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 × 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

Table 1
Improvement of local statistical models using LRS and APC techniques on the GREMLIN benchmark dataset.

Method	*Separation* ≥ 5			*Separation* ≥ 18		
	Top 10	Top L/10	Top L/5	Top 10	Top L/10	Top L/5
MI	0.30	0.25	0.22	0.31	0.25	0.20
MI_APC	0.37	0.34	0.32	0.48	0.43	0.38
MI_LRS	**0.70**	**0.67**	**0.61**	**0.68**	**0.64**	**0.55**
COV	0.29	0.24	0.21	0.29	0.24	0.20
COV_APC	0.40	0.36	0.33	0.50	0.44	0.38
COV_LRS	**0.70**	**0.66**	**0.58**	**0.69**	**0.63**	**0.53**
OMES	0.24	0.20	0.17	0.25	0.21	0.17
OMES_APC	0.29	0.25	0.22	0.33	0.27	0.23
OMES_LRS	**0.64**	**0.60**	**0.51**	**0.64**	**0.55**	**0.45**

Table 2
Improvement of local statistical models using LRS and APC techniques on the CASP11 benchmark dataset.

Method	*Separation* ≥ 5			*Separation* ≥ 18		
	Top 10	Top L/10	Top L/5	Top 10	Top L/10	Top L/5
MI	0.27	0.21	0.18	0.25	0.20	0.17
MI_APC	0.28	0.23	0.21	0.37	0.31	0.26
MI_LRS	**0.56**	**0.50**	**0.44**	**0.56**	**0.47**	**0.40**
COV	0.24	0.20	0.18	0.25	0.20	0.16
COV_APC	0.26	0.25	0.22	0.36	0.31	0.27
COV_LRS	**0.53**	**0.47**	**0.41**	**0.52**	**0.45**	**0.38**
OMES	0.18	0.16	0.13	0.17	0.14	0.12
OMES_APC	0.20	0.19	0.16	0.26	0.21	0.17
OMES_LRS	**0.55**	**0.48**	**0.38**	**0.51**	**0.41**	**0.30**

Table 3
Improvement of global statistical models using LRS and APC techniques on the GREMLIN benchmark dataset.

Method	*Separation* ≥ 5			*Separation* ≥ 18		
	Top 10	Top L/10	Top L/5	Top 10	Top L/10	Top L/5
mfDCA	0.63	0.58	0.51	0.60	0.54	0.46
mfDCA_APC	0.70	0.67	0.63	0.70	0.68	0.62
mfDCA_LRS	**0.72**	**0.70**	**0.64**	**0.71**	**0.69**	**0.63**
plmDCA	0.68	0.65	0.59	0.66	0.62	0.54
plmDCA_APC	0.72	0.70	0.66	0.73	0.70	0.65
plmDCA_LRS	**0.75**	**0.73**	**0.68**	**0.75**	**0.71**	**0.66**
PSICOV	0.70	0.67	0.60	0.67	0.62	0.55
PSICOV_APC	**0.71**	0.68	0.63	0.68	0.63	0.58
PSICOV_LRS	**0.71**	**0.69**	**0.64**	**0.69**	**0.65**	**0.59**

Table 4
Improvement of global statistical models using LRS and APC techniques on the CASP11 benchmark dataset.

Method	*Separation* ≥ 5			*Separation* ≥ 18		
	Top 10	Top L/10	Top L/5	Top 10	Top L/10	Top L/5
mfDCA	0.47	0.41	0.36	0.47	0.38	0.31
mfDCA_APC	0.55	0.47	0.42	0.53	0.46	0.42
mfDCA_LRS	**0.60**	**0.50**	**0.46**	**0.55**	**0.48**	**0.43**
plmDCA	0.49	0.46	0.42	0.49	0.45	0.39
plmDCA_APC	0.55	0.53	0.48	**0.57**	0.52	0.46
plmDCA_LRS	**0.58**	**0.55**	**0.50**	**0.57**	**0.54**	**0.47**
PSICOV	0.57	0.52	0.46	0.54	0.49	0.42
PSICOV_APC	0.58	0.54	0.48	0.56	0.50	0.43
PSICOV_LRS	**0.59**	**0.55**	**0.50**	**0.57**	**0.51**	**0.45**

correlation measure, called COV, based on empirical covariance.

Similar to OMES, COV measures the difference between observed versus expected frequency of residue pairs; however, COV calculates the l_1 norm of the frequency difference.

$$COV_{ij} = \sum_{a,b \in AA} \left| f_{ij}(a,b) - f_i(a) f_j(b) \right| \qquad (1)$$

For the sake of simplicity, we put the details of MI and OMES in the supplementary information.

It should be pointed out that in the study, MSAs have already been re-weighted before calculating residue correlations. The reason is that the re-weighting strategy has also been reported helpful for improving contact prediction *precision* [16,27,10], although it was designed to mitigate potential redundancy in MSA [15,1].

2.2. Low-rank and sparse matrix decomposition

Suppose we have already calculated a residue correlation matrix M, where the entry M_{ij} quantifies the correlation between residues at the *i*th and *j*th sites using one of the measures mentioned above. We decomposed the matrix M into two components, i.e., $M = \mathscr{L} + S$, where $\mathscr{L}$ is a low-rank matrix describing background correlations, and S is a sparse matrix describing true correlations. Thus, the non-zero entries in S—for instance, S_{ij}—represents the propensity of the contact between the *i*th and *j*th residues.

The optimal decomposition can be determined via solving the following optimization problem [4].

$$\text{minimize} \quad \|\mathscr{L}\|_* + \lambda \|S\|_1 \qquad (2a)$$

$$\text{subject to} \quad \mathscr{L} + S = M \qquad (2b)$$

The objective function consists of two terms. The first term measures the rank of $\mathscr{L}$ using its nuclear norm, i.e., the sum of its singular values, and generally the smaller of the nuclear form, the lower rank of $\mathscr{L}$. The second term measures the sparsity of S using its l_1 norm, and the smaller of the l_1 norm, the more sparse of S. In the study, the objective function was minimized by virtue of the *inexact augmented Lagrange multiplier* technique [20], and the parameter λ was introduced to balance the two terms. In particular, the sparsity of S was emphasized by setting a large λ, while the low-rank of $\mathscr{L}$ was emphasized by setting a small λ. The optimal setting of λ was obtained based on a training dataset.

3. Results and discussion

In our experiments, the PSICOV dataset having 150 single domain monomeric proteins [16] was utilized to train the parameter λ, and the GREMLIN dataset having 329 proteins [18] was utilized as testing set to evaluate the LRS technique. We also evaluated the performance of the LRS technique on CASP11 targets.

To avoid the biases incurred by overlap between training dataset and testing dataset, the similar proteins shared by training dataset and testing dataset were removed. Here, the criterion of sequence similarity was set as sequence identity over 25%, which is commonly adopted in previous studies [23,32]. After this overlap removing process, the training set contains a total of 131 proteins (see supplementary file for the list).

*For e*ach protein in training and testing datasets, true contacts have been annotated between two residues with a C_β–C_β (C_α *in the case of* Glycine *residues*) *distance of less than* 8*Å*. The cutoff distance of 8*Å* is widely adopted in the community of protein structure prediction as well as CASP competition; however, no consensus has been reached on the cutoff distance setting [18,16]. For a thorough understanding of the prediction performance, we repeated our experiments with two other cutoff distance settings, namely 7.5*Å* and 8.5*Å* (Tables 3 and 4 in the supplementary file).

In our study, the performance of the contacts prediction was

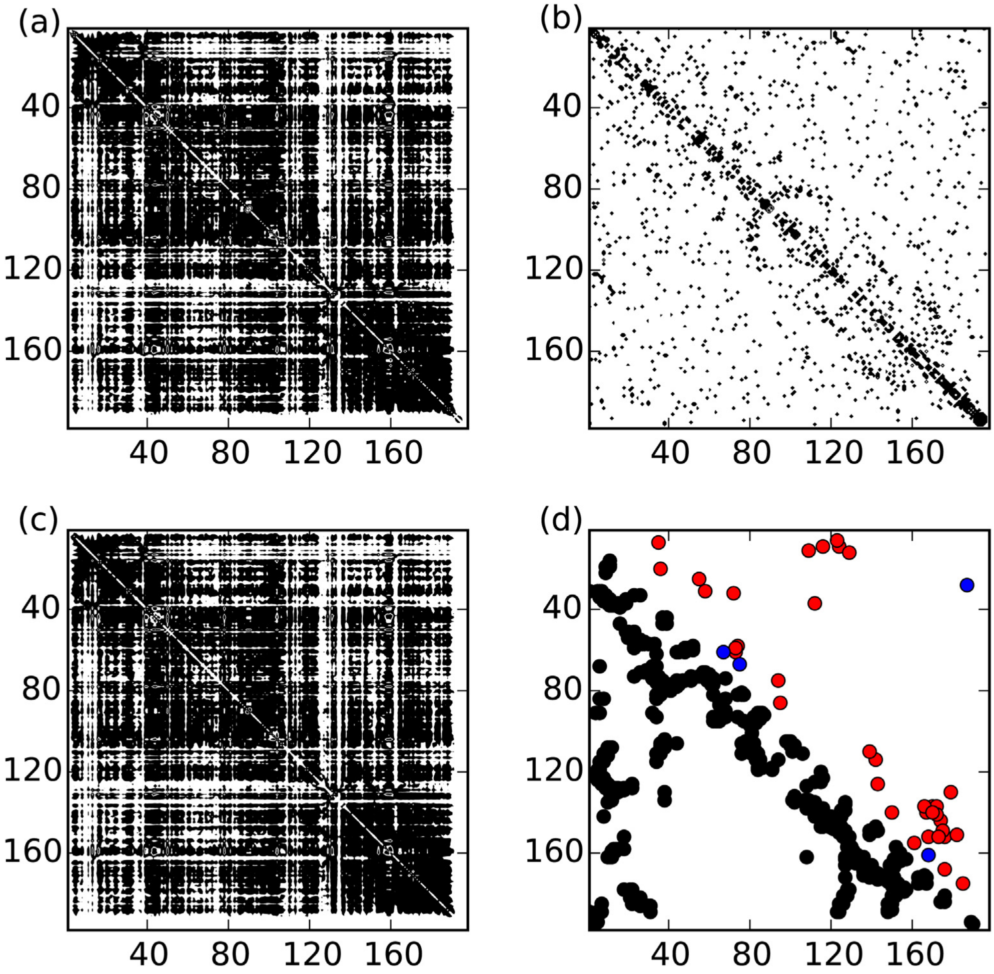

Fig. 1. Running process of the MI_LRS approach on protein 1chdA. (a) The original MI matrix M; (b) The sparse component *S*; (c) The low-rank component $\mathscr{L}$; and (d) The contact map predicted from the largest *L*/5 entries in *S*, where the true contacts were drawn in black in the lower triangle, the true positives in red, and the false positives in blue. The prediction *precision* is 0.90. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

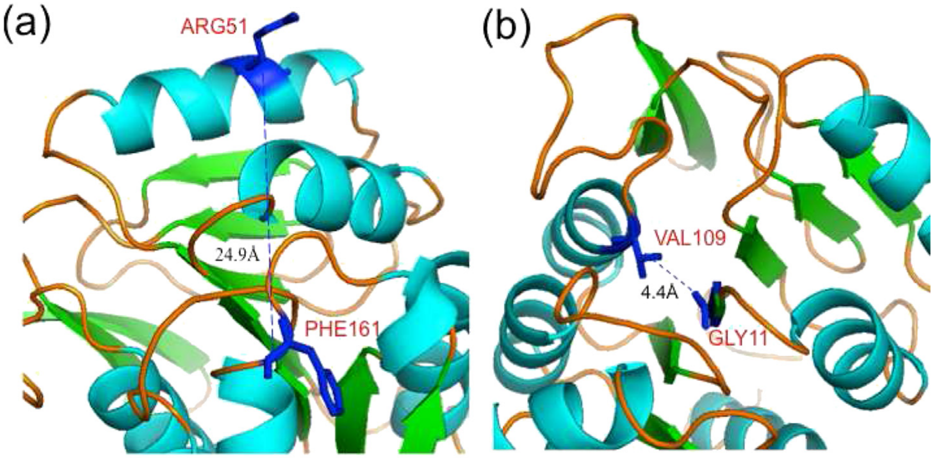

Fig. 2. Two examples of residue pairs showing the running process of MI_LRS approach. (a) Non-contact residue pair ARG51-PHE161 with C_β distance of 24.9 Å. The pair is ranked 6th according to the original MI score $M_{51,161} = 0.577$, and is ranked 368th according to the S score $S_{51,161} = 0.001$. Thus, the pair was not annotated as a contact by LRS technique. (b) Contact residue pair GLY11-VAL109 with C_β distance of 4.4 Å. The pair is ranked 670th according to the original MI score $M_{11,109} = 0.262$, and is ranked 3rd according to the S score $S_{11,109} = 0.163$. Thus, the LRS technique successfully reports this pair as a contact.

evaluated using the mean prediction *precision* (*also known as accuracy*), i.e. the fraction of predicted contacts that are true [16,27,23,18].

Before showing performance analysis results, we first presented protein 1chdA as a concrete example of how the LRS technique works. Then we investigated the improvements on local statistical

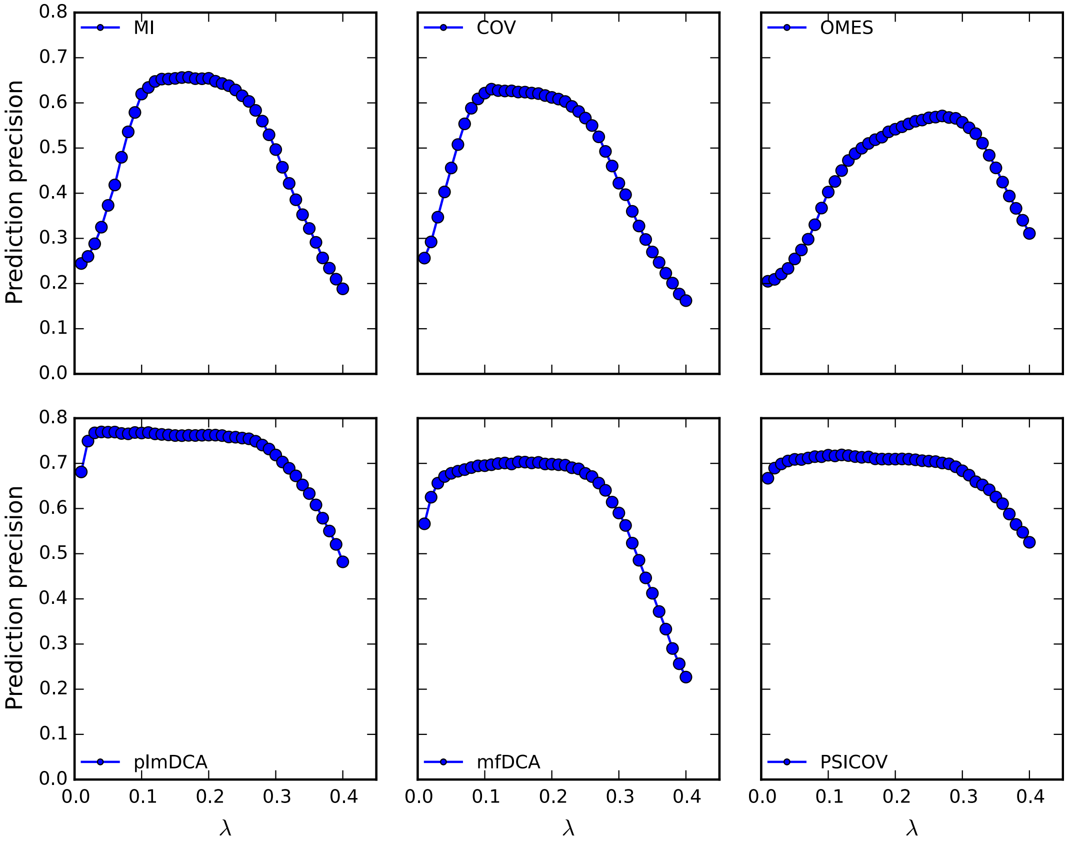

Fig. 3. The effects of parameter λ on prediction *precision* (measured in top *L*/10 predicted contacts with the sequence separation threshold of 5 AA). Training dataset: PSICOV benchmark dataset.

models and global statistical models by using the LRS technique in terms of contact prediction *precision*. We further compared the contribution of the LRS technique with that of the popular APC technique. In fact, LRS can be treated as an extension of APC. The relationship between LRS and APC techniques was elucidated in the supplementary file.

3.1. Running process of the LRS technique: a case study

In Fig. 1, the running process of the LRS technique is illustrated using the concrete example of protein 1chdA. More specifically, the original MI correlation matrix M is shown in Fig. 1(a), with pixels drawn in grayscale proportional to correlation values. Then M was decomposed into a low-rank component $\mathscr{L}$ (Fig. 1(c)) plus a sparse component *S* (Fig. 1(b)). Finally, the largest *L*/5 entries (with sequence separation threshold of 5AA) in the matrix S were identified and further converted as a contact map, shown in Fig. 1(d). By comparing with the true contacts of 1chdA (drawn in black in the lower triangle of Fig. 1(d)), we observed that MI_LRS achieved a contact prediction *precision* of 0.90, which is considerably higher than MI_APC (prediction *precision*: 0.58).

Take ARG51-PHE161 as an example of non-contact residue pair (Fig. 2 (a)). According to the original MI score, it was ranked as the 6th ($M_{51,161}$=0.577), and was thus incorrectly reported as a contact by the MI score. Contrastly, we estimated $S_{51,161} = 0.001$ by using the LRS technique. The residue pair, ranked 368th according to S, was not reported as a contact by LRS.

Another pair, GLY11-VAL109, serves as an example of contacting residues (Fig. 2 (b)). Compared with other residue pairs, the pair has relatively small MI value $M_{11,109} = 0.262$ with a rank of 670th. Thus, the pair was not reported by the MI score. However, the LRS technique estimates $S_{11,109} = 0.163$ and ranks the pair the 3rd. The pair, outstanding from other pairs in terms of S, was correctly reported as a contact by LRS.

3.2. Improving local statistical models by using LRS

Table 1 and Table 2 summarize the performance of local statistical models—including MI, COV, and OMES—on GREMLIN and CASP11 datasets, respectively. Besides the original version of the models, the variants equipped with the LRS technique and APC correction were also evaluated. Following the contact prediction conventions, we filtered out short distance contacts under two settings of sequence separation thresholds (5 AA, and 18 AA), and reported the *precisions* of top 10, *L*/10 and *L*/5 predicted contacts.

The tables reveal substantial improvements on local statistical models by employing the LRS technique. Take top *L*/10predictions with the sequence separation threshold of 5 AA on GREMLIN dataset as an example. The original MI model exhibited a prediction precision of only 0.25; this precision level was considerably improved to 0.67 via using the LRS technique. Moreover, the LRS technique is significantly superior to the APC correction in terms of prediction precision (MI_LRS: 0.67 versus MI_APC: 0.34). Similar observations were reported on COV and OMES models.

As shown in Table 2, the prediction *precision* is relatively low over the CASP11 targets compared to that over the GREMLIN dataset. This might be attributed to the difference in MSA quality. More specifically, the median of the number of non-redundant homologs in the MSAs of the proteins in the GREMLIN and CASP11 dataset is 1760 and 720 respectively. However, the outperformance of LRS over APC was still observed even if the MSA quality is low.

Together, these results imply the existence of strong background correlations in the local statistical models, and only when background correlations are removed can local statistical models yield accurate prediction.

3.3. Improving global statistical models by using LRS

For global statistical models, we again investigated the contributions of the LRS technique. The experimental results on GREMLIN dataset and CASP11 targets, shown in Tables 3 and 4, represented a difference from the local statistical models. Take top L/10 predicted contacts on GREMLIN benchmark as an example, the incorporation of APC and LRS techniques contributed an *precision* improvement of approximately 5–12% to plmDCA and mfDCA models, and around 1–3% to PSICOV model. The *precision* improvements are smaller than that for local statistical models. Nevertheless, LRS still performed better than APC consistently.

3.4. Parameter tuning and its effects on prediction precision

The parameter λ is used to balance the sparsity of S and low-rank of L; thus, a proper setting of λ is important to prevent from overemphasizing any of the two sides. To obtain the optimal setting, we trained the parameter λ on PSICOV dataset. As shown in Fig. 3, too large or too small λ values lead to poor performance. In addition, the prediction *precisions* are roughly the same for a large range of settings, which implies the robustness of the LRS technique.

It should be pointed that the experiments using two other cutoff distance settings, namely 7.5$Å$ and 8.5$Å$, revealed similar observations to that using the cutoff distance of 8$Å$ (Tables 3 and 4 in the supplementary file).

In summary, our experimental results suggested that LRS significantly improved the contact prediction *precision*. Moreover, LRS consistently outperformed APC. Thus, the LRS technique greatly facilitated protein contact prediction by removing background correlations.

Conflict of interests

All authors declare no conflict of interest.

Acknowledgment

This study was funded by the National Basic Research Program of China (973 Program) (2012CB316502, 2015CB910303), the National Nature Science Foundation of China (11175224, 11121403, 31270834, 61272318, 31171262, 31428012, 31471246) and the Open Project Program of State Key Laboratory of Theoretical Physics (No. Y4KF171CJ1). This work made use of the eInfrastructure provided by the European Commission co-funded project CHAIN-REDS (GA no 306819).

We greatly appreciate Sergey Ovchinnikov for sharing with us the GREMLIN benchmark and providing details on how to use it. We also thanks Torsten Julich for his invaluable comments and suggestions.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.bbrc.2016.01.188.

References

[1] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.

[2] M. Andreatta, S. Laplagne, S. C. Li, S. Smale, 2013, Prediction of residue-residue contacts from protein families using similarity kernels and least squares regularization. arXiv preprint arXiv:1311.1301.

[3] C.B. Anfinsen, Studies on the Principles that Govern the Folding of Protein Chains, 1972.

[4] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM JACM 58 (3) (2011) 11.

[5] D.K. Chiu, T. Kolodziejczak, Inferring consensus structure from nucleic acid sequences, Comput. Appl. Biosci. CABIOS 7 (3) (1991) 347–352.

[6] D. de Juan, F. Pazos, A. Valencia, Emerging methods in protein co-evolution, Nat. Rev. Genet. 14 (4) (2013) 249–261.

[7] P. Di Lena, K. Nagata, P. Baldi, Deep architectures for protein contact map prediction, Bioinformatics 28 (19) (2012) 2449–2457.

[8] S.D. Dunn, L.M. Wahl, G.B. Gloor, Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction, Bioinformatics 24 (3) (2008) 333–340.

[9] J. Eickholt, J. Cheng, Predicting protein residue–residue contacts using deep networks and boosting, Bioinformatics 28 (23) (2012) 3066–3072.

[10] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: using pseudolikelihoods to infer potts models, Phys. Rev. E 87 (1) (2013), 012707.

[11] A.A. Fodor, R.W. Aldrich, Influence of conservation on calculations of amino acid covariance in multiple sequence alignments, Proteins Struct. Funct. Bioinforma. 56 (2) (2004) 211–221.

[12] U. Gobel, C. Sander, R. Schneider, A. Valencia, Correlated mutations and residue contacts in proteins, Proteins Struct.Funct. Genet. 18 (4) (1994) 309–317.

[13] M.M. Gromiha, S. Selvaraj, Inter-residue interactions in protein folding and stability, Prog. Biophys. Mol. Biol. 86 (2) (2004) 235–277.

[14] N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein sectors: evolutionary units of three-dimensional structure, Cell 138 (4) (2009) 774–786.

[15] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. 89 (22) (1992) 10915–10919.

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

[17] D.T. Jones, T. Singh, T. Kosciolek, S. Tetchner, MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins, Bioinformatics 31 (7) (2015) 999–1006.

[18] H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era, Proc. Natl. Acad. Sci. 110 (39) (2013) 15674–15679.

[19] I. Kass, A. Horovitz, Mapping pathways of allosteric communication in groel by analysis of correlated mutations, Proteins Struct. Funct. Bioinforma. 48 (4) (2002) 611–617.

[20] Z. Lin, M. Chen, Y. Ma, 2010, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055.

[21] J.-X. Liu, Y.-T. Wang, C.-H. Zheng, W. Sha, J.-X. Mi, Y. Xu, Robust PCA based method for discovering differentially expressed genes, BMC Bioinforma. 14 (Suppl 8) (2013) S3.

[22] J. Ma, S. Wang, Z. Wang, J. Xu, MRFalign: protein homology detection through alignment of markov random fields, PLoS Comput. Biol. 10 (3) (2014) e1003500.

[23] J. Ma, S. Wang, Z. Wang, J. Xu, Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning, in: Research in Computational Molecular Biology, Springer, 2015, pp. 218–221.

[24] D.S. Marks, T.A. Hopf, C. Sander, Protein structure prediction from sequence variation, Nat. Biotechnol. 30 (11) (2012) 1072–1080.

[25] L. Martin, G.B. Gloor, S. Dunn, L.M. Wahl, Using information theory to search for co-evolving residues in proteins, Bioinformatics 21 (22) (2005) 4116–4124.

[26] M. Michel, S. Hayat, M.J. Skwark, C. Sander, D.S. Marks, A. Elofsson, Pconsfold: improved contact predictions improve protein models, Bioinformatics 30 (17) (2014) i482–i488.

[27] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D.S. Marks, C. Sander, R. Zecchina, J.N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. 108 (49) (2011) E1293–E1301.

[28] O. Noivirt, M. Eisenstein, A. Horovitz, Detection and reduction of evolutionary noise in correlated mutation analysis, Protein Eng. Des. Sel. 18 (5) (2005) 247–253.

[29] I. Shindyalov, N. Kolchanov, C. Sander, Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng. 7 (3) (1994) 349–358.

[30] M.J. Skwark, D. Raimondi, M. Michel, A. Elofsson, Improved contact predictions using the recognition of protein like contact patterns, PLoS Comput. Biol. 10 (11) (2014) e1003889.

[31] Y. Wang, D. Yang, M. Deng, Low-rank and sparse matrix decomposition for genetic interaction data, BioMed Res. Int. 2015 (2015).

[32] Z. Wang, J. Xu, Predicting protein contact map using evolutionary and physical constraints by integer programming, Bioinformatics 29 (13) (2013) i266–i273.

[33] K.R. Wollenberg, W.R. Atchley, Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap, Proc. Natl. Acad. Sci. 97 (7) (2000) 3288–3291.