Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 1146–1153, February 1993
Colloquium Paper

This paper was presented at a colloquium entitled "Molecular Recognition," organized by Ronald Breslow, held September 10 and 11, 1992, at the National Academy of Sciences, Washington, DC.

Molecular recognition analyzed by docking simulations: The aspartate receptor and isocitrate dehydrogenase from *Escherichia coli*

(protein docking/drug design/energy minimization/substrate binding/receptor signaling)

BARRY L. STODDARD AND DANIEL E. KOSHLAND, JR.

Division of Biochemistry, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720

ABSTRACT **Protein docking protocols are used for the prediction of both small molecule binding to DNA and protein macromolecules and of complexes between macromolecules. These protocols are becoming increasingly automated and powerful tools for computer-aided drug design. We review the basic methodologies and strategies used for analyzing molecular recognition by computer docking algorithms and discuss recent experiments in which (*i*) substrate and substrate analogues are docked to the active site of isocitrate dehydrogenase and (*ii*) maltose binding protein is docked to the extracellular domain of the receptor, which signals maltose chemotaxis.**

Why Protein Docking?

The terms "rational drug design" and "computer-aided drug design" refer in their most specific sense to the systematic exploration of the three-dimensional structure of a macromolecule of pharmacological importance, in order to design potential ligands that might bind to the target with high affinity and specificity. This goal is largely carried out through docking protocols, which quantitate the affinity between the macromolecule and a ligand bound in specific locations and conformations. This discipline is currently used to examine molecular recognition with some success for the following purposes:

(*i*) The screening of a large number of small molecule species for binding activity against a single target molecule (1–4). A number of data bases of small molecule structures currently exists for such docking searches, including the Cambridge Structural Database (5), the Fine Chemicals Directory by Molecular Design Limited (4), and the Chemical Abstracts registry.

(*ii*) Detailed statistical and energetic analyses of an individual small molecule and its binding interactions to a specific macromolecular site (6–8). Such an analysis often begins where the initial computational screening of many candidates ends, allowing us to quantitate the binding of individual compounds and to design and test closely related molecules that exploit the architecture and specificity of the protein binding site. Computational paradigms are usually needed, which use more robust conformational searches and energy calculations than those used for rapid screening of large data bases. For this paper we discuss substrate analogue binding studies pursued through a combination of computational techniques and actual enzymatic analysis, using isocitrate dehydrogenase (IDH) as a model system.

(*iii*) The determination of the structure of protein–protein complexes (9–12). This is one of the most important emerging problems in structural biochemistry, due to the rapidly increasing number of structures being solved and the even more rapidly increasing number of gene products being identified, characterized, and sequenced that recognize and associate with one another. It is also one of the most difficult problems, due to the challenge of performing actual structural analyses of large multiprotein complexes and of computationally modeling structures with such a large degree of topographical and thermodynamic complexity. We discuss in this paper recent results from computationally predicting the protein–protein interactions between the tar protein, which has been shown to be a membrane-bound receptor that mediates both aspartate and maltose chemotaxis, and the maltose binding protein (MBP; which binds to the receptor).

Modeling and Simplifying the System

To determine and characterize molecular recognition and binding by a large macromolecule, simplified computational strategies currently must be followed in order to keep the calculations within reason. The simplifications that are used are severalfold.

(*i*) *Rigid body docking searches.* The number of possible conformational isomers of a macromolecule of even limited size is so large that the target molecule is usually treated initially as a collection of unmoving atoms (7, 10). In addition, for most data base-screening algorithms, the small molecule probe structures are also held rigid (a compromise that reduces the success rate of identifying potential drugs by an unknown amount). During the refinement of the low-energy docking solutions, some macromolecular dynamic motion is sometimes allowed in the region of the docked complex. The methods that use rigid atoms reduce the computational demand of docking searches but can be misleading since virtually all substrates and other ligands to macromolecular surfaces induce conformational changes upon binding. Agard and coworkers (13, 14) have shown that modeling algorithms that make use of multiple side-chain rotamers provide an energy calculation that is powerful in predicting conformational variation in the active site. Such calculations should allow the design and manipulation of engineered enzyme–

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: IDH, isocitrate dehydrogenase; MBP, maltose binding protein.

ligand complexes through empirical energy evaluations (13, 14).

(ii) Reduction of macromolecular structural information. The size of the target macromolecule in a docking simulation can range from several hundred atoms for short lengths of nucleic acid sequence (15) to 2000–3000 for the full oligomeric structure of a small ligand-binding domain of a membrane-bound receptor (12) and up to 6000–10,000 for an enzymatic catalyst such as human immunodeficiency virus protease and thymidylate synthase (4) or isocitrate dehydrogenase. Strategies for reducing this massive amount of data include representing individual protein residues or side chains and ligand atoms by space-filling spheres possessing various charge or polar characteristics (9, 10) or representing the clefts and binding pockets across the macromolecular surfaces with sets of filling spheres (4, 16), a process that is analogous to making a wax mold of a keyhole, and then comparing the mold with known molecules. Both of these strategies have lead to docking protocols that operate in response to complementarity between the surface structures of protein and ligand. In the first case, the energy of nonbonded contact between two molecules is calculated directly; and in the second case, the physical match and superimposition of three-dimensional models of the ligands and the corresponding binding clefts in the protein surface are calculated. Alternatively, a macromolecular structure may be reduced to a series of three-dimensional grids of molecular affinity potentials, which correspond to specific atom types in the docked ligand probe (6, 7), allowing rapid energy evaluation during docking simulations while maintaining large search space and a relatively robust energy calculations, including terms that take into account nonbonding, electrostatic, and possible hydrogen-bonding interactions.

(iii) Reduction of available search area. Given the complexity and sheer size of a macromolecular surface, it is desirable to restrict the possible area of complex formation for at least one of the two species, provided that such a simplification is based on data that does not undermine the probability of obtaining a correct bound solution. This strategy has been used with success in deriving protein–protein complex structures involving antibodies, for which the variable recombination sites may be used in place of the full protein structure (10), enzymes, for which the active site is used (4, 7), and macromolecular structures, for which genetic information related to the binding event allows selection of specific surfaces or peptide sequences to use as a set of probe ligands (12). This final strategy has been used to predict the structure of a receptor–protein complex, with a comparison between the individual docked peptides and the same sequences in the actual protein structure used to help distinguish correct from incorrect docked solutions.

(iv) Energy calculations and evaluation. Unlike a number of manual docking methods, which explore a limited subset of bound positions and conformations by sophisticated energy evaluation methods, most automated docking protocols require the use of simplified energetic models to keep the computational demands within reason. Therefore, most methods combine a simplified model of the surfaces to be docked (as described above) with a straightforward method of quantitating bound energies. For example, a crude measurement of the complementarity of surface contours and/or charges is followed by a minimization of the calculated binding energy by using a variety of protocols,* including "Metropolis," or simulated annealing minimization (17), or a least-squares minimization (18), which usually finds the closest local energy minima for each docked solution.

The energy assessment strategies can be divided roughly into three groups, all of which have been used with a large amount of overlap in the computational docking applications reported by a number of investigators.

Strategy 1: geometric analysis, in which the calculated docking affinity is proportional to interface complementarity, total buried surface area, and/or van der Waals contact potentials (4, 9, 10, 16). The usual term to be minimized in docking protocols that model the spatial complementarity of the ligand and the protein surface is the Lennard–Jones potential, or nonbonded van der Waals contact potential:

$$E = A/d^n - B/d^m, \quad [1]$$

where d is proportional to the distance between nonbonded residues, A and B are scale factors, and n and m are exponential terms that influence the distance-dependent degree of attractive and repulsive energies between spheres and therefore the maximum and minimum nonbonded contact distances influenced and allowed by van der Waals forces, respectively. In addition, some geometric docking protocols have also attempted to model energies of desolvation upon complex formation (9).

Some protocols such as the program DOCK (4, 16) attempt to optimize the physical similarity of the docked ligand and available macromolecular binding sites by representing the cleft as a group of overlapping spheres and then searching for molecules that are capable of forming a close three-dimensional match to those spheres. This method generally offers fast computational run times and allows the incorporation of electrostatic and hydrogen bonding terms and minimization methods as needed.

Strategy 2: electrostatic analysis, in which the calculated docked affinity is primarily related to the sum of electrostatic interaction energies. Methods that attempt to align and optimize the partial charges of the ligand and protein atoms have existed for as long as the previously discussed geometry-driven methods. Salemme (19) predicted the structure for a complex of the intermolecular electron-transfer cytochromes c and b_5 by optimizing the complementary charge and steric interactions between the two molecules by using a least-squares optimization and manual fitting procedure. More recently, Warwicker (20) examined the interface structure of several complexes, including trypsin–bovine pancreas trypsin inhibitor, anti-lysozyme Fab–lysozyme, and cytochrome c–cytochrome c peroxidase; he found highly favorable interacting regions in the interface as determined from reduced charge contour maps of the protein surfaces. Finally, Shoichet and Kuntz (11) have found that evaluation of total interaction energy and electrostatic interaction energy of protein complexes is somewhat more successful at discriminating between correct and incorrect docked solu-

*During a Metropolis, or simulated annealing minimization, the temperature (or energy) of the system during the ensuing minimization is slowly decreased from a starting high value while allowing the model to vary in a random manner, and the probability of accepting a step of higher energy than the previous step is given by $P(\Delta E) = \exp(\Delta E/kBT)$, where kB is Boltzmann's constant and ΔE is the difference in energy from one step to the next. Thus high initial energies in the system will accept most random steps, allowing the substrate molecule to sample all energetic bound states; as the temperature decreases, the probability of accepting a high energy step diminishes and a global energy minimum is eventually reached. A least-squares minimization is a straightforward protocol for the minimization of a structural parameter, such as the agreement between calculated and observed diffraction terms or the superposition of a ligand structure onto a space-filled model of a binding site; such a method often uses a conjugant gradient calculation and sparse matrix sampling protocol to assess the agreement between observed and calculated parameters during minimization and is an efficient method for sampling local minima surrounding an initial structural model.

tions (as determined from crystallographic analysis) than methods that rely on surface burial, solvation free energy, packing, and mechanism-based filtering. This advantage may be most marked in the analysis of protein–protein complexes, due to the larger surface areas involved, which may hinder geometry-dependent methods.

Strategy 3: molecular affinity analysis, in which the total interaction energy is related to the sum of interactions of specific atomic probe groups with target protein groups, as calculated through a summation of independent energy terms for Lennard–Jones, electrostatic, and hydrogen bond interactions. This method attempts to combine the best features of the steric- and electrostatic-dependent docking protocols described above, while also reducing the surface of the target protein to a series of atom-dependent "affinity grids" (6). These grids are calculated over a preselected volume and surface area of the protein in such a manner as to assign a potential affinity of the protein for each atom type in the ligand molecule at periodic points throughout the protein's three-dimensional structure. This algorithm, as shown in Eqs. 2 and 3, thus places weight on both the steric fit of ligand and protein surface and on the chemical properties of individual atoms through the calculation of electrostatic potentials in the model (ion pairs, partial charge dipoles, and hydrogen bonds).†

$$E_{xyz} = \sum E_{lj} + \sum E_{elec} + \sum E_{hb} \quad [2]$$

$$= \sum(A/d^{12} - B/d^{6}) + \sum(pq/K\gamma[1/d((\gamma - \varepsilon)/(\gamma + \varepsilon))/(d^2 + 4s_p s_q{}^{1/2})] + \sum[C/d^6 - D/d^4]\cos^m\theta \quad [3]$$

Goodsell and Olson (7) have linked the molecular affinity analysis algorithm with the Metropolis technique (17) of conformational searching to obtain an efficient procedure for docking small ligands to macromolecules. Tests using a number of crystallographically well-categorized systems, including phosphocholine/McPC-603 (an antibody), chymotrypsin/*N*-formyl-tryptophan, and lysozyme/*N*-acetylglucosamine yield lowest energy solutions with rms deviation from the crystallographic coordinates for bound substrate ranging from 0.94 Å to 4.01 Å. Recently we have used the technique successfully to model widely different problems, enzyme–substrate and protein–protein interactions.

IDH—a Difficult-to-Dock Active Site

IDH from *Escherichia coli* [isocitrate:NADP$^+$ oxidoreductase (decarboxylating), E.C. 1.1.1.42] catalyzes the conversion of isocitrate to α-ketoglutarate and CO_2. The enzyme is dependent on NADP and on bound metal (usually Mg^{2+}) and lies at an important branch point in carbohydrate metabolism. The enzyme is completely inactivated by phosphorylation of an active site serine residue, which controls net flow of metabolites between the Krebs cycle and the glyoxylate bypass. This form of metabolic regulation is critical for the survival of *E. coli* on nutrients such as acetate.

The structure of IDH has previously been solved at 2.5 Å resolution as apoenzyme, as phosphorylated apoenzyme, as a binary complex of isocitrate and Mg^{2+}, and as a complex with NADP in the absence of substrate and metal (21–24). The enzyme is a dimer of 416 residues per subunit and contains a single catalytic metal per monomer, which is tightly chelated by two conserved aspartate residues and by bound isocitrate. The substrate molecule is bound in the active site primarily through interactions between its free carboxylate groups and several conserved basic residues. Modeling of the binary complex of isocitrate and Mg^{2+} in the active site indicates that phosphorylation of Ser-113 prevents substrate binding through direct electrostatic interactions between the phosphate oxygens and the γ-carboxylate of isocitrate (22).

Isocitrate is bound in the active site through a collection of electrostatic interactions between its three carboxyl groups and single hydroxyl group and a large number of conserved electophilic groups (Fig. 1), including Arg-119, Arg-129, Arg-153, Ser-113, Lys-230, Asn-115, and Mg^{2+}. The K_m is 10 μM. In the original structural solution of the isocitrate–Mg^{2+} binary complex bound in the active site, modeling of the bound substrate was difficult to accomplish due to the relatively symmetric structure of the negatively charged isocitrate (only the hydroxyl breaks the symmetry) and of the binding site itself (22). It was necessary to solve the structure of the complex in the presence of Mg^{2+} and then Mn^{2+} (in a separate experiment) in order to calculate difference maps that locate the bound metal ion and thereby properly orient the isocitrate molecule. This was due to the featureless, uniform appearance of the electron density in substrate/metal difference maps and because it is possible to orient isocitrate in the binding pocket in more than one orientation and still create reasonable contacts with the surrounding protein.

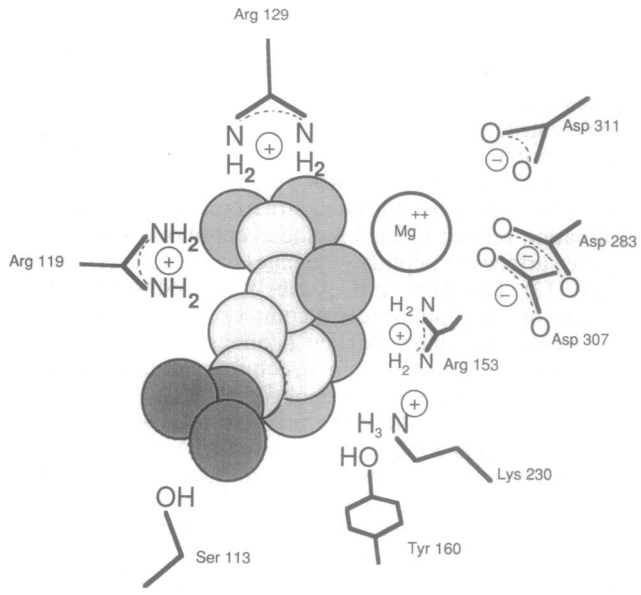


FIG. 1. Complex of isocitrate in the active site of IDH, with residues involved in substrate binding (22). The substrate molecule is bound in the active site through interactions between its 3-carboxyl and 1-hydroxyl group and various conserved polar groups as shown. The labile carboxyl of isocitrate, which is eliminated through a putative endiolate-intermediate mechanism (shown below) to generate α-ketoglutarate and CO_2, is hydrogen bonded to Lys-230′ and Tyr-160′. The terminal γ-carboxylate (which is missing in D-malate) is shown as darkly shaded spheres hydrogen-bonded to Ser-113. Carbons are shown as white spheres, and oxygens are shown as lightly shaded spheres relative to γ-carboxylate.

†Terms for the potentials described in the molecular affinity analysis protocol: xyz, lj, elec, and hb denote the total interaction potential, and the Lennard–Jones, electrostatic, and hydrogen-bond potentials, respectively; A and B are atom-specific weighting terms reflecting the repulsive and attractive nonbonded contact potentials; p and q are electrostatic charges on paired atoms from ligand and protein; d is the distance between those atoms; K is a combination of geometrical factors and constraints on electrostatic fields; γ and ε are constants reflecting the distance dependence of electrostatic potentials on the dielectric nature of the environment; C and D are geometric and distance-dependent constants for hydrogen-bond potentials; θ is the bond angle between donor, proton, and acceptor; and m is a scale factor related to the hydrogen bond angle.

Table 1. Docked solutions and energies of substrates with IDH and agonists of the MBP–receptor complex

Protein target	Ligand	Energy*	Redundancy†	rms‡
IDH	Isocitrate	−92	1 of 5	1.95
		−89	1 of 5	3.02
		−87	1 of 5	3.67
		−82	1 of 5	4.38
		−81	1 of 5	4.40
	Malate	−36	3 of 3	1.80
	Methyl malate	22	3 of 3	1.93
	Ethyl malate	126	3 of 3	1.72
Aspartate receptor	Aspartate	−48	6 of 10	0.96
	MBP peptide 1	2411	3 of 5	0.79
	MBP peptide 2	−15	4 of 5	2.10

*Calculated energy in kcal/mol.

†Number of runs yielding the docked solution vs. total number of independent docking runs. For example, isocitrate was docked five times, to give five separate, closely related solutions of different energies; the lowest energy solution has the closest agreement with the crystallographic solution of the isocitrate complex. Malate, methyl malate, and ethyl malate were docked three times, beginning with the compound in the same orientation as the previously solved structure of isocitrate.

‡The rms for isocitrate and aspartate in Å is calculated against previously solved crystallographic complexes. The rms for malate and its derivatives is against the equivalent atoms in the crystallographic structure of the isocitrate–IDH complex. The rms for peptides from MBP is compared to the same atoms in the crystal structure of MBP after superposition of peptide 2 over residues 340–348.

Not surprisingly, computational docking of isocitrate and of related molecules to the IDH active site provides a challenge in terms of the specificity of the protocol used and the ability of the energy evaluation to distinguish between correct and incorrect orientations of bound substrate. In an initial experiment, the molecular affinity-based protocol (7) was used to determine the precision of calculated docking to IDH, by comparing the predicted binding of isocitrate and malate with the crystallographically solved structure of the enzyme bound with isocitrate–Mg^{2+} (22) and with the binding constants determined through initial-slope kinetic studies. Partial charges were assigned to each atom in the enzyme structure by using the molecular modeling program QUANTA, and we then calculated three-dimensional grids of molecular affinity potentials, which encase a preselected region of the protein surface and interior volume. The grids were 15 Å on each side, sectioned every 0.5 Å, and centered on the crystallographically identified binding site. The metal ion was included in the enzyme structure, but with no explicitly modeled solvent or bound ligands in the active site. For isocitrate, all five possible torsion bonds were allowed to rotate freely. The individual carboxyl groups on the substrate were restrained from contacting one another in order to avoid cyclization in the conformational search algorithm. The initial energy of the system was 100 kcal/mol, and 300 cycles of simulated annealing minimization were performed, reducing the energy by 1% each cycle. A maximum of 20,000 rejected or accepted steps were allowed in each cycle.

Docking of isocitrate, using multiple runs with random initial starting orientations, produces multiple docked conformations, which all overlap one another, but offer different combinations of interactions between the carboxyl groups of the docked substrate and the enzyme active site (Table 1). The calculated bound energies range from −92 kcal/mol to −82 kcal/mol. The absolute values of these energies are not accurate, as they are dependent on the environmental and charge parameters used to set up the simulation. However, the lowest energy solution, with a difference of about 3% in calculated interaction energy from the next lowest energy solution, is the one most closely matching the crystallographic structure of bound isocitrate (with the metal chelated by the C1 carboxyl and the hydroxyl group of isocitrate and the γ-carboxyl interacting with Ser-113; Fig. 2). Thus the structure that agrees with the crystal structure was the one calculated to have the lowest binding energy, but the difference with the next lowest energy structure would make the choice marginal in the absence of supporting evidence.

Perhaps even more interesting are the results from docking experiments in which analogues of isocitrate, with various chemical substitutions at their γ-carboxyl group, are computationally docked to the active site. Removal of the carboxyl group altogether, to produce the molecule D-malate, produces a less favorable calculated energy of binding of −36 kcal/mol. The ratio of the computationally derived binding energy for isocitrate to that for malate is in close agreement with the actual ratio, which can be determined from the experimental Michaelis binding constants for the two molecules (A. Dean, A. Shiau, and D.E.K., unpublished results), implying that the energy evaluation algorithm copes well with the simple loss of attractive van der Waals potential and the formation of a "hole" in the binding site, which occurs when a wild-type carboxyl group is removed.

Docking of methyl, ethyl, and propyl malate produces far less accurate results when the molecules are initially placed

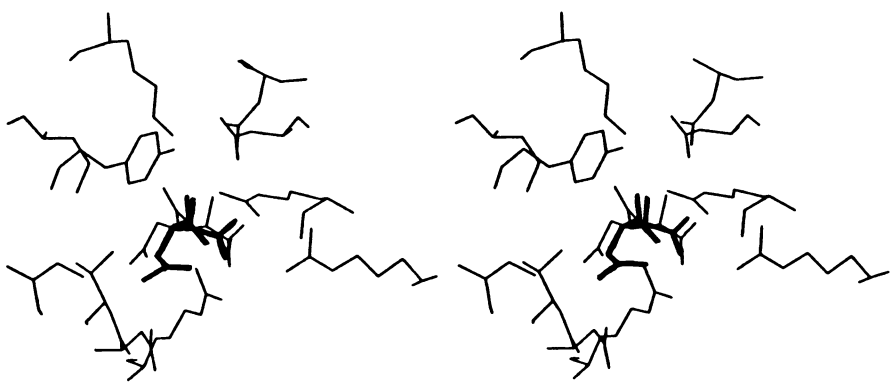


FIG. 2. Stereoview of isocitrate bound in the active site of IDH. The crystallographic structure of the complex is shown as thin bonds, and the best computationally docked solution of isocitrate is superimposed and shown as the molecule with thick bonds. The actual orientation of isocitrate, with the metal complexed to the C1 carboxylate and the hydroxyl group, is reproduced by the docking protocol. In the next best docking solution, the substrate is rotated about the vertical axis by almost 180°, with the metal chelated by the γ-carboxyl.

in the isocitrate binding orientation and then subjected to a simulated annealing docking experiment as described above. The calculated binding energies after docking and conformational minimization increase sequentially as methyl groups are added to the end of the small molecule. In this case, steric clash between the substrate analogue and the enzyme active site residues prohibits an effective search for a global minimum. Starting from random orientations of the substrate analogues produces some lowering of these energies, but produces final orientations that do not agree with the crystallographic or computational structure of the isocitrate complex and that are probably not actually formed during catalysis. Kinetic studies show that in actuality these substituted malate compounds are substrates for the enzyme and bind *more* tightly than malate itself (A. Dean, A. Shiau, and D.E.K., unpublished results) and only about 10 times worse than isocitrate, with a V_{max} reduced by 2–3 orders of magnitude. In other words, during real turnover, the enzyme has the flexibility necessary to accommodate a moderately large number of extra atoms (methyl groups) on the substrate and still allow effective binding and turnover. Most docking protocols, which as previously described use a rigid-body approximation for the protein structure, do not allow proper modeling of this flexibility, which is an enormously important component of enzymatic structure and function. In terms of protocols where large structural data bases of small molecules are systematically screened for potential binding and therapeutic activity, this means that there is an inherent bias against molecules bulkier than naturally occurring substrate molecules, a fact that lowers the "hit" rate of such screening methods by an undetermined amount.

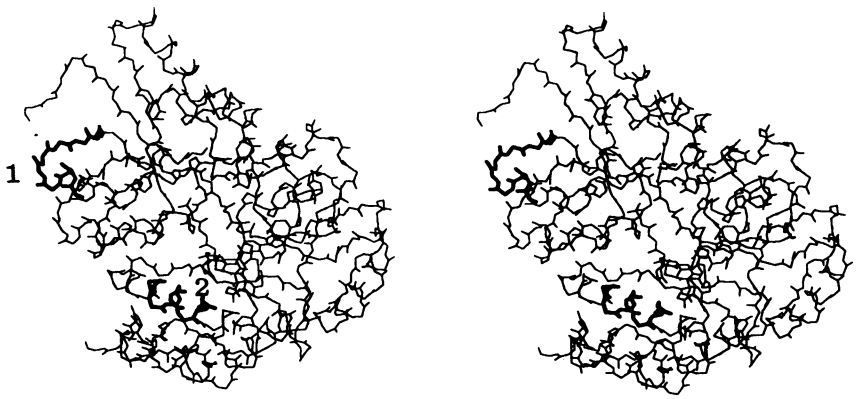
A Receptor–Agonist Complex—Attacking the Protein–Protein Docking Problem

One of the most striking results from the recent application of docking algorithms to the prediction of ligand–protein structures is that protocols utilizing a mix of energy evaluation methods and search patterns display a relatively high success rate at finding the correct structure when multiple runs are performed and the program is allowed to have several opportunities at finding energetic docked minima. In two recent, authoritative reports (10, 11), roughly 80–90% of those systems analyzed yielded such results (typical problems used as test cases include trypsin–bovine pancreas trysin inhibitor, serine protease–ovomucoid third domain, and lysozyme–Fab complexes).

However, these same reports are less successful in differentiating between a "correct" result and alternatives that display similar interaction energies, interface complementarity, and buried surface areas, but at different locations on the protein surface. These results indicate that final energy minimization of the docked solutions usually distinguishes to a limited degree between "true" and "false" positives, but with very small calculated energy differences, so that obviously unique, correct solutions are rare (≈10% of test systems reported). We have recently reported a strategy, termed "binary docking", which combines genetic information about a specific protein–protein association with the techniques reviewed above to produce a model of interaction between a periplasmic transport protein and its membrane-bound receptor (12). The method is designed to provide a check for internal consistencies, which helps to validate the quality of the resulting three-dimensional model.

The tar protein was originally shown to be a membrane-bound transducer for aspartate chemotaxis (25), and the same receptor from *E. coli* was subsequently shown to bind MBP also. The periplasmic domain of the *Salmonella* receptor has been solved crystallographically (26), and it is a straightforward procedure to deduce the structure of the *E. coli* receptor to which it has high homology. The site of interaction of the receptor and MBP is unknown, although mutants of both proteins that eliminate maltose taxis (27–29) are known. Using clues from these mutational studies, we selected two octapeptides in the regions of MBP that had sites of mutations that affected maltose chemotaxis (Fig. 3). Peptides from each region were then generated, docked to the receptor, and energy-minimized to their optimal positions on the aspartate receptor by using the molecular-affinity simulated annealing procedure of Goodsell and Olson (7). Large protein grids (30 Å^3) were purposely calculated in order to allow the "substrates" to achieve a random walk over a large surface of the receptor periplasmic domain.

In an initial test application of the automated docking protocol with the receptor model as a target, a model of aspartate (zwitterionic form) with partial atomic charges in an



MBP Peptide 1:

Gln - Val - Ala - Ala - [Thr] - Gly - [Asp] - Gly - Pro
49 53 55

MBP Peptide 2:

Trp - Tyr - Ala - Val - Arg - [Thr] - Ala - Val - Ile
340 345

FIG. 3. MBP and the peptides identified genetically as important for maltose taxis and for binding to the aspartate receptor. Peptide 1 is a loop region in the N-terminal domain of the receptor; peptide 2 is a length of helix in the C-terminal domain. [Modified figure reproduced with permission from ref. 12 (copyright Macmillan Magazines Ltd).]

FIG. 4. Stereoview of aspartate bound to its receptor. The crystallographic structure of the complex is shown as thin bonds, and the best computationally docked solution of aspartate is superimposed and shown as the molecule with thick bonds. The rms difference between the two bound models is 0.96 Å.

arbitrary initial conformation (which was unrestrained during docking) was docked to the protein. The results of our docking runs show that the calculated bound location and conformation of aspartate within the receptor structure matches the crystallographically observed aspartate-bound structure to within an rms deviation of 0.96 Å (Fig. 4). This matches our results with isocitrate and IDH (rms deviation = 1.95 Å), indicating that even with a large global search area, docking algorithms are capable of locating ligand binding sites.

In the case of MBP and the aspartate receptor, the size of the docked species is obviously far too large to pursue a direct automated docking solution. However, two peptide sequences were selected based on mutations that eliminated maltose taxis. These peptides were then used as the ligands to find their binding sites on the receptor. The Protein Data Bank coordinate files for each peptide with partial charges appended for each atom were generated from the crystal structure of the intact protein. In this manner the large tertiary structure of MBP is reduced to a pair of small molecule species. These two peptides were then used as substrate molecules in independent docking runs against the structure of the aspartate receptor. For each peptide, five separate docking minimizations were performed as described. Both peptides from MBP produce several docked positions and orientations, with one solution in each case dominating the total number of runs, both in terms of final docked energy and in the number of times that solution was produced (Table 1). However, the differences in energy between the minimum and the next best alternates are quite small (<5%), a fact that mirrors the results seen by other investigators during protein–protein docking experiments.

Peptide 2, corresponding to residues 340–348 in MBP (helix 13 in the C-terminal domain), gives the same docked

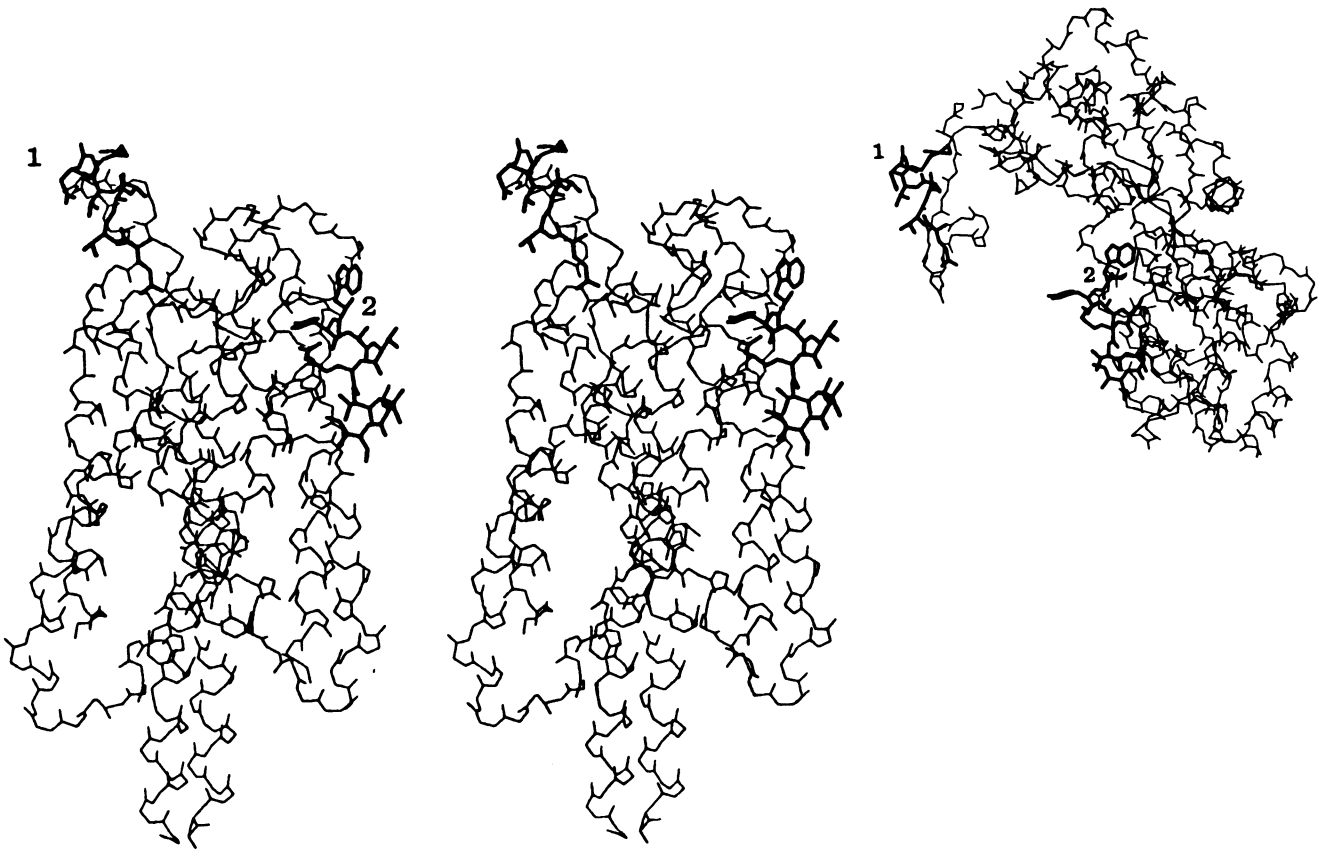


FIG. 5. (*Left*) Stereoview of the two peptides from MBP (thick bonds) in their computationally docked positions on the receptor (thin bonds). Peptide 2 (from helix 13 of the MBP C-terminal domain) packs into the dimer interface against two receptor helices. (*Right*) The structure of MBP in its conformation from the final, minimized complex structure is shown for comparison. [Reproduced with permission from ref. 12 (copyright Macmillan Magazines Ltd).]

solution in four of five runs and a bound energy of −15.2 kcal/mol, packing against the parallel helices 2 and 4′ of the receptor dimer (Fig. 5), in close contact with residues 73–85 and 148′–152′ of tar protein (regions previously implicated as being involved in MBP interactions). Peptide 1, corresponding to a loop region in the N-terminal domain of MBP encompassing residues 49–57, minimized to the same solution in three of five runs. The final energy, however, is quite high (2411 kcal/mol), which may reflect the fact that this peptide interacts in the docked structure with disordered external loops of the receptor, which are not resolved clearly in the x-ray structure. It is in close contact with residues 73′–81′ of the receptor, which have also been shown genetically to bind MBP.

The strategy of docking independent peptides from MBP has two important advantages. First, the "goodness" and reproducibility of unique docked solutions for each individual peptide can be assessed by examining final bound energies and the number of runs that produce the same result. Second, we can assess the accuracy of the docked solutions by three criteria for each peptide: (*i*) by comparing the calculated final positions, conformations, and distances between the two peptides (derived from independent runs) with those found in the crystallographic structure of MBP; (*ii*) by examining the quality of the protein–protein complex structure after superimposing MBP on the docked peptides; and (*iii*) by the agreement with separate genetic evidence that identifies regions on the receptor that should be found to interact with specific residues in MBP. This three-part validation process is important for assessing the correctness of the predicted model and for overcoming the ambiguities inherent in macromolecular docking.

Assessment of the docking solutions. The two peptides that were docked independently are both located on the surface of the receptor, with an approximate distance between one another of 30.0 Å (the C^{α} of Thr-53 to the C^{α} of Thr-345). The distance between the same two residues in MBP (complexed to maltose) is 29.3 Å. Superposition of the backbone atoms of residues 340–348 in MBP on peptide 2, which is located on the receptor surface after docking, leads to an almost direct overlap of the sequence of peptide 1 with its corresponding residues in MBP (rms total = 2 Å). The intact MBP, oriented by overlapping the two docked peptides, produces an excellent initial model, which suffers from steric conflict between the two proteins in only two locations involving surface loops of MBP and tar protein (residues 37–45 of MBP clash with tar loop 1 residues 68–86; MBP residues 610–615 overlap helix 3 of tar residues 130–135). Restrained minimization of the resulting complex structure produces a final solution (Fig. 6) of good geometry and energy.

The docked structure of MBP and the aspartate receptor. The internal consistencies within the binary docking of MBP to the receptor is enough to convince most investigators of the basic correctness of the model and of the power of computer docking algorithms. However, a second area of validation exists, which reinforces this conclusion: agreement of the predicted structure with all the genetic and chemical evidence that exists regarding binding of the two

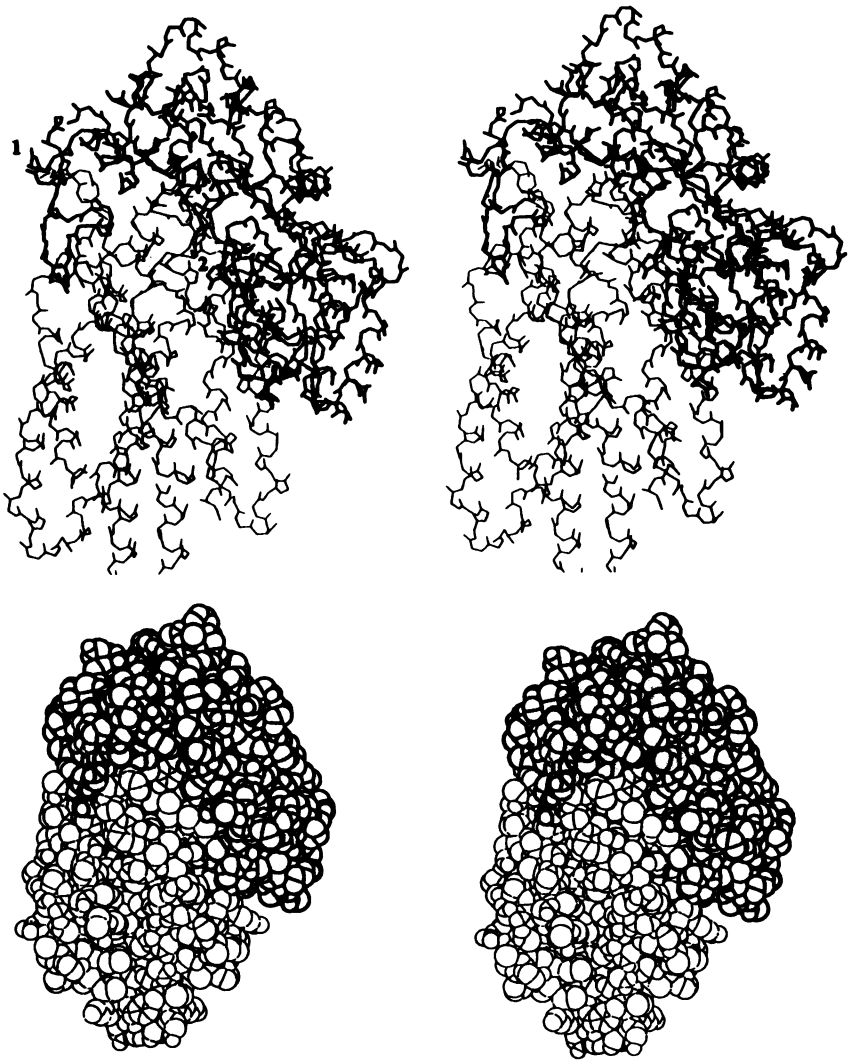


FIG. 6. Stereo protein backbone (*Upper*) and space-filling (*Lower*) diagrams of the complex of MBP bound to the dimeric structure of the periplasmic domain of the aspartate receptor, as generated from the computational docking of the genetically selected peptides shown in Figs. 3 and 5. The receptor domain is oriented with the ends of the helices that extend into the membrane pointing down. [Modified figure reproduced with permission from ref. 12 (copyright Macmillan Magazines Ltd).]

receptors to one another. The structure of the protein complex reported here contains a number of details that agree with various analyses and observations of the binding protein, the receptor, and maltose chemotaxis. One of the clearest results of several different genetic experiments on maltose chemotaxis, which is supported by the model presented here, shows that, of the five residues that directly bind aspartate, one (Arg-73) is also important for response to maltose (11, 19). In addition, mutation of Arg-73 acts to suppress the deleterious effect of mutating residues 53 and 55 in MBP. In the docked model of the protein–receptor complex, Arg-73 is the only one of these residues on the receptor that is directly involved in binding interactions to MBP, through hydrogen bonds to the backbone of the binding protein. Equally important, the side chain is located 3–4 Å from residues 53 and 55, which in turn contact a series of receptor residues involved in protein binding. Thus the complementation of mutations of residue 73 of the receptor with mutations of residues 53 and 55 of MBP is consistent with the model of the interactions based on the docking calculations.

One of the most intriguing aspects of the aspartate receptor is the fact that it can simultaneously and independently modulate responses to both aspartate and maltose (30). It is probable that this would necessitate sequential binding of both aspartate and MBP, which would lead to independent, additive signaling events. Crystallographic analysis of the receptor periplasmic domain (26), in conjunction with binding studies in our laboratory, seems to support half-of-sites binding of aspartate and negative cooperativity between the aspartate binding sites (D. Milligan, H. P. Biemann, and D.E.K., unpublished results). In the current models of receptor signaling, this binding event initiates a conformational change within the receptor structure, causing conformational changes in the protein dimer (26, 30, 31).

In the complex of MBP to the receptor reported here, the binding protein associates with the external loop regions of the tar protein, and primarily with the helices 2 and 4′ at one side of the dimer interface. In the complex, one of the two available aspartate binding sites is buried in the protein interface, whereas the second is completely solvent accessible. This would imply that aspartate is capable of binding a single accessible site and promoting chemotaxis before or after the binding of MBP. In the case of a mechanism involving conformational changes within individual monomers, it is easy to visualize signaling and adaptation induced within one receptor subunit by a single bound aspartate and a second independent signal transduced through the second subunit in response to the binding of MBP. The tar protein may have evolved a pattern of half-of-sites binding for aspartate and subsequent signaling through a single subunit in order to prevent saturation of the signaling potential of the receptor by high aspartate concentrations, which would negate the ability of the receptor to display a separate, additive signal in response to a second stimulating ligand such as maltose-bound MBP.

In conclusion, high-resolution structural studies of proteins, when combined with accurate dynamic modeling and energy calculation methods, seem capable of providing the information necessary for the prediction of complex binding events, such as protein–protein interactions. Such docking predictions can be performed with only a most basic knowledge of the chemical structure of the ligands of interest when they are small molecules and with the help of genetic evidence when large macromolecules are used. The ability to predict computationally and examine the binding of molecules ranging from single amino acids to 60,000-kDa (or larger) proteins is an elegant testimony to the combined power of protein crystallography, computational biophysics, and molecular biology.

We thank Drs. David Goodsell and Art Olson of Scripps Clinic for making computer code available for the molecular-affinity automated docking protocol described in the text and for advice and assistance. We also thank Dr. Antony Dean for the results of kinetic experiments on IDH. Crystallographic coordinates for MBP were provided by Dr. Florante Quiocho. Preprints of genetic experiments on MBP were provided by Dr. Michael Manson. B.L.S. was supported as a fellow of the Helen Hay Whitney Biomedical Research Foundation during the research described in this paper; this work was also supported by the National Institutes of Health (Grant NIDDK 09765).

1. DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D. & Venkataraghavan, R. (1988) *J. Med. Chem.* **31**, 722–729.
2. Shoichet, B. K., Bodian, D. L. & Kuntz, I. D. (1992) *J. Comput. Chem.* **13**, 380–397.
3. Meng, E. C., Shoichet, B. K. & Kuntz, I. D. (1992) *J. Comput. Chem.* **13**, 505–524.
4. Kuntz, I. D. (1992) *Science* **257**, 1078–1082.
5. Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., MacRey, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991) *J. Chem. Inf. Comput. Sci.* **31**, 187–204.
6. Goodford, P. J. (1985) *J. Med. Chem.* **28**, 849–857.
7. Goodsell, D. S. & Olson, A. J. (1990) *Proteins: Struct. Funct. Genet.* **8**, 195–202.
8. Leach, A. R. & Kuntz, I. D. (1992) *J. Comput. Chem.* **13**, 730–748.
9. Wodak, S. J. & Janin, J. (1978) *J. Mol. Biol.* **124**, 323–342.
10. Cherfils, J., Duquerroy, S. & Janin, J. (1991) *Proteins: Struct. Funct. Genet.* **11**, 271–280.
11. Shoichet, B. K. & Kuntz, I. D. (1991) *J. Mol. Biol.* **221**, 327–346.
12. Stoddard, B. L. & Koshland, D. E., Jr. (1992) *Nature (London)* **358**, 774–776.
13. Bone, R., Silen, J. L. & Agard, D. A. (1989) *Nature (London)* **339**, 191–195.
14. Wilson, C., Mace, J. E. & Agard, D. A. (1991) *J. Mol. Biol.* **220**, 495–506.
15. Goodsell, D. & Dickerson, R. E. (1986) *J. Med. Chem.* **29**, 727–733.
16. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982) *J. Mol. Biol.* **161**, 269–288.
17. Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983) *Science* **220**, 671–680.
18. Hendrickson, W. A. & Konnert, J. H. (1980) in *Computing in Crystallography*, eds. Diamond, R., Ramaseshan, S. & Venkatesan, K. (Indian Inst. Sci., Bangalore), pp. 13.01–13.23.
19. Salemme, F. R. (1976) *J. Mol. Biol.* **102**, 563–568.
20. Warwicker, J. (1989) *J. Mol. Biol.* **206**, 381–395.
21. Hurley, J. H., Thorsness, P. E., Ramalingam, V., Helmers, N. H., Koshland, D. E., Jr., & Stroud, R. M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 8635–8639.
22. Hurley, J. H., Dean, A. M., Sohl, J. L., Koshland, D. E., Jr., & Stroud, R. M. (1990) *Science* **249**, 1012–1016.
23. Hurley, J. H., Dean, A. M., Thorsness, P. E., Koshland, D. E., Jr., & Stroud, R. M. (1990) *J. Biol. Chem.* **265**, 3599–3607.
24. Hurley, J. H., Dean, A. M., Koshland, D. E., Jr., & Stroud, R. M. (1991) *Biochemistry* **30**, 8671–8678.
25. Wang, E. A. & Koshland, D. E., Jr. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7157–7161.
26. Milburn, M. V., Prive, G. G., Milligan, D. L., Scott, W. G., Yeh, J., Jancarik, J., Koshland, D. E. & Kim, S.-H. (1991) *Science* **254**, 1342–1347.
27. Gardina, P., Conway, C., Kossmann, M. & Manson, M. J. (1992) *J. Bacteriol.* **174**, 1528–1536.
28. Kossmann, M., Wolff, C. & Manson, M. D. (1988) *J. Bacteriol.* **170**, 4516–4521.
29. Manson, M. D. & Kossmann, M. (1986) *J. Bacteriol.* **165**, 34–40.
30. Mowbray, S. L. & Koshland, D. E., Jr. (1987) *Cell* **50**, 171–180.
31. Milligan, D. & Koshland, D. E., Jr. (1991) *Science* **254**, 1651–1654.
32. Spurlino, J. C., Lu, G.-Y. & Quiocho, F. A. (1991) *J. Biol. Chem.* **266**, 5202–5219.