BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 2 2012, pages 184–190
doi:10.1093/bioinformatics/btr638

Sequence analysis

Advance Access publication November 17, 2011

PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments

David T. Jones[1,*], Daniel W. A. Buchan[1], Domenico Cozzetto[1] and Massimiliano Pontil[2]

[1]Department of Computer Science, Bioinformatics Group and [2]Department of Computer Science, Centre for Computational Statistics and Machine Learning, University College London, Malet Place, London WC1E 6BT, UK

Associate Editor: Mario Albrecht

ABSTRACT

Motivation: The accurate prediction of residue–residue contacts, critical for maintaining the native fold of a protein, remains an open problem in the field of structural bioinformatics. Interest in this long-standing problem has increased recently with algorithmic improvements and the rapid growth in the sizes of sequence families. Progress could have major impacts in both structure and function prediction to name but two benefits. Sequence-based contact predictions are usually made by identifying correlated mutations within multiple sequence alignments (MSAs), most commonly through the information-theoretic approach of calculating mutual information between pairs of sites in proteins. These predictions are often inaccurate because the true covariation signal in the MSA is often masked by biases from many ancillary indirect-coupling or phylogenetic effects. Here we present a novel method, PSICOV, which introduces the use of sparse inverse covariance estimation to the problem of protein contact prediction. Our method builds on work which had previously demonstrated corrections for phylogenetic and entropic correlation noise and allows accurate discrimination of direct from indirectly coupled mutation correlations in the MSA.

Results: PSICOV displays a mean precision substantially better than the best performing normalized mutual information approach and Bayesian networks. For 118 out of 150 targets, the L/5 (i.e. top-L/5 predictions for a protein of length L) precision for long-range contacts (sequence separation >23) was ≥ 0.5, which represents an improvement sufficient to be of significant benefit in protein structure prediction or model quality assessment.

Availability: The PSICOV source code can be downloaded from http://bioinf.cs.ucl.ac.uk/downloads/PSICOV

Contact: d.jones@cs.ucl.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 20, 2011; revised on November 8, 2011; accepted on November 14, 2011

1 INTRODUCTION

Residue–residue contacts are known to play critical roles in maintaining the native fold of proteins and guiding protein folding (Gromiha and Selvaraj, 2004). Such residue contacts are usually well separated with regards to the primary sequence but display close proximity within the 3D structure. Although different geometric criteria for defining contacts have been given in the literature, typically, contacts are regarded as those pairs of residues where the C-β atoms approach within 8 Å of one another (C-α in the case of Glycine) (Fischer *et al.*, 2001; Graña *et al.*, 2005a, b; Hamilton and Huber, 2008).

It has long been observed that with sufficient correct information about a protein's residue–residue contacts, it is possible to elucidate the fold of the protein (Gobel *et al.*, 1994; Olmea and Valencia, 1997). However, accurate prediction of intrachain residue–residue contacts from sequence data alone remains an open problem. A solution would yield benefits for a range of endeavours including fold recognition, *ab initio* protein folding, 3D model quality assessment and *de novo* protein design.

The majority of successful approaches for contact prediction attempt to extract contact information from the content of a multiple sequence alignment (MSA); usually through the simple identification of correlated mutations (Ashkenazy and Kliger, 2010; Gobel *et al.*, 1994; Neher, 1994; Pollock and Taylor, 1997) or by calculating the mutual information (MI) between columns in the MSA (Burger and van Nimwegen, 2010; Dunn *et al.*, 2008). The underlying rationale rests on the fact that any given contact critical for maintaining the fold of a protein will constrain the physicochemical properties of the amino acids involved. Should a given contacting residue mutate and potentially perturb the properties of the contact, then its contacting partner will be more likely to mutate to a physicochemically complementary amino acid residue, to ensure the native fold of the protein remains stabilized. Turning this observation around, pairs of residues seen to co-evolve in tandem and thus preserving their relative physiochemical properties, are likely candidates to form contacts. Such linked mutational events are often referred to simply as 'correlated mutations'. The physicochemical similarity of residue pairs is typically scored with the McLachlan matrix (McLachlan, 1971), although recent work has called into question its use (Burger and van Nimwegen, 2010; Lena *et al.*, 2011). To date, a wide variety of information theory and machine learning algorithms have been applied to the problem of correlated mutation analysis including MI, Neural Networks, Support Vector Machines and linear regression models (Fariselli *et al.*, 2001; Hamilton *et al.*, 2004; MacCallum, 2004; Martin *et al.*, 2005; Pollastri and Baldi, 2002; Punta and Rost, 2005; Shao and Bystroff, 2003; Xue *et al.*, 2009; Yuan, 2005).

*To whom correspondence should be addressed.

A thorough review of the currently available methods can be found in Horner *et al.* (2008)

Despite significant attention, what success there has been in predicting structurally important contacts has generally been rather modest, and progress in the field has remained slow (Ezkurdia *et al.*, 2009). Even the best methods display low accuracies and, while the predictions are better than random, an accuracy between 20% and 40% is typical (Burger and van Nimwegen, 2010; Fariselli *et al.*, 2001; Hamilton *et al.*, 2004; Pollastri and Baldi, 2002; Punta and Rost, 2005). This low accuracy, indicating a large number of false positives in the prediction, is a consequence of two further features of the MSA: additional phylogenetic residue correlations and linked chains of covariance. Both these factors add considerable noise to the signal contained in the MSA (Lapedes *et al.*, 1999). Many prediction methods attempt to enrich the prediction set by filtering out the false positives using simple heuristic rules. The most direct method being to filter out contacts in excess of the expected number of contacts each residue can make (Yuan, 2005). Additionally, statistical methods such as bootstrapping (Olmea and Valencia, 1997), estimating the background phylogenetic noise (Dunn *et al.*, 2008) or the use of ancillary predictions such as secondary structure prediction (Shao and Bystroff, 2003) have also been applied with a degree of success.

As already mentioned, for purely sequence-based approaches to contact prediction, there are two main sources of noise in the analysis of correlated mutations: phylogenetic bias and indirect coupling effects (Lapedes *et al.*, 1999). The latter problem of determining direct from indirect coupling effects in correlation mutation analysis seems to have received the lesser amount of attention in the literature. Lapedes *et al.* related the problem of decoupling mutation correlations in sequence alignments to the inverse Ising problem in statistical physics, and proposed a solution based on maximization of entropy. This idea has been further refined using a message-passing algorithm (Weigt *et al.*, 2009). Recently, Burger and van Nimwegen (2010) used a computationally efficient Bayesian network approach to tackle the same indirect coupling problem.

In this article, we propose the use of sparse inverse covariance estimation techniques (Meinshausen and Bühlmann, 2006) to deal with the coupling effects and test our method on a benchmark set of experimentally determined protein structures. These graphical inference techniques are simple and yet remarkably powerful, and while they have been applied to other areas of computational biology such as gene network discovery (Friedman *et al.*, 2008), they have not previously been applied to sequence analysis problems.

2 METHODS

2.1 Mutual information

The most common method for identifying correlated mutations in MSAs is to calculate the MI between two sites:

$$\mathrm{MI} = \sum_{ab} f(A_i B_j) \log \frac{f(A_i B_j)}{f(A_i) f(B_j)} \tag{1}$$

where $f(A_i B_j)$ is the observed relative frequency of amino acid pair ab at columns ij, $f(A_i)$ is the observed relative frequency of amino acid type a at column i and $f(B_j)$ is the observed frequency of amino acid type b at column j. The usefulness of MI to predict protein contacts can be further enhanced by normalization to take into account bias in the sequence family being analysed (Dunn *et al.*, 2008). Both entropic bias and some measure of phylogenetic bias can be removed by this kind of normalization. Entropic bias refers to the false signals that arise from either having insufficient sequences to properly sample residue types or from extremes of conservation (sites with very high or very low conservation can lead to spurious predictions of contacts). Phylogenetic bias refers to the false signals due to functionally related clusters of residues appearing to co-evolve according to the structure of the underlying evolutionary tree. Despite doing a reasonable job of correcting for entropic and phylogenetic bias, such normalization cannot help reduce the effects of chaining within the protein's contact graph (i.e. indirect coupling effects where direct coupling between sites AB and BC can result in observed correlations between AC, even though no direct interaction exists between AC). Here we attempt to correct for these effects using sparse inverse covariance estimation.

2.2 Inferring directly coupled sites using covariance

The starting point of our method is to consider an alignment with m columns and n rows, where each row represents a different homologous sequence and each column a set of equivalent amino acids across the evolutionary tree, with gaps considered as an additional amino acid type. We can compute a $21m$ by $21m$ sample covariance matrix as follows:

$$S_{ij}^{ab} = \frac{1}{n} \sum_{k=1}^{n} (x_i^{ak} - \bar{x}_i^a)(x_j^{bk} - \bar{x}_j^b) \tag{2}$$

where x_i^{ak} is a binary variable ($x \in \{0,1\}$) indicating the presence or absence of amino acid type a at column i in row k and x_j^{bk} the equivalent variable for observing residue type b at column j in row k. This calculation of covariance based on binary amino acid variables is similar to that used by Halabi *et al.* (2009) to determine independent evolutionary units. Based on the standard identity for covariance of $Cov(X,Y) = E(XY) - E(X)E(Y)$, and the expectation of a binary variable $E(x)$ being equivalent to the probability of a positive observation $p(x=1)$, this simplifies to the following expression based on the observed marginal frequencies of amino acids ($f(A_i)$ and $f(B_j)$) and amino acid pairs ($f(A_i B_j)$) at the given sites in the set of aligned sequences:

$$S_{ij}^{ab} = E(x_i^a x_j^b) - E(x_i^a)E(x_j^b) = f(A_i B_j) - f(A_i) f(B_j) \tag{3}$$

Any individual element of this matrix gives the covariance of amino acid type a at position i with amino acid type b at position j. By calculating the matrix inverse of the covariance matrix, the *precision* or *concentration* matrix (Θ) is obtained, from which a matrix of partial correlation coefficients for all pairs of variables can be calculated as follows:

$$\rho_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} \tag{4}$$

In the simplest case, a partial correlation coefficient can be calculated between two random variables with the controlling effect of a third random variable taken into account. The partial correlation matrix above, however, gives the correlations between all pairs of variables with the controlling effects of all other variables taken into account [e.g. see Bühlmann and van de Geer (2011); Chapter 13, Section 13.4]. Here, the partial correlation matrix gives the correlation between any pair of amino acids at any two sites, *conditional on the frequencies of amino acids at all other sites*. Thus, assuming the sample covariance matrix can in fact be inverted, the inverse covariance matrix provides information on the degree of direct coupling between pairs of sites in the given MSA. Off-diagonal elements of the inverse covariance matrix which are significantly different from zero are indicative of pairs of sites which have strong direct coupling (and are likely to be in direct physical contact in the native structure).

Unfortunately, the empirical covariance matrices produced in this application are guaranteed to be singular due to the fact that not every amino acid will be observed at every site, even in very large families, and thus there will be more variables than observations. Similar issues occur in the areas of finance (in calculating correlations between stock prices) and in gene

network reconstruction where the number of observed variables is again often less than the dimensionality of the problem. Although different approaches have been proposed to allow inverse covariance estimation where the sample covariance matrix cannot be directly inverted, one of the most powerful techniques is that of sparse inverse covariance estimation. In the absence of other constraints on obtained solutions, the expected sparsity of the inverse covariance matrix itself provides a powerful self-contained constraint on the obtained solution. In general terms, where an inverse covariance estimate is constrained to be sparse, the non-zero terms tend to more accurately relate to correct positive correlations in the true inverse covariance matrix. The expectation of sparsity in this application is well justified from observations of contacts in known protein structures, where on average only around 3% of all residue pairs are observed to be in direct contact.

2.3 Sparse inverse covariance estimation

The problem of sparse inverse covariance estimation has previously been studied by different authors, for example see Banerjee *et al.* (2008); Friedman *et al.* (2008); Meinshausen and Bühlmann (2006); Yuan and Lin (2007) and the additional references therein. Here, we follow the formulation of Banerjee *et al.* (2008), which is known as the *graphical Lasso* method, and the implementation provided by Friedman *et al.* (2008). We summarize the main idea behind this method and comment on how the algorithm of (Friedman *et al.*, 2008) can be used to solve the problem.

Let S be the empirical covariance matrix computed from a sequence of d-dimensional vectors, $x^1,\ldots,x^n$, sampled from some fixed but unknown probability distribution. Matrix S can be computed as $S_{ij}=\frac{1}{n}\sum_{k=1}^{n}(x_i^k-\bar{x}_i)(x_j^k-\bar{x}_j)$, for every $i,j=1,\ldots,d$, where $\bar{x}$ is the empirical mean. The graphical Lasso is a statistical method which estimates the inverse covariance of the data by minimizing the objective function:

$$\sum_{ij=1}^{d} S_{ij}\Theta_{ij} - \log\det\Theta + \rho\sum_{ij=1}^{d}|\Theta_{ij}| \tag{5}$$

In this function, the $d\times d$ matrix Θ is required to be symmetric and positive definite. The first two terms in (5) can be interpreted as the negative log-likelihood of the inverse covariance matrix Θ under the assumption that the data distribution is a multivariate Gaussian.

The third term in (5) is the ℓ_1-norm of matrix Θ, a special kind of regularization or penalty term. One main insight behind such a regularizer is that it favours *sparse* solutions, in the sense that many of the components of the positive definite matrix $\hat{\Theta}$ which minimizes (5) will be zero. The amount of sparsity in $\hat{\Theta}$ is controlled by the positive parameter ρ, which needs to be chosen by the user. Typically, as ρ increases the number of zero components of $\hat{\Theta}$ increases, eventually reaching the point where all the components are equal to zero.

The need for a sparse inverse covariance matrix is well understood when the data probability distribution is a multivariate Gaussian distribution with covariance Σ. In this case, it is well known that $\Sigma^{-1}_{ij}=0$ if and only if the variables i and j are conditionally independent. Hence, if we know *a priori* that many pairs of variables are conditionally independent, it seems appropriate to estimate the inverse covariance by the graphical Lasso method. The parameter ρ can be selected to reach a desired sparsity level, provided this is known *a priori* .

An important rationale for sparse estimation comes from the observation that in many practical applications the size of the matrix Θ is much larger than the number of data points n, but the underlying inverse covariance matrix which we wish to estimate is known to be sparse. Under this assumption and certain technical conditions, it has been shown that the solution $\hat{\Theta}$ of the above problem is a good estimate of the inverse covariance. More importantly, the sparsity pattern of $\hat{\Theta}$, namely the set of non-zero entries of this matrix, is close to the sparsity pattern of Σ^{-1}, thereby providing a valuable tool for selecting the pairs of conditionally dependent variables, see for example Bühlmann and van de Geer (2011) and the associated references therein.

We find the solution $\hat{\Theta}$ by solving the dual optimization problem (Banerjee *et al.*, 2008) and using the block coordinate descent technique described in Friedman *et al.* (2008). This formulation also allows one to use the more general penalty term $\sum_{ij} r_{ij}|\Theta_{ij}|$ where the r_{ij}'s are some positive parameters, which incorporate prior knowledge on the relative important pairs of components.

A final point worth noting here is that while all of the above is under the assumption that input data distribution is multivariate Gaussian, it has been shown by Banerjee *et al.* (2008) that this dual optimization solution also applies to binary data (as is the case in this application).

2.4 Shrinking the sample covariance matrix

Although the graphical Lasso approach works well in dealing with singular or poorly conditioned sample covariance matrices, the time taken to reach convergence can be problematic in some cases (e.g. families with few sequences or highly conserved regions). To speed up convergence, particularly in the worst cases, we condition the sample covariance matrix by shrinking towards a highly structured unbiased estimator:

$$S' = \lambda F + (1-\lambda)S \tag{6}$$

where F is the structured estimator matrix and $\lambda\in[0,1]$ is the so-called shrinkage parameter. The shrinkage target we used was $F=diag(\bar{S},\bar{S},\ldots,\bar{S})$ i.e. the identity matrix scaled by the mean sample variance (mean of the sample covariance matrix diagonal values). Although various approaches have been proposed to choose the ideal shrinkage parameter [e.g. Ledoit and Wolf (2003)], these methods either do not apply to binary variables or are based on unsuitable shrinkage targets. Here, therefore, we use the simple *ad hoc* approach of gradually increasing λ until the adjusted covariance matrix is no longer singular i.e. has no remaining negative eigenvalues (tested by Cholesky decomposition). Given the large number of dimensions, this shrinking procedure is generally much faster than the more common approach of truncating negative eigenvalues after finding all the eigenvalues and eigenvectors of the matrix.

Having conditioned the covariance matrix by shrinking, the graphical Lasso algorithm is then used to compute its sparse inverse.

2.5 Final processing

To arrive at the final predictions of contacting residues, for alignment columns i and j, the ℓ_1-norm is calculated for the 20×20 submatrix of Θ corresponding to the 20×20 amino acid types ab observed in the two alignment columns (contributions from gaps are ignored):

$$S_{ij}^{\text{contact}} = \sum_{ab}|\Theta_{ij}^{ab}| \tag{7}$$

To calculate a final score which has reduced entropic and phylogenetic bias, we can correct the raw precision norms $S_{ij}^{contact}$ using the same average product correction (APC) used to adjust MI for background effects as described by Dunn *et al.* (2008). Thus, the final PSICOV score for positions i and j is given by:

$$PC_{ij} = S_{ij}^{\text{contact}} - \frac{\bar{S}_{(i-)}^{\text{contact}}\bar{S}_{(-j)}^{\text{contact}}}{\bar{S}^{\text{contact}}} \tag{8}$$

where $\bar{S}_{(i-)}^{contact}$ is the mean precision norm between alignment column i and all other columns, $\bar{S}_{(-j)}^{contact}$ is the equivalent for alignment column j, and $\bar{S}^{contact}$ is the mean precision norm across the whole alignment. Finally, this background corrected score can easily be converted into an estimated positive predictive value by fitting a logistic function to the observed distribution of scores.

2.6 Experimental details

A program to implement this method, called PSICOV (Protein Sparse Inverse COVariance), was written in C and linked to the glasso

Fortran code as obtained from the CRAN archive (http://cran.r-project.org/web/packages/glasso). Good initial results were obtained with the lasso regularization parameter ρ set to a constant value of 0.001. However, we find that slightly better results can be obtained (at the expense of a 2–3× increase in calculation time) by iteratively adjusting ρ to achieve a target density of 3% (i.e. 3% non-zero terms in the final precision matrix). This target density was chosen to roughly correspond to the expected fraction of contacting residue pairs in globular protein domains. We tried both the exact and approximate algorithms implemented in the glasso code and found the quicker to calculate approximate solution (Meinshausen and Bühlmann, 2006) was often just as good as (and occasionally better than) the exact solution. However, with recent upgrades in the glasso code (V1.7 and later), the speed advantage of the approximation over the exact method is now only around 50% and so all results presented here are based on exact solutions.

MIp values were calculated as described in Dunn *et al.* (2008). Results for the method of Burger and van Nimwegen (2010) were generated using Perl scripts and C++ source code kindly provided by the authors. The software as provided by Burger and van Nimwegen employs the same product correction used in PSICOV and MIp, but it does not include the knowledge-based priors discussed in their paper.

Alignments were generated for Pfam families with ≥1000 sequences [http://pfam.sanger.ac.uk, Finn *et al.* (2010)] where a highly resolved (resolution ≤1.9 Å) X-ray crystallographic structure was available, the target protein was known to be a biological monomer, and where the structure comprised a single copy of the Pfam domain. Target sequences shorter than 50 residues and longer than 275 were not considered, giving a final total of 150 chains (listed in Supplementary Material). The actual target sequences were derived from the C-α ATOM records in the PDB files [http://www.pdb.org, Berman *et al.* (2000)]. As the seed alignments in Pfam are often derived from structure-based alignments, our alignments were generated automatically using the jackhmmer program, which is part of the HMMER 3.0 package (http://hmmer.org). For each of the 150 PDB-derived sequences, three iterations of jackhmmer, searching against the UNIREF100 data bank (Magrane and the UniProt Consortium, 2011), were used to find and align homologues. In the final alignments, duplicate rows (i.e. sequences 100% identical over the length of the alignment) and columns with gaps in the target sequence were deleted so that the number of columns in each alignment equalled the target sequence length. Numbers of distinct sequences in each alignment ranged from 511 (AraC-like ligand binding domain) to 74 836 (response regulator receiver domain). A full list of the 150 targets used, along with PDB codes, Pfam identifiers, chain lengths and numbers of aligned sequences is provided as Supplementary Material. Full datasets will be made available alongside the program source code.

In computing the marginal relative frequencies [$f(A_iB_j)$, $f(A_i)$ and $f(B_i)$], simple BLOSUM-style sequence weighting (Henikoff and Henikoff, 1992) and pseudocount regularisation are used. The sequence weighting was carried out with a threshold of 62% sequence identity (as used in the construction of the standard BLOSUM62 matrix), and a pseudocount of 1 was used in all experiments. The same weighting and pseudocount procedures were also applied to the calculation of MIp values.

For benchmarking purposes, a true contact was defined as any pair of residues where the C-β to C-β distance (C-α to C-α distance in the case of glycine) was <8 Å. This corresponds to the standard contact prediction evaluation criteria used in the biennial CASP experiment (Ezkurdia *et al.*, 2009). In addition, all atom contacts were used, where a contact is defined when any pair of heavy atoms in the two residues are closer than <6 Å.

3 RESULTS AND DISCUSSION

Figure 1, Tables 1 and 2 summarize the performance of PSICOV in terms of precision (alternatively known as positive predictive value in some texts) compared to the Bayesian approach of Burger and van Nimwegen (2010) (B&vN), and MIp when applied to the benchmark set of 150 proteins as described in the Section 2. Results are also subdivided by sequence separation of predicted pairs, as it is expected that contacting residue pairs far apart in the sequence will be harder to predict than those close together. Figure 1 also shows the relative effects of sequence weighting on both PSICOV and MIp. It is quite clear that PSICOV is greatly superior to the best MI approach in predicting contacts. There is very little difference in results across all four methods for the top 100 ranked predictions at lower sequence separations, but this is only to be be expected as there are a limited number of true contacts to find in those ranges. The large gap between MIp and the two direct coupling approaches for sequence range 5–9 suggests that indirect coupling effects have a stronger effect for residue pairs close in the sequence. Possibly this is a result of the rigidity of secondary structure elements amplifying indirect coupling effects, and we are keen to investigate that hypothesis in future work.

The earlier Bayesian network approach also performs better than MIp, in agreement with the original results of Burger and van Nimwegen (2010), but for long range contact prediction, it does not outperform MIp with sequence weighting. The importance of applying sequence weighting to the calculation

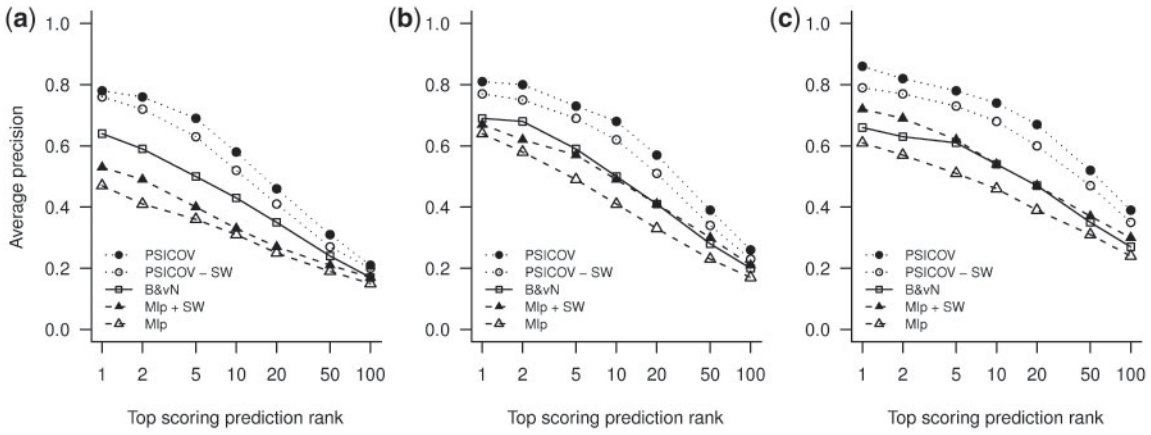

Fig. 1. Average precision of the top *n* ranked predictions. The graphs compare performance between normalized MI scoring as per Dunn *et al.* (2008) (MIp), normalized MI scoring with sequence weighting (MIp + SW), the method of Burger and van Nimwegen (2010) (B&vN) and our own PSICOV method both with (PSICOV) and without (PSICOV-SW) sequence weighting. The three panels indicate the average precision at sequence separations between 5 and 9 residues (**a**), between 10 and 23 residues (**b**) and >23 residues (**c**).

Table 1. Mean precision values for the top-L or top-L/2 contacts divided by sequence separation ranges where the C-β-C-β distance <8 Å

	$[i-j]>4$				$[i-j]>8$			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV	0.43	0.56	0.69	0.74	0.4	0.54	0.67	0.73
B&vN	0.32	0.41	0.52	0.6	0.29	0.39	0.51	0.58
MIp + SW	0.29	0.34	0.42	0.48	0.3	0.38	0.46	0.53
	$[i-j]>11$				$[i-j]>23$			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV	0.39	0.52	0.66	0.73	0.33	0.46	0.62	0.69
B&vN	0.28	0.37	0.5	0.56	0.24	0.33	0.45	0.5
MIp + SW	0.3	0.38	0.48	0.54	0.27	0.35	0.45	0.51

Table 2. Mean precision values for the top-L or top-L/2 contacts divided by sequence separation ranges where the any heavy atom distance is <6 Å

	$[i-j]>4$				$[i-j]>8$			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV	0.5	0.63	0.76	0.82	0.46	0.6	0.74	0.81
B&vN	0.37	0.47	0.59	0.67	0.34	0.44	0.57	0.64
MIp + SW	0.37	0.43	0.53	0.6	0.35	0.43	0.53	0.6
	$[i-j]>11$				$[i-j]>23$			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
PSICOV	0.44	0.58	0.73	0.8	0.38	0.52	0.68	0.76
B&vN	0.33	0.43	0.56	0.63	0.28	0.38	0.51	0.58
MIp + SW	0.35	0.44	0.55	0.62	0.32	0.4	0.52	0.58

of MI was previously highlighted by Buslje *et al.* (2009), and similarly, PSICOV also benefits from sequence weighting. Nevertheless, even without sequence weighting, PSICOV still significantly outperforms all the other methods across all sequence separation ranges.

Looking at the rank-50 results for the hardest case of sequence separation >23, PSICOV is able to predict over 50% of true contacts compared to just under 40% in the case of MIp (and around 20% in the case of raw MI). This level of accuracy is likely to be of very significant benefit in protein structure prediction applications. In this respect, Table 1 gives a more accurate view of performance, where precision values are given for the top L, L/2, L/5 and L/10 predicted contacts, where L is the length of the target protein. These thresholds are commonly used in independent assessments of contact prediction methods (Ezkurdia *et al.*, 2009). Although, for simplicity, C-β distances are most commonly used in benchmarking contact prediction methods, a more realistic criterion is to define contacts between any heavy atom in the two amino acids. Table 2 gives the equivalent benchmark data for all heavy atom contacts. Again, PSICOV produces substantially more correct predictions than the other tested methods across all sequence separation ranges and for all ranking subsets. For the all-atom contact definition, we find that for 59% of the targets the L precision is >0.5. With 1 predicted contact per residue and a precision of over 50%, this information would be sufficient to narrowly constrain the fold of a given protein.

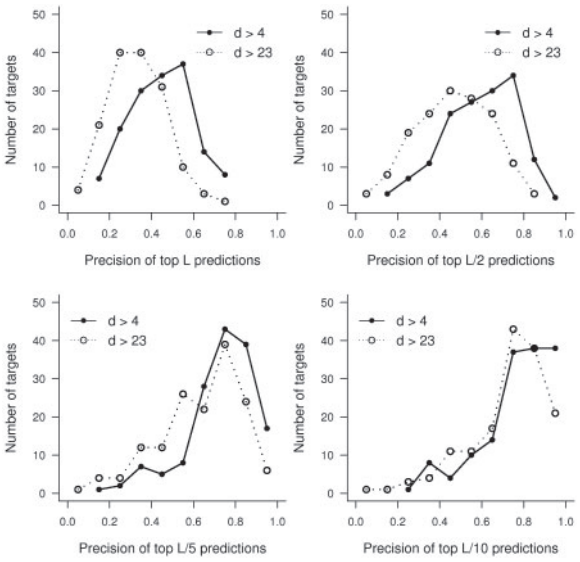

Fig. 2. Top-L, top-L/2, top-L/5 and top-L/10 precision values for all 150 benchmark targets for sequence separation >4 and >23 residues.

Average performance only provides part of the picture, of course. Of perhaps more interest is the distribution of benchmark results across the set of targets. Figure 2 shows the range of precision values for all 150 targets, and from this we see that the L precision distribution is fairly symmetric, with around 25 poor results (L precision <0.3) and a similar number of excellent results (L precision >0.6). For the smaller subsets of predicted contacts, as expected, the precision distributions skew to higher values, with 85% of targets having an L/10 precision of 0.6 or better. Unfortunately, it is difficult to find a single factor that determines prediction accuracy. No correlations were observed between prediction accuracy and sequence length or number of gaps in the alignment. A weak correlation (Spearman $\rho = 0.3$) was observed with mean alignment entropy (basically the degree of sequence conservation). As might be expected, there was a moderate correlation with the number of aligned sequences (Spearman $\rho = 0.596$). No stronger correlation was observed (Spearman $\rho = 0.588$) with the effective number of aligned sequences (i.e. taking the sequence redundancy into account). Figure 3 shows these two relationships.

Finally, to look at the added value of the graphical Lasso procedure over and above the widely used MIp measure, Figure 4 shows the L/2 precision for PSICOV compared with that of MIp. Clearly the treatment of indirect coupling effects by the graphical Lasso has a substantial positive effect on prediction accuracy in almost every case. In five cases, the graphical Lasso analysis seems to slightly reduce accuracy, although it is not clear to us what is special about these cases. The worst case example is that of one of the shortest targets 1M8A-A (61 residues) where PSICOV predicts 17 contacts, compared with 13 in the case of MIp. This is hardly a significant difference, but it is still apparent that the direct coupling analysis is not helping in this case, for whatever reason.

Looking at the performance of PSICOV both in terms of average performance and performance in the best cases, it is clear that it can provide spatial constraints for protein modelling beyond those that can be obtained by even the best machine learning-based contact prediction approaches. In the CASP8 experiment, the best contact prediction server had a precision of 30% for the top L/5 predicted contacts (sequence separation >23) over a common subset of nine targets (Ezkurdia *et al.*, 2009). Although the benchmark set in this case is different, PSICOV achieves an average accuracy of 62% using the same criteria. It is also possible that PSICOV could be further improved by incorporating it in a more general machine learning-based contact prediction

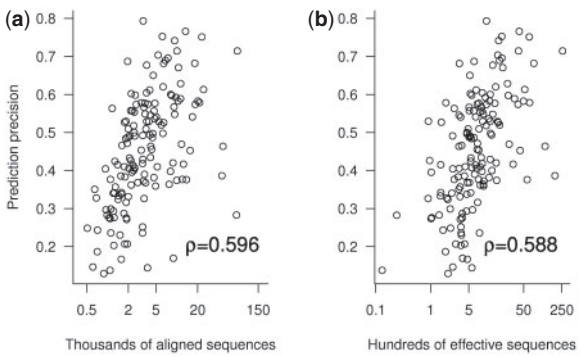

Fig. 3. These graphs show the correlation between precision (top-L predicted contacts) with the number of sequences aligned (**a**) or the effective number of aligned sequences taking weighting into account (**b**).

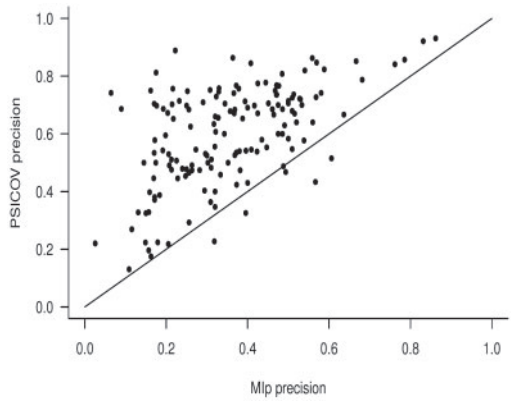

Fig. 4. PSICOV top-L/2 precision plotted against normalized sequence-weighted mutual information (MIp) precision for the 150 benchmark targets (line $x = y$ shown for reference).

method that also takes into account information such as predicted secondary structure and solvent accessibility.

Although the problem of indirect versus direct coupling in residue covariation has been known for some time, surprisingly few approaches have been proposed to tackle it. Weigt *et al.* (2009) focus on the problem of predicting protein–protein interactions, and their method is extremely computationally intensive, requiring several days of CPU time to evaluate just 60 pre-selected alignment columns in bacterial two-component systems. Recently, the same group have tried adapting the method to predicting intrachain contacts with apparently good results (C.Sander and D.Marks, personal communication). The method of Burger and van Nimwegen (2010) has actually been applied to the task of intrachain contact prediction, and here we have benchmarked it directly against PSICOV on the same set of alignments. This Bayesian network method is far less computationally demanding than the message-passing approach of Weigt *et al.* (2009), and is also substantially faster than PSICOV, taking a median of 2 mins per target compared with a median of 30 mins per target for PSICOV (range 1–240 mins). Despite the obvious speed advantage of the Bayesian network approach, it is clear from the results shown that PSICOV is the more precise method across all sequence separations and residue rankings.

As high-throughput sequencing rapidly expands the sizes of protein families, the applicability of contact prediction from large MSAs will clearly be expected to increase. Nevertheless, it is important to consider the targets which performed poorly in our benchmark and yet had large numbers of homologous sequences available (Fig. 3). These cases are unlikely to be improved by simply collecting further sequence data. Although PSICOV is able to effectively disentangle indirect coupling effects, some alignments are impossible to analyse due to functional or structural noise. A good example of a problem case would be a family of homo-multimers, where it would be impossible to determine which correlations are due to interchain contacts and which are due to intrachain contacts. Other problems would include high conservation in some families, particularly conservation around ligands or co-factors.

4 CONCLUSION

We have demonstrated that the graphical Lasso procedure, previously applied successfully to other graphical inference problems such as gene network reconstruction, performs excellently in the task of identifying directly coupled covarying columns in large MSAs, which are indicative of residue–residue contacts in protein families. For 44% of the targets, contact prediction was excellent with a precision >0.5 for the longest-range top-L/2 predicted contacts (i.e. >50% correctly predicted long-range contacts per residue). At this level of accuracy, predicted contacts should be invaluable in determining protein folds; e.g. in protein model quality assessment or decoy selection (Miller and Eisenberg, 2008). Both the mean and peak performance of PSICOV is substantially higher than that achieved by either the recently proposed Bayesian network approach of Burger and van Nimwegen (2010) or the APC corrected mutual information approach of Dunn *et al.* (2008), and it is likely that when further combined with other predicted structural information such as predicted secondary structure and solvent exposure that PSICOV will be able to reach even higher levels of accuracy.

Although PSICOV is able to deal with many of the statistical problems in contact prediction from large MSAs, there remain several practical issues. First, sequence weighting becomes more important as the sizes of families increase. It turns out that for large families, sequence weighting occupies the bulk of the computation time as it is an $O(n^2)$ procedure. More efficient sequence weighting schemes do exist [e.g. Henikoff and Henikoff (1994)], but we found these schemes performed poorly in this application, probably because they tend to overweight misaligned or incorrect sequences. Secondly, the practical difficulties of accurately and automatically aligning tens or hundreds of thousands of protein sequences cannot be underestimated. Better alignment methods for large sequence families are certain to offer improvements in contact prediction accuracy.

Although the standard graphical Lasso approach has been used in PSICOV, other sparse learning algorithms could easily be applied to the same problem. One interesting possibility, that we are currently investigating, is to make use of a group Lasso approach (Ma *et al.*, 2007) to cluster together covariances relating to different residue pairs. In the current approach, each potential contact is scored by summing 20×20 matrix values relating to the individual amino acid pairs in the two alignment columns. In a grouped Lasso approach, rather than treating the matrix values in the 20×20 submatrix as independent, they could be processed as a group of related variables. Overall, we would hope that more tailored sparse learning algorithms, better alignment algorithms, along with better sequence weighting and regularization schemes will significantly improve the results obtained from methods such as PSICOV in the future.

ACKNOWLEDGEMENTS

We are grateful to Chris Sander, Debbie Marks and Willie Taylor for useful discussions. We also thank Erik van Nimwegen and Lukas Burger for providing code to implement their Bayesian network approach.

Funding: UK Biotechnology and Biological Sciences Research Council (Grant reference: BB/F010451/1; DTJ & DWAB); UK Engineering and Physical Sciences Research Council (Grant reference: EP/H027203/1; MP); Marie Curie Intra European Fellowship within the 7th European Community Framework Programme (Grant Agreement Number: PIEF-GA-2009-237292 to DC).

Conflict of Interest: none declared.

REFERENCES

Ashkenazy,H. and Kliger,Y. (2010) Reducing phylogenetic bias in correlated mutation analysis. *Protein Eng. Des. Sel.*, **23**, 321–326.

Banerjee,O. *et al.* (2008) Model selection through sparse maximum likelihood estimation. *J. Mach. Learn. Res.*, **9**, 485–516.

Berman,H. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bühlmann,P. and van de Geer,S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg.

Burger,L. and van Nimwegen,E. (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.*, **6**, e1000633.

Buslje,C.M. *et al.* (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, **25**, 1125–1131.

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Ezkurdia,I. *et al.* (2009) Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins*, **77** (Suppl. 9), 196–209.

Fariselli,P. *et al.* (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Eng.*, **14**, 835–843.

Finn,R.D. *et al.* (2010) The pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Fischer,D. *et al.* (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins*, **45** (Suppl. 5), 171–183.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, **9**, 432–441.

Gobel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Graña,O. *et al.* (2005a) CASP6 assessment of contact prediction. *Proteins*, **61** (Suppl. 7), 214–224.

Graña,O. *et al.* (2005b) EVAcon: a protein contact prediction evaluation service. *Nucleic Acids Res.*, **33**, W347–W351.

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

Halabi,H. *et al.* (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell*, **138**, 774–786.

Hamilton,N. and Huber,T. (2008) An introduction to protein contact prediction. *Methods Mol. Biol.*, **453**, 87–104.

Hamilton,N. *et al.* (2004) Protein contact prediction using patterns of correlation. *Proteins*, **56**, 679–684.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Henikoff,S. and Henikoff,J.G. (1994) Position-based sequence weights. *J. Mol. Biol.*, **243**, 574–578.

Horner,D.S. *et al.* (2008) Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief. Bioinform.*, **9**, 46–56.

Lapedes,A.S. *et al.* (1999) Correlated mutations in protein sequences: Phylogenetic and structural effects. In *Proceedings of the AMS/SIAM Conference on Statistics in Molecular Biology and Genetics*, Monograph Series of the Institute for Mathematical Statistics, Hayward, CA, pp. 236–256.

Ledoit,O. and Wolf,M. (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance*, **10**, 603–621.

Lena,P.D. *et al.* (2011) Is there an optimal substitution matrix for contact prediction with correlated mutations? *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 1017–1028.

Ma,S. *et al.* (2007) Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, **8**, 60.

MacCallum,R.M. (2004) Striped sheets and protein contact prediction. *Bioinformatics*, **20** (Suppl. 1), i224–i231.

Magrane,M. and the UniProt Consortium (2011) UniProt knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.

Martin,L.C. *et al.* (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.

McLachlan,A.D. (1971) Tests for comparing related amino-acid sequences. cytochrome c and cytochrome c 551 . *J. Mol. Biol.*, **61**, 409–424.

Meinshausen,N. and Bühlmann,P. (2006) High dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, **34**, 1436–1462.

Miller,C.S. and Eisenberg,D. (2008) Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, **24**, 1575–1582.

Neher,E. (1994) How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA*, **91**, 98–102.

Olmea,O. and Valencia,A. (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des.*, **2**, S25–S32.

Pollastri,G. and Baldi,P. (2002) Prediction of contact maps by giohmms and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, **18** (Suppl. 1), S62–S70.

Pollock,D.D. and Taylor,W.R. (1997) Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng.*, **10**, 647–657.

Punta,M. and Rost,B. (2005) PROFcon: novel prediction of long-range contacts. *Bioinformatics*, **21**, 2960–2968.

Shao,Y. and Bystroff,C. (2003) Predicting interresidue contacts using templates and pathways. *Proteins*, **53** (Suppl. 6), 497–502.

Weigt,M. *et al.* (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA*, **106**, 67–72.

Xue,B. *et al.* (2009) Predicting residue-residue contact maps by a two-layer, integrated neural-network method. *Proteins*, **76**, 176–183.

Yuan,Z. (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics*, **6**, 248.

Yuan,M. and Lin,Y.L. (2007) Model selection and estimation in the gaussian graphical model. *Biometrika*, **91**, 19–35.