Graph Cuts and Protein Design

Adi Mukund Jennifer Zou

1 The Protein Design Problem

The protein design problem can be described as follows. Given a set of backbone coordinates $\mathbf{c} = (\vec{c_1}, \vec{c_2}, \ldots, \vec{c_n})$ and rotamer library R of length r, identify the optimal rotamer assignment sequence $\vec{r} = (r_1, r_2, \ldots, r_n)$, $r_i \in R$, $1 \leq i \leq n$ according to an energy function $E(\vec{r})$. This assignment sequence is known as the global minimum energy conformation (GMEC).

2 Graph Labeling

The graph-cut formulation of the protein design problem represents the set of all residues as a sparse interaction graph, a modification of the residue interaction graph $G = (V, \mathcal{E})$ in which each vertex represents a protein residue and each edge represents an interaction between two such residues. The sparse graph is created by omitting a set of interactions $\mathcal{E}^*$ to create the graph $G^* = (V, \mathcal{E} - \mathcal{E}^*)$.

The label graph G_l is generated by adding a set of vertices V_r representing all of the rotamers in the rotamer library R to the vertex set V^* of G_* and connecting each vertex in V^* to each vertex in V_r to generate the vertex set V_l and corresponding edge set $\mathcal{E}_l$. The set of all edges connecting vertices in V_r and V^* is $\mathcal{E}_r$. Edges from $\mathcal{E}_l$ are then pruned until every vertex in V^* is connected to exactly 1 vertex in V_r (that is, until every member of V^* is represented in $\mathcal{E}_r$ exactly once) in order to generate a labeling for the G where each vertex in V^* is connected to a vertex in V_r representing its optimal rotameric assignment.

The energy of the graph can then be computed by finding

$$E_G = \sum_{r \in V_r} \sum_{v \in V^*} \phi(r, v) + \sum_{v \in V^*} \sum_{\substack{w \in V^* \\ w \neq v}} \psi(v, w) \tag{1}$$

where $\phi(r, v)$ represents the cost of assigning label r to vertex v (intuitively, the internal energy of residue v when assigned rotamer r) and $\psi(v, w)$ represents the pairwise interaction between vertices v and w given an initial labeling.

Thus, finding the GMEC is equivalent to minimizing equation 1 over the set $\mathcal{L}$ of all possible labelings.

3 ILP Transformation

4 Approximation Bounds