Designing a Peptide Ligand with Dead End Elimination

Kyle Roberts

4/21/08

1 Introduction

In protein redesign one of the main challenges is the ability to evaluate many different conformations of the redesigned protein in order to determine the best structure. One approach to this problem is to provably get rid of bad protein conformations using dead-end elimination (DEE). DEE was first developed for application to protein structure evaluation by Desmet et al. [1]. Since then there have been many advances in both the pruning criteria and its applications. My project aim is to extend Donald Lab DEE algorithm implementation in order to design peptide ligands. Specifically, I apply my extension to the cystic fibrosis transmembrane conductance regulator (CFTR) and cystic fibrosis associated ligand (CAL) system.

2 Biological Background

Cystic fibrosis is caused by a single amino acid deletion in CFTR. CFTR is a transmembrane chloride channel that is located on the apical surface of epithelial cells. The mutation in CFTR leads to its rapid degradation and a lack of expression in the cell membrane. Mutated CFTR is still able to function, but is degraded before it can get inserted into the membrane. It has been hypothesized that a possible treatment for cystic fibrosis would be to restore CFTR insertion into the membrane [2].

PDZ domains are ubiquitous peptide-binding modules that mediate protein-protein interactions involved in cell trafficking and localization. Specific PDZ
domain containing proteins, including CAL, have been shown to interact directly or indirectly with CFTR. Additionally, overexpression of CAL has been shown to lead to the degradation of CFTR in a dose-dependent fashion. Wolde et al. further showed through RNA interference that knockdown of CAL can lead to increased cell-surface expression of mutant CFTR [2](Fig 1). This is very exciting because it means that it might be possible to treat cystic fibrosis by inhibiting CAL. Thus, the goal of this project is to design a ligand that will bind to the PDZ domain of CAL.

3 Design/Computational Approach

In order to design a ligand to bind CAL, the NMR structure of CAL bound to the C-terminus of CFTR is considered. Mutations to the known ligand are computationally explored and then ranked by a pairwise energy function. In practice, the number of mutations to be considered leads to a search space which is too large to computationally enumerate. To make the search
tractable the space is first discretized using a set of rotamers. Rotamers are discrete conformations that side chains can adopt. Next, dead-end elimination is used in order to prune all possible structures that could not possibly be part of the global minimum energy conformation (GMEC). Finally, the A\(^*\) algorithm is used to enumerate the remaining conformations in order from lowest to highest energy.

The Donald lab has already implemented these design algorithms and that implementation was used in this project \[4\]. The Donald lab implementation only functions for ligands of amino acid length one. This project required that the ligand length be eight, so the Donald lab code was extended to incorporate this feature.

### 3.1 Dead End Elimination

Dead End Elimination is an algorithm that is used to prune rotamers from a proposed protein structure that cannot possibly be part of the GMEC. DEE uses a pruning criteria in order to eliminate these rotamers. The Donald lab code incorporates four different pruning criteria.

**Simple Goldstein:** The simple Goldstein criterion states that a given rotamer \(i_r\) can be pruned from the search space if there is another rotamer \(i_t\) such that \(i_t\) always contributes a lower energy than \(i_r\). This can be seen in Fig 2A where \(i_{t1}\) and \(i_{t2}\) always contribute a lower energy than \(i_r\), so \(i_r\) can be pruned from further consideration.

**Goldstein Pairs:** This criteria is similar to the simple Goldstein criterion, but instead of a single rotamer getting pruned, a pair or rotamers can be pruned if there is always another pair of rotamers that contribute a lower energy.

**Conformational Splitting:** The conformational splitting criterion first partitions the conformation space into \(k\) partitions. A rotamer \(i_r\) can be pruned if for each partition there is a rotamer \(i_{tk}\) such that \(i_{tk}\) always contributes a lower energy to the conformation than \(i_r\). This is shown in Fig 2B. Here the conformation space is divided into two partitions and \(i_{t1}\) is always lower than \(i_r\) in the first partition, and \(i_{t2}\) in the second partition. Thus, \(i_r\) can be pruned by the splitting criteria. Note, that \(i_r\) cannot be pruned with the Goldstein criterion using either \(i_{t1}\) or \(i_{t2}\).

**DACS:** The DACS criterion partitions the conformational space similar to the conformational splitting criterion. The DACS criterion saves information that conformational splitting throws away because, if a rotamer cannot
be pruned from all partitions in DACS, the rotamer is pruned from all the partitions it can be. Then each partition is analyzed separately and the local GMEC for each partition is analyzed to find the overall GMEC. This can be seen in Fig 2C. Here \( i_r \) can be pruned from \( P_1 \) but not \( P_2 \). Conformational splitting would not prune \( i_r \) at all, but DACS prunes \( i_r \) from \( P_1 \) and keeps \( i_r \) in \( P_2 \).

3.2 \( \text{A*} \)

After all the conformations are pruned by DEE that can be, \( \text{A*} \) is used to enumerate the remaining conformations. The \( \text{A*} \) search algorithm deterministically identifies the least-cost path in a tree from the root to a leaf node [5]. This tree can be constructed to represent the protein conformations not pruned by DEE that need to be enumerated. The path costs of the tree can be designed to represent the costs of incorporating that rotamer into the structure. Thus, \( \text{A*} \) is used to traverse the lowest cost path in order to determine the lowest energy structure [5].

4 Results

One main result of this project was gaining more understanding of the DEE and \( \text{A*} \) implementation. During this project there were significant problems relating to the use of the energy function and \( \text{A*} \) enumeration. This required me to further understand these areas of the software implementation in order to diagnose the problem.

Specifically, I addressed problems related to deciding when to evaluate the energy of different atoms. In the original code, when the template energy is calculated the ligand is not considered and is merely deleted from the calculation at that time. In order for the ligand to be allowed to mutate I had to include the backbone atoms in the template energy calculation. This caused a double counting of the ligand energy that had to be addressed but was ultimately solved.

In addition, I learned that the energy function is very sensitive to atom position. When a residue is mutated to a different type, it is ambiguous how to align the new residue at the given position. Using the \( C\alpha-C\beta \) vector was originally used, but required a special case for glycine. In the later versions of the code the use of the \( C\alpha-C' \) vector was explored. Interestingly, when this
Figure 2: DEE criteria. (A) Simple Goldstein criteria. (B) Conformational Splitting. (C) DACS. See text for more description. (A) and (B) taken from [3] and (C) taken from [4]
vector was used it resulted in slightly different atomic coordinates than the first method. While the differences in atomic coordinates were minor, when summed over the entire protein these changes caused drastic changes in the algorithm’s solutions.

Before applying the algorithm to the CFTR-CAL system, the previous design of phenylalanine activating domain of gramicidin synthetase (pdb #:1AMU) was used as a test case. The previous results were compared to the results from the extended code from this project in order to confirm that the code was still behaving as desired. The extended code was able to reproduce the results from the previous design.

The algorithm was then applied to the CFTR-CAL system. As a preliminary analysis, the design algorithm was run to compute the pairwise energy matrix considering all 8 ligand residues as mutable positions. Test DEE runs were conducted for allowing all 8 residues to mutate to any amino acid. Unfortunately, there is very little pruning for the end of the ligand. It seems that there are not enough nearby atoms in order to prune any of the rotamers that are on the end of the ligand (Fig 3). Since this left many rotamers that had to be considered by A*, the DEE process was fairly slow for all 8 residues. By incorporating biological information some possible mutations can be eliminated.

Since PDZ domains bind the motif x(S/T)x(I/L/V) the first ligand amino acid was restricted to (I/L/V) mutations and the third ligand was reduced to either serine or threonine. In addition to these restrictions, the last 3 residue positions were not allowed to mutate in order to speed up the analysis process for early runs. Sufficient pruning was obtained with these mutation limitations such that DEE runs finish in minutes.

DEE runs were conducted allowing no minimization, only side-chain minimization or only backbone minimization. Table 1 shows the energies for the minimized structures obtained from each method.

The energies from all of these methods are very similar. I should note that the redesigned ligand is missing its C-terminal O (it is not included in the redesign) which in the original structure forms an important H-bond. This is being worked on, but is not currently implemented in the redesign process.

The sequences found by each method are very similar. Interestingly, the original NMR structure has the best energy yet none of the methods rank the wild-type structure at the very top of their sequences. Upon closer inspection of the GLN to GLU mutation, the GLN rotamer in the original structure is
Figure 3: NMR structure of CAL bound to the C-terminus of CFTR. The C-terminus of CFTR is shown in orange.

<table>
<thead>
<tr>
<th></th>
<th>Sequence</th>
<th>DEE Run Energy (Kcal/mol)</th>
<th>Built PDB Energy (Kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMR Structure</td>
<td>E E V Q D T R L</td>
<td>NA</td>
<td>-230.69</td>
</tr>
<tr>
<td>No Minimization</td>
<td>E E V R S R L</td>
<td>-226.7</td>
<td>-230.44</td>
</tr>
<tr>
<td>Side-Chain Min</td>
<td>E E V E R T R L</td>
<td>-230.8</td>
<td>-229.84</td>
</tr>
<tr>
<td>Backbone Min</td>
<td>E E V E R T R L</td>
<td>-232.4</td>
<td>-229.84</td>
</tr>
</tbody>
</table>

Table 1: Ligand amino acid sequence predictions and their associated energies. The DEE run energy represents the energy that was obtained using either no minimization, side-chain minimization or backbone minimization during the DEE run. The Built PDB Energy is the energy of the generated structure after it was rebuilt and the whole structure was allowed to minimize.
Figure 4: Gln rotamer. This shows the rotamer for GLN (all red) in the library overlaid on top of the original CAL structure. Note that the H’s on the rotamer are coming off the N such that the rotamer is 180 degrees away from the actual side chain position.

not represented in the rotamer library. The N and O on the GLN are actually 180 degrees rotated ($\chi_3$ angle) from the rotamer that is present in the rotamer library (Fig 4). Thus, this shows the reliance that DEE has on the initial rotamer library. Looking at the ASP to ARG mutation, seems to reveal the same reasoning for this mutation. The ASP in the rotamer library has the O-O vector perpendicular to that in the original structure. In addition, the mutated ARG makes some long range contacts that the ASP can’t reach.

Through analysis in KiNG using probe dots, it seems that the redesigned ligand forms stronger $\beta$-sheet contacts than the original structure. Although, the ARG and LEU side-chain moities form more H-bonds in the original structure than in the redesigned structure.

In addition to redesigning the ligand, this project aimed to generate testable data that could be compared to experimental measurements. Binding measurements have been made in vitro for several CAL ligands. In order
to compare my results to the experimental results I ran DEE runs where the sequence was fixed but the rotamers were flexible. Thus, we would hope that the predicted energies from these runs should be able to correlate with the experimental data (Of course this might take a lot of work and we would most likely need to account for entropy and a better unbound structure before these results resembled true experimental data.)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>No Min (Kcal/mol)</th>
<th>SideChain Min (Kcal/mol)</th>
<th>Backbone Min (Kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CANGLMQTSKL</td>
<td>-199.39137</td>
<td>-201.30228</td>
<td>-200.16449</td>
</tr>
<tr>
<td>CHRTCYLVTLQ</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>CREKKFSTIL</td>
<td>-9.269205</td>
<td>-180.37839</td>
<td>-177.12985</td>
</tr>
<tr>
<td>CEVYLLNSTTL</td>
<td>-199.80173</td>
<td>-203.13985</td>
<td>-203.13036</td>
</tr>
<tr>
<td>CDGQMQLVTSL</td>
<td>-200.93214</td>
<td>-204.55664</td>
<td>-201.99214</td>
</tr>
<tr>
<td>CLPHPHSTTRV</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>CAPIAGFSSFV</td>
<td>-170.36308</td>
<td>-181.18483</td>
<td>-183.30212</td>
</tr>
<tr>
<td>CHPHSHSTTRV</td>
<td>-174.7577</td>
<td>-188.09058</td>
<td>-189.74332</td>
</tr>
<tr>
<td>CQPRTLQWSPV</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>CGNDPDRGTSI</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>CYHKGVSLSHV</td>
<td>-182.35423</td>
<td>-190.55286</td>
<td>-195.38289</td>
</tr>
<tr>
<td>CRPGFASESKV</td>
<td>-181.88568</td>
<td>-189.05798</td>
<td>-188.39902</td>
</tr>
<tr>
<td>CTEEEVDTRVL</td>
<td>-217.92957</td>
<td>-223.76346</td>
<td>-224.73553</td>
</tr>
<tr>
<td>CALDTSHTGTSI</td>
<td>-180.91606</td>
<td>-185.95477</td>
<td>-190.28566</td>
</tr>
<tr>
<td>CKKMPSESIDV</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 2: Predicted energies of ligand protein complexes for experimentally validated sequences. If the energy is NA that means that all the rotamers were pruned and an output structure was not generated.

5 Discussion

These initial predictions do not seem to be that promising. The minimum energy structure from the different methods all have sequences that are very similar to the wild-type. In addition, they all seem to have very similar energies suggesting that they would be mediocre competitive inhibitors. However, it is possible that redesigning the 6th, 7th, and 8th positions in the ligand will bring about new possible enhancements and specificity that could increase...
the binding of the designed ligand. Also, including the terminal oxygen on the ligand could lead to better ligand designs as well.

There are still many improvements that could yield better design results. First, increasing backbone flexibility could improve the ligand design. Ligands are generally much more flexible than proteins, so being able to incorporate increased flexibility would yield more ligand conformations and more possible ligand designs. Since ligands are more flexible it would also be helpful to possibly extend the rotamer library or at least update it, since ligands could possibly have more of the lower population rotamers as compared to protein side chains.

Incorporating non-natural amino acids could also improve the ligand design process. Including D-amino acids or other non-natural amino acids would not only improve the possible designs that could make up a good binding ligand, but are also more resistant to proteases.

The aim of this project was to extend a DEE implementation to incorporate ligand design. The code was successfully extended, and the design was applied to the CFTR-CAL system. Currently, the success of the results are unclear. The designed sequences are very close to the wild-type sequence and the predicted energies are almost identical. This suggests that these should be able to bind to CAL, but needs experimental validation to confirm. Whether the calculated energies correlate with the collected experimental binding energies will further support or reject the current designs.

References


