PROTEINS: Structure, Function, and Genetics 8:195–202 (1990)

Automated Docking of Substrates to Proteins by Simulated Annealing

David S. Goodsell and Arthur J. Olson
Department of Molecular Biology, Research Institute of Scripps Clinic, La Jolla, California 92037

ABSTRACT **The Metropolis technique of conformation searching is combined with rapid energy evaluation using molecular affinity potentials to give an efficient procedure for docking substrates to macromolecules of known structure. The procedure works well on a number of crystallographic test systems, functionally reproducing the observed binding modes of several substrates.**

Key words: simulated annealing, computer-aided drug design, substrate docking

INTRODUCTION

Computer-aided drug design (CADD) is potentially one of the most useful medical and industrial applications of macromolecular crystallography. A major stumbling block in the development of rational drug design schemes has been the lack of a simple procedure for docking potential drug candidates to a target macromolecule. In any docking scheme, two conflicting requirements must be balanced: the desire for a robust computational procedure and the desire to keep the computational demands at a reasonable level. An ideal docking procedure would find the globally most favorable site of binding of the trial molecule, with its many degrees of conformational freedom, using the most realistic energy evaluation model available. This ideal procedure must at the same time stay at a level of computational complexity which most structural researchers are willing to deal with, for instance, with time and resource commitments similar to a crystallographic refinement. To conform to these computational limits, reported techniques resort to some means of simplifying the docking process. The manual docking methods currently in widespread use,[1–4] because of their need for intensive user interactivity, only explore a limited subset of conformations, but allow the use of sophisticated energy evaluation methods. Automated methods, using exhaustive search[5–7] or distance geometry methods[8,9] allow the exploration of a large area of conformational space, but require the use of simplified energetic models to keep the computational needs within bound. We describe a procedure combining a Monte Carlo search technique with rapid energy evaluation using molecular affinity potentials. It combines the advantages of both of these methods—robust energy evaluation, large search space and reasonable computational cost. For the specific task of docking a flexible substrate to the binding site of a static protein, the technique is powerful. The researcher's input is minimized; he need only specify the region of the protein containing the binding site and place the substrate, in an arbitrary conformation and orientation, somewhere in the space near the binding site. The procedure will take these relatively unbiased starting conditions and produce a docked conformation.

METHODS

The Metropolis method,[10,11] also known as simulated annealing, is used for conformational and positional searching. The trial molecule performs a random walk in the space around the target protein; the protein is static throughout the simulation. At each time step of the simulation, a small random displacement in each of the degrees of freedom of the substrate is performed: in rigid body translation and rotation and in each torsion angle. At the new position and conformation, the energy of interaction is evaluated, as described below, and the new energy is compared to the energy of the previous step. If the new energy is lower, the step is immediately accepted. If the new energy is higher, the result is treated probabilistically, with the result dependent on a user defined temperature (T). The probability of acceptance is given by

$$P(\Delta E) = \exp(-\Delta E/k_B T)$$

where ΔE is the difference in energy and k_B is Boltzmann's constant. High temperatures will accept nearly all steps, and lower temperatures provide more stringent conformational selection.

The binding conformation is successively refined as the simulation procedes by control of the temperature. The simulation is broken up into a number of cycles, each at a constant temperature and composed of a large number of individual steps. The tempera-

Received December 14, 1989; revision accepted April 16, 1990.

Address reprint requests to Dr. Arthur J. Olson, Department of Molecular Biology, Research Institute of Scripps Clinic, 10666 North Torrey Pines Road, La Jolla, CA 92037.

ture is reduced at the beginning of each cycle, such that

$$T_i = gT_{i-1}$$

where T_i is the temperature of step i and g is a constant between 0 and 1. Each cycle begins with the trial molecule in the conformation of lowest energy found in the previous cycle. This protocol has shown better convergence than the previously reported method of carrying the last conformation of each cycle over to the next.[11]

At each step of the simulation, the energy of interaction of substrate and protein is evaluated using molecular affinity potentials, as described by Goodford.[12] The protein is embedded in a three-dimensional grid and a probe atom is sequentially placed at each grid point. The energy of interaction of this single atom with the protein is assigned to the grid point. An affinity grid is calculated for each type of atom in a typical substrate—for carbon, oxygen, nitrogen, and hydrogen—as well as a grid of electrostatic potential using a point charge of +1 as the probe.

During the simulation, these grids are used as look-up tables to evaluate the interaction energy rapidly. For each atom in the substrate, the interaction energy is found by trilinear interpolation of the affinity values of the eight grid points surrounding the atom position. The electrostatic interaction energy is evaluated similarly, by interpolating the value of the electrostatic potential and multiplying by the charge on the atom (the electrostatic term is evaluated separately to allow finer control of the substrate atomic charges). The calculation is proportional only to the number of atoms in the substrate, not to any function of the number of protein atoms.

The separation of the molecular affinity grids from the docking simulation provides a useful modularity to the procedure. The grid calculations are performed only once for the target macromolecule, so any level of sophistication may be used: from constant dielectrics to finite difference methods and from 6–12 functions to distributions based on observed binding sites.[13]

RESULTS

Simulation Parameters

Both the Monte Carlo simulation parameters and the weights applied to the electrostatic, hydrogen bond, and repulsion/dispersion terms of the affinity potentials were determined empirically in tests of phosphocholine binding to the immunoglobulin McPC 603. Phosphocholine binding is an ideal test system, as the crystallographically observed binding mode depends on both steric and electrostatic interactions.[14] The choline end of the substrate binds deep in a triangular pocket in an area of high carbon affinity potential (Fig. 1A). A glutamic acid at the base of this pocket forms a favorable electrostatic interaction with the choline amine. The phosphate end is not sterically constrained, forming hydrogen bonds and charge interactions with tyrosine and arginine groups protruding from the antibody (Fig. 1B and C). A search procedure that uses only shape or only charge complementarity would not properly dock this substrate.

For these tests, protein and substrate coordinates were taken from the Brookhaven Protein Data Bank (PDB) listing 2MCP. The crystallographic substrate coordinates were rotated by 180° about the crystallographic y axis, flipping the molecule end-for-end, and translated from 5 to 15 Å away from the binding site to give the substrate starting coordinates. The two central torsion angles were allowed to rotate and were started 90° away from their crystallographic orientations. The outer two torsion angles were held fixed in their crystallographic staggered conformations. A set of simulation parameters was obtained that yielded bound conformations reproducing the observed mode of phosphocholine binding. The parameters are all physically reasonable (described in Fig. 1 and Table I) and were used unchanged for the other three systems.

Docking of Phosphocholine to McPC 603

We performed a set of 10 docking simulations on the antibody–antigen system, using the final simulation parameters and starting the substrate 5 Å from the active site. Seven unique, but similar, conformations were obtained, shown superimposed on the crystallographically observed conformation[14] in Figure 2A. The top three ranked conformations are shown in Figure 2B and the interaction energies and rms differences from the crystallographic conformation may be found in Table I. All of the docked conformations place the charged phosphate and choline ends in their proper binding sites, but differ in details. The top ranked conformation, which was obtained independently four times, functionally reproduces the crystallographic solution.

Docking of *N*-Formyl-tryptophan to Chymotrypsin

Chymotrypsin provided a second system where binding specificity is dependent on both electrostatic and steric factors. The virtual substrate *N*-formyltryptophan binds in the active site, with its indole moiety slotted into a hydrophobic groove and with its carboxyl hydrogen bonded to several side chain and main chain atoms on the side of the binding cleft.[15]

Protein coordinates were taken from PDB listing 4CHA and substrate coordinates were created from a GLY-TRP dipeptide extracted from the protein structure. The substrate was started 3 Å outside of the chymotrypsin active site and all torsion angles but the peptide linkage allowed to rotate. In a set of

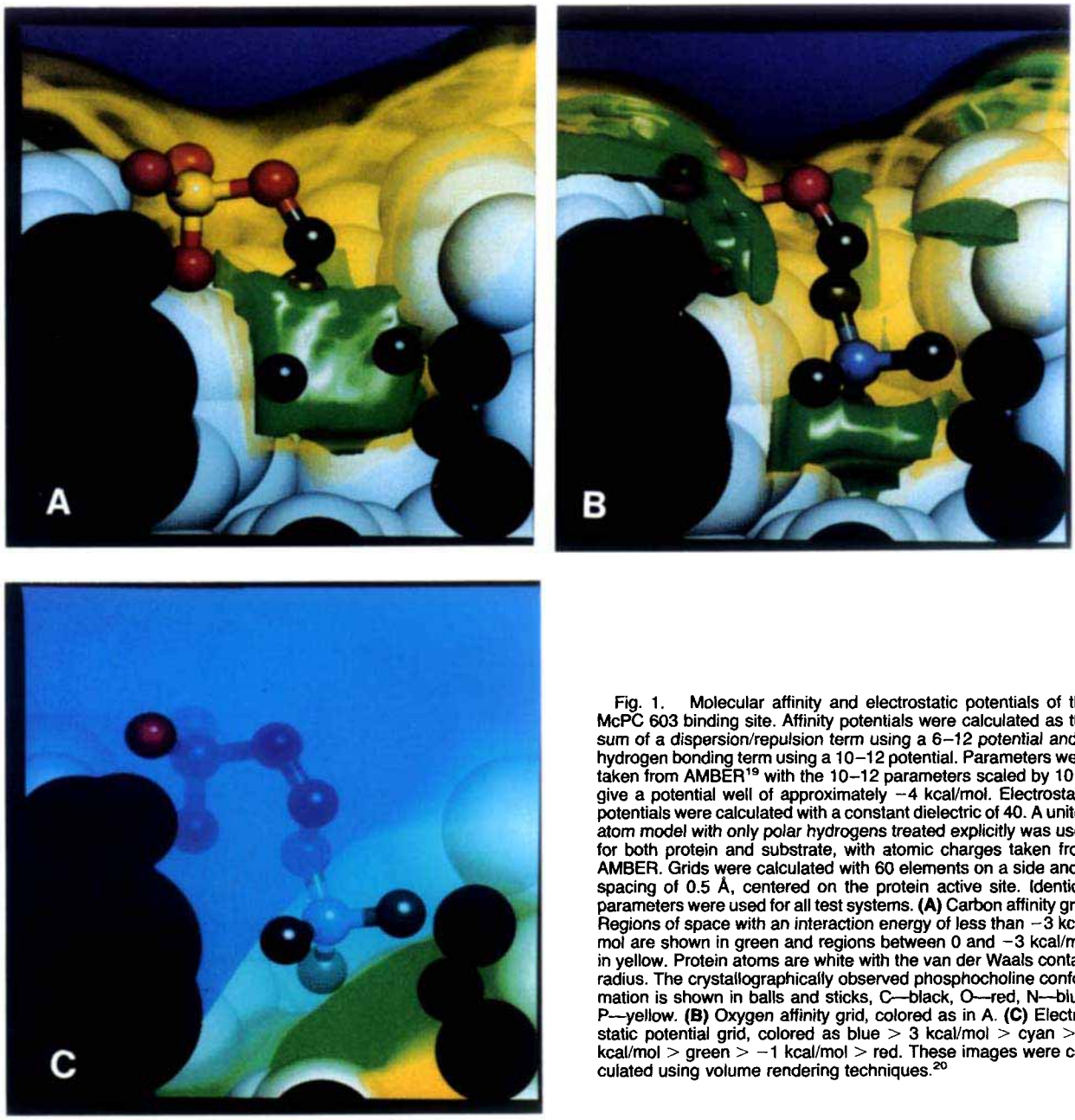


Fig. 1. Molecular affinity and electrostatic potentials of the McPC 603 binding site. Affinity potentials were calculated as the sum of a dispersion/repulsion term using a 6–12 potential and a hydrogen bonding term using a 10–12 potential. Parameters were taken from AMBER[19] with the 10–12 parameters scaled by 10 to give a potential well of approximately −4 kcal/mol. Electrostatic potentials were calculated with a constant dielectric of 40. A united atom model with only polar hydrogens treated explicitly was used for both protein and substrate, with atomic charges taken from AMBER. Grids were calculated with 60 elements on a side and a spacing of 0.5 Å, centered on the protein active site. Identical parameters were used for all test systems. **(A)** Carbon affinity grid. Regions of space with an interaction energy of less than −3 kcal/mol are shown in green and regions between 0 and −3 kcal/mol in yellow. Protein atoms are white with the van der Waals contact radius. The crystallographically observed phosphocholine conformation is shown in balls and sticks, C—black, O—red, N—blue, P—yellow. **(B)** Oxygen affinity grid, colored as in A. **(C)** Electrostatic potential grid, colored as blue > 3 kcal/mol > cyan > 1 kcal/mol > green > −1 kcal/mol > red. These images were calculated using volume rendering techniques.[20]

five simulations, the three top ranked conformations accurately conform to the crystallographic result (Fig. 2C). The indoles are all virtually identical (the rms differences from the crystallographic conformation for only indole atoms are 0.36, 0.32, and 0.38 Å) and the carboxyls occupy the same hydrogen bonding position. The position of the formyl group is not accurately placed by the docking procedure, with conformations 90 to 120° away from the crystallographic conformation, around the C_α–N bond. This conformation is not well constrained by the protein, however, as no specific interactions for the formyl carbonyl are within reach after the substrate carboxyl and indole find their prefered spots.

Binding of *N*-Acetylglucosamines to Lysozyme

Lysozyme provided a more difficult test of the method, as the binding is due mainly to steric and hydrogen bond effects. Lysozyme has a long cleft providing several sites for sugar binding. We attempted to reproduce the crystallographic modes of binding of two anomers of *N*-acetylglucosamine[16]: the α and β anomers bind with their acetyl groups in the same position, but with the sugar rings in different sites.

Protein coordinates were taken from PDB listing 6LYZ and substrate coordinates were taken from the crystal structure of the binary complex.[16] Simula-

TABLE I. Results of Simulated Annealing*

System	Rank	Number of observations†	Energy (kcal/mol)	rms diff. (Å)	Simulation time (min)‡
McPC 603: phosphocholine	1	4	−44.71	0.97	36
	2	1	−44.70	0.99	
	3	1	−44.45	1.06	
	4	1	−43.27	1.70	
	5	1	−43.06	1.76	
	6	1	−42.78	1.92	
	7	1	−41.69	2.17	
Chymotrypsin: *N*-formyltryptophan	1	1	−75.57	1.40	59
	2	1	−75.43	1.32	
	3	1	−74.62	1.46	
	4	1	−74.35	5.87	
	5	1	−72.42	2.68	
Lysozyme: α-*N*-acetylglucosamine	1	2	−46.50	4.01	49
	2	1	−46.38	4.00	
	3	2	−46.36	4.00	
	4	1	−46.18	3.96	
	5	3	−46.17	3.99	
	6	1	−44.31	5.45	
Lysozyme: β-*N*-acetylglucosamine	1	1	−47.55	0.94	49
	2	3	−47.18	1.99	
	3	2	−46.99	1.12	
	4	1	−46.87	1.95	
	5	1	−46.71	1.95	
	6	1	−46.68	1.98	
	7	1	−44.39	5.53	
Aconitase : sulfate	1	1	−52.64	4.75	12
	2	1	−52.29	4.76	
	3	2	−52.29	0.95	
	4	1	−52.28	4.76	
	5	1	−52.26	0.98	
	6	2	−52.24	0.98	
	7	1	−51.78	0.90	
	8	1	−51.27	4.94	

*Each simulation was composed of 50 cycles containing a maximum of 30,000 accepted or 30,000 rejected steps, whichever limit was reached first. Simulations were started at a very high temperature ($k_BT = 100$ kcal/mol). Applying a value of $g = 0.9$ at the end of each cycle yielded a final temperature of $k_BT = 0.5$ kcal/mol. Displacements at each step were a random fraction of 0.2 Å for translation and 5° for rigid body and torsional rotations.

†Conformations were scored as identical if within 5° in rigid body and torsional rotation angles and within 0.2 Å in translation.

‡Computation times are given per simulation, on a Convex C1 computer.

tions were started with the substrates centered between the α and β binding sites and rotated 180° from the crystallographic orientations. Two torsion angles were varied: the acetyl-sugar linkage and the hydroxymethyl-sugar bond at the 5 position. The peptide linkage and the three other hydroxyl-sugar bonds were held fixed. Ten simulations were performed with each anomer; Figures 2D and E show the top three ranked conformations of each. The β conformation was well reproduced, with only the lowest of the 10 not conforming to the crystallographic result. The α anomer prefered the β binding mode over its experimentally observed mode, with all but the least favorable conformation finding the β site. This incorrect result could be due to the use of the native lysozyme structure. In the crystallographic structure of the complex, several sidechains are observed to move upon binding. Alternatively, the simple model we use for hydrogen bonding may not be accurate enough to orient this type of substrate, which lacks the strong electrostatic character of the other test systems.

The Active Site of Aconitase

We chose the recently solved structure of aconitase[17] as the final test system, as it contains a bound sulfate, as well as providing an interesting system for trying a molecule not yet experimentally observed: its substrate, citrate. The active site of aconitase is unusual, being completely buried in the protein. It is lined with a number of arginine and histidine groups and contains a [4Fe–4S] cluster at one end, providing many sites of binding for the three carboxyl groups of citrate.

Protein coordinates were kindly provided by David Stout. Simulations of sulfate and citrate were

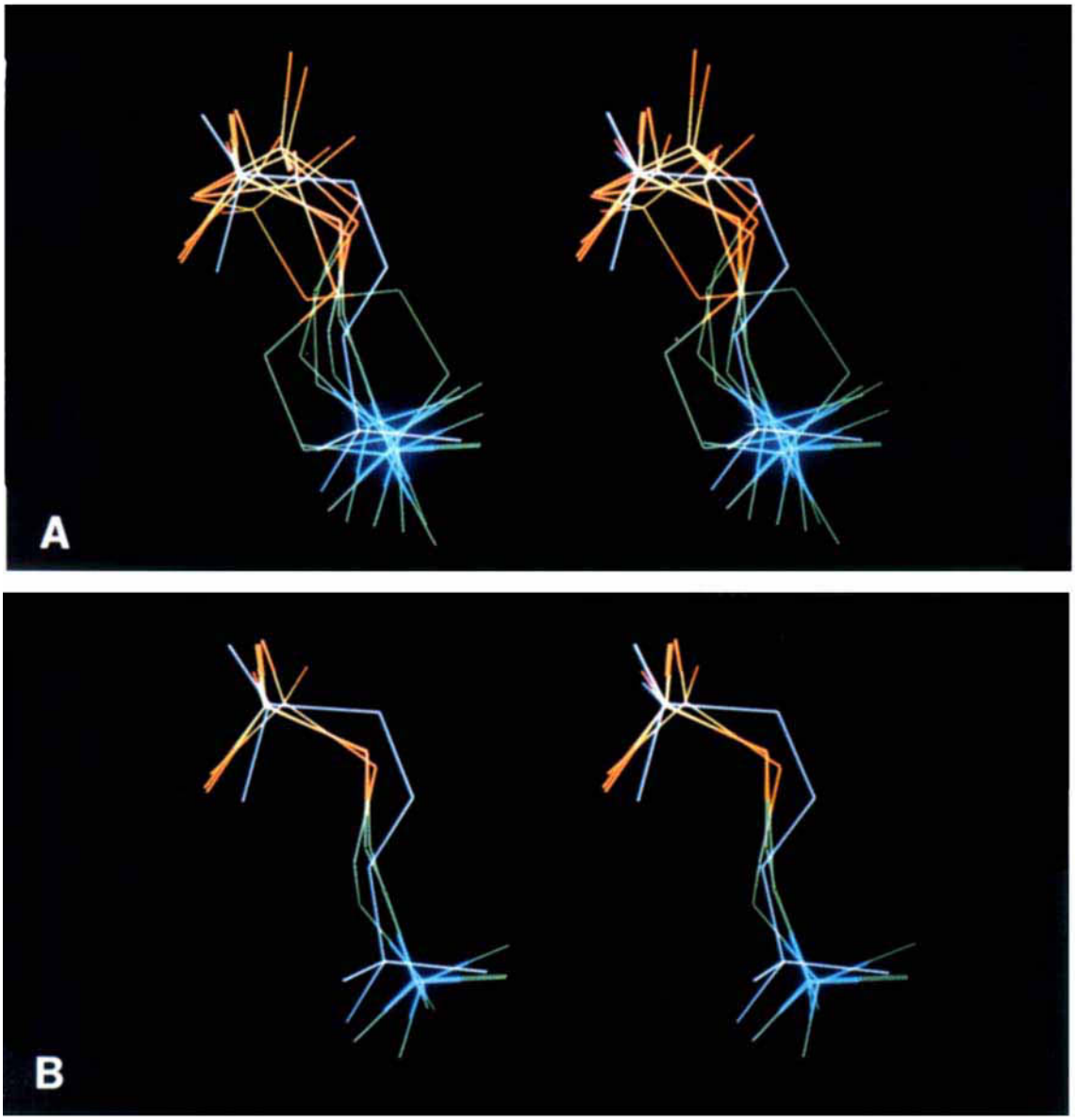
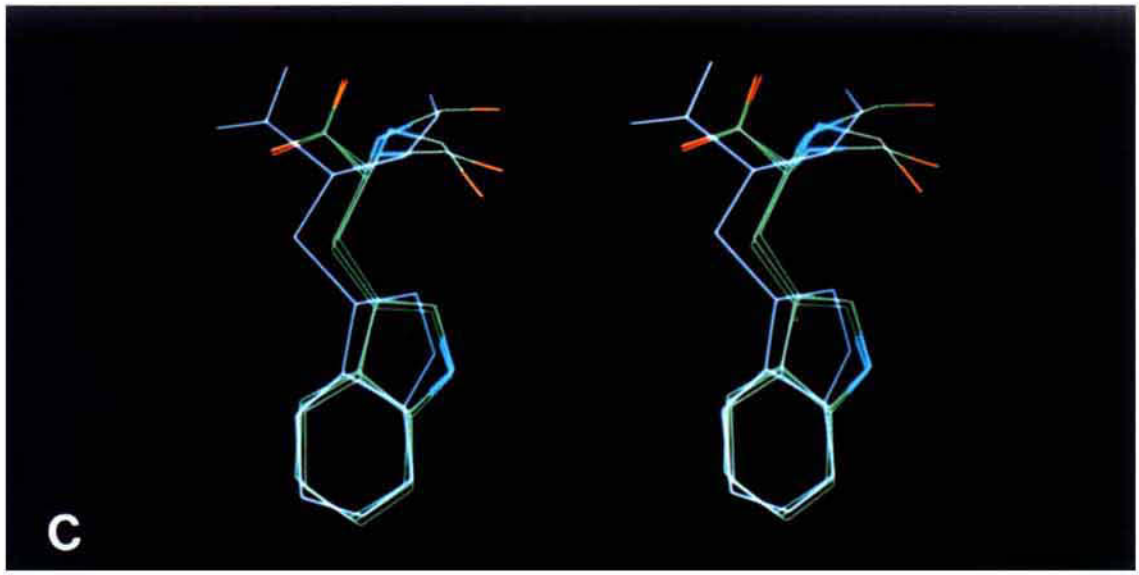


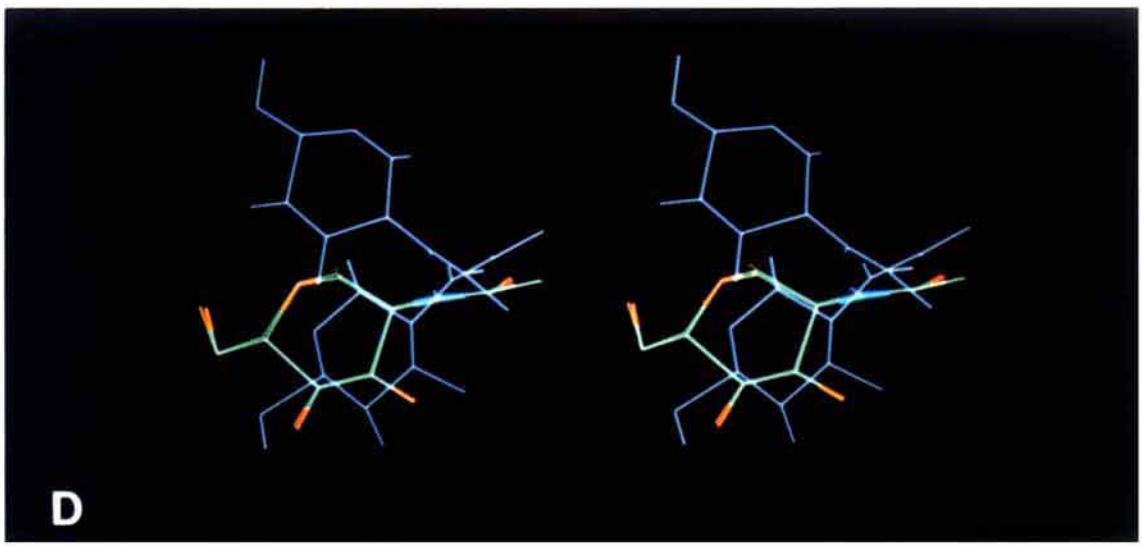
Fig. 2. Results of test simulations. In each image, the simulated conformations (C—green, O—red, N—blue, P,S—yellow) are shown relative to the crystallographic conformations (purple). **(A)** Ten simulations of phosphocholine binding to McPC 603. **(B)** Top three ranked phosphocholine conformations. **(C)** Top three ranked conformations of *N*-formyl-tryptophan binding to chymotrypsin. **(D)** Top three ranked conformations of α-*N*-acetylglucosamine binding to lysozyme. The crystallographic conformation of the α anomer is at the top and beta on the bottom. **(E)** Top three ranked conformations of β-*N*-acetylglucosamine. **(F)** Ten simulations of sulfate binding to aconitase. The active site iron–sulfur cluster is shown at lower left and the crystallographic sulfate position at upper right. **(G)** Top ranked citrate binding conformation in aconitase. These images were created with GRANNY.[21]

started with the substrate centered between the iron–sulfur cluster and the crystallographic sulfate site. Iron atoms were given a potential of 10 kcal/mol at 3 Å to simulate coordination.

Two unique sites of sulfate binding were found: the crystallographic site and a site coordinated to the iron–sulfur cluster (Fig. 2F). The two conformations were each found by four different simulation paths, yielding binding modes that are functionally identical, but which differ by rotation about the sulfur atom, placing different oxygen atoms in the favorable binding sites. The iron site is filled by a water or hydroxyl molecule in the crystal structure, which was not included in the simulation. In the crystallographic-like conformation the two sites of hydrogen bonding are accurately reproduced by simulation.

The lowest energy citrate conformation stretches between the crystallographic sulfate site, which is surrounded by arginines, and the iron site (Fig. 2G).





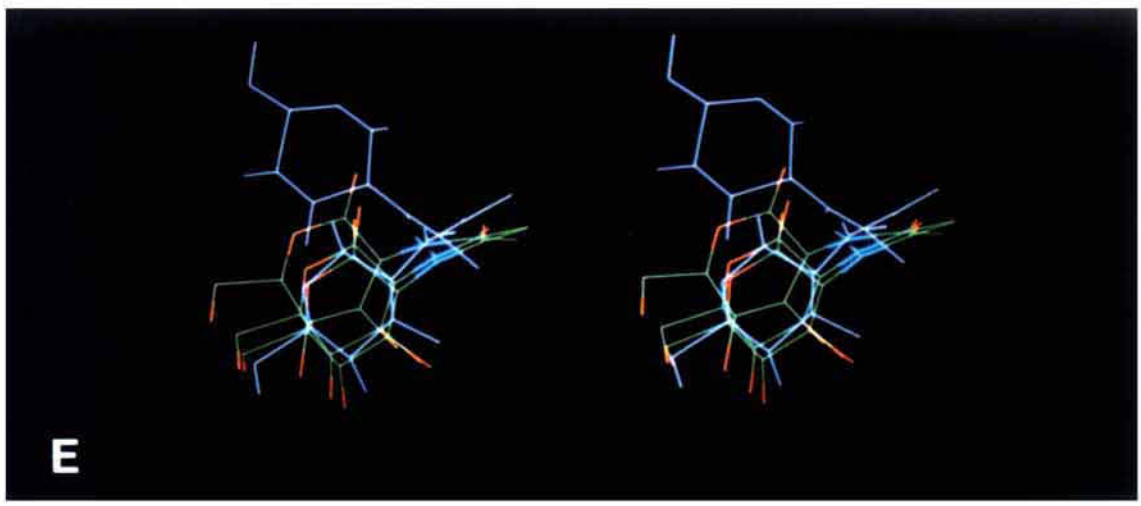
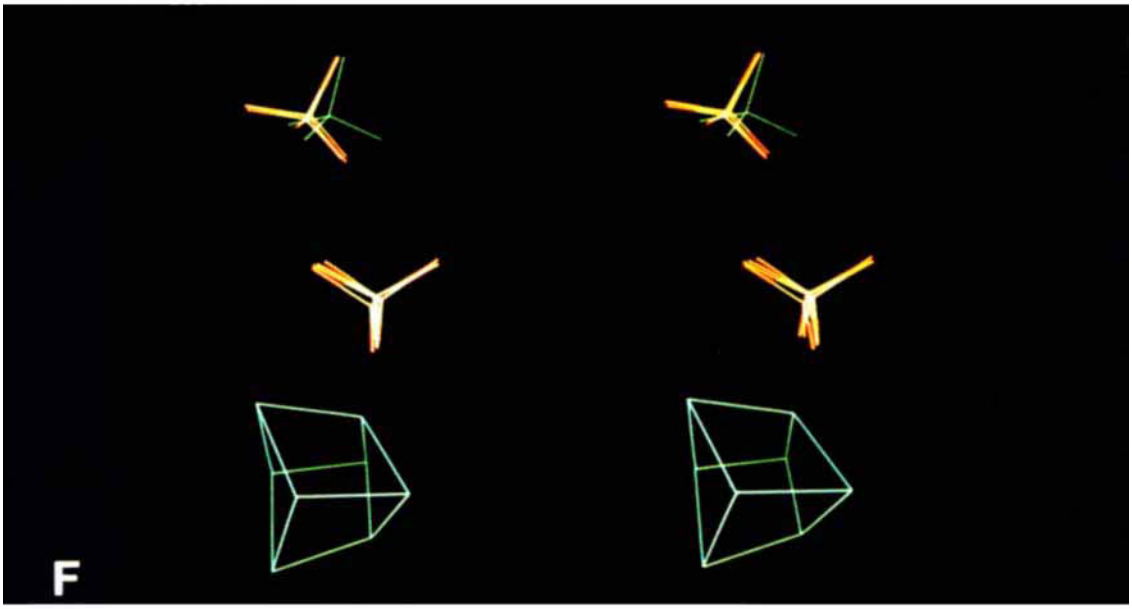


Fig. 2 C–E. Legend appears on page 199.



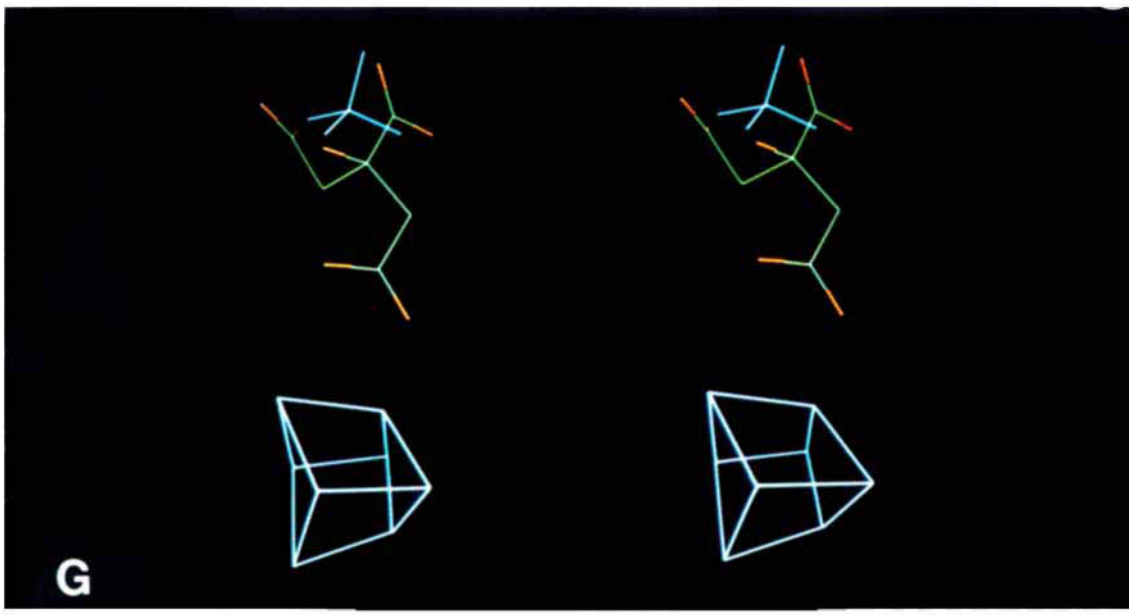


Fig. 2 F,G. Legend appears on page 199.

One citrate carboxyl bridges the iron and a histidine. From spectroscopic studies, two distinct modes of substrate binding have been postulated[18]: the "citrate mode," with the central β carboxyl of the substrate coordinated to the iron cluster, and the "isocitrate mode," coordinated at the terminal α carboxyl. The conformation obtained in the simulation is consistent with the latter mode.

DISCUSSION

Monte Carlo techniques are often criticized because they cannot absolutely guarantee a global solution. We have used tests of internal consistency and consistency with experiment to validate our results. The top ranked solutions in the test systems were consistently found in similar or identical final conformations, often reaching them by different simulation paths (for example, the different sulfate rotations). In addition, the top ranked solutions in all but the lysozyme:α-*N*-acetylglucosamine system conform to the experimentally observed binding mode. The rms difference of approximately 1 Å in these top solutions is reasonable, given the diffuse nature of the affinity potential—the substrate func-

tional groups are each placed at their proper binding loci.

The current technique has one important limitation: the use of a static protein structure. We are extending the procedure in two ways to address this problem. The affinity grids may be calculated using a soft potential, smearing the atoms over a larger region of space based on their mobility, from temperature factors or dynamics simulations. Alternatively, several grids may be calculated, each with a slightly different protein conformation, from molecular dynamics or normal mode analysis. These grids may be scrolled through during the simulation, using the grid number as another parameter to be varied.

The grid calculation and docking programs will be made available to the public, through the authors.

ACKNOWLEDGMENTS

We wish to thank David Stout for the use of the aconitase coordinates before publication and I. Saira Mian and David Case for helpful discussions. This work was funded in part by Grant DRG 972 from the Daymon Runyon–Walter Winchell Cancer Research Fund and NIH Grant GM3-5-221144.

REFERENCES

1. For an extensive bibliography of computational studies, see: Stedmann, N.J., Morris, G.M., Atkinson, P.J. Bibliography of theoretical calculations in molecular pharmacology. J. Mol. Graphics 5:211–222, 1987.
2. Beddell, C.R., Goodford, B.J., Norrington, F.E., Wilkinson, S., Wootten, R. Compounds designed to fit a site of known structure in human hemoglobin. Br. J. Pharmacol. 57:201–209, 1976.
3. Freudenreich, C., Samama, J.-P., Biellmann, J.-F. Design of inhibitors from the three-dimensional structure of alcohol dehydrogenase. Chemical synthesis and enzymatic properties. J. Am. Chem. Soc. 106:3344–3353, 1984.
4. Blaney, J.M., Jorgensen, E.C., Connolly, M.L., Ferrin, T.E., Langridge, R., Oatley, S.J., Burridge, J.M., Blake, C.C.F. Computer graphics in drug design: Molecular modelling of thyroid hormone-prealbumin interactions. J. Med. Chem. 25:785–790, 1982.
5. Wodak, S.J., Janin, J. Computer analysis of protein-protein interaction. J. Mol. Biol. 124:323–342, 1978.
6. Santavy, M., Kypr, J. A fast computational algorithm for finding an optimum geometrical interaction of two macromolecules. J. Mol. Graphics 2:47–49, 1984.
7. Goodsell, D., Dickerson, R.E. Isohelical analysis of DNA groove-binding drugs. J. Med. Chem. 29:727–733, 1986.
8. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E. A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. 161:269–288, 1982.
9. Billeter, M., Havel, T.F., Kuntz, I.D. A new approach to the problem of docking two molecules: the ellipsoid algorithm. Biopolymers 26:777–793, 1987.
10. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. Equation-of-state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092, 1953.
11. Kirkpatrick, S., Gelatt, C.D., Jr., Vecchi, M.P. Optimization by simulated annealing. Science 220:671–680, 1983.
12. Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important molecules. J. Med. Chem. 28:849–857, 1985.
13. Tintelnot, M. and Andrews, P. Geometries of functional group interactions in enzyme-ligand complexes: Guides for receptor modelling. J. Comp.-Aided Mol. Design 3:67–84, 1989.
14. Segal, D.M., Padlan, E.A., Cohen, G.H., Rudikoff, S., Potter, M., Davies, D.R. The three-dimensional structure of a phosphorylcholine-binding mouse immunoglobulin Fab and the nature of the antigen binding site. Proc. Natl. Acad. Sci. U.S.A. 71:4298–4302, 1974.
15. Steitz, T.A., Henderson, R., Blow, D.M. Structure of crystalline alpha-chymotrypsin. J. Mol. Biol. 46:337–348, 1969.
16. Imoto, T., Johnson, L.N., North, A.C.T., Phillips, D.C., Rupley, J.A. Vertebrate lysozymes. In: "The Enzymes," Vol. 7 (P.D. Boyer, ed.). New York: Academic Press, 1972.
17. Robbins, A.H., Stout, C.D. Structure of activated aconitase: formation of the [4Fe-4S] cluster in the crystal. Proc. Natl. Acad. Sci. U.S.A. 86:3639–3643, 1989.
18. Emptage, M.H. Aconitase, evolution of the active-site picture. In: "Metal Clusters in Proteins" (L. Que, Jr., ed.), ACS Symposium Series 372. Washington, D.C.: American Chemical Society, 1988: 343–371.
19. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., Weiner, P. A new force field for molecular mechanical simulation of nucleic acid and protein. J. Am. Chem. Soc. 106:765–784, 1984.
20. Goodsell, D.S., Mian, I.S., Olson, A.J. Rendering volumetric data in molecular systems. J. Mol. Graphics 7:41–47, 1989.
21. Connolly, M.L., Olson, A.J. GRANNY, a companion to GRAMPS for the real-time manipulation of macromolecular models. Comput. Chem. 9:1–6, 1985.