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

Ryan H. Lilien[*,†,‡] Brian W. Stevens[‡,§] Amy C. Anderson[‡,¶,**] Bruce R. Donald[*,‡,¶,‖,**]

Abstract: Realization of novel molecular function requires the ability to alter molecular complex formation. Enzymatic function can be altered by changing enzyme-substrate interactions via modification of an enzyme's active site. A redesigned enzyme may either perform a novel reaction on its native substrates or its native reaction on novel substrates. A number of computational approaches have been developed to address the combinatorial nature of the protein redesign problem. These approaches typically search for the global minimum energy conformation among an exponential number of protein conformations. We present a novel algorithm for protein redesign, which combines a statistical mechanics-derived ensemble-based approach to computing the binding constant with the speed and completeness of a branch-and-bound pruning algorithm. In addition, we developed an efficient deterministic approximation algorithm, capable of approximating our scoring function to arbitrary precision. In practice, the approximation algorithm decreases the execution time of the mutation search by a factor of ten. To test our method, we examined the Phe-specific adenylation domain of the non-ribosomal peptide synthetase gramicidin synthetase A (GrsA-PheA). Ensemble scoring, using a rotameric approximation to the partition functions of the bound and unbound states for GrsA-PheA, is first used to predict binding of the wildtype protein and a previously described mutant (selective for leucine), and second, to switch the enzyme specificity toward leucine, using two novel active site sequences computationally predicted by searching through the space of possible active site mutations. The top scoring *in silico* mutants were created in the wet-lab and dissociation / binding constants were determined by fluorescence quenching. These tested mutations exhibit the desired change in specificity from Phe to Leu. Our ensemble-based algorithm which flexibly models both protein and ligand using rotamer-based partition functions, has application in enzyme redesign, the prediction of protein-ligand binding, and computer-aided drug design.

Categories and Subject Descriptors: J.3 [Life and Medical Sciences]: Biology and genetics

General Terms: Algorithms, Measurement, Design, Experimentation

Keywords: Protein design, Enzyme design, Protein flexibility, Protein-ligand binding, Molecular ensemble, Non-ribosomal peptide synthetase, Fluorescence binding assay

Abbreviations used: DTT, dithiothreitol; GMEC, global minimum energy conformation; GrsA-PheA, gramicidin synthetase A - phenylalanine adenylation domain; LB, luria broth; NRPS, non-ribosomal peptide synthetase; PCR, polymerase chain reaction; PMSF, phenylmethanesulfonyl fluoride; RMSD, root mean square distance; WT, wildtype

[*]Dartmouth Computer Science Department, Hanover, NH 03755.
[†]Dartmouth Medical School, Hanover, NH 03755.
[‡]Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755.
[§]Dartmouth Department of Biochemistry, Hanover, NH 03755.
[¶]Dartmouth Department of Chemistry, Hanover, NH 03755.
[‖]Dartmouth Department of Biology, Hanover, NH 03755.
[**]Corresponding authors, Bruce.R.Donald@dartmouth.edu and Amy.C.Anderson@dartmouth.edu. This work is supported by grants to B.R.D. from the National Institutes of Health (R01 GM-65982), and the National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, EIA-9802068, and EIA-0305444), and to A.C.A. from the NSF (0133469) and the Research Corporation (RI0857).

RECOMB'04, March 27–31, 2004, San Diego, California, USA.

1 Introduction

In a variety of fungi and parasites, non-ribosomal peptide synthetase (NRPS) enzymes complement the traditional ribosomal peptide synthesis pathway. NRPS enzymes produce peptide-like products via the incorporation of both standard and non-standard amino acid precursors. Unlike the ribosome, many NRPS enzymes methylate or epimerize their amino acid substrates, join them with peptide or ester bonds, and sometimes cyclize their final product. NRPS products include natural antibiotics (e.g., penicillin, vancomycin), antifungals, antivirals, anticancer therapeutics, immunosuppressants, and siderophores. Enzymes of the NRPS pathway have multiple domains with individual functions acting in an assembly-line fashion (Figure 1). It is believed that the substrate specificity of the NRPS enzymes is dictated primarily by the 'gatekeeper' adenylation (A) domain that binds and acylates the incoming amino acid, forming an amino-acyl adenylate [49, 7, 46]. Recent evidence also indicates that the condensation (C), thiolation (T), and epimerization (E) domains may carry some specificity as well albeit to a lesser extent [1, 53, 14, 27].

Enzyme redesign of NRPS enzymes offers the opportunity to reëngineer biosynthetic pathways, greatly increasing the number and types of NRPS products. Therefore, the interest in redesigning NRPS enzymes is motivated by the long-range goal of reprogramming the enzymatic pathway to achieve combinatorial biosynthe-

sis, and the development of new libraries of antibiotics [5]. We explore the idea of reprogramming NRPS enzymes by introducing K^*, an ensemble-based protein redesign algorithm, to analyze and redesign the phenylalanine adenylation domain of the NRPS enzyme gramicidin synthetase A (GrsA-PheA).

1.1 Computational Protein Design

A variety of computational protein redesign efforts have recently been reported. Incorporation of molecular flexibility into protein design is essential; every previous structure-based protein design algorithm has included some notion of flexibility [51, 21, 20, 18, 3, 36, 28, 22, 47]. Many protein design algorithms treat the peptide backbone as rigid and model amino acid side-chain flexibility with a rotamer library containing a discrete set of side-chain conformations [30, 41].

Design algorithms have assumed that for a given protein sequence, folding and binding can best be predicted by examining the single global minimum energy conformation (GMEC). Unfortunately, protein design by searching for the GMEC using rotamers and a pairwise energy function on a rigid peptide backbone has recently been shown to be NP-hard [39]. As a result, a number of heuristic (random sampling, neural network, genetic algorithm) GMEC-based approaches for protein design have been reported [51, 21, 20, 18, 31]; however, the dominant algorithm for assisting in the GMEC search has been dead-end elimination (DEE) [12, 25, 40]. Given a protein backbone, a set of allowable mutations, and a rotamer library, DEE employs a number of sophisticated conformation pruning techniques to prune conformations that are provably not part of the GMEC. Typically, DEE will eliminate the vast majority of mutation sequences; remaining sequences can subsequently be scored and ranked. Growing evidence supports the hypothesis that protein-ligand binding can involve a number of low-energy bound states [11, 42, 35, 55, 56, 33]. Therefore, we have developed a scoring method for protein-ligand redesign based on molecular ensembles. Molecular ensembles have been successfully utilized in structure-based drug design: most commonly, molecular docking is performed against each member of an ensemble or a unified ensemble model and an average or best interaction energy between the protein and ligand may be retained [8, 37, 26, 4, 23, 6].

1.2 Previous NRPS Redesign

NRPS enzyme redesign methods can be divided into two main techniques, *domain-swapping* and *active site modification through site-directed mutagenesis*. Domain-swapping techniques do not require computational analysis nor knowledge of molecular structure; NRPS enzymes are modified by swapping an adenylation domain of an existing NRPS enzyme for an adenylation domain from a second, different NRPS enzyme (carrying a different substrate specificity) [50, 45, 13, 34]. Results of domain-swapping experiments led to the hypothesis that the disruption of native domain:domain interfaces vitiates the proper transfer of synthesis intermediates thereby degrading catalytic efficiency [27]. Emphasis on the importance of domain:domain interactions and domain specificity has directed domain-swapping work to include simultaneous cloning of A, C, and T domains, demonstrating increased yield [13, 34].

The second method for NRPS redesign, active site modification through site-directed mutagenesis, utilizes structural information of the GrsA-PheA enzyme (1AMU [9]). Sequence alignment of GrsA-PheA with 160 other known adenylation domains supports the hypothesis that NRPS adenylation domains, specific for different amino acid substrates, share a similar overall structure differing mainly in the composition of residues lining the active site [49, 16, 7]. A "signature sequence" can be derived for each adenylation domain by extracting those residues that align with the

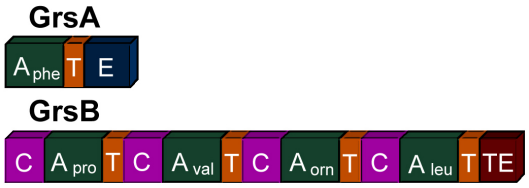

Figure 1: Gramicidin S Synthetase is composed of two NRPS proteins, GrsA (3 domains) and GrsB (13 domains). Gramicidin S is produced in an assembly-line manner where two D-Phe-L-Pro-L-Val-L-Orn-L-Leu peptides are joined and cyclized. (A: Adenylation, T: Thiolation (Peptidyl Carrier Protein), E: Epimerization, C: Condensation, TE: Thioesterase)

structurally-determined substrate binding pocket of the GrsA-PheA crystal structure. By abstracting away from the GrsA-PheA crystal structure, mutations are suggested for a given amino acid substrate by sequence comparison alone. Using signature sequences, Stachelhaus, *et al.* [49] mutated two adenylation domains: PheA, successfully switching the substrate specificity from Phe to Leu; and a second adenylation domain that naturally accepts Asp to accept Asn. More recently, Eppelmann, *et al.* [16] changed the Glu adenylation domain of surfactin synthetase A to accept Gln.

In summary, previous NRPS redesign methods include domain-swapping and site-directed mutagenesis from active site signature sequences. Our method, active site manipulation by site-directed mutagenesis from a computational mutation search utilizing ensemble docking, adds to the armamentarium of techniques available for protein redesign and confers some significant advantages over existing NRPS redesign methods. Signature sequence methods project active site information into a consensus sequence, thus losing structural information. Because molecular structure is not explicitly considered during redesign, successful redesign is more difficult if there are significant structural differences affecting the overall active site shape between the A domains accepting the natural and target substrates. Our method builds mutations into the high-resolution structure of the wildtype enzyme, thus mitigating potential problems arising from these structural differences. In contrast to domain-swapping, our method is more likely to preserve the NRPS enzyme's native modular structure, thus maintaining crucial specific domain:domain interface regions. Finally, unlike either signature sequence or domain-swapping techniques, our method can propose mutations for substrates for which no existing adenylation domain sequences are known.

When considering the use of molecular ensembles for protein design, a major challenge has been the development of ensemble-based redesign algorithms that efficiently prune mutations and conformations. In this paper we introduce the K^* method which generalizes Boltzmann-based scoring to ensembles and applies the result to protein design. The following contributions are made in this work:

1. Introduction of the first ensemble-based protein redesign algorithm;
2. Development of ϵ-approximation algorithms for K^* capable of pruning the vast majority of conformations from more computationally expensive consideration thereby reducing execution time and making the mutation search computationally feasible;
3. The use of K^* to reproduce known adenylation domain binding experiments;
4. The use of K^* to predict novel mutation sequences capable of switching substrate specificity of GrsA-PheA;

5. Confirmation of the K^* method by the creation of predicted protein mutants in the wetlab and testing of their specificity by fluorescence quenching binding assays.

2 Methods

2.1 Ensemble Scoring Method (K^*)

Protein-ligand binding in a single mutation is modeled using the following K^* equation:

$$K^* = \frac{q_{PL}}{q_P q_L}. \quad (1)$$

K^* is derived to be an approximation to the true association (binding) constant K_A by expressing each species' chemical potential as a function of the species partition function q [19, 32] and solving for the equilibrium condition (full details are provided in Appendix A). Unfortunately, it is not currently possible to compute exact partition functions for complex molecular species. This would require integrating an exact energy function over a molecule's entire conformational space. We therefore approximate these partition functions with the use of rotamerically-based conformational ensembles:

$$q_{PL} = \sum_{b \in B} \exp(-E_b/RT), \quad q_P = \sum_{f \in F} \exp(-E_f/RT),$$
$$q_L = \sum_{l \in L} \exp(-E_l/RT), \quad (2)$$

where B, F, and L represent rotamer-based ensembles for the bound protein-ligand complex (PL), the free protein (P), and the free ligand (L) respectively, E_s is the energy of conformation s, R is the gas constant, and T is the temperature in Kelvin. The accuracy with which K^* approximates K_A is proportional to the accuracy of the partition function approximation used.

2.2 Mutation Search

When applying K^* to a protein-ligand system a number of choices must be made with respect to ensemble generation and single-structure scoring. *Single-structure scoring* is the method by which each individual member of an ensemble is scored. These individual scores are then combined using Eqs. (1, 2) to compute an ensemble score. The choices made in ensemble scoring should strike a balance between fidelity to the underlying physical biochemistry, and computational feasibility even in the inner loop of a combinatorially expensive search. A more detailed single structure scoring model takes longer to evaluate and typically cannot be used with large or complex molecules. We present one implementation of the K^* scoring function; however, alternate schemes for both ensemble generation and single-structure scoring can be explored and utilized within the K^* framework. First, in this section, we present a brute-force algorithm that does not utilize filters on mutation space nor pruning of conformation space. The algorithm is then extended to utilize both mutation space filters and conformation space pruning in Section 2.3. The specific application of K^* to redesign GrsA-PheA is described in Section 3.

In the general case, for each allowable active site sequence, K^* is computed using the following three steps.

Step 1: Ensemble Generation. Molecular ensembles are generated by fixing the protein backbone and using the Lovell *et al.* [30] rotamer library to vary side-chain conformation. Each flexibly-modeled residue is allowed to sample all allowable rotamers. Steric clash is determined by computing the distance that two atoms' AMBER van der Waals' (vdW) spheres penetrate each other. Conformations containing any pair of atoms with more than 1.5Å of steric clash are discarded. Finally, if the volume of the active site is large relative to the size of the ligand, multiple translations of the ligand in the active site should be generated and included in the bound ensemble.

Step 2: Ensemble Scoring (Single-Structure Scoring). In the brute-force algorithm, all conformations that pass step 1 are energy-minimized using our implementation of the AMBER energy function (containing electrostatic, vdW, and dihedral terms) [54, 10]. In our model, hydrogen atoms are added to all residues by the LEAP module of the AMBER distribution and are used in computing the electrostatic but not vdW energies. We perform a constrained minimization (analogous to voxel minimization [52, 43]) on each rotameric conformation where side chain dihedrals on flexible residues (including the ligand when present) may move by up to $\pm 9^\circ$ and, for the bound states, the entire ligand may rotate and translate in the active site. This allows our algorithm to sample a larger region of conformation space while not allowing one rotamer to minimize into another.

Step 3: K^* Scoring. The three partition functions q_{PL}, q_P, and q_L are computed separately. The energy minimized scores from Step 2 are used to compute each partition function which is then combined to compute K^* (Eq. 1).

After computing K^* for each mutation, the top mutations (those with the largest K^*) are examined graphically and selected for testing in the wetlab. It is worth noting that although the AMBER scoring function we use for single structure scoring only contains enthalpic terms, the K^* method of ensemble scoring encompasses conformational entropy through use of the partition function over ligand and side chain conformations.

2.3 Efficient Algorithms for Mutation Search

Because protein redesign is NP-Hard [39] there is most likely no way, in the worst case, to avoid having to potentially examine an exponential number of conformations. That is to say, the run time of a protein design algorithm that returns the optimal mutation sequence (under a given metric) is likely to be inherently exponential. In contrast, a random sampling mutation search algorithm can run in sub-exponential time, yet the mutation sequence returned is not guaranteed to be optimal. The combinatorial nature of protein redesign is exacerbated when utilizing an ensemble-based scoring function because multiple low-energy states must be considered for each mutation sequence. Therefore, in designing an ensemble-based mutation search algorithm it is necessary both to prune unlikely conformations and mutations, and also to reduce the runtime constant. Consequently, our mutation search algorithm utilizes the following methods. First, volume and steric filters prune a combinatorial number of conformations from consideration. Thus, analogous to DEE, our pruning techniques eliminate the majority of conformations early in the mutation search. Second, our K^* approximation algorithm quickly discards the majority of remaining conformations by placing provable bounds on each conformation's contribution to the partition function; only those conformations that significantly contribute to the partition function are further considered. This algorithm reduces the average amount of time spent examining each conformation and is essential in practice; without such optimizations the ensemble-based mutation search would not be possible.

Sequence-Space Filters. A residue type filter restricts the mutation search to include only a subset of amino acids based on compatibility with the target substrate. A volume filter removes mutations that significantly over- or under-pack the substrate-bound active site relative to the wildtype.

Deterministic Approximation Algorithm. In a Boltzmann distribution, conformations with large energies are not likely to be assumed and only contribute a vanishingly small amount to the partition function. We therefore prune conformations from consideration (and hence minimization) when we know that they will con-

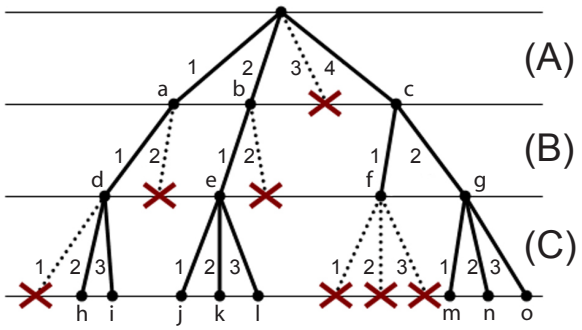

Figure 2: An Example Conformation Tree. The rotamers of flexible residue i are represented by the branches at depth i. Internal nodes of a conformation tree represent partially-assigned conformations. Red $\times$s represent nodes of the conformation tree where steric clash has been identified among a partially assigned conformation. All children of $\times$ nodes are pruned and not considered.

tribute only a small percentage to the total partition function. In Section 2.3.2 we show that the true partition function can be provably approximated to arbitrary precision.

2.3.1 Conformation Generation and the Steric Filter

Active site rotameric conformations are generated by traversing a conformation tree in a depth-first search order. In a conformation tree (Figure 2), the rotamers of flexible residue i are represented by the branches at depth i. For example, in Figure 2, residue A has 4 rotamers, residue B has 2 rotamers, and residue C has 3 rotamers. Internal nodes of a conformation tree represent partially-assigned conformations. For example, in Figure 2, node **e** represents the partially-assigned conformation where residue A has assumed rotamer 2 and residue B has assumed rotamer 1; no rotamer has yet been assigned for residue C. Nodes of the conformation tree are visited in a depth-first search order. As each node is visited, conformations with more than 1.5Å of pre-minimization steric overlap are not further considered, thereby pruning a branch of the conformation tree. For example, when generating the children of node **b**, rotamer 2 is not assigned to residue B since it causes steric clash. That branch of the search tree is pruned and not considered further. Steric clash identified at higher levels of the conformation tree prunes more conformations than the identification of steric clash at lower nodes of the conformation tree. Pruning at depth i eliminates $O(c^{n-i})$ conformations, where c is the average number of rotamers per amino acid type and n is the total number of flexible residues.

2.3.2 Intra-mutation Pruning

We now derive the energetic-based pruning method and quantify the total error accrued by ignoring pruned conformations when computing a single partition function. Because this technique is applied during the computation of a single partition function for a single mutation it is called *intra-mutation pruning*. We show that any desired approximation accuracy to the true partition function can be guaranteed.

Assume that a partition function is to be computed over n sterically allowed conformations. Let $C_k = \{c_1, c_2, \ldots, c_k\}$ be the subset containing the first k conformations such that C_n contains all the conformations. Let q_k be the partial partition function computed by evaluating the first k conformations,

$$q_k = \sum_{c \in C_k} \exp(-E_c/RT).$$

Let q_k^* be an approximation to q_k determined by examining a subset of states S_k, such that $S_k \subseteq C_k$,

$$q_k^* = \sum_{s \in S_k} \exp(-E_s/RT).$$

If S_k contains most of the low energy conformations of C_k then q_k^* will represent a good approximation to q_k. Let p_k be the partition function of the pruned residues $(C_k - S_k)$ equal to the difference $q_k - q_k^*$ such that

$$p_k = \sum_{s \in (C_k - S_k)} \exp(-E_s/RT).$$

One method of ensuring that q_n^* is a good approximation to q_n is to maintain an invariant throughout the computation requiring that q_k^* be a good approximation to q_k. Therefore, we maintain that at any point k, q_k^* must be a good approximation to q_k, that is,

$$q_k^* \geq (1 - \epsilon) q_k \quad \forall k \leq n. \tag{3}$$

Here, ϵ is the desired approximation constant $(\epsilon < 1)$. When Eq. (3) holds, we say that q_k^* is an *ϵ-approximation* to q_k. We can maintain Eq. (3) by ensuring that $p_k \leq q_k \epsilon$. Since we know that $q_k^* \leq q_k$, we can also maintain Eq. (3) and therefore that q_k^* remains an ϵ-approximation to q_k by ensuring that

$$p_k \leq q_k^* \epsilon. \tag{4}$$

To determine a pruning criterion, we assume that we have already considered k conformations and that S_k contains the subset of fully evaluated conformations. To prune conformation c_{k+1} we must first guarantee that after pruning c_{k+1}, the value q_{k+1}^* is an ϵ-approximation to q_{k+1}. We know that

$$q_{k+1} = q_k^* + p_k + \exp(-E_{c_{k+1}}/RT).$$

If we prune c_{k+1} then

$$p_{k+1} = p_k + \exp(-E_{c_{k+1}}/RT) \tag{5}$$

$$q_{k+1}^* = q_k^*. \tag{6}$$

To maintain the invariant we need to ensure $p_{k+1} \leq q_{k+1}^* \epsilon$ which can be rewritten using Eqs. (5, 6) as

$$p_k + \exp(-E_{c_{k+1}}/RT) \leq q_k^* \epsilon \tag{7}$$

Solving Eq. (7) for $E_{c_{k+1}}$, we get that

$$E_{c_{k+1}} \geq -RT \ln\left(q_k^* \epsilon - p_k\right). \tag{8}$$

Therefore, if the energy of conformation c_{k+1} satisfies Eq. (8) then we can prune conformation c_{k+1} while maintaining $q_{k+1}^* \geq (1 - \epsilon) q_{k+1}$ (that is, q_{k+1}^* is an ϵ-approximation to q_{k+1}). Of course, the purpose of pruning is to avoid the computationally expensive energy minimization required to obtain $E_{c_{k+1}}$. In practice, we use a *pairwise energy matrix* to compute a lower bound on $E_{c_{k+1}}$ (full details are in Appendix D) and compare this bound to Eq. (8). Because we don't compute $E_{c_{k+1}}$, we cannot maintain an exact value for p_k. Therefore, the lower bounds on $E_{c_{k+1}}$ are used to compute p_k^*, an upper bound on p_k, namely

$$p_k \leq p_k^* = \sum_{s \in (C_k - S_k)} \exp(-B(s)/RT),$$

where $B(s)$ returns a lower bound on the energy of conformation s. Because $-RT \ln\left(q_k^* \epsilon - p_k^*\right) \geq -RT \ln\left(q_k^* \epsilon - p_k\right)$, the approximation p_k^* can be used to determine the practical pruning criteria, that is: *conformation c_{k+1} can be pruned* if

$$B(c_{k+1}) \geq -RT \ln\left(q_k^* \epsilon - p_k^*\right). \tag{9}$$

This leads to the following lemma:

LEMMA 1. *If Eq. (9) is satisfied by conformation c_{k+1}, then conformation c_{k+1} can be pruned and q_{k+1}^* is guaranteed to be an ϵ-approximation to q_{k+1}.*

Let $n \leftarrow$ Number of Rotameric Conformations
Let $c \leftarrow$ Rotameric Conformations
Initialize: $q^* \leftarrow 0,\ p^* \leftarrow 0$
for $k = 1$ to n
    if B(c_k) $\leq -RT \ln (q^* \epsilon - p^*)$
        $q^* \leftarrow q^* + \exp(-\texttt{ComputeMinEnergy}(c_k)/RT)$
    else
        $p^* \leftarrow p^* + \exp(-\text{B}(c_k)/RT)$
Return q^*

Figure 3: INTRA-MUTATION PRUNING. **q^* is the running approximation to the partition function, p^* is an upper bound on the partition function of the pruned conformations. The function B($\cdot$) computes a lower energy bound for the given conformation. The function `ComputeMinEnergy`($\cdot$) returns the energy of the energy-minimized conformation as computed using steepest-descent minimization and our implementation of the AMBER energy function (as described in *Step 2* of Section 2.2). At the end, q^* represents an ϵ-approximation to the true partition function q such that $q^* \geq (1-\epsilon)q$.**

Maintaining the invariant that after considering k conformations q_k^* is an ϵ-approximation to q_k leads to the following lemma:

LEMMA 2. *If q_k^* is an ϵ-approximation to q_k for all k $(0 < k \leq n)$ then by induction, at the end of the computation, q_n^* will be an ϵ-approximation to q_n.*

The intra-mutation pruning algorithm is shown in Figure 3.

When using K^* to perform a mutation search we can bootstrap the pruning condition for improved efficiency (by caching partition functions, we can exploit K^* bounds from other mutations in the same search). Our search algorithm has the desirable property that provably accurate ϵ-approximations are computed for top-ranking mutations, while the bounds we can prove on the quickly-computed K^* values for lower-ranked mutations do not enjoy the same degree of accuracy. Our algorithm requires an ϵ-approximation only for those mutations with $K^* \geq \gamma \max_{i \leq m} K_i^*$, where K_i^* is the K^* value of the i^{th} mutation, m is the total number of mutations, and γ $(0 < \gamma \leq 1.0)$ is a user-specified constant defining a range of K^* values that must be computed accurately. Since the value of the largest (best) K^* is not known during the mutation search, we use K_o^*, the largest K^* value encountered so far. Similarly to Eq. (9), after examining k of n total conformations of a partition function q_{PL}, evaluation of conformation c_{k+1} can be skipped if

$$B(c_{k+1}) \geq -RT \ln (\psi_k - p_k^*), \tag{10}$$

where $\psi_k = \max(\epsilon K_o^* q_L \gamma q_P, q_k^* \epsilon)$. The full derivation of Eq. (10) is in Appendix C.

3 Results & Discussion

3.1 Structural Model

Our structural model employs the previously solved structure of GrsA-PheA (1AMU) [9] and consists of the 9 active site residues (D235, A236, W239, T278, I299, A301, A322, I330, C331) (Figure 4), the 30 residues with at least one atom within 8Å of the active site, termed the steric shell, the amino acid substrate, and the AMP cofactor. The steric shell allows us to compute the interaction energy of the active site residues with neighboring regions of the protein and constrains the active site residues from assuming conformations that would sterically clash with the body of the PheA protein.

3.2 Comparison to Wildtype PheA

We performed a series of experiments to confirm the proper implementation of the steric filter, the rotamer library, the AMBER energy function, and the minimization algorithm. The first test was designed to confirm that we could find an accepted (crystallographically confirmed) conformation for Phe in the PheA active site. The bound partition function q_{PL} was computed for Phe in the GrsA-PheA wildtype (WT) protein. The energy calculated for the best minimized rotamer structure (i.e., the lowest computed energy) and that calculated for the crystal structure are within 5% of each other and have a non-hydrogen atom RMSD of 0.66Å (Figure 4A). In the bound molecular ensemble, approximately 39 conformations of the PheA:Phe complex have energies within 5% of the minimum (Figure 4B). This observation supports the hypothesis that multiple structures have energies contributing to the weighted ensemble. This test confirmed that we were able to generate structures compatible with the X-ray structure and therefore demonstrated the feasibility of both the rotamer search strategy with minimization and the scoring scheme.

3.3 T278M/A301G Double Mutant

To further test our model we simulated the biochemical activity assays of L-Phe and L-Leu against wildtype PheA and the T278M/A301G double mutant [49]. The T278M/A301G double mutant was designed by signature sequence homology modeling by Stachelhaus *et al.* [49] to be similar to a known Leu adenylation domain. K^* scores were computed for each substrate in each active site and compared with activity assays performed by Stachelhaus *et al.* [49]. Because the experimentally measured solvation energy of Leu and Phe are similar (2.3 vs. 2.4 kcal/mol) [15, 17], in these experiments we chose not to determine q_L computationally but rather treat the q_L values of Leu and Phe as equivalent. Stachelhaus *et al.* normalized the activity of each protein such that the substrate with the most activity was assigned a specificity of 100%. The wildtype enzyme has a normalized specificity of 100% for Phe and approximately 10% for Leu while the T278M/A301G double mutant has a normalized specificity of approximately 40% for Phe and 100% for Leu [49]. Our normalized K^* results closely agree with these specificity scores. For the wildtype enzyme, PheA has a normalized K^* for Phe of 100% and for Leu of 6%. The double mutant enzyme has a K^* of 10% for Phe and 100% for Leu. Although K^* is a binding constant approximation, its results qualitatively agree with the activity assays of Stachelhaus *et al.*

3.4 Redesign for Leu

A K^* mutation search was performed to redesign GrsA-PheA to bind and adenylate Leu instead of Phe. The 9 active site residues, 30 active site neighboring residues, substrate, and AMP were modeled as described above. We performed a 2-residue mutation search, where any 2 of the 9 active site residues were allowed to mutate to any of the hydrophobic residues (GAVLIFYWM) (2916 possible mutations, appx. 6.8×10^8 conformations). Each mutation was checked against the volume filter. Active sites that were over- or under-packed relative to Phe in the PheA wildtype by more than 30Å^3 were eliminated. 1011 mutations (35% of the total), containing 1.98×10^8 conformations, passed the volume filter and were fully evaluated. Of the 822,061 conformations that pass the steric pruning step, 742,116 (90.3%) are pruned based on minimum energy bounds (Eq. 10) leaving only 79,945 conformations of the original 6.8×10^8 that were then energy minimized and scored (Table 1). The 2-residue mutation search took less than 1 day on a cluster of 18 1.6GHz Athlon processors. Without energetic pruning approximately 10 times as many conformations would require minimization taking approximately 10 times longer to execute. Compu-

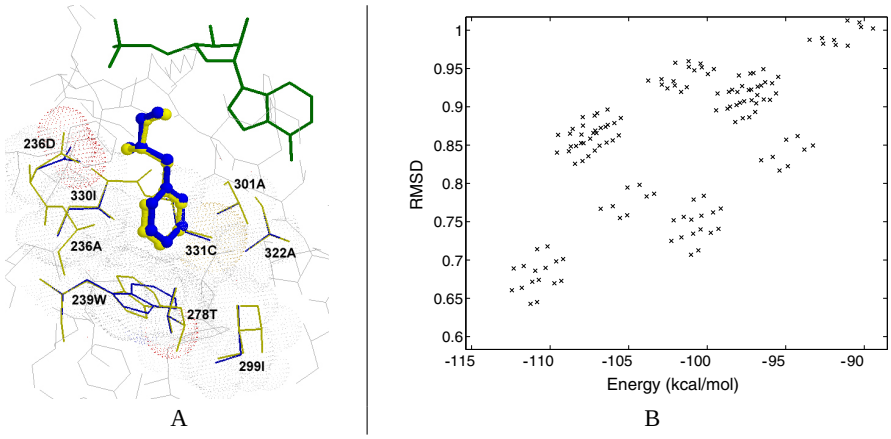

Figure 4: Comparison of K^*-Predicted and Crystal Structures of GrsA-PheA. (A) The lowest-energy ensemble member of Phe in WT GrsA-PheA is shown with the crystal structure (RMSD: 0.66Å). Active site residues are yellow (predicted) and blue (crystal), Phe substrate is ball and stick, AMP is green, and residues immediately surrounding the active site (the steric shell) are grey wireframe. The energy-minimized rotamer-based prediction accurately reproduces the crystallographic conformation. The main structural difference occurs in 239W where a small difference in the (C_α-C_β-C_γ) bond angle (118.3 (resp. 115.1) degrees in the crystal structure (resp. prediction)) prevents the rotamer-based conformation from more closely matching the crystallographic conformation. (B) Conformational energy (kcal/mol) vs. RMSD for conformations in the K^* generated ensemble for Phe in wildtype GrsA-PheA. RMSDs are computed between each ensemble conformation and the crystal structure [9] using all non-hydrogen atoms. Low-energy conformations have a lower RMSD than high energy conformations and conformations with low RMSDs have lower energies than conformations with larger RMSDs.

tationally, for comparison, a three-point mutation search has been run in 10 days; consequently, due to the computation required in minimizing each rotameric conformation, execution without pruning is impractical. In practice, the accuracy of the computed solution is significantly higher than that guaranteed by the approximation. When a 3% approximation (accuracy = 97%) is requested, the accuracy achieved is over 99% which suggests that we may relax the pruning criteria and still maintain an excellent approximation.

The two mutation sequences with the best K^* scores are A301G/I330W and A301G/I330F: these novel mutations are unknown in nature and have never been tested before. The lowest energy predicted conformations of the bound ensemble for the best two mutation sequences are shown in Figure 5. The first mutation in both sequences, A301G, sterically allows for the difference in position of the C_δ atoms between Phe and Leu. Residue 301G appears in 69% of the top 40 K^*-ranked mutations (Figure 5C) and is also present in all 19 known native Leu adenylation domains [7]. The second mutation (I330W, I330F) fills the bottom of the substrate binding pocket accounting for the difference in size between Phe and Leu. Both mutations I330W and I330F form a staggered stacked ring structure [44] with the existing residue 239W (Figure 5). The previously-reported T278M/A301G mutant [49] is ranked 12th out of 2916 by K^*, thus demonstrating that a known Leu binding mutation is ranked highly in our mutation search.

The GrsA-PheA gene was cloned into the QE60 vector using PCR. *E. coli* M15 cell lines were transformed with the constructed plasmid. These genes where then modified to incorporate the two desired mutations, 301G/330W and 301G/330F. The reader can find complete expression and purification details in Appendix E. For both mutations, the presence of Trp239 in the active site allowed us to determine the dissociation constants for substrate binding (K_D) by measuring the change in fluorescence of Trp at 340nm after titrating substrate and exciting at 280nm. The dissociation constant K_D is inversely proportional to the binding constant $K_A = 1/K_D$. Hence, a smaller K_D is associated with tighter binding. Both mutations clearly exhibit stronger binding for Leu than for Phe (Table 2), and the K_D measured for Leu in both redesigned proteins is approximately half that for Phe, strongly demonstrating the success of the protein redesign.

	Conformations Remaining	Pruning Factor (%)	Pruning Type
Initial	6.8×10^8	-	-
Volume Filter	1.98×10^8	3.43 (70.9)	C
Steric Filter	8.22×10^5	240.86 (99.6)	C
Energy Filter	7.99×10^4	10.28 (90.3)	CF

Table 1: Conformational Pruning. The initial number of conformations for the GrsA-PheA 2-residue Leu mutation search is shown with the number of conformations remaining after the application of volume, steric, and energy (Eq. 10) pruning. The pruning factor represents the ratio of the number of conformations present before and after the given pruning stage. The pruning-% (in parentheses) represents the percentage of remaining conformations eliminated by the given pruning stage. The combined pruning factor of all filters is 8510. The pruning type column indicates if the pruning represents a combinatorial (C) or a constant factor (CF) speedup.

4 Comparison to GMEC Search

For comparison, we explored whether the mutation sequences suggested by the K^* mutation search were indeed different from those that would have been found using a GMEC type search. We therefore compared the K^*-based mutation sequence ranking to two non-ensemble based scoring techniques. The first alternative mutation scoring method scores each mutation sequence by the lowest minimum energy bound conformation (among all allowable rotameric conformations for the given mutation sequence). We refer to this brute-force approach as the *minimum energy (ME)* technique. The second alternative scoring method simulates the rankings that would be returned by a DEE-type search. This method consists of two stages. The first stage is an initial pruning step based on the lowest-energy bound rotameric conformation (no energy minimization is performed). In the second stage, a ranking is

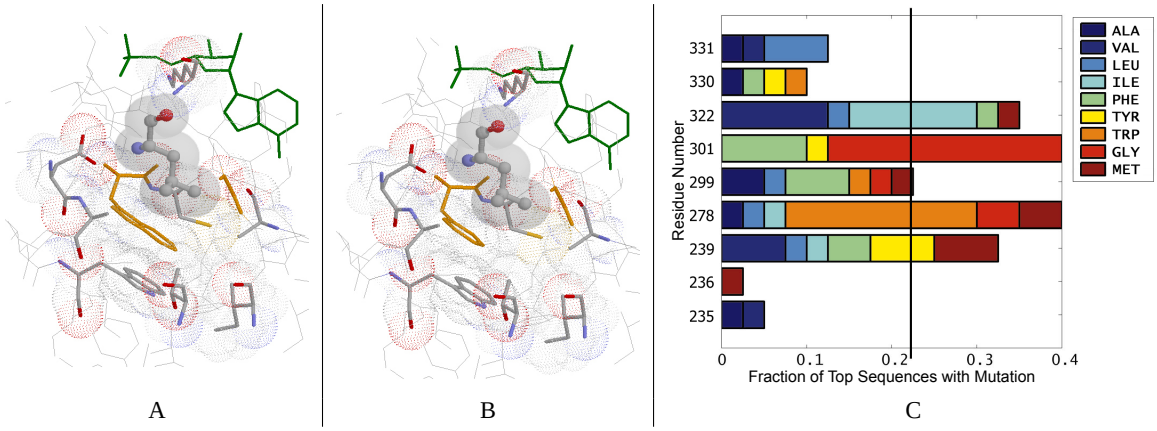

Figure 5: The top two K^*-predicted mutations and the mutation frequency in the 40 top-ranking mutations. Shown are the lowest energy ensemble members of the bound partition functions for (A) A301G/I330W and (B) A301G/I330F. Residue 301G sterically allows for the Leu C_δ atoms while residues 330W and 330F both stack on residue 239W and serve to fill a void at the bottom of the active site created by the difference in size between Phe and Leu. Mutated residues are shown in orange. (C) The fraction of the top 40 K^*-ranked sequences involving the specified residues. If mutations were randomly distributed one would expect that each residue would mutate 2/9 (22.2%) of the time (indicated by the vertical black line). Therefore, residues 235, 236, 330, and 331 tend to assume the wildtype amino acid in the K^*-predicted distribution.

computed for each remaining mutation based on an energy-minimized binding energy ($E_{bound} - E_{unbound}$). We therefore refer to this method as the the *minimized binding energy (MBE)* technique. In computing the MBE ranking, the top 10% scoring mutations from the first stage are scored in stage two. The rankings produced by this method therefore approximate those of a DEE-type search in that the vast majority of mutation sequences are pruned in the DEE stage based on the lowest energy bound rotameric conformation, and unpruned mutations are subsequently scored using a more sophisticated method. To compare the mutation sequence scoring technique rather than the energy function, both the ME and MBE techniques use the same implementation of the AMBER empirical energy function and steepest descent energy minimization used by K^*.

The 3 scoring methods were applied to all 1011 2-residue mutation sequences that passed the volume filter during Leu redesign. Of the top 40 K^*-ranked mutations, only 2 (5%) appear among the top 40 ME-ranked mutations and only 7 (17.5%) appear among the top 100 ME-ranked mutations. Conversely, of the top 40 ME-ranked mutations, only 2 (5%) appear among the top 40 K^*-ranked mutations and only 8 (20%) appear among the top 100 K^*-ranked mutations. When compared to the MBE method, of the top 40 K^*-ranked mutations only 10 (25%) appeared in either the top 40 or the top 100 MBE-ranked mutations. Conversely, of the top 40 MBE-ranked mutations, 10 (25%) appear among the top 40 K^*-ranked mutations and 23 (57.5%) appear among the top 100 K^*-ranked mutations. Perhaps most interesting is the result that neither of the top two K^*-ranked mutations (A301G/I330W and A301G/I330F) were found among the top 100 ME- or MBE-ranked mutations. Furthermore, the previously reported (T278M/A301G) Leu binding mutation [49] is ranked 80th by the ME method, 3rd by the MBE method, and 12th by the K^* method.

Most of the top K^*-, ME-, and MBE-ranked mutations remain biologically untested, thus precluding an exhaustive comparison of the mutations predicted by the three scoring techniques; however, we can conclude that the top mutation sequences returned by K^* are different from those returned by either ME or MBE. The tested top K^* mutations, shown in this paper to have Leu binding specificity (by wetlab experiments), provide evidence that the K^* rankings provide an additional and effective method for ranking mutation sequences.

Substrate	Dissociation Constant (μM) K_D WT	301G/330W	301G/330F
L-Phe	26	91	52
L-Leu	50	44	31

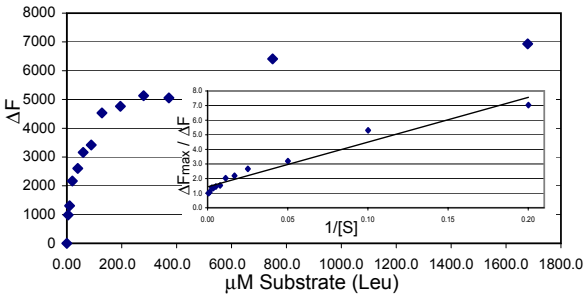

Table 2: (Top) Dissociation Constants measured by fluorospectrophotometry for amino acid binding to GrsA-PheA and the novel mutants 301G/330W and 301G/330F. A lower dissociation constant is associated with tighter binding. (Bottom) Change in fluorescence vs. Leu concentration for 301G/330F. (Inset) Reciprocal plot of the fluorescence quenching data and the linear regression (correlation coefficient of 0.977) showing a dissociation constant of 31μM.

5 Conclusions

The K^* ensemble scoring method presented here was successfully applied to redesign the Phe adenylation domain of gramicidin synthetase A. This represents the first use of an ensemble-based scoring function for enzyme redesign. Despite the inherently exponential nature of ensemble-based scoring, a deterministic approximation algorithm for computing each partition function enables sufficient pruning to make the search feasible. The redesigned enzymes demonstrate a specificity switch from Phe to Leu in binding affinity and we are now pursuing enzyme activity assays to determine the rate of amino-acyl adenylate formation for the designed proteins. Our ensemble-based mutation search algorithm represents a novel and effective alternative to both domain-swapping and the use of signature sequences for NRPS adenylation domain modification.

Many previous modeling algorithms have used biophysically-motivated scoring functions to rank results (i.e., LUDI score [2],

DOCK score [24, 29]). Although based on biophysical phenomenon, these scores often do not provide accurate absolute binding information but rather are useful in predicting relative binding. K^* represents a similar type of scoring function. At present, K^* can best provide relative binding to rank mutations for a given ligand. In the future, we hope to enhance K^* to provide more information on absolute binding.

While GMEC-based approaches remain the dominant algorithm for protein design, because of their ability to handle the design of large proteins, we propose that, when active site flexibility and multiple binding modes must be considered for the redesign of a moderate-sized system, K^* represents an accurate and feasible approach to ensemble-based redesign. Although in this paper we demonstrate K^* as a stand-alone algorithm, in larger systems DEE can be used to reduce the exponential number of mutation sequences and conformations prior to K^* scoring. In this manner, K^* could be used to efficiently rank those mutation sequences that survive DEE pruning, leading to a DEE-K^* hybrid search. Enhancements to our pruning methods should increase both the fraction of sequence space searched during protein design and the size of active site for which K^* is feasible.

It would be interesting to extend our algorithm to create 'sloppy' adenylation domains capable of adenylating several types of amino acids thereby facilitating combinatorial biosynthesis. Such enzymes could, potentially, play a role in drug synthesis analogous to "generic operations" in computer science. Modified synthesis pathways will create multiple final synthesis products, each demonstrating slight variations on the designed product [5]. Such biosynthetic combinatorial diversity should prove useful during the lead-discovery phase of pharmaceutical development.

Acknowledgments We thank Drs. L. Wang, C. Bailey-Kellogg, T. Lozano-Pérez, R. Mettu and B. Tidor, Mr. C. Langmead, Mr. A. Yan, Ms. E. Werner-Reiss, and all members of the Donald and Anderson Labs for helpful discussions and comments on drafts. We are grateful to Dr. H. Higgs for use of his ISS PC1 spectrofluorometer.

6. REFERENCES

[1] BELSHAW, P., WALSH, C., AND STACHELHAUS, T. Aminoacyl-CoAs as probes of condensation domain selectivity in nonribosomal peptide synthesis. *Science 284* (1999), 486–489.

[2] BÖHM, H. On the use of LUDI to search the fine chemicals directory for ligands of proteins of known three-dimensional structure. *J. Comput. Aided Mol. Des. 8* (1994), 623–632.

[3] BOLON, D., AND MAYO, S. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA 98* (2001), 14274–14279.

[4] BOUZIDA, D., REJTO, P., ARTHURS, S., COLSON, A., FREER, S., GEHLHAAR, D., LARSON, V., LUTY, B., ROSE, P., AND VERKHIVKER, G. Computer simulations of ligand-protein binding with ensembles of protein conformations: A monte carlo study of HIV-1 protease binding energy landscapes. *Int. J. Quantum Chem. 72* (1999), 73–84.

[5] CANE, D., WALSH, C., AND KHOSLA, C. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science 282* (1998), 63–68.

[6] CARLSON, H. Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol. 6* (2002), 447–452.

[7] CHALLIS, G., RAVEL, J., AND TOWNSEND, C. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol. 7* (2000), 211–224.

[8] CLAUSSEN, H., BUNING, C., RAREY, M., AND LENGAUER, T. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol. 308* (2001), 377–395.

[9] CONTI, E., STACHELHAUS, T., MARAHIEL, M., AND BRICK, P. Structural basis for the activation of phenylalanine in the nonribosomal biosynthesis of Gramicidin S. *EMBO J. 16* (1997), 4174–4183.

[10] CORNELL, W., CIEPLAK, P., BAYLY, C., GOULD, I., MERZ, K., FERGUSON, D., SPELLMEYER, D., FOX, T., CALDWELL, J., AND KOLLMAN, P. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc. 117* (1995), 5179–5197.

[11] DECANNIERE, K., TRANSUE, T., DESMYTER, A., MAES, D., MUYLDERMANS, S., AND WYNS, L. Degenerate interfaces in antigen-antibody complexes. *J. Mol. Biol. 313* (2001), 473–478.

[12] DESMET, J., MAEYER, M., HAZES, B., AND LASTERS, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356* (1992), 539–542.

[13] DOEKEL, S., AND MARAHIEL, M. Dipeptide formation on engineered hybrid peptide synthetases. *Chem. Biol. 7* (2000), 373–384.

[14] EHMANN, D., TRAUGER, J., STACHELHAUS, T., AND WALSH, C. Aminoacyl-SNACs as small-molecule substrates for the condensation domains of nonribosomal peptide synthetases. *Chem. Biol. 7* (2000), 765–772.

[15] EISENBERG, D., AND MCLACHLAN, A. Solvation energy in protein folding and binding. *Nature 319* (1984), 199–203.

[16] EPPELMANN, K., STACHELHAUS, T., AND MARAHIEL, M. Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry 41* (2002), 9718–9726.

[17] FAUCHÈRE, J., AND PLIŠKA, V. Hydrophobic parameters Π of amino-acid side chains from the partitioning of n-acetyl-amino-acid amides. *Eur. J. Med. Chem. Chim. Ther. 18* (1983), 369–375.

[18] HELLINGA, H., AND RICHARDS, F. Construction of new ligand binding sites in proteins of known structure: I. Computer-aided modeling of sites with pre-defined geometry. *J. Mol. Biol. 222* (1991), 763–785.

[19] HILL, T. *Statistical Mechanics: Principles and Selected Applications*. McGraw-Hill Book Company, Inc., New York., New York, 1956, ch. 1, "Principles of Classical Statistical Mechanics", pp. 1–17.

[20] JARAMILLO, A., WERNISCH, L., HÉRY, S., AND WODAK, S. Automatic procedures for protein design. *Comb. Chem. High Throughput Screen. 4* (2001), 643–659.

[21] JIN, W., KAMBARA, O., SASAKAWA, H., TAMURA, A., AND TAKADA, S. De novo design of foldable proteins with smooth folding funnel: Automated negative design and experimental verification. *Structure 11* (2003), 581–591.

[22] KEATING, A., MALASHKEVICH, V., TIDOR, B., AND KIM, P. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. *Proc. Natl. Acad. Sci. USA 98* (2001), 14825–14830.

[23] KNEGTEL, R., KUNTZ, I., AND OSHIRO, C. Molecular docking to ensembles of protein structures. *J. Mol. Biol 266* (1997), 424–440.

[24] KUNTZ, I., BLANEY, J., OATLEY, S., LANGRIDGE, R., AND FERRIN, T. A geometric approach to macromolecular-ligand interactios. *J. Mol. Biol. 161* (1982), 269–288.

[25] LASTERS, I., AND DESMET, J. The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Protein Eng. 6* (1993), 717–722.

[26] LILIEN, R., SRIDHARAN, M., HUANG, J., BUSHWELLER, J., AND DONALD, B. Computational screening studies for Core Binding Factor Beta: Use of multiple conformations to model receptor flexibility. Poster - 8th International Conference on Intelligent Systems for Molecular Biology ISMB-2000.

[27] LINNE, U., DOEKEL, S., AND MARAHIEL, M. Portability of epimerization domain and role of peptidyl carrier protein on epimerization activity in nonribosomal peptide synthetases. *Biochemistry 40* (2001), 15824–15834.

[28] LOOGER, L., DWYER, M., SMITH, J., AND HELLINGA, H. Computational design of receptor and sensor proteins with novel functions. *Nature 423* (2003), 185–190.

[29] LORBER, D., AND SHOICHET, B. Flexible ligand docking using conformational ensembles. *Protein Sci. 7* (1998), 938–950.

[30] LOVELL, S., WORD, J., RICHARDSON, J., AND RICHARDSON, D. The penultimate rotamer library. *Proteins 40* (2000), 389–408.

[31] MARVIN, J., AND HELLINGA, H. Conversion of a maltose receptor into a zinc biosensor by computational design. *PNAS 98* (2001), 4955–4960.

[32] MCQUARRIE, D. *Statistical Mechanics*. Harper & Row, New York, 1976.

[33] MONTFORT, W., PERRY, K., FAUMAN, E., FINER-MOORE, J., MALEY, G., HARDY, L., MALEY, F., AND STROUD, R. Structure, multiple site binding, and segmental accommodation in thymidylate synthase on binding dUMP and an anti-folate. *Biochemistry 29* (1990), 6964–6977.

[34] MOOTZ, H., SCHWARZER, D., AND MARAHIEL, M. Construction of hybrid peptide synthetases by module and domain fusions. *Proc. Natl. Acad. Sci. USA 97* (2000), 5848–5853.

[35] MURTHY, K., WINBORNE, E., MINNICH, M., CULP, J., AND DEBOUCK, C. The Crystal Strucutres at 2.2-Å Resolution of Hydroxyethylene-based Inhibitors Bound to Human Immunodeficiency Virus Type 1 Protease Show the Inhibitors Are Present in Two Distinct Orientations. *J. Biol. Chem. 267* (1992), 22770–22778.

[36] OFFREDI, F., DUBAIL, F., KISCHEL, P., SARINSKI, K., STERN, A., VAN DE WEERDT, C., HOCH, J., PROSPERI, C., FRANÇOIS, J., MAYO, S., AND MARTIAL, J. De novo backbone and sequence design of an idealized α/β-barrel protein: evidence of stable tertiary structure. *J. Mol. Biol. 325* (2003), 163–174.

[37] ÖSTERBERG, F., MORRIS, G., SANNER, M., OLSON, A., AND GOODSELL, D. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins 46* (2002), 34–40.

[38] PHILIPPOPOULOS, M., AND LIM, C. Exploring the dynamic information content of a protein NMR structure: Comparison of a molecular dynamics simulation with the NMR and X-Ray structures of *Escherichia coli* ribonuclease HI. *Proteins 36* (1999), 87–110.

[39] PIERCE, N., AND E., W. Protein design is *NP*-hard. *Protein Eng. 15* (2002), 779–782.

[40] PIERCE, N., SPRIET, J., DESMET, J., AND MAYO, S. Conformational splitting: a more powerful criterion for dead-end elimination. *J. Comput. Chem. 21* (2000), 999–1009.

[41] PONDER, J., AND RICHARDS, F. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol. 193* (1987), 775–791.

[42] RAAG, R., AND POULOS, L. Crystal structures of cytochrome P-450CAM complexed with camphane, thiocamphor, and adamantane: factors controlling P-450 substrate hydroxylation. *Biochemistry 30* (1991), 2674–2684.

[43] RIENSTRA, C., TUCKER-KELLOGG, L., JARONIEC, C., HOHWY, M., REIF, B., MCMAHON, M., TIDOR, B., LOZANO-PÉREZ, T., AND GRIFFIN, R. De novo determination of peptide structure with solid-state magic-angle spinning NMR spectroscopy. *Proc. Natl. Acad. Sci. USA 99* (2002), 10260–10265.

[44] SAMANTA, U., PAL, D., AND CHAKRABARTI, P. Packing of aromatic rings against tryptophan residues in proteins. *Acta Cryst. D 55* (1999), 1421–1427.

[45] SCHNEIDER, A., STACHELHAUS, T., AND MARAHIEL, M. Targeted alteration of the substrate specificity of peptide synthetases by rational module swapping. *Mol. Gen. Genet. 257* (1998), 308–318.

[46] SCHWARZER, D., FINKING, R., AND MARAHIEL, M. Nonribosomal peptides: from genes to products. *Nat. Prod. Rep. 20* (2003), 275–287.

[47] SHIFMAN, J., AND MAYO, S. Modulating calmodulin binding specificity through computational protein design. *J. Mol. Biol. 323* (2002), 417–423.

[48] STACHELHAUS, T., AND MARAHIEL, M. Modular structure of peptide synthetases revealed by dissection of the multifunctional enzyme GrsA. *J. Biol. Chem. 270* (1995), 6163–6169.

[49] STACHELHAUS, T., MOOTZ, H., AND MARAHIEL, M. The specificiy-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol. 6* (1999), 493–505.

[50] STACHELHAUS, T., SCHNEIDER, A., AND MARAHIEL, M. Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science 269* (1995), 69–72.

[51] STREET, A., AND MAYO, S. Computational protein design. *Structure 7* (1999), R105–R109.

[52] TUCKER-KELLOGG, L. *Systematic Conformational Search with Constraint Satisfaction*. PhD thesis, Massachusetts Institute of Technology, 2002.

[53] WEBER, T., BAUMGARTNER, R., RENNER, C., MARAHIEL, M., AND HOLAK, T. Solution structure of PCP, a prototype for the peptidyl carrier domains of modular peptide synthetases. *Structure 8* (2000), 407–418.

[54] WEINER, S., KOLLMAN, P., CASE, D., SINGH, U., GHIO, C., ALAGONA, G., PROFETA, S., AND WEINER, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc. 106* (1984), 765–784.

[55] WOJTCZAK, J., CODY, V., LUFT, J., AND PANGBORN, W. Structures of human transthyretin complexed with thyroxine at 2.0Å resolution and $3', 5'$-Dinitro-*N*-acetyl-L-thyronine at 2.2Å resolution. *Acta Cryst. D 52* (1996), 758–765.

[56] ZHOU, G., FERRER, M., CHOPRA, R., KAPOOR, T., STRASSMAIER, T., WEISSENHORN, W., SKEHEL, J., OPRIAN, D., SCHREIBER, S., HARRISON, S., AND WILEY, D. The structure of an HIV-1 specific cell entry inhibitor in complex with the HIV-1 gp41 trimeric core. *Bioorg. Med. Chem. 8* (2000), 2219–2228.

APPENDIX

In Appendix A, K^* is derived from first principles by examining the sum of chemical potentials at chemical equilibrium. Details of our structural model are provided in Appendix B. Appendix C presents the derivation of inter-mutation pruning (Eq. 10). A description of how a bound is computed on a conformation's minimum energy is presented in Appendix D. Finally, in Appendix E we provide details on the cloning, mutation, expression, and purification of our novel mutant proteins as well as details of the fluorescence quenching experiments.

A Detailed Derivation of K^*

K^* represents a biophysically-motivated scoring function over molecular ensembles. By using the Boltzmann probability distribution, K^* satisfies the Ergodic hypothesis and can be proved to approximate the true association (binding) constant K_A. If K^* is computed using exact partition functions then K^* will equal K_A. In practice, we sample conformation space, replacing the continuous integral with a discrete summation and use a molecular mechanics scoring function to compute the energy of each conformation. Hence, our algorithm represents an approximation to the true association (binding) constant.

We describe K^* for the protein-ligand binding reaction $P + L \rightleftharpoons PL$, where P represents the protein and L can represent either a small molecule in protein-ligand binding or a complete protein in the case of protein-protein binding. Our scoring method represents an approximation to the association constant, K_A, by

$$K^* = \frac{\sum_{b \in B} \exp(-E_b/RT)}{\sum_{l \in L} \exp(-E_l/RT) \sum_{f \in F} \exp(-E_f/RT)}, \qquad (11)$$

where B is the set of bound protein states, F is the set of unbound (free) protein states, L is the set of unbound ligand states, E_s is the energy of conformation s, R is the gas constant, and T is the temperature in Kelvin. We will now motivate this equation and describe our physically derived approximation. We first note that for the enzyme/ligand system the true association (binding) constant is defined as:

$$K_A = \frac{[PL]}{[P][L]}.$$

It is known in statistical mechanics [19, 32] that at chemical equilibrium the sum of the chemical potentials, μ, is equal to zero. In our ligand binding example,

$$\mu_P + \mu_L - \mu_{PL} = 0 \tag{12}$$

where μ_P, μ_L, and μ_{PL} are the chemical potentials for the free protein, free ligand, and protein-ligand complex respectively. The chemical potential, μ_J, for a species J of indistinguishable particles is

$$\mu_J = -kT \ln\left(\frac{q_J(V,T)}{N_J}\right), \tag{13}$$

where k is Boltzmann's constant and $q_J(V,T)$ is the partition function for the N_J molecules of species J at constant volume V and temperature T. The partition function includes all allowable states of a system. By substituting the chemical potentials Eq. (13) into the equilibrium condition Eq. (12) we obtain the result that at equilibrium:

$$\frac{q_{PL}(V,T)}{q_P(V,T)q_L(V,T)} = \frac{N_{PL}}{N_P N_L} = K_A. \tag{14}$$

Thus the association constant K_A is the quotient of the individual species partition functions. Unfortunately, it is not currently possible to compute exact partition functions for a complex molecular species. This would require integrating an exact energy function over a molecule's entire conformational space. We therefore approximate these partition functions with the use of rotamerically-based conformational ensembles. The partition function is approximated by our rotamerically-based conformations as:

$$q_{PL} = \sum_{b\in B} \exp(-E_b/RT), \quad q_P = \sum_{f\in F} \exp(-E_f/RT),$$
$$q_L = \sum_{l\in L} \exp(-E_l/RT), \tag{15}$$

where B, F, and L represent our rotamer based ensembles for the bound protein-ligand complex, the free protein, and the free ligand conformations respectively. When combined this gives the binding constant approximation K^* in Eq. (11).

The accuracy with which K^* approximates K_A is proportional to the accuracy of the partition function approximation used. There are two components to an ensemble-based approximation to the partition function: the set of conformations used in the ensemble, and the method used to score each conformation. While a molecular ensemble can be generated by multiple techniques (rotamers, multiple NMR structures, multiple crystal structures, molecular dynamics) [38, 30, 23], it is important that the ensemble sample all appropriate regions of conformation space. For example, an ensemble of structures generated from NMR experiments on an apo protein may not sample regions of protein conformation space that are compatible with ligand binding. Rotameric-based ensembles have the potential to sample this space more evenly. When ensembles containing a large number of conformations are used it is important to choose an energy function that can be computed efficiently.

B Details of the Structural Model

Our model consists of a portion of the GrsA-PheA protein (pdb: 1AMU [9]) including the active site and a shell of surrounding residues (termed the *steric shell*). The residues of the active site modeled as flexible using rotamers and subject to energy minimization include: 235D, 236A, 239W, 278T, 299I, 301A, 322A, 330I, and 331C. The steric shell was selected to include all residues not modeled as flexible and that contain at least one atom within 8Å of the active site. The steric shell residues include: 186Y, 188I, 190T, 210L, 213F, 214F, 230A, 234F, 237S, 238V, 240E, 243M, 279L, 300T, 302G, 303S, 320I, 321N, 323Y, 324G, 325P, 326T, 327E, 328T, 329T, 332A, 333T, 334T, 515N, and 517K. In addition to the active site flexible and steric shell residues the model also includes the substrate and AMP.

Flexible residues are represented by rotamers from the Lovell *et al.* rotamer library [30]. Each rotameric based conformation in B, F, and L is minimized by steepest descent minimization using the AMBER energy function (electrostatic, vdW, and dihedral energy terms) [54, 10] and is then combined using Eqs. (11, 15) above.

C Detailed Derivation of Inter-mutation Pruning

In Section 2.3.2 we described conditions under which a conformation could be pruned when computing a single partition function for a single mutation. When performing a mutation search, we can bootstrap the pruning condition for improved efficiency. As in Section 2.3.2, pruned conformations are not energy-minimized, thereby saving time in the overall mutation search. We will show how, in a mutation search, the ϵ-approximation pruning conditions derived below make use of the partition functions previously computed for other, different mutation sequences evaluated earlier in that search. Therefore, we call this pruning *inter-mutation pruning*. The intuition is that we assume a lower bound on the partition function q_n^* that allows us to prune more conformations earlier in the search. Inter-mutation pruning can only be applied during the computation of a bound partition function, q_{PL}; the unbound partition function q_P must be computed using the intra-mutation pruning method of Section 2.3.2.

During a mutation search, our primary goal is to compute a provably accurate ϵ-approximation for the top-ranking mutations while quickly computing K^* values for lower-ranked mutations that do not require the same degree of accuracy. As each mutation is examined in the mutation search, it suffices to compute an ϵ-approximation for only those mutation sequences with $K^* \geq \gamma \max_{i\leq m} K_i^*$, where K_i^* is the K^* value of the i^{th} mutation, m is the total number of mutations, and γ $(0 < \gamma \leq 1.0)$ is a user-specified constant defining a range of K^* values that must be computed accurately. Setting $\gamma = 1.0$ will compute an ϵ-approximation for only the best scoring K^* value. Setting $\gamma = 0.0$ will compute an ϵ-approximation for all K^* values. We typically set $\gamma = 0.01$ which causes the approximation algorithm to compute ϵ-approximations for all mutation sequences with K^* scores within two orders of magnitude of the best. Since the value of $\max_{i\leq m} K_i^*$ is not known during the mutation search, we compute ϵ-approximations for all $K^* \geq \gamma K_o^*$, where K_o^* is the largest (best) K^* value seen thus far in the mutation search. By definition, all values of K_o^* satisfy the inequality $K_o^* \leq \max_{i\leq m} K_i^*$ (in other words, all local maxima must be less than or equal to the global maximum). As a result, by computing an ϵ-approximation for all mutations with $K^* \geq \gamma K_o^*$ we will have computed ϵ-approximations for all mutations with $K^* \geq \gamma \max_{i\leq m} K_i^*$. When

$$K^*(1-\epsilon) \leq \tilde{K}^* \leq K^* \frac{1}{1-\epsilon}, \tag{16}$$

we say *$\tilde{K}^*$ is an ϵ-approximation to K^**. To prove that the computed $\tilde{K}^*$ is an ϵ-approximation to K^* we first show that inter-mutation pruning can compute an ϵ-approximation for q_{PL} and then combine the result with the intra-mutation pruning of Section 2.3.2. The following proof builds from the ideas used in the intra-mutation pruning of Section 2.3.2. Assume that we've computed q_P (Eq. 2) using intra-mutation pruning and now want to efficiently compute q_{PL} (Eq. 2). As stated in the previous paragraph, it is only necessary to compute q_{PL} accurately for mutation sequences with corresponding K^* values that are larger than our minimum accepting score (γK_o^*). That is, we require an ϵ-approximation to q_{PL} when

$$\frac{q_{PL}}{q_P q_L} > \frac{q'_{PL}}{q'_P q'_L}\gamma, \tag{17}$$

where q_{PL}, q_P, and q_L are the partition functions used to compute K^* and q'_{PL}, q'_P, and q'_L are the partition functions used to compute K_o^*. Since we are performing a mutation search to find good mutation sequences for a single ligand, we know $q'_L = q_L$. Therefore, it is only necessary to compute q_{PL} accurately when

$$q_{PL} > \frac{q'_{PL}}{q'_P}\gamma q_P. \tag{18}$$

For notational convenience, we define $K^\dagger = \frac{q_{PL}}{q_P}$ and $K_o^\dagger = \frac{q'_{PL}}{q'_P}$.

PROPOSITION 1. *The algorithm in Figure 6 computes an ϵ-approximation q^*_{PL} for a bound partition function, q_{PL}, when $q_{PL} > K_o^\dagger \gamma q_P$. If $q_{PL} \leq K_o^\dagger \gamma q_P$ then an $(\epsilon+\delta)$-approximation ($\delta \geq 0$) is computed.*

Proof: To prove Proposition 1 we consider two cases. Case 1, Equation (18) holds thus requiring an ϵ-approximation. Case 2, Equation (18) does not hold thus requiring only an $(\epsilon+\delta)$-approximation. We derive pruning criteria for Case 1. The pruning criteria will compute a correct ϵ-approximation for Case 1 and will compute an $(\epsilon+\delta)$-approximation ($\delta \geq 0$) for Case 2. We will show that the $(\epsilon+\delta)$-approximation holds only for partition functions falling into Case 2 and that these are situations for which we do not require an ϵ-approximation (see Proposition 2).

After computing an ϵ-approximation to the partition function, it must be the case that $q^*_n \geq (1-\epsilon)q_n$ which implies that $p_n \leq \epsilon q_n$. If we assume Eq. (18) holds (Case 1) then $q_n \geq K_o^\dagger \gamma q_P$ and we can conservatively conclude that $p_n \leq \epsilon q_n$ if

$$p_n \leq \epsilon K_o^\dagger \gamma q_P. \tag{19}$$

In reality, p_n can be as large as ϵq_n but during the conformation search we don't yet know the value of q_n. Therefore, given p_k, we can prune conformation, c_{k+1} if p_{k+1} remains less than $\epsilon K_o^\dagger \gamma q_P$ thereby satisfying Eq. (19), i.e.,

$$p_{k+1} = p_k + \exp\left(-B(c_{k+1})/RT\right) \leq \epsilon K_o^\dagger \gamma q_P. \tag{20}$$

If we solve for $B(c_{k+1})$ then the pruning criterion becomes

$$B(c_{k+1}) \geq -RT \ln\left(\epsilon K_o^\dagger \gamma q_P - p_k\right). \tag{21}$$

Because $p_k \leq p_k^*$, Eq. (21) can be rewritten as

$$B(c_{k+1}) \geq -RT \ln\left(\epsilon K_o^\dagger \gamma q_P - p_k^*\right). \tag{22}$$

Eq. (22) is the same as Eq. (9) if $(\epsilon K_o^\dagger \gamma q_P)$ is substituted for $(q_k^* \epsilon)$. Therefore, when computing q_{PL} during a mutation search we use the pruning criterion

$$B(c_{k+1}) \geq -RT \ln\left(\psi_k - p_k^*\right), \tag{23}$$

Let $n \leftarrow$ Number of Rotameric Conformations
Let $c \leftarrow$ Rotameric Conformations
Initialize: $\psi \leftarrow \epsilon K_o^\dagger \gamma q_P,\ q^* \leftarrow 0,\ p^* \leftarrow 0$
for $k = 1$ to n
  $\psi \leftarrow \texttt{Max}\left(\psi, q^* \epsilon\right)$
  if $\text{B}(c_k) \leq -RT \ln\left(\psi - p^*\right)$
    $q^* \leftarrow q^* + \exp\left(-\texttt{ComputeMinEnergy}(c_k)/RT\right)$
  else
    $p^* \leftarrow p^* + \exp\left(-\text{B}(c_k)/RT\right)$
Return q^*

Figure 6: INTER-MUTATION PRUNING **used in computing the bound partition function q_{PL}. The values q^* and p^* and the functions B$(\cdot)$ and** ComputeMinEnergy$(\cdot)$ **are as described in Figure 3. For all mutations with $q_{PL} \geq K_o^\dagger \gamma q_P$, the computed q^* will represent an ϵ-approximation to the true partition function q such that $q^* \geq (1-\epsilon)q$.**

where $\psi_k = \max\left(\epsilon K_o^\dagger \gamma q_P, q_k^* \epsilon\right)$ and p_k^* is an upper bound on the partition function of the pruned conformations $(C_k - S_k)$ as described in Section 2.3.2.

LEMMA 3. *When Eq. (18) holds, any conformation c_{k+1} that satisfies Eq. (23) can be pruned during computation of the bound partition function while maintaining the invariant that q_k^* is an ϵ-approximation to q_k (where q_k is q_{PL} computed through conformation k and q_k^* is q^*_{PL} computed through conformation k).*

By maintaining the invariant throughout computation of the bound partition function the following lemma holds:

LEMMA 4. *When a bound partition function is computed for a mutation sequence satisfying Eq. (18) while maintaining the invariant that q_k^* is an ϵ-approximation to q_k for all k $(0 < k \leq n)$ then, by induction, at the end of the computation, q_n^* will be an ϵ-approximation to q_n.*

The inter-mutation pruning algorithm is shown in Figure 6.

We have shown that when $q_{PL} > K_o^\dagger \gamma q_P$ (Case 1) pruning using Eq. (23) will produce an ϵ-approximation q^*_{PL}. When $q_{PL} \leq K_o^\dagger \gamma q_P$ it is possible that Eq. (23) will prune the wrong conformations resulting in $q^*_{PL} < (1-\epsilon)q_{PL}$. Thus, the computed q^*_{PL} will be too small and we will not have computed an ϵ-approximation. However, by definition q_{PL} will have been from a mutation whose corresponding $K^\dagger \leq K_o^\dagger \gamma$ $(K^* \leq K_o^* \gamma)$ (Case 2) and thus (by Proposition 1) it was not necessary to compute an ϵ-approximation. These partition functions are easily identified by their magnitude. This completes the proof of Proposition 1. □

We now have the necessary tools to prove Proposition 2.

PROPOSITION 2. *An ϵ-approximation, $\tilde{K}^*$, is computed for K^* when $K^* > \gamma K_o^*$.*

Proof: We start by determining bounds on the computed K^* for Case 1 and Case 2. In both cases the unbound partition function q_P^* was computed according to the intra-mutation pruning of Section 2.3.2 and the bound partition function q^*_{PL} was computed according to the inter-mutation pruning described above in this section. Note that because each term in the summation of each partition function (Eq. 2) is positive, all approximations q_n^* computed by omitting terms must be less than q_n. Thus for both intra- and inter-mutation pruning, when q_n^* is an ϵ-approximation to q_n we know that $q_n \geq q_n^* > (1-\epsilon)q_n$. Therefore,

$$\begin{aligned} & q_P \geq q_P^* > (1-\epsilon)q_P \qquad \text{and} \\ \text{Case 1)} \quad & q_{PL} \geq q^*_{PL} > (1-\epsilon)q_{PL} \\ \text{Case 2)} \quad & q_{PL} \geq q^*_{PL} > 0. \end{aligned}$$

The resulting $\tilde{K}^\dagger$ (the approximation to $K^\dagger$) for high-scoring mutations (Case 1) will fall within the range

$$\left[K^\dagger(1-\epsilon), K^\dagger \frac{1}{1-\epsilon}\right],$$

which implies that $\tilde{K}^*$ (the approximation to K^*) lies in the range

$$\left[K^*(1-\epsilon), K^* \frac{1}{1-\epsilon}\right],$$

as desired. For lower-scoring mutations (i.e., mutations with $K^* \leq \gamma K_o^*$) (Case 2), the resulting $\tilde{K}^\dagger$ will fall within the range

$$\left[0, K^\dagger \frac{1}{1-\epsilon}\right],$$

which implies that $\tilde{K}^*$ lies in the range

$$\left[0, K^* \frac{1}{1-\epsilon}\right].$$

Therefore an ϵ-approximation, $\tilde{K}^*$ is computed for K^* when $K^* > \gamma K_o^*$ and an $(\epsilon+\delta)$-approximation ($\delta \geq 0$) is computed when $K^* \leq \gamma K_o^*$. This completes the proof of Proposition 2. □

D Computing a Bound on a Conformation's Minimum Energy

In our structural model, we treat some residues as rigid while others have a rigid backbone but flexible side-chains. If we let R represent the set of all rigid atoms (all atoms of the steric shell and backbone atoms of the flexible residues) then for a system with k flexible residues the energy of this system can be computed as

$$E_0 = A_{00} + \sum_{i \leq k} A_{0i} + \sum_{i \leq k} A_{i0} + \sum_{i \leq k} \sum_{i < j \leq k} A_{ij}, \qquad (24)$$

where A is a precomputed residue-indexed pairwise *energy matrix*. A residue-indexed pairwise energy matrix is a $(k+1) \times (k+1)$ matrix composed of energy terms describing each residue's interaction energy with the backbone, itself, and all other residues. A detailed description of the matrix elements is as follows: A_{00} is the sum of all energy terms exclusively involving atoms in R, A_{0i} is the sum of all energy terms involving at least one atom of residue i and at least one atom from set R, A_{i0} is the sum of all energy terms involving only atoms of residue i (intra-residue energy), and A_{ij} is the sum of all energy terms involving at least one atom from each of residues i and j.

Similarly, one can define the energy E_m of a minimized conformation, using a different matrix M of pairwise energy terms evaluated on the energy minimized conformation:

$$E_m = M_{00} + \sum_{i \leq k} M_{0i} + \sum_{i \leq k} M_{i0} + \sum_{i \leq k} \sum_{i < j \leq k} M_{ij}, \qquad (25)$$

where M_{00}, M_{0i}, M_{i0}, and M_{ij} are the analogues of A_{00}, A_{0i}, A_{i0}, and A_{ij} except that they are computed based on the positions of the atoms in the energy minimized structure.

During both intra- and inter-mutation pruning (see Section 2.3.2 and Appendix C) a lower bound on the energy of the energy-minimized conformation is required. If we replace M_{00}, M_{0i}, M_{i0}, and M_{ij} in Eq. (25) with lower bounds D_{00}, D_{0i}, D_{i0}, and D_{ij} such that $D_{00} \leq M_{00}$, $D_{i0} \leq M_{i0}$, $D_{0i} \leq M_{0i}$, and $D_{ij} \leq M_{ij}$, then we can compute a bound:

$$E_b = D_{00} + \sum_{i \leq k} D_{0i} + \sum_{i \leq k} D_{i0} + \sum_{i \leq k} \sum_{i < j \leq k} D_{ij}, \qquad (26)$$

such that $E_b \leq E_m$.

Because E_b is simply the sum of $O(k^2)$ pairwise energy terms, if a precomputed residue-indexed lower-bound pairwise energy matrix is available then E_b can be computed in time $O(k^2)$. The use of a precomputed residue-based pairwise energy matrix thus avoids the costly computation of $O(a^2)$ energy terms, where a is the total number of atoms in the system and $k \ll a$.

In our implementation of the K^* algorithm, we precompute a residue-indexed lower-bound pairwise energy matrix V over all rotamers for each residue. This matrix contains a lower bound (V_{ij}) on the energy for all allowed pairs of rotamers as well as lower bounds V_{00}, V_{0i}, and V_{i0} on the shell-shell, shell-residue, and residue-self energies (respectively as described above). The matrix we use in computation is thus slightly different than those described in Eqs. (24, 25, 26). The matrix V contains lower-bound energy terms for all rotamers for all residues. Therefore, the matrix has size $(km+1) \times (km+1)$ where k is the number of flexibly-modeled residues and m is the number of allowed rotamers (spanning all residue types). When computing an energy bound, terms corresponding to the currently assigned rotamers are used in a manner similar to those described for Eqs. (24, 25, 26).

To prevent one rotamer from minimizing into another, the maximum dihedral movement allowed during energy minimization is bounded (as described in *Step 2* of Section 2.2). As a result, one can easily compute the terms of matrix V by examining all pairs of residues in their active site specified relative orientations. The lower-bound energy matrices are precomputed before K^* evaluation or a mutation search is performed.

E Cloning, Mutation, Expression, Purification, and Fluorescence

Cloning. GrsA-PheA was cloned from GrsA by PCR as described previously [48]. PCR reactions were performed using PfuTurbo DNA Polymerase (Stratagene) per manufacturer's directions. PCR products and pQE-60 were digested with 10U NcoI and 10U BamHI for 2 hrs at 37°C (50mM Tris-HCl pH 8.0, 10mM $MgCl_2$, 100mM NaCl). PCR products and linearized vector were gel purified and recovered using the QIAquick gel extraction kit (Qiagen), and ligated (2U T4 ligase, 50mM Tris-HCl pH 7.6, 10mM $MgCl_2$, 1mM ATP, 1mM DTT, 5% (w/v) PEG-8000, 1 hr at 24°C). *Escherichia coli* M15(pRep4) were transformed with the PheA pQE-60 construct and selected by growth on Luria Broth (LB) supplemented with 50 μg/ml ampicillin and 30 μg/ml kanamycin.

Mutation. Mutations were introduced by site directed mutagenesis using the QuikChange Site-Directed Mutagenesis Kit (Qiagen). Protocols were carried out per manufacturers instructions with the following primers: 301G: `CGTTAATTACAGGAGGCTCAGCTACC` ($T_m = 72°C$) 330F: `CCTACGGAAACAACTTTTGTGCGACTACA TGG` ($T_m = 76°C$) 330W: `GGCCCTACGGAAACAACTTGGTGTGC GACTACATGG` ($T_m = 77°C$).

Expression. 1 Liter of LB was inoculated and grown at 37°C until OD_{600} = 1.5-1.8. IPTG was added (1.5mM) and cultures grown for an additional 3 hr. Cells were harvested by centrifugation at 2,500g for 40 min, resuspended in 20 mL LB, pelleted by centrifugation at 2,000g for 30 min, and frozen at -80°C.

Purification. Cells were thawed, resuspended in buffer (20mM Tris-HCl pH 7.4, 50mM NaCl, 50μM PMSF) and lysed by sonication. Cell extract was clarified by centrifugation at 40,000g for 40 min and 0.45μm syringe-driven filtration. PheA-His_6 was purified by Ni^{2+} affinity chromatography using a gradient of 0-100mM imidazole. Pure PheA-His_6 was dialyzed in 50mM Hepes pH 7.6.

Fluorescence. Each protein solution (50mM Hepes pH 7.6) was supplemented with 2mM dithiothreitol (DTT), 100mM NaCl, and 10mM $MgCl_2$. The excitation wavelength was 280nm. Substrate was titrated into the protein solution and fluorescence quenching was measured at 340nm.