Flexible Protein–Ligand Docking by Global Energy Optimization in Internal Coordinates

Maxim Totrov and Ruben Abagyan*
The Skirball Institute of Biomolecular Medicine, Biochemistry Department of New York University Medical College, New York, New York

ABSTRACT **Eight protein–ligand complexes were simulated by using global optimization of a complex energy function, including solvation, surface tension, and side-chain entropy in the internal coordinate space of the flexible ligand and the receptor side chains [Abagyan, R.A., Totrov, M.M. J. Mol. Biol. 235: 983–1002, 1994]. The procedure uses two types of efficient random moves, a pseudobrownian positional move [Abagyan, R.A., Totrov, M.M., Kuznetsov, D.A. J. Comp. Chem. 15:488–506, 1994] and a Biased-Probability multitorsion move [Abagyan, R.A., Totrov, M.M. J. Mol. Biol. 235: 983–1002, 1994], each accompanied by full local energy minimization. The best docking solutions were further ranked according to the interaction energy, which included intramolecular deformation energies of both receptor and ligand, the interaction energy, surface tension, side-chain entropic contribution, and an electrostatic term evaluated as a boundary element solution of the Poisson equation with the molecular surface as a dielectric boundary. The geometrical accuracy of the docking solutions ranged from 30% to 70% according to the relative displacement error measure at a 1.5 Å scale. Similar results were obtained when the explicit receptor atoms were replaced with a grid potential. Proteins, Suppl. 1:215–220, 1997.**

Key words: molecular recognition; ligand binding; flexible docking

INTRODUCTION

Theoretical prediction of the association of flexible ligands with protein receptors requires efficient sampling of the conformational space of a flexible ligand, a sufficiently accurate energy function and an efficient way to account for the receptor flexibility (reviewed in Refs. 1–4). Flexible docking schemes can be based on incremental construction of the docked conformation from separately docked rigid pieces[5–7] or on a limited discrete set of ligand conformations.[8,9] Considering the entire continuously flexible ligand molecule with molecular dynamics can be used to sample the conformational space of relatively small compounds.[1,10–12] Monte Carlo methods allow to increase the sampling efficiency by making larger conformational rearrangements.[13,14] Typically sampling is performed by making random changes of one angle by a random value.[14,15] Caflish et al.[15] used Monte Carlo combined with local energy minimization after each random change of a ligand torsion (receptor assumed to be rigid), as suggested by Li and Scheraga[16] for peptide structure prediction.

The continuous flexible docking procedure in internal coordinate space of both the ligand and the side chains of protein receptor was first introduced in 1994 and applied to predict the association of two helices.[17] This method attempted to globally optimize a rather complex energy function simultaneously with ligand and receptor rearrangements (each followed by local energy minimization) rather then refine a set of solutions generated with rigid ligand molecules and with a simpler energy function. Later, the side-chain entropy and the MIMEL approximation of the solvation energy were added to the globally optimized objective function,[18] these terms being evaluated after each local minimization as outlined in a “double energy” scheme.[17] The ICM docking procedure correctly docked lysozyme and its antibody in full atom representations with flexible side-chain association and reached a discrimination of 19 kcal/mole between the correct lowest energy conformation and the next false solution.[19] Later, the association of β-lactamase and its inhibitor[2,20] were correctly predicted with a similar energy discrimination gap, this time under blind prediction conditions.

In this article, we apply the ICM docking method to small flexible ligands that are globally energy-optimized together with the active site side chains by using the double energy scheme. Additionally, we use an accurate boundary element solution of the Poisson equation to evaluate the 30 best docking solutions for each compound.

*Correspondence to: Dr. Ruben Abagyan, The Skirball Institute of Biomolecular Medicine, Biochemistry Department of NYU MC, 540 First Avenue, New York, NY 10016.
E-mail: abagyan@earth.med.nyu.edu

Received 6 May 1997; Accepted 1 August 1997

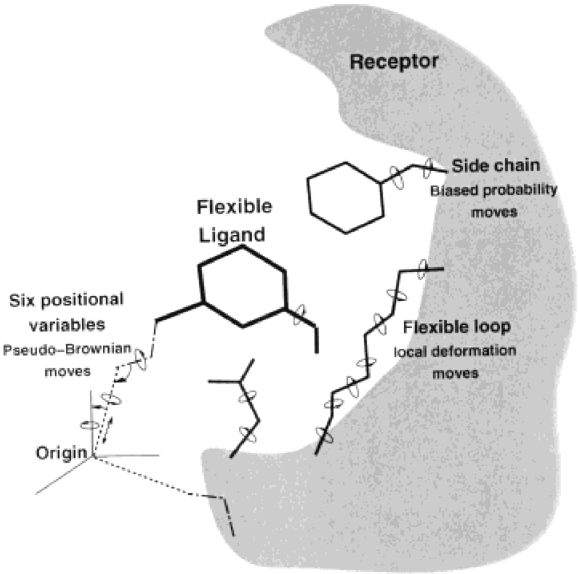

Fig. 1. ICM docking setup with flexible ligand and explicit flexible receptor. Most of the receptor variables are fixed.

METHOD

The ICM method describes both the relative positions of two molecules and their conformations by a uniform set of internal variables. Any subset of internal variables can be subjected to local or global energy minimization procedures. Docking of flexible ligands into a flexible receptor requires three groups of free variables: positional variables of the ligand, intramolecular variables of the ligand and the torsion angles of the active site side chains (Fig. 1). Flexible loops can also be sampled simultaneously with the ligand (e.g., in antibodies). All the other variables are fixed to accelerate energy evaluation and sampling. The global minimization procedure involves a random change of the internal variables followed by local energy minimization and selection by the Metropolis criterion. Pseudobrownian random moves change the position of the ligand molecule as a whole with a certain amplitude (here we used 2 Å), as well as randomly rotate it around its center of gravity by an angle close to the translation amplitude over the radius of gyration.[17] Internal torsion angles of the ligand are randomly changed one at a time, with an amplitude of 180°. Coupled groups of receptor side-chain torsion angles were sampled with biased probability moves.[18]

Once the set of free variables is defined, the ICM global energy optimization was performed from multiple starting points. The number of starting points depends on the size of a ligand and here we used six random starting points. The energy optimization routine consisted of the following iterative steps[17]:

1. Make a random conformational change of three possible types (Figure 1: loops were not considered here).
2. Perform local energy optimization of the vacuum ECEPP3 energy[21] with a distant-dependent dielectric constant $\epsilon = 4r$.
3. Evaluate surface-based solvation energy and entropic contribution from the receptor side chains and add it to the ECEPP3 energy.
4. Apply Metropolis et al.[22] selection criterion at a certain temperature T and make another step.

Geometrically different (as evaluated by the root-mean-square displacement [RMSD] of the ligand atoms) and low-energy conformations were accumulated in the conformational stack as described in Ref. 23. At the end of simulations, the conformational stacks were merged and the 30 best energy conformations were ranked with a more rigorous evaluation of the electrostatic free energy. Electrostatic free energy was calculated by a numerical solution to the Poisson equation by using the boundary element algorithm.[24] Our implementation of the boundary element algorithm uses the accurate analytical molecular surface build by the fast contour buildup method.[25] The ECEPP charges[26] were used for the protein atoms; charges of the ligand atoms were calculated with the Gaussian program.[27]

RESULTS

The techniques developed were tested on the docking prediction targets in the CASP-2 (critical assessment of structure prediction techniques) protein structure prediction contest. For the docking simulations, eight ligand–protein complexes were proposed (Table I). We made predictions for all eight complexes. For each of the targets, the coordinates for a complex of the protein with some other ligand(s) were found in the protein structure database (PDB), which allowed us to establish the approximate locations of the binding sites as a first step of the prediction. Next, three-dimensional models of the ligands had to be built. The chemical structures of the ligands were available in the form of the connectivity tables. Since the experimental 3D coordinates for the ligands were not available, we built the models in the ICM program[28] from the fragments of the compounds found in the Cambridge structural database (CSD)[29] with known three-dimensional (3D) structures. To find those, CSD was searched for the compounds with chemical structures similar to the chemical structure of the ligand. The third step was the assignment of partial charges to the individual

Fig. 2. Predicted docking conformations are shown in red and conformations determined by x-ray crystallography are shown in green. Analytical molecular surface of protein receptors was generated with the contour-buildup method[25] as implemented in the ICM program.[28]

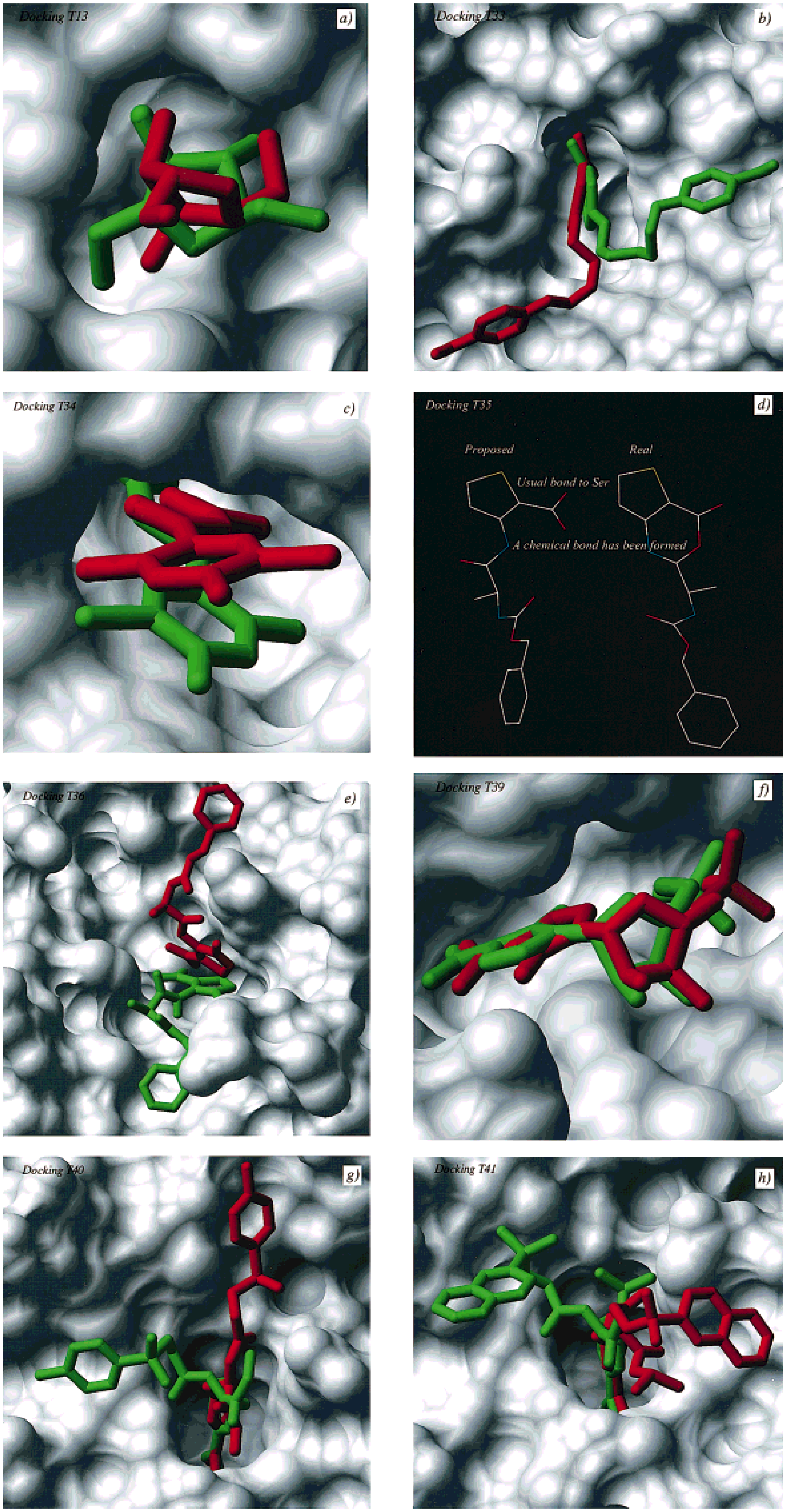

Fig. 2 (legend on preceding page).

TABLE I. Results for the Docking of Eight Ligands to Their Receptors Evaluated by All Heavy Atom RMSD and the Relative Displacement Error*

Target	Ligand	Receptor (PDB template code)	Site	Restraints	RMSD[†]	Fraction correct (%)[‡]
t13	Methyl alpha-D-arabinofuranoside	Concanavalin A (5cna)	Pocket	No	3.5	49.6
t33	Pentamidine	Pancreatic trypsin (2tbs)	Pocket	Tip[§]	9.27	51.7
t34	Amiloride	Pancreatic trypsin (2tbs)	Pocket	Tip	4.2	48.1
t35[¶]	SBA[‖]	Pancreatic elastase (1inc)	Covalent	Chem. bond	10.6	31.2
t36	SBB[‖]	Pancreatic elastase (1inc)	Covalent	Chem. bond	10.7	35.6
t39	Aica-riboside phosphate	Fructose bis-phosphotase (1fpd)	Pocket	No	1.8	70.1
t40	INH[‖]	Pancreatic trypsin (2tbs)	Pocket	Tip	6.7	49.7
t41	INI[‖]	Pancreatic trypsin (2tbs)	Pocket	Tip	7.8	44.6

*Runtimes for simulations with fully flexible receptor side chains and ligand varied from 5 to 15 hours. From Ref. 30, with permission.

[†] Cartesian RMSD was calculated for all ligand heavy atoms with the receptor models superimposed.

[‡] Fraction correct, or 100% relative displacement error[30] is calculated for all N heavy atoms of a ligand by using this formula: 100% $(L/N)\,\Sigma(L + D_{ii})^{-1}$, where D_{ii} is the deviation of the model atom i from the corresponding atom in the reference structure, and the scale parameter $L = 1.5$ Å.

[§] Tip indicates a distance restraint imposed on the carbon atom of the guanyl group.

[¶] Predictions were misled by the wrong chemical structure of the t35 ligand suggested for predictions.

[‖] We use abbreviations suggested by the CASP2 organizers. SMILES strings of these compounds can be found at http://PredictionCenter.llnl.gov/casp2/targets.html.

atoms of the ligand, which were needed for the subsequent energy calculations. This was done with the help of the quantum-chemical program package Gaussian.[27] A CNDO hamiltonian was used to obtain the ligand atomic charges that are the most consistent with the standard ECEPP3 charges used for the protein molecule. The fourth and central step of the procedure was global energy optimization of the ligand–protein complex. The ligand was placed in the vicinity of the binding site of the protein, and the system was subjected to the ICM docking procedure described above. During the procedure, torsion angles of the ligand and of the protein side chains in a 7-Å vicinity of the binding site were randomly changed. Each random change was followed by up to 100 steps of local conjugate–gradient minimization. New conformations were accepted or rejected according to the Metropolis criterion by using the temperature of 600 K. Several independent Monte Carlo runs of 300,000 energy evaluations were done for each ligand to ensure the convergence of the optimization.

In the last step, putative solutions accumulated in the conformational stacks were reevaluated by using a more precise solvation electrostatic energy approximation based on the boundary element solution of the Poisson equation. The solution that scored best in this energy approximation was taken as the answer and submitted. When the experimental structures became available, we were able to check the predictions. In most cases, the parts of the ligand inside the binding center were predicted with good accuracy. Relatively large deviations occurred only for atoms that were outside of the binding center. We used relative displacement error[30] (RDE) as well as RMSD to evaluate our solutions. The results are summarized on the Table I. The high RMSD values for several complexes are somewhat misleading, because in fact only about half of the atoms of these ligands have large deviations, as the RDE measure correctly suggests. In the case of target 35, elastase/elastase inhibitor, the actual structure of the ligand has undergone chemical changes that were impossible to predict.

After the CASP2 meeting, an attempt was made to predict the same complexes with the explicit receptor replaced by the grid potential. Four types of potentials were precalculated: the van der Waals potential for a hydrogen atom probe, the van der Waals potential for a heavy atom probe (generic carbon of 1.7 Å radius was used), an electrostatic potential from the receptor atoms, and the hydrogen-bonding potential calculated as spherical gaussians centered at the ideal putative donor and/or acceptor sites. Simulations took only about 5 minutes per compound and results similar to the results of the full-atom simulations were obtained. While this approach does not allow the explicit receptor flexibility, it might be preferred when the calculation speed is crucial, for example, in database scanning.

DISCUSSION

Accurate prediction of protein–ligand association requires inclusion of the ligand flexibility and protein surface flexibility in the docking procedure as well as precise evaluation of the interaction energy. The docking technique described here allows continuous and efficient sampling of internal torsions of the ligand and receptor side chains as well as sampling of the variables that define the mutual orientation of the receptor and ligand within the same Monte Carlo-based global optimization framework. The pseudobrownian random rearrangements are different from other schemes of random positional sampling, such as local minimizations from multiple starting points,[13] or random translations and rotations (e.g., Ref. 15), and has the advantage of imitating local ligand rearrangements. The proposed biased-probability sampling method[18] for all surface side chains in the vicinity of the active site is much more efficient than either discrete sampling[9] or changing one side-chain torsion angle at a time.[15,31] This method also can be used to sample the ligand if its conformational preferences in the form of continuous distributions are preliminary generated or evaluated by using the database.

However, even if the global optimization of the ligand/side-chain subsystem is fast and convergent, deformations of the backbone may still be crucial to docking with detailed atomic models. An adequate simulation of the backbone flexibility simultaneously with the ligand docking is still out of reach for the current computational approach. To some extent, softening the potential (e.g., Refs. 32 and 33) or using an approximate grid potential,[12,14,34] which is less steep than the realistic van der Waals repulsion, may be a practical way of overcoming this problem. Furthermore, simulations with the grid potential are much faster than the explicit flexible docking simulations and can be used for scanning large databases. Clearly, the choice between the explicit receptor model or the grid potential model depends on the docking problem and the available computer time.

In this work the receptor side chains were sampled together with the ligand. Previously we found that for protein–protein docking this approach leads to a better discrimination between the correct and incorrect solutions.[19,20] It is unclear, however, that in this work this flexibility was essential.

The energy function optimized with the procedure included a detailed vacuum energy complemented with the surface-based solvation and side-chain entropies. Since we intended to compare different conformations of the same ligand rather than binding affinities of different ligands, we did not estimate the ligand entropy loss.[35] However, inclusion of the side-chain entropies into global optimization[18] may be essential for discrimination between putative binding sites, since these contributions can reach 2 kcal/mol per residue.

Numerical solutions of the Poisson or the Poisson-Boltzmann equations (reviewed in Refs. 36 and 37) provide the most accurate representation of the electrostatic solvation component of the ligand-binding energy and can be added to the molecular mechanical force field to rank the docking solution.[38–40] We ranked the 30 best solutions by using a more accurate evaluation of the electrostatic free energy calculated with the boundary element algorithm.[24,41,42] However, even these energies could not identify the correct positions of the solvent exposed parts of the long ligands. Technically, explicit water molecules could have been sampled together with the ligand, but explicit solvation can only be adequately considered within the framework of molecular dynamics.

Although the smaller compounds were predicted reasonably well, the relatively poor quality of prediction for the longer ligands suggests that the part of the ligand outside the binding pocket might not have a strong preference toward any one conformation. Presumably, the experimental structure in these cases is defined by a fine balance of energy terms, which is still beyond the accuracy of the available energy approximations, or even perhaps by the crystallographic packing. The presence of many alternative configurations for such parts of the ligand molecule among the low-energy conformations accumulated during the simulations also suggests that the energy minimum for them is less well defined. Some of these alternative configurations are closer to the native conformation, but also have significantly higher energy then the lowest-energy conformation, suggesting that sampling of the conformational space of the ligand is sufficient. Further improvement in the free energy evaluation is necessary to achieve better docking precision for the weakly bound groups.

REFERENCES

1. Rosenfeld, R., Vajda, S., DeLisi, C. Flexible docking and design. Annu. Rev. Biophys. Biomol. Struct. 24:677–700, 1995.
2. Janin, J. Protein–protein recognition. Prog. Biophys. Mol. Biol. 64:145–166, 1995.
3. Bamborough, P., Cohen, F.E. Modeling protein–ligand complexes. Curr. Opin. Struct. Biol. 6:236–241, 1996.
4. Lengauer, T., Rarey, M. Computational methods for biomolecular docking. Curr. Opin. Struct. Biol. 6:402–406, 1996.
5. Gulukota, K., Vajda, S., Delisi, C. Peptide docking using dynamic programming. J. Comp. Chem. 17:418–428, 1996.
6. Welch, W., Ruppert, J., Jain, A.J. Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. Chem. Biol. 3:449–462, 1996.
7. Rarey, M., Kramer, B., Lengauer, T., Klebe, G. A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. 261:470–489, 1996.
8. Kearsley, S.K., Underwood, D.J., Sheridan, R.P., Miller, M.D. Flexibases: A way to enhance the use of molecular docking methods. J. Comput. Aided. Mol. Design 8:565–82, 1994.
9. Leach, A.R. Ligand docking to proteins with discrete side-chain flexibility. J. Mol. Biol. 235:345–356, 1994.
10. Miranker, A., Karplus, M. Functionality maps of binding sites: A multiple copy simultaneous search method. Proteins 11:29–34, 1991.
11. DiNola, A., Raccatano, D., Berendsen, H. Molecular dynamics simulation of the docking of substrates to proteins. Proteins 19:174–182, 1994.
12. Luty, B.A., Wasserman, Z.R., Stouten, P.F.W., Hodge, C.N., Zacharias, M., McCammon, J.A. A molecular mechanics/grid method for evaluation of ligand-receptor interactions. J. Comput. Chem. 16:454–464, 1995.
13. Hart, T.N., Read, R.J. A multiple-start Monte Carlo docking model. Proteins 13:206–222, 1992.
14. Goodsell A.S., Olson A.J. Automated docking of substrates to proteins by simulated annealing. Proteins 8:195–202, 1990.
15. Caflisch, A., Niederer, P., Anliker, M. Monte Carlo docking of oligopeptides to proteins. Proteins 13:223–230, 1992.
16. Li, Z., Scheraga, H.A. Monte Carlo-minimization approach

to the multiple-minima problem in protein folding. Proc. Natl. Acad. Sci. U.S.A. 84:6611–6615, 1987.

17. Abagyan, R.A., Totrov, M.M., Kuznetsov, D.A. ICM: a new method for structure modeling and design: Applications to docking and structure prediction from the distorted native conformation. J. Comp. Chem. 15:488–506, 1994.
18. Abagyan, R.A., Totrov, M.M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J. Mol. Biol. 235:983–1002, 1994.
19. Totrov, M.M., Abagyan, R.A. Detailed ab initio prediction of lysozyme–antibody complex with 1.6A accuracy. Nature Struct. Biol. 1:259–263, 1994.
20. Strynadka, N.C.J., Eisenstein, M., Katchalski-Katzir, E., et al. Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. Nature Struct. Biol. 3:233–239, 1996.
21. Nemethy, G., Gibson, K.D., Palmer, K.A., et al. Energy parameters in polypeptides. 10. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J. Phys. Chem. 96:6472–6484, 1992.
22. Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, N.M., Teller, A.H., Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092, 1953.
23. Abagyan, R.A., Argos, P. Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. J. Mol. Biol., 225:519–532, 1992.
24. Zauhar, R.J., Morgan, R.S. A new method for computing the macromolecular electric potential. J. Mol. Biol. 186:815–820, 1985.
25. Totrov, M.M., Abagyan, R.A. The contour–buildup algorithm to calculate the analytical molecular surface. J. Struct. Biol. 116:138–143, 1996.
26. Momany, F.A., McGuire, R.F., Burgess, A.W., Scheraga, H.A. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. J. Phys. Chem. 79:2361–2381, 1975.
27. Frisch, M.J., Trucks, G.W., Schlegel, H.B., et al. Gaussian 94 (Revision A.1). A..Gaussian 94 (Revision A.1). Gaussian, Inc., Pittsburgh, PA, 1995.
28. ICM software manual. Version 2.5, Molsoft, LLC, 1996.
29. Allen, F.H., Kennard, O. 3D Search and research using the Cambridge structural database. Chem. Design Autom. News 8:1, 31–37, 1993.
30. Abagyan, R.A., Totrov, M. Contact area difference (CAD): A robust measure to evaluate accuracy of protein models. J. Mol. Biol. 268:678–685, 1997.
31. Hartl, F.U., Martin, J. Protein folding in the cell: The role of molecular chaperones Hsp70 and Hsp80. Annu. Rev. Biophys. Biomol. Struct. 21:293–322, 1992.
32. Jiang, F., Kim, S.-H. “Soft docking”: Matching of molecular surface cubes. J. Mol. Biol. 219:79–102, 1991.
33. Bacon, D.J., Moult, J. Docking by least-squares fitting of molecular surface patterns. J. Mol. Biol. 225:849–858, 1992.
34. Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J. Med. Chem. 28:849–875, 1985.
35. Novotny, J., Bruccoleri, R.E., Saul, F.A. On the attribution of binding-energy in antigen–antibody complexes Mcpc-603. Biochemistry 28:4735–4749, 1989.
36. Davis, M.E., McCammon, J.A. Electrostatics in biomolecular structure and dynamics. Chem. Rev. 90:509–521, 1990.
37. Honig, B., Nicholls, A. Classical electrostatics in biology and chemistry. Science 268:1144–1149, 1995.
38. Shoichet, B.K., Kuntz, I.D. Protein docking and complementarity. J. Mol. Biol. 221:327–346, 1991.
39. Zacharias, M., Luty, B.A., Davis, M.E., McCammon, J.A. Combined conformational search and finite-difference Poisson-Boltzmann approach for flexible docking: Application to an operator mutation in the lambda repressor-operator complex. J. Mol. Biol. 238:455–465, 1994.
40. Jackson, R.M., Sternberg, M.J.E. A continuum model for protein–protein interactions: Application to the docking problem. J. Mol. Biol. 250:258–275, 1995.
41. Juffer, A.H., Botta, E.F.F., van Keulen, B.A.M., van der Ploeg, A., Berendsen, H.J.C. The electric potential of a macromolecule in a solvent: A fundamental approach. J. Comput. Phys. 97:144–171, 1991.
42. Bharadwaj, A., Windemuth, A., Sridharan, S., Honig, B., Nicholls, A. The fast multipole boundary element method for molecular electrostatics: An optimal approach for large system. J. Comp. Chem. 16:898–910, 1995.