Bounding A Protein's Free Energy In Lattice Models Via Linear Programming

Robert Carr[1], **William E. Hart**[1], **Alantha Newman**[2]

Keywords: Protein structure prediction, linear programming, lattice models, HP model

1 Introduction

The established HP lattice 2D and 3D models have been useful abstractions in understanding protein structure prediction. In these models, a protein folds to maximize H-H contacts (minimize free energy). We analyze and compare integer programming models for the 2D lattice, whose linear relaxations provide non-trivial upper bounds on the maximum number of contacts. These bounds can be used in a branch-and-bound approach to solve the problem optimally and could potentially be used to obtain improved approximation algorithms. In particular, we seek to beat the simple combinatorial bound that arises from the lattice being bipartite.

2 Problem formulation

The Hydrophilic-Hydrophobic (HP) model, introduced by Dill [4], abstracts the dominant force in protein folding: the hydrophobic interaction. The hydrophobicity of an amino acid measures its affinity for water, and the hydrophobic amino acid residues of a protein form a tightly clustered core. In the HP model, each amino acid is classified as an H (hydrophobic) or a P (hydrophilic). The model further simplifies the problem by restricting the feasible foldings to the 2D or 3D square lattice. An optimal conformation for a string of amino acid residues in the HP model is the one that maximizes the number of H-H contacts, which are formed by pairs of H's that occupy adjacent lattice points but are not adjacent on the string.

3 Our approach

We discuss discrete optimization approaches to the problem of protein folding in the Hydrophobic-Hydrophilic (HP) model. We formulate several different integer programs for the problem of protein folding in the 2D HP model and compare the relative strengths of their respective linear programming relaxations. One way to measure the quality of an integer program for a maximization problem is to determine the upper bound guaranteed by its linear relaxation. A linear programming relaxation provides an upper bound on a maximum integral solution and can be solved much more efficiently than an integer program. In general, the tighter (better) the bound provided by the linear relaxation, the higher the quality of the integer programming formulation.

Such methods have been posed previously as a potential approach to protein folding in lattice models [3, 5]. However, the strengths of the proposed LP relaxations were not addressed. For example, we prove that the linear programming relaxation for a natural integer program (described in [3]) provides a solution with value at least twice as much as

[1]Discrete Algorithms and Math Department, Sandia National Laboratories, Albuquerque, NM. E-mail: `rdcarr, wehart@cs.sandia.gov`

[2]CSAIL, MIT, Cambridge, MA. E-mail: `alantha@theory.lcs.mit.edu`

the simple combinatorial upper bound for *every* string. However, a strengthened version of this linear program with *backbone constraints* provides a bound that is provably no worse than the simple combinatorial upper bound. We propose additional constraints that may further strengthen these linear programs.

4 Experimental results

For our experiments, we used benchmarks for the problem in the 2D HP model that were taken from: www.cs.sandia.gov/ tech_reports/compbio/tortilla-hp-benchmarks.html. We ran one of our linear programs (LP_3 in [2]) on the following strings:

1. `hphpphhphpphphhpphph`
2. `hhpphpphpphpphpphpphpphh`
3. `pphpphhpppphhpppphhpppphh`
4. `ppphhpphhpppppphhhhhhhpphhpppphhpphpp`
5. `pphpphhpphhpppppphhhhhhhhhhpppppphhpphhpphpphhhhh`
6. `hhhpphphphpphphphpph`

String	length	upper bound	LP_3	Opt
1	20	11	10.67529996	9
2	24	11	11	9
3	25	8	8	8
4	36	16	14.89908257	14
5	48	25	24.88770748	22
6	20	11	10.76264643	10

5 Discussion

The challenge that we introduce here is to compute better upper bounds for the 2D folding problem using linear programming or otherwise. Our integer and linear programming models provide a promising direction for solving the 2D folding problem to optimality using branch-and-bound. However, because of the large size of the linear program (i.e. number of variables), we likely need tighter linear programming bounds to make these techniques practical.

Another possible application of our integer and linear programming formulations is to find actual foldings that are better than those obtained in approximation algorithms but perhaps not provably optimal. Backofen has used exact methods from constraint logic programming to obtain compact conformations, i.e. solutions, for these folding problems [1]. If we can further constrain our integer programs to the solution space of compact foldings, then we may be able to reduce the time needed to find a solution.

References

[1] Rolf Backofen, "Optimization Techniques for the Protein Structure Prediction Problem", *Ph.D. Thesis, Ludwig-Maximilians-Universität München* (2000).

[2] Robert Carr, William E. Hart, and Alantha Newman, "Discrete Optimization Models for Protein Folding", Technical Report, Sandia National Laboratories, 2003.

[3] V. Chandru, A. DattaSharma, and V. S. A. Kumar, "The algorithmics of folding proteins on lattices", *Discrete Applied Mathematics* (2003) Vol. 127(1):145-161.

[4] K. A. Dill, "Dominant Forces in Protein Folding, *Biochemistry* (1990) Vol. 29:7133-7155.

[5] H. J. Greenberg, W. E. Hart, and G. Lancia, "Opportunities for Combinatorial Optimization in Computational Biology", *INFORMS Journal of Computing, To appear.*