

Lecture 7

Computational Protein Design

This lecture presents the automated protein design and experimental validation of a novel sequence, as described in [1].

1 Introduction

Given a 3-D backbone structure, the *protein design* problem is to find an optimal sequence that satisfies the physical chemical potential functions and stereochemical constraints. Protein design is an “*inverse folding problem*”, and fundamental for understanding the protein function.

2 Overview of Methodology

Following is the methodology used in [1]:

Given a backbone fold of a target structure, [1] first developed an automated side-chain selection algorithm to (1) screen all possible amino acid sequences, and (2) find the optimal sequence and side-chain orientations. Then the experimental validation by using NMR is applied to evaluate the computed optimal sequence.

3 Algorithm Design

Input: Backbone fold (Zif268), represented by structure coordinates.

Output: Optimal sequence (FSD-1)

Overview:

- (1) The algorithm considers specific interactions between (a) side-chain and backbone and (b) side chain and side chain.
- (2) The algorithm scores a sequence arrangement, based on a van der Waals potential function, solvation, hydrogen bonding, and secondary structure propensity [1].
- (3) The algorithm considers a discrete set of rotamers, which are all allowed conformers of each side chain.
- (4) The algorithm applies the *dead-end elimination* (DEE) to prune rotamers that are inconsistent with the global minimum energy solution of the system.

Details:

The inputs of the algorithm are structure coordinates of the target motif's backbone, such as N, C_α, C and O atoms, and C_α-C_β vectors. The residue positions in the protein structure are partitioned into *core*, *surface*, and *boundary* classes. The set of possible amino acids at the core positions is {Ala, Val, leu, Ile, Phe, Tyr, Trp}. The set of amino acids considered at the surface positions is {Ala, Ser, Thr, His, Asp, Asn, Glu, Gln, Lys, Arg}. The combined set of both core and surface amino acids are considered for the boundary positions.

Notes: The total number of possible amino acid sequences is equal to the product of possible amino acids at each residue position. For instance, suppose that there are 7 possible amino acids at one core position, and 16 possible amino acids at each of 7 boundary positions, and 10 possible amino acids at each of 18 surface positions. Then searching space consists of $7 * 16^7 * 10^{18} = 1.88 * 10^{27}$ possible amino acid sequences.

The algorithm is divided into two phases:

Phase 1 (Scoring): The different scoring functions are defined for core, surface and boundary residues separately. The scoring function for core residues uses “a van der Waals potential to account for steric constraints and an atomic solvation potential favoring the burial and penalizing the exposure of nonpolar surface area” [1]. The surface residues apply a hydrogen-bond potential and secondary structure propensities, and a van der Waals potential. The residues at the boundary positions utilize a combination of both core and surface scoring functions.

Phase 2 (Pruning): The algorithm applies DEE to find and eliminate rotamers that are dead-ending with the global minimum energy solution. A rotamer r at the residue position i will be eliminated (or dead-ending) if there is another rotamer t at the same position such that:

$$E(i_r) - E(i_t) + \sum_j \min_s [E(i_r j_s) - E(i_t j_s)] > 0,$$

where $E(i_r)$ and $E(i_t)$ represent rotamer-template energies, while $E(i_r j_s)$ and $E(i_t j_s)$ represent rotamer-rotamer energies for rotamers i and j .

Results:

Figure 1 shows the comparison of optimal computed sequence FSD-1 computed and the target sequence Zif268, while Figure 2 gives the structure comparisons between these two sequence.

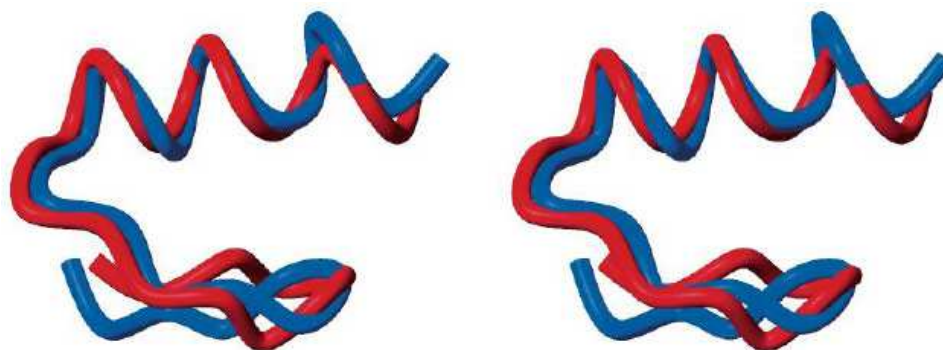


Figure 2: Structure comparisons of computed sequence FSD-1 and the target sequence Zif268 [1]. Comparison of the FSD-1 structure (blue) and the design target (red). Stereoview of the best-fit superposition of the restrained energy minimized average NMR structure of FSD-1 and the backbone of Zif268. Residue 3 to 26 are shown.

4 Complexity Analysis

Let n denote the number of residues, and r denote the (maximum) number of possible rotamers for each residue.

We first analyze the time complexity of DEE pruning in Phase 2. For each rotamer at a specific residue position i , it takes $O(nr)$ to search all r possible amino acids in all other $n - 1$ positions to find $\sum_j \min_s [E(i_r j_s) - E(i_t j_s)]$. Comparisons with other rotamer at the same position i take $r \cdot O(nr) = O(nr^2)$ time. Since we need to consider all possible rotamers at every position i , the total DEE pruning takes $n \cdot r \cdot O(nr^2) = O(n^2 r^3)$ time.

Although the DEE pruning method in Phase 2 takes polynomial time, unfortunately, the scoring in Phase 1 takes exponential time r^n , since we have to compare each set of possible amino acids in each residue position.

Notes: In fact, the optimization problem in protein design has been proved NP-hard.

5 Experimental Validation

The solution structure of computed sequence FSD-1 was obtained by using 2D ^1H NMR spectroscopy. Sample NMR spectra are shown in Figure 3. X-PLOR plus the standard protocols for hybrid distance geometry-simulated annealing was used to calculate the structure. Figure 4 and 5 show an ensemble of 41 structures that are consistent with good geometry and distance constraints within a small tolerance.

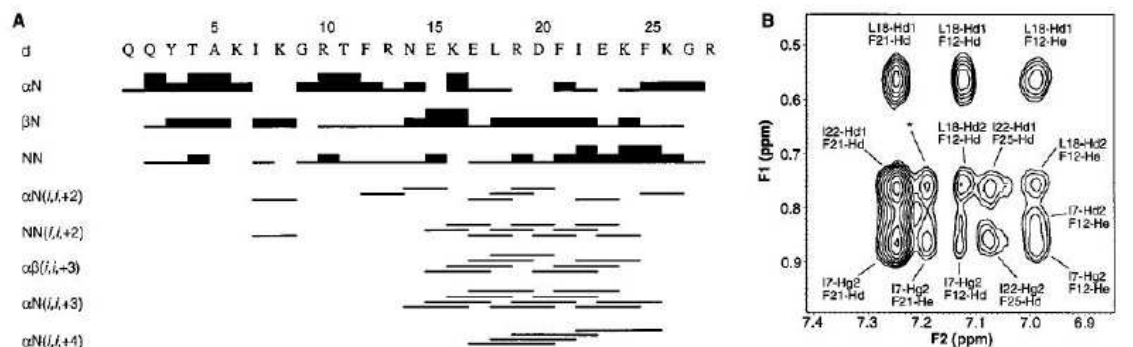


Fig. 4. NCE contacts for FSD-1. **(A)** Sequential and short-range NCE contacts. The *d* denotes a contact between the indicated protons. All adjacent residues are connected by H α -HN, HN-HN, or H β -HN NCE cross-peaks. The helix (residues 15 to 26) is well defined by short-range connections, as is the hairpin turn at residues 7 and 8. **(B)** Representative NOE contacts from aromatic to methyl protons. Several long-range NOEs from Ile⁷ and Phe¹² to the helix help define the fold of the protein. The starred peak has an ambiguous F1 assignment, Ile²² Hd1 or Leu¹⁹ Hd2.

Figure 3: The NMR spectra of computed sequence FSD-1 [1].

Table 1. NMR structure determination: distance restraints, structural statistics, and atomic root-mean-square (rms) deviations. $\langle SA \rangle$ are the 41 simulated annealing structures, SA is the average structure before energy minimization, $\langle SA \rangle_r$ is the restrained energy minimized average structure, and SD is the standard deviation.

Distance restraints		
Intraresidue		97
Sequential		83
Short range ($ i - j = 2$ to 5 residues)		59
Long range ($ i - j > 5$ residues)		35
Hydrogen bond		10
Total		284
Structural statistics		
rms deviations	$\langle SA \rangle \pm SD$	$\langle SA \rangle_r$
Distance restraints (\AA)	0.043 ± 0.003	0.038
Idealized geometry		
Bonds (\AA)	0.0041 ± 0.0002	0.0037
Angles (degrees)	0.67 ± 0.02	0.65
Impropers (degrees)	0.53 ± 0.05	0.51
Atomic rms deviations (\AA) [*]		
	$\langle SA \rangle$ versus $SA \pm SD$	$\langle SA \rangle$ versus $\langle SA \rangle_r \pm SD$
Backbone	0.54 ± 0.15	0.69 ± 0.16
Backbone + nonpolar side chains [†]	0.99 ± 0.17	1.16 ± 0.18
Heavy atoms	1.43 ± 0.20	1.90 ± 0.29

^{*}Atomic rms deviations are for residues 3 to 26, inclusive. Residues 1, 2, 27, and 28 were disordered ($\langle \phi, \psi, \text{angular order parameters} \rangle < 0.75$) and had only sequential and $|i - j| = 2$ NCEs. [†]Nonpolar side chains are from residues Tyr³, Ala⁵, Ile⁷, Phe¹², Leu¹⁹, Phe²¹, Ile²², and Phe²⁵, which constitute the core of the protein.

Figure 4: NMR structure determination of FSD-1 [1].

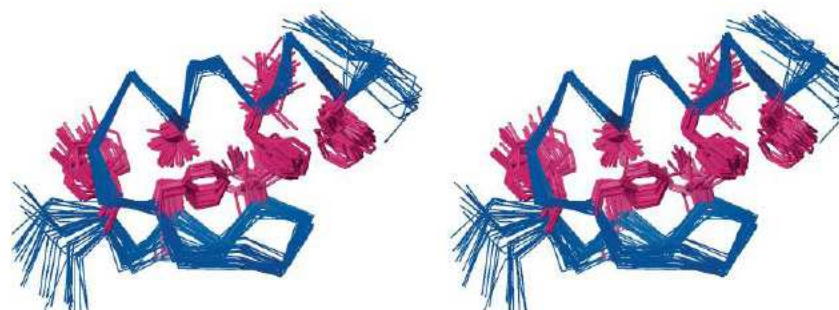


Fig. 5. Solution structure of FSD-1. Stereoview showing the best-fit superposition of the 41 converged simulated annealing structures from X-PLOR (37). The backbone C α trace is shown in blue and the side-chain heavy atoms of the hydrophobic residues (Tyr³, Ala⁵, Ile⁷, Phe¹², Leu¹⁹, Phe²¹, Ile²², and Phe²⁵) are shown in magenta. The amino terminus is at the lower left of the figure and the carboxyl terminus is at the upper right of the figure. The structure consists of two antiparallel strands from positions 3 to 6 (back strand) and 9 to 12 (front strand), with a hairpin turn at residues 7 and 8, followed by a helix from positions 15 to 26. The termini, residues 1, 2, 27, and 28 have very few NOE restraints and are disordered.

Figure 5: Calculated structures of FSD-1 [1].

References

- [1] B. I. Dahiyat and S. L. Mayo. De Novo Protein Design: Fully Automated Sequence Selection *Science*, October 3; 278 (5335):82, 1997.
- [2] Niles Pierce and Erik Winfree. Protein Design is NP-Hard. *Protein Engineering*, v15, pp. 779-782, 2002.