Graph-Based Representations of the Protein Design Problem

Adi Mukund Jennifer Zou

1 Modified cluster expansion

Cluster expansion (CE) can be used to approximate the energy of proteins from sequence features or cluster functions (CFs). The algorithm fits a set of coefficients called effective cluster interaction (ECI) values that reflect the energetic contribution of a CF. After defining a set of candidate CFs, an iterative algorithm called CLEVER can be used to determine which CFs contribute significantly to total energy.

Even though low-order interactions between residues are assumed to contribute more to total energy, high-order interactions can affect prediction accuracy of CE. [**?**] However, inclusion of many high-order interactions in CE is computationally intense and may lead to overfitting.

The triplet option in CLEVER considers all triple cluster functions (CFs) that are spanned by three pair CFs. [**?**] This filter chooses triplets CFs that are collinear with the pair CFs and may neglect important interactions. Rather than choosing triplet CFs from the pair CFs, we hope to use strongly connected components in sparse residue interaction graphs to create high-order interaction terms that match the optimal substructure of graphs. Alternatively, terms can be generated by finding articulation points within branch decompositions of the sparse residue graphs. Residues within subgraphs when articulation points are removed could potentially have significant high order interactions.

The size of strongly connected components in sparse residue graphs is generally too large to include in cluster expansions. The distribution could be shifted by sparsifying the graphs, but this may introduce additional bias. Large strongly connected components could potentially be broken up into several CFs, but it is difficult to determine how this will impact the accuracy the CE.

Tasks for the remaining of the semester include

1. Implement branch decomposition
2. Run cluster expansion with new interaction terms
3. Compare results with CLEVER triplet option

References

[1] Grigoryan G, et al. Ultra-fast evaluation of protein energies directly from sequence. PLoS Computational Biology 63,0551-0563 (2006).

[2] Hahn S, et al. Identifying and reducing error in cluster-expansion approximations of protein energies. Wiley Online Library (2010).

[3] Negron C, Keating A. Multistate Protein Design Using CLEVER and CLASSY. Methods in Enzymology,Vol. 523.