


RESEARCH ARTICLE

Open Access



COUSCOus: improved protein contact prediction using an empirical Bayes covariance estimator

Reda Rawi^{1*} , Raghvendra Mall¹, Khalid Kunji¹, Mohammed El Anbari², Michael Aupetit¹, Ehsan Ullah¹ and Halima Bensmail¹

Abstract

Background: The post-genomic era with its wealth of sequences gave rise to a broad range of protein residue-residue contact detecting methods. Although various coevolution methods such as PSICOV, DCA and plmDCA provide correct contact predictions, they do not completely overlap. Hence, new approaches and improvements of existing methods are needed to motivate further development and progress in the field. We present a new contact detecting method, COUSCOus, by combining the best shrinkage approach, the empirical Bayes covariance estimator and GLasso.

Results: Using the original PSICOV benchmark dataset, COUSCOus achieves mean accuracies of 0.74, 0.62 and 0.55 for the top $L/10$ predicted long, medium and short range contacts, respectively. In addition, COUSCOus attains mean areas under the precision-recall curves of 0.25, 0.29 and 0.30 for long, medium and short contacts and outperforms PSICOV. We also observed that COUSCOus outperforms PSICOV w.r.t. Matthew's correlation coefficient criterion on full list of residue contacts. Furthermore, COUSCOus achieves on average 10% more gain in prediction accuracy compared to PSICOV on an independent test set composed of CASP11 protein targets. Finally, we showed that when using a simple random forest meta-classifier, by combining contact detecting techniques and sequence derived features, PSICOV predictions should be replaced by the more accurate COUSCOus predictions.

Conclusion: We conclude that the consideration of superior covariance shrinkage approaches will boost several research fields that apply the GLasso procedure, amongst the presented one of residue-residue contact prediction as well as fields such as gene network reconstruction.

Keywords: Residue-residue contact prediction, Shrinkage, GLasso

Background

A multiple sequence alignment (MSA) of orthologous protein sequences not only carries evolutionary sequence information, but also information about functional and structural constraints imposed on the three-dimensional (3D) structure of a protein. Conserved or slightly mutated columns indicate important protein positions for protein stability and function. Additionally, non-conserved positions may also play key roles in maintaining the functionality when accompanied by compensatory mutations at other positions [1, 2]. It is of high interest to

develop methods predicting coevolution patterns from MSAs, because coevolving positions mainly involve protein positions proximal in 3D structure [3] and they serve as a valuable source of distance constraints in protein structure [4–7] as well as in protein complex interface predictions [8, 9].

Due to the substantial increase in sequence data in the post-genomic era, a broad range of methods have been introduced for detecting residue-residue contacts from MSAs in the past decades. Mutual information (MI) was one of the first metrics to be applied for contact prediction from MSAs [10, 11]. An improved MI version that corrects for background noise and phylogenetic effects (MIp) has been introduced by Dunn et al. [12]. Recent methodological improvements are

*Correspondence: redarawi411@gmail.com

¹Computational Science and Engineering, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

Full list of author information is available at the end of the article

able to distinguish between direct and indirect couplings and have demonstrated enormous accuracy in predicting real couplings and coevolution. Such methods include a Bayesian network approach (BvN) [13], Direct Coupling Analysis (DCA) [14, 15], Protein Sparse Inverse COVariance (PSICOV) [16], and pseudolikelihood approaches implemented in plmDCA [17] and GREMLIN [18].

Most recently, new hybrid methods have been developed, amidst many others such as DNCON [19], PConsC [20], CoinDCA [21] or MetaPSICOV [22], where contact detecting methods are combined along with protein physiochemical features to provide more accurate contact predictions.

In the present study, we developed Contact predictiOn Using Shrunked COvariance (COUSCOus), a residue-residue contact detecting method approaching contact inference in a similar manner as PSICOV, by applying the sparse inverse covariance estimation technique introduced by Meinshausen and Bühlmann [23]. Here, we used a different covariance matrix shrinkage approach, the empirical Bayes covariance estimator, which has been shown by Haff to be the best estimator in a Bayesian framework [24], especially dominating estimators of the form aS , such as the smoothed covariance estimator applied in PSICOV. By analysing the original PSICOV benchmark test set [16] and proteins from the Critical Assessment of techniques for protein Structure Prediction 11 (CASP11) experiments, we show that COUSCOus significantly outperforms PSICOV. Furthermore, we designed a simple random forest (RF) meta-classifier that includes contact detecting techniques and sequence-derived physiochemical features and showed that replacing PSICOV with COUSCOus enhances the prediction outcome.

Methods

Dataset

The benchmark dataset used in this study is the original PSICOV test set introduced by Jones et al. [16]. We used the same alignments without modification as have been made available to ensure comparability. However, for validation we selected 146 out of 150 single domain monomeric proteins because the latest PSICOV version V2.1b3 was unable to provide contact predictions on the remaining four test cases, when used with default parameters. This is due to the insufficient number of effective sequences within the four alignments.

The second test set consists of 37 proteins of the CASP11 experiment (see Additional file 1). We selected only those proteins where the latest version of PSICOV successfully provided predictions, to make a fair comparison. The training set introduced by Jones et al. [22] was used to build the RF meta-classifier.

Coevolution analysis methods

The residue-residue contact prediction metrics applied in this study are MI [10, 11], MIp [12], OMES [25], BvN [13], DCA [15] and PSICOV [16]. The resulting coevolution between pairs of amino-acids using MI, MIp and OMES were calculated using *Evol* module of Prody [26]. BvN results were generated using Perl scripts and C++ source code kindly provided by the authors [13]. PSICOV results were calculated using the code available online [16]. DCA results were obtained using the fast and free software version FreeContact introduced by Kaján et al. [27]. Methodological details for the different methods may be found in the original studies.

COUSCOus

Pre-processing

In our approach, we generate a sample covariance matrix S from the input MSA. The MSAs are composed of n orthologous protein sequences where each sequence represents a row. Each protein sequence is made of m amino acids as a result of which we have L columns per alignment row. The size of the covariance matrix S is $21L \times 21L$. This is because we compute the marginal single site frequencies $f(A_i)$ and $f(B_j)$ of observed amino acid types (20 natural occurring amino acids and a gap) in columns i and j and their corresponding pair site frequencies $f(A_iB_j)$:

$$S_{ij}^{ab} = f(A_iB_j) - f(A_i)f(B_j) \quad (1)$$

Interestingly the precision matrix Θ which is the inverse of the covariance matrix S will contain the partial correlations of all pairs of variables taking into consideration the effects of all other variables. Hence, the non-zero entries of Θ will provide the extent of direct coupling between any two pairs of amino acids at sites i and j .

Yet, due to the fact that we are generating a covariance matrix S out of MSAs representing homologous protein sequences where not all amino acids are present at each site of the MSA, it is certain that S is singular and not directly invertible. Several approaches have been proposed to approximate the precision matrix in such cases. The most powerful and widely used technique is the sparse inverse covariance estimation using the graphical lasso (GLasso) [28].

GLasso

We briefly summarise the basic motivation and algorithm. Consider matrix $X = [X_1, \dots, X_p]$ where X_i is a random vector of length n with covariance matrix Σ and precision matrix $\Theta = \{\theta_{ij}\}_{1 \leq i, j \leq p}$. Further, let S denote the empirical covariance matrix obtained from the data. The estimation of the precision matrix Θ is challenging when it is sparse. Interestingly, this task is closely related to selection of graphical models.

Let $G = (V, E)$ be a graph representing conditional independence relations between components of X . G is composed of a set of vertices V with p components $\{X_1, \dots, X_p\}$ and an edge set E of ordered pairs (i, j) , with $(i, j) \in E$, if an edge between X_i and X_j exists. The edge between X_i and X_j is excluded from the edge set E if and only if X_i and X_j are independent given all other components $\{X_k, k \neq i, j\}$. Assuming that the raw data X is multivariate gaussian ($X \sim N(\mu, \Sigma)$), the conditional independence between X_i and X_j given all other components is equivalent to zero in the precision matrix ($\theta_{ij} = 0$) as shown in [29]. Hence, for gaussian distributions recovering the structure of graph

G is equivalent to the estimation of the support of the precision matrix.

The precision matrix Θ can then be estimated using a L_1 penalised log-likelihood approach. The GLasso algorithm, introduced by Friedman et al. [28], efficiently computes the solution by:

$$\hat{\Theta}_{\text{GLasso}} := \arg \min_{\Theta > 0} \{ \text{tr}(S\Theta) - \log \det(\Theta) + \lambda \|\Theta\|_1 \}, \quad (2)$$

with tr as trace, $\|\Theta\|_1$ as the sum of the absolute values of the elements in Θ and λ as a positive tuning parameter to control the sparsity.

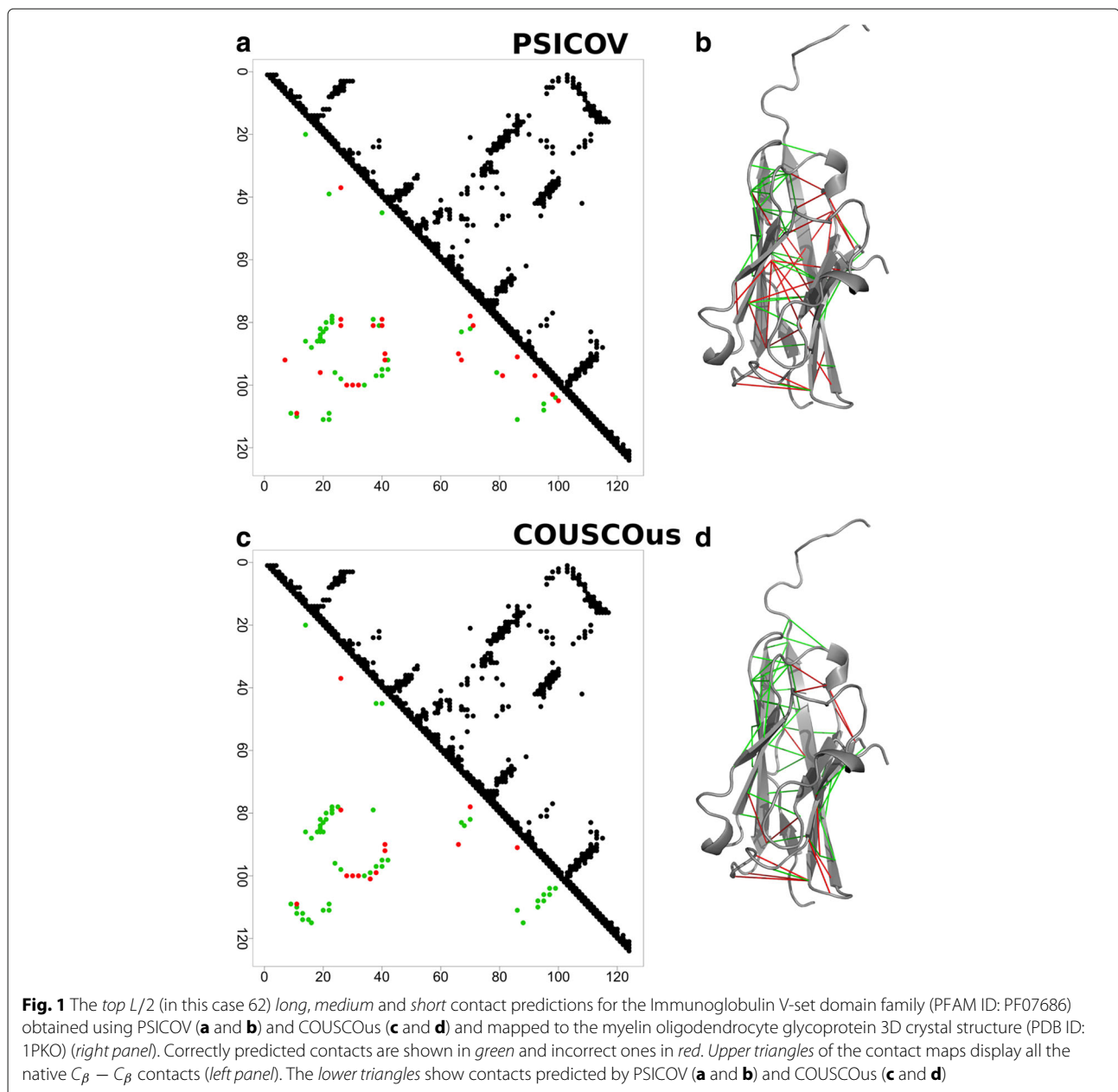


Table 1 Mean accuracies of COUSCOus vs. PSICOV on PSICOV benchmark dataset

	Accuracy top-L/10			Accuracy top-L/5			Accuracy top-L		
	Long	Medium	Short	Long	Medium	Short	Long	Medium	Short
PSICOV	0.6724	0.5709	0.4876	0.5816	0.4401	0.3716	0.3016	0.1787	0.1589
COUSCOus	0.7394	0.6151	0.5509	0.6494	0.4837	0.4037	0.3341	0.1892	0.1664

Higher mean accuracies in bold

Empirical Bayes covariance estimator

The most natural estimator of the covariance Σ is the sample covariance matrix S . The estimator is optimal in the classical settings with large number of samples and fixed low dimensions ($n > p$). However, it performs poorly in high-dimensional settings ($n \ll p$), see Johnstone [30]. The GLasso approach operates very well in this context, but the computational time required to reach convergence can be large in some cases such as for protein families with low number of sequences. As an alternative to the natural estimator S , several shrinkage estimators have been proposed in the literature [31, 32]. They take a weighted average of the sample covariance matrix S , with a suitable chosen target diagonal matrix. Jones et al. applied a smoothed covariance estimator that shrinks the matrix towards the shrinkage target $F = \text{diag}(\bar{S}, \bar{S}, \dots, \bar{S})$ [16]. In this work, we applied the empirical Bayes estimator proposed by Haff [24]:

$$\hat{\Sigma} = S + \frac{p-1}{n \text{tr}(S)} I_p, \quad (3)$$

where I_p represents the identity matrix of order p . In a Bayesian framework, it has been proven by Haff that this estimator is the best estimator of the form $a(S + ut(u)C)$, with $0 < a < 1/(n-1)$ and $u = 1/\text{tr}(S^{-1}C)$. Here $t(\cdot)$ is non-increasing and C an arbitrary positive definite matrix. It dominates estimators of the form aS by a substantial amount. More precisely, it has been proven that under the L_2 loss, the uniform reduction in the risk function is at least $100 \frac{p+1}{n+p} \%$. In this study, we performed the shrinkage until the adjusted covariance matrix $\hat{\Sigma}$ is no longer singular, i.e. is a positive-definite matrix. The adjusted covariance matrix $\hat{\Sigma}$ was finally applied in the GLasso algorithm to obtain the sparse precision matrix, that contains the degree of direct coupling between any pair of amino acids.

APC correction

The coevolution pair list is generated identically to the PSICOV final processing. For each MSA column pair i and j we compute the L_1 -norm out of the corresponding 20×20 submatrix in Θ (only contributions of the 20 amino acid types are considered):

$$C_{ij} = \sum_{ab} |\Theta_{ij}^{ab}| \quad (4)$$

Furthermore, we adjust the coupling score by the average product correction (APC), previously applied for MI by Dunn et al. [12] to reduce entropic and phylogenetic bias:

$$COUSCOus_{ij} = C_{ij} - \frac{\bar{C}_{(i-)}\bar{C}_{(-j)}}{\bar{C}} \quad (5)$$

with $\bar{C}_{(i-)}$ as the mean precision norm of column i and all other columns, $\bar{C}_{(-j)}$ as the corresponding for column j and \bar{C} as the mean precision norm of all coupling scores.

Random forest meta-classifier

As previously indicated, new hybrid methods that combine coevolution detecting tools with other sources of information such as protein physiochemical features outperforms single methods like PSICOV. However, we are convinced that improvements of single residue-residue contact detecting methods can boost new emerging hybrid techniques. We designed a RF meta-classifier that includes several contact prediction methodologies along with a small number of sequence-derived physiochemical features.

In particular, we built a RF classifier using the training set alignments from the MetaPSICOV study [22]. In total, we used 336 protein alignments where PSICOV was able to successfully provide contact predictions. The RF was trained using the following features:

Table 2 Mean X_d values of COUSCOus and PSICOV on PSICOV benchmark dataset

	X_d top-L/10			X_d top-L/5			X_d top-L		
	Long	Medium	Short	Long	Medium	Short	Long	Medium	Short
PSICOV	0.2694	0.2751	0.2239	0.2518	0.2564	0.2068	0.1930	0.1864	0.1422
COUSCOus	0.2816	0.2788	0.2302	0.2718	0.2646	0.2295	0.2239	0.2058	0.1671

Higher mean values in bold

Table 3 Mean AUC_{pr} values

	Long	Medium	Short
PSICOV	0.2150	0.2630	0.2715
COUSCOus	0.2447	0.2930	0.3014

Higher mean values in bold

- Contact detecting methodologies MI, MIp, BvN, PSICOV or COUSCOus, FreeContact and CCMpred
- Secondary structure and solvent exposure probabilities derived from PSIPRED [33]
- Shannon entropy using R [34] package bio3d [35]
- Hydrophobicity using R package Interpol [36]
- Amino acid physiochemical properties

The RF meta-classifier was trained using 500 trees with ten features and a max-depth of eight. We performed five-fold cross-validation while training the classifier and optimised the area under the curve (AUC) metric for performance.

Evaluation metrics and distance descriptions

The problem of predicting protein residue-residue contacts is well-known to be an extremely difficult one as on average only 3% of all possible residue pairs in known protein structures are identified to be real contacts. In the latest CASP11 challenge [37], this problem was tackled by dividing the contact prediction task into two categories. First, evaluation of predicted contacts using quality metrics like Accuracy and X_d [38] on reduced lists (RL). Second, evaluation of predicted contacts using quality metrics like Matthew's correlation coefficient (MCC) [39] and area under the precision-recall (AUC_{pr}) for full lists

(FL). The RL are usually defined by considering the top L/n predicted contacts where L is the length of the evaluation target or protein sequence and n is a small integer (e.g. 1, 5 or 10). RL metric accuracy is calculated as $\frac{TP}{TP+FP}$ where TP defines a correctly predicted contact and FP an incorrectly predicted contact. The second RL metric X_d represents the difference between the distance distributions of the predicted contacts and all pairs distance distributions in the 3D target structure. It is defined as

$$X_d = \sum_{i=1}^{15} \frac{P_{ip} - P_{ia}}{di \cdot 15},$$

with P_{ia} and P_{ip} are the percentages of pairs included in the i^{th} bin for the whole target and predicted contacts respectively. Additional details can be found in [38]. The FL metrics used in this study are AUC_{pr} as it is a robust metric for unbalanced classes and the Matthew's correlation coefficient [39] to evaluate all residue pairs for contact prediction.

Results and discussion

We first illustrate as an example in Fig. 1 the spatial proximity of the predicted contacts obtained by PSICOV (a and b) and COUSCOus (c and d) for the Immunoglobulin V-set domain (Protein family database [40] (PFAM) ID: PF07686). The upper triangles of the presented contact maps (Fig. 1a and c) display the native contacts. A residue-residue pair is hereby considered to be in contact if the two amino acids are proximal in the 3D structure, in particular if their $C_\beta-C_\beta$ (C_α in the case of glycine) distance is less than 8 Å ngström (Å). The lower triangles show the $L/2$ contact predictions (in this case 62) obtained by using either PSICOV or COUSCOus. Correctly predicted contacts are coloured in green and incorrect ones in red. Further, we mapped the top $L/2$ predictions to

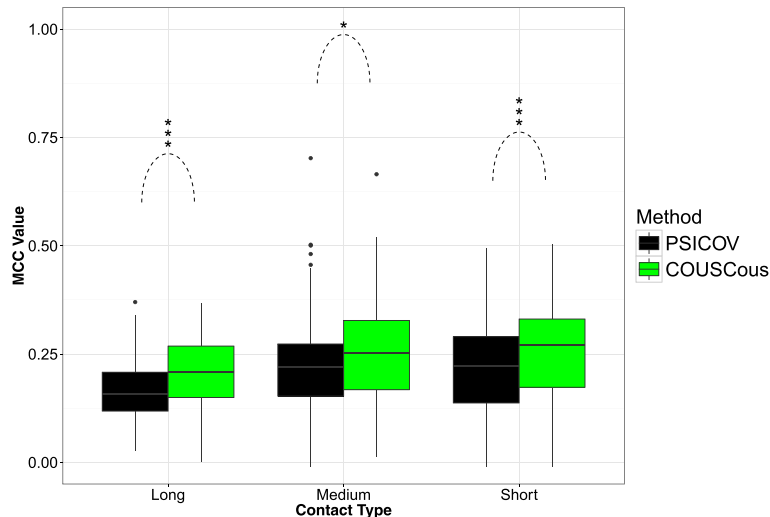


Fig. 2 MCC distributions for PSICOV benchmark proteins in case of long, medium and short range contacts predicted by PSICOV and COUSCOus. The stars represent statistical significance where * is used to represent P -value < 0.05 and *** is used to represent P -values < 0.001

the structure of the myelin oligodendrocyte glycoprotein (Protein Data Bank [41] (PDB) ID: 1PKO), solved at 1.45 Å resolution (Fig. 1b and d). Out of 62 possible contacts, COUSCOus correctly predicted 49 (accuracy: 0.79) compared to 39 (accuracy: 0.63) by PSICOV resulting in higher accuracy of COUSCOus. The figure shows (b) that the incorrect identified pairs are mainly located in loop regions at the top and the bottom. Hence, the pairs may still have distances less than 8Å considering that the unordered regions are not static as illustrated by a crystal structure. In contrast, incorrect predicted pairs from PSICOV (Fig. 1d), are distributed over the entire protein.

The performance of COUSCOus was evaluated on the original PSICOV benchmark test set using standard assessment metrics applied in CASP: accuracy, X_d , MCC and AUC_{pr} (see Methods). We distinguished between three types of contacts: short ($6 \leq$ residue separation < 12), medium ($12 \leq$ residue separation < 24) and long range contacts (residue separation ≥ 24).

We list in Table 1 the mean accuracies of COUSCOus and PSICOV for the top- $L/10$, top- $L/5$ and top- L for long, medium and short range predicted contacts on the original PSICOV benchmark set. For the top- $L/10$ predictions COUSCOus is 10, 8 and 13% more accurate than PSICOV for long, medium and short range contacts. Similarly, for the top- $L/5$ predictions we observed the similar gains in accuracy when using COUSCOus. For the top- L predictions we observed different accuracies for the three types of contacts. The gain in accuracy of 11% when using COUSCOus was similar for the long range contacts but the gain dropped to 6 and 5% for medium and short range contacts.

The second evaluation metric applied in this study, the X_d score, estimates the deviation of the distribution of distance in the RL sets ($L/10$, $L/5$ or L) of contacts from the distribution of the distances in all residue pairs within the protein (see Methods). Table 2 summarises the average X_d scores for COUSCOus and PSICOV. For the top- L predictions COUSCOus is more accurate than PSICOV on long, medium and short range contacts (16, 10 and 18%). For the top- $L/10$ and top- $L/5$ predictions we observed smaller improvements in X_d score ranging from 1 to 11%. Moreover, we compared the performance of COUSCOus and PSICOV on FL; considering all possible residue pairs. In Table 3 we summarise the mean AUC_{pr} values of the precision recall (PR) curves for COUSCOus and PSICOV. COUSCOus outperforms PSICOV with gains in accuracy of 14, 11 and 11% for long, medium and short range contacts, respectively.

We also performed an exhaustive analysis of COUSCOus and PSICOV w.r.t. MCC for long, medium and short range contact predictions. In Figure 2 we illustrate via box plot the distribution of the MCC values for the two prediction methods. It is apparent from the box plots

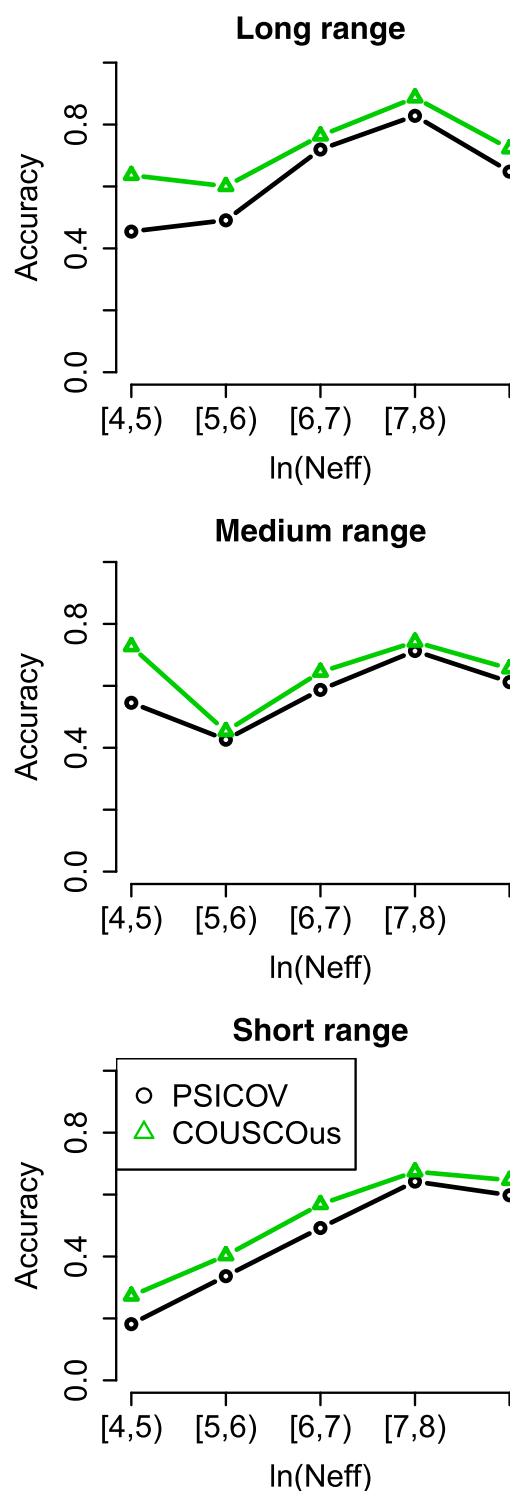


Fig. 3 Dependence of the performance of PSICOV and COUSCOus on the effective number of sequences (N_{eff}) in the MSAs. The performance is evaluated using accuracies for the top $L/10$ long, medium and short contacts. The solid line represents the averaged accuracies of the test set binned into five different categories of N_{eff} ($\ln(N_{eff})$: [4, 5], [5, 6], [6, 7], [7, 8], [8, 10]). COUSCOus outperforms PSICOV independent of the $\ln(N_{eff})$ in the test set

Table 4 Mean accuracies of COUSCOus vs. PSICOV on the CASP11 benchmark dataset

	Accuracy top-L/10			Accuracy top-L/5			Accuracy top-L		
	Long	Medium	Short	Long	Medium	Short	Long	Medium	Short
PSICOV	0.6687	0.5809	0.5278	0.5872	0.4809	0.4229	0.3383	0.2373	0.1820
COUSCOus	0.7385	0.6335	0.5965	0.6610	0.5285	0.4636	0.3828	0.2477	0.1958

Higher mean accuracies in bold

that COUSCOus is superior in predicting real residue contacts. To further test for significance we performed a t-test, after successfully testing for normal distribution and variance homogeneity, on the *MCC* distributions for different contact ranges. COUSCOus outperforms PSICOV on all types of contacts significantly, with *P*-values of 6×10^{-7} , 1.2×10^{-2} and 6×10^{-4} for long, medium and short range contacts, respectively.

Next, we analysed the dependence of the performance of PSICOV and COUSCOus with regard to the size of the protein family. In this case, we used the number of effective sequences in a MSA as comparison metric to account for the fact that highly similar homologous do not provide any additional contact information than a single one. Similar to Ma et al. [21] we grouped the test set members into five categories by $\ln(N_{eff})$: [4, 5), [5, 6), [6, 7), [7, 8), [8, 10), and calculated the averaged *L/10* accuracies for each group. Figure 3 shows clearly that COUSCOus outperforms PSICOV regardless of the $\ln(N_{eff})$ on long, medium and short range contacts. In addition, we tested the performance of COUSCOus and PSICOV on an independent test set from the latest CASP11 experiment. In Table 4 we show the mean accuracies of COUSCOus and PSICOV for the top-*L/10*, top-*L/5* and top-*L* for long, medium and short range predicted contacts. For the top-*L/10* predictions COUSCOus is 10, 9 and 13% more accurate than PSICOV for long, medium and short range contacts, respectively. Similarly, for the top-*L/5* predictions we observed gains in accuracy of 13, 10 and 9% when using COUSCOus. For the top-*L* predictions we observed different accuracies for the three types of contacts. A gain in accuracy of 13% is observable when using COUSCOus for the long range contacts but the gain dropped to 4 and 8% for medium and short range contacts.

Next, we designed two experiments using a RF meta-classifier where we combine contact detecting tools with protein sequence derived features. In the first case we

used predictions of PSICOV as a feature in the meta-classifier and in the second case we replaced PSICOV predictions with COUSCOus predictions. The classifier including COUSCOus results as a feature outperforms the classifier including PSICOV results for the top-*L/10* and top-*L/5* predictions on all types of contacts except for the top-*L* predictions for long range contacts (see Table 5).

In Additional file 2 we illustrate the mean accuracies of different contact detecting techniques. Our newly developed technique COUSCOus (green upper triangle) outperforms PSICOV (black points) and is equally well as FreeContact (red lower triangle) on all contact types. The best single contact detecting tool is CCMpred (magenta rectangle). Our simple RF meta-classifier that combines 6 single residue-residue contact detecting tools along with 4 sequence-derived features outperforms all single methods. However, MetaPSICOV, a multi-stage neural network hybrid method that combines five coevolution techniques along with a broad range of sequence-derived features is still the best performing method.

Discussion

In this work, we assessed the performance of our newly developed method COUSCOus in predicting residue-residue contacts from MSAs. In particular, the performance was tested in comparison to PSICOV on the original PSICOV benchmark test set as well as on CASP11 targets. On the RL sizes COUSCOus outperformed PSICOV substantially, with on average 10% gain in prediction accuracy for all types of contacts. Moreover, COUSCOus proved to be superior over PSICOV on FL sizes with average *AUC_{pr}* gains of 12%. With regard to *MCC* scores COUSCOus is even significantly outperforming PSICOV, illustrated with the help of box plots and hypothesis tests (see also Additional file 3). Further, we reported that COUSCOus's gain in accuracy is independent of the number of effective sequences in a given MSA.

Table 5 Mean accuracies of RF meta-classifier including COUSCOus or PSICOV as a feature on the PSICOV benchmark dataset

	Accuracy top-L/10			Accuracy top-L/5			Accuracy top-L		
	Long	Medium	Short	Long	Medium	Short	Long	Medium	Short
RF-PSICOV	0.7846	0.6946	0.6547	0.7047	0.5500	0.5140	0.3991	0.2439	0.2212
RF-COUSCOus	0.7881	0.7065	0.6618	0.7112	0.5674	0.5191	0.3984	0.2453	0.2225

Higher mean accuracies in bold

The main motivation of this work was to highlight that improvements of single residue-residue contact detecting tools, in this case PSICOV, might lead to improvements of new hybrid methods that combine contact detecting techniques with physiochemical and other sequence derived protein features. As proof of concept, we showed with the help of a simple RF meta-classifier that PSICOV should be replaced in hybrid classifiers by COUSCOus.

Conclusion

Jones et al. [16] demonstrated in their initial work that GLasso in principle performs excellently in identifying directly coupled columns within a MSA. In the present study, we highlighted that the application of a different shrinkage approach than the one used in PSICOV, the empirical Bayes covariance estimator, in combination with GLasso substantially increased the contact precision. The theoretically shown superiority of the empirical Bayes covariance estimator over simpler smoothed covariance estimators of the form aS is also valid within this application of contact detection from MSAs.

Furthermore, it is worth mentioning that other research fields that apply the GLasso procedure, such as gene network reconstruction, may also benefit from applying other shrinkage techniques. We are keen to investigate the effect of shrinkage in other graphical inference problems in future work.

Another important application that we are keen to investigate in future is the de novo structure prediction of proteins or protein complexes using COUSCOus or a hybrid classifier, including COUSCOus contact predictions as distance constraints, similarly to what have been applied in EVFold [4] and EVComplex [8], or PconsFold [42].

Additional files

Additional file 1: A complete list of all CASP11 protein codes applied in this study. (PDF 72.9 kb)

Additional file 2: Mean accuracies of several contact detecting methods for the top- $L/10$, $L/5$ and L predictions for all contact types applying the PSICOV benchmark dataset. (PDF 85.6 kb)

Additional file 3: Scatterplot comparing the accuracies of the top L contacts of PSICOV to COUSCOus, using sequence separation ≥ 6 . (PDF 78 kb)

Abbreviations

3D: three-dimensional; Å: Ångström; AUC: area under the curve; BvN: Bayesian network approach; COUSCOus: Contact predictiOn Using ShrInked COvariance; CASP11: Critical Assessment of techniques for protein Structure Prediction 11; DCA: Direct Coupling Analysis; FL: full lists; GLasso: graphical lasso; MCC: Matthew's correlation coefficient; MI: mutual information; MIp: corrected mutual information; MSA: multiple sequence alignment; PDB: Protein Data Bank; PFAM: Protein family database; PR: precision recall; PSICOV: Protein Sparse Inverse COvariance; RF: Random forest; RL: reduced lists

Acknowledgements

We thank Erik van Nimwegen and Lukas Burger for providing code to implement their Bayesian network approach.

Funding

Not applicable.

Availability of data and materials

Implementation of COUSCOus is available as R package (<https://cran.rstudio.com/web/packages/COUSCOus/>). Raw data (alignments) and other applied software can be found as referred in the Methods section.

Authors' contributions

RR contributed to the concept, programmed COUSCOus, coordinated the project and wrote the manuscript; RM programmed the RF and assisted in the analysis and writing; KK, MA, and EU assisted in the analysis, reviewed and edited the manuscript; ME and HB contributed to the concept, reviewed and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

¹Computational Science and Engineering, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar. ²Division of Biomedical Informatics, Sidra Medical and Research Center, Doha, Qatar.

Received: 17 June 2016 Accepted: 1 December 2016

Published online: 15 December 2016

References

1. Yanofsky C, Horn V, Thorpe D. Protein structure relationships revealed by mutual analysis. *Science (New York NY)*. 1964;146:1593–4.
2. Fitch WM, Markowitz E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet*. 1970;4(5):579–93.
3. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet*. 2013;14(4):249–61.
4. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE*. 2011;6(12):e28766.
5. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. *Nat Biotechnol*. 2012;30(11):1072–80.
6. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*. 2012;149(7):1607–21.
7. Kosciolk T, Jones DT. De Novo Structure Prediction of Globular Proteins Aided by Sequence Variation-Derived Contacts. *PLoS ONE*. 2014;9(3):e92197.
8. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*. 2014;3:e03430.
9. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*. 2014;3:e02030.
10. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*. 2005;44(19):7156–65.
11. Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. *Bioinformatics (Oxford England)*. 2005;21(22):4116–24.
12. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics (Oxford England)*. 2008;24(3):333–40.
13. Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol*. 2010;6(1):e1000633.
14. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA*. 2009;106(1):67–72.

15. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA*. 2011;108(49):E1293–301.
16. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford England)*. 2012;28(2):184–90.
17. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlinear Soft Matter Phys*. 2013;87(1):012707.
18. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA*. 2013;110(39):15674–9.
19. Eickholt J, Cheng J. Predicting protein residue-residue contacts using deep networks and boosting. *Bioinformatics*. 2012;28(23):3066–72.
20. Skwark MJ, Abdel-Rehim A, Elofsson A. PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics*. 2013;29(14):1815–6.
21. Ma J, Wang S, Wang Z, Xu J. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*. 2015;31(21):3506–13.
22. Jones DT, Singh T, Kosciolk T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*. 2015;31(7):999–1006.
23. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. *Ann Stat*. 2006;34(3):1436–62.
24. Haff LR. Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix. *Ann Stat*. 1980;8(3):586–97.
25. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins Struct Funct Genet*. 2002;48(4):611–7.
26. Bakan A, Dutta A, Mao W, Liu Y, Chennubhotla C, Lezon TR, et al. Evol and ProDy for bridging protein sequence evolution and structural dynamics. *Bioinformatics*. 2014;30(18):2681–3.
27. Kaján L, Hopf TA, Kalaš M, Marks DS, Rost B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinforma*. 2014;15(1):85.
28. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008;9(3):432–41.
29. Lauritzen SL. *Graphical Models*, 1st ed. Oxford: Oxford University Press; 1996.
30. Johnstone IM. On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *Ann Stat*. 2001;29(2):295–327.
31. James W, Stein C. Estimation with quadratic loss. In: *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* Berkeley: University of California Press; 1961. p. 361–379.
32. Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J Empir Financ*. 2003;10(5):603–21.
33. Jones DT. Protein secondary structure prediction based on position-specific matrices. *J Mol Biol*. 1999;292:195–202.
34. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna; 2014. <http://www.R-project.org/>.
35. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Cavas LSD. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*. 2006;22(21):2695–6.
36. Heider D, Hoffmann D. Interpol: An R package for preprocessing of protein sequences. *BioData Min*. 2011;4(1):16.
37. Park H, DiMaio F, Baker D. CASP11 refinement experiments with ROSETTA. *Proteins*. 2016;84:1097–0134.
38. Izarzugaza JMG, Graña O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. *Proteins Struct Funct Bioinforma*. 2007;69(S8):152–8.
39. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta Protein Struct*. 1975;405(2):442–51.
40. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42(D1):D222–D230.
41. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235–42.
42. Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A. PconsFold: improved contact predictions improve protein models. *Bioinformatics*. 2014;30(17):i482–i488.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

