 1, 2 and 3 $\pm$ 0.5 Å, respectively. The distorted receptor models were used as targets for docking 47 ligands co-crystallized with trypsin. Details concerning the preparation of distorted models of trypsin and ligand selection are presented elsewhere.[18] We compared the results of Q-Dock decoy-docking with the results reported for all-atom docking by Kim and Skolnick.[18]

Finally, the performance of Q-Dock was evaluated for weakly homologous protein models used as targets for docking flexible ligands. From the set of 318 proteins, for which the results of the Dolores method were reported,[27] we selected 206 proteins up to 300 residues in length. Protein structure modeling consisted of template identification followed by an assembly/refinement procedure. First, for each target protein weakly homologous structure templates were selected from a nonredundant PDB library by our threading algorithm PROSPECTOR_3,[34,35] which was designed to identify analogous as well as homologous templates. We note that only threading templates with a sequence similarity to the target protein <35% were used in the modeling procedure. Subsequently, threading templates were submitted to TASSER,[36–38] a coarse-grained template assembly/refinement procedure guided by tertiary restraints extracted from threading templates. Weakly homologous protein models were then taken as targets for the prediction of ligand binding sites using FINDSITE, a method that identifies ligand-binding sites based on binding site similarity among superimposed groups of template structures identified from threading.[22] Ligand-binding sites predicted by FINDSITE were used to extract pocket-specific protein–ligand restraints from the threading templates to support low-resolution docking of flexible ligands into the theoretical receptor structures using Q-Dock.

Q-Dock Force Field

To quantitatively describe protein–ligand interactions, a combined knowledge-based potential was derived from the regularities observed in training protein–ligand complexes. The generic part of the force field (E_{GEN}) consists of four energy terms that account for different energetic contributions. E_{CP} (contact potential) accounts for the attractive and repulsive interactions between protein residues and ligand functional groups, i.e. it favors a specific orientation of a small molecule in the binding pocket. The surface-dependent terms E_{SL} and E_{SP} are in general less specific, scaled to the portion of the accessible solvent area of ligand functional groups and binding pocket residues that become buried upon complex formation. The differences in the accessible solvent area in the complexed and fully solvated states are used to express the burial likelihood for ligand functional groups and binding pocket residues. Moreover, we include a bias to the expected number of contacts, E_{CN} (spatially neighboring residues), for ligand functional groups. Finally, to ensure the best native-like recognition capability, the force field parameters were optimized against the ensemble of ligand decoys and the energy terms were combined with optimized weight factors.

Reduced Model of Protein–Ligand Complexes

A knowledge-based potential implemented in Q-Dock was developed for simplified models of ligands and receptor proteins. We employed the following coarse-grained representation of protein–ligand complexes: Protein residues are represented by Cα atoms and single points at their side-chain centers of mass. For glycine residues, only the Cα positions are used. Ligand molecules are first decomposed into 17 chemical groups, which are listed in Table 1. A single effective point is then placed at the center of mass of each group. Since conformational space for protein–ligand interactions is defined continuously, a repulsive potential is essential to account for the volume exclusion among a ligand and a protein. We defined two repulsion shells: a ligand group – side-chain repulsion shell S_{ij}^R and a ligand group—backbone repulsion shell B_j^R. The pair-specific repulsive shell S_{ij}^R was defined as the minimum distance between effective points of the side chain center of mass of amino acid i and ligand functional group j. For each effective ligand point, a backbone repulsion shell B_j^R is defined as the minimum observed distance from any Cα atom in crystal structures of protein–ligand complexes. The excluded volume between units is approximated by a strong energy penalty when the distance between them is below the cutoff values of S_{ij}^R or B_j^R.

Ligand—Side-Chain Contact Potential

For each pair of amino acid i and ligand functional group j, a unique contact shell was defined. The limiting values for the pair-specific S_{ij}^C were calculated for the protein–ligand complexes present in the dataset using the Matthew's correlation coefficient, MCC:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (1)$$

where TP and TN is the number of true positives and true negatives and FP and FN is the number of false positives and false negatives, respectively. TP, TN, FP, and FN were obtained by comparison to the interatomic interactions calculated for all-atom models. A residue and a ligand functional group are defined to be in contact if any of their heavy atoms were found to be in contact as reported by the LPC algorithm,[29] which is

based on the inter-atomic contact surface analysis. For each pair of effective points i and j, a pair-specific contact shell S_{ij}^C is determined by a distance cutoff that maximizes MCC.

The limiting distances were subsequently used to extract the observed number of contacts between a given pair of amino acid i and ligand functional group j in the training set of protein–ligand complexes (n_{ij}). The observed number of contacts is then compared to that expected in a reference state where there are no specific interactions:

$$n_{ij}^0 = N \times x_i \times x_j \tag{2}$$

where n_{ij}^0 is the expected number of contacts between amino acid i and ligand functional group j, N is the total number of contacts between any pair of protein–ligand effective points, and x_i and x_j are the mole fraction of units i and j in the training set, respectively. For protein residues, the mole fractions are calculated with respect to surface residues only. A surface residue is defined having $\geq$30% of its total surface exposed. We used POPS-A[39] for the solvent accessible area calculations.

The potential of mean force P^C between amino acid i and ligand functional group j is simply given by:

$$P_{ij}^c = -\ln \frac{n_{ij}}{n_{ij}^0} \tag{3}$$

Non-Polar Surface-Dependent Potential

The change in a solvent accessible area upon complex formation is accounted for as a surface-dependent potential. The non-polar surface-dependent potential P^S is based on the differences in the accessible solvent area of a ligand functional group or a protein residue in the complexed and fully solvated states:[40]

$$P_i^S = -\ln \frac{g_i(\mathrm{ASA_C})}{g_i(\mathrm{ASA_S})} \tag{4}$$

where g_i is the probability distribution of the solvent accessible area attached to unit i in the complexed state ($\mathrm{ASA_C}$) compared to the solvated state ($\mathrm{ASA_S}$). The distribution function g is calculated for ligand groups and amino acids by a statistical analysis of the protein–ligand complexes present in the training set. For proteins, only binding pocket residues are taken into consideration. The solvent accessible area of coarse-grained models of both ligands and proteins was approximated by the modified method of Wodak and Janin[41,42] (the details are given in the Appendix).

Contact Number

A bias to the expected number of neighboring residues for each ligand functional group is incorporated into the force field as

$$E_{\mathrm{CN}} = \sum_{j=1}^{L} |N_j - N_j^0| \tag{5}$$

where L is the total number of effective points in the ligand molecule, N_j is the observed number of contacting residues (calculated using the pair-specific contact shell S_{ij}^C) and N_j^0 is the expected number of neighbors (the mean value calculated for protein–ligand complexes in the training set).

Generation of Decoys

The energy parameters as well as the energy weight factors were optimized against an ensemble of decoy conformations. For each protein–ligand complex, an ensemble of nonredundant flexible decoys was constructed as follows: In the first step, 10^9 ligand orientations were created. A sphere of 7 Å radius centered on the center of mass of the ligand in the native conformation was imposed, such that if a ligand molecule leaves the sphere it will enter through the opposite side. Subsequently, the number of ligand variations was reduced by using hard-sphere steric potentials S_{ij}^R and B_j^R to account for volume exclusion between the ligand and the protein. To avoid the overaccumulation of some ligand orientations, a pairwise position similarity cutoff was used to ensure that the RMSD of any pair of decoys is larger than 3.5 Å. In addition, for each 20 non-native decoys (RMSD from native $>$3.5 Å), one native-like conformation (RMSD from native $\leq$3.5 Å) was generated and included into the decoy ensemble to account for the ligand distribution around the native position.

Parameter Optimization

Similarly to Genetic Algorithms, Evolution Strategies (ESs) are algorithms which imitate the principles of natural evolution as a method to solve parameter optimization problems.[43,44] ESs are random strategies, and as such are particularly robust and cope well with a large number of variables, or rugged objective functions. We employed the ES algorithm to improve the native-like recognition capability by the optimization of the force field parameters against the ensemble of ligand decoys. For each energy term, its parameters were optimized independently using the values derived from the statistical analysis as the initial set. The objective function to minimize (G) was the combination of the correlation between the energy function and the RMSD from the native ligand position (CC), the Z-score (the dimensionless ratio of the first and second moments of the energy distribution within the native-like pool and the decoy pool) and the B-score (the fraction of decoys with an energy higher than that of at least one native-like conformation):

$$G = \frac{1}{1 + \frac{1}{N}\sum_{p=1}^{N} \mathrm{CC}_p} \times \frac{1}{1 + \frac{1}{N}\sum_{p=1}^{N} Z\text{-score}_p} \times \frac{1}{1 + \frac{1}{N}\sum_{p=1}^{N} B\text{-score}_p} \tag{6}$$

where N is the total number of training protein–ligand complexes, and CC_p, $Z\text{-score}_p$, $B\text{-score}_p$ are the coefficients calculated for a complex p.

Weight Optimization

It was already shown for reduced protein models that the combined energy with optimized weight factors has higher correla-

tion coefficients and native-like recognition ability than a naïve combination of energy terms (all the weight factors set to 1) and each of the single energy terms alone.[45] We used this observation to optimize the energy weight factors. The optimization was done using the CERN MINUIT package.[46] Similar to the optimization of force field parameters, this procedure minimizes the objective function G as defined by eq. (6).

Ligand Move Set

We allow rotational and translational freedom of a small molecule within a restricted area of the receptor protein. A spherical distribution is sampled to generate random vectors, located on a spherical surface. To speed up the conformational space sampling, their normalized components[47] ($|v|^2 = x_1^2 + x_2^2 + \ldots + x_6^2 = 1$) are used as the scaling factors of the translational (1.0 Å) and rotational (10°) steps of a random walk. For each protein–ligand complex, we also allow for the random perturbation of the ligand's internal conformation sampled according to a uniform distribution.

Similar to other docking algorithms that employ a pre-docking generation of multiple ligand conformations,[5,48] ligand flexibility in Q-Dock is accounted for by docking an ensemble of ligand discrete conformations into the receptor protein. First, the set of conformations is generated for the all-atom ligand representations using the torsion angles as the degrees of freedom. Torsion angles are identified with the aid of the Autotors program available from AutoDock.[6,49] The number of states for each ligand dihedral angle depends on the hybridization of the linked atoms: three states (60, 180, 300°) are considered for two sp^3 hybridized atoms, two states (0, 180°) for two sp^2 hybridized atoms and 12 states (starting from 0° with 30° step) for all other combinations.[5] Conformations with steric clashes (when the distance between two nonbonded atoms <2 Å) are excluded. Moreover, a structural similarity cutoff is imposed to ensure that any two ligand conformations in the ensemble have a RMSD >1 Å. Subsequently, all-atom ligand representations are decomposed into 17 chemical groups, see Table 1, and a single effective point is placed at the center of mass of each group.

Energy Minimization

For a reasonable force field, a ligand native pose should appear as the lowest energy conformation. To determine the deviation of the lowest energy pose from experiment, we performed simple low-resolution energy minimization using the Simplex method[50] starting from the crystal structures. Energy minimization was carried out for training as well as benchmark protein–ligand complexes using the statistical and optimized sets of parameters with optimized energy weight factors.

Binding Mode Optimization (Docking)

To efficiently explore the conformational space in docking simulations, we used Replica Exchange Monte Carlo (REMC).[51-53] The temperature range was chosen such that at the lowest temperature a protein–ligand complex is stable in the native structure, whereas at the highest temperature, a ligand freely explores conformational space. A 7 Å radius sphere is imposed to prevent the ligand molecule from moving too far from the binding site in the high temperature replicas. Q-Dock utilizes 16 replicas where each is created by randomly choosing the position of a ligand in the vicinity of the binding pocket. The simulations consist of 100 attempts at replica exchange and 100 MC steps between replica swaps. The lowest energy ligand conformation identified in all replica trajectories is taken as the final model.

Pocket-Specific Protein–Ligand Potential

To improve docking accuracy particularly against low-quality protein models, we incorporated into the force field a pocket-specific protein–ligand interaction potential that is derived from weakly homologous (<35% sequence identity to a target protein) threading holo-templates. First, structure templates are identified by the threading algorithm PROSPECTOR_3[34,35] and used to predict ligand-binding sites and binding residues by recently developed FINDSITE algorithm.[22] A short overview of FINDSITE is provided in the Appendix. To derive a pocket-specific protein–ligand interaction potential, we used binding pockets predicted for each target protein by FINDSITE. Protein–ligand contacts are calculated for all threading templates that share a top-ranked predicted binding site. These are used to extract the observed number of contacts between a binding residue corresponding to position k in the target sequence (the chemical properties of binding residues are ignored) and ligand functional group of type j. Subsequently, the expected number of contacts in a reference state is calculated as in eq. (2). Then, a pocket-specific potential of mean force E_{PS} between a binding residue at position k in the target sequence and a ligand functional group of type j is given by eq. (3), but now averaged over the FINDSITE identified ligands and functional groups. The total energy now becomes the sum of weighted generic energy terms (E_{GEN}) and the pocket-specific energy (E_{PS}):

$$E_{TOT} = E_{GEN} + w_{PS}E_{PS} \tag{7}$$

The weight w_{PS} was optimized using the objective function G [eq. (6)] over the subset of 426 proteins ≤400 residues in length selected from the training set. During the optimization of w_{PS}, the generic weights were kept fixed at previously optimized values. Native-like recognition capability was then separately assessed for the subsets of proteins ≤400 residues selected from the training (426 cases) and benchmark complexes (400 cases).

Reconstruction of All-Atom Models and Simple High-Resolution Refinement

The final models obtained from Q-Dock simulations can be easily transformed into their all-atom representation. Reconstruction consists of the translation and rotation of all-atom ligand structures and the adjustment of dihedral angles so that the centers of mass of the functional groups overlap exactly with those predicted by the low-resolution docking simulation. The rebuilt protein–ligand complexes are subsequently refined by a simple energy minimization procedure using Amber8[54] with the all-atom force field ff03[55] used for proteins in conjunction with the general Amber force field,[56] GAFF, for ligand molecules.

Hydrogen atoms are added by the Open Babel package.[57] To speed up ligand parameterization, partial charges on ligands atoms were approximated by the Gasteiger-Marsili[58] formalism. A Coulombic potential on a 1 Å grid was calculated by LEaP (Amber8) in order to place chloride or sodium ions at positions of the highest or lowest electrostatic potential around a protein–ligand complex to neutralize it. Long-range non-bonded interactions were truncated using a 12 Å cutoff (electrostatic and vdW). The protein was kept fixed during the simulation, whereas the conformation of the ligand is energy minimized in 250 cycles of steepest-descent followed by 250 cycles of a conjugate gradient procedure.

Results

Ligand—Side-Chain Contact Potential

The averaged interactions between ligand functional groups and surface residues in the nonredundant library of 818 training protein–ligand complexes were used as a reference state for the calculation of a log odds potential that expresses the likelihood of interaction between ligand groups and protein residues. The average value of the MCC [defined in eq. (1)] for contacts using the reduced representation as compared to a detailed atomic model is 0.8, which suggests that the extracted contacts between effective points in reduced models reproduce well the real contacts between ligands and receptor proteins observed in all-atom structures. In general, favorable and unfavorable interactions between amino acids and ligand functional groups are found to be consistent with their physicochemical properties.

Non-Polar Surface-Dependent Potential

Solvent effects are accounted for as a nonpolar surface-dependent potential. We observed that a very small portion of hydrophobic groups surface remain solvent accessible, rather, the complete burial of hydrophobic groups is strongly favorable. Simultaneously, a "partially" buried state is favorable for most hydrophilic groups. The optimization procedure significantly enhances the preferences of polar and nonpolar functional groups. Similar characteristics are observed for binding pocket residues.

Contact Number

For the statistically derived set of parameters, the contact number simply expresses the average number of neighboring residues calculated for training protein–ligand complexes. The optimization procedure caused a significant increase in the expected contact number and corresponds to a strong penalty for ligand conformations that partially form a complex with the receptor protein (characterized by fewer contacts compared to the native conformation).

Minimization of Native Complexes

An accepted quality measure for the results of docking small molecules into the receptor proteins is the root-mean-square deviation, RMSD, from the ligand position in the complex crystal structure.[31-33] As a consequence of the imperfections of the force field as well as experimental deficiencies affecting reference conformations, often the energy minimum does not exactly correspond to the native conformation.[40] Nevertheless, for a reasonable force field, the lowest energy pose of a ligand should not deviate substantially from the native conformation. To ascertain the deviation from experiment when Q-Dock's force field is used, we performed a simple energy minimization, starting from the crystal structure. The simulations were carried out separately for each force field parameter set using the optimized weights factors for energy terms. The results obtained for the training and the benchmark set are shown in Figure 1. The minimization procedure slightly shifted down the central tendency of energy (Fig. 1A) and causes an acceptable deviation from the crystal structures. In most cases, the lowest energy ligand positions do not deviate by more than 2.0 Å from the experimental structure (Fig. 1B) and preserve >90% the of the native protein–ligand contacts (Fig. 1C).

Native-Like Recognition Capability

The quality of the native-like discriminatory power of Q-Dock was assessed by the correlation between the energy and RMSD from the native ligand position (CC), the relative energy gap between native-like structures and the ensemble of non-native decoys (*Z*-score), and the fraction of decoys with an energy higher than at least one native-like structure (*B*-score). The summary of native-like recognition capability is presented in Table 2. The parameters optimized on ligand decoys exhibit considerably higher discriminatory power than the statistically derived potential. Furthermore, the optimization of weight factors improved native-like recognition capability. Finally, the slight difference between the coefficients calculated for the training and benchmarking set excludes possible specificity toward the training complexes. Thus, in all subsequent calculations, only results for the optimized parameters are reported.

Weight Optimization for Pocket-Specific Restraints

To further improve docking accuracy against the crystal structures as well as low-quality predicted receptor structures, a pocket-specific protein–ligand interaction potential (E_{PS}) was derived from weakly homologous (<35% sequence identity to a target protein) threading holo-templates and combined with the generic potential derived from the regularities observed in crystal structures of the training complexes [see eq. (7)]. The value of $w_{PS} = 3.1$ was found to maximize the native-like recognition capability (see Table 2).

Docking Results for Receptor Crystal Structures

The performance of Q-Dock was evaluated for 34 protein–ligand complexes for which comparative assessments of all-atom algorithms for flexible ligand docking were reported.[31-33] The crystal structures of proteins were taken as targets for flexible ligand docking using the optimized generic parameters set (E_{GEN}) as well as the generic potential combined with the pocket-specific threading restraints (E_{TOT}). No proteins with >35% sequence

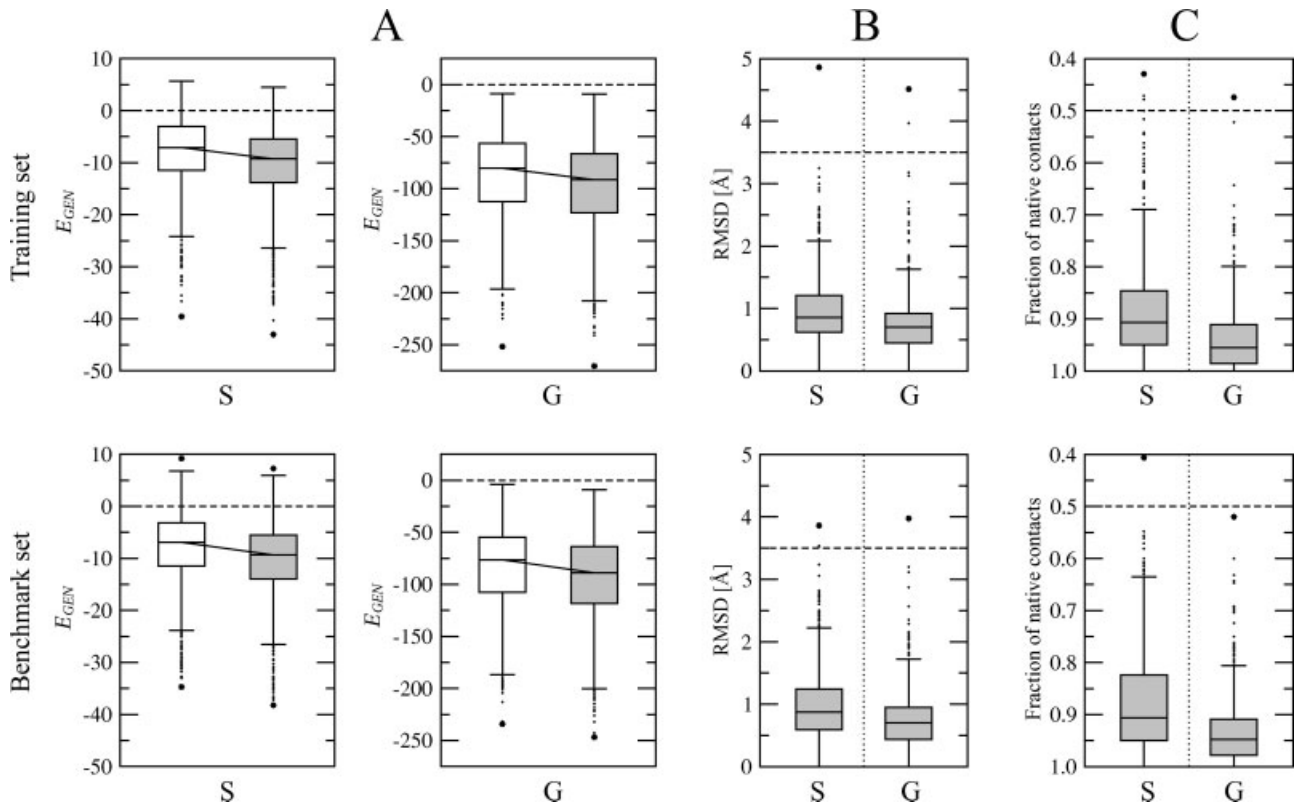

Figure 1. Distribution of the combined generic energy E_{GEN} (A), RMSD from the native structure (B) and the fraction of preserved native contacts (C) for minimized low-resolution models of training (top panel) and benchmark (bottom panel) protein–ligand complexes. White and gray boxes denote the crystal and minimized complexes, respectively. The results are presented for the two different sets of parameters used in the minimization procedure: statistically derived (S) and optimized (G) potentials. Boxes end at the quartiles Q_1 and Q_3; a horizontal line in a box is the median. "Whiskers" point at the farthest points that are within 3/2 times the interquartile range. Outliers, minima, and maxima are presented as dots and stars, respectively.

identity to targets are in the training dataset. The top docked solutions obtained from Q-Dock simulations were transformed into their all-atom representations and refined by a simple energy minimization in an all-atom force field.

The results of docking simulations evaluated by the RMSD from the crystal structure are presented in Table 3. The overall performance of Q-Dock is comparable to many all-atom approaches; the average RMSD calculated for all-atom models reconstructed from top-ranked docked conformations obtained using (E_{GEN}) and (E_{TOT}) is 3.90 and 3.03 Å, respectively. In Figure 2, we show the examples of the energy versus the RMSD correlation for neuraminidase (Fig. 2A, PDB-ID: 1nsc), intestinal fatty acid binding protein (Fig. 2B, PDB-ID: 1icm), ribonuclease T1 (Fig. 2C, PDB-ID: 1rhl) and thermolysin (Fig. 2D, PDB-ID: 3tmn). With the pocket-specific restraints (E_{TOT}), the global minimum is frequently closer to the native ligand pose. Moreover, the higher correlation between the energy and RMSD speeds up the convergence of the binding mode optimization.

Furthermore, we found that the high-resolution refinement improved the quality of the final models reconstructed from low-resolution images provided by Q-Dock. This is particularly pronounced for already well-docked solutions. Indeed, the vast majority of models with a RMSD <3.5 Å move toward the native pose. Next, we assessed the ability to select the native-like ligand conformation from the ensemble used to mimic ligand flexibility. Interestingly, native-like ligand conformers are often observed in the top docked solutions, even if the internal ligand energy was not evaluated and no energy minimization was applied. The average internal ligand RMSD from the native conformation calculated for the models reconstructed from top-ranked Q-Dock solutions obtained using (E_{GEN}) and (E_{TOT}) is 0.75 and 0.62 Å, respectively, whereas the average RMSD calculated for all conformations present in the ligand ensembles is 1.80 Å.

Examples of successful all-atom refinement are shown for neuraminidase and carboxypeptidase A in Figure 3. For neuraminidase (Fig. 3A, PDB-ID: 1nnb), the RMSD of the inhibitor 2-deoxy 2,3-dehydro-*N*-acetyl neuraminic acid rebuilt from low-resolution Q-Dock's top model is 3.99 Å. The final RMSD calculated for the inhibitor after all-atom refinement is 2.13 Å. The lowest energy pose of a phosphonate in the active site of carboxypeptidase A (Fig. 3B, PDB-ID: 6cpa) reported by Q-Dock

Table 2. Summary of Q-Dock's Native-Like Recognition Capability for a Large Set of Ligand Decoys.

		Statistical set of parameters[c]		Optimized set of parameters[d]			
		Training set	Benchmark set	Training set		Benchmark set	
Energy[a]	Coefficient[b]	(full)[g]	(full)[g]	(full)[g]	(≤400)[g]	(full)[g]	(≤400)[g]
E_{CP}	*CC*	0.30	0.26	0.48	0.51	0.47	0.49
	Z-score	1.18	1.01	1.90	2.12	1.81	1.96
	B-score	0.92	0.90	0.97	0.98	0.96	0.97
E_{SL}	*CC*	0.29	0.27	0.31	0.33	0.29	0.33
	Z-score	0.67	0.64	0.91	1.02	0.84	0.96
	B-score	0.89	0.88	0.92	0.93	0.91	0.92
E_{SP}	*CC*	0.27	0.27	0.50	0.52	0.51	0.52
	Z-score	0.88	0.86	1.85	1.99	1.88	1.95
	B-score	0.92	0.91	0.97	0.97	0.97	0.97
E_{CN}	*CC*	0.27	0.29	0.46	0.48	0.47	0.48
	Z-score	0.89	0.90	1.61	1.77	1.63	1.73
	B-score	0.89	0.90	0.96	0.96	0.96	0.96
E_{PS}	*CC*	–	–	–	0.48	–	0.47
	Z-score	–	–	–	2.47	–	2.40
	B-score	–	–	–	0.97	–	0.97
		E_{GEN}	E_{GEN}	E_{GEN}	E_{TOT}	E_{GEN}	E_{TOT}
E_{GEN}/E_{TOT}	*CC*[e]	0.38	0.34	0.53	0.59	0.52	0.58
	Z-score[e]	1.30	1.26	1.99	2.39	1.97	2.28
	B-score[e]	0.90	0.90	0.97	0.98	0.97	0.98
	CC[f]	0.42	0.41	0.54	0.64	0.53	0.63
	Z-score[f]	1.34	1.33	2.04	3.01	2.01	2.89
	B-score[f]	0.93	0.92	0.98	0.99	0.97	0.99

[a]Energy terms: $E_{CP,}$ ligand—side-chains contact potential; E_{SL}, $E_{SP,}$ non-polar surface-dependent potential for ligand groups and binding pocket residues, respectively, $E_{CN,}$ contact number; $E_{PS,}$ pocket-specific energy; $E_{GEN,}$ combined generic energy terms; $E_{TOT,}$ combined generic terms with pocket-specific restraints.
[b]*CC,* correlation coefficient between energy and RMSD to native structure; *Z*-score, the relative energy gap between native-like structures and the ensemble of non-native decoys; *B*-score, fraction of decoys with energy higher than at least one native-like structure.
[c]Derived from the statistical analysis of the training complexes.
[d]Optimized over the training set decoys and objective function *G*.
[e]Calculated for the naïve weight factors of energy terms.
[f]Calculated for the optimized weight factors.
[g]Calculated for the complete set (full) of proteins or the subset of proteins ≤400 residues in length.

corresponds to a RMSD of 2.84 Å. High-resolution refinement shifted the ligand toward the native pose with a final RMSD of 1.18 Å.

Docking Results for Deformed Receptor Structures

In this experiment, we used 291 distorted models of bovine trypsin with a 1, 2 and 3 ± 0.5 Å Cα RMSD from the crystal structure as targets for low-resolution flexible ligand docking using Q-Dock. Forty-seven different ligands known experimentally to bind to trypsin were docked into each deformed receptor structure. No high-resolution refinement was applied. The results were compared to those reported for all-atom decoy-docking using AutoDock and FlexX.[18] Figure 4 presents the accuracy of flexible ligand docking against the set of distorted receptor structures in terms of the fraction of correctly predicted specific native contacts and binding residues (nonspecific contacts). We find that Q-Dock is far less sensitive to the deformation of the receptor protein than all-atom approaches. The average fraction of binding residues predicted by AutoDock, FlexX, and Q-Dock for 1/2/3 Å RMSD decoys is 0.85/0.57/0.36, 0.70/0.43/0.26 and 0.87/0.68/0.62, respectively. Moreover, the average fraction of specific native contacts recovered by AutoDock, FlexX and Q-Dock for 1/2/3 Å RMSD decoys is 0.75/0.46/0.27, 0.59/0.33/0.19, and 0.62/0.47/0.42, respectively. In the case of the most distorted receptor structures (Cα RMSD of 3 ± 0.5 Å), Q-Dock was capable to predict on average 25–35% more binding residues and 15–20% more specific native contacts than all-atom approaches.

Docking Results for Receptor Models

The weakly homologous protein models used in this study were generated by a threading-based protein structure prediction pro-

Table 3. Comparison of RMSD values for the top models from all-atom and coarse-grained flexible ligand docking.

	All-atom docking								Q-Dock (E_{GEN}/E_{TOT})[a]	
PDB ID	AutoDock[b]	DOCK[b]	FlexX[b]	ICM[b]	GOLD[b]	SODOCK[c]	T10[d]	T20[d]	Rebuilt[e]	Refined[f]
1abe	0.16	1.87	0.55	0.36	0.18	0.25	0.56	0.56	2.92/2.63	2.33/2.41
1abf	0.48	3.25	0.76	0.61	0.50	0.31	0.68	0.70	3.28/3.80	3.19/4.01
1apt	1.89	8.06	5.95	0.88	8.82	2.18	5.72	4.79	0.93/0.71	0.77/0.69
1apu	9.10	7.58	8.43	2.02	10.70	1.42	1.32	1.32	2.08/0.91	1.75/0.73
1cbx	1.33	3.13	1.32	0.82	1.87	7.12	0.62	0.62	1.33/4.50	1.17/4.53
1cil	5.81	2.78	3.52	2.00	6.04	2.80	1.86	1.86	4.91/4.91	5.04/5.04
1cnx	10.90	7.35	6.83	2.09	6.32	7.15	6.20	6.20	9.21/9.21	9.36/9.36
1etr	4.61	6.66	7.26	0.87	5.99	1.14	1.09	1.09	1.59/1.31	0.56/0.53
1ets	5.06	3.93	2.11	6.22	2.39	2.15	1.97	1.97	0.83/3.68	0.72/2.66
1ett	8.12	1.33	6.24	0.99	1.30	2.57	0.82	0.82	3.28/2.50	2.60/2.51
1nnb	0.92	4.51	0.92	1.09	0.84	0.71	1.67	3.97	3.99/2.46	2.13/2.32
1nsc	1.40	4.86	6.00	1.80	1.02	0.89	1.47	1.40	4.69/0.82	4.24/0.45
1nsd	1.20	4.51	1.56	1.04	0.96	0.47	1.85	1.85	5.14/0.81	4.68/0.86
1okl	8.54	5.65	4.22	3.03	3.55	1.52	2.84	2.84	4.63/6.62	4.97/4.97
1pph	5.14	3.91	3.27	1.44	4.23	0.92	4.00	0.53	6.55/6.67	6.92/6.88
1tng	0.62	0.86	1.08	0.71	1.89	2.32	0.70	0.69	3.33/3.27	3.58/1.72
1tni	2.61	5.26	2.73	3.40	4.93	3.92	2.22	2.22	5.99/3.82	5.40/4.50
1tnj	1.21	1.56	1.73	2.17	1.90	2.12	1.42	1.50	7.21/5.32	8.16/4.89
1tnk	1.69	1.87	1.70	2.53	3.08	1.50	1.16	1.14	3.00/4.55	2.17/4.44
1tpp	1.80	3.25	1.95	1.71	2.33	1.65	2.43	2.53	0.56/2.62	1.88/2.46
2ifb	3.09	1.43	8.94	1.04	2.61	1.91	2.09	5.19	3.14/1.29	3.03/0.96
3cpa	8.30	8.30	9.83	1.60	4.96	1.37	2.22	2.22	1.51/1.26	1.41/1.14
3ptb	0.80	0.59	1.11	0.49	1.09	0.34	0.56	0.54	8.85/3.27	9.02/2.60
3tmn	4.51	7.09	5.30	1.36	3.96	4.10	3.65	3.65	6.20/2.05	6.13/1.03
5abp	0.48	3.89	4.68	0.88	0.59	0.23	0.48	0.51	3.66/3.13	3.81/2.94
5tln	5.34	1.39	6.33	1.42	1.60	9.18	1.21	1.21	5.50/1.25	5.16/1.07
6cpa	8.30	8.30	9.83	1.60	4.96	1.11	4.00	4.00	2.84/2.61	1.18/1.33
6tmn	8.72	7.78	4.51	2.60	8.54	2.99	2.21	2.21	2.17/1.79	1.41/1.05
Average[g]	**4.00**	**4.32**	**4.24**	**1.67**	**3.47**	**2.30**	**2.04**	**2.08**	**3.90/3.14**	**3.67/2.79**
1icm	1.80	3.99	2.94	1.11	2.30	5.26			3.62/2.28	3.42/1.20
1icn	3.99	3.88	2.95	1.35	2.05	7.79			3.13/3.19	2.35/2.54
1gsp	2.67	1.16	3.71	0.54	0.70	0.54			4.72/2.83	4.63/2.28
1tnl	0.41	2.08	3.74	1.93	1.61	0.46			4.77/5.20	4.23/5.36
1rhl	0.96	0.71	1.15	3.53	1.08	0.86			4.45/1.02	4.43/0.69
1rls	0.98	1.75	4.33	0.79	1.16	0.68			2.47/0.73	1.68/0.65
Average[h]	**3.62**	**3.96**	**4.04**	**1.65**	**3.12**	**2.35**			**3.90/3.03**	**3.63/2.67**

In Q-Dock simulations we employed the generic part of the force field (E_{GEN}) as well as the generic potential combined with pocket-specific threading restraints (E_{TOT}).
[a]Obtained using E_{GEN} or E_{TOT} energy function.
[b]Reported by Bursulaya et al.[31]
[c]Reported by Chen et al.[32]
[d]Reported by Taufer et al.[33]
[e]Calculated for all-atom structures reconstructed from the reduced models.
[f]Calculated for the reconstructed structures further refined in an all-atom force field.
[g]Average over first 28 complexes (1abe to 6tmn).
[h]Average over all 34 complexes (1abe to 1rls).

cedure that consists of structure template identification by PROSPECTOR_3[34,35] followed by template assembly/refinement using TASSER.[36–38] Subsequently, the modeled protein structures were submitted to ligand-binding site prediction using a recently developed FINDSITE algorithm that can accurately identify binding sites in experimentally solved protein structures as well as in approximate, theoretical models.[22] Here, FINDSITE predictions were used to derive a pocket-specific potential for each target protein. We provided Q-Dock with the modeled receptor structures, predicted binding sites and pocket-specific restraints and carried out flexible ligand docking simulations employing (E_{TOT}) [eq. (7)] as an objective function in ligand

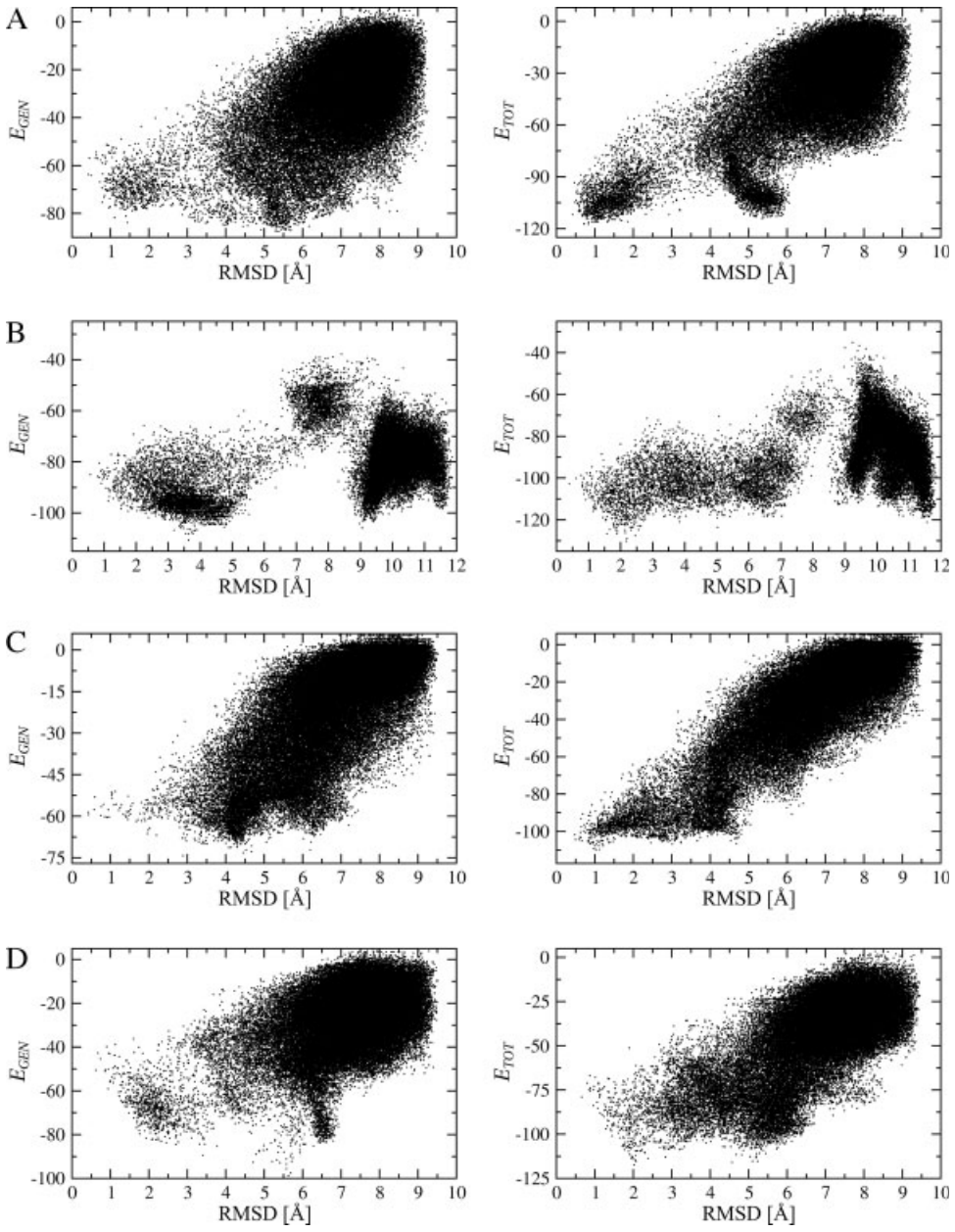

Figure 2. Energy plotted as a function of RMSD for REMC trajectories collected for neuraminidase (A), intestinal fatty acid binding protein (B), ribonuclease T1 (C), and thermolysin (D). The simulations were carried out using the optimized generic parameters set (E_{GEN}) as well as the generic potential combined with the pocket-specific threading restraints (E_{TOT}).

binding mode optimization and the selection of final models. Furthermore, to evaluate the improvement of docking accuracy against low-to-moderate quality protein models resulting from including pocket-specific restraints, we performed simulations using (E_{GEN}) (optimized generic energy terms only) instead of (E_{TOT}). The performance of Q-Dock was then compared with the results reported for Dolores which is another low-resolution approach that docks rigid ligand structures into receptor proteins.[27] The results obtained for the set of 206 target proteins evaluated in terms of the fraction of predicted specific protein–ligand contacts as well as the fraction of recovered binding residues are shown in Figure 5.

We note the higher accuracy of ligand-binding site prediction using FINDSITE compared to the grid-based method implemented in Dolores;[27] the fraction of proteins with at least one native specific contact is 0.73 for Dolores and is 0.92 and 0.93

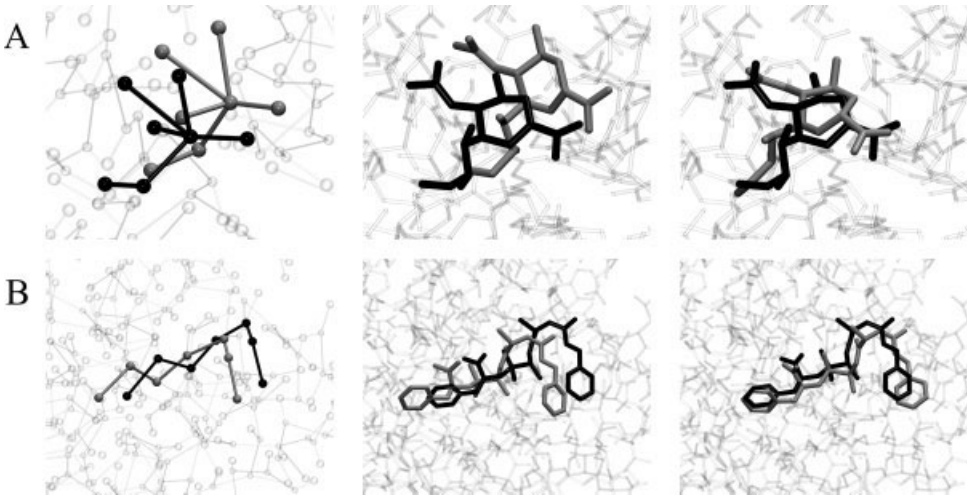

Figure 3. Examples of a high-resolution refinement for neuraminidase (A) and carboxypeptidase A (B). Low-resolution images representing Q-Dock top-ranked solutions, all-atom models rebuilt from coarse-grained models and refined structures are presented in left, middle and right column, respectively. The native and predicted ligand pose is colored black and grey, respectively. Receptor proteins are shown as transparent balls/sticks (reduced models) and sticks (all-atom models).

for Q-Dock employing (E_{GEN}) and (E_{TOT}), respectively. In general, Q-Dock predicts considerably more specific protein–ligand contacts than Dolores, especially if pocket-specific restraints are applied (Fig. 5, circles). For example, the fraction of proteins with $\geq$50% of recovered specific native contacts is 0.07, 0.30 and 0.46 for Dolores and Q-Dock employing (E_{GEN}) and (E_{TOT}), respectively. Interestingly, the fraction of predicted binding residues depends entirely on the accuracy of ligand-binding site prediction and not the presence of pocket-specific restraints (Fig. 5, squares). The restraints support the recovering of specific contacts as the result of the improved ability to predict a "true" ligand binding mode in the putative binding site of the receptor model.

Discussion

Despite progress in protein structure prediction, theoretical protein models frequently have structural inaccuracies in their side-chain and backbone coordinates when compared to experimentally determined structures. Since all-atom docking approaches were found to be highly sensitive to the structural distortions of the ligand binding region,[16,17,59] they are inapplicable to such models. This deficiency has motivated the development of protocols capable of docking small molecules into the structurally distorted ligand-binding sites using low-resolution docking techniques.[26,27,60,61] In this spirit, we have developed Q-Dock, an approach that effectively utilizes low-quality protein structures as targets for flexible ligand docking. The force field implemented in Q-Dock combines two classes of energy terms: generic knowledge-based potentials derived from the regularities observed in crystal protein–ligand complexes and pocket-specific potentials extracted for each target protein from ligand-bound forms of weakly homologous structure templates. The combined knowledge-based potential implemented in Q-Dock was derived from the statistics of crystal protein–ligand complexes and further optimized to increase the native-like recognition capability. The resulting potentials for low-resolution modeling of protein–ligand interactions seem to make good physical sense; they can be rationalized in terms of fundamental ligand–protein interactions including ionic interactions, hydrogen bonds, aromatic stacking or hydrophobic interactions.

Self-docking utilizing crystal structures of receptor proteins as targets for flexible ligand docking revealed that the accuracy of Q-Dock is comparable to all-atom approaches; in most cases, the native-like structures appear as the lowest-energy conformations. Furthermore, the low-resolution models can be transformed back into their all-atom representations and efficiently refined even by a simple all-atom minimization. For the vast majority of reasonably well-docked conformations reported by Q-Dock, the high-resolution refinement procedure considerably improved the quality of final models. Thus, low-resolution modeling serves as a valuable initial step for a more detailed structural analysis, as well as a complement to experimental and computational data obtained by other techniques.[26,62] Moreover, the results obtained by docking of the ensemble of discrete ligand conformations into receptor proteins shows that ligand flexibility can be successfully included in low-resolution docking. Despite the fact the ligand internal energy was ignored, native-like ligand conformers were frequently observed in top docked solutions.

The main practical advantage of a coarse-grained docking methodology, such as Q-Dock, is the possibility of utilizing low-quality receptor structures routinely produced by proteome-scale protein structure modeling projects. Our decoy-docking study of flexible ligands against the distorted receptor models revealed

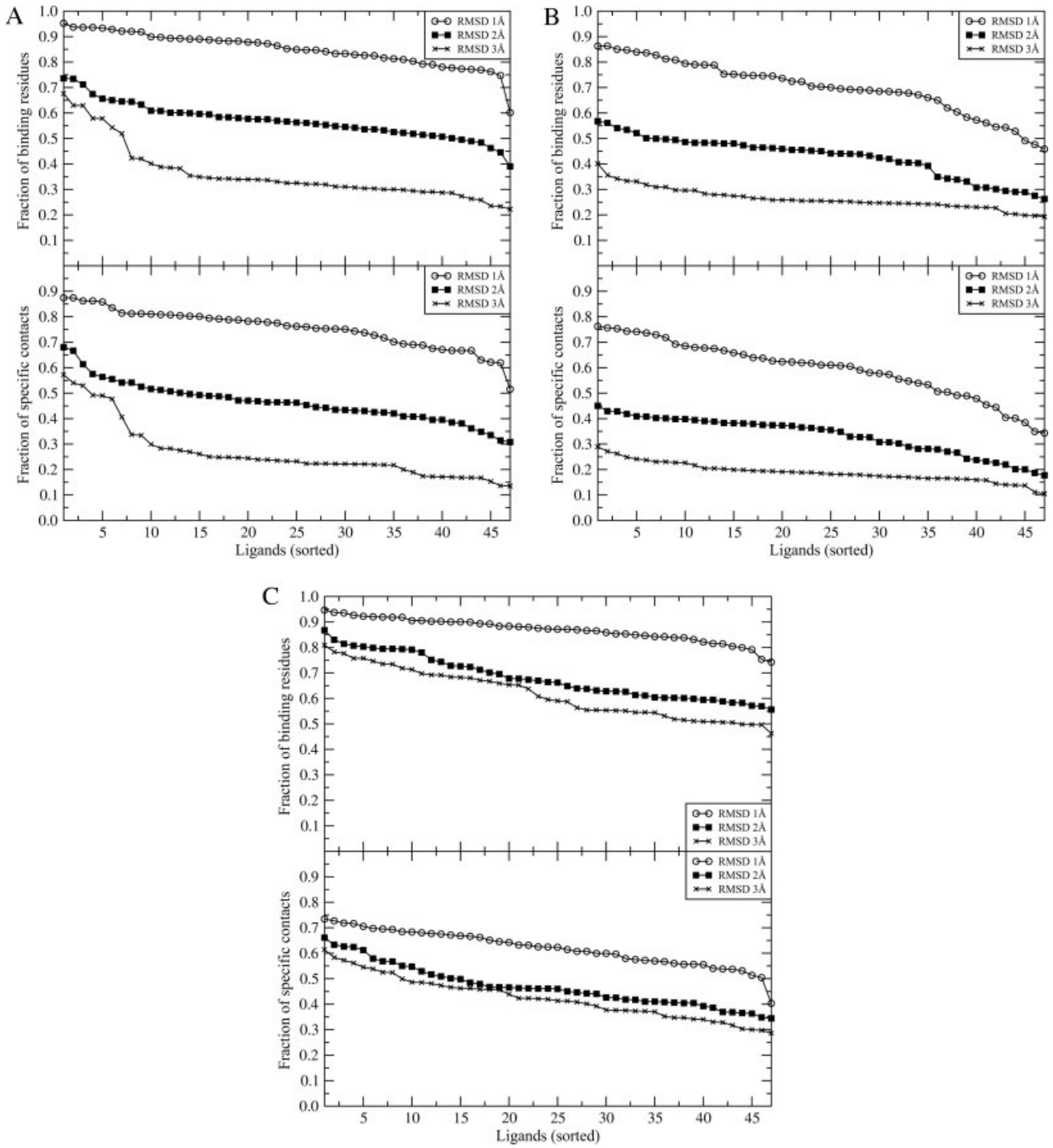

Figure 4. Comparison of the flexible ligand docking results for 47 different ligands and deformed structures of trypsin with Cα RMSD of 1, 2 and 3 ± 0.5 Å obtained using AutoDock (A), FlexX (B) and Q-Dock (C). Top and bottom plots show the fraction of predicted binding residues and the fraction of recovered specific protein–ligand contacts, respectively.

that Q-Dock recovers on average 25–35% more binding residues and 15–20% more specific native contacts than all-atom approaches. In more than 90% of the cases, at least one ligand-binding residue was correctly predicted. Moreover, in almost one-third of the cases, the fraction of recovered specific protein–ligand contacts was ≥50%.

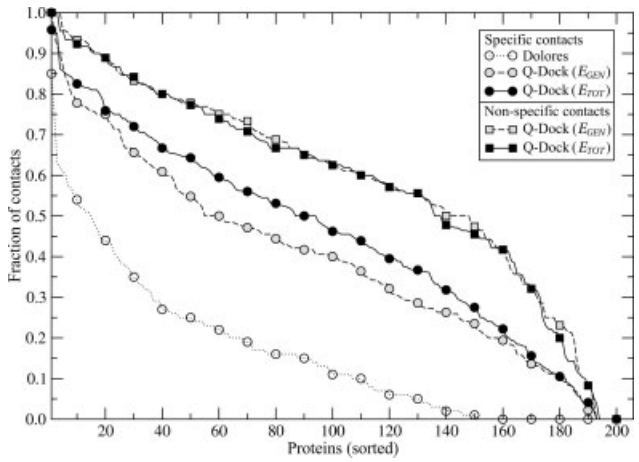

Figure 5. Fraction of predicted specific and nonspecific (binding residues) native contacts identified by Dolores method and Q-Dock using weakly homologous protein models as targets for docking small ligands. Flexible ligand docking simulations by Q-Dock were carried out employing only generic energy terms with the optimized set of parameters (E_{GEN}) as well as using generic terms combined with pocket-specific restraints derived from threading templates (E_{TOT}).

To full advantage of predicted binding regions, we proposed a pocket-specific protein–ligand interaction potential derived from weakly homologous structure templates selected by threading that can be used as valuable supplementary restraints in ligand docking against low-quality receptor structures. This yields a 6.3 times higher success rate of Q-Dock compared to the previously published Dolores method.[27]

The tolerance to structural inaccuracies in receptor models clearly enhances the importance of protein models as reliable targets for virtual screening or structure-based drug design. Q-Dock represents a practical tool for utilizing the rapidly growing number of theoretically predicted protein structures in experiments that require an effective flexible ligand docking procedure.

Acknowledgments

The authors gratefully acknowledge Dr RyangGuk Kim for providing the set of deformed receptor structures and all-atom docking results.

Appendix

Solvent Accessible Surface Estimation

To calculate the accessible surface area (ASA), we used an analytical approximation approach adapted from Wodak and Janin.[42] This fast and reliable analytical model expresses the ASA as a function of interatomic distances only, and works at both the atomic and residue levels. It has been shown that when it is applied to simplified models of proteins[41,63] and nucleic acids,[63] it reproduces the surface area calculated by accurate all-atom algorithms. The total solvent accessible area of a molecule is expressed as the sum of the ASA attached to all of its atoms:

$$\mathrm{ASA} = \sum_{i=1}^{N} A_i \tag{A1}$$

For a given atom i, the following expression can be applied to account for the intersecting spheres of the neighboring atoms:

$$A_i = S_i \prod_{i \neq i}^{N} \left(1 - \frac{p_i p_{ij} b_{ij}(r_{ij})}{S_i}\right) \tag{A2}$$

where S_i is the accessible solvent area of isolated atom i, $b_{ij}(r_{ij})$ is the area cut out by the overlap of the atom j at a distance $r_{ij} = |r_i - r_j|$, and p_i, p_{ij} are the empirical correction factors.

The ASA of isolated atom i with radius R_i can be calculated using a solvent probe with radius R_{SP} (usually equal to 1.4 Å[64]) as follows:

$$S_i = 4\pi(R_i + R_{\mathrm{sp}}) \tag{A3}$$

The area cut out of A_i by atom j can be calculated from

$$b_{ij}(r_{ij}) = \begin{cases} \pi(R_i + R_{sp})(R_i + R_j + 2R_{sp} - r_{ij})\left(1 + \frac{R_j - R_i}{r_{ij}}\right) \\ \quad \text{if } r_{ij} < R_i + R_j + 2R_{sp} \\ 0 \quad \text{otherwise} \end{cases} \tag{A4}$$

where R_i and R_j are the radii of atom i and j, respectively.

Originally, the method was tested for the all-atom as well as reduced representations of protein structures considering Cα atoms only. In our approach, the surface area is estimated based on the positions of the Cα atoms and centers of mass of residue side chains and ligand functional groups. The initial full set of parameters for the 20 amino acids (radii R, empirical correction factors p_i and p_{ij}) were taken from Cavallo et al.[39] The radii for ligand functional groups were obtained by statistical analysis of isolated ligand functional groups present in the set of protein-ligand complexes used in this study:

$$R_i = \sqrt{\frac{\langle S_i \rangle}{4\pi}} \tag{A5}$$

where R_i is the estimated radius of ligand group i and $\langle S_i \rangle$ is the average surface of the isolated group i, as calculated for all-atom models by LPC.[29]

Subsequently, the initial set of parameters for protein residues and ligand functional groups was submitted to an optimization procedure to minimize the variance of ASA calculated for reduced models of protein–ligand complexes from the ASA calculated for their all-atom models by POPS-A[39] and LPC.[29] Since in our model it is the residues in contact with a ligand that are important for protein–ligand interactions, the parameters for pro-

teins were optimized over binding pocket residues only. For the optimized set of parameters, the accessible surface area calculated for all-atom models is reproduced by this coarse-grained method, considering ligands and proteins individually as well as in the complexed state with an average correlation coefficient of 0.94. The approximation of accessible surface area seems to be well suited for the practical use in low-resolution docking simulations using Q-Dock.

Prediction of Ligand Binding Sites by FINDSITE

To predict ligand-binding sites in protein models and to derive a pocket-specific potential for protein–ligand interactions, we used the recently developed FINDSITE approach that detects ligand-binding sites based on the binding site similarity across superimposed groups of threading templates.[22] FINDSITE not only works well for crystal structures but also exhibits a good tolerance to structural inaccuracies in modeled protein structures (up to a global backbone RMSD from the crystal structure of 8–10 Å); thus it is particularly well suited for ligand-binding site prediction in weakly homologous protein models. FINDSITE employs template identification, structure superimposition and binding sites clustering as follows: First, for a given target sequence, structure templates are selected from a nonredundant PDB library by the threading program PROSPECTOR_3.[34,35] PROSPECTOR_3 evaluates the score significance in terms of the *Z*-score of the sequence assigned to a given structure based on the average of the best alignment given by Dynamic Programming over the template library. FINDSITE requires threading templates with *Z*-scores $\geq$4. For the purpose of benchmarking, from the threading templates reported by PROSPECTOR_3 we used only those that have <35% sequence identity to the target protein. Subsequently, structures that contain a bound ligand molecule are identified and superimposed onto a reference structure using the structural alignment algorithm TM-align.[65] In this study, we used TASSER-generated[36–38] models as reference structures for the template superimposition. Upon superimposition, the centers of mass of ligands bound to threading templates are clustered. Then each cluster represents one putative binding site. Finally, the predicted binding sites are ranked according to the number of threading templates that share a common binding pocket. For each target protein, we selected a top-ranked predicted ligand-binding site for ligand docking and the derivation of a potential for pocket-specific protein–ligand interactions.

References

1. Gilson, M. K.; Zhou, H. X. Annu Rev Biophys Biomol Struct 2007, 36, 21.
2. Hermann, J. C.; Marti-Arbona, R.; Fedorov, A. A.; Fedorov, E.; Almo, S. C.; Shoichet, B. K.; Raushel, F. M. Nature 2007, 448, 775.
3. Joseph-McCarthy, D.; Baber, J. C.; Feyfant, E.; Thompson, D. C.; Humblet, C. Curr Opin Drug Discov Dev 2007, 10, 264.
4. Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. J Comput Aided Mol Des 2001, 15, 411.
5. Meiler, J.; Baker, D. Proteins 2006, 65, 538.
6. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. J Comput Chem 1998, 19, 1639.
7. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. J Mol Biol 1996, 261, 470.
8. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. J Med Chem 2004, 47, 3032.
9. Perola, E.; Walters, W. P.; Charifson, P. S. Proteins 2004, 56, 235.
10. Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. J Med Chem 2006, 49, 5912.
11. Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. J Med Chem 2005, 48, 962.
12. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Proteins 2004, 57, 225.
13. Bissantz, C.; Bernard, P.; Hibert, M.; Rognan, D. Proteins 2003, 50, 5.
14. Enyedy, I. J.; Ling, Y.; Nacro, K.; Tomita, Y.; Wu, X.; Cao, Y.; Guo, R.; Li, B.; Zhu, X.; Huang, Y.; Long, Y. Q.; Roller, P. P.; Yang, D.; Wang, S. J Med Chem 2001, 44, 4313.
15. Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. J Med Chem 2005, 48, 5448.
16. McGovern, S. L.; Shoichet, B. K. J Med Chem 2003, 46, 2895.
17. Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. J Med Chem 2004, 47, 45.
18. Kim, R. G.; Skolnick, J. J Comput Chem doi: 10.1002/jcc.20893.
19. Zhang, Y.; Devries, M. E.; Skolnick, J. PLoS Comput Biol 2006, 2, e13.
20. Kryshtafovych, A.; Venclovas, C.; Fidelis, K.; Moult, J. Proteins 2005, 61(Suppl 7), 225.
21. Moult, J.; Fidelis, K.; Rost, B.; Hubbard, T.; Tramontano, A. Proteins 2005, 61(Suppl 7), 3.
22. Brylinski, M.; Skolnick, J. Proc Natl Acad Sci USA 2008, 105, 129.
23. Huang, S. Y.; Zou, X. Proteins 2007, 66, 399.
24. Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. J Med Chem 2004, 47, 5076.
25. Zhao, Y.; Sanner, M. F. Proteins 2007, 68, 726.
26. Vakser, I. A. Biopolymers 1996, 39, 455.
27. Wojciechowski, M.; Skolnick, J. J Comput Chem 2002, 23, 189.
28. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
29. Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M. Bioinformatics 1999, 15, 327.
30. Willett, P.; Winterman, V. A. Quant Struct Act Relat 1986, 5, 18.
31. Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., III. J Comput Aided Mol Des 2003, 17, 755.
32. Chen, H. M.; Liu, B. F.; Huang, H. L.; Hwang, S. F.; Ho, S. Y. J Comput Chem 2007, 28, 612.
33. Taufer, M.; Crowley, M.; Price, D. J.; Chien, A. A.; Brooks, C. L. Concurrency Comp: Practice Exp 2005, 17, 1627.
34. Skolnick, J.; Kihara, D. Proteins 2001, 42, 319.
35. Skolnick, J.; Kihara, D.; Zhang, Y. Proteins 2004, 56, 502.
36. Zhang, Y.; Arakaki, A. K.; Skolnick, J. Proteins 2005, 61 (Suppl 7), 91.
37. Zhang, Y.; Skolnick, J. Biophys J 2004, 87, 2647.
38. Zhang, Y.; Skolnick, J. Proc Natl Acad Sci USA 2004, 101, 7594.
39. Cavallo, L.; Kleinjung, J.; Fraternali, F. Nucleic Acids Res 2003, 31, 3364.
40. Gohlke, H.; Hendlich, M.; Klebe, G. J Mol Biol 2000, 295, 337.
41. Wodak, S. J.; Janin, J. J Mol Biol 1978, 124, 323.
42. Wodak, S. J.; Janin, J. Proc Natl Acad Sci USA 1980, 77, 1736.

43. Bäck, T.; Hoffmeister, F.; Schwefel, H. P. In Proceedings of the 4th International Conference on Genetic Algorithms; Belew, R. K.; Booker, L. B., Eds.; Morgan Kaufmann: San Diego, CA, 1991; pp. 2–9.
44. Bäck, T.; Schwefel, H. P. Evol Comp 1993, 1, 1.
45. Zhang, Y.; Kolinski, A.; Skolnick, J. Biophys J 2003, 85, 1145.
46. James, F. MINUIT, Reference Manual, Version 94.1; CERN: Geneva, Switzerland, 1998.
47. Knuth, D. E. In The Art of Computer Programming: Seminumerical Algorithms; Addison-Wesley: Reading, Massachusetts, 1997; pp. 135–136.
48. Lorber, D. M.; Shoichet, B. K. Protein Sci 1998, 7, 938.
49. Goodsell, D. S.; Morris, G. M.; Olson, A. J. J Mol Recognit 1996, 9, 1.
50. Nelder, J. A.; Mead, R. Comput J 1964, 7, 308.
51. Fukunishi, H.; Watanabe, O.; Takada, S. J Chem Phys 2002, 116, 9058.
52. Sugita, Y.; Kitao, A.; Okamoto, Y. J Chem Phys 2000, 113, 6042.
53. Zhang, Y.; Skolnick, J. J Chem Phys 2001, 115, 5027.
54. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. R.; Cheatham, T. E.; DeBolt, I. S.; Ferguson, D.; Seibel, G.; Kollman, P. A. Comp Phys Commun 1995, 91, 1.
55. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. J Comput Chem 2003, 24, 1999.
56. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. J Comput Chem 2004, 25, 1157.
57. Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. J Chem Inf Model 2006, 46, 991.
58. Gasteiger, J.; Marsili, M. Tetrahedron Lett 1978, 34, 3181.
59. Murray, C. W.; Baxter, C. A.; Frenkel, A. D. J Comput Aided Mol Des 1999, 13, 547.
60. Bindewald, E.; Skolnick, J. J Comput Chem 2005, 26, 374.
61. Schafferhans, A.; Klebe, G. J Mol Biol 2001, 307, 407.
62. Vakser, I. A. Protein Eng 1995, 8, 371.
63. Fraternali, F.; Cavallo, L. Nucleic Acids Res 2002, 30, 2950.
64. Lee, B.; Richards, F. M. J Mol Biol 1971, 55, 379.
65. Zhang, Y.; Skolnick, J. Nucleic Acids Res 2005, 33, 2302.