Mean-Field Approaches to Protein Evolution

Kevin S. Brown[1], Christopher A. Brown[2]

***Short Abstract* — There are a variety of methods that attempt to infer networks of positional correlations in proteins from multiple sequence data. Recently, promising results have been achieved by using the q-state Potts model to describe the probability distribution of symbols in the sequences. Mean-field techniques can then be used to compute site-site couplings given observed site-site correlations. We compare four mean-field methods for protein contact prediction and find that while techniques beyond lowest order appear extremely promising in principle, the number of sequences needed to gain this advantage in practice is well beyond anything currently available in typical multiple sequence alignments.**

I. Introduction

WITHIN an evolving protein, a complex network of amino acid correlations is embedded that drives residue substitutions at single sites. Allosteric protein regulation can be achieved by conformational changes induced by substrate or ligand binding, which can then propagate to distant sites, inducing long-range spatial correlations [1]. In addition, long-range interactions can exist between charged residues [2] or may reflect other modes of long-range energetic coupling in the protein [3].

In order to infer this network of interactions from an alignment of related protein sequences, a variety of methods under the names *correlated substitution analysis* have been developed [4-7]. If one adopts a pairwise probability model constrained by the observed pair correlations, maximum entropy yields the famous Ising model for binary variables [8] and the q-state Potts model for discrete multistate variables [9,10]. In either case, one then wishes to "invert" the model, in which spin-spin couplings are inferred given observed pair correlations. These couplings can be computed to successively higher order using high-temperature (small-correlation) expansions [11,12].

II. Results

We derive four mean-field approximations to the couplings in the inverse q-state Potts model in order to compare their efficacy for correlated substitution analysis. The four methods are: naïve mean-field (NMF), independent pairs (IP), Thouless-Anderson-Palmer (TAP), and Sessak-Monasson (SM). Higher-order methods, particularly SM, perform far better than NMF in recovering known couplings for the q-state Potts model when the correlations are computed by enumeration of the states. Unfortunately, when Monte Carlo data are used to compute the pair correlations, this advantage evaporates unless extremely large numbers of Potts chains are available. For protein data, this means the alignment must consist of an enormous number of sequences, more than will typically be available. We confirm these conclusions using a large set of high-quality alignments (>1200) from the Pfam database [13].

III. Conclusion

For analysis of evolutionary dynamics in proteins, inverse q-state Potts models beyond the naïve mean field approximation are unlikely to generate better predictions of cofluctuating residues. This may not be the case for genetic sequences in which q is much smaller (4 vs. 21 in proteins). However, once reproducibility – essentially statistical robustness [14] – is considered it is possible that higher-order methods may have additional advantages; this remains an open question.

[1]Biomedical Engineering, University of Connecticut, Storrs, CT, 06340. E-mail: kevin.s.brown@uconn.edu
[2]Palomidez LLC, Cambridge, MA 02139. E-mail: chris.al.brown@gmail.com

References

[1] Horovitz A, et al. (2001) Allostery in chaperonins. *J Struct Chem* **135**, 104-114.
[2] Lowenthal R, et al. (1993) Long-range surface charge-charge interactions in proteins: comparison of experimental results with calculations from a theoretical method. *J Mol Biol* **232**, 574-583.
[3] Lockless S, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295-299.
[4] Horner D, Pirovano W, Pesole G (2007) Correlated substitution analysis and the prediction of amino acid structural contacts. *Briefings in Bioinformatics* **9**, 46-56.
[5] Kass I, Horovitz A (2002) Mapping pathways of allosteric communication in groEL by analysis of correlated mutations. *Proteins* **48**, 611-617.
[6] Gobel U, et al. (1994) Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317.
[7] Atchley W, et al. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol* **17**, 164-178.
[8] Schneidman E, et al. (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007-1012.
[9] Weigt M et al. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *PNAS* **106**, 67-72.
[10] Morcos F et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many proteins. *PNAS* **108**, E1293-E1301.
[11] Sessak V, Monasson R (2009) Small-correlation expansions for the inverse ising problem. *J Phys A* **42**, 055001.
[12] Tanaka T (1998) Mean-field theory of Boltzmann machine learning. *Phys Rev E* **58**, 2302-2310.
[13] Finn R et al. (2011) The Pfam protein families database. *Nucleic Acids Res* **38**, D211-D222.
[14] Brown C, Brown K (2010) Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, Oh, My! *PLoS ONE* **5**, e10779.