Protein Design by Sampling an Undirected Graphical Model of Residue Constraints

John Thomas[1], Naren Ramakrishnan[2], Chris Bailey-Kellogg[3]

Protein engineering seeks to produce amino acid sequences with desired characteristics, such as specified structure [1] or function [4]. This is a difficult problem due to interactions among residues; choosing an amino acid type at one position may constrain the possibilities at others, in order for the resulting protein to have proper structure and activity. To account for the dependence of some residues and take advantage of the independence of others, we have developed a new approach to protein design based on undirected probabilistic graphical models (Fig. 1). Our approach first constructs a graphical model that encodes residue constraints, and then uses the model generatively to produce new sequences optimized to meet the constraints. We focus here on constraints due to residue *coupling*, common pairs of amino acid types at particular pairs of positions, also known as correlated mutations or co-evolving residues. Recently, Ranganathan and colleagues showed that accounting for residue coupling, in addition to conservation, was to some extent both necessary and sufficient for viability of new WW domains [6, 5].

We have previously developed an approach for learning an undirected graphical model encapsulating conservation and coupling constraints in a protein family [7]. Our model provides a formal probabilistic semantics for reasoning about amino acid choices, defining a probability distribution function measuring how well a new sequence satisfies coupling constraints observed in the extant sequences of a family. Thus in order to design high-quality novel sequences, we can optimize for their likelihood under the model. Furthermore, our model explicates dependence and independence relationships between residue positions, so that we may reason about the impact of an amino acid choice at one position on those at others.

While sampling from an undirected model is difficult in general, we have developed two complementary algorithms that effectively sample the constrained sequence space. *Constrained shuffling* generates a fixed number of high-likelihood sequences that are reflective of the amino acid composition of a given family. A set of shuffled sequences is iteratively improved so as to increase their mean likelihood under the model. *Component sampling* explores the high-likelihood regions of the space and yields a user-specified number of sequences. Sequences are generated by sampling the cliques in a graphical model according to their likelihood, while maintaining neighborhood consistency. In contrast to the approach used by Ranganathan and colleagues, which simply seeks to reproduce the aggregate degree of coupling without regard to the quality of the individual sequences, our methods utilize a graphical model to generate sequences that meet the observed constraints, thereby improving the chances the designed sequences will be folded and functional. Theoretical results show that both of our methods properly sample the underlying sequence distribution.

We have applied our sampling algorithms in a study of WW domains, small proteins that assist in protein-protein interactions by binding to proline rich targets. We first showed that likelihood under our graphical model, trained on 42 wild-type WW domains, is predictive of foldedness for the new sequences designed by Ranganathan and colleagues, achieving a classification power of 0.8. We then generated novel putative WW domains optimizing the predicted likelihood. Both methods generated sequences with likelihoods near those of the wild-type WW domains, while being relatively novel and diverse (Fig. 2). The designed sequences serve as hypotheses for further biological study.

Our learning and sampling methods are applicable to a wide variety of protein families that may be targets for protein design. While multiple sequence alignments provide fundamental information on sequence constraints, our models may also incorporate additional structural or functional information. By including functional subclass information [7], we can design proteins with specific functional properties. By incorporating energetic constraints on side-chain interactions [2] we can design proteins with favorable predicted free energy.

[1]Department of Computer Science, Dartmouth College, Hanover, NH, USA. E-mail: jthomas@cs.dartmouth.edu
[2]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA. E-mail: naren@cs.vt.edu
[3]Department of Computer Science, Dartmouth College, Hanover, NH, USA. E-mail: cbk@cs.dartmouth.edu

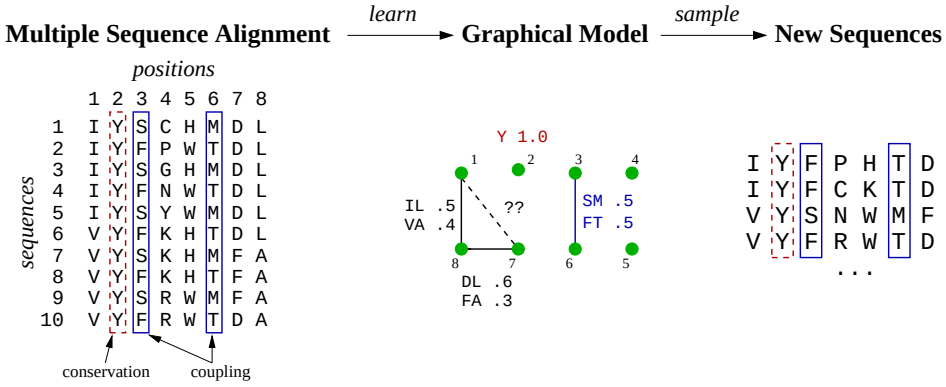

Figure 1: Given a multiple sequence alignment for a protein family (left), conservation and coupling constraints are inferred and summarized into a graphical model (middle) which captures conditional independence relationships through its edges. Through its clique potentials (not shown here), the model captures probability distributions for subsets of residues. Sampling from the model (right) then yields new sequences that obey the underlying constraints.

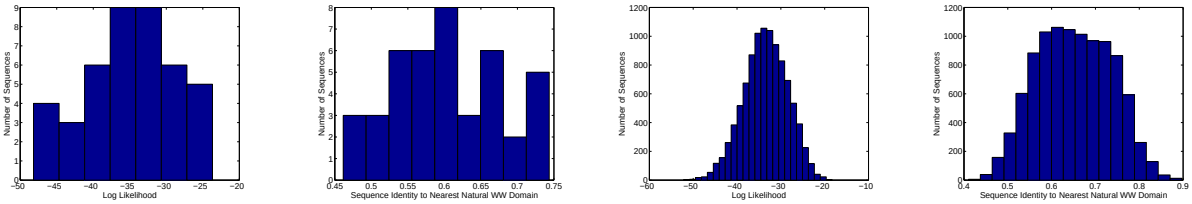

Figure 2: The log likelihood distribution (a, c), and sequence identity to the nearest natural WW domains (b, d), for the 42 and 10000 sequences generated by constrained shuffling and component sampling, respectively. The average log likelihood scores for the designed sequences are -34.69 with a standard deviation of 6.46 (constrained shuffling) and -33.48 with a standard deviation of 5.02 (component sampling). The wild-type WW domain sequences have an average log likelihood score under the model of -32.65 with standard deviation 4.93.

References

[1] B.I. Dahiyat and S.L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, Oct 1997.

[2] H. Kamisetty, E.P. Xing, and C.J. Langmead. Free energy estimates of all-atom protein structures using generalized belief propagation. In *Proc. RECOMB*, pages 366–380, Apr 2007.

[3] S.W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, Oct 1999.

[4] L.L. Looger, M.A. Dwyer, J.J. Smith, and H.W. Hellinga. Computational design of receptor and sensor proteins with novel functions. *Nature*, 423(6936):185–190, May 2003.

[5] W.P. Russ, D.M. Lowery, P. Mishra, M.B. Yaffee, and R. Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437(7058):579–583, Sep 2005.

[6] M. Socolich, S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, Sep 2005.

[7] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Graphical models of residue coupling in protein families. *IEEE Trans. Comp. Biol. and Bioinf.*, 2007. In press. Preprint available at: `http://www.cs.dartmouth.edu/~cbk/papers/tcbb07.pdf`.