Biochemical and Biophysical Research Communications 472 (2016) 217–222

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Improving residue–residue contact prediction via low-rank and sparse decomposition of residue correlation matrix

Haicang Zhang [a, b, 1], Yujuan Gao [c, 1], Minghua Deng [c, d, e], Chao Wang [a, b], Jianwei Zhu [a, b], Shuai Cheng Li [f], Wei-Mou Zheng [g, **], Dongbo Bu [a, *]

[a] Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Bejing, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Center for Quantitative Biology, Peking University, Beijing, China
[d] School of Mathematical Sciences, Peking University, Beijing, China
[e] Center for Statistical Sciences, Peking University, Beijing, China
[f] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[g] Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 14 January 2016
Accepted 30 January 2016
Available online 23 February 2016

Keywords:
Protein contacts prediction
Correlation analysis
Background correlation removal
Low-rank and sparse matrix decomposition

ABSTRACT

Strategies for correlation analysis in protein contact prediction often encounter two challenges, namely, the indirect coupling among residues, and the background correlations mainly caused by phylogenetic biases. While various studies have been conducted on how to disentangle indirect coupling, the removal of background correlations still remains unresolved. Here, we present an approach for removing background correlations via low-rank and sparse decomposition (LRS) of a residue correlation matrix. The correlation matrix can be constructed using either local inference strategies (e.g., mutual information, or MI) or global inference strategies (e.g., direct coupling analysis, or DCA). In our approach, a correlation matrix was decomposed into two components, i.e., a low-rank component representing background correlations, and a sparse component representing true correlations. Finally the residue contacts were inferred from the sparse component of correlation matrix.

We trained our LRS-based method on the PSICOV dataset, and tested it on both GREMLIN and CASP11 datasets. Our experimental results suggested that LRS significantly improves the contact prediction *precision*. For example, when equipped with the LRS technique, the prediction *precision* of MI and mfDCA increased from 0.25 to 0.67 and from 0.58 to 0.70, respectively (Top L/10 predicted contacts, sequence separation: 5 AA, dataset: GREMLIN). In addition, our LRS technique also consistently outperforms the popular denoising technique APC (average product correction), on both local (MI_LRS: 0.67 vs MI_APC: 0.34) and global measures (mfDCA_LRS: 0.70 vs mfDCA_APC: 0.67). Interestingly, we found out that when equipped with our LRS technique, local inference strategies performed in a comparable manner to that of global inference strategies, implying that the application of LRS technique narrowed down the performance gap between local and global inference strategies. Overall, our LRS technique greatly facilitates protein contact prediction by removing background correlations.

An implementation of the approach called COLORS (improving COntact prediction using LOw-Rank and Sparse matrix decomposition) is available from http://protein.ict.ac.cn/COLORS/.

1. Introduction

In natural environment, a protein usually adopts a specific tertiary structure determined primarily by its amino acid sequence [3]. Under chemical and physical effects, some residues are spatially close to others, forming a set of residue–residue contacts. These contacts are known to be responsible for stabilizing the native protein folds [13]. The accurate prediction of residue–residue

* Corresponding author.
** Corresponding author.
E-mail addresses: zheng@itp.ac.cn (W.-M. Zheng), dbu@ict.ac.cn (D. Bu).
[1] The first two authors contributed equally to this paper.

http://dx.doi.org/10.1016/j.bbrc.2016.01.188
0006-291X/

contacts can provide distance information among residues, which should greatly helps both free modeling [24,26] and template-based modeling strategies [22] for protein structure prediction.

A large variety of approaches have been proposed for residue–residue contact prediction, including supervised-learning approaches [7,9,32,30] and purely sequence-based approaches [5,8,27,6]. Typically, a purely sequence-based approach begins with building multiple sequence alignment (MSA) for a target protein, and then identifies possible residue–residue contacts through correlated mutation analysis [29,12]. The underlying principle is that residue–residue contacts, generally being responsible for stabilizing protein structure, tend to be held during evolutionary history of the protein; thus, if a residue in contact mutates, its contacting partner is expected to accordingly mutate to maintain the contact. This coevolution between contacting residues commonly appear as correlations between the corresponding columns in MSA of the target protein (hereafter called *true correlation*); the correlation among MSA columns, in turn, can be explored to infer residue contacts.

Two difficulties are involved in the purely sequence-based strategy for correlation analysis [16,24]. First, the true correlations are generally blurred by transitive correlations, also known as *indirect coupling*. More precisely, suppose the *i*th residue correlates with the *j*th residue, and the *j*th residue correlates with the *k*th residue; in this situation, even if the *i*th residue does not contact with the *k*th residue, correlation might still be observed between them due to transitive effects. Second, the intrinsic background correlations usually interfere with the identification of coevolution signals. The background correlations come from at least two sources: (1) During the phylogenetic history of a certain protein family, mutations occurring in an ancestral protein will be inherited by all of its descendants. Thus, almost all residue pairs appear to have some degree of correlations purely caused by phylogenetic biases. (2) The highly variable columns in MSAs usually lead to relatively high level of both random and non-random correlations among these columns [8], which forms another source of the background correlations. The background correlations, as well as the indirect coupling, often confound the correlation analysis and subsequent contact prediction.

Recently, there have been significant progresses in overcoming the indirect coupling difficulty. For example, mfDCA employed the mean field technique for direct coupling analysis [27], while plmDCA exploited the pseudo-likelihood maximization technique to achieve the same objective [10,18]. Another approach, called sparse inverse covariance estimation (PSICOV), models MSA using a Gaussian distribution, and estimates partial correlations by inverting the empirical covariance matrix through graphical lasso technique [16]. Following this strategy, Andreatta et al. proposed to utilize the least-square technique to speed up the inversion of empirical covariance matrices [2]. Note that an MSA usually consists of proteins with divergent sequences but similar folds, Ma et al. successfully applied the group graphical lasso technique into direct coupling analysis of MSA [23]. These approaches were known as “global” since correlated residues are treated dependent on each other; in contrast, the “local” statistical inference models—for instance, MI [25] and OMES [11]—treat a certain residue pair independent of others [24].

Besides these efforts to overcome indirect coupling, a few methods have been developed for removing the background correlations caused by phylogenetic biases. In particular, it has been reported that the exclusion of highly similar sequences helps reduce phylogenetic biases [25]. Bootstrapping and other randomization methods [33,28] were also found effective in reducing phylogenetic biases. Also promising is the average product correction (APC) technique. APC was originally designed to efficiently estimate the expected levels of background noise arising from phylogenetic sources [8], and currently the APC technique is widely used as a post-processing procedure in both local and global inference strategies. The existing approaches have proven to be relatively successful on various proteins; however, the removal of background correlation still remains a challenge to the correlation analysis of MSA.

In this study, we present a novel approach that employs the low-rank and sparse matrix decomposition (LRS) technique for removing background correlations. The approach distinguishes true correlations from background correlations according to their different characteristics, i.e., the sparsity of true correlations, and the low-rank characteristic of background correlations. On one side, the number of contacts in a L-length protein was estimated as ~$0.05 \times L^2$ [17]. This number is substantially small when considering the total L^2 possible contacting residue pairs, and thus implying the considerable sparsity of true correlations. On the other side, the first mode (principal component) of a correlation matrix describes the “coherent” correlations among all positions caused by phylogenetic biases [14]. In fact, the APC technique is essentially equivalent to removing the first mode of a correlation matrix, which implicitly assumes the rank of background correlation as 1 (see supplementary). However, besides the first mode, the phylogenetic biases might also contribute to other modes especially when MSA are constructed from proteins segregated into subfamilies [14]. Here, we adopted the similar but more general assumption of background correlations being low-rank and performed LRS to self-adaptively separates true correlations from background correlations.

It should be pointed out that the LRS technique, also known as robust principle component analysis (PCA), has been widely applied in the field of computational vision analysis [4] and gene expression analysis [21,31]. As far as we know, this is the first time that the LRS technique has been applied for protein contact prediction.

We evaluated LRS technique on GREMLIN dataset and CASP11 targets as well. The evaluation results suggested that by using the LRS technique, the contact prediction *precision* was significantly improved regardless of whether local or global inference models were used.

2. Methods

To apply the LRS technique for protein contacts prediction, we first built a matrix to measure correlations among residues in the target protein. The residue correlation measure can be calculated by using local statistical models (e.g., MI and OMES) or global statistical models (e.g., DCA and PSICOV). Next, by using the LRS technique, we decomposed the residue correlation matrix into a low-rank component plus a sparse component. The sparse component was then used to infer residue–residue contacts in the target protein.

We will describe the residue correlation matrix construction in Section 2.1, thereafter describe the LRS technique in Section 2.2.

2.1. Residue correlation matrix construction

A variety of measures have been proposed to evaluate the correlation between any two residues in a protein. The correlation measures are usually derived from MSA information by using local statistical models or global statistical models. The correlation matrices reported by mfDCA [27], PSICOV [16], and plmDCA [10] were used as representatives of global statistical models. As for local statistical models, we focused on the widely-used MI [25] and OMES [19] correlation measures. In addition, we designed another