Appendix I

Combinatorial Methods for Small Molecule Placement in Computational Enzyme Design

The text of this appendix was adapted from a manuscript coauthored with J. Kyle Lassila, Heidi K. Privett, and Stephen L. Mayo.

Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L., Combinatorial methods for small-molecule placement in computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **2006,** *103* (45), 16710–16715.

Abstract

The incorporation of small molecule transition state structures into protein design calculations poses special challenges because of the need to represent the added translational, rotational, and conformational freedoms within an already difficult optimization problem. Successful approaches to computational enzyme design have focused on catalytic side-chain contacts to guide placement of small molecules in active sites. We describe a process for modeling small molecules in enzyme design calculations that extends previously described methods, allowing favorable small molecule positions and conformations to be explored simultaneously with sequence optimization. Because all current computational enzyme design methods rely heavily on sampling of possible active site geometries from discrete conformational states, we tested the effects of discretization parameters on calculation results. Rotational and translational step sizes as well as side-chain library types were varied in a series of computational tests designed to identify native-like binding contacts in three natural systems. We find that conformational parameters, especially the type of rotamer library used, significantly affect the ability of design calculations to recover native binding site geometries. We describe the construction and use of a crystallographic conformer library, and find that it more reliably captures active-site geometries than traditional rotamer libraries in the systems tested.

Introduction

As catalysts, enzymes offer advantageous properties including dramatic rate enhancements, complete control over absolute stereochemistry, and nontoxic biodegradation. Yet a fundamental limiting factor in the use of enzymes for chemical synthesis, bioremediation, therapeutics, and other applications is the availability of enzymes with the required activities, specificities, and tolerances to reaction conditions. It is therefore a major goal of computational protein design to be able to reliably create completely new protein catalysts with specific properties on demand.

A catalyst by definition must reduce the energy barrier for formation of the transition state. To design transition-state-stabilizing interactions, computational protein design groups have incorporated transition-state or high-energy intermediate state structures into design calculations. These efforts have yielded experimentally verified new catalytic proteins.[1–3] However, substantial challenges still prevent routine or reliable design of enzymes. One major challenge is in finding energy functions that are fast enough for large calculations but that still provide informative approximations of electrostatic and desolvation effects in the protein environment.[4, 5] This paper focuses on another fundamental challenge, the need to represent the large translational, rotational, and conformational freedoms of a small molecule within already astronomically large sequence design calculations.

Here we define protein design as the selection of amino acid sequences such that the resulting protein occupies a given three-dimensional fold and has desired functional properties. Earlier experiments sought to redesign full protein sequences or confer

increased thermostability,[6, 7] but newer work has successfully introduced other properties including catalytic activity, conformational specificity, ligand affinity, and even novel protein folds.[1–3, 8–10] In these examples, side-chain placement algorithms were used to select from a set of discrete, probable side-chain rotamers using energy functions tuned to produce thermostable proteins. These calculations represent difficult optimization problems[11] and they can also be large—a sample calculation performed on a typical enzyme active site yields more than 10^{65} possible sequence combinations, even when excluding movements of the small molecule.

The computational demands of sequence selection prevent ligand positioning using standard docking procedures, which often approximate or neglect side-chain flexibility.[12] Approaches developed specifically for the purpose of enzyme and binding site design have introduced other schemes to limit the calculation size. Looger et al. used stationary, inflexible ligand poses in a large number of individual protein design calculations and demonstrated experimentally that several of the resulting proteins had high ligand affinity.[9] Lilien *e*t al. reported and experimentally validated an ensemble-based method that allows ligand translation and rotation simultaneously with side-chain optimization but only permits mutation of two or three amino acid positions at a time.[13] Chakrabarti et al. described a method for sequence design that neglects conformational and positional ligand flexibility and has not been experimentally tested.[14, 15]

To design new enzyme active sites, a ligand placement method must be able to select side chains in many positions and must consider rotational, translational, and conformational freedom of the small molecule. Previously, methods for the design of catalytic proteins treated high-energy-state structures of the reacting molecules as

extensions of contacting amino acid side-chain rotamers. A two-step procedure was utilized, where ligands, anchoring side chains, and other catalytic side chains were placed through a geometric screening procedure and surrounding side chains were designed in a second step.[1, 16–18] We have developed a process for ligand placement in computational protein design calculations that expands upon previous work and that allows ligand rotation, translation, and conformational freedom to be explored combinatorially within the sequence design calculation itself. The implementation of ligand placement procedures within the context of the pairwise-decomposable protein design framework makes it possible to use a single energy function that can be parameterized as needed to reproduce experimental data.

We tested both a simple rotational and translational process for ligand placement as well as the previously used targeted ligand placement approach. A contact-based screening method is described that allows selection of ligand positions and confomations compatible with catalytic contacts. Test calculations in three systems, *E. coli* chorismate mutase, *S. cerevisiae* triosephosphate isomerase, and *S. avidinii* streptavidin, suggest that the success of ligand placement procedures can be quite sensitive to conformational sampling parameters, including rotational and translational step sizes and the types of rotamer libraries used. We evaluated the efficacy of two standard rotamer libraries and two crystallographic conformer libraries. Traditional rotamers are constructed from canonical χ angles determined by statistical analysis of the PDB,[19–21] whereas conformers have Cartesian coordinates taken directly from high-resolution structures.[22, 23] Conformer libraries may allow more accurate modeling because they are not limited to ideal geometries and their sizes can be tuned more easily and naturally.[22, 23] In our tests, a

backbone-independent conformer library recovered wild-type-like active site geometries more successfully than the other libraries, despite smaller size.

Results and discussion

We have implemented and tested a process for incorporation of small molecules into computational protein design calculations. The procedure is general and may be used to place ground-state ligands or transition-state structures. It is also amenable to multi-state design methods that seek to explicitly reflect the energy difference between reactant and transition states or between alternative ligands.

General calculation procedure

Each ligand placement calculation comprised five steps. In the first step, a large number of discrete variations of ligand coordinates was created. Initial sets of orientations were created by one of two methods, either simple rotation and translation or a targeted placement approach, both of which are discussed in more detail in subsequent sections. In the tests described here, each set of ligand variations contained 10^6–10^9 members, reflecting rotational and translational movement as well as internal conformational flexibility.

Next, the large number of substrate orientations was reduced to a manageable number (< ~ 20,000) using both a simple hard-sphere steric potential to check for backbone clashes and a set of user-defined geometric criteria for side-chain/ligand contacts. In this work, geometric criteria were defined to reflect the distances, angles, and torsions characteristic of important catalytic contacts observed in the crystal

structures (Figure 1). In designing an enzyme with no naturally existing precedent, ideal contact geometries would be based on chemical intuition and/or quantum mechanical calculations. The geometric criteria were applied as follows. For every ligand variation, each of the geometric criteria was tested for satisfaction by contacts from any possible amino acid side-chain conformation in all designed protein positions. If a ligand variation was not able to make at least one of each type of user-specified contact, that ligand variation was discarded from the set. After geometric and steric pruning, the ligand variations remaining were only those theoretically capable of making each of the user-specified contacts.

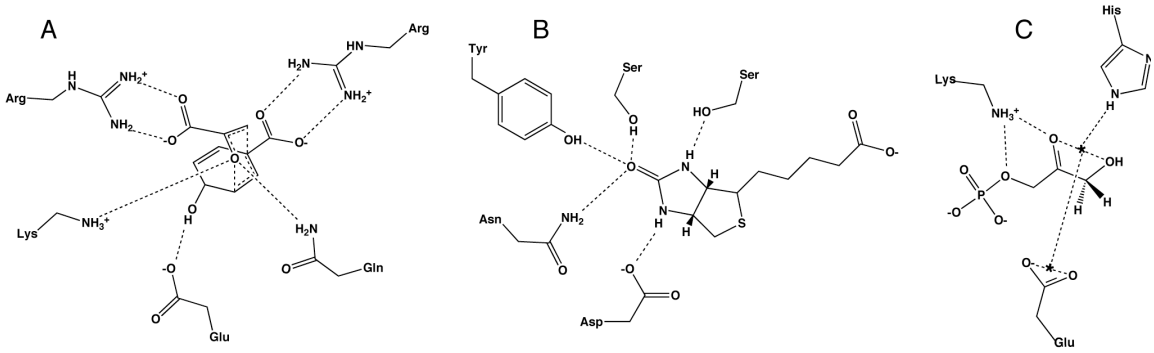

Figure 1: Contact geometries specified in small-molecule pruning step. Ranges of distances, angles, and torsions were allowed that included the crystallographic geometries. (*A*) Chorismate mutase. (*B*) Biotin in streptavidin. (*C*) Triosephosphate isomerase Michaelis complex. Asterisks indicate pseudoatoms used in geometry definitions.

In the third step, pairwise energies for all side-chain/side-chain, side-chain/backbone, backbone/ligand, and side-chain/ligand interactions were calculated using the full force field. In our work, this normally includes a scaled van der Waals[1] term,[24] hydrogen-bonding and electrostatic terms,[25] and a solvation potential.[26, 27]

The fourth step is an optional energy biasing that favors side-chain/ligand contacts deemed necessary for catalysis or binding. This energy biasing step helps to overcome the shortcomings of molecular mechanics energy functions as well as the inherent limitation of treating a multi-state design problem—*differential* stabilization of transition state relative to substrate in protein versus solution—using single-state design algorithms. As methods for modeling electrostatics and solvation and for designing over multiple states improve, the need for this biasing step should be reduced. Previous work utilized selective application of solvation energy or an additional search algorithm step[9] for the same purpose. We favor the use of adjustable bias energies that can be tailored for specific purposes and investigated as a design variable.

To implement the bias, user-specified energies were added or subtracted from pairwise side-chain/ligand interaction energies. We use the energy bias under two regimes, one for normal design calculations and another for rapid assessment of catalytic residue arrangements within a protein scaffold. In normal design calculations, a small energy benefit is simply applied to favor specified types of side-chain/ligand contacts. Alternatively, to quickly identify potential catalytic residues, exaggerated energetic benefits and penalties are applied together. A very large energy benefit is given for desired types of pairwise interactions (100 kcal/mol was used in the test cases reported here). An even larger energy penalty (10,000 kcal/mol here) is applied to all other pairwise side-chain/ligand interactions, except when the side chain is alanine or glycine. In other words, the energy penalty forces all designed side chains to alanine or glycine unless they participate in user-specified catalytic contacts with the ligand. Although this process clearly does not yield physically relevant energetics, it offers a useful tool to

investigate the catalytic conformational space within a binding pocket. The tests performed here to study the effect of sampling parameters on calculation results took advantage of this second approach. Calculations performed to demonstrate sequence selection utilized the normal design approach of applying a simple energy benefit to catalytic contacts.

Finally, in the fifth step, optimal sequences were identified using the FASTER[28, 29] or HERO[30] search methods. In the test cases described here, the result reported is the lowest-energy sequence with the maximal number of specified contacts.

Rotation-translation search

Simple rotation and translation can be used to fill the active site with an initial set of ligand variations in the first step of the process described. Because discrete steps must be used to rotate and translate the ligand, we evaluated the sensitivity of the calculation results to rotational and translational step sizes. A series of calculations was performed using an alanine-containing active-site background, as discussed in step 4 above. We first tested different rotational step sizes using the crystallographic translational starting position with three initial random rotations. Backbone-dependent and backbone-independent rotamer and conformer libraries were tested. Each side-chain library was tested with and without inclusion of the specific crystallographic side-chain rotamers from the structure under examination.

As seen in Table 1, the results of these calculations (in terms of both RMSD relative to crystallographic position and number of wild-type contacts) were strongly dependent on the both the rotational step size and the rotamer library used. In the case of

chorismate mutase, only the backbone-independent conformer library was able to find native-like geometry and contacts. Figure 2 shows results from this library with the 5° step size. When the crystallographic rotamers were included in the calculation, however, all four libraries returned native-like results. It should be noted that none of the three test case structures were included in the set of structures used to create the conformer libraries. The backbone-independent conformer library appeared the most consistently successful with the other two test cases as well, although it showed strong dependence on rotational step size in streptavidin.

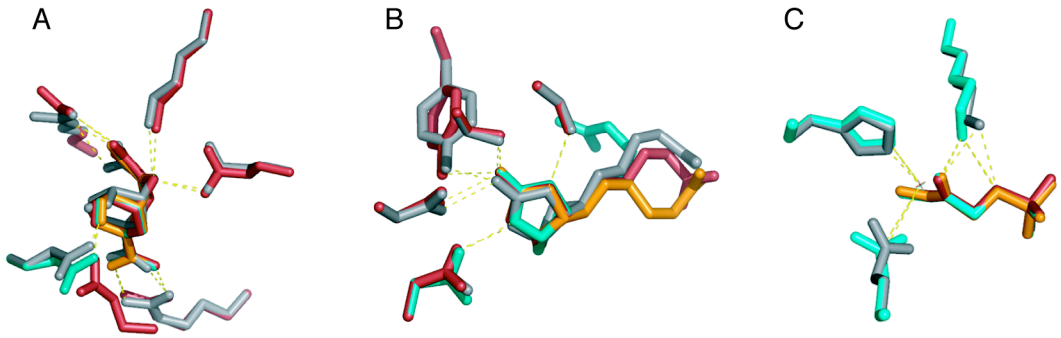

Figure 2: Sample results from test calculations presented in Table 1. Crystallographic side chains and ligands are shown in gray. Results from three trials using different initial random rotational positions are shown in red, teal, and orange. In cases where three colors are not visible, the selected rotamers from two or more calculations were identical. Results are shown from calculations with 5° rotation and the backbone-independent conformer library. (*A*) Chorismate mutase. An alternate backbone position was chosen for a glutamate-hydroxyl contact in one trial (red side chain, lower left). (*B*) Biotin in streptavidin. Note that the biotin pentanoic acid moiety samples different conformations in the calculation and the surrounding side chains were not designed. (*C*) Triosephosphate isomerase.

Table 1: RMSD and number of wild-type contacts as a function of rotational step size and rotamer library[a,b]

Chorismate mutase

	Rotational step size				
Rotamer Library[c]	30°	20°	15°	10°	5°
Conformer: bb-ind	-	-	0.61 ± 0.03 (4.0)	0.55 ± 0.05 (4.0)	0.47 ± 0.04 (4.7)
with xtal rotamers	-	-	0.61 ± 0.03 (4.0)	0.55 ± 0.05 (4.0)	0.47 ± 0.04 (4.7)
Rotamer: bb-ind	-	-	3.88 ± 0.37 (0.0)	2.88 ± 1.44 (0.0)	3.01 ± 1.61 (0.0)
with xtal rotamers	-	-	1.57 ± 1.70 (2.7)	0.51 ± 0.00 (4.0)	0.52 ± 0.01 (4.0)
Conformer: bb-dep	-	-	3.66 ± 0.11 (1.0)	3.59 ± 0.08 (1.0)	3.60 ± 0.09 (1.0)
with xtal rotamers	-	1.67 ± 1.78 (3.3)	1.57 ± 1.83 (3.7)	0.60 ± 0.08 (4.3)	0.54 ± 0.06 (5.0)
Rotamer: bb-dep	-	-	-	-	-
with xtal rotamers	-	-	-	0.49 ± 0.04 (4.3)	0.52 ± 0.01 (4.0)

Streptavidin-Biotin

	Rotational step size				
Rotamer Library[c]	30°	20°	15°	10°	5°
Conformer: bb-ind	-	-	-	-	0.27 ± 0.09 (5.0)
with xtal rotamers	-	0.24 ± 0.09 (5.0)	0.24 ± 0.07 (5.0)	0.26 ± 0.06 (5.0)	0.20 ± 0.13 (5.0)
Rotamer: bb-ind	-	-	0.77 ± 0.42 (2.3)	0.60 ± 0.14 (3.0)	0.60 ± 0.05 (2.7)
with xtal rotamers	0.37 ± 0.17 (5.0)	0.24 ± 0.09 (5.0)	0.24 ± 0.07 (5.0)	0.26 ± 0.06 (5.0)	0.30 ± 0.17 (5.0)
Conformer: bb-dep	-	-	-	0.25 ± 0.12 (5.0)	0.20 ± 0.07 (5.0)
with xtal rotamers	-	0.24 ± 0.09 (5.0)	0.24 ± 0.07 (5.0)	0.22 ± 0.03 (5.0)	0.29 ± 0.09 (4.0)
Rotamer: bb-dep	-	-	-	0.82 ± 0.28 (2.3)	0.66 ± 0.02 (3.0)
with xtal rotamers	-	0.24 ± 0.09 (5.0)	0.24 ± 0.07 (5.0)	0.26 ± 0.06 (5.0)	0.16 ± 0.06 (5.0)

Triosephosphate isomerase

	Rotational step size				
Rotamer Library[c]	30°	20°	15°	10°	5°
Conformer: bb-ind	-	1.87 ± 1.07 (0.7)	3.59 ± 2.28 (1.0)	0.28 ± 0.07 (3.0)	0.24 ± 0.05 (3.0)
with xtal rotamers	-	1.31 ± 0.29 (1.0)	1.95 ± 2.28 (1.3)	0.27 ± 0.06 (3.0)	0.15 ± 0.02 (3.0)
Rotamer: bb-ind	5.09 ± 0.05 (0.3)	0.60 ± 0.12 (1.7)	0.55 ± 0.25 (2.3)	0.34 ± 0.04 (2.3)	0.25 ± 0.08 (3.0)
with xtal rotamers	5.06 ± 0.05 (0.3)	0.60 ± 0.12 (2.0)	0.37 ± 0.04 (3.0)	0.25 ± 0.04 (3.0)	0.15 ± 0.02 (3.0)
Conformer: bb-dep	-	-	-	-	-
with xtal rotamers	-	-	-	-	0.15 ± 0.02 (3.0)
Rotamer: bb-dep	3.28 ± 0.73 (1.7)	0.60 ± 0.12 (1.7)	0.37 ± 0.05 (2.3)	0.31 ± 0.04 (2.3)	0.25 ± 0.08 (3.0)
with xtal rotamers	3.28 ± 0.73 (2.3)	0.60 ± 0.12 (2.3)	0.37 ± 0.05 (3.0)	0.29 ± 0.03 (3.0)	0.15 ± 0.02 (3.0)

[a] Dashes indicate that required contacts were not satisfied in at least one of three trials.
[b] Values are non-hydrogen-atom RMSD in Ångstroms relative to crystallographic ligands or bicyclic ring atom RMSD relative to crystallographic ligand for biotin (i.e., the pentanoic acid moiety was not considered in biotin RMSDs). Averages and standard deviations from three random initial positions are reported. Numbers in parentheses are the number of contacts where the amino acid position was the same as in the wild-type structure, averaged over the three trials. Maximum possible number of wild-type contacts: chorismate mutase, 5; streptavidin, 5; triosephosphate isomerase, 3
[c] bb-ind: backbone-independent, bb-dep: backbone-dependent

Next, we tested various combinations of rotational and translational step sizes starting from random initial ligand positions and using only the backbone-independent conformer library (Figure 3, Table 2). The crystallographic rotamers from the structures under investigation were not included in these calculations. The results show that, subject to the constraints imposed by the geometries defined in the pruning step and the biasing step, more than one combination of rotational and translational step size is viable for each test case and the sensitivity of the result to step size varies among the test cases.

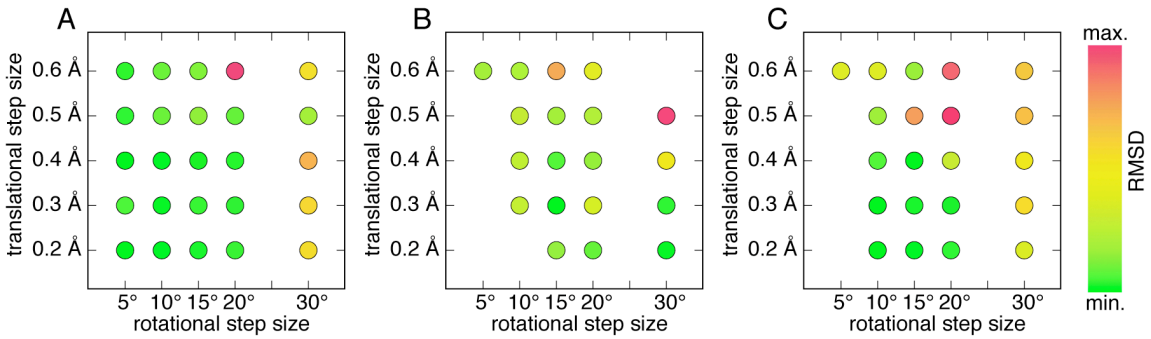

Figure 3: Effect of rotational and translational step sizes. Each spot represents the average of three trials with initial random starting positions. Missing points indicate that one or more trials could not identify wild-type-like contacts or else that the calculation was prohibitively large; no calculations were performed using a 25° rotational step size. Colors indicate non-hydrogen atom RMSD as described in the tables. (*A*) Chorismate mutase (min., 0.53 Å; max., 2.61 Å) (*B*) Streptavidin-biotin (min., 0.57 Å; max., 2.05 Å) (*C*) triosephosphate isomerase (min., 0.44 Å; max., 5.64 Å)

The rotation/translation tests were performed using three initial random starting positions for each system. The starting positions were created by randomly rotating and translating the ligand within a 1 $Å^3$ box around the ligand centroid (or the centroid of the bicyclic ring system in biotin). Using the same atom comparisons as described in the tables, the nine initial positions had RMSDs relative to crystallographic positions of

between 2.1 Å and 4.5 Å, with an average of 3.2 Å. These tests do not provide full, unbiased searches of the active sites. Full active site searches could be conducted using this method by performing separate calculations for grid points distributed evenly through the active site. Given the time required to perform these smaller calculations (Table 2), searching an entire active site using rotational and translational perturbations would be computationally expensive. For example, examining a 3.6 x 3.6 x 3.6 Å grid using the 10° and 0.3 Å step sizes would require an estimated 324 hours on a 16-processor cluster for placement of ligands and catalytic side chains in the chorismate mutase active site. Thus, for initial positioning of a ligand within an active site, rotational and translational placement is inefficient. However, the ability to adjust small molecule position and conformation simultaneously with side-chain optimization should be extremely valuable for refining an initial position identified from a coarser search method.

Table 2: RMSD and number of wild-type contacts as a function of rotational and translational step sizes[a,b]

Chorismate mutase

Translational step size (Å)	Rotational step size 30°	20°	15°	10°	5°	Time (10°, hours)[c]
0.6	1.69 ± 1.54 (2.3)	2.61 ± 1.67 (1.3)	0.77 ± 0.10 (4.3)	0.73 ± 0.02 (4.0)	0.61 ± 0.06 (4.7)	3
0.5	0.91 ± 0.20 (3.7)	0.72 ± 0.07 (4.0)	0.83 ± 0.06 (3.3)	0.74 ± 0.05 (4.0)	0.60 ± 0.13 (4.3)	10
0.4	2.02 ± 1.99 (2.3)	0.60 ± 0.04 (4.7)	0.59 ± 0.13 (4.0)	0.57 ± 0.12 (4.3)	0.53 ± 0.13 (4.3)	11
0.3	1.73 ± 1.51 (2.3)	0.61 ± 0.07 (4.3)	0.62 ± 0.15 (4.3)	0.58 ± 0.07 (4.0)	0.65 ± 0.04 (4.0)	12
0.2	1.71 ± 1.53 (2.3)	0.62 ± 0.10 (4.0)	0.60 ± 0.09 (4.0)	0.54 ± 0.07 (4.0)	0.56 ± 0.05 (4.0)	33

Streptavidin-biotin

Translational step size (Å)	Rotational step size 30°	20°	15°	10°	5°	Time (10°, hours)[c]
0.6	-	1.16 ± 0.60 (3.7)	1.67 ± 1.02 (3.7)	0.88 ± 0.44 (4.3)	0.84 ± 0.48 (4.3)	5
0.5	2.05 ± 0.59 (1.7)	0.91 ± 0.44 (5.0)	0.84 ± 0.61 (5.0)	0.99 ± 0.91 (3.7)	-	18
0.4	1.32 ± 1.39 (3.7)	0.80 ± 0.09 (5.0)	0.67 ± 0.28 (5.0)	0.96 ± 0.72 (3.7)	-	19
0.3	0.63 ± 0.16 (5.0)	1.08 ± 0.49 (5.0)	0.57 ± 0.21 (5.0)	1.03 ± 0.48 (4.3)	-	18
0.2	0.60 ± 0.32 (5.0)	0.70 ± 0.34 (5.0)	0.80 ± 0.24 (5.0)	-	-	-

Triosephosphate isomerase

Translational step size (Å)	Rotational step size 30°	20°	15°	10°	5°	Time (10°, hours)[c]
0.6	3.80 ± 2.14 (0.3)	5.22 ± 0.32 (0.0)	1.29 ± 0.91 (1.3)	2.39 ± 2.54 (1.7)	2.40 ± 2.58 (2.0)	0.4
0.5	3.92 ± 1.94 (0.0)	5.64 ± 0.45 (0.3)	4.47 ± 1.45 (0.0)	1.33 ± 1.01 (1.7)	-	2
0.4	3.13 ± 1.77 (0.3)	1.96 ± 1.05 (2.0)	0.47 ± 0.24 (1.7)	0.78 ± 0.66 (3.0)	-	2
0.3	3.44 ± 1.96 (0.3)	0.59 ± 0.18 (2.0)	0.60 ± 0.29 (2.3)	0.46 ± 0.11 (3.0)	-	2
0.2	2.33 ± 1.80 (0.7)	0.68 ± 0.10 (2.3)	0.49 ± 0.12 (3.0)	0.44 ± 0.11 (3.0)	-	5

[a] Dashes indicate that required contacts were not satisfied in at least one of three trials or that the calculation was too large to complete.

[b] Values are non-hydrogen atom RMSD in Ångstroms relative to crystallographic ligands or bicyclic atom RMSD relative to crystallographic ligand for biotin (i.e., the pentanoic acid moiety was not considered in biotin RMSDs). Averages and standard deviations from three random initial positions are reported. Numbers in parentheses are the number of contacts where the amino acid position was the same as in the wild-type structure, averaged over the three trials. Maximum possible number of wild-type contacts: chorismate mutase, 5; streptavidin, 5; triosephosphate isomerase, 3

[c] Wall clock time; calculations performed on a 16-processor cluster

Targeted ligand placement

A second approach places the small molecule with reference to a contacting side chain (Figure 4). In this approach, one or more small molecule variations are placed for every rotamer of the selected contacting side chain in every putative active-site position. This process has the advantage that ligand poses are targeted more efficiently to orientations that are able to make productive side-chain contacts. Previous computational enzyme design work utilized similar approaches.[1, 16, 17] In contrast to previous methods, however, our procedure does not maintain any association between the targeting rotamer and the small molecule—once the set of ligand conformations and orientations is constructed in step one, the ligand variations are all subjected to pruning, pairwise energy calculations, and optimization as independent entities in the calculation. An implication of this procedure is that a ligand may engage in a catalytic contact with a rotamer, amino acid, or protein position that differs from those of the side-chain rotamer that was originally used to place that ligand.

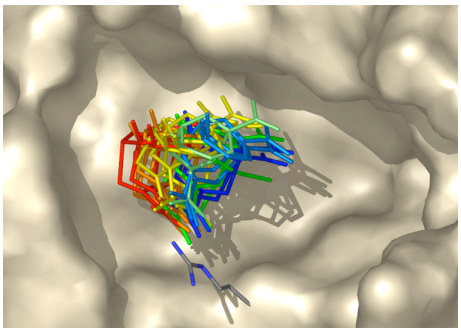

Figure 4: Targeted placement procedure. For a given side-chain rotamer, small molecule ligands are placed such that they are able to meet specified geometric criteria. This is repeated for every possible conformation of the amino acid at every designed position. Shown is a subset of orientations of a chorismate mutase transition-state structure in contact with one conformation of arginine.

We tested the effect of four types of side-chain libraries on the ability of a targeted placement process to find wild-type-like ligand positions and contacts. For the three test cases, the following side-chain contacts were used to anchor the ligand: chorismate mutase, C11 carboxylate to arginine; streptavidin, N1 to aspartate; triosephosphate isomerase, O2 and O3 to histidine. For each contact type, variations were allowed in the geometry of the contact, including the contacting atoms (NH1-NH2 vs. NE-NH1 for arginine) and variations in defined distances, angles, and dihedrals of the contact.

As with the rotational and translational search, success in achieving native active-site conformations was highly dependent on the side-chain library used (Table 3). Only the backbone-independent conformer library yielded results for all three test cases that were comparable to those with crystallographic rotamers included. Using that library, all three systems returned all wild-type contacts with low ligand RMSD relative to the crystallographic position. As with the rotation/translation search, the chorismate mutase case showed the strongest sensitivity to rotamer library. Inspection of the structures revealed that an arginine side chain (Arg 28) occupies a conformation in the inhibitor-bound, active enzyme structure that was not well approximated in the other rotamer libraries.

The targeted placement approach allowed a thorough and directed search of active-site conformational space, including between 10^6 and 10^9 small-molecule orientations and conformations spread throughout the active site. In contrast to the rotation/translation method, a full active-site search took between one and eighteen hours to complete using the backbone-independent conformer library and no initial starting

position was required. This method offers an efficient first step for defining active-site geometry in a new protein scaffold. One shortcoming is that it may be difficult to sample the many geometrical variations of a flexible hydrogen-bonding interaction. For example, the 972 variations in guanidino-carboxylate contact geometry sampled in the chorismate mutase case are probably adequate to reflect flexibility in this relatively rigid dual hydrogen-bonding interaction. A less restrained interaction, however, such as a serine hydrogen bonding with a sterically unrestricted ligand carbonyl oxygen, results in a compromise between maintaining a manageable calculation size and modeling contact flexibility. One solution is to use a targeted method to find an initial ligand position within the binding site and then, in a second calculation, optimize both active-site packing and fine rotational and translational placement of the ligand.

Table 3: Results from targeted placement procedure as a function of rotamer library

Chorismate mutase

Rotamer library[a]	log(initial ligand variations)	RMSD (Å)[b] (WT contacts)	Time (hours)[c]
Conformer: bb-ind	7.88	0.60 (5)	16
with xtal rotamers	7.88	0.68 (3)	18
Rotamer: bb-ind	8.18	3.61 (0)	51
with xtal rotamers	8.18	0.66 (4)	62
Conformer: bb-dep	7.64	3.62 (1)	8
with xtal rotamers	7.64	0.68 (4)	9
Rotamer: bb-dep	7.76	2.31 (1)	14
with xtal rotamers	7.76	0.66 (4)	16

Streptavidin-biotin

Rotamer library[a]	log(initial ligand variations)	RMSD (Å)[b] (WT contacts)	Time (hours)[c]
Conformer: bb-ind	7.07	0.64 (5)	1.4
with xtal rotamers	7.07	0.64 (5)	1.4
Rotamer: bb-ind	7.20	0.54 (4)	3.5
with xtal rotamers	7.20	0.34 (4)	3.4
Conformer: bb-dep	6.35	0.37 (5)	0.2
with xtal rotamers	6.35	0.54 (4)	0.2
Rotamer: bb-dep	7.17	3.50 (0)	2.6
with xtal rotamers	7.17	0.19 (5)	2.8

Triosephosphate isomerase

Rotamer library[a]	log(initial ligand variations)	RMSD (Å)[b] (WT contacts)	Time (hours)[c]
Conformer: bb-ind	7.31	0.49 (3)	1.3
with xtal rotamers	7.31	0.49 (3)	1.3
Rotamer: bb-ind	7.78	0.46 (3)	8.7[d]
with xtal rotamers	7.78	0.46 (3)	87[d]
Conformer: bb-dep	6.82	7.51 (0)	0.3
with xtal rotamers	6.82	0.78 (3)	0.3
Rotamer: bb-dep	7.58	0.51 (3)	4.3[d]
with xtal rotamers	7.58	0.51 (3)	4.9[d]

[a] bb-ind, backbone-independent; bb-dep, backbone-dependent

[b] RMSDs calculated as described in Table 1. Maximum possible number of wild-type contacts: chorismate mutase, 5; streptavidin, 5; triosephosphate isomerase, 3

[c] Wall clock time; calculations performed on a 16-processor cluster

[d] Calculation was performed as a series of smaller calculations.

Sequence design

The computational tests described in the previous sections were designed to evaluate the effects of calculation parameters on recovery of native enzyme geometries, and the design of active-site residues was limited to catalytic side chains. However, the general procedure described here is equally amenable to full active-site design calculations.

In previously published work, 18 active site residues of *E. coli* chorismate mutase were redesigned simultaneously with rotational and translational relaxation of the transition-state structure from the starting crystallographic position.[31] The six predicted mutations were experimentally investigated and some were found to confer increased catalytic efficiency or thermostability.[31] A detrimental mutation predicted in the study underscored the importance of continued work on energy functions. In the calculation that motivated this experimental work, the initial starting position of the small molecule was taken from the crystal structure and a limited degree of rotational and translational optimization was employed.

We performed a test calculation to demonstrate that small molecules can be placed simultaneously with full active-site side-chain optimization, without reference to any known starting position. In a sample calculation using *E. coli* chorismate mutase, the targeted placement method was used to identify 10^7 small-molecule variations. In this example, after the geometric pruning step and elimination of variants with backbone steric clashes, 155 small-molecule variations remained. These variants were evaluated combinatorially with ten different side-chain identities in twelve active-site positions.

Using FASTER for optimization, the calculation took approximately 9 hours to complete on a 16-processor cluster with about 70% of the total calculation time consumed in calculating a surface-area-based solvation term.

Conclusions

The described procedures allow the incorporation of small-molecule placement directly into sequence design calculations. The test calculations performed suggest that the results of computational enzyme design processes can be quite sensitive to calculation parameters, including the rotamer library used and the coarseness of ligand positioning. These results emphasize that the conformational space of a calculation must be explored before meaningful conclusions can be reached about energy functions.

Given that we still have much to learn about the complex relationship between protein structure and catalytic activity,[32, 33] luck and choice of system may continue to play a role in the success of *de novo* computational enzyme design efforts for some time. However, the power of computational enzyme design to stringently evaluate our understanding of the energetics of catalysis should not be overlooked. Experimental feedback gained from both successful and unsuccessful designs will make it possible to critically examine energy functions for modeling active sites. Employing quality transition-state structures derived from *ab initio* calculations and experimental evidence will help computational design experiments to provide more meaningful information about the effectiveness of energy functions. The use of large side-chain structural libraries and fine movements of transition-state structures will help to reduce errors from conformational sampling. Backbone relaxation and multi-state design will offer other

important tools to improve the value of design calculations. Finally, the construction of gene libraries or large numbers of computationally designed variants has great potential for overcoming the shortcomings of enzyme design models,[34] but results from these experiments will be most useful for furthering our understanding of catalysis and design if both active and inactive variants are reported. By critically evaluating current methods for computational enzyme design, we will move closer to a deeper and more practically useful understanding of the sequence determinants of enzyme activity in the future.

Methods

Structures and charges

PDB files were used without minimization (*E. coli* chorismate mutase, 1ecm;[35] *S. avidinii* streptavidin, 1mk5;[36] *S. cerevisiae* triosephosphate isomerase, 1ney).[37] Hydrogens were added with Reduce.[38]

A library of ligand internal conformations was created for each system as follows. Chorismate mutase: An HF/6-31G* *ab initio* transition-state structure[39] was used with only one variation—the O4 hydroxyl proton was allowed to occupy three positions, 60°, 180°, and -35°, defined by the H-C-O-H dihedral angle. The minima in a torsional profile at the HF/6-31G* level were at approximately 180° and -35°, and 60° was included as an option because hydrogen-bonding patterns in chorismate mutases from other species suggested population of that region of torsional space. Streptavidin: Four rotatable bonds in biotin were allowed to occupy three positions each (60°, -60°, 180° for sp3-sp3 bonds and 30°, 90°, 150° for the symmetric carboxylate group). Thirty-four conformations were excluded because of high internal energy calculated using the van

der Waals component of the DREIDING force field.[40] Triosephosphate isomerase: The pdb structure used was the Michaelis complex with the substrate dihydroxyacetone phosphate. In ground-state dihydroxyacetone phosphate, two rotatable bonds (defined by the P-O-C-C and C-C-O-H dihedral angle) were allowed to occupy three positions each (60°, -60°, 180°). Three conformations were excluded because of high internal DREIDING van der Waals energy.

Ligand atomic charges were obtained by fitting charges to electrostatic potential from HF/6-31G* single-point energy calculations using[19] the transition-state structure (chorismate mutase) or crystallographic ground-state structure (biotin, dihydroxyacetone phosphate). *Ab initio* calculations and charge determinations were performed using Spartan (Wavefunction, Inc.) or Jaguar (Schrödinger, Inc.).

Side-chain rotamer libraries

Standard backbone-dependent and backbone-independent rotamer libraries were used with expansion by one standard deviation about $\chi 1$ and $\chi 2$.

Crystallographic conformer libraries were prepared using coordinates from 149,813 side chains selected from 1011 unique structures. A clustering algorithm was developed based on ideas described by Shetty et al.[22] and is described briefly here. Every side-chain conformation from the raw data set is assigned to exactly one cluster. Each cluster is represented by the centroid, which is the member with coordinates closest to the average coordinates of all cluster members. A conformer library consists of a list of all of the cluster representatives and their coordinates. In our clustering algorithm, clusters are assigned through discrete clustering moves: *Switch* allows a single raw conformer to

leave one cluster and join another; *Merge* combines two clusters into one; *Split* allows a raw conformer to start a new cluster on its own. These moves are depicted in Figure 5.

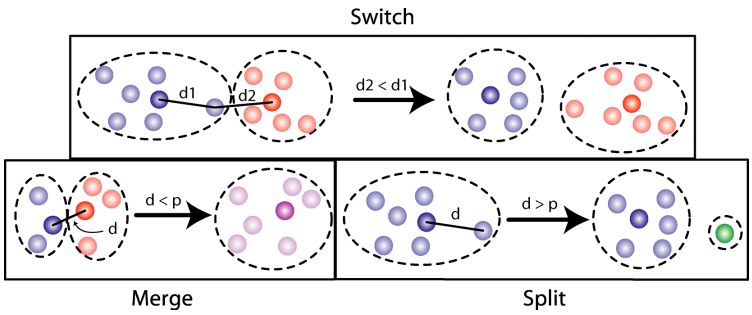

Figure 5: **The three clustering moves are illustrated by showing the state of a sample system before and after the move is performed.** Each dot represents a single side-chain conformation taken from the PDB. Distances represent side-chain RMSDs between pairs of conformers. Dots sequestered together by a dashed line and colored the same are members of the same cluster. Darker-colored dots denote cluster representatives.

RMSDs between pairs of conformers are compared to determine whether or not to apply a particular move. *Switch* is applied so that each raw conformer is a member of the cluster whose centroid is closest to it. *Merge* and *Split* are applied based on the value of the clustering parameter *p*: two clusters are merged if their centroids are within *p* of each other, whereas a conformer splits off and starts a new cluster if the closest centroid of any existing cluster is farther than *p* from it. The clustering moves are applied as follows until the number of clusters converges:

1. Start with a small number of clusters (1 was used in this work), and randomly assign a single raw conformer to each as the sole member and cluster representative.
2. Assign each raw conformer in the data set to the cluster whose centroid is closest.
3. While the number of clusters is not converged:
    a. Iteratively attempt to *Merge* pairs of clusters until no cluster can be further merged.
    b. For each conformer C:
        i. Measure the distance *d* between C and the centroid of every existing cluster.
        ii. If the distance *d* to the closest cluster centroid is greater than *p*, *Split* C off as its own cluster.
        iii. Else, *Switch* C to the closest cluster.
        iv. Recompute the centroid for every cluster that has changed membership.

The algorithm allows the construction of both backbone-dependent and backbone-independent libraries to custom sizes by using clustering factor *p* to define the desired degree of similarity between independent conformers. In this work, clustering factors of 0.3 Å and 1.0 Å were used for backbone-dependent and backbone-independent rotamer libraries, respectively.

For all calculation types, conformer libraries were smaller than the standard rotamer libraries. As an example, the number of side-chain conformations for the chorismate mutase calculations described in Table 3 were as follows: backbone-independent rotamer, 14229; backbone-independent conformer, 5955; backbone-dependent rotamer, 7945; and backbone-dependent conformer, 5539.

Calculation parameters

All non-Gly, non-Pro residues reasonably within the natural active sites were included in calculations. Residues with any atom within a 5 Å radius from any atom in the crystallographic ligands were included, less those residues separated from the natural ligand by backbone elements and plus a few adjacent residues not within the 5 Å cutoff. The positions designed were (all in chain A unless otherwise designated): chorismate mutase, 28, 32, 35, 39, 46, 47, 48, 51, 52, 55, 81, 84, 85, 88, 7B, 11B, 14B, 18B; streptavidin, 23, 24, 25, 27, 43, 45, 46, 47, 49, 50, 79, 86, 88, 90, 92, 108, 110, 112, 128, 130; and triosephosphate isomerase, 10, 12, 95, 97, 165, 170, 211, 230.

In ligand placement test cases, designed residues were restricted to ligand-contacting residues or alanine as follows: Arg, Lys, Gln, Glu, or Ala in chorismate mutase; Ser, Asn, Tyr, Asp, or Ala in streptavidin; and Glu, His, Lys, or Ala in triosephosphate isomerase. Four calculations on triosephosphate isomerase were run as smaller component calculations, as indicated in Table 2, because of prohibitive size as a single calculation.

Energy functions and optimization

Energy functions included scaled van der Waals,[24] hydrogen-bonding, and electrostatic terms.[25] A surface-area-based solvation potential[27] was used in sequence design calculations but not for ligand placement, where solvation energy would have been heavily outweighed by geometric considerations. Sequences were optimized with respect to the energy function using FASTER[28, 29] or HERO.[30] On occasion, a top-ranked sequence contained more than one instance of a given specified geometric contact, owing to the energy benefit applied for these contacts. In these cases, Monte Carlo[41, 42] was used to sample around the global minimum energy sequence, and the top-ranked sequence with a single instance of each geometric contact was reported.

References

1. Bolon, D. N.; Mayo, S. L., Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** *98* (25), 14274–14279.

2. Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Rothlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D., De novo computational design of retro-aldol enzymes. *Science* **2008,** *319* (5868), 1387–1391.

3. Rothlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D., Kemp elimination catalysts by computational enzyme design. *Nature* **2008,** *453* (7192), 190–U4.

4. Mendes, J.; Guerois, R.; Serrano, L., Energy estimation in protein design. *Current Opinion in Structural Biology* **2002,** *12* (4), 441–446.

5. Vizcarra, C. L.; Mayo, S. L., Electrostatics in computational protein design. *Current Opinion in Chemical Biology* **2005,** *9* (6), 622–626.

6. Dahiyat, B. I.; Mayo, S. L., De novo protein design: Fully automated sequence selection. *Science* **1997,** *278* (5335), 82–87.

7. Malakauskas, S. M.; Mayo, S. L., Design, structure and stability of a hyperthermophilic protein variant. *Nature Structural Biology* **1998,** *5* (6), 470–475.

8. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D., Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003,** *302* (5649), 1364–1368.

9. Looger, L. L.; Dwyer, M. A.; Smith, J. J.; Hellinga, H. W., Computational design of receptor and sensor proteins with novel functions. *Nature* **2003,** *423* (6936), 185–190.

10. Shimaoka, M.; Shifman, J. M.; Jing, H.; Takagi, L.; Mayo, S. L.; Springer, T. A., Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nature Structural Biology* **2000,** *7* (8), 674–678.

11. Pierce, N. A.; Winfree, E., Protein design is NP-hard. *Protein Engineering* **2002,** *15* (10), 779–782.

12. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W., A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design* **2002,** *16* (3), 151–166.

13. Lilien, R. H.; Stevens, B. W.; Anderson, A. C.; Donald, B. R., A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the

substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *Journal of Computational Biology* **2005,** *12* (6), 740–761.

14. Chakrabarti, R.; Klibanov, A. M.; Friesner, R. A., Sequence optimization and designability of enzyme active sites. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (34), 12035–12040.

15. Chakrabarti, R.; Klibanov, A. M.; Friesner, R. A., Computational prediction of native protein ligand-binding and enzyme active site sequences. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (29), 10153–10158.

16. Dwyer, M. A.; Looger, L. L.; Hellinga, H. W., Computational design of a biologically active enzyme. (Retracted article: See vol. 319, p. 569, 2008). *Science* **2004,** *304* (5679), 1967–1971.

17. Dwyer, M. A.; Looger, L. L.; Hellinga, H. W., Computational design of a biologically active enzyme. (Retraction of vol. 304, p. 1967, 2004). *Science* **2008,** *319* (5863), 569–569.

18. Hellinga, H. W.; Richards, F. M., Construction of New Ligand-Binding Sites in Proteins of Known Structure .1. Computer-Aided Modeling of Sites with Predefined Geometry. *Journal of Molecular Biology* **1991,** *222* (3), 763–785.

19. Dunbrack, R. L.; Cohen, F. E., Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science* **1997,** *6* (8), 1661–1681.

20. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C., The penultimate rotamer library. *Proteins—Structure Function and Genetics* **2000,** *40* (3), 389–408.

21. Ponder, J. W.; Richards, F. M., Tertiary Templates for Proteins - Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes. *Journal of Molecular Biology* **1987,** *193* (4), 775–791.

22. Shetty, R. P.; de Bakker, P. I. W.; DePristo, M. A.; Blundell, T. L., Advantages of fine-grained side chain conformer libraries. *Protein Engineering* **2003,** *16* (12), 963–969.

23. Xiang, Z. X.; Honig, B., Extending the accuracy limits of prediction for side-chain conformations. *Journal of Molecular Biology* **2001,** *311* (2), 421–430.

24. Dahiyat, B. I.; Mayo, S. L., Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America* **1997,** *94* (19), 10172–10177.

25. Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L., Automated design of the surface positions of protein helices. *Protein Science* **1997,** *6* (6), 1333–1337.

26. Lazaridis, T.; Karplus, M., Effective energy function for proteins in solution. *Proteins—Structure Function and Genetics* **1999,** *35* (2), 133–152.

27. Street, A. G.; Mayo, S. L., Pairwise calculation of protein solvent-accessible surface areas. *Folding & Design* **1998,** *3* (4), 253–258.

28. Allen, B. D.; Mayo, S. L., Dramatic performance enhancements for the FASTER optimization algorithm. *Journal of Computational Chemistry* **2006,** *27* (10), 1071–1075.

29. Desmet, J.; Spriet, J.; Lasters, I., Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **2002,** *48* (1), 31–43.

30. Gordon, D. B.; Hom, G. K.; Mayo, S. L.; Pierce, N. A., Exact rotamer optimization for protein design. *Journal of Computational Chemistry* **2003,** *24* (2), 232–243.

31. Lassila, J. K.; Keeffe, J. R.; Oelschlaeger, P.; Mayo, S. L., Computationally designed variants of Escherichia coli chorismate mutase show altered catalytic activity. *Protein Engineering Design & Selection* **2005,** *18* (4), 161–163.

32. Benkovic, S. J.; Hammes-Schiffer, S., A perspective on enzyme catalysis. *Science* **2003,** *301* (5637), 1196–1202.

33. Kraut, D. A.; Carroll, K. S.; Herschlag, D., Challenges in enzyme mechanism and energetics. *Annual Review of Biochemistry* **2003,** *72*, 517–571.

34. Bolon, D. N.; Voigt, C. A.; Mayo, S. L., De novo design of biocatalysts. *Current Opinion in Chemical Biology* **2002,** *6* (2), 125–129.

35. Lee, A. Y.; Karplus, P. A.; Ganem, B.; Clardy, J., Atomic-Structure of the Buried Catalytic Pocket of Escherichia-Coli Chorismate Mutase. *Journal of the American Chemical Society* **1995,** *117* (12), 3627–3628.

36. Hyre, D. E.; Le Trong, I.; Merritt, E. A.; Eccleston, J. F.; Green, N. M.; Stenkamp, R. E.; Stayton, P. S., Cooperative hydrogen bond interactions in the streptavidin-biotin system. *Protein Science* **2006,** *15* (3), 459–467.

37. Jogl, G.; Rozovsky, S.; McDermott, A. E.; Tong, L., Optimal alignment for enzymatic proton transfer: Structure of the Michaelis complex of triosephosphate isomerase at 1.2-angstrom resolution. *Proceedings of the National Academy of Sciences of the United States of America* **2003,** *100* (1), 50–55.

38. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C., Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology* **1999,** *285* (4), 1735–1747.

39. Wiest, O.; Houk, K. N., On the Transition-State of the Chorismate-Prephenate Rearrangement. *Journal of Organic Chemistry* **1994,** *59* (25), 7582–7584.

40. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., Dreiding — a Generic Force-Field for Molecular Simulations. *Journal of Physical Chemistry* **1990,** *94* (26), 8897–8909.

41. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E., Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **1953,** *21* (6), 1087–1092.

42. Voigt, C. A.; Gordon, D. B.; Mayo, S. L., Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology* **2000,** *299* (3), 789–803.