BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 5 2013, pages 580–587
doi:10.1093/bioinformatics/btt005

Structural bioinformatics

Advance Access publication January 10, 2013

Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure

Nikolas S. Burkoff, Csilla Várnai and David L. Wild*

Systems Biology Centre, Senate House, University of Warwick, Coventry, CV4 7AL, UK

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The problem of *ab initio* protein folding is one of the most difficult in modern computational biology. The prediction of residue contacts within a protein provides a more tractable immediate step. Recently introduced maximum entropy-based correlated mutation measures (CMMs), such as direct information, have been successful in predicting residue contacts. However, most correlated mutation studies focus on proteins that have large good-quality multiple sequence alignments (MSA) because the power of correlated mutation analysis falls as the size of the MSA decreases. However, even with small autogenerated MSAs, maximum entropy-based CMMs contain information. To make use of this information, in this article, we focus not on general residue contacts but contacts between residues in β-sheets. The strong constraints and prior knowledge associated with β-contacts are ideally suited for prediction using a method that incorporates an often noisy CMM.

Results: Using contrastive divergence, a statistical machine learning technique, we have calculated a maximum entropy-based CMM. We have integrated this measure with a new probabilistic model for β-contact prediction, which is used to predict both residue- and strand-level contacts. Using our model on a standard non-redundant dataset, we significantly outperform a 2D recurrent neural network architecture, achieving a 5% improvement in true positives at the 5% false-positive rate at the residue level. At the strand level, our approach is competitive with the state-of-the-art single methods achieving precision of 61.0% and recall of 55.4%, while not requiring residue solvent accessibility as an input.

Availability: http://www2.warwick.ac.uk/fac/sci/systemsbiology/research/software/

Contact: D.L.Wild@warwick.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on September 27, 2012; revised on December 17, 2012; accepted on January 2, 2013

1 INTRODUCTION

The problem of *ab initio* protein folding is one of the most difficult in modern computational biology. The prediction of residue contacts within a protein provides a more tractable immediate step, and these contacts can be used as a guide to generate the tertiary structure of the protein.

Correlated mutation (CM) methods, first pioneered by Valencia and colleagues (Gobel *et al*., 1994), take a multiple sequence alignment (MSA) profile of evolutionarily related proteins and attempt to predict residues that have co-evolved. If residues have co-evolved, this may imply proximity in the native structure. For example, if a small residue increases in size by mutating, a proximal residue may have to reduce in size to retain the viability of the fold.

Many CM methods have been developed using Pearson correlation coefficients (Gobel *et al*., 1994), adaptions of Mutual Information (Dunn *et al*., 2008; Lee and Kim, 2009), perturbation methods (Dekker *et al*., 2004) and Dynamic Bayesian networks (Burger and van Nimwegen, 2010).

A recently developed correlated mutation measure (CMM), the *direct information* (Morcos *et al*., 2011; Weigt *et al*., 2008), is a global measure that is derived from modelling the entire MSA, specifically defining the probability of each sequence being a member of the MSA. This distribution shares the same low-order moments as the MSA, and the maximum entropy principle (Jaynes, 2007) is used to fully specify the distribution. Marks *et al*. (2011), Sułkowska *et al*. (2012) and Hopf *et al*. (2012) have used this measure to successfully aid the folding of a diverse range of proteins. However, like the majority of CM studies, these authors focused on a small number of proteins for which there is a large high-quality MSA because all CMMs suffer as the size of the MSA decreases (Olmea and Valencia, 1997). A key distinction of this work is that we focus on a wide selection of proteins that have a variety of sizes of MSAs. We also automate the generation of MSAs and do not rely on a large high-quality MSA being available.

In an attempt to improve the power of CM methods, the Dynamic Bayesian network of Burger and van Nimwegen (2010) incorporates primary-sequence distance into an informative prior for the model. The incorporation of this knowledge substantially improves the results. Inspired by this, we have chosen to predict the lateral pairs of residues in interacting β-strands, β-contacts, using a CMM. β-contacts are associated with strong constraints, for example, sequential pairs of residues form β-contacts and residues can only be in β-contact with up to two other residues. These constraints mean β-contacts are ideally suited for prediction using a CMM—the noise associated with the CMM is compensated for by incorporating the strong β-contact constraints.

The prediction of β-contacts can be used to aid tertiary structure prediction (Podtelezhnikov and Wild, 2009; Ruczinski *et al*., 2002), explore energy landscapes (Burkoff *et al*., 2012), in designing proteins (Kortemme *et al*., 1998; Smith and Regan, 1995) and understanding protein folding pathways (Mandel-Gutfreund *et al*., 2001; Merkel and Regan, 2000).

*To whom correspondence should be addressed.

We highlight BetaPro, the work of Cheng and Baldi (2005), which uses a three-stage method to predict β-topologies and was the first method to take into account the global nature of β-topologies. Firstly, a 2D recurrent neural network is used to generate a residue-level pairing map. Secondly, a dynamic programming algorithm is applied to this map to derive strand-level pseudo binding energies and finally, a graph matching algorithm is used to predict strand contacts.

There are a variety of other existing methods for β-contact prediction. They include the use of statistical potentials (Hubbard and Park, 1994), information theoretic approaches (Steward and Thornton, 2002), integer linear optimization (Rajgaria *et al.*, 2010), hybrid neural network-probabilistic models (Aydin *et al.*, 2011) and Markov logic networks (MLNs; Lippi and Frasconi, 2009).

In this article, we have developed a global probabilistic model for β-contact prediction, inspired by the secondary structure models of Schmidler (2002), which can be used to predict both residue- and strand-level interactions. We have integrated this model with a CMM, similar in nature to direct information, and using this model on a standard dataset, significantly outperform the recurrent neural network of BetaPro and are competitive with the best single methods currently available. Unlike these methods, our approach does not require additional information such as residue solvent accessibility to be entered as an input to the model. In common with other methods, we assume the native secondary structure is known. However, our framework can be easily extended to predict both secondary structure and β-contacts simultaneously, and this is the focus of our current work.

2 METHODS

2.1 Data set

In this work, we use the set of 916 proteins from Cheng and Baldi (2005) (CB916). The proteins share no >15–20% sequence identity, and the set consists of 187 516 residues, of which 48 996 are strand residues, which are involved in 31 638 β-contacts.

Most CM analysis procedures focus primarily on proteins for which there is a large good-quality MSA, often a large PFAM alignment (Sonnhammer *et al.*, 1997). We wanted to develop a method that will take advantage of this information where it exists, and yet is applicable even if the CM analysis is not useful, or indeed there is no MSA, which can be the case for newly sequenced proteins, such as those selected as targets in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) community-wide experiment. Therefore, our method of generating MSAs is extremely general.

We generate MSAs following a similar method to Saqi *et al.* (1999). For each sequence, we run PSI-Blast (Altschul *et al.*, 1997) for two iterations (Evalue = 0.005) against the Non-Redundant database, keeping all sequences that share at least 30% identity to the profile constructed after the first PSI-Blast iteration, similar to the procedure recommended in Ashkenazy *et al.* (2009). We then perform a global–local alignment using GLsearch (Pearson, 2000) to trim the sequences PSI-Blast found. We then use CD-Hit (Li *et al.*, 2001; Li and Godzik, 2006) to cluster the trimmed sequences at the 98% threshold and use Muscle (Edgar, 2004) (maxiters = 2) to generate MSAs. Finally, we removed columns of the MSA that were gaps in our target sequence and any row that contained >33% gaps. There is an enormous variation in the number of sequences in the alignments: six proteins have no homologues, one-fifth have <100 homologues and 7% have >2000 (see Fig. 1).

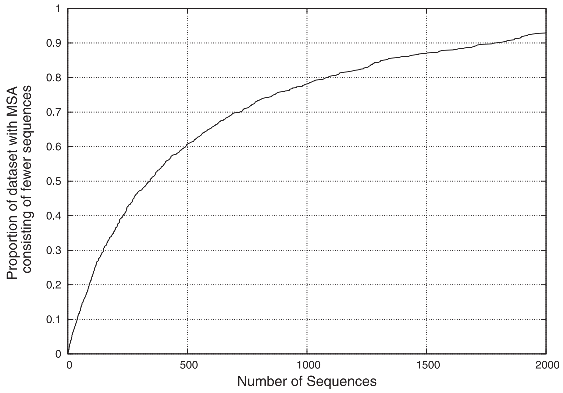

Fig. 1. The number of sequences in the 916 MSAs varies enormously. For example, 60% of the MSA have <500 sequences

2.2 Maximum entropy-based CM measure

CMMs based on maximum entropy modelling (also called Direct Coupling Analysis) (Marks *et al.*, 2011; Morcos *et al.*, 2011; Weigt *et al.*, 2008) aim to distinguish between *direct* and *indirect* correlations. Direct correlations arise owing to proximity in the native structure of the protein and are of primary interest in contact prediction; indirect correlations are caused by other reasons, such as the fact that correlations are transitive, and are the cause of the poor performance of many CMMs.

The idea is to model the entire family of evolutionarily related proteins, assigning probability mass over all possible (fixed-length) sequences, including those that have not been observed. From this global model, measures can be developed to model the strength of the direct correlations between pairs of residues. This idea is formalized below.

Given an MSA containing M sequences for a protein of length N, we define $f_i(A_i)$ as the observed frequency of residue A_i occurring in position i of the MSA and $f_{ij}(A_i, A_j)$ as the observed frequency of both residue A_i occurring in position i and residue A_j occurring in position j of the MSA. Given any sequence $\mathbf{A} = A_1, A_2, \ldots, A_N$, we model the probability of it occurring in the MSA by a distribution $P(\mathbf{A}) = P(A_1, A_2, \ldots, A_N)$ However, there are q^N possible different sequences (where q is the size of the alphabet of amino acids) and only $M \ll q^N$ sequences in the MSA. The sparsity of the data and the number of sequences imply that it is impractical for detailed use. However, we would like our model to match the empirical low-order moments given by the MSA. Specifically we would like

$$P_i(A_i) = f_i(A_i) \quad \text{and} \quad P_{ij}(A_i, A_j) = f_{ij}(A_i, A_j)$$

where $P_i(.)$ is the marginal distribution for position i and $P_{ij}(.,.)$ is the (joint) marginal distribution (We have not added pseudo-counts or weighted sequences) marginal distribution of positions i and j (we have not added pseudo-counts or weighted sequences).

Among the valid distributions P satisfying these constraints, using the maximum entropy principle (Jaynes, 2007), we favour $\mathfrak{P}$, the distribution that has maximum entropy, S:

$$\mathfrak{P} = \text{argmax}_P[S(P)] \equiv \text{argmax}_P\left\{-\sum_{\mathbf{A}} P(\mathbf{A}) \log[P(\mathbf{A})]\right\}$$

and solving this optimization problem using Lagrange multipliers leads to the distribution

$$\mathfrak{P}(A_1, \ldots, A_N) \propto \exp\left[-\sum_{1 \le i \le j \le N} e_{ij}(A_i, A_j) + \sum_{1 \le i \le N} h_i(A_i)\right]$$

for some pair-interaction energies $e_{ij}(A_i, A_j)$ and local fields $h_i(A_i)$ (Weigt *et al.*, 2008). See the Supplementary Data for further details.

The maximum entropy distribution can be viewed as a Potts model on an underlying complete graph, where the nodes represent the residue positions, the ‘spins’ correspond to the amino acid types and the edges describe the pairwise interactions, whose strengths are described by the pairwise interaction energies e_{ij}. A related model for protein families, using Markov random fields (Balakrishnan *et al.*, 2011), can also be viewed as a Potts model. However, instead of the underlying graph being complete, an optimal subgraph is chosen that aims to fully explain the correlations and conditional independencies within the underlying protein family.

To generate the maximum entropy distribution $\mathfrak{P}$, we use a statistical machine learning technique, contrastive divergence (Hinton, 2002). This work represents the first application of this approach to the modelling of protein MSAs. For a given set of $e_{ij}(A_i, A_j)$ and $h_i(A_i)$, we use contrastive divergence to approximate the marginal distributions $P_i(.)$ and $P_{ij}(.,.)$ and use gradient descent to update e_{ij} and h_i. We iterate this procedure to convergence. For a protein of 75 residues, the procedure takes ~10 minutes on a single core of an Intel Core i7 processor, and for a protein of 350 residues, the procedure takes ~2.5 h. Further details are found in the Supplementary Data.

Once we have calculated the distribution $\mathfrak{P}$, we define our CMM, $\mathcal{D}$. For each pair of residues (*i, j*), we define $\mathcal{D}(i,j)$ as follows:

$$\mathcal{D}(i,j) = \sum_{A_i, A_j} \mathfrak{P}^D_{ij}(A_i, A_j) \log \frac{\mathfrak{P}^D_{ij}(A_i, A_j)}{f_i(A_i) f_j(A_j)}$$

where

$$\mathfrak{P}^D_{ij}(A_i, A_j) \propto f_i(A_i) f_j(A_j) \exp[-e_{ij}(A_i, A_j)].$$

This is a modified version of the *Direct Information* previously used to predict protein contacts (Marks *et al.*, 2011; Weigt *et al.*, 2008). The Direct Information measure itself was tried but produced slightly poorer results than $\mathcal{D}$. See the Supplementary Data for more details.

To show the power of $\mathcal{D}$, for each protein in the dataset, we took the top $N/2$ ranked $\mathcal{D}(i,j)$, where N is the length of the protein (we remove those for which $|i-j| \leq 4$ from the analysis) and calculated the contact ratio: the proportion of these pairs of residues whose C_α distance is ≤ 8 Å. The contact ratio versus log(M) is shown in Figure 2 (Top). Figure 2 (Bottom) shows the average C_α distance of these $N/2$ predicted contacts. These figures show that there is a lot of information contained within $\mathcal{D}$, especially as M increases.

However, using randomly chosen contacts of known structures, it has been shown that one needs around a quarter to two-fifths of contacts to be able to successfully regenerate the native structure (Duarte *et al.*, 2010; Sathyapriya *et al.*, 2009; Vendruscolo *et al.*, 1997). Marks *et al.* (2011) and Hopf *et al.* (2012) have shown that if a protein has a large number of sequences in its MSA, then maximum entropy-based CM analysis, together with predicted secondary structure is enough to successfully reconstruct the tertiary structure of the protein. In these articles, the authors take the highest-ranked correlated pairs of residues to be incorporated into distance constraints used to generate initial all-atom conformations of the protein. Simulated annealing, relaxing these distance constraints throughout the simulation, is then used to generate final three-dimensional structures.

However, as shown by Figures 1 and 2, a large number of proteins have only a small MSA and CMMs by themselves are unlikely to be able to provide a large enough number of contacts to successfully fold the protein. For example, Marks *et al.* (2011) restrict their attention to proteins whose MSA has at least 1000 sequences, and usually significantly more. Nevertheless, even an alignment with $M = e^6 \approx 400$ sequences produces an average contact ratio of ~0.15, which still contains lots of information (for an average protein, the contact ratio for randomly chosen contacts is ~0.03). In contrast to these other studies, we investigate whether one can make use of this evolutionary information. We propose to use $\mathcal{D}$ to improve the prediction of β-contacts, for which there is a large amount of structural knowledge, which can be incorporated as prior beliefs within a Bayesian statistical framework. The following sections describe the new β-strand Bayesian model we have developed and how we couple $\mathcal{D}(i,j)$ to it.

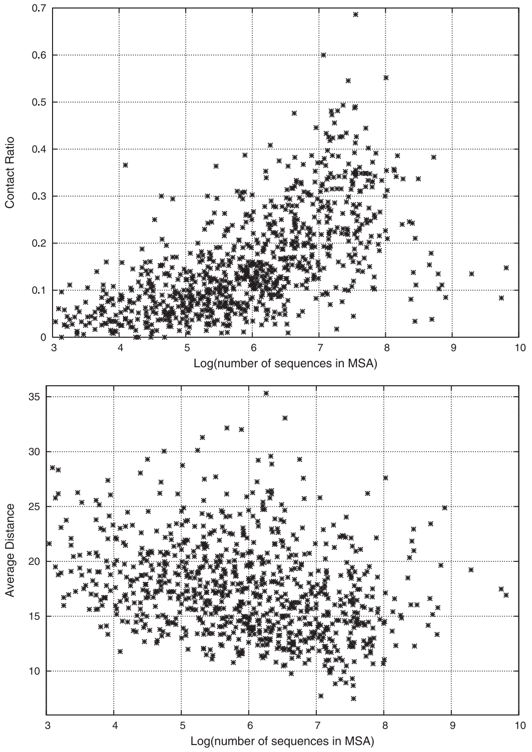

Fig. 2. Top: The proportion of the top $N/2$ ranked $\mathcal{D}(i,j)$ in contact (contact ratio) versus the number of sequences in the MSA for each protein in the CB916 dataset. A contact is defined as the C_α distance being $\leq$ 8 Å. Bottom: The average C_α distance of the top $N/2$ ranked $\mathcal{D}(i,j)$. Two outliers [at (3.4, 37.9) and (4.5, 42.1)] are not shown

2.3 β-Topology model

Given a primary sequence $\mathbf{R} = \{R_1, R_2, \ldots, R_N\}$ and its secondary structure $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$, where S_i is the secondary structure of residue *i*, residues R_i and R_j are defined to be a β-contact if they are a lateral pair within two interacting β-strands. For example, in Figure 3, residues 6 and 53 are a parallel β-contact and residues 44 and 53 are an antiparallel β-contact. We define $\mathcal{I}$ to be the set of β-contacts. Specifically $(i,j,1) \in \mathcal{I}$ if residues R_i and R_j are a parallel β-contact, $(i,j,-1) \in \mathcal{I}$ if residues R_i and R_j are an antiparallel β-contact and $(i,j,0) \in \mathcal{I}$ if either residue R_i or R_j is an isolated β-bridge. We say $(i,j) \in \mathcal{I}$ if $(i,j,1)$, $(i,j,0)$ or $(i,j,-1) \in \mathcal{I}$.

The general framework we are using (from Schmidler, 2002) allows inference for $\mathbf{S}$ and $\mathcal{I}$ given $\mathbf{R}$. Following the Bayesian method, we require a prior $\mathbb{P}(\mathbf{S}, \mathcal{I}) = \mathbb{P}(\mathcal{I}|\mathbf{S})\mathbb{P}(\mathbf{S})$ and a likelihood $\mathbb{P}(\mathbf{R}|\mathbf{S}, \mathcal{I})$. Using Bayes’ theorem, these yield the posterior of interest $\mathbb{P}(\mathbf{S}, \mathcal{I}|\mathbf{R}) \propto \mathbb{P}(\mathbf{R}|\mathbf{S}, \mathcal{I})\mathbb{P}(\mathbf{S}, \mathcal{I})$.

In this work, we assume the secondary structure is fixed. Specifically $\mathbb{P}(\mathbf{S}) = 1$, if $\mathbf{S}$ is the secondary structure assignment given by DSSP (Kabsch and Sander, 1983)—we map residues labelled E and B to E, strand residues, and all other labels to C, non-strand residues. For clarity we suppress the dependency on $\mathbf{S}$, i.e. $\mathbb{P}(\mathcal{I}|\mathbf{S}) = \mathbb{P}(\mathcal{I})$. A focus of our current work is to extend the model to allow joint inference for $\mathbf{S}$ and $\mathcal{I}$.

Definitions: Viewing $\mathcal{I}$ as a collection of individual residue contacts does not easily allow the incorporation of the structure of β-contacts into a model; therefore, we model $\mathcal{I}$ as a set of interacting strand segments, following (Chu *et al.*, 2006). The set of residue contacts in $\mathcal{I}$ can be uniquely determined by specifying which strand segments interact and for each pair of interacting strands specifying their direction, alignment and position of any bulges. We formalize these terms below.

The strand residues of a protein can be represented as a set of distinct strand segments (For some proteins, DSSP defines two separate strand segments immediately adjacent in sequence. For example 'EEEB'. For a fair comparison with BetaPro we define a strand segment as a contiguous block of strand residues. However, this is not necessary for our model). For example, Figure 3 shows 4 strand segments (E_1,E_2,E_3,E_4). In this protein, there is a single sheet, and in this simple case, the strand interactions can be described by a permutation ϕ of the set of strand segments. Specifically $\phi(1, 2, \ldots, m) = [\phi(1), \phi(2), \ldots, \phi(m)]$ and implies segment $E_{\phi(r)}$ and $E_{\phi(r+1)}$ interact for $r = 1, 2, \ldots, m-1$. In Figure 3, $\phi(1, 2, 3, 4) = (3, 4, 1, 2)$. In more complicated cases, the sheet structure cannot be described by a permutation. For example, if there is more than one sheet, if strands are involved in more than two interactions or if there is a cycle (for example in β-barrels, where every strand interacts with two partners).

Following the terminology in (Ruczinski *et al.*, 2002), we say there is a jump between segments E_r and E_{r+1} if E_r and E_{r+1} are not interacting. In Figure 3, there is a jump between segments E_2 and E_3 and no other jumps. We define the jump pattern J as the set of r for which E_r and E_{r+1} are not interacting; in Figure 3, the jump pattern $J = \{2\}$. See Figure 4(a–d) for further examples of ϕ and J.

We introduce d_{rs} to describe the direction of interaction, specifically $d_{rs} = 1$ if interacting segments E_r and E_s are a parallel strand interaction and $d_{rs} = -1$ if the segments are antiparallel. In Figure 3, $d_{34} = d_{12} = -1$ and $d_{14} = 1$. If either E_r or E_s is an isolated β-bridge, then $d_{rs} = 0$.

The variable a_{rs} is used to define the shift between strands. For parallel interactions, a_{rs} describes the shift between the final residues of both strands. For example, in Figure 3, $a_{14} = 0$ because $(8, 55) \in \mathcal{I}$. If E_1 was shifted up by one residue, so that $(8, 54) \in \mathcal{I}$, then a_{14} would equal $+1$. Conversely, if E_1 was shifted down by two residues, so that residue $(6, 55) \in \mathcal{I}$, then a_{14} would equal -2. For antiparallel interactions, a_{rs} describes the shift between the end of the strand earlier in the sequence and the beginning of its interacting partner (i.e. between residues 8 and 13 for a_{12} in Fig. 3).

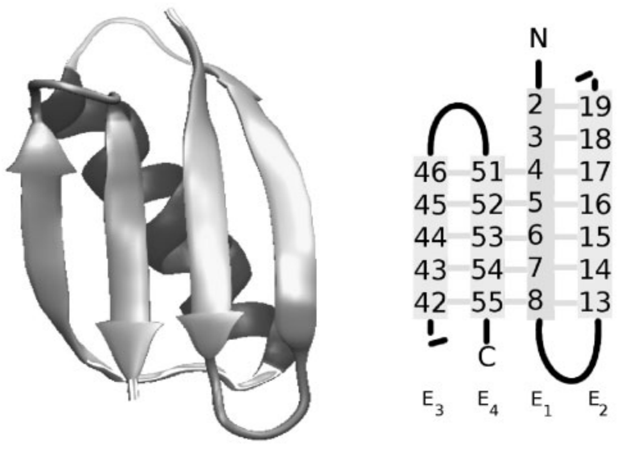

Fig. 3. Left: The structure of protein G (1PGA). Right: The β-topology for protein G. The numbers are the positions in the sequence of the strand residues and the horizontal lines are β-contacts. Residues 6 and 53 are a parallel β-contact and residues 53 and 44 are an antiparallel β-contact. Hence $(6, 53, 1)$ and $(44, 53, -1) \in \mathcal{I}$. The 4-strand segments E_1,E_2,E_3,E_4 are ordered from left to right in the sheet E_3,E_4,E_1,E_2, and hence the permutation ϕ, which permutes (1, 2, 3, 4) to (3, 4, 1, 2) describes the set of interactions. There is a jump between segments E_2 and E_3. Segments E_1 and E_2 are an antiparallel interaction and hence $d_{12} = -1$, and segments E_1 and E_4 are parallel and so $d_{14} = 1$

Restricting the number of bulges to at most one per β-strand interaction (which is the case in 98.6% of cases), we can define $b_{rs} = 0$ if there is no bulge or $b_{rs} = k$ if residue k is the β-bulge. There are no bulges in the sheet shown in Figure 3. Figure 4 shows the values of $\{d_{rs}, a_{rs}, b_{rs}\}$ for different interacting segments.

Prior for $\mathcal{I}$, $\mathbb{P}(\mathcal{I})$: There is a huge amount of structure in β-topologies and the challenge for a Bayesian statistician is to try and capture this while being able to efficiently calculate posterior probabilities and not overfitting the model. Rather than aim for the most probable β-topology, we calculate $\mathbb{P}((i,j) \in \mathcal{I}|\mathbf{R})$, producing a *probability contact map*, analogous to the output from BetaPro's Neural Network. Unlike other statistical models (Aydin *et al.*, 2011), we do not take the output from BetaPro's Neural Network as an input to our model.

We take advantage of the framework of Bayesian inference, which allows us to exercise our scientific judgement and experience concerning parameters that we expect to be of particular importance, and by specifying how these are plausibly related.

We model the interacting β-strands as a single sheet defined by a permutation ϕ, as described above. Although our approach does not model more than one sheet per protein, we can predict multiple sheets (see Figure 8). More complicated models involving partitioning the segments into different sheets were tried, but these did not improve the results. We only allow a single bulge per strand interaction.

Our prior is defined as

$$\mathbb{P}(\mathcal{I}) = \mathbb{P}(\phi) \prod_{r,s} \mathbb{P}(d_{rs})\mathbb{P}(a_{rs}|d_{rs})\mathbb{P}(b_{rs}|a_{rs}, d_{rs})$$

where the product is over all segments E_r and E_s that are interacting, given permutation ϕ, and we have suppressed the dependence of everything on the secondary structure $\mathbf{S}$. The set $\{\phi, d_{rs}, a_{rs}, b_{rs}\}$ gives a unique set of residue contacts $(i,j) \in \mathcal{I}$, and if $\mathcal{I}$ cannot be described by a set $\{\phi, d_{rs}, a_{rs}, b_{rs}\}$, then $\mathbb{P}(\mathcal{I}) = 0$. We define the distance τ_{rs} as the number of residues between segments E_r and E_s. For example, in Figure 3, $\tau_{12} = 4$, and we define l_r as the number of residues in segment E_r.

- $\mathbb{P}(\phi)$: The probability of a specific permutation depends on all the distances τ_{rs} and the lengths of all the strands l_r. However, incorporating all this information leads to an exponential number of parameters. In the dataset, 50% of interacting strands are adjacent in sequence (and 42% of adjacent strands are interacting), so one of the most important things we would like the distribution to capture is whether adjacent strands are in contact. For these reasons, in our

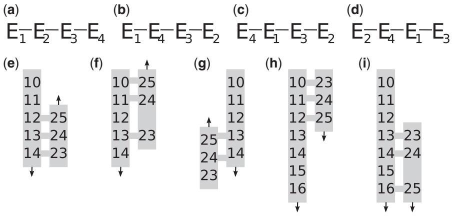

Fig. 4. **(a–d)** Examples of different values of ϕ: (a) $\phi = (1, 2, 3, 4), J = \{\}$; (b) $\phi = (1, 4, 3, 2), J = \{1\}$; (c) $\phi = (4, 1, 3, 2), J = \{1, 3\}$; (d) $\phi = (2, 4, 1, 3), J = \{1, 2, 3\}$. **(e–i)** Examples of different $\{d_{rs}, a_{rs}, b_{rs}\}$: (e) $\{-1, 0, 0\}$; (f) $\{-1, -1, 12\}$; (g) $\{-1, 1, 0\}$; (h) $\{1, -4, 0\}$; (i) $\{1, 0, 15\}$

model, all ϕ that share the same jump pattern J are equally likely, and the probability $\mathbb{P}(r \in J|\tau_{rr+1}, l_r, l_{r+1})$ is independent for each r. $\mathbb{P}(r \in J|\tau_{rr+1}, l_r, l_{r+1})$ is taken from the training set by counting occurrences. For small l and τ, we take values directly from the training set, and for larger l and τ, owing to sparsity of data, we collapse the data into a small number of bins. We added a pseudo-count to smooth the data from the training dataset.

- If either l_r or $l_s = 1$, then $\mathbb{P}(d_{rs} = 0) = 1$, otherwise $\mathbb{P}(d_{rs} = 1|\tau_{rs})$ is a piecewise linear function of τ_{rs}, fitted from the training set.
- $\mathbb{P}(a_{rs}|d_{rs})$: There is an inherent asymmetry in our definition of a_{rs}; in the case of parallel strands, we are measuring the shift from a perfect alignment of the ends of the segments, not the beginnings. In proteins, it is found that the shift measured from a perfect alignment of *at least one* end of the segments is small. Compare a_{rs} for Figures 4(h–i). Previous work has not taken this into account (Chu *et al.*, 2006), which leads to a drop in performance. Therefore, we model $\mathbb{P}(a_{rs}|d_{rs})$ as a mixture (equally weighted) of the distributions $\mathbb{P}_{d_{rs}}(a_{rs})$ and $\mathbb{P}_{d_{rs}}(\hat{a}_{rs})$, where $\hat{a}_{rs}$ is the shift required from aligning the beginnings of the segments to get the same residue contacts as a shift of a_{rs} produces from aligning the ends of the strands. These distributions are taken from the training set. An analogous procedure is followed for the antiparallel case.
- $\mathbb{P}(b_{rs} \neq 0|d_{rs}, a_{rs}) = \mathbb{P}(b_{rs} \neq 0)$ is taken from the training set, and if there is a bulge, there is a uniform probability over all residues involved in the interaction that they are a bulge (hence the dependence on d_{rs} and a_{rs}—to know which residues can be the bulge).

Likelihood: $\mathbb{P}(\mathbf{R}|\mathcal{I})$

$$\mathbb{P}(\mathbf{R}|\mathcal{I}) \propto |\mathcal{I}|^{(u|E|-1)} \exp(-v|\mathcal{I}|) \prod_{(i,j,d_{ij})\in\mathcal{I}} \mathbb{L}(R_i, R_j|d_{ij})$$

where the joint likelihood $\mathbb{L}(R_i, R_j|d_{ij})$ is approximated from the limited training set by the product of the conditionals, $\mathbb{P}(R_i|R_j, d_{ij})$ and $\mathbb{P}(R_j|R_i, d_{ij})$, where $\mathbb{P}(.|R_j, d_{ij})$ is the distribution of amino acids in contact with the residue type of R_j in the direction of d_{ij}. $|\mathcal{I}|$ is the number of contacts and $|E|$ is the number of β-residues. The distributions $\mathbb{P}(.|R_j, d_{ij})$ are taken from the training set, and u and v are constants to be determined.

We have chosen this likelihood because of its simplicity. More complicated dependencies, such as letting R_i depend on $R_{j\pm1}$, were tried, but did not noticeably improve the results. We include a gamma distribution on the number of contacts into the likelihood because, without this term, the likelihood is a product of $2|\mathcal{I}|$ numbers smaller than one, and so actively penalises against contacts. We include $|E|$ so that the mean and variance of the gamma distribution depend on the number of β-residues, which allows the model to control the total number of contacts. This is important as $|\mathcal{I}|$ and $|E|$ are strongly correlated. The constants u and v were fitted using an empirical Bayes approach, and set to 18 and 12, respectively. See Supplementary Data for more details.

2.4 Integrating CM measure with the β-topology model

In this work, we perform inference on both the posterior distribution $\mathcal{P}_1(\mathcal{I}|\mathbf{R}) \propto \mathbb{P}(\mathcal{I})\mathbb{P}(\mathbf{R}|\mathcal{I})$ and, by adapting the concept of a 'product of experts' (Hinton, 1999, 2002), on a distribution that couples $\mathcal{D}(i,j)$ to the β-topology model. A product of experts allows different probabilistic models of the same data to be combined together by multiplying the probabilities together and renormalizing. An advantage of this method is that each model ('expert') can focus on different aspects of the underlying problem, and that regions of space with high probability mass must satisfy each of the experts, owing to the multiplication of their probabilities.

A product of experts has been successfully used for secondary structure prediction (Chu *et al.*, 2006), where there were separate experts for segmental dependency and strand and helical capping signals. In the present case, we have a distribution for inference of $\mathcal{I}$ given strand pattern $\mathcal{P}_1$, and a distribution for inference of $\mathcal{I}$ given $\mathcal{D}$, a distribution proportional to $\exp[\omega(\mathcal{D},\mathcal{I})]$, described below. Adapting the idea of a product of experts distribution, we use a product of distributions $\mathcal{P}_2(\mathcal{I}|\mathbf{R}) \propto \mathcal{P}_1(\mathcal{I}|\mathbf{R}) \exp[\omega(\mathcal{D},\mathcal{I})]$. When $\mathcal{P}_2(\mathcal{I})$ is large, $\mathcal{I}$ must satisfy both the strand pattern model of $\mathcal{P}_1$ and the CMM $\exp[\omega(\mathcal{D},\mathcal{I})]$. (Formally, $\mathcal{P}_2(\mathcal{I}|\mathbf{R}) = \mathcal{P}(\mathcal{I}|\mathbf{R},\mathcal{D}) \propto \mathbb{P}(\mathcal{I}|\mathbf{R})\mathbb{P}(\mathcal{I}|\mathcal{D})$ and $\mathbb{P}(\mathcal{I}|\mathcal{D}) = \exp[\omega(\mathcal{D},\mathcal{I})]/\sum_i \exp\ [\omega(\mathcal{D},\mathcal{I}_i)]$ where the sum is over the (finite) set of possible $\mathcal{I}_i$.)

Correlated mutation measure, $\exp[\omega(\mathcal{D},\mathcal{I})]$*:* As previously described, $\mathcal{D}(i,j)$ is a measure of how strongly residues in columns i and j co-vary, and a large $\mathcal{D}(i,j)$ suggests residues in columns i and j have co-evolved, and may imply a β-contact between R_i and R_j. This information can be incorporated into the inference as a CMM $\exp[\omega(\mathcal{D},\mathcal{I})]$. The better $\mathcal{I}$ and $\mathcal{D}(i,j)$ fit the larger the value of ω. The formal description of ω follows. We define

$$\emptyset_i = (j : R_j \text{ is a residue in a different strand to residue } R_i).$$

As a concrete example, Figure 5a shows a protein with three strands, residues 3–5, 12–14 and 23–25, where, for example, $\emptyset_4 = \{12, 13, 14, 23, 24, 25\}$ and $\emptyset_{23} = \{3, 4, 5, 12, 13, 14\}$.

In β-sheets, the side chains of residues j and $j \pm 2$ are near each other in space, and so if $\mathcal{D}(i, j \pm 2)$ are large, this may also imply a contact between R_i and R_j. For a particular set of contacts $\mathcal{I}$ and residue R_i, we define the score $\chi(i,\mathcal{I})$ as the mean of the set $\{\mathcal{D}(i,j) : j \in \mathcal{I}_i\}$ where

$$\mathcal{I}_i = \begin{cases} \emptyset_i & \text{if} \nexists j : (i,j) \in \mathcal{I} \\ \bigcup_{j:(i,j)\in\mathcal{I}} \{j-2, j, j+2\} & \text{otherwise} \end{cases}$$

As a concrete example, Figure 5b shows a specific instance of $\mathcal{I}$, and in this case $\mathcal{I}_5 = \emptyset_5, \mathcal{I}_4 = \{12, 14, 16\}$ and $\mathcal{I}_{13} = \{1, 3, 5, 22, 24, 26\}$. The larger $\chi(i,\mathcal{I})$ the better $\mathcal{D}$ and $\mathcal{I}$ fit for residue R_i. However, for different residues R_i, the mean and variance of the set of values $\{\mathcal{D}(i,j) : j \in \emptyset_i\}$ differ wildly and so $\chi(i,\mathcal{I})$ needs to be standardized before being used. For this standardization we take the sample mean μ_i and standard deviation σ_i of the set $\{\mathcal{D}(i,j) : j \in \emptyset_i\}$. So the standardized score, for residue R_i and interaction set $\mathcal{I}$ is then defined as

$$Z(i,\mathcal{I}) = \frac{\chi(i,\mathcal{I}) - \mu_i}{\sigma_i}$$

Defining $\mathcal{I}_{\text{native}}$ as the crystal structure β-contacts defined by DSSP, Figure 6 shows the empirical distribution of $Z(i,\mathcal{I}_{\text{native}})$ over all residues involved in at least one β-contact from the dataset. A much larger mass has positive score than a negative score, implying native contacts have, on average, a larger value for Z.

We then define

$$\omega(\mathcal{D},\mathcal{I}) = \log M \sum_i Z(i,\mathcal{I})$$

where the sum is over all i for which R_i are strands and M = number of sequences in MSA; so that proteins with larger MSA attach more importance to ω.

3 RESULTS AND DISCUSSION

We performed 10-fold cross validation using the same folds as Cheng and Baldi (2005). To estimate posterior probabilities, we used importance sampling. We generated 1 million independent

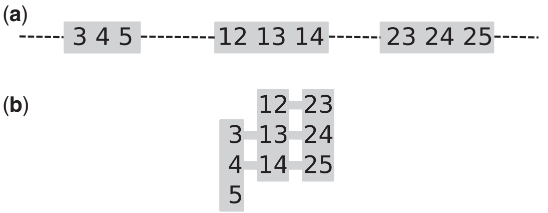

Fig. 5. (**a**) A protein with three strands, residues 3–5, 12–14 and 23–25. (**b**) A specific set of contacts $\mathcal{I} = \{(3, 13), (4, 14), (12, 23), (13, 24), (14, 25)\}$. See the text for how the standardized score is calculated for this example

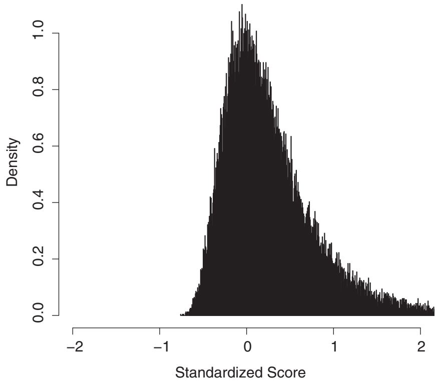

Fig. 6. The empirical distribution of the standardized score, $Z(i, \mathcal{I}_{\text{native}})$ (described in the main text), of all residues involved in at least one β-contact from the dataset

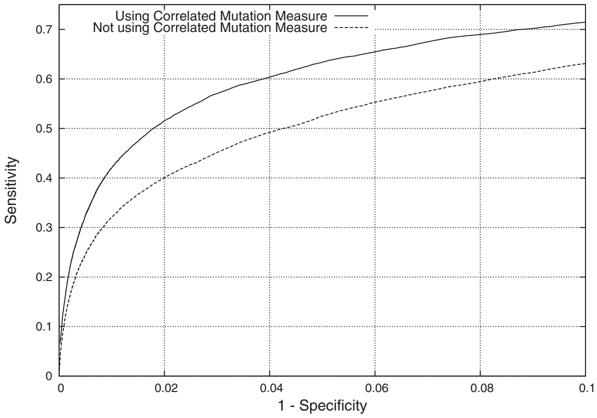

Fig. 7. ROC curves for the posterior both unweighted (dashed) and weighted (solid)

samples from the prior $\mathbb{P}(\mathcal{I}), \{\mathcal{I}\}$ and use these to generate a probability contact map:

$$\mathbb{P}((i,j) \in \mathcal{I}|\mathbf{R}) \approx \sum_{\{\mathcal{I}\}} \mathbb{I}[(i,j) \in \mathcal{I}] \frac{\mathbb{P}(\mathbf{R}|\mathcal{I})}{\sum_{\{\mathcal{I}\}} \mathbb{P}(\mathbf{R}|\mathcal{I})}$$

where $\mathbb{I}$ is the indicator function and $\mathbb{P}(\mathbf{R}|\mathcal{I})$ is the likelihood described above (in the case with the CMM we replace $\mathbb{P}(\mathbf{R}|\mathcal{I})$ by $\mathbb{P}(\mathbf{R}|\mathcal{I}) \exp[\omega(\mathcal{D}, \mathcal{I})]$. See Supplementary Data for further details. We repeated this 50 times and took the mean of the 50 values to generate a single result.

We first quantify the effect of incorporating the CMM into our model. We can take the output of our model and discretize the results, taking as our β-contacts, all $(i,\ j)$ such that $\mathbb{P}((i,j) \in \mathcal{I}|\mathbf{R})$ is larger than a threshold value. Taking different threshold values, Figure 7 shows the receiver operating characteristic (ROC) curve for β-contacts using both the posterior without the CMM $\mathcal{P}_1$ (dashed) and the model using the CMM $\mathcal{P}_2$ (solid). Using the CMM has significantly improved the results. For example, there is a 10% improvement in the number of true positives at the 5% false-positive rate. Figure 7 clearly shows that we have successfully used the evolutionary information, shown to exist in Figures 2 and 6, to improve the prediction of β-contacts.

We can also compare our model with existing β-contact prediction methods. For example, Table 1 shows a comparison with the Neural Network output of the first stage of BetaPro. The results quoted are AUC (Area Under Curve), the true-positive (TP) rate at 5% false positives (FP), TP at the break even point (BEP—when the total number of predicted β-contacts is equal to the true number of β contacts) and the correlation coefficient $\gamma = (\text{TPxTN} - \text{FPxFN}) / \sqrt{(\text{TP}+\text{FN})(\text{TP}+\text{FP})(\text{TN}+\text{FN})(\text{TN}+\text{FP})}$ at the BEP. This table shows that without the CMM, we produce poorer results than BetaPro. This is to be expected as $\mathcal{P}_1$ is a single sequence method, in contrast to BetaPro that inputs the whole MSA into its neural network. The addition of our CMM improves our method, producing better results than BetaPro.

Unlike some existing models, including BetaPro, our model is completely probabilistic, which enables us to predict both residue-level contacts and strand interactions simultaneously, rather than the latter needing a post processing step. Given strands E_r and E_s, they are defined to be interacting if there exist any β-contact between a residue in strand E_r and a residue in strand E_s. Using our model, we find the following:

$$\mathbb{P}(E_r, E_s \text{ interact}|\mathbf{R}) \approx \sum_{\{\mathcal{I}\}} \mathbb{I}(E_r, E_s \text{ interact}) \frac{\mathbb{P}(\mathbf{R}|\mathcal{I})}{\sum_{\{\mathcal{I}\}} \mathbb{P}(\mathbf{R}|\mathcal{I})}$$

Figure 8 shows the results for two proteins, the N-terminal domain of the yeast HSP90 chaperone [1A4H (left)] and the tetramerization domain of the Shal voltage-gated potassium channel [1NN7 (right)]. For these proteins, our model correctly predicted all strand level interactions and it is interesting to note that for 1NN7, two separate β-sheets are correctly predicted (strands {5,6}, {3,4,1,2} are distinct β-sheets), despite our model not explicitly modelling multiple sheets.

By thresholding the strand interaction probabilities at different values, we can generate a Precision [P = TP/(TP + FP)] versus Recall [R = TP/(TP + FN)] graph for strand interactions, shown in Figure 9. This figure again shows the improvement of the results when we use our CMM.

Table 2 shows a comparison of the strand interactions results for our model, the final output of BetaPro and a MLN (Lippi and Frasconi, 2009). For the comparison, we have only included independent methods and not those such as MLN-2S (Lippi and Frasconi, 2009) or those found in Aydin *et al.* (2011), which are hybrid approaches that combine results from more than one method. The results quoted for our model use the specific probability threshold of 0.45; however, taking the threshold at any

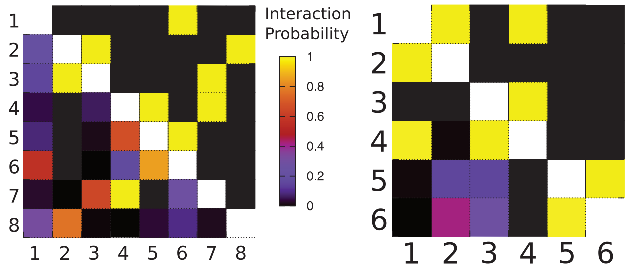

Fig. 8. Contact maps for the strand level for proteins 1A4H (left) and 1NN7 (right). Above the main diagonal, the native (true) strand interactions are shown in yellow, and below the diagonal, $\mathbb{P}(E_r, E_s \text{ interact}|\mathbf{R})$ using $\mathcal{P}_2$ is shown. For protein 1NN7, it is interesting to note that two separate β-sheets are correctly predicted, despite our model not explicitly modelling multiple sheets

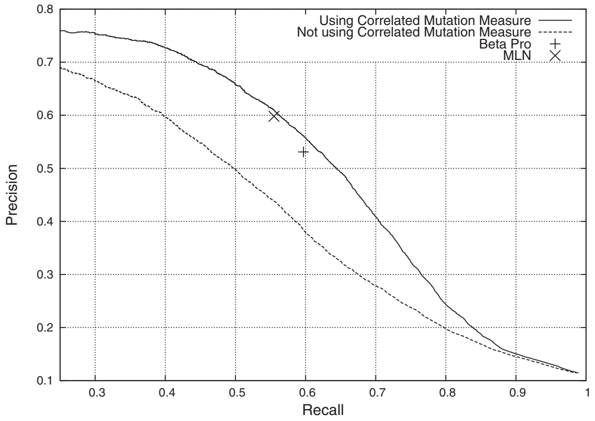

Fig. 9. Precision versus Recall graph for strand interactions. As a comparison, a naive algorithm always pairing adjacent strands yields $P = 0.42$ and $R = 0.50$. The results from the final output of BetaPro and a Markov Logic method (Lippi and Frasconi, 2009) are also displayed for comparison

Table 1. Comparison of the model with the output of BetaPro's Neural Network

Method	AUC	TP at 5% FPR	TP at BEP	γ at BEP
$\mathcal{P}_1$ (no CMM)	0.85	53%	36%	0.34
$\mathcal{P}_2$ (with CMM)	0.89	63%	44%	0.43
BetaPro	0.86	58%	41%	0.4

See text for further details. FPR = false positive rate.

value between 0.25 and 0.63 produces an $F_1 = 2PR/(P + R)$ statistic equal to or above the value found by BetaPro.

The results of our model are clearly better than BetaPro and competitive with MLN. This is an impressive result, as unlike these methods we do not require the additional information of the solvent accessibility of the residues as an input. We also do not require the secondary structure of the non-strand residues, which is important to the MLN method. The only information we use is the maximum entropy-based CMM $\mathcal{D}$ ($\mathcal{P}_1$ is a single sequence method). $\mathcal{D}$ is as useful as providing the entire MSA as a set of 20-dimensional vectors of probabilities as input to a neural or MLN. This may be because providing the columns of the MSA as independent input vectors captures the wrong information; although certain residue pairs are more likely to form β-contacts (for example, pairs of hydrophobic residues in the core of a protein), the individual pairing preferences are not especially strong, and proteins do not seem to have strong evolutionary pressure to maintain favourable pairings between strands (Mandel-Gutfreund *et al.*, 2001).

Table 2. Comparison of strand level statistics of our model ($\mathcal{P}_2$), the final output of BetaPro and the MLN method of Lippi and Frasconi (2009)

Statistic	$\mathcal{P}_2$	BetaPro	MLN
P	61.0	53.1	59.8
R	55.4	59.7	55.5
F_1	58.1	56.2	57.6
γ	0.532	0.508	0.528
Chains with $F_1 \geq 70.0$	35.0	31.7	33.7

Apart from γ, statistics are shown as percentages.

Also, just considering the specific residue types, rather than how they co-vary, suffers from the problem of transitivity: if E_r is paired with both E_s and E_t, then it is often the case E_t and E_s themselves contain residues with favourable pairings, as they both favourably interact with E_r.

For our method to be useful for proteins with unknown structure, it is important to test our method with predicted secondary structure. In the Supplementary Data, we have presented results for the CASP 2010 set of proteins using both known and predicted strand structure, and in both cases our method compares favourably with BetaPro.

4 CONCLUSION AND FURTHER WORK

In this article, we have used a statistical machine learning approach known as contrastive divergence to efficiently calculate a Maximum Entropy distribution that models the evolutionarily related family of a protein and have used this to calculate a CMM to predict residue contacts. We have coupled this measure to a probabilistic model of β-strand interactions to produce a state-of-the-art β-contact predictor that can be used even if a poor quality or no MSA is available. The current focus of our work is to allow joint inference of β-contacts and secondary structure by incorporating a semi-segmental Markov model to model the secondary structure of proteins (Chu *et al.*, 2006; Schmidler *et al.*, 2000).

Unlike other recent CM studies, we have focused on proteins that do not necessarily have large enough MSAs to enable full tertiary structure determination using a CM approach. However, our strand interaction prediction can be incorporated into a tertiary structure prediction method. For example, our previously published work describes a coarse-grained protein model that uses a physically meaningful energy function, biased by a harmonic potential on β-contacts to enable the protein to fold (Burkoff *et al.*, 2012; Podtelezhnikov and Wild, 2008). Using our strand prediction method to predict β-contacts enables this model to be used for protein tertiary structure prediction.

Further details and specific examples are shown in the Supplementary Data.

ACKNOWLEDGEMENT

We thank John Skilling for helpful discussions.

Funding: We acknowledge support from the Leverhulme Trust (grant F/00 215/BL to N.S.B, C.V. and D.L.W.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search proteins. *Nucleic Acids Res.*, **25**, 3389–3402.

Ashkenazy,H. *et al.* (2009) Optimal data collection for correlated mutation analysis. *Proteins*, **74**, 545–555.

Aydin,Z. *et al.* (2011) Bayesian models and algorithms for protein β-sheet prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **8**, 395–409.

Balakrishnan,S. *et al.* (2011) Learning generative models for protein fold families. *Proteins*, **79**, 1061–1078.

Burger,L. and van Nimwegen,E. (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.*, **6**, e1000633–51.

Burkoff,N.S. *et al.* (2012) Exploring the energy landscapes of protein folding simulations with bayesian computation. *Biophysical. J.*, **102**, 878–886.

Cheng,J. and Baldi,P. (2005) Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, **21**, i75–i84.

Chu,W. *et al.* (2006) Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **3**, 98–113.

Dekker,J.P. *et al.* (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.

Duarte,J.M. *et al.* (2010) Optimal contact definition for reconstruction of contact maps. *BMC Bioinformatics*, **11**, 283.

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

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.*, **32**, 1792–1797.

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

Hinton,G.E. (1999) Products of experts. In: *Proceedings of the Ninth International Conference on Artificial Neural Networks*, Vol. 1. University of Edinburgh, UK, pp. 1–6.

Hinton,G.E. (2002) Training products of experts by minimizing contrastive divergence. *Neural Comput.*, **14**, 1771–1800.

Hopf,T.A. *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.

Hubbard,T.J. and Park,J. (1994) Use of β-strand interaction pseudo potentials in protein structure and modelling. In: *Proceedings of the 27th Hawaii Int'l Conf. System Sciences*. Maui, HI, USA, pp. 336–344.

Jaynes,E.T. (2007) *Probability Theory: The Logic of Science*. CUP, Cambridge, UK.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kortemme,T. *et al.* (1998) Design of a 20-amino acid, three-stranded β-sheet protein. *Science*, **281**, 253–256.

Lee,B.-C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li,W. *et al.* (2001) Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics*, **17**, 282–283.

Lippi,M. and Frasconi,P. (2009) Prediction of protein β-residue contacts by Markov logic networks with grounding-specific weights. *Bioinformatics*, **25**, 2326–2333.

Mandel-Gutfreund,Y. *et al.* (2001) Contributions of residue pairing to beta-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel beta-strands. *J. Mol. Biol.*, **305**, 1145–1149.

Marks,D.S. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.

Merkel,J.S. and Regan,L. (2000) Modulating protein folding rates in vivo and in vitro by side chain interactions between the parallel beta strands of green flluorescent protein. *J. Biol. Chem.*, **275**, 29200–29206.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA*, **108**, E1293–E1301.

Olmea,O. and Valencia,A. (1997) Improving contact prediction by the combination of correlated mutations and other sources of sequence information. *Fold. Des.*, **2**, S25–S31.

Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.

Podtelezhnikov,A.A. and Wild,D.L. (2008) Crankite: a fast polypeptide backbone conformation sampler. *Source Code Biol. Med.*, **3**, 12.

Podtelezhnikov,A.A. and Wild,D.L. (2009) Reconstruction and stability of secondary structure elements in the context of protein structure prediction. *Biophys. J.*, **96**, 4399–4408.

Rajgaria,R. *et al.* (2010) Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD. *Proteins*, **78**, 1825–1846.

Ruczinski,I. *et al.* (2002) Distribution of beta sheets in proteins with application to structure prediction. *Proteins*, **48**, 85–97.

Saqi,M.A.S. *et al.* (1999) Protein analyst—a distributed object environment for protein sequence and structure analysis. *Bioinformatics*, **15**, 521–522.

Sathyapriya,R. *et al.* (2009) Defining an essence of structure determining residue contacts in proteins. *PLoS Comput. Biol.*, **5**, e1000584.

Schmidler,S.C. (2002) Statistical models and monte carlo methods for protein structure prediction. In: PhD Thesis, Stanford University, Stanford, CA, USA.

Schmidler,S.C. *et al.* (2000) Bayesian segmentation of protein secondary structure. *J. Comput. Biol.*, **7**, 232–248.

Smith,C.K. and Regan,L. (1995) Guidelines for protein design: the energetics of β sheet side chain interactions. *Science*, **270**, 980–982.

Sonnhammer,E.L.L. *et al.* (1997) Pfam: a comprehensive database of protein families based on seed alignments. *Proteins*, **28**, 405–420.

Steward,R.E. and Thornton,J.M. (2002) Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins Struct. Funct. Genet.*, **48**, 178–191.

Sułkowska,J. *et al.* (2012) Genomics-aided structure prediction. *Proc. Natl Acad. Sci. USA*, **109**, 10340–10345.

Vendruscolo,M. *et al.* (1997) Recovery of protein structure from contact maps. *Fold. Des.*, **2**, 295–306.

Weigt,M. *et al.* (2008) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA*, **106**, 67–72.