A Novel Minimized Dead-End Elimination Criterion and Its Application to Protein Redesign in a Hybrid Scoring and Search Algorithm for Computing Partition Functions over Molecular Ensembles

Ivelin Georgiev[1,⋆], Ryan H. Lilien[1,2,3,⋆], and Bruce R. Donald[1,3,4,5,⋆⋆]

[1] Dartmouth Computer Science Department, Hanover, NH 03755, USA
[2] Dartmouth Medical School, Hanover, NH
[3] Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH
[4] Dartmouth Department of Chemistry, Hanover, NH
[5] Dartmouth Department of Biological Sciences, Hanover, NH
Bruce.R.Donald@dartmouth.edu

Abstract. Novel molecular function can be achieved by redesigning an enzyme's active site so that it will perform its chemical reaction on a novel substrate. One of the main challenges for protein redesign is the efficient evaluation of a combinatorial number of candidate structures. The modeling of protein flexibility, typically by using a rotamer library of commonly-observed low-energy side-chain conformations, further increases the complexity of the redesign problem. A dominant algorithm for protein redesign is Dead-End Elimination (DEE), which prunes the majority of candidate conformations by eliminating *rigid* rotamers that provably are not part of the Global Minimum Energy Conformation (GMEC). The identified GMEC consists of rigid rotamers that have not been energy-minimized and is referred to as the *rigid-GMEC*. As a post-processing step, the conformations that survive DEE may be energy-minimized. When energy minimization is performed after pruning with DEE, the combined protein design process becomes heuristic, and is no longer provably accurate: That is, the rigid-GMEC and the conformation with the lowest energy among all energy-minimized conformations (the *minimized-GMEC*, or *minGMEC*) are likely to be different. While the traditional DEE algorithm succeeds in not pruning rotamers that are part of the rigid-GMEC, it makes no guarantees regarding the identification of the minGMEC. In this paper we derive a novel, provable, and efficient DEE-like algorithm, called *minimized-DEE* (*MinDEE*), that guarantees that rotamers belonging to the minGMEC will not be pruned, while still pruning a combinatorial number of conformations. We show that MinDEE is useful not only in identifying the minGMEC, but also as a filter in an ensemble-based scoring and search algorithm for protein redesign that exploits energy-minimized conformations. We compare our results both to our previous computational predictions of protein designs and to biological activity assays of predicted protein mutants. Our provable and efficient minimized-DEE algorithm is applicable in protein redesign, protein-ligand binding prediction, and computer-aided drug design.

⋆ These authors contributed equally to the work.

⋆⋆ This work is supported by grants to B.R.D. from the National Institutes of Health (R01 GM-65982), and the National Science Foundation (EIA-0305444).

A. Apostolico et al. (Eds.): RECOMB 2006, LNBI 3909, pp. 530–545, 2006.

1 Introduction

Computational Protein Design. The ability to engineer proteins has many biomedical applications. Novel molecular function can be achieved by redesigning an enzyme's active site so that it will perform its chemical reaction on a novel substrate. A number of computational approaches to the protein redesign problem have been reported. To improve the accuracy of the redesign, protein flexibility has been incorporated into most previous structure-based algorithms for protein redesign [30, 14, 13, 12, 1, 18, 15]. A study of bound and unbound structures found that most structural changes involve only a small number of residues and that these changes are primarily side-chains, and not backbone [22]. Hence, many protein redesign algorithms use a rigid backbone and model side-chain flexibility with a rotamer library, containing a discrete set of low-energy commonly-observed side-chain conformations [19, 25]. The major challenge for redesign algorithms is the efficient evaluation of the exponential number of candidate conformations, resulting not only from mutating residues along the peptide chain, but also by employing rotamer libraries. The development of pruning conditions capable of eliminating the majority of mutation sequences and conformations in the early, and less costly, redesign stages has been crucial.

Non-ensemble-based algorithms for protein redesign are based on the assumption that protein folding and binding can be accurately predicted by examining the GMEC. Since identifying the GMEC using a model with a rigid backbone, a rotamer library, and a pairwise energy function is known to be NP-hard [24], different heuristic approaches (random sampling, neural network, and genetic algorithm) have been proposed [30, 14, 13, 12, 20]. A provable and efficient deterministic algorithm, which has become the dominant choice for non-ensemble-based protein design, is Dead-End Elimination (DEE) [6]. DEE reduces the size of the conformational search space by eliminating *rigid* rotamers that provably are not part of the GMEC. Most important, since no protein conformation containing a dead-ending rotamer is generated, DEE provides a combinatorial factor reduction in computational complexity.

When energy minimization is performed after pruning with DEE, the process becomes heuristic, and is no longer provably accurate: a conformation that is pruned using rigid-rotamer energies may subsequently minimize to a structure with lower energy than the rigid-GMEC. Therefore, the traditional DEE conditions are not valid for pruning rotamers when searching for the lowest-energy conformation among all energy-minimized rotameric conformations (the *minimized-GMEC*, or *minGMEC*).

NRPS Redesign and K^*. Traditional ribosomal peptide synthesis is complemented by non-ribosomal peptide synthetase (NRPS) enzymes in some bacteria and fungi. NRPS enzymes consist of several domains, each of which has a separate function. Substrate specificity is generally determined by the adenylation (A) domain [28, 3, 27]. Among the products of NRPS enzymes are natural antibiotics (penicillin, vancomycin), antifungals, antivirals, immunosuppressants, and antineoplastics. The redesign of NRPS enzymes can lead to the synthesis of novel NRPS products, such as new libraries of antibiotics [2]. The main techniques for NRPS enzyme redesign are *domain-swapping* [29, 26, 7, 21], *signature sequences* [28, 8, 3], and *active site manipulation from a structure-based mutation search utilizing ensemble docking* (the K^* method [17]).

The K^* algorithm [17] has been demonstrated for NRPS redesign, but is a general algorithm that is, in principle, capable of redesigning any protein. K^* is an ensemble-based scoring technique that uses a Boltzmann distribution to compute partition functions for the bound and unbound states of a protein. The ratio of the bound to the unbound partition function is used to compute a provably-good approximation (K^*) to the binding constant for a given sequence. A volume and a steric filter are applied in the initial stages of a redesign search to prune the majority of the conformations from more expensive evaluation. The number of evaluated conformations is further reduced by a provable ε-approximation algorithm. Protein flexibility is modeled for *both* the protein *and* the ligand using energy-minimization and rotamers [17].

Contributions of the Paper. Boltzmann probability implies that low-energy conformations are more likely to be assumed than high-energy conformations. The motivation behind energy minimization is therefore well-established and algorithms that incorporate energy minimization often lead to more accurate results. However, if energy minimization is performed *after* pruning with DEE, then the combined protein design process is heuristic, and not provable. We show that a conformation pruned using rigid-rotamer energies may subsequently minimize to surpass the putative rigid-GMEC.

We derive a novel, provable, and efficient DEE-like algorithm, called *minimized-DEE* (*MinDEE*), that guarantees that no rotamers belonging to the minGMEC will be pruned. We show that our method is useful not only in (a) identifying the minGMEC (a non-ensemble-based method), but also (b) as a filter in an ensemble-based scoring and search algorithm for protein redesign that exploits energy-minimized conformations. We achieve (a) by implementing a MinDEE/A^* algorithm in a search to switch the binding affinity of the Phe-specific adenylation domain of the NRPS Gramicidin Synthetase A (GrsA-PheA) towards Leu. The latter goal (b) is achieved by implementing MinDEE as a combinatorial filter in a hybrid algorithm,[1] combining A^* search and our previous work on K^* [17]. The experimental results, based on a 2-point mutation search on the 9-residue active site of the GrsA-PheA enzyme, confirm that the new Hybrid MinDEE-K^* algorithm has a much higher pruning efficiency than the original K^* algorithm. Moreover, it takes only *30 seconds* for MinDEE to determine which rotamers can be provably pruned. We make the following contributions in this paper:

1. Derivation of MinDEE, a novel, provable, and efficient DEE-like algorithm that incorporates energy minimization, with applications in both non-ensemble- and ensemble-based protein design.
2. Introduction of a MinDEE/A^* algorithm that identifies the minGMEC and returns a set of low-energy conformations;
3. Introduction of a hybrid MinDEE-K^* ensemble-based scoring and search algorithm, improving on our previous work on K^* [17] by replacing a constant-factor with a combinatorial-factor provable pruning condition; and
4. The use of our novel algorithms in a redesign mutation search for switching the substrate specificity of the NRPS enzyme GrsA-PheA; we compare our results to previous computational predictions of protein designs and to biological activity assays of predicted protein mutants.

[1] For brevity, we will henceforth refer to this algorithm as the *Hybrid MinDEE-K^** algorithm.

2 Derivation of the Minimized-DEE Criterion

2.1 The Original DEE Criterion

In this section we briefly review the *traditional-DEE* theorem [6, 23, 11]. Traditional-DEE refers to the original DEE, which is not provably correct when used in a search for the minGMEC. The total energy, E_T, of a given rotameric-based conformation can be written as $E_T = E_{t'} + \sum_i E(i_r) + \sum_i \sum_{j>i} E(i_r, j_s)$, where $E_{t'}$ is the template self-energy (i.e., backbone energies or energies of rigid regions of the protein not subject to rotamer-based modeling), i_r denotes rotamer r at position i, $E(i_r)$ is the self energy of rotamer i_r (the intra-residue and residue-to-template energies), and $E(i_r, j_s)$ is the non-bonded pairwise interaction energy between rotamers i_r and j_s. The rotamers assumed in the rigid-GMEC are written with a subscript g. Therefore i_g is the rotamer assumed in the rigid-GMEC at position i. The following two bounds are then noted: for all i, j $(i \neq j)$, $\max_{s \in R_j} E(i_t, j_s) \geq E(i_t, j_g)$ and $\min_{s \in R_j} E(i_g, j_s) \leq E(i_g, j_g)$, where R_j is the set of allowed rotamers for residue j. For clarity, we will not include R_j in the limits of the $\max$ and $\min$ terms, since it will be clear from the notation from which set s must be drawn. The DEE criterion for rotamer i_r is defined as:

$$E(i_r) + \sum_{j \neq i} \min_s E(i_r, j_s) > E(i_t) + \sum_{j \neq i} \max_s E(i_t, j_s). \tag{1}$$

Any rotamer i_r satisfying the DEE criterion (Eq. 1) is provably not part of the rigid-GMEC $(i_r \neq i_g)$, and is considered 'dead-ending.' Extensions to this initial DEE criterion allow for additional pruning while maintaining correctness with respect to identifying the rigid-GMEC [6, 10, 11, 23].

2.2 DEE with Energy Minimization: MinDEE

We now derive generalized DEE pruning conditions which can be used when searching for the minGMEC. The fundamental difference between traditional-DEE and MinDEE

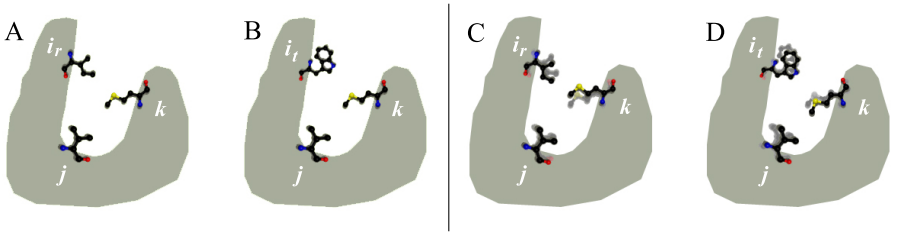

Fig. 1. Energy-Minimized DEE. Without energy minimization the swapping of rotamer i_r for i_t (A to B) leaves unchanged the conformations and self and pairwise energies of residues j and k. When energy minimization is allowed, the swapping of rotamer i_r for rotamer i_t (C to D) may cause the conformations of residues j and k to minimize (i.e., move) to form more energetically favorable interactions (from the faded to the solid conformations in C and D).

is that the former enjoys significant independence among multiple energy terms during a rotamer swap (Fig. 1). Therefore, to be provably correct, MinDEE must account for a range of possible energies. The conformation of a residue may change during energy minimization, however we constrain this movement to a region of conformation space called a *voxel* to keep one rotamer from minimizing into another. In this framework, the voxel $\mathcal{V}(i_r)$ for rotamer i_r is simply all conformations of residue i within a $\pm\theta$ range around each rotamer dihedral when starting from the rotamer[2] i_r. We similarly define the voxel $\mathcal{V}(i_r, j_s)$ for the pair of rotamers i_r and j_s to be the region of conformation space $\mathcal{V}(i_r) \times \mathcal{V}(j_s)$. Next, we can define the *maximum*, *minimum*, and *range* of voxel energies: $E_{\oplus}(i_r) = \max_{z \in \mathcal{V}(i_r)} E(z)$; $E_{\ominus}(i_r) = \min_{z \in \mathcal{V}(i_r)} E(z)$; and $E_{\oslash}(i_r) = E_{\oplus}(i_r) - E_{\ominus}(i_r)$. Analogous definitions exist for pairwise terms (Fig. 4 in [9, p. 25]). For a given protein, we define a *rotamer vector* $A = (A_1, A_2, \cdots, A_n)$ to specify the rotamer at each of the n residue positions; $A_i = r$ when rotamer r is assumed by residue i. We then define the *conformation vector* $A^\bullet = (A_1^\bullet, A_2^\bullet, \cdots, A_n^\bullet)$ such that $A_i^\bullet$ is the conformation of residue i in the voxel-constrained minimized conformation, i.e., $A_i^\bullet \in \mathcal{V}(A_i)$ and $A^\bullet = (A_1^\bullet, A_2^\bullet, \cdots, A_n^\bullet) = \operatorname{argmin}_{B=(B_1,B_2,\cdots,B_n) \in \prod_{i=1}^{n} \mathcal{V}(A_i)} E(B)$, where $E(B)$ is the energy of the system specified by conformation vector B. For the energy-minimized conformation starting from rotamer vector A, we define the self-energy of rotamer i_r as $E_{\odot}(i_r|A) = E(A_i^\bullet)$ and the pairwise interaction energy of the rotamer pair i_r, j_s as $E_{\odot}(i_r, j_s|A) = E(A_i^\bullet, A_j^\bullet)$ where $E(A_i^\bullet)$ is the self-energy of residue i in conformation $A_i^\bullet$ and $E(A_i^\bullet, A_j^\bullet)$ is the pairwise energy between residues i and j in conformations $A_i^\bullet$ and $A_j^\bullet$. We can then express the minimized energy of A, $E_{\scriptscriptstyle T}(A)$ as: $E_{\scriptscriptstyle T}(A) = E_{t'} + \sum_i E_{\odot}(i_r|A) + \sum_i \sum_{j>i} E_{\odot}(i_r, j_s|A)$. Let G represent the rotamer vector that minimizes into the minGMEC and $E_{\scriptscriptstyle T}(G)$ be the energy of the minGMEC. Let $G_{i_g \to i_t}$ be the rotamer vector G where rotamer i_g is replaced with i_t. We know that $E_{\scriptscriptstyle T}(G_{i_g \to i_t}) \geq E_{\scriptscriptstyle T}(G)$, so we can pull residue i out of the two summations, obtaining:

$$\begin{aligned} E_{t'} + E_{\odot}(i_t|G_{i_g \to i_t}) + \sum_{j \neq i} E_{\odot}(i_t, j_g|G_{i_g \to i_t}) + \sum_{j \neq i} E_{\odot}(j_g|G_{i_g \to i_t}) \\ + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\odot}(j_g, k_g|G_{i_g \to i_t}) \geq E_{t'} + E_{\odot}(i_g|G) \\ + \sum_{j \neq i} E_{\odot}(i_g, j_g|G) + \sum_{j \neq i} E_{\odot}(j_g|G) + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\odot}(j_g, k_g|G). \end{aligned} \tag{2}$$

The $E_{t'}$ terms (Sec. 2.1) are independent of rotamer choice, are equal, and can be canceled. We make the following trivial upper and lower-bound observations:

$$E_{\odot}(i_t|A) \leq E_{\oplus}(i_t);\ E_{\odot}(i_t, j_g|A) \leq \max_{s \in R_j} E_{\oplus}(i_t, j_s); \tag{3}$$

$$E_{\odot}(j_g|A) \leq E_{\oplus}(j_g);\ E_{\odot}(j_g, k_g|A) \leq E_{\oplus}(j_g, k_g); \tag{4}$$

$$E_{\ominus}(i_g) \leq E_{\odot}(i_g|A);\ \min_{s \in R_j} E_{\ominus}(i_g, j_s) \leq E_{\odot}(i_g, j_g|A); \tag{5}$$

$$E_{\ominus}(j_g) \leq E_{\odot}(j_g|A);\ E_{\ominus}(j_g, k_g) \leq E_{\odot}(j_g, k_g|A). \tag{6}$$

[2] The voxel for each rotamer can be multi-dimensional, depending on the number of dihedrals.

Substituting Eqs. (3-6) into Eq. (2), we obtain:

$$E_{\oplus}(i_t) + \sum_{j \neq i} \max_s E_{\oplus}(i_t, j_s) + \sum_{j \neq i} E_{\oplus}(j_g) + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\oplus}(j_g, k_g) \geq$$
$$E_{\ominus}(i_g) + \sum_{j \neq i} \min_s E_{\ominus}(i_g, j_s) + \sum_{j \neq i} E_{\ominus}(j_g) + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\ominus}(j_g, k_g). \quad (7)$$

We now define the MinDEE criterion for rotamer i_r to be:

$$E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s) - \sum_{j \neq i} \max_s E_{\oslash}(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_{\oslash}(j_s, k_u) >$$
$$E_{\oplus}(i_t) + \sum_{j \neq i} \max_s E_{\oplus}(i_t, j_s). \quad (8)$$

Proposition 1. *When Eq. (8) holds, rotamer i_r is provably not part of the minGMEC.*

Proof. When Eq. (8) holds, we can substitute the left-hand side of Eq. (8) for the first two terms of Eq. (7), and simplify the resulting equation to:

$$E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s) - \sum_{j \neq i} \max_s E_{\oslash}(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_{\oslash}(j_s, k_u)$$
$$+ \sum_{j \neq i} E_{\oslash}(j_g) + \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\oslash}(j_g, k_g) > E_{\ominus}(i_g) + \sum_{j \neq i} \min_s E_{\ominus}(i_g, j_s). \quad (9)$$

We then substitute the following two bounds $\sum_{j \neq i} \max_s E_{\oslash}(j_s) \geq \sum_{j \neq i} E_{\oslash}(j_g)$ and $\sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_{\oslash}(j_s, k_u) \geq \sum_{j \neq i} \sum_{k \neq i, k > j} E_{\oslash}(j_g, k_g)$ into Eq. (9) and reduce: $E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s) > E_{\ominus}(i_g) + \sum_{j \neq i} \min_s E_{\ominus}(i_g, j_s)$. Thus, when the MinDEE pruning condition Eq. (8) holds, $i_r \neq i_g$ and we can provably eliminate rotamer i_r as not being part of the minGMEC. □

MinDEE (unlike traditional-DEE) accounts for energy changes during minimization (addends 3-4 in Eq. 8). Using precomputed energy bounds, the MinDEE condition (Eq. 8) can be computed as efficiently as the traditional-DEE condition (Eq. 1). Here we presented a generalization of traditional-DEE, to obtain an initial MinDEE pruning criterion. Analogously to the traditional-DEE extensions [6, 10, 11, 23], we also derived extensions to MinDEE to improve its pruning efficiency (see Appendix A in [9]).

3 Minimized-DEE/A^* Search Algorithm (Non-ensemble-Based Redesign)

3.1 Traditional-DEE with A^*

In [16], an A^* branch and bound algorithm was developed to compute a number of low-energy conformations for a single mutation sequence (i.e., a single protein). In this algorithm, traditional-DEE was first used to reduce the number of side-chain

conformations, and then surviving conformations were enumerated in order of conformation energy by expanding sorted nodes of a conformation tree.[3] The following derivation of the DEE/A^* combined search closely follows [16]. The A^* algorithm scores each node in a conformation tree using a scoring function $f = g + h$, where g is the cost of the path from the root to that node (the energy of all self and pairwise terms assigned through depth d) and h is an estimate (lower bound) of the path cost to a leaf node (a lower bound on the sum of energy terms involving unassigned residues). The value of g (at depth d) can be expressed as $g = \sum_{i=1}^{d}(E(i_r) + \sum_{j=i+1}^{d} E(i_r, j_s))$. The lower bound h can be written as $h = \sum_{j=d+1}^{n} E_j$, where n is the total number of flexible residues and $E_j = \min_s(E(j_s) + \sum_{i=1}^{d} E(i_r, j_s) + \sum_{k>j}^{n} \min_u E(j_s, k_u))$. The A^* algorithm maintains a list of nodes (sorted by f) and in each iteration replaces the node with the smallest f value by an expansion of the children of that node, until the node with the smallest f value is a leaf node, corresponding to a fully-assigned conformation. To reduce the branching factor of the conformation tree, the DEE algorithm is used to preprocess the set of allowed rotamers. If low-energy conformations within E_w of the GMEC are to be returned by the DEE/A* search, then the DEE criterion (Eq.1) must be modified to only eliminate rotamers that are provably not part of any conformation within E_w of the GMEC: $E(i_r) - E(i_t) + \sum_{j \neq i} \min_s E(i_r, j_s) - \sum_{j \neq i} \max_s E(i_t, j_s) > E_w$.

3.2 MinDEE with A^*

The traditional-DEE/A^* algorithm [16] can be extended to include energy minimization by substituting MinDEE for traditional-DEE. So that no conformations within E_w of the minGMEC are pruned, the MinDEE equation (Eq. 8) becomes:

$$E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s) - \sum_{j \neq i} \max_s E_{\oslash}(j_s) - \sum_{j \neq i} \sum_{k \neq i, k > j} \max_{s,u} E_{\oslash}(j_s, k_u) - E_{\oplus}(i_t) - \sum_{j \neq i} \max_s E_{\oplus}(i_t, j_s) > E_w \; . \tag{10}$$

We modify the definition of the A^* functions g and h to use the minimum energy terms $E_{\ominus}(\cdot)$: $g = \sum_{i=1}^{d}(E_{\ominus}(i_r) + \sum_{j=i+1}^{d} E_{\ominus}(i_r, j_s))$, and $h = \sum_{j=d+1}^{n} E_j$, where $E_j = \min_s \left(E_{\ominus}(j_s) + \sum_{i=1}^{d} E_{\ominus}(i_r, j_s) + \sum_{k>j}^{n} \min_u E_{\ominus}(j_s, k_u)\right)$. A lower bound on the minimized energy of the partially-assigned conformation is given by g, while a lower bound on the minimized energy for the unassigned portion of the conformation is given by h. Thus, the MinDEE/A^* search generates conformations in order of increasing *lower bounds* on the conformation's *minimized* energy.

We combine our modified MinDEE criterion (Eq. 10) with the modified A^* functions g and h above in a provable search algorithm for identifying the minGMEC and obtaining a set of low-energy conformations. First, MinDEE prunes the majority of the conformations by eliminating rotamers that are provably not within E_w of the minGMEC. The remaining conformations are then generated in order of increasing *lower bounds*

[3] In a 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 and each leaf node represents a fully-assigned conformation (see Fig. 3 in [17, p. 745]).

on their minimized energies. The generated conformations are energy-minimized and ranked in terms of increasing *actual* minimized energies.

MinDEE/A^* must guarantee that upon completion all conformations within E_w of the minGMEC are returned. However, the minGMEC may not be among the top A^* conformations if the lower bound on its energy does not rank high. We therefore derive the following condition for halting the MinDEE/A^* search. Let $B(s)$ be a lower bound on the energy of conformation s (see Appendix B in [9], which describes how lower energy bounds are precomputed for all rotamer pairs) and let E_m be the current minimum energy among the minimized conformations returned so far in the A^* search.

Proposition 2. *The MinDEE/A^* search can be halted once the lower bound $B(c)$ on the energy of the next conformation c returned by A^*, satisfies $B(c) > E_m + E_w$. The set of returned conformations is guaranteed to contain every conformation whose energy is within E_w of the energy of the minGMEC. Moreover, at that point in the search, the conformation with energy E_m is the minGMEC.*

The proof of Proposition 2 can be found in [9, Sec. 3.2]. Using both MinDEE and A^*, our algorithm obtains a combinatorial pruning factor by eliminating the majority of the conformations, which makes the search for the minGMEC computationally feasible. MinDEE/A^* incorporates energy minimization with provable guarantees, and is thus more capable of returning conformations with lower energy states than traditional-DEE.

4 Hybrid MinDEE-K^* Algorithm (Ensemble-Based Redesign)

We now present an extension and improvement to the original K^* algorithm [17] by using a version of the MinDEE criterion plus A^* branch-and-bound search. The K^* ensemble-based scoring function approximates the protein-ligand binding constant with the following quotient: $K^* = \frac{q_{PL}}{q_P q_L}$, where q_{PL}, q_P, and q_L are the partition functions for the protein-ligand complex, the free (unbound) protein, and the free ligand, respectively. A partition function q over a set (ensemble) of conformations S is defined as $q = \sum_{s \in S} \exp(-E_s/RT)$, where E_s is the energy of conformation s, T is the temperature in Kelvin, and R is the gas constant. In a naive K^* implementation, each partition function would be computed by a computationally-expensive energy minimization of all rotamer-based conformations. However, because the contribution to the partition function of each conformation is exponential in its energy, only a subset of the conformations contribute significantly. By identifying and energy-minimizing *only* the significantly-contributing conformations, a provably-accurate ε-approximation algorithm substantially improved the algorithm's efficiency [17]. The MinDEE criterion must be used in this algorithm because the K^* scoring function is based on *energy-minimized* conformations. Since pruned conformations never have to be examined, the Hybrid MinDEE-K^* algorithm provides a combinatorial improvement in runtime over the previously described constant-factor ε-approximation algorithm [17].

MinDEE (Eq. 8) can prune rotamers across mutation sequences.[4] By pruning *across* mutations with MinDEE, we risk pruning conformations that could otherwise contribute

[4] A *mutation sequence* specifies an assignment of amino-acid type to each residue in a protein.

substantially to the partition functions, thus violating our provably-good partition function approximation (Sec. 4.1). Hence, we derive a modified version of MinDEE, called *Single-Sequence* MinDEE (*SSMinDEE*), that is capable of pruning rotamers only within a single mutation sequence; the MinDEE criterion (Eq. 8) is valid for SSMinDEE.

4.1 Efficient Partition Function Computation Using A^* Search

Using A^* with SSMinDEE, we can generate the conformations of a rotamerically-based ensemble in order of increasing lower bounds on the conformation's minimized energy. As each conformation c is generated from the conformation tree, we compare a lower bound[5] $B(c)$ on its conformational energy to a moving *stop-threshold* and stop the A^* search once $B(c)$ becomes greater than the threshold, since all remaining conformations are guaranteed to have minimized energies above the stop-threshold. We now prove that a partial partition function q^* computed using only those conformations with energies below (i.e., better than) the stop-threshold will lie within a factor of ε of the true partition function q. Thus, q^* is an ε-approximation to q, i.e., $q^* \geq (1-\varepsilon)q$.

During the application of the MinDEE criterion (Eq. 8), we can easily piggyback the computation of a lower bound B_{i_r} on the energy of all conformations that contain a pruned rotamer i_r: $B_{i_r} = E_{t'} + E_{\ominus}(i_r) + \sum_{j \neq i} \min_s E_{\ominus}(j_s) + \sum_{j \neq i} \min_s E_{\ominus}(i_r, j_s) + \sum_{j \neq i} \sum_{k \neq i, k > j} \min_{s,u} E_{\ominus}(j_s, k_u)$. Now, let E_0 be the minimum lower energy bound among all conformations containing at least one pruned rotamer, $E_0 = \min_{i_r \in S} B_{i_r}$, where S is the set of pruned *rotamers*. E_0 can be precomputed during the MinDEE stage and prior to the A^* search. Let p^* be the partition function computed over the set P of pruned *conformations*, so that $p^* \leq k \exp(-E_0/RT)$, where $|P| = k$. Also, let X be the set of conformations not pruned by MinDEE and let q^* be the partition function for the top m conformations already returned by A^*; let q' be the partition function for the n conformations that have not yet been generated, all of which have energies above E_t, so that $q' \leq n \exp(-E_t/RT)$; note that $|X| = m + n$. Finally, let $\rho = \frac{\varepsilon}{1-\varepsilon}$. We can then guarantee an ε-approximation to the full partition function q using:

Proposition 3. *If the lower bound $B(c)$ on the minimized energy of the $(m+1)^{\text{st}}$ conformation returned by A^* satisfies $B(c) \geq -RT\left(\ln(q^*\rho - k\exp(-E_0/RT)) - \ln n\right)$, then the partition function computation can be halted, with q^* guaranteed to be an ε-approximation to the true partition function q, that is, $q^* \geq (1-\varepsilon)q$.*

Proof. The full partition function q is computed using all conformations in both P and X: $q = q^* + q' + p^*$. Thus, $q \leq q^* + n\exp(-E_t/RT) + k\exp(-E_0/RT)$. Hence, $q^* \geq (1-\varepsilon)q$ holds if $q^* \geq (1-\varepsilon)(q^* + n\exp(-E_t/RT) + k\exp(-E_0/RT))$. Solving for E_t, we obtain the desired stop-threshold:

$$-RT\left(\ln(q^*\rho - k\exp(-E_0/RT)) - \ln n\right) \leq E_t. \tag{11}$$

We can halt the search once a conformation's energy lower bound becomes greater than the stop-threshold (Eq. 11), since then q^* is already an ε-approximation to q. □

[5] Efficiently computed as a sum of precomputed pairwise minimum energy terms (see Appendix B in [9]).

```
Initialize: n ← Number of Rotameric Conformations; q* ← 0
while (n > 0)
      c ← GetNextAStarConf()
      if B(c) ≤ −RT (ln(q*ρ − k exp(−E0/RT)) − ln n)
            q* ← q* + exp (−ComputeMinEnergy(c)/RT)
            n ← n − 1
      else  Return q*
if q*ρ < k exp(−E0/RT)
      RepeatSearch(q*, ρ, k, E0)
else  Return q*
```

Fig. 2. Intra-Mutation Filter for Computing a Partition Function with Energy Minimization Using the A* Search. q^* is the running approximation to the partition function. The function $\mathtt{B}(\cdot)$ computes the energy lower bound for the given conformation (see Appendix B in [9]). The function $\mathtt{ComputeMinEnergy}(\cdot)$ returns a conformation's energy after energy minimization. The function $\mathtt{GetNextAStarConf}()$ returns the next conformation from the A* search. The function $\mathtt{RepeatSearch}(\cdot)$ sets up and repeats the mutation search if an ε-approximation is not achieved after the generation of all A^* conformations; the search is repeated at most once. Upon completion, q^* represents an ε-approximation to the true partition function q, such that $q^* \geq (1-\varepsilon)q$.

If at some point in the search, the stop-threshold condition has not been reached and there are no remaining conformations for A^* to extract ($n = 0$), then $q' = 0$ by definition, and $q = q^* + p^*$. Hence, if $q^*\rho \geq k\exp(-E_0/RT)$, then $q^* \geq (1-\varepsilon)q$ is already an ε-approximation to q; otherwise, the set of pruned rotamers must be reduced to guarantee the desired approximation accuracy (see Fig. 2 and [9, p. 13] for details).

Proposition 3 represents an *intra-mutation* energy filter for pruning within a single mutation sequence (Fig. 2). For an analogous provable partition-function approximation for pruning *across* mutation sequences (so that conformations for a given sequence can be pruned based on the K^* scores computed for other sequences), see [9, Sec. 4.2].

Table 1. Conformational Pruning with Hybrid MinDEE-K^*. 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, single-sequence minimized-DEE, steric, and energy (with A^*) pruning. The A^* energy filter is based on the ε-approximation algorithm in Sec. 4.1. 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.

	Conf. Remaining	Pruning Factor (%)
Initial	6.8×10^8	-
Volume Filter	2.04×10^8	3.33 (70.0)
SSMinDEE Filter	8.83×10^6	23.12 (95.7)
Steric Filter	5.76×10^6	1.53 (34.7)
A^* Energy Filter	2.78×10^5	20.7 (95.2)

We now have all the necessary tools for our ensemble-based Hybrid MinDEE-K^* algorithm. The volume filter (Sec. 5) in the original K^* is applied first to eliminate under- and over-packed mutation sequences; this is followed by the combinatorial SS-MinDEE filter and the A^* energy filter using the ε-approximation algorithm above (see Table 1). A steric filter (Sec. 5), similar to the one in [17], prevents some high-energy conformations (corresponding to steric clashes) with good lower bounds from being returned by A^*, gaining an additional combinatorial speedup. Only the conformations that pass all of these filters are energy-minimized and used in the computation of the partition function for the conformational ensemble. Finally, the K^* score for a given mutation is computed as the ratio of the bound and unbound partition functions. Hybrid MinDEE-K^* efficiently prunes the majority of the mutation sequences and conformations from more expensive full energy-minimization (see Appendix B in [9]), while still giving provable guarantees on the accuracy of its score predictions.

5 Methods

Structural Model. Our structural model is the same as the one used in the original K^* [17]. In our experiments, the structural model consists of the 9 active site residues (D235, A236, W239, T278, I299, A301, A322, I330, C331) of GrsA-PheA (PDB id: 1AMU) [4], the steric shell (the 30 residues with at least one atom within 8 Å of a residue in the active site), the amino acid substrate, and the AMP cofactor. Flexible residues are represented by rotamers from the Lovell *et al.* rotamer library [19]. Each rotameric-based conformation is minimized using steepest-descent minimization and the AMBER energy function (electrostatic, vdW, and dihedral energy terms) [31, 5]. For full details of our structural model, see [9, Sec. 5].

Energy Precomputation for Lower Bounds, B$(\cdot)$. The MinDEE criterion (Eq. 8) uses both min and max *precomputed* energy terms to determine which rotamers are not part of the minGMEC. There is no need to re-compute the min and max energies every time Eq. (8) is evaluated. See Appendix B in [9] for a detailed discussion.

Approximation Accuracy. We use $\varepsilon = 0.03$, thus guaranteeing that the computed partial partition functions will be at least 97% of the corresponding full partition functions.

Filters. *Volume filter:* Mutation sequences that are over- or under-packed by more than 30Å^3 compared to the wildtype PheA are pruned; *Steric filter*: Conformations in which a pair of atoms' vdW radii overlap by more than 1.5Å prior to minimization are pruned; *Sequence-space filter*: The active site residues are allowed to mutate to the set (GAVLIFYWM) of hydrophobic amino acids; *MinDEE*: We use an implementation of the MinDEE analog to the simple coupled Goldstein criterion ([10] and Fig. 4d in [9]).

6 Results and Discussion

In this section, we compare the results of GMEC-based protein redesign without (traditional-DEE/A^*) and with (MinDEE/A^*) energy minimization. We also compare the

redesign results when energy minimization is used without (MinDEE/A^*) and with (Hybrid MinDEE-K^*) conformational ensembles. We further compare our ensemble-based redesign results both to our previous computational predictions of protein designs and to biological activity assays of predicted protein mutants.

Comparison to Biological Activity Assays. Similarly to [17], we simulated the biological activity assays of L-Phe and L-Leu against the wildtype PheA enzyme and the double mutant T278M/A301G [28]. In [28], T278M/A301G was shown to have the desired switch of specificity from Phe to Leu by performing activity assays. The activity for both the wildtype and the mutant protein sequences was normalized, so that the substrate with the larger activity was assigned a specificity of 100%, while the other substrate was assigned specificity relative to the first one. The wildtype PheA had a specificity of 100% for Phe and approximately 7% for Leu; the double mutant had a specificity of 100% for Leu and approximately 40% for Phe. The computed Hybrid MinDEE-K^* normalized scores qualitatively agreed with these results, showing the desired switch of specificity for T278M/A301G. The wildtype sequence had a normalized K^* score of 100% for Phe and 0.01% for Leu; the double mutant had a normalized score of 100% for Leu and 20% for Phe.

Comparison to Traditional-DEE. For comparison, the simple coupled Goldstein traditional-DEE criterion [10] was used in a redesign search for changing the specificity of the wildtype PheA enzyme from Phe to Leu, using the experimental setup in Sec. 5. A comparison to the rotamers in the minGMEC A236M/A322M (see MinDEE/A^* results below), revealed that 2 of these 9 rotamers were in fact *pruned* by traditional-DEE. As an example, the minGMEC was energy-minimized from a conformation that included rotamer 5 [19] of Met at residue 236. This particular rotamer (χ angles -177°, 180°, and 75°) was pruned by traditional-DEE. We then energy-minimized A236M/A301G, the rigid-GMEC obtained by traditional-DEE/A^*, and determined that its energy was higher (by appx. 5 kcal/mol) than the energy for the minGMEC obtained by MinDEE/A^*. Moreover, a total of 104 different conformations minimized to a lower energy than the rigid-GMEC. These results confirm our claim that traditional-DEE is not provably-accurate with energy-minimization; they also show that conformations pruned by traditional-DEE may minimize to a lower energy state than the rigid-GMEC.

Hybrid MinDEE-K^*. The experimental setup for Leu redesign with Hybrid MinDEE-K^* is as described in Sec 5. The 2-point mutation search took approximately 10 hours on a cluster of 24 processors. Only 30% of the mutation sequences passed the volume filter, while MinDEE pruned over 95% of the remaining conformations. The use of the ε-approximation algorithms reduced the number of conformations that had to be subsequently generated and energy-minimized by an additional factor of twenty (see Table 1). A brute-force version of Hybrid MinDEE-K^* that did not utilize any of the filters, would take approximately 2,450 times longer (appx. 1,023 days).

The two top-scoring sequences are A301G/I330W and A301G/I330F for both Hybrid MinDEE-K^* and the original K^* [17]. These novel mutation sequences were tested in the wetlab and were shown to have the desired switch of specificity from

Phe to Leu [17]. Moreover, the other known successful redesign T278M/A301G [28] is ranked 4^{th}. Furthermore, all of the top 17 Hybrid MinDEE-K^* sequences contain the mutation A301G, which is found in all known native Leu adenylation domains [3]. These results show that our algorithm can give reasonable predictions for redesign.

To compare the efficiency of Hybrid MinDEE-K^* and the original K^*, we measured the number of fully-evaluated conformations. The original K^* (using the better minimizer of Hybrid MinDEE-K^*, see Appendix B in [9]) fully-evaluated approx. 30% more conformations than the 2.78×10^5 evaluated by Hybrid MinDEE-K^* (Table 1). Thus, Hybrid MinDEE-K^* is much more efficient at obtaining the desired results.

MinDEE/A^*. We now discuss results from our non-ensemble-based experiments using MinDEE/A^*. To redesign the wildtype PheA enzyme so that its substrate specificity is switched towards Leu, we used the experimental setup described in Sec. 5. The MinDEE filter on the bound protein:ligand complex pruned 206 out of the 421 possible rotamers for the active site residues, reducing the number of conformations that were subsequently supplied to A^* by a factor of 2,330. We then extracted and minimized all conformations over the 2-point mutation sequences using A^* until the halting condition defined in Proposition 2 was reached, for $E_w = 8.5$ kcal/mol. A total of 813 conformations, representing 45 unique mutation sequences, had actual minimized energies within 8.5 kcal/mol of the minGMEC energy. The top-ranked MinDEE/A^* sequence is A236M/A322M; the minGMEC is obtained from this sequence. The entire redesign process took approximately 14 days on a single processor, with more than $120,000$ extracted conformations before the search could be provably halted. Thus, the provable accuracy of the results comes at the cost of this computational overhead. Note, however, that a redesign effort without a MinDEE filter and a provable halting condition would be computationally infeasible.

Like A301G/I330W and A301G/I330F, the top 5 MinDEE/A^* sequences are unknown in nature. To assess the switch of specificity from Phe to Leu, we extracted the minimum-energy conformation for these top 5 Leu-binding sequences. Each of these 5 conformations was then energy-minimized when bound to Phe. Whereas the Leu-bound energies were negative and low, the corresponding Phe-bound energies were positive and high. Thus, the top sequences are predicted to bind more stably to Leu, as desired.

Only 9 of the 45 MinDEE/A^* sequences passed the Hybrid MinDEE-K^* volume filter. Moreover, only 5 of the MinDEE/A^* sequences were found in the top 40 Hybrid MinDEE-K^* sequences, indicating that ensemble-scoring yields substantially different predictions from single-structure scoring using the minGMEC, where only the minimized *bound* state of a *single* conformation is considered (see Fig. 3 in [9, p. 20]). We can conclude that, currently, MinDEE appears useful as a filter in the Hybrid MinDEE-K^* algorithm; however, the incorporation of additional information, such as a comparison to negative design (the energies to bind the wild-type substrate), may promote MinDEE as a valuable stand-alone non-ensemble-based algorithm for protein redesign.

7 Conclusions

When energy-minimization is required, the traditional-DEE criterion makes no guarantees about pruning rotamers belonging to the minGMEC. In contrast, a rotamer is only pruned by MinDEE if it is provably not part of the minGMEC. We showed experimentally that the minGMEC can minimize to lower energy states than the rigid-GMEC, confirming the feasibility and significance of our novel MinDEE criterion. When used as a filter in *ensemble-based* redesign, MinDEE efficiently reduced the conformational and sequence search spaces, leading both to predictions consistent with previous redesign efforts and novel sequences that are unknown in nature. Our Hybrid MinDEE-K^* algorithm showed a significant improvement in pruning efficiency, as compared to the original K^* algorithm. Redesign searches for two other substrates, Val and Tyr, have also been performed, confirming the generality of our algorithms.

Protein design using traditional-DEE uses neither ensembles nor rotamer minimization. In our experiments, we reported the relative benefits of incorporating ensembles and energy-minimization into a provable redesign algorithm. A major challenge for protein redesign algorithms is the balance between the efficiency and accuracy with which redesign is performed. While the ability to prune the majority of mutation/conformation search space is extremely important, increasing the accuracy of the model is a prerequisite for successful redesign. It would be interesting to implement finer rotamer sampling and more accurate (and hence more expensive) energy functions, remove bias in the rotamer library by factoring the Jacobian into the partition function over torsion-angle space, and incorporate backbone flexibility. An accurate and efficient algorithm for redesigning natural products should prove useful as a technique for drug design.

Acknowledgments. We thank Prof. A. Anderson, Dr. S. Apaydin, Mr. J. MacMaster, Mr. A. Yan, Mr. B. Stevens, and all members of the Donald Lab for helpful discussions and comments.

References

1. D. Bolon and S. Mayo. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA*, 98:14274–14279, 2001.
2. D. Cane, C. Walsh, and C. Khosla. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science*, 282:63–68, 1998.
3. G. Challis, J. Ravel, and C. Townsend. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.*, 7: 211–224, 2000.
4. E. Conti, T. Stachelhaus, M. Marahiel, and P. Brick. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of Gramicidin S. *EMBO J.*, 16:4174–4183, 1997.
5. W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.
6. J. Desmet, M. Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

7. S. Doekel and M. Marahiel. Dipeptide formation on engineered hybrid peptide synthetases. *Chem. Biol.*, 7:373–384, 2000.
8. K. Eppelmann, T. Stachelhaus, and M. Marahiel. Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry*, 41:9718–9726, 2002.
9. I. Georgiev, R. Lilien, and B. R. Donald. A novel minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. Technical Report 570, Dartmouth Computer Science Dept., http://www.cs.dartmouth.edu/reports/abstracts/TR2006-570, 2006.
10. R. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.*, 66:1335–1340, 1994.
11. D. Gordon and S. Mayo. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.*, 19:1505–1514, 1998.
12. H. Hellinga and F. Richards. Construction of new ligand binding sites in proteins of known structure: I. Computer-aided modeling of sites with pre-defined geometry. *J. Mol. Biol.*, 222:763–785, 1991.
13. A. Jaramillo, L. Wernisch, S. Héry, and S. Wodak. Automatic procedures for protein design. *Comb. Chem. High Throughput Screen.*, 4:643–659, 2001.
14. W. Jin, O. Kambara, H. Sasakawa, A. Tamura, and S. Takada. De novo design of foldable proteins with smooth folding funnel: Automated negative design and experimental verification. *Structure*, 11:581–591, 2003.
15. A. Keating, V. Malashkevich, B. Tidor, and P. Kim. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. *Proc. Natl. Acad. Sci. USA*, 98:14825–14830, 2001.
16. A. Leach and A. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33:227–239, 1998.
17. R. Lilien, B. Stevens, A. Anderson, and B. R. Donald. 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*, 12(6–7):740–761, 2005.
18. L. Looger, M. Dwyer, J. Smith, and H. Hellinga. Computational design of receptor and sensor proteins with novel functions. *Nature*, 423:185–190, 2003.
19. S. Lovell, J. Word, J. Richardson, and D. Richardson. The penultimate rotamer library. *Proteins*, 40:389–408, 2000.
20. J. Marvin and H. Hellinga. Conversion of a maltose receptor into a zinc biosensor by computational design. *PNAS*, 98:4955–4960, 2001.
21. H. Mootz, D. Schwarzer, and M. Marahiel. Construction of hybrid peptide synthetases by module and domain fusions. *Proc. Natl. Acad. Sci. USA*, 97:5848–5853, 2000.
22. R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman. Side-chain flexibility in proteins upon ligand binding. *Proteins*, 39(3):261–8, 2000.
23. N. Pierce, J. Spriet, J. Desmet, and S. Mayo. Conformational splitting: a more powerful criterion for dead-end elimination. *J. Comput. Chem.*, 21:999–1009, 2000.
24. N. Pierce and E. Winfree. Protein design is *NP*-hard. *Protein Eng.*, 15:779–782, 2002.
25. J. Ponder and F. Richards. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193: 775–791, 1987.
26. A. Schneider, T. Stachelhaus, and M. Marahiel. Targeted alteration of the substrate specificity of peptide synthetases by rational module swapping. *Mol. Gen. Genet.*, 257:308–318, 1998.
27. D. Schwarzer, R. Finking, and M. Marahiel. Nonribosomal peptides: from genes to products. *Nat. Prod. Rep.*, 20:275–287, 2003.

28. T. Stachelhaus, H. Mootz, and M. Marahiel. The specificiy-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, 6:493–505, 1999.
29. T. Stachelhaus, A. Schneider, and M. Marahiel. Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science*, 269:69–72, 1995.
30. A. Street and S. Mayo. Computational protein design. *Structure*, 7:R105–R109, 1999.
31. S. Weiner, P. Kollman, D. Case, U. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.