The Influence of Variations of Ligand Protonation and Tautomerism on Protein−Ligand Recognition and Binding Energy Landscape

Nikolay P. Todorov,*,† Philippe H. Monthoux,‡ and Ian L. Alberts†

De Novo Pharmaceuticals Ltd., Vision Park, Histon, Cambridge CB49ZR, United Kingdom, and Theory of Condensed Matter Physics Group, Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge CB3 0HE, United Kingdom

Received February 28, 2005

We investigate the influence of variations of ligand protonation and tautomeric states on the protein−ligand binding energy landscape by applying the concept of structural consensus. In docking simulations, allowing full flexibility of the ligand, we explore whether the native binding mode could be successfully recovered using a non-native ligand protonation state. Here, we consider three proteins, dihydrofolate reductase, transketolase, and α-trichosanthin, complexed with ligands having multiple tautomeric forms. We find that for the majority of protonation and tautomeric states the native binding mode can be recovered without a great loss of accuracy.

INTRODUCTION

Recent advances in computational methods for drug discovery have proved valuable when coupled to 3D structural information of receptors derived from X-ray crystallographic and NMR studies, and several success stories have been reported.[1] One of the areas in which significant computational advances have been realized is in the virtual screening of molecular databases.[2] These methods are aimed at facilitating the process of finding novel lead compounds and utilize algorithms for the molecular docking of a large number of organic compounds into the active sites of proteins of medical interest.[3−7] Ideally, such methods generate a reliable structure of the protein−ligand complex as well as the provision of an accurate estimate of the binding energy.

Molecular docking approaches require an understanding of the concept of receptor recognition, which represents a great challenge in molecular biology.[8] This is primarily a consequence of two factors. The first is the rugged nature of the protein−ligand binding energy landscape arising from the competition between protein−ligand interactions, solvent effects, and entropy changes, which causes structural prediction to be a very difficult task.[9] The second issue concerns the nature of the energy functions used in molecular docking. It is not feasible to use exact energy functions, and thus, approximations are incorporated to rapidly quantify the molecular binding energy during the simulations. As a result of the lack of a complete comprehension of the physics and principles of protein−ligand binding, the accuracy of these approximate functions could be limited.[10−12]

Theories to describe energy landscapes have been actively developed in the field of protein folding with the aim of understanding the general principles, which lead to a protein adopting its native fold.[13,14] In this viewpoint, naturally occurring proteins have a prominent energy gap in their energy spectrum between the lowest energy structure and alternative conformations[15] and their energy surface is funnel-shaped.[16−19] These properties are important not only for the thermodynamic stability but also for the kinetic accessibility of the ground state within a reasonable biological time scale.[20] Non-native-like proteins do not exhibit such energy landscape, spectrum, and folding behavior. The folding process is dependent on the "frustration" of the system. A frustrated energy landscape has many local minima, arising from competing geometric and energetic features[21] that could reduce the efficiency of the folding process. Naturally occurring proteins are believed to follow pathways of minimal frustration and efficiently fold to their native state.[16,22]

It has been suggested that protein−ligand molecular recognition bears notable similarities to protein folding in terms of the energy landscape and spectrum.[23] Ligands that are native to the complex (receptor-specific ligands) exhibit minimally frustrated pathways on their energy landscape, leading to a stable binding mode, while non-native ligands will span a frustrated landscape, leading to many alternative binding modes.[24] Native ligands should fulfill both thermodynamic stability criteria and kinetic accessibility in terms of a funnel-shaped binding energy landscape. The latter property has been analyzed in terms of a *structural consensus* derived from multiple docking simulations. This descriptor allows differentiation between nativelike and non-native-like ligands, because many random start conformations of the native ligand regularly dock to the stable binding mode, whereas nonreceptor specific ligands will proceed to alternative metastable binding modes. It has been observed that the structural consensus has a better discriminating ability than the absolute binding energy.[24] Considerations from energy landscape theory have also led to the derivation of a criterion for optimal ligand binding specificity[25] requiring the maximization of the ratio of the energy gap and the energy variation.

* Corresponding author phone: +44 1223 238058; fax: +44 1223 238088; e-mail: nikolay.todorov@denovopharma.com.
† De Novo Pharmaceuticals Ltd.
‡ University of Cambridge.

10.1021/ci050071n CCC: $33.50 © 2006 American Chemical Society
Published on Web 03/10/2006

The successful identification of native ligands and their binding modes in protein active sites depends crucially on the quality of the approximations employed to characterize the form of the real protein−ligand binding energy landscape. Significant effort has been focused on the development of approximate functions for describing protein−ligand interactions.[10−12] The performance of such scoring functions has a critical impact on the success of molecular docking approaches; however, it is well-known that the accuracy of such functions can vary when applied to different molecular systems. It has been suggested that appropriate approximate functions should conserve the funnel-shaped energy landscape in order to enhance the efficiency of the docking search algorithm as well as correctly identify the native binding mode as the lowest-energy or global minimum structure. The results of molecular docking simulations typically find that the native binding mode corresponds to a low-energy structure on the binding energy landscape, but not necessarily the lowest energy configuration.[10−12] This observation implies that the success rate for finding the native binding mode will be increased when a small set of unique binding modes within a certain energy threshold above the lowest energy state is considered.[26]

Tautomeric and protonation states involve only small variations of the ligand topology compared to those of distinct ligands. This leads to the following question: how important is the influence of these structural differences on the binding energy landscape, the thermodynamic and kinetic accessibility of the ground state, and most importantly, the prediction of the native binding mode? Multiple protonation and tautomeric forms of molecules have rarely been addressed in molecular docking approaches and the investigation of protein−ligand interactions.[27] Here, we examine the effect of the neglect and incorporation of these structural features on native binding mode determination and the practical approximation of the exact binding energy. For each ligand, we generate multiple protonation and tautomeric states and conduct multiple docking simulations for each state within the corresponding protein binding site. Our aim is to assess whether the native binding mode could be found for any protonation state or whether there are more strict requirements. We find that for the majority of protonation and tautomeric states the native binding mode can be recovered without a great loss of accuracy. We relate this result to the close similarity of the binding energy landscapes of different tautomers.

METHODS

Tautomer Generation. Tautomeric and protonation states of the ligands were generated using a previously described more-general atom-assignment method,[28] tuned to produce only the tautomeric and ionic forms of the specified ligand.

The atom-assignment method uses single-atom templates (a library of 28), with defined atom types and bond orders. A template consists of a central atom and the bonds incident upon it; bond orders and dummy electron lone pairs are specified in the description of the template. Atom types frequently found in drug molecules are considered in several geometric and ionization states.

Each node in the ligand structure is examined, and a set of matching templates from the library is compiled. The geometric states and atomic identities of the template and the node of the ligand structure must match; for example, planar carbon-atom templates are assigned to planar carbon atoms in the ligand structure and so forth. Hydrogen atoms and electron lone pairs in the ligand structure are matched by both hydrogen atoms and lone pairs from the templates.

Next, atom types are assigned exhaustively using a depth-first search over the templates assigned to each atom. Initially, the first template of the first atom is assigned, followed by the assignment of the first template of the second atom, and so on. When a new template is considered, a check is made to ascertain that the orders of the bonds that are suggested are not in conflict with the orders of the bonds previously assigned. If a template is rejected, the next of the templates available for the vertex is tested. If none of the suggestions are accepted, the algorithm backtracks to the previous atom and replaces the currently used template with the next in the list, and the same procedure is repeated. The search continues until all possible structures are generated.

Degenerate structures can result from symmetry in the ligand. The most common situation, where symmetry plays a role, is where rotationally invariant terminal atoms are set to either hydrogen atoms or lone pairs. Also, alternative patterns on conjugated aromatic systems result in equivalent structures. Structures are canonically numbered, and the unique representatives are identified.

Thus, following the protocol of the method, hybridization states of the non-hydrogen atoms were kept fixed, protonation patterns of nitrogen and oxygen atoms were changed combinatorially, and unique protonation and tautomeric forms were generated and collected.

Protein−Ligand Docking. Search methods in molecular docking must rapidly explore the energy landscape of the protein−ligand interaction, and to this end, a genetic-algorithm-based method GOLD (version 2.0) was used to perform ligand docking[6,7] with standard parameter configurations as specified by Nissink et al.[29] In incorporating the genetic algorithm (see Holland[30] and Goldberg[31]) into molecular docking applications, the population consists of ligand molecules and the position, orientation, and the conformation of each ligand copy is encoded into the chromosome as well as the orientations of rotatable hydroxyl groups in the protein. Once the chromosomes are decoded, the interaction energy with the receptor is calculated, as well as the internal ligand energy, and used to calculate the fitness score. Recombination and mutations are used to create the individuals forming the next generation, and as the population is evolved for a specified number of generations, the fitness score gradually increases.

Binding Mode Prediction. To assess the importance of the tautomeric state of the ligand for successful docking and recognition, we determine the number of tautomers yielding low-energy docking solutions closer than specified root-mean-square-deviation (RMSD) values to the crystal configuration of the ligand, that is the number of tautomers for which a docking solution is found within a given RMSD to the ligand crystal structure at a specified binding-energy threshold above the lowest-energy solution. Several energy thresholds above the lowest-energy solution are adopted to account for the uncertainty in the binding energy values as a consequence of the approximations used.

Table 1. Protein−Ligand Crystal Complexes

PDB code	protein	ligand	resolution	reference
4dfr	dyhydrofolate reductase	methotrexate	1.7 Å	27
1trk	transketolase	thiamine diphosphate	2.0 Å	28
1mrk	α-trichosanthin	formycin	1.6 Å	29

Lowest Binding Energy. To gain further insight into the changes of the binding energy landscape concurrent with the variations in the protonation and tautomeric states of the ligand, for each tautomer, we investigated the behavior of various descriptors of the landscape: the lowest binding energy, the stability gap in the energy spectrum, and the structural consensus.[24]

Energy Spectrum. The energy spectrum is generated by applying the concept of energetic collapse.[23,24] In this approach, the lowest-energy docked structure is taken as the reference binding mode. The energy values for other docking solutions that are within a certain RMSD cutoff of the reference mode are collapsed to the energy value associated with the reference mode. Of the remaining docking solutions, the structure with the next lowest energy represents an alternative binding mode, and the collapsing procedure is conducted relative to this binding mode. The process is repeated until all docking solutions have been considered.

Energy Gap. The energy gap for each tautomer is calculated as the difference between the lowest-energy binding mode and the average energy of all of the binding modes as designated by the energetic collapse procedure. A higher value for the energy gap implies greater thermodynamic stability of the lowest-energy binding mode.

Structural Consensus. The structural consensus of the docking simulations for each protonation state and tautomer is derived from the average RMSD of the docked structures relative to a reference structure, which can correspond to either the lowest-energy docking solution or the crystal structure of the ligand. The structural consensus is defined by

$$\langle C \rangle_{N/M} = \sum_{i=1,N} sqrt(\Delta r_i^2)/(N-1)$$

where $sqrt(\Delta r_i^2)$ represents the RMSD deviation of docked solution i relative to the reference structure. The RMSD deviations of the lowest N energy docked structures out of M docking simulations are incorporated in the summation. In this work, $M = 100$ and $N = 50$ were used, giving a structural consensus value denoted as $\langle C \rangle_{50/100}$.

RESULTS

Three high-resolution protein−ligand complexes (Table 1) were selected from the GOLD/Astex data set[29] for which the ligands had a large number of tautomeric and ionization states. The complex with Protein Databank (PDB) code 4dfr[32] contains dihydrofolate reductase from *Escherichia coli* cocrystallized with ligand methotrexate. Dihydrofolate reductase catalyzes the reduction of dihydrofolate to tetrahydrofolate. This is an important reaction in the biosynthetic pathway of nucleic acids. The complex 1trk contains transketolase from *Saccaromyces cerevisiae* crystallized with cofactor thiamine diphosphate.[33] Transketolase is involved in the nonoxidative steps of the pentose phosphate pathway. Complex 1mrk contains α-trichosanthin, which is a ribosome-inactivating protein,[34] complexed with formycin.

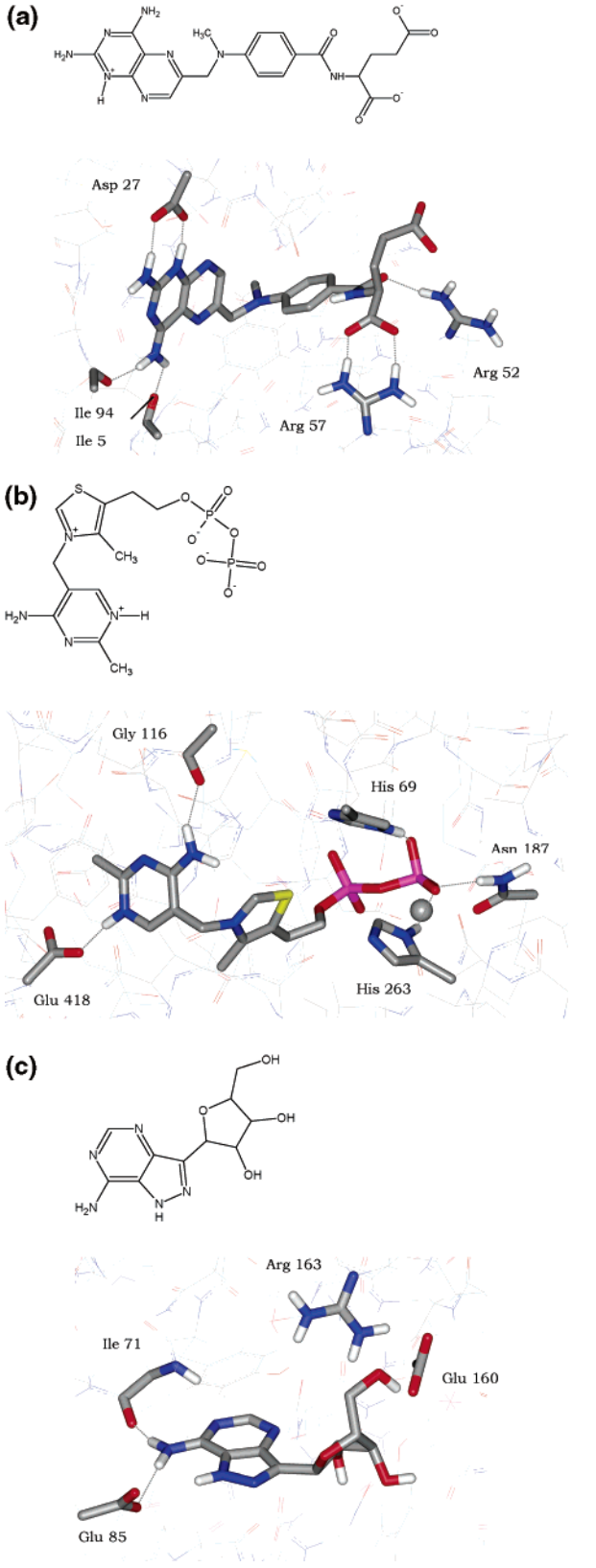

Figure 1. Structures of ligands used in this study: (a) methotrexate, (b) thiamine diphosphate, and (c) formycin.

The three ligands are depicted in Figure 1, where some important residues and hydrogen bonds are highlighted as

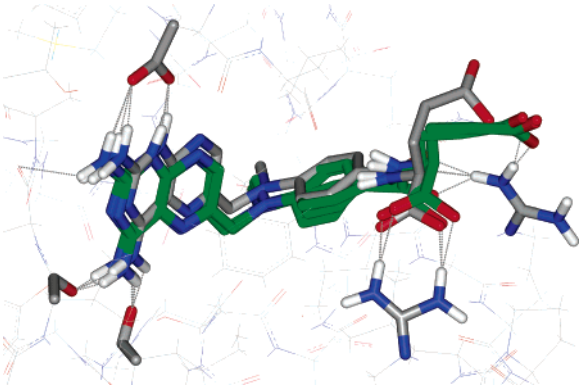

Figure 2. Top three solutions (green) for the native form of methotrexate and the crystal form.

well. Methotrexate makes hydrogen bonds with Ile 5, Asp 27, Arg 52, Arg 57, and Ile 94 and also with water molecules HOH405 and HOH403, which, however, had been removed from the structure following the docking protocol.[29] There are also a number of hydrophobic residues in the vicinity of the ligand (Ala 6, Ala 7, Leu 28, Ile 50, and Leu 54). Thiamine diphosphate is involved in hydrogen bonding with His 69, Gly 116, Asn 187, and His 263 and also binds to a Ca atom in the site. There are also a number of hydrophobic residues present in the site, for example, Ala 33, Ile 191, Pro 117, Thr 190, and Ile 250. Formycin makes hydrogen bonds with Ile 71 and Glu 85; only very few of the potential hydrogen bonds it has have the capacity to form. Some protein residues, namely, Glu 160 and Arg 163, are close to the ligand and offer alternative hydrogen-bonding possibilities. Some hydrophobic residues are also present in the site (Ile 155 and Phe 83).

The ligands were extracted from the Protein Databank files, and sets of tautomers were generated. A total of 192 tautomers and protomers were generated for methotrexate (4dfr), 42 for thiamine diphosphate (1trk), and 32 for formycin (1mrk). We use as a reference the tautomers suggested in the original publications describing the crystal complexes[32−34] and used by Nissink et al.[29] in the docking study and refer to them as the native tautomers.

Each ligand tautomer is docked 100 times to the corresponding protein, with GOLD parameter settings of 3000 generations and populations of 100 individuals, that is, 30 000 genetic operations and pose evaluations per run as suggested by Nissink et al.[29] These settings produced consistent results with well-populated clusters of solutions for all ligand tautomers. For example, Figure 2 shows the top three solutions for the native form of methotrexate. The binding mode differs only slightly from the crystal; the ligand conformations have changed in order to make an extra hydrogen bond to Arg 52. Experiments using 50 000 or 100 000 genetic operations per docking run, performed for several of the complexes, did not produce different binding modes or lower energies.

Binding Mode Prediction. The distributions of binding modes and docking solutions for the complexes are shown in Table 2. Each cell in the table lists the number of tautomers for which a docking solution is found within a given RMSD to the ligand crystal structure at a specified binding-energy threshold above the lowest-energy solution. Different energy thresholds above the lowest-energy solution are adopted to account for the uncertainty in the binding energy values as a consequence of the approximations used.

Table 2. Distribution of Docking Solutions[a]

ΔE/RMSD	1.0 Å	2.0 Å	3.0 Å
0.0 kJ/mol	7, 2, 1	178, 37, 5	192, 42, 24
1.0 kJ/mol	16, 4, 2	183, 38, 17	192, 42, 28
2.0 kJ/mol	32, 9, 5	188, 38, 22	192, 42, 31
3.0 kJ/mol	49, 13, 8	189, 38, 24	192, 42, 32
4.0 kJ/mol	75, 16, 11	190, 38, 24	192, 42, 32
5.0 kJ/mol	99, 16, 14	191, 40, 25	192, 42, 32

[a] Each cell in the table lists the number of tautomers (4dfr, 1trk, 1mrk) for which a docking solution is found within a given RMSD to the crystal (e.g., 2.0 Å) and a given binding-energy threshold above the lowest (e.g., 3.0 kJ/mol).

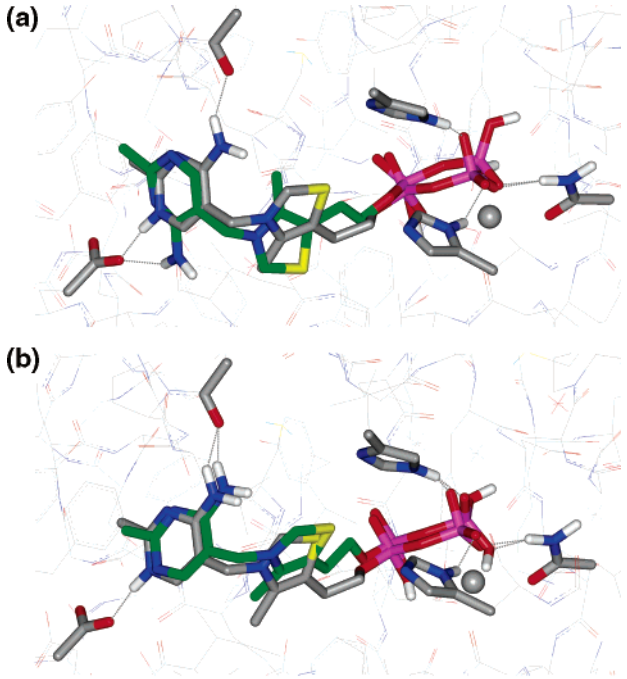

Figure 3. Docking solutions of thiamine diphosphate (1trk): (a) best-ranking solution and (b) fifth-best-ranking solution.

For 4dfr and 1trk, the best-ranking docking solutions (the row with $\Delta E = 0.0$ kJ/mol in Table 2) of all of the tautomers were within 3.0 Å of the native binding mode and for the majority of the tautomers within 2.0 Å. The variation of ligand protonation and tautomeric states has a limited effect on protein−ligand recognition for these complexes. The results for the tautomers of the 1mrk ligand, formycin, were less clearly cut: eight out of the 32 tautomers had their best-ranking solution with an RMSD above 3.0 Å. Of these, we note that the lowest-energy solution for one of the tautomers had an RMSD value greater than 6.5 Å; the other seven were within 4.0 Å.

However, if a small tolerance in the binding energy values is allowed by considering solutions within 3.0 kJ/mol of the lowest-energy docked structure, we find solutions within 3.0 Å of the native crystal structure binding mode for all of the tautomers of all three complexes, with a majority within 2.0 Å. For example, Figure 3 shows the best and the fifth-best binding mode for thiamine diphosphate (1trk) with energy difference of less than 1 kJ/mol. While solutions 1−4 deviate from the crystal binding mode, solution 5 is close and within a very small energy threshold. Given that the function

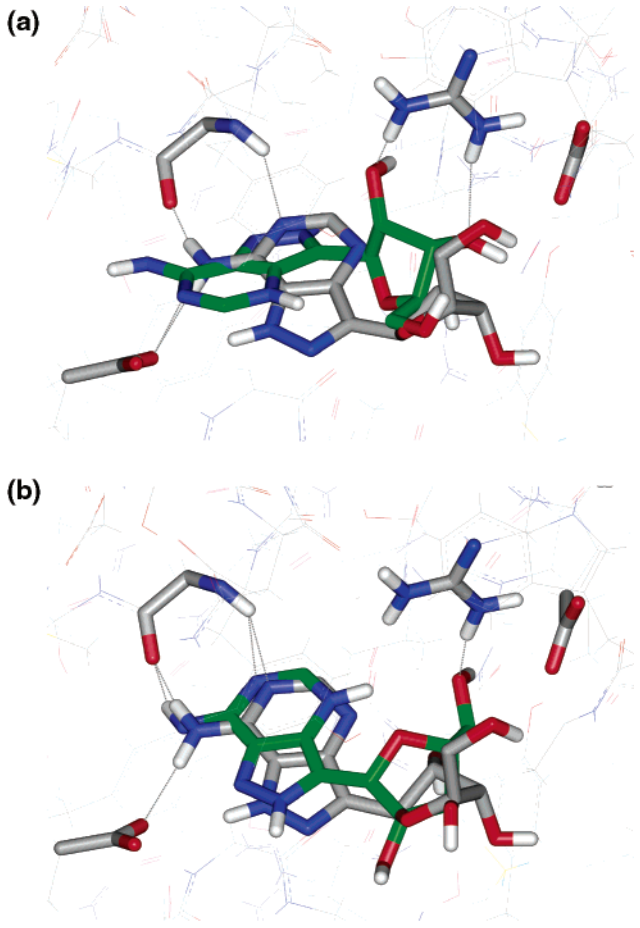

Figure 4. Docking solutions of formycin (1mrk): (a) best-ranking solution and (b) second-best-ranking solution.

deployed to calculate the binding energy is not exact, our results suggest that the energy landscape does not change much from tautomer to tautomer. The changes in the landscape we observe are within the accuracy that an approximate energy function can achieve. Even when the native tautomer state of a ligand is docked, solutions within similar binding energy thresholds still have to be considered in order to represent the native binding mode accurately. For a minority of tautomers, the RMSD did not get too close to the crystal binding mode, even when the energy threshold is allowed (Figure 4 shows two low-energy solutions for a formycin tautomer), but generally, it appears that protein−ligand recognition is to a large extent insensitive to the changes in the ligand tautomeric and protonation states.

How does the ligand manage to maintain the binding mode? An examination of the docking results indicates that most of the protein−ligand hydrogen-bond pairs are kept intact in the majority of the tautomers. This constrains the ligand to adopt a certain binding mode. It is possible to maintain the same hydrogen-bond network because many of the hydrogen-bond groups in the ligand (i) have the ability to form more than one hydrogen bond (e.g., hydroxyl), but only one of the potential hydrogen bonds is utilized in the crystal mode, and (ii) have conformational flexibility, and (iii) the terminal hydrogen-bonding groups are involved in the interaction. When the tautomeric or protonation form of the group is changed, the group could accommodate the change. If a sufficient number of hydrogen bonds is maintained, a small number of mismatched pairs could be

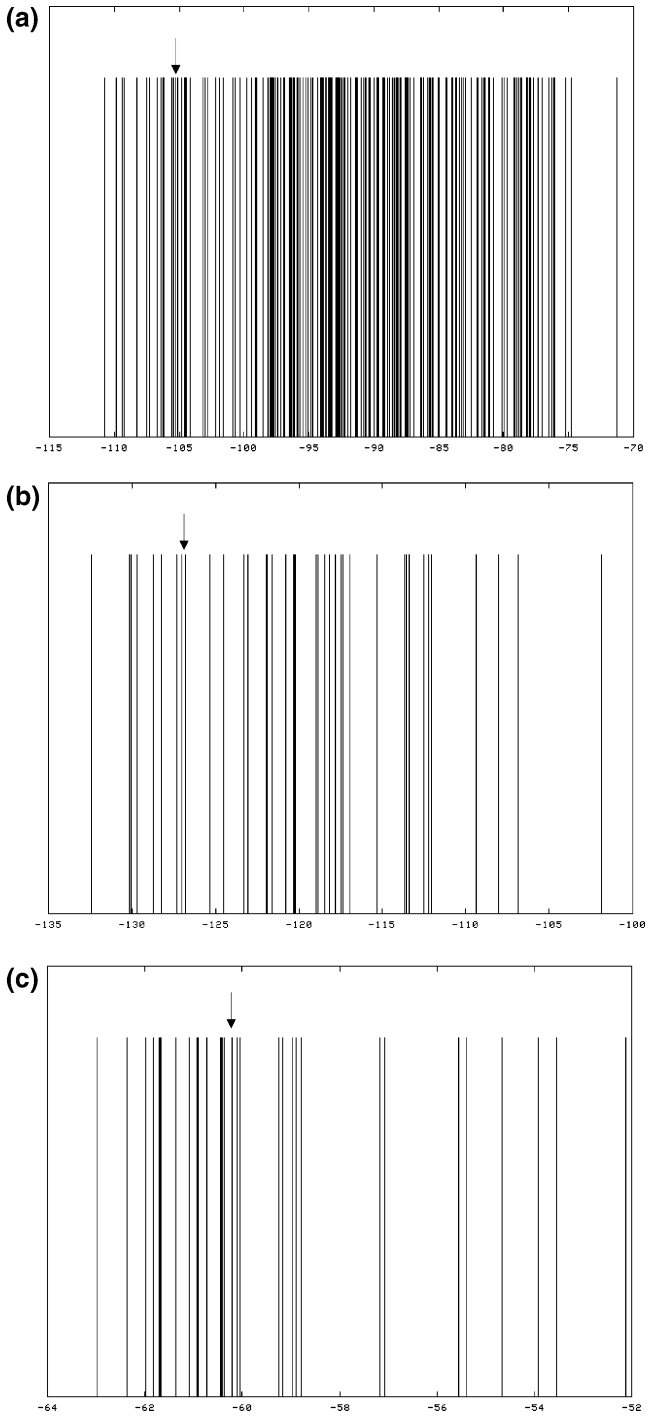

Figure 5. Lowest-energy distributions: (a) 4dfr (−106.67), (b) 1trk (−127.02), and (c) 1mrk (−60.45). The values in parentheses are the values for the native forms.

tolerated without a change in the binding mode (Figure 3b). This is also helped by the inaccuracies in the accounting for hydration effects in the scoring function. Formycin, on the other hand, is more likely to adopt alternative binding modes because (i) some of the terminal hydrogen-bonding groups are not involved, which results in a smaller number of hydrogen bonds in the crystal, (ii) all potential hydrogen bonds in those groups are utilized, and (iii) the groups are less flexible (Figure 4).

Lowest Binding Energy. The distributions of lowest binding energy values for all tautomers of the three complexes are shown in Figure 5. The general observation is

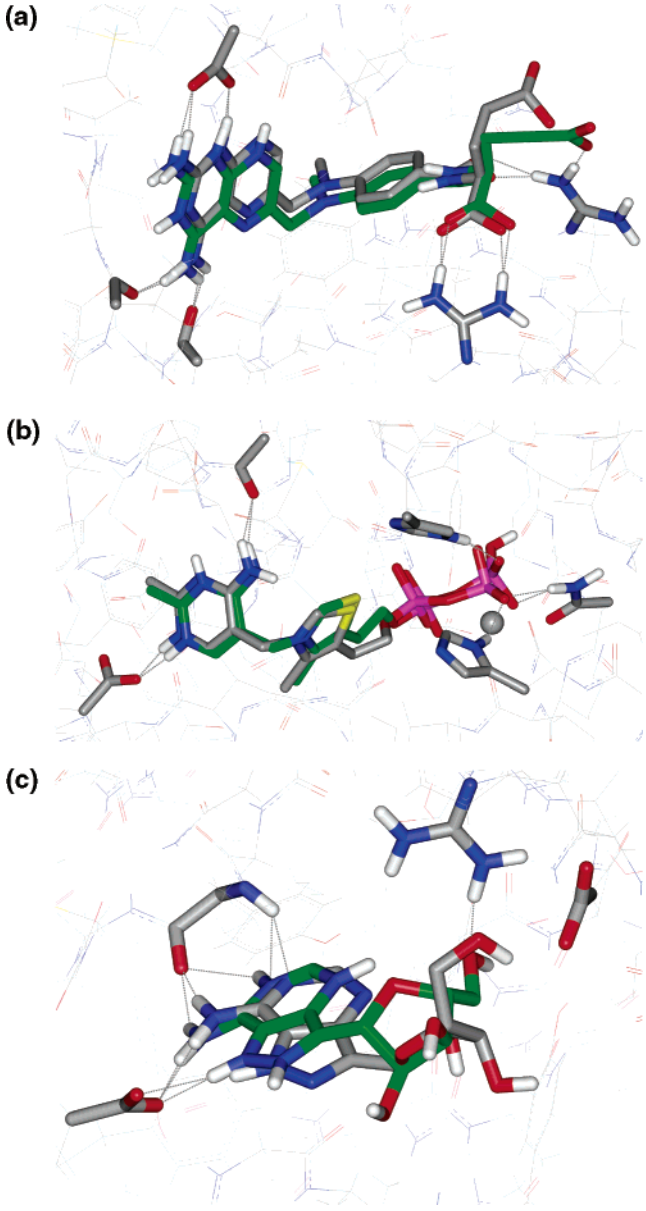

Figure 6. Tautomers with lowest energy (green): (a) methotrexate, (b) thiamine diphosphate, and (c) formycin.

that this parameter does not distinguish one dominant tautomer from the other forms. In all three cases, the values for the suggested native forms lie within about 5.0 kJ/mol of the lowest-energy tautomer. For example, in the case of methotrexate, the lowest binding energy ranges from about −110 to −70 kJ/mol, whereas the suggested native form yields a value of −106.7 kJ/mol (Figure 5a). If we examine the tautomers with the lowest energies (Figure 6), we note the contribution of the following factors: (i) the same hydrogen-bonding pattern as that for the suggested native tautomers is maintained; (ii) the differences in the protonation pattern occur at positions where no hydrogen bonds are made with the protein and, thus, have only a small effect on the binding energy; (iii) there is an involvement of groups with more potential hydrogen bonds than are actually made in the crystal; (iv) there is a flexibility of these hydrogen-bonding groups.

It is interesting to note that the positions where the protonation pattern in methotrexate differs between the

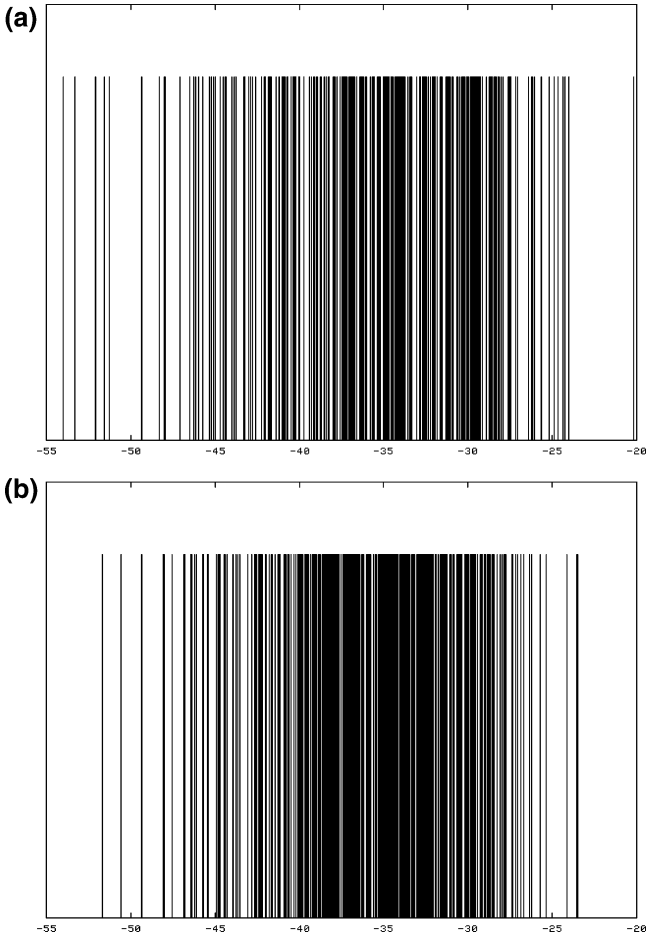

Figure 7. Energy spectra of two tautomers of formycin: (a) the native form and (b) the form with the highest RMSD of the lowest-energy solution.

suggested native form and the tautomer with the lowest-energy are involved in interactions with water molecules that have been neglected in the docking protocol.[29]

Energy Spectrum. As an illustration of the energy spectra generated by this procedure, we chose to depict the spectra of the two tautomers of formycin that were most likely, in our estimation, to exhibit significant differences: the native form of formycin and the tautomer with the highest RMSD for the lowest-energy solution relative to the ligand crystal structure (6.71 Å in this case vs 1.97 Å for the lowest-energy docked solution of the suggested native form). The representation of the spectra from the docking simulations was rather poor in the case of 1mrk. The docking runs are focused on the low-energy states, and higher states are not sampled sufficiently. In fact, less than 10 unique solutions were generated for all of the tautomers of formycin. This is because the formycin ligand, in this case, is quite rigid, and most of the docking attempts converge to similar solutions. To obtain spectra for the two chosen tautomeric forms of formycin with an improved representation of the higher energy states, we ran 3000 short docking simulations (250 generations, populations of 25 individuals). The spectra (Figure 7) appeared consistent with the random-energy model[16,17] with discrete low-energy states and a continuous band of higher-energy solutions. The spectra for the tautomeric forms of 4dfr and 1trk using the original GOLD docking settings showed more unique solutions than for 1mrk (over 20 solutions for each tautomer of 1trk and over 50 for

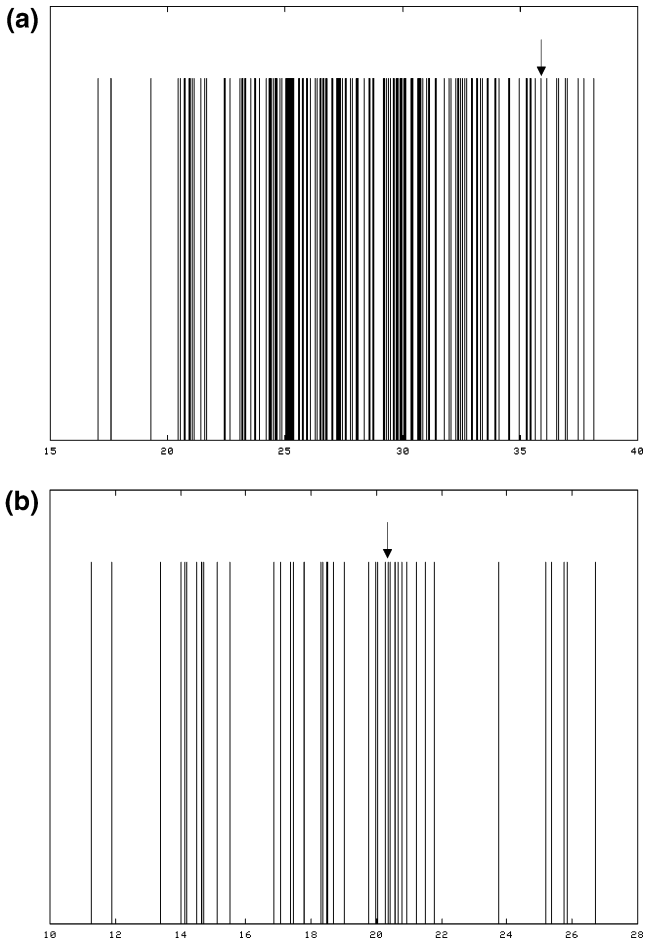

Figure 8. Energy-gap distributions: (a) 4dfr (36.99) and (b) 1trk (20.43). The values in parentheses are the values for the native forms.

4dfr, which is the most flexible of the three ligands) but still required better sampling of the higher energy states.

Energy Gap. Figure 8 shows the distributions for all tautomers of 4dfr and 1trk. The corresponding plot for 1mrk could not be presented because an insufficient number of states were generated to allow a significant distribution of the energy values, as discussed above. The energy gaps for the native forms of 4dfr and 1trk rank quite highly, although there are higher-ranking tautomers in each case. For 4dfr, the energy gap ranges from about 16 to 38 kJ/mol, with a value of 36.99 kJ/mol for the native form (Figure 8a); the range for 1trk is 11−27 kJ/mol, and the native structure value is 20.43 kJ/mol (Figure 8b).

Structural Consensus. The distributions of the structural consensus $\langle C \rangle_{50/100}$ values, using the lowest energy structure as a reference, for all tautomers for each of the three complexes are displayed in the distributions in Figure 9. A low value of $\langle C \rangle_{50/100}$ implies that docking simulations tend to converge to a common solution. The range of $\langle C \rangle_{50/100}$ values is relatively narrow in each case (less than 3.0 Å), and in general, the $\langle C \rangle_{50/100}$ structural consensus for the native form shows a small preference for this state. For example, in the case of 4dfr, $\langle C \rangle_{50/100}$ values range from 1.2 to 3.5 and the native form gives a value of 1.58 (Figure 9a). Similar results are obtained if the crystal structure of the ligand is used as the reference structure for evaluating the structural consensus values (data not shown).

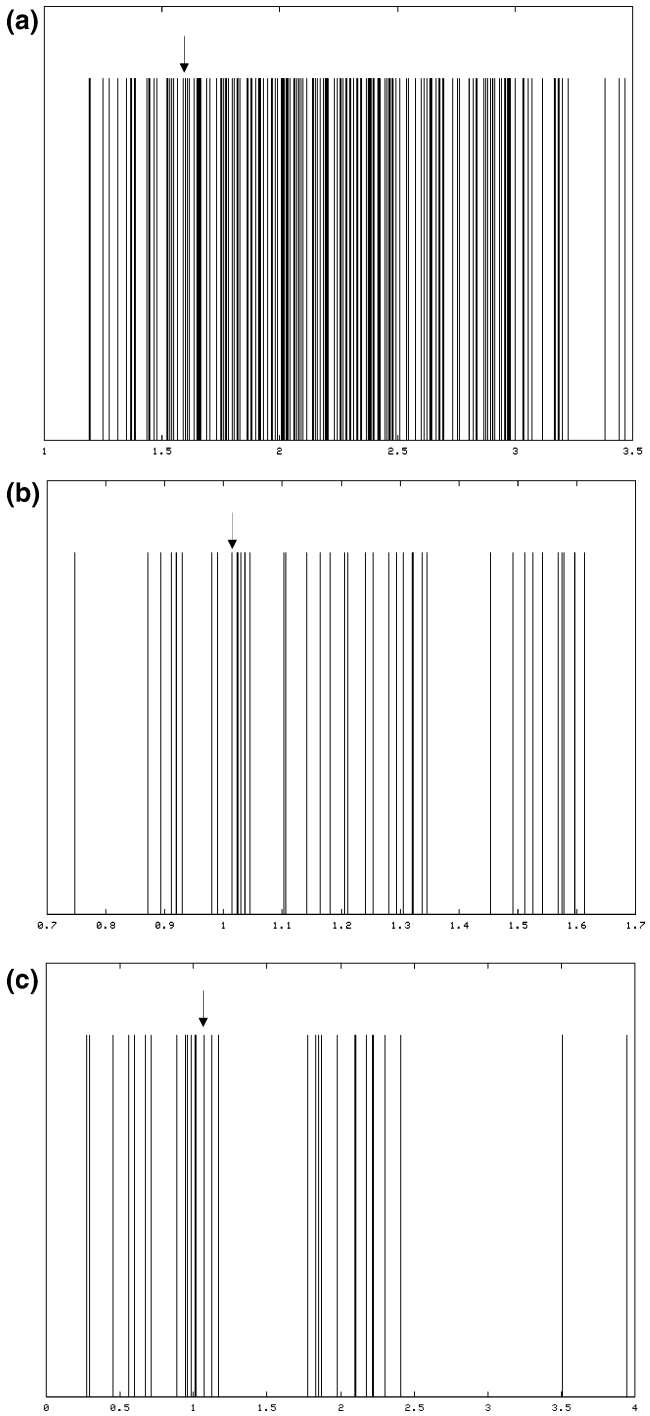

Figure 9. Structural-consensus $\langle C \rangle_{50/100}$ distributions: (a) 4dfr (1.58), (b) 1trk (1.01), and (c) 1mrk (1.13). The values in parentheses are the values for the native forms. The lowest-energy solution is used as a reference structure.

Figure 10 shows the variation of $\langle C \rangle_{N/100}$ values for the suggested native tautomeric form of methotrexate for N ranging from 1 to 100. In this plot, $\langle C \rangle_{N/100}$ has a relatively low value early in the series until about $N = 35$, from where a sharp increase is observed, which is a result of pronounced structural deviations in the docked solutions. The value of the structural consensus parameter depends not only on the shape of the energy landscape but also on the efficacy of the configuration space search method used for docking. To explore the latter aspect, docking simulations were conducted for the native methotrexate tautomer with both a larger and

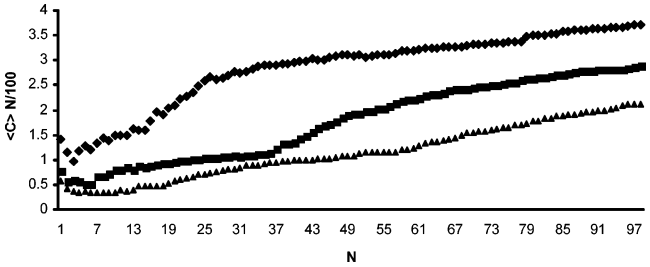

Figure 10. $\langle C \rangle_{N/100}$ values for the methotrexate native state with different numbers of genetic operation values in GOLD: 20 000 (diamond), 30 000 (square), and 50 000 (triangle).

a smaller number of genetic operations, and the $\langle C \rangle_{N/100}$ series derived from these runs are also shown in Figure 10. We note that the longer simulation (i.e., the increased number of genetic operations) leads to lower structural consensus values with a less-pronounced steepness in the curve, which arises later in the series. This suggests that more exhaustive docking simulations tend to converge to a common binding mode.

CONCLUSIONS

In this work, we investigated the dependence of docking results and ligand−receptor specificity on the tautomeric and protonation states of the ligand. Multiple tautomers and ionization states of three ligands, methotrexate, thiamine diphosphate, and formycin, were generated and docked into the corresponding receptor binding sites.

The binding modes arising from the docking simulations were analyzed. The RMSD values of the low-energy solutions were found to be, in general, independent of the choice of tautomer used in the docking procedure. The modal RMSD range is 1−2 Å for methotrexate and thiamine diphosphate and 2−3 Å for formycin, relative to the corresponding ligand crystal structure. Most of the protein−ligand hydrogen-bond pairs are kept intact in the majority of the tautomers and, thus, restrict the ligand to maintain the binding mode. This is due to the fact that many of the hydrogen-bond groups in the ligand are flexible and have the ability to form several hydrogen bonds, but not all of the potential hydrogen bonds are utilized in the crystal. The terminal hydrogen-bonding groups are often involved in the interaction and constrain the binding mode the ligand can adopt.

The underlying protein−ligand binding energy landscapes for the three complexes were characterized by several descriptors: the lowest-energy docking solution, the energy gap, and the structural consensus. We found some variation in the values of these descriptors over all of the tautomeric states. The tautomers with the lowest energies have the same hydrogen-bonding pattern as that of the suggested native tautomers, and the differences in the protonation pattern occur at positions where no hydrogen bonds are made with the protein and, thus, have only a limited effect on the binding energy. Also, we note the involvement of flexible hydrogen-bonding groups capable of forming a greater number of hydrogen bonds than that actually made in the crystal.

The binding specificity,[25] defined as the ratio of the energy gap and the standard deviation of the energies of all of the unique binding modes, was also determined and again showed no preference for the suggested native form. Furthermore, the energy spectra for all of the tautomers of the three complexes were quite similar and appeared to be described by the classical random-energy model.[16,17] From these results, we conclude that the above descriptors do not show any considerable preference for any particular tautomer. Given the diversity of the three ligands and the different number of tautomeric forms in each case, we suggest that this may be a general phenomenon. It appears that the degree of frustration on the energy landscape of alternative tautomeric forms is quite similar. These results are compatible with the premise of a minimally frustrated funnel-shaped energy landscape with few low-energy metastable states, such that complex formation consistently proceeds to the stable binding mode irrespective of the starting configuration and protonation state. This is believed to be related to the kinetic accessibility of the stable binding mode and suggests that there are multiple routes along the molecular recognition landscape leading to the native binding configuration. The thermodynamic stability and kinetic accessibility of the native binding mode are not significantly disrupted by the variations in the tautomeric form of the ligand.

This work may have important consequences for future drug discovery studies. A ligand participating in strong interactions with a receptor binding site does not necessarily entail a significant energy gap to alternative binding modes. Effective site-specific (native) ligand binding appears to necessitate the existence of a stable, unique native structure in conjunction with a minimally frustrated energy landscape which is quite insensitive to changes of tautomeric and protonation states.

REFERENCES AND NOTES

(1) Leach, A. R. *Molecular Modelling: Principles and Applications*. Adison-Wesley Longman: Reading, Massachusetts, 1996.

(2) Brooijmans, N.; Kuntz, I. D. Molecular Recognition and Docking Algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335−373.

(3) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule−Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(4) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(5) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: Fast, Fully Flexible Automated Docking of Flexible Ligands to Protein Binding Sites. *Chem. Biol.* **1996**, *3*, 449−462.

(6) Jones, G.; Willet, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.

(7) Jones, G.; Willet, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(8) McCammon, J. A. Theory of Biomolecular Recognition. *Curr. Opin. Struct. Biol.* **1998**, *8*, 245−249.

(9) Totrov, M.; Abagyan, R. Detailed ab initio Prediction of Lysozyme−Antibody Complex with 1.6 Å Accuracy. *Nat. Struct. Biol.* **1994**, *1*, 259−263.

(10) Stahl M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(11) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(12) Pérez, C.; Ortiz, A. R. Evaluation of Docking Functions for Protein−Ligand Docking. *J. Med. Chem.* **2001**, *44*, 3768−3785.

(13) Schymkowitz, J. W. H.; Rousseau, F.; Serrano, L. Surfing on Protein Folding Energy Landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15846−15848.

(14) Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Rokhsar, D. S. Pathways for Protein Folding: Is a New View Needed? *Curr. Opin. Struct. Biol.* **1998**, *8*, 68−79.

(15) Shakhnovich, E.; Gutin, A. Formation of Unique Structure in Polypeptide Chains: Theoretical Investigation with the Aid of a Replica Approach. *Biophys. Chem.* **1989**, *34*, 187−199.
(16) Bryngelson, J. D.; Wolynes, P. G. Spin Glasses and the Statistical Mechanics of Protein Folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524−7528.
(17) Bryngelson, J. D.; Onochic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 167−195.
(18) Lazaridis, T.; Karplus, M. "New View" of Protein Folding Reconciled with the Old Through Multiple Unfolding Simulations. *Science* **1997**, *278*, 1928−1931.
(19) Dobson, C. M.; Karplus, K. The Fundamentals of Protein Folding: Bringing Together Theory and Experiment. *Curr. Opin. Struct. Biol.* **1999**, *9*, 92−101.
(20) Levinthal, C. Are There Pathways for Protein Folding? *J. Chem. Phys.* **1968**, *65*, 44−45.
(21) Shea, J.-M.; Onuchic, J. N.; Brooks, C. L. Probing the Folding Free Energy Landscape of the src-SH3 Protein Domain. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16064−16068.
(22) Onuchic, J. N.; Wolynes, P. G.; Luthey-Shulten, Z.; Socci, N. D. Towards an Outline of the Topography of a Realistic Protein Folding Funnel. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3626−3630.
(23) Verkhivker, G. M.; Rejto, P. A. A Mean Field Model of Ligand−Protein Interactions: Implications for the Structural Assessment of Human Immunodeficiency Virus Type 1 Protease Complexes and Receptor-Specific Binding. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 60−64.
(24) Rejto, P. A.; Verkhivker, G. M. Unraveling Principles of Lead Discovery: From Unfrustrated Energy Landscapes to Novel Molecular Anchors. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8945−8950.
(25) Wang, J.; Verkhivker, G. M. Energy Landscape Theory, Funnels, Specificity and Optimal Criterion of Biomolecular Binding. *Phys. Rev. Lett.* **2003**, *90*, 188101.
(26) Källblad, P.; Mancera, R. L.; Todorov, N. P. Assessment of Multiple Binding Modes in Ligand−Protein Docking. *J. Med. Chem.* **2004**, *47*, 3334−3337.
(27) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in Computer-Aided Drug Design. *J. Recept. Signal Transduct. Res.* **2003**, *23*, 361−371.
(28) Todorov, N. P.; Dean, P. M. A Branch-and-Bound Method for Optimal Atom-Type Assignment in de novo Ligand Design. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 335−349.
(29) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L. J. C.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein−Ligand Interaction. *Proteins* **2002**, *49*, 457−471.
(30) Holland, J. H. *Adaptation in Natural and Artificial System*. University of Michigan Press: Ann Arbor, Michigan, 1975.
(31) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*. Adison-Wesley: Reading, Massachustts, 1989.
(32) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. Crystal Structures of *Escherichia coli* and *Lactobacillus casei* Dihydrofolate Reductase Refined at 1.7 Å Resolution. I. General Features of Binding of Methtrexate. *J. Biol. Chem.* **1982**, *257*, 13650−13562.
(33) Nikkola, M.; Lindqvist, Y.; Schneider, G. Refined Structure of Transketolase from *Saccaromyces cerevisiae* at 2.0 Å Resolution. *J. Mol. Biol.* **1994**, *238*, 387−404.
(34) Huang, Q.; Liu, S. Y.; Tang, Y.; Jin, S.; Wang, Y. Studies on Crystal Structures, Active-Centre Geometry and Denaturating Mechanism of Two Ribosome Inactivating Proteins. *Biochem. J.* **1995**, *309*, 285−298.

CI050071N