Accurate contact prediction in proteins: using pseudolikelihoods with gap term to infer Potts models

Christoph Feinauer[1,4], **Marcin J. Skwark**[2,3], **Erik Aurell**[2,3,4,5]

[1] Department of Applied Science and Technology, Politecnico di Torino
[2] Department of Information and Computer Science, Aalto University
[3] COIN, The Finnish Centre of Excellence in Computational Inference Research
[4] NETADIS, Marie Curie FP7 Initial Training Network
[5] Department of Computational Biology, KTH Royal Institute of Technology

1 Introduction

Proteins are crucial for the existence of life as we know it. Their function is tightly coupled to their three-dimensional structure, which is difficult to determine experimentally.

Computational protein structure prediction *ab initio* is one of the longest standing challenges in structural biology. Initial methods for prediction showed little success, when tested blindly, because of insurmountable dimensionality of the unrestrained search space.

The knowledge of which amino acids in a protein interact with each other provides sufficient information to predict the structure of a protein. However, until recently, contact prediction in proteins did not produce sufficient information to aid significantly in protein structure predictions. The most successful methods for contact predictions were based on identifying correlated mutations between pairs of residues. The introduction of predictors using methods derived from statistical physics, inferring direct couplings J given the observed evolutionary data, [2] have been shown to significantly increase the accuracy for proteins with many homologous sequences [1] derived from next generation sequencing experiments.

2 Direct Coupling Analysis using pseudolikelihood maximization for the Potts model (plmDCA)

A possible approach of inferring parameters $\Theta = \{J_{ij}(a,b), h_i(a)\}^{i,j,a,b}$ in the Potts model $P(D|\Theta) = \prod_{b=1} P(\underline{a}^b|\Theta)$ given the data $D = \{\underline{a}^b\}$ is maximum-likelihood inference:

$$\Theta^* = \arg\max_{\Theta} \prod_{b=1}^{B} P(\underline{a}^b|\Theta) \tag{1}$$

Here, the probability $P(\underline{a}^b|\Theta)$ can be obtained by dividing the Boltzmann factor of a sequence $e^{-H(\underline{a}^b)}$ by a normalization constant, the partition function.

This is intractable for any reasonable system size because the calculation of the probabilities becomes computationally too expensive. An alternative approach is pseudolikelihood maximization,

$$\Theta_i^* = \arg\max_{\Theta_i} \prod_{b=1}^{B} P(a_i^b|\underline{a}^b_{/i}, \Theta_i), \tag{2}$$

where $\Theta_i = \{\underline{J}_i, h_i\}$ are all parameters appearing in the conditional probability distribution for node i.

The solution to this restrained problem can be solved by gradient descent methods because the probabilities given a set of parameters can be computed explicitly.

The number of parameters is still large and scales as N^2q^2. To avoid overfitting, a regularization term is introduced and the final function to be maximized is:

$$\sum_{b=1}^{B} \log P(a_i^b|\underline{a}^b_{/i}, \Theta_i) - \lambda_J \sum_{i<j} \sum_{a,b} J_{ij}(a,b)^2 - \lambda_h \sum_i \sum_a h_i(a)^2 \tag{3}$$

The resulting parameters will be a trade-off between explaining the data well and setting as many parameters to insignificantly small values as possible.

3 Prediction accuracy

For successful contact-assisted protein folding, one needs roughly the same amount of well-predicted contacts as the amount of amino acids in protein chain. In comparison to the methods based on mean-field approximation, pseudo-likelihood maximization is consistently at least *15% more accurate* in terms of positive predictive value (see section 5 and [3])

On average plmDCA is most effective while working on alignments including also distantly related proteins, as Potts model in this case seems to be more resilient to the inherent noise in such alignments than Ising model of mean-field methods or methods based on partial correlation coefficients, such as PSICOV [2]). Such alignments tend to contain significant amount of unaligned residues (gaps), especially in case of the distantly homologous protein sequences.

4 A Data-Driven Addition to the Hamiltonian: Gap Stretch Parameters (gplmDCA)

Current version of plmDCA does not explicitly account for gaps, but treats them as 21^{st} amino acid. Consequently, in case of regions with low coverage, amino acids flanking the gaps are assigned inordinately strong couplings, leading to mispredictions.

An intuitive way to reduce the gap-bias of the predictions is a set of parameters describing stretches of gaps of length m at a position i:

$$H_{GAPS} = H_{plmDCA} + \sum_{m=1}^{N} \sum_{i=2}^{N-m} \gamma_i^{(m)} I[a_{i-1} \neq q] I[a_i : a_{i+m-1} = q] I[a_{i+m} \neq q], \tag{4}$$

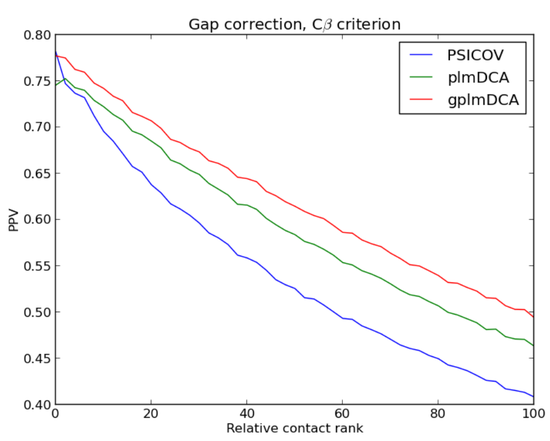
where $I[X = Y]$ are indicator functions that are 1 if their argument is true, and 0 otherwise, and q is the symbol denoting a gap.

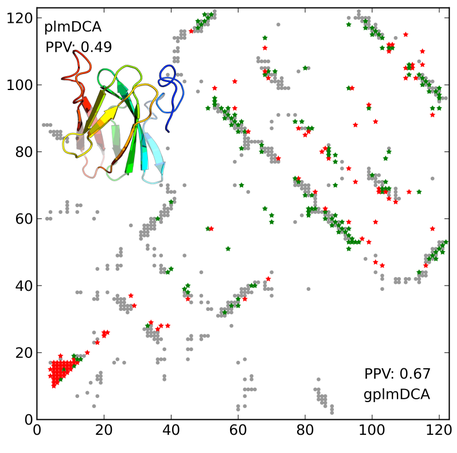
Any term of the first sum constitutes a multi-node interaction of order $m+2$ and introduces $N-m-1$ new parameters. The complete number of new parameters is of the order of N^2 instead of order N^2q^2 for a general multi-node interaction. This addition allows the system to explain gap stretches, clearly not the effect of a two-node interaction, independently of the couplings. These should therefore then be, intuitively, less tainted by them.

Given the straightforward inference technique of plmDCA based on a gradient method, any addition to the Hamiltonian can easily be included in the implementation as long as the derivative with respect to any parameter can be calculated. This can be done by eye from Equation 4. The additional running time can be kept small calculating the results of the indicator functions beforehand — at the expense of a larger memory use of about two times the size of the alignment.

5 Performance improvement and impact

Based on a set of 212 proteins of known structure and alignments produced by HHblits [4], gplmDCA provides more accurate predictions than plmDCA in over 85% of cases and as-good-or-better in over 95% of cases. In comparison to PSICOV (world-class competitive method [2]), plmDCA is better in 70% of cases, while gplmDCA is better in *90% of cases* and slightly worse in only 5%.





As previously shown, spatial couplings inferred by methods like plmDCA can be used to accurately predict unknown protein structures [5]. Increased accuracy provided by gplmDCA should allow for tackling greater range of proteins, and coupled with the rapidly growing amount of protein sequence information, ultimately solving majority of protein structures *in silico*.

References

[1] M. Ingaberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell. **Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models**. *Phys Rev E Stat Nonlin Soft Matter Phys*, **87**(1-1):012707, Jan 2013.

[2] D. Jones, D. Buchan, D. Cozzetto, and M. Pontil. **PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.** *Bioinformatics*, **28**(2):184–190, Jan 2012.

[3] M.J. Skwark, A. Abdel-Rehim, and A. Elofsson. **PconsC: combination of direct information methods and alignments improves contact prediction** *Bioinformatics*, **29**(14):1815–1816, May 2013.

[4] M. Remmert, A. Biegert, A. Hauser, and J. Söding. **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.** *Nat Methods*, **9**(2):173–175, Feb 2012.

[5] T.A. Hopf, L.J. Colwell, R. .Sheridan, B. Rost, C. Sander, D.S Marks **Three-dimensional structures of membrane proteins from genomic sequencing.** *Cell*, **149**(7), 1607-1621.

Christoph Feinauer
Doctoral student, NETADIS Marie Curie ITN
Politecnico di Torino
Email: christoph.feinauer@polito.it

Marcin J. Skwark
Postdoctoral researcher, COIN Centre of Excellence
Aalto University
Email: marcin.skwark@aalto.fi

Erik Aurell
Professor
Royal Institute of Technology, Aalto University
Email: erik.aurell@aalto.fi