In silico protein design: the implementation of Dead-End Elimination algorithm

CS 273 Spring 2005: Project Report

Tyrone Anderson[2], Yu Bai1[3], and Caroline E. Moore-Kochlacs[2]

[1]Biophysics program, [2]Department of Biomedical Informatics, [3]Department of Biochemistry, Stanford University, CA, 94305

Contacts: yubai@stanford.edu, tanderso@stanford.edu, caromk@stanford.edu

Abstract

In silico protein design has emerged as a powerful method for understanding the underlying physical principles that dictate protein folding and function. Given a desired protein structure, computational protein design aims to select the best stable sequence(s), minimizing the energy of the designed protein. The possible solution space is an exponential function of the protein length. Various efficient search algorithms for selecting the rotamer configurations have been developed to reduce the search space. In our project, we implement one of these algorithms, Dead-End Elimination (DEE), which selects sequences (and their associated rotamer configurations) that are likely to achieve a desired fold. The sequences of the cold-shock protein core were predicted.

Introduction

Proteins are nano-machines engineered by evolution that carry out catalytic, structural, and signaling functions essential to life. Scientists across fields like molecular biology, chemistry, engineering, and material science are working towards a predictive molecular understanding of protein functions. The primary structure of proteins is conserved in genetic material, which determines the protein's sequence of amino acids. To understand the function of a protein, one must understand the sequence-structure relationship beyond this primary structure. The function of proteins is largely dictated by its secondary (and tertiary and quartenary) structure, specifically, how the linear sequence of amino acids folds into a complex 3-D structure.

While great advances are being made in the field of structure prediction, its complement, the challenge of protein design has also experienced many breakthroughs in recent years (Park 2004). Structure prediction starts with the primary sequences and predict the corresponding 3-D structure. Protein design starts with a target structure and searches for protein sequences that are compatible with the folds of the target structure. Protein design and protein prediction use many of the same theoretical inputs and the physical potentials to evaluate the sequence-structure compatibility. Through protein design, one can learn much about the molecular nature of interactions between amino acids by testing the assumptions and techniques used during the prediction-synthesis-analysis cycle. The insights learned from these efforts will help elucidate how simple protein modules come together to create complex protein assemblies.

The design of protein sequences requires evaluating an extraordinarily large number of sequences for their ability to 'fit' a given structure. Efforts have been taken to reduce solution space. In 1987, Ponder and Richards introduced the idea of using a library of rotamers, the clustered, discrete side-chain conformations for a certain amino acid, for the possible configurations of amino acids in the search. Their insight drew from the observation that side-chain torsions fall into N-dimensional clusters. Today, the use of rotamer libraries significantly improves both the speed and accuracy of sequence design.

Still, the combinatorial complexity remains exponential for these rotameric choices. Further reductions in complexity come from the use of special algorithms to identify the GMEC (Global Minimum Energy Conformation) of a given fold (Gordon 1999, Hellinga 1998, Street & Mayo 1999). Searching algorithms can be divided into two categories, stochastic and deterministic. The stochastic algorithms, such as Monte Carlo, genetic algorithms (Desjarlais 1995; Jones 1994), and the mean-field optimization (Koehl 1994), sample the solutions semi-randomly; however, there is no guarantee that these algorithms will explore solutions near the global energy minimum. The deterministic algorithms, on the other hand, apply rejection criteria to eliminate the vast majority of combinatorial possibilities before formally considering them as GMEC candidates.

In this project, we implemented Dead-End Elimination algorithm (Looger 2001), a pruning method, to re-design the core region of the major cold shock protein (1MJC.pdb). The DEE algorithm was applied to rotamer clusters with a size of one and two. Significant reduction of searching space enabled the sequence evaluation to finish in days. The predicted sequences are 50% identical with the native sequence and have energy scores highly similar to the native sequence.

Materials and Methods

The program was coded in Perl and executed on the BioX-cluster at Stanford University. The algorithm contains rotamer library manipulations (Tyrone Anderson), scoring functions (Caroline E. Moore-Kochlacs), and DEE algorithm implementations including backtrack sequence constructions and final exhaustive searching (Yu Bai).

Rotamer Library Manipulation

The rotamer library was obtained from the website supporting scap (a side-chain prediction program by Jason A. Xiang at the Homig lab at Columbia University). The library from scap contained more data than necessary for protein prediction. We converted the lines corresponding to atoms relevant to our project (nitrogen, oxygen, different types of carbon) into a 3-D array containing the residues, their corresponding atoms and relative coordinates.

Code was written to manipulate the rotamer library contained in this 3-D array. Given a specific residue (e.g. 'R' for arginine) and an associated rotamer number (e.g. 2 for the 2nd rotamer specific to the arginine residue), the desired rotamer (atoms and relative coordinates) was retrieved to be added to the backbone. Before the rotamer can be added, we orient the rotamer correctly onto the backbone. The backbone and rotamer were translated to the origin (0, 0, 0). After this translation, the backbone was fixed and the rotamer was rotated around the X-axis. To rotate, the coordinates of each of the atoms in the rotamer was multiplied by a 3x3 rotation matrix (1) specific to the axis being rotated around, the X-axis in this case.

$$\mathsf{R}_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \quad (1)$$

The rotamer is then rotated around the Z-axis using the same technique, with the appropriate matrix for Z-axis rotations (2).

$$\mathbf{R}_z(\gamma) = \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

These two rotations were followed by a second translation returning the alpha carbon to its original (x, y, z) position and translating the other rotamer atoms accordingly. The final result returned by the function manipulating the desired rotamer was a 2-D array containing the coordinates for each atom of the rotamer. The rotamer was already aligned with the backbone.

Scoring Functions

The scoring function for a particular rotamer-backbone configuration is calculated as the energy of that configuration. Protein design typically considers a variety of energy terms (Gordon, 1999), the most common of which are: (1) van der Waals for packing specificity, (2) hydrogen bonding, typically represented by an angle dependent, 12-10 hydrogen bond potential, (3) electrostatics which guard against destabilizing interactions between like charged residues, (4) internal coordinate terms for the 'bonded' energies, (5) solvation energy for protein-solvent interactions, and (6) entropy, which assumes that the conformational space is completely restricted in the folded state. For simplicity, we only looked at three of these terms, specifically, the van der Waals potential, the electrostatics term, and the solvation energy.

Van der Waals potential represents the interaction between two uncharged atoms. Two atoms are mildly attractive as they approach from a distance and repulsive as they approach too close. The resulting term prefers native-like folded states with well-organized cores over disordered or molten-globule states (Gordon, 1999). The standard approximation is the 12-6 Lennard-Jones potential, given by:

$$E_{vdW} = D_0\left[\left(\frac{R_0}{R}\right)^{12} - 2\left(\frac{R_0}{R}\right)^{6}\right] \quad (3)$$

Where D_0 is the well depth, R is the distance between atoms, and R0 is the van der Waals radii. The well depth and van der Waals radii are taken from the CHARMM19 parameter set (Neria et al 1996). To account for the fixed backbone and rotamer set, we use a dampened variation of the potential (Kuhlman and Baker, 2004):

$$E_{atr} = \sum_{i}^{natom}\sum_{j>i}^{natom}\left[\left(\frac{R_0}{R}\right)^{12} - 2\left(\frac{R_0}{R}\right)^{6}\right] \quad if\ \frac{R_0}{R} < 1.12$$
$$E_{rep} = \sum_{i}^{natom}\sum_{j>i}^{natom}\left(10.0 - 11.2\left(\frac{R}{R_0}\right)\right) \quad if\ \frac{R_0}{R} \geq 1.12 \quad (4)$$

The electrostatic term is used for stability. At moderate temperatures, folding and functional interactions may be the more significant role of electrostatics (Gordon, 1999). The term guards against destabilizing interactions between like-charged residues. We use a simple implementation of Coulomb's Law (though possibly strong Bayesian versions conditioning on the probability of two

amino acids in close proximity also exist, see Kuhlman & Baker, 2004). In equation (5), Q_i and Q_j are the charges on the charges on two atoms, R is the distance between two atoms, and ε is the dielectric constant.

$$E_{elec} = 322.0637\left(\frac{Q_i Q_j}{\epsilon R}\right) \quad (5)$$

Modeling solvation effects represents how the hydrophobic effect drives folding. Computing explicitly all protein-solvation interactions is too computationally expensive so we use an approximation from from Lazaridis and Karplus (1999).

$$E_{solv} = \sum_{i}^{natom}\left(\Delta G_i^{ref} - \sum_{j>i}^{natom}\left\{\frac{2\Delta G_i^{free}}{4\pi\sqrt{\pi}\lambda_j r_{ij}^2}\exp(-d_{ij}^2)V_i + \frac{2\Delta G_j^{free}}{4\pi\sqrt{\pi}\lambda_j r_{ij}^2}\exp(-d_{ij}^2)V_j\right\}\right) \quad (6)$$

Where dij = distance between atoms, rij = van der Waals radii, Vi = atomic volume
ΔGref = reference solvation free energy, ΔGfree = solvation free energy of free (isolated) group
λ = correlation length.

Though the electrostatics and solvation terms were implemented, they were not used in the design of our results protein. Even with these terms and the terms we did not implement, the model for energy is not complete and many researchers use ad hoc models, mixing directly physical terms with Bayesian terms based on PDB statistic, weighted according to what has worked best in the past. In all simulations, the van der Waals potential is the dominant term, and valid as an approximation.

DEE Algorithm Implementation

The first DEE elimination criterion, proposed by Desmet et al., has the form:

$$E(c)\text{-}E(c') + \sum_{j=1..p, j\neq r} \min_{js} E(c, j_s) - \sum_{j=1..p, j\neq r} \max_{js} E(c', j_s) > 0 \quad (7)$$

Where *c* and *c'* are query and comparison rotamer clusters over a residue cluster *r*, respectively, i.e., a query cluster *c* can be eliminated if the minimum energy it obtains by interacting with the conformational background is larger than the maximum possible energy of the comparison cluster can have. A further improvement of the selection criteria (Goldstein 1994), called the "Goldstein criteria" is

$$E(c)\text{-}E(c') + \sum_{j=1..p, j\neq r} \min_{js} [E(c, j_s) - E(c', j_s)] > 0 \quad (8)$$

That is, the conformational background is obtained by fixing each residue in the conformation that most favors *c* relative to *c'*.

In this project, we applied "Goldstein criteria". For the purpose of protein core design, only hydrophobic residues are considered (V, I, L, F, A, and W). At each core position of the protein of interest, the energies $E(c)$ and $E(c')$ are evaluated against the complete input structure, which includes the backbone structure as well as non-core side chains. $E(c, j_s)$ and $E(c', j_s)$ are interaction energies against the individual rotamers in the conformational background. Given the limitation of storage and speed, we first executed DEE for clusters of single rotamers (cycle.pm), then upon the selected single rotamer clusters constructed the clusters of rotamer pairs (cycle2.pm) and employed a second order DEE.

After the second order DEE iteration converges, an exhaustive search procedure (exsearch.pm, seqcreator.pm, chainrot.pm) was applied to all possible sequences in the reduced search space. For each sequence, the energy of the most stable rotamer conformation was reported. Ranks for sequences were based on these energies.

Results

We examined the performance of the design program in predicting the sequences of natural proteins. Our first test was re-design the core region of the major cold-shock protein (1MJC.pdb) (Fig. 1).

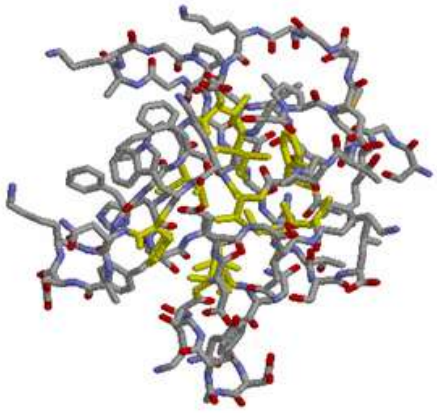

Figure 1. The core region (yellow) of the cold-shock protein provides a backbone scaffold for the sequence design.

Designing the core region reduced the search space to combinations of the rotamers corresponding to hydrophobic amino acids (V, I, L, F, A and sometimes W). Even so, given this set size for five or six hydrophobic sidechain types, and a core consisting of 10-20 positions, a naïve search for an optimal core sequence and structure has significant combinatorial complexity: more than 10^{18} structural solutions exist with roughly 10^{10} sequence combinations. After applying DEE with the rotamer cluster size of one (1st order DEE, script *cycle.pm*), the searching space is clearly reduced (Fig 2). The reduction of the searching space was also seen in a more complex case, a full length protein with 20 residues.

A

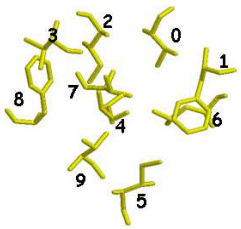

Figure 2: Comparison of the searching space (residues are labeled in A) before (B) and after (C) DEE with rotamer cluster size of one.

B

Residue 0							
A	1						
F	1	2	3				
I	1	2	3				
L	1	2					
V	1	2					
W	1	2	3	4	5	6	7
...							
Residue 3							
A	1						
F	1	2	3				
I	1	2	3				
L	1	2					
V	1	2					
W	1	2	3	4	5	6	7
Residue 8							
A	1						
F	1	2	3				
I	1	2	3				
L	1	2					
V	1	2					
W	1	2	3	4	5	6	7

C

Residue 1							
A	1						
F	1	2					
I	1	2	3				
L	1	2					
V	1	2					
W	1	2	5	6			
...							
Residue 3							
A	1						
F	1	3					
I	1	2	3				
L	1	2					
V	1	2					
W	1	3	7				
Residue 8							
A	1						
F	1	2	3				
I	1	2	3				
L	1	2					
V	2						
W	5	6	7				

The size of the rotamer clusters was extended to two (2nd order DEE, script *cycle2.pm*). By choosing a query cluster consisting of two rotamers associated with two residues, the corresponding comparison cluster consists of a different rotamer pair at the same residue pair. The initial set of the pair-wise rotamers were constructed upon the reduced rotamer space after the 1st order DEE. The pair-wise rotamer cluster set converged after 3 rounds of iterations.

Upon the converged pair-wise rotamer cluster set, a backtracking algorithm (*seqcreator.pm*) constructed all possible sequences given the pair-wise rotamers that are candidates for the GMEC. A total of approximately 3000 sequences, with on average ~200 corresponding rotamer combinations (*chainrot.pm*) were generated for the cold-shock protein core region. Finally, an exhaustive search routine, *exsearch.pm*, was applied to evaluate all the possible sequences above. Each sequence was scored and ranked according to its lowest energy rotamer combination.

Seq.	E_{Score}
N: V F I V V I L V F V	-46.47
1. I F I I I I L I F V	-53.58
2. I F I I V I L I F V	-52.48
3. V F I I I I L I F V	-51.70
4. I F I I I I L V F V	-50.72
5. V F I I V I L I F V	-50.53
6. I F I I V I L V F V	-49.63
7. I F V V I I L I F V	-49.34
8. V F V I I I L I F V	-49.23
9. I F V I I I L V F V	-48.92
10. V F I I I I L V F V	-48.88
. . .	

Fig.3 The native (red) and the top 10 Predicted sequences for the cold-shock protein core. The residues that are identical to the native sequence are shaded.

The top 10 lowest energy sequences are shown in Figure 3. Though only Van de waals potential was considered in the energy scoring function, the predictions are quite promising: the predicted sequences are 50% identical to the native one. The small range of values for total energies indicate the subtlety involved in determining an optimal core sequence. Experimental studies have shown that, for a given protein, many core sequences may exist that yield stably folded functionally active proteins (Lim & Sauer, 1989; Richards & Lim, 1993). Thus, the native sequence is not expected to be predicted exactly. However, sequences similar to the native are expected if the program is to be useful for protein design.

Though reasonable sequence prediction is obviously important, correct prediction of the spatial orientation of the core residues is also expected if our methodology is to be confirmed. In Figure 4 we show the predicted structure of the lowest energy core sequence for the major cold shock protein compared to the crystal structure of the native protein. While this particular model has three sequence differences from the native, it is difficult to distinguish the structures. Experimental studies of core variants have shown that in many cases the orientations of mutated side chains are preserved even when the sequence is altered (Baldwin et al., 1993).

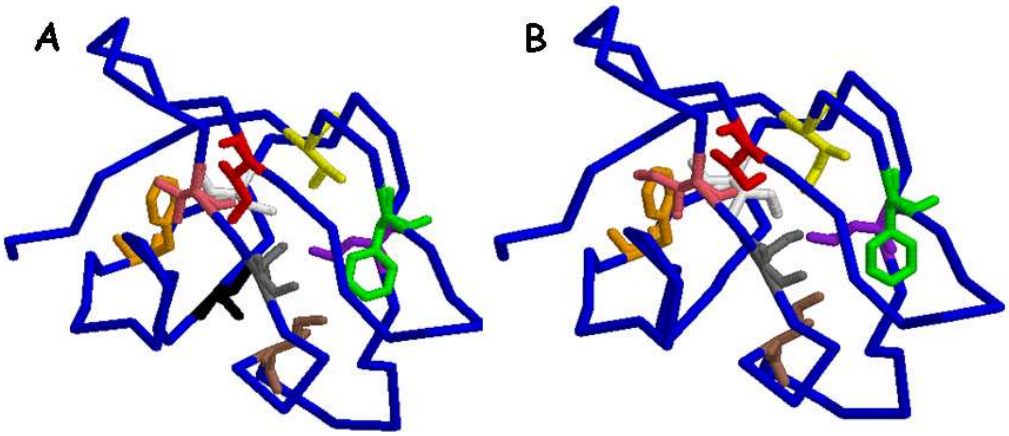

Figure 4: Comparison of the core structure in wildtype(A) and re-designed (B) cold shock protein. The lowest energy sequence in Fig. 3 is used for the re-designed core structure. Residues involved in the core are color coded: 9(yellow), 12(green),21(red), 30(pink),32(gray),37(brown), 45(purple),51(white), 53 (orange), 67 (black).

As an additional test of the sequence prediction, we performed the predictions with the Rosetta *ab inito* program which utilizes Monte Carlo simulations to explore the low energy sequences. The re-designed sequence is almost the same with the native one except an Ala replaces the Leu at 7th position. Considering the scoring function we applied mainly focus on the core packing (Van de waals potential), the lower rank of the native sequence in our prediction indicates that other interactions besides packing, such as solvations energy of burying a certain residue, may be important for a protein core formation and needs to be considered in our future work.

Discussion

Applying 1st and 2nd order DEE in this project dramatically decreased the computational costs required for searching sequence and rotamer conformational spaces. Nevertheless, the searching speed can be further enhanced by 1) pre-ranking the rotamer clusters according to their elimination likehood (Gordon & Mayo 1998) starting with higher elimination possibilities; 2) comparison cluster focusing: a rotamer cluster *c* can be eliminated if the smallest rotamer super-cluster containing *c* are eliminated (Looger 2001), and 3) the stronger, generalized elimination criteria: the residues used to compare the query and comparison clusters are added to the conformational background in groups instead of singly (Looger 2001). Above modifications would be employed in future work.

Despite the success of DEE in protein design seen in this work and elsewhere (Desjarlais 1998), like all pruning-type algorithms, the implementation of DEE requires the use of discrete representations of the backbone and sidechains. In addition, it is restricted to energy terms that can be written as the sum of individual and pair-wise energy terms. In some cases, these limitations might be overly restrictive for the problem at hand, necessitating use of other sampling methods, such as MD (molecular dynamic simulation), MC (Monte Carlo) or GA (genetic algorithms). For example, one could apply more rigorous energy calculations and allow small-scale conformation rearrangement via MD simulations to achieve the true energy minima of a certain conformational state rooted from the rotamer library. Another possible improvement is including an ensemble of backbone structures in

the design process. In vivo, the native protein is subject to dynamic relaxations; in other work, allowing backbone flexibility has been shown to considerably improve the prediction accuracy (Harbury 1998, Larson 2002).

Conclusion

We developed a protein design program that implements DEE algorithm. The result that the core sequences and structures are well predicted for the cold shock protein implies that the program could indeed be applied to the de novo design of a protein core. Future improvements of the design program will involve using more efficient elimination strategies and considering the flexibility of the target backbones.

References

1) Baldwin, EP; Hajiseyedjavadi, 0; Baase, WA; & Matthews, BW. 1993. *Science* 262:1715-1718.
2) Desjarlais, JR & Clarke, ND. 1998.*Curr. Opin. Struct. Biol.* 8, 471-475.
3) Desmet, J; De Maeyer, M; & Hazes, B. *Nature* 356,539-542. 1992.
4) Gordon, DB, Marshall SA, Mayo SL. 1999. *Curr. Op. In Struct. Biol.* 9, 509-513.
5) Gordon, DB & Mayo, SL. 1999. *Structure Fold. Design*. 7, 1089-1098.
6) Hellinga, HW. 1998. *Nature* Struct. Biol. 5, 525-527.
7) Jones, DT. 1994. *Protein Sci.* 3, 567-574.
8) Koehl, P & Delarue, M. 1994. J. *Mol. Biol.* 239, 249-275.
9) Kuhlman B, Baker D. 2000. *Proc Natl Acad Sci USA*. 97, 10383-8.
10) Lazaridis, T. & Karplus, M. 1999. *Proteins: Struct. Func. Genet.* 35, 133-152.
11) Lim, WA & Sauer, RT. 1989. *Nature* 339:31-36.
12) Looger, LL & Hellings, HW. 2001. *JMB*, 429-445.
13) Neria, E., Fischer, S. & Karplus, M. 1996. *J. Chem. Phys*. 105, 1902-1921.
14) Park, S; Stowell , XF; Wang, W; Yang, X; & Saven, JG. 2004. *Annu. 10. Rep. Prog. Chem.* Sect. C, 100, 195–236.
15) Ponder JW, Richards FM. *J Mol Biol.* 1987 Feb 20;193(4):775-91.
16) Street, A. G. & Mayo, S. L. 1999. *Structure Fold. Design*. 7, R105-R109.
17) Richards FM, Lim WA. 1993. *Q Rev Biophys* 26:423-498.
18) http://honiglab.cpmc.columbia.edu/programs/sidechain