

Lecture 8

Nonribosomal Code & Protein Design

Protein redesign plays an important role in the realization of novel molecular functions and drug design. In this lecture, we introduce a novel algorithm for protein redesign problem, 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 [1].

1 Nonribosomal Peptide Synthetase (NRPS) Enzyme

Nonribosomal peptide synthetase (NRPS) enzymes, usually produced by microorganisms like bacteria and fungi, complement the traditional ribosomal peptide synthesis pathway. They are also the sources of hundreds of peptide-like products with pharmaceutical properties, including natural antibiotics, antifungals, antivirals, anticancer therapeutics, immunosuppressants, and siderophores. Enzymes of the NRPS pathway have multiple domains with individual functions acting in an assembly-line fashion. It is commonly believed that the substrate specificity of the NRPS enzymes is dictated primarily by the “gatekeeper” adenylation (A), and recent evidence also indicates that the condensation (C), thiolation (T), and epimerization (E) domains may carry some specificity as well.

Previous NRPS enzyme redesign methods can be divided into two main techniques, *domain-swapping* and *active site modification through site-directed mutagenesis*. Domain-swapping techniques modify NRPS enzymes by swapping an adenylation domain of an existing NRPS enzyme for an adenylation domain from a second, different NRPS enzyme. Active site modification through site-directed mutagenesis utilizes structural information of the GrsA-PheA enzyme. From sequence alignment of GrsA-PheA with 160 other known adenylation domains, a “signature sequence” was derived for each adenylation domain by extracting those residues that align with the structurally determined substrate binding pocket of the GrsA-PheA crystal structure.

2 Methods

First, [1] developed an ensemble scoring method K^* to model the protein-ligand binding in a single mutation. Since it is not currently possible to compute exact partition functions for complex molecular species, K^* approximates these partition functions with the use of rotamerically-based conformational ensembles.

When applying K^* to a protein-ligand system, a number of choices must be made with respect to ensemble generation and single-structure scoring, where *single-structure scoring*

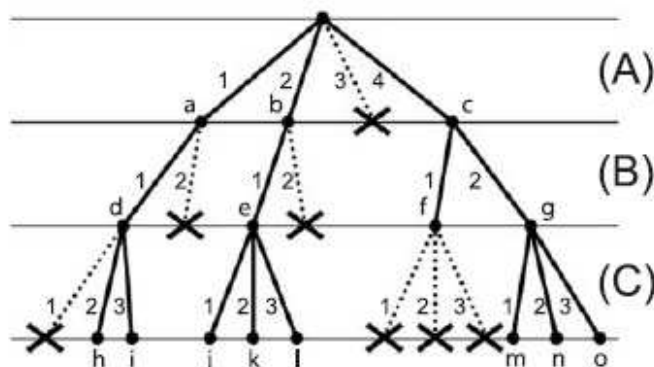


Figure 1: An example conformation tree. The rotamers of flexible residue i are represented by the branches at depth i . Internal nodes of a conformation tree represent partially-assigned conformations. X's represent nodes of the conformation tree where steric clash has been identified among a partially assigned conformation. All children of X nodes are pruned and not considered.

is the method by which each individual member of ensemble is scored. The choices made in ensemble scoring should strike a balance between fidelity to the underlying physical biochemistry and computational feasibility. [1] indicates a brute-force algorithm for mutation search: (1) generate all the molecular ensembles by fixing the protein backbone and using rotamer library to vary side-chain conformation; (2) Use an AMBER energy function to energy-minimize each generated conformation; (3) Compute each partition function and then combine them to obtain an overall K^* score.

To search the optimal mutation sequence more efficiently, mutation space filters and conformation space pruning are employed. Mutation space filters indicate to two filters: sequence-space filter, a residue type filter which restricts the mutation search to include only a subset of amino acids based on compatibility with the target substrate, and volume filter, which removes mutations that significantly over- or underpack the substrate-bound active site relative to the wildtype. These two filters prune a combinatorial number of conformations from consideration and eliminate the majority of conformations early in mutation space. Moreover, since conformations with large energies are unlikely to be assumed and contribute only a vanishingly small amount to the partition function, it is reasonable to prune such conformations from consideration. [1] employed a DEE strategy that generates conformations and prunes unlike conformations and mutations (Figure 1).

By ignoring the pruned conformations, a so-called *intra-mutation pruning* is then applied during the computation of a single conformation for a single mutation. The intra-mutation pruning algorithm is capable of guaranteeing any desired approximation accuracy to the true partition function as shown in Figure 2. That is, a conformation c_{k+1} can be pruned if it satisfies $B(c_{k+1}) \geq -RT \ln(q_k^* \epsilon - p_k^*)$, and by induction, if all pruned conformations satisfy this condition, then at the end of the computation, q_{k+1}^* will be ϵ -approximation to q_n , the

```

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  $B(c_k) \leq -RT \ln(q^* \epsilon - p^*)$ 
     $q^* \leftarrow q^* + \exp(-\text{ComputeMinEnergy}(c_k)/RT)$ 
  else
     $p^* \leftarrow p^* + \exp(-B(c_k)/RT)$ 
Return  $q^*$ 

```

Figure 2: Intra-mutation pruning. Here q^* is the running approximation to the partition function, and 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 $\text{ComputeMinEnergy}(\cdot)$ returns the energy of the energy-minimized conformation as computed using steepest-descent minimization and our implementation of the AMBER energy function. At the end, q^* represents an ϵ -approximation to the true partition function q such that $q^* \leq (1 - \epsilon)q$.

real partition function.

In summary, four levels of approximation are used in computing K^* scores. (1) A rotamer library is used to model side chain conformations. (2) Sterically disallowed rotamer-based conformations are combinatorially pruned. (3) Intra- and inter-mutation pruning methods are used to skip evaluation of conformations not required to compute ϵ -approximations to the partition function and K^* values. (4) In computing partition functions and K^* values, the algorithm starts with discrete rotamers and then performs bounded minimization.

3 Results

The work in [1] constructed a structural model based on the previously solved structure of GrsA-PheA (1AMU) which consists of 9 active site residues. Several tests were implemented to validate the algorithm. One is to find an accepted conformation for Phe in the PheA active site to confirm the rotamer search strategy with minimization and the scoring scheme. Another one is to simulate the biochemical activity assays of L-Phe and L-Leu against wildtype PheA and the T278M/A301G double mutation, and the results qualitatively agree with the activity assays of Stachelhaus *et al* [2]. Moreover, A K^* mutation search was performed to redesign GrsA-PheA to bind the adenylate Leu instead of Phe. Two novel mutations were reported that are unknown in nature and have never been tested before, and a known Leu binding mutation is ranked highly in the mutation search results.

References

- [1] Lilien RH, Stevens BW, Anderson AC, Donald BR. 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. *J Comput Biol.* 2005 Jul-Aug;12(6):740-61.
- [2] Stachelhaus, T., Mootz, H., and Marahiel, M. 1999. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* 6, 493-505.