Ensemble Learning for Correlated Substitution Analysis

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. However, method accuracy is inconsistent from sequence alignment to sequence alignment, depends strongly on sequence preprocessing and method parameters, and the predicted networks from different methods show little overlap. We use ensemble learning to combine the results of multiple scoring methods. When tested on a large set of alignments, the ensemble method outperforms the individual scoring methods in the ensemble.**

I. Introduction

EMBEDDED in an evolving protein is a complex network of amino acid correlations. The constraints induced by this network of correlated fluctuations drive residue substitutions at single sites. Correlations can be strong even between pairs of residues widely separated in the folded structure because of allostery [1], charged interactions [2], or other forms of energetic coupling [3].

In order to infer this correlation network from multiple sequence data, many methods under the names *correlated substitution analysis* have been developed. The methods for scoring pairs of residues for high correlation include chi-squared tests [4], explicit likelihood [5], variants of mutual information [6,7], and maximum entropy models [8].

Unfortunately, method accuracy – as assessed by comparison to protein contact maps – is inconsistent from sequence alignment to sequence alignment and can be strongly dependent on other preprocessing steps and scoring parameters. In addition, predicted networks from different methods often show relatively poor overlap [9].

II. Results

In some machine learning problems, combining several models yields better results than can be achieved by any individual model [10]. We use an ensemble approach to blend the results of multiple correlated substitution scoring methods.

We first convert the set of scores each method assigns into a set of ranks. The ranks are then aggregated, using a metric similar to those employed for web meta-search engines [11]. We combined nine different scoring methods on a large (~ 3500) set of high-quality protein alignments from the Pfam database [12]. The methods used include both newer, more sophisticated methods [7,8,9] and older, simpler ones [4,6]. The ensemble approach shows a marked improvement in scoring accuracy when compared to the individual ensemble members.

III. Conclusion

A large amount of effort in the field of correlated substitution analysis is directed towards developing ever more complicated, and hopefully more accurate, scoring methods. Our mixture-of-experts results suggest that even older, relatively simple methods can still yield impressive predictions when properly blended. Given that many scoring methods are derived from others, in the future it might be desirable to blend the models with a more sophisticated scheme [13].

[1]Biomedical Engineering, University of Connecticut, Storrs, CT. E-mail: kevin.s.brown@uconn.edu

[2]Palomidez, LLC. 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] Kass I, Horovitz A (2002) Mapping pathways of allosteric communication in groEL by analysis of correlated mutations. *Proteins* **48**, 611-617.

[5] Dekker J, Fodor A, Aldrich R, Yellen G (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinormatics* **20**, 1565–1572.

[6] Atchley W, et al. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol* **17**, 164-178.

[7] Little D, Chen L, Shiu S (2009) Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation, and catalytic coordination in protein evolution. *PLoS One* **4**, e4762.

[8] Morcos F et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many proteins. *PNAS* **108**, E1293-E1301.

[9] 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.

[10] Sollich P, Krogh A (1996) Learning with ensembles: how overfitting can be useful. *Advances in Neural Information Processing Systems 8*, 190-196.

[11] Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the web. *Proceedings of the 10th international conference on World Wide Web*, D613-622.

[12] Finn R et al. (2011) The Pfam protein families database. *Nucleic Acids Res* **38**, D211-D222.

[13] Wolpert D (1992) Stacked generalization. *Neural Networks* **5**, 241-259.