Bioinformatics Advance Access published January 23, 2015

bbcontacts: prediction of β-strand pairing from direct coupling patterns

Jessica Andreani [1,2,*] and Johannes Söding [1,2,*]

[1]Gene Center, LMU Munich, Feodor-Lynen-Strasse 25, 81377 Munich, Germany
[2]Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

Associate Editor: Prof. Anna Tramontano

ABSTRACT

Motivation: It has recently become possible to build reliable *de novo* models of proteins if a multiple sequence alignment (MSA) of at least 1000 homologous sequences can be built. Methods of global statistical network analysis can explain the observed correlations between columns in the MSA by a small set of directly coupled pairs of columns. Strong couplings are indicative of residue-residue contacts, and from the predicted contacts a structure can be computed. Here, we exploit the structural regularity of paired β-strands that leads to characteristic patterns in the noisy matrices of couplings. The β-β contacts should be detected more reliably than single contacts, reducing the required number of sequences in the MSAs.

Results: bbcontacts predicts β-β contacts by detecting these characteristic patterns in the 2D map of coupling scores using two hidden Markov models (HMMs), one for parallel and one for antiparallel contacts. β-bulges are modeled as indel states. In contrast to existing methods, bbcontacts uses predicted instead of true secondary structure. On a standard set of 916 test proteins, 34% of which have MSAs with <1000 sequences, bbcontacts achieves 50% precision for contacting β-β residue pairs at 50% recall using predicted secondary structure and 64% precision at 64% recall using true secondary structure, while existing tools achieve around 45% precision at 45% recall using true secondary structure.

Availability: bbcontacts is open source software (GNU Affero GPL v3) available at https://bitbucket.org/soedinglab/bbcontacts

Contact: jessica.andreani@mines.org; soeding@mpibpc.mpg.de

Supplementary information: available at *Bioinformatics* online.

1 INTRODUCTION

Methods for protein structure prediction can be classified into template-based and *de novo* methods. The first group model the structure for a query protein based on a sequence alignment with a homologous template protein of known structure. This class of methods is by far the most widely used for its speed and reliability. Many protein families lack a 3D template, and therefore much effort has been invested into developing methods for *de novo* protein structure prediction.

The most successful *de novo* prediction methods, such as ROSETTA (Leaver-Fay *et al.*, 2011), are based on complex, knowledge-based scoring functions and structural fragment assembly. Even though in roughly a quarter of the cases the top methods can produce models with the correct fold (Tai *et al.*, 2014), their practical usefulness is severely limited by the difficulty of predicting which models are correct.

An alternative *de novo* approach relies on the observation that correlated mutations between pairs of MSA columns could predict physical contacts between residues (Göbel *et al.*, 1994). Furthermore, it was realized that only few correctly predicted residue-residue contacts ($\sim 10\%$ of the number of residues) are sufficient to predict the correct protein fold (Skolnick *et al.*, 1997; Kim *et al.*, 2014). However, it was only recently that statistical methods were applied to the MSAs that could distinguish direct couplings between MSA columns from mere transitive correlations (Weigt *et al.*, 2009; Marks *et al.*, 2011). This allowed for the first time the reliable *de novo* prediction of structures for proteins with many homologs (Marks *et al.*, 2011; Hopf *et al.*, 2012; Nugent and Jones, 2012). In the past few years, these methods have been further improved by applying different approaches of direct coupling analysis (Jones *et al.*, 2012; Kamisetty *et al.*, 2013; Ekeberg *et al.*, 2013). Yet, obtaining reliable structural models requires large numbers of homologous sequences, still severely limiting the scope of these methods.

Here, our goal is to increase the reliability of contact predictions by detecting patterns in the matrix of predicted couplings corresponding to interactions between secondary structure elements. We focus on the case of β-β contacts because of their strongly constrained spatial arrangement. β-sheets are composed of regularly arranged pairs of interacting β-strands. The interaction between two extended β-strands is defined on the basis of regular patterns of hydrogen bonds, connecting residues in two different strands in either a parallel or an antiparallel fashion (Kabsch and Sander, 1983). The prediction of contacts between β-residues has applications in protein design (Smith and Regan, 1995; Kortemme, 1998), in the study of folding characteristics (Merkel and Regan, 2000; Kamat and Lesk, 2007) and in *de novo* protein structure prediction (Ruczinski *et al.*, 2002; Klepeis and Floudas, 2003). According to a recent study, while direct contact predictions have similar average precision for mainly-α and mainly-β proteins, the structural models obtained using these predictions as restraints are more accurate for mainly-α proteins (Michel *et al.*, 2014).

A variety of methods have been developed for the prediction of β-β contacts. An early method used statistical potentials for pairs of interacting β-strand residues (Hubbard, 1994). Baldi *et al.* (2000)

[*] to whom correspondence should be addressed

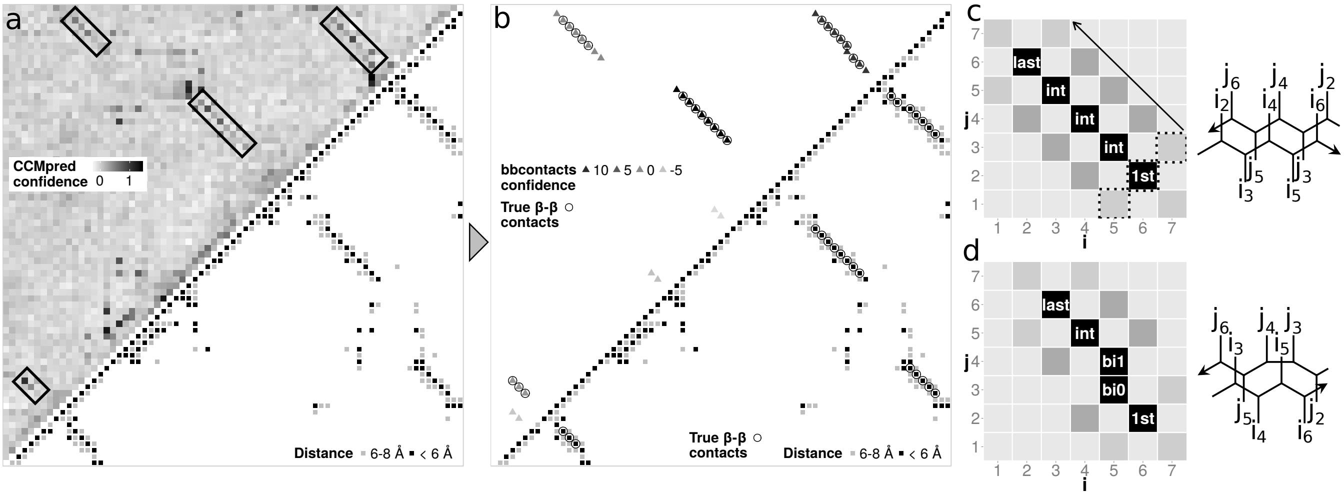

Fig. 1. (a) CCMpred coupling matrix (upper-left) and coarse Cβ-Cβ distance matrix (lower-right) for one domain of PDB structure 3dk9, with four boxed regions containing patterns created by antiparallel β-strands. (b) Upper-left: β-β contacts predicted by bbcontacts using predicted secondary structure (triangles). The Viterbi score of the local alignment is the confidence value. Lower-right: coarse Cβ-Cβ distance matrix. The true β-β contacts (annotated by DSSP) are shown as open circles. (c-d) Schematic diagrams describing (c) a regular antiparallel interaction between two β-strands and the associated pattern and (d) an antiparallel interaction involving a β-bulge and the associated pattern, displaying a shift in the main diagonal. The HMM states associated to the main diagonal are annotated in patterns (c-d) ("int" is short for internal and "bi0", "bi1" for bulge i_0, bulge i_1; see also Figure 2). In (c), the dotted lines delineate the three couplings – one main diagonal and two secondary diagonal couplings – entering into the emission probability calculation for cell (6,2). The arrow indicates the direction of Viterbi score calculation and decoding for antiparallel contacts.

were the first to predict β-β contacts with neural networks. Cheng and Baldi (2005) introduced the important idea of exploiting the topology of β-sheets in their BetaPro method by filtering out the solutions incompatible with the specific geometry of β-sheets. In MLN and MLN-2S, Lippi and Frasconi (2009) used Markov logic networks to incorporate structure-based constraints directly into the learning process.

Two recent methods use correlated mutation signatures to predict β-sheets. CMM (Burkoff *et al.*, 2013) integrates these with a β-topology model. BCov (Savojardo *et al.*, 2013) processes predicted coupling scores with integer programming to enforce topological constraints. CMM, BCov and MLN-2S display the best β-β contact prediction performances so far (Savojardo *et al.*, 2013).

Finally, some methods aiming to predict all protein contacts (not only β-β contacts) also use topological information, in particular related to β-sheet organisation: CMAPpro (Di Lena *et al.*, 2012), PhyCMAP (Wang and Xu, 2013), PconsC2 (Skwark *et al.*, 2014). PconsC2 reports the highest contact prediction accuracies to date. It takes direct coupling scores from the PconsC meta-predictor as input and trains random forests on local 11x11 windows in the coupling matrix to predict the contact state of the central cell.

Existing β-β contact prediction methods have used known instead of predicted secondary structure. It is unclear, however, how they would perform in practice when substituting true with predicted secondary structure. Here, we describe a method to predict β-β contacts that, even though it makes use of predicted instead of true secondary structure, achieves better performance than previous methods. We designed two hidden Markov models (HMMs) for parallel and antiparallel β-β contacts that integrate signals from the predicted couplings and the predicted secondary structure.

2 METHODS

2.1 General approach

Similarly to contact maps or distance maps, which can be used to describe a protein structure or compare two structures (Holm and Sander, 1996), direct coupling predictions can be mapped on a two-dimensional grid. Each cell in the resulting matrix contains the strength of the predicted coupling, indicating whether the two corresponding positions in the protein are reliably predicted to be in direct physical contact. We call such representations "matrices of predicted couplings" or simply "coupling matrices".

Interactions between β-strands create conspicuous patterns in the coupling matrices, linked to the regularity of their 3D structural arrangement. Figure 1a shows a coupling matrix for a domain of PDB structure 3dk9, with four highlighted antiparallel β-β contacts. A schematic antiparallel pattern is displayed in Figure 1c together with a diagram showing the corresponding contacts between β-residues. In short, β-strand interactions create a diagonal stretch of strong couplings between the closest residues. This stretch is perpendicular (respectively parallel) to the diagonal of the coupling matrix for antiparallel (respectively parallel) β-strands. Most often, given a contact between residues i and j on this main diagonal, increased couplings can also be observed between i and $j \pm 2$ and between j and $i \pm 2$, as the corresponding side-chains are close in space and point in the same direction. Such couplings form "secondary diagonals" on both sides of the main diagonal of the pattern.

Interactions between β-strands are very regular, but β-bulges constitute a frequent type of irregularity disrupting the regular alternation of side-chain direction (Richardson *et al.*, 1978; Chan *et al.*, 1993; Craveur *et al.*, 2013). The most frequent β-bulges arise from the insertion of a residue between successive hydrogen bonds connecting two β-strands. Such β-bulges induce a shift in the main diagonal of the pattern by one position, as illustrated in Figure 1d. In some rarer cases, a β-bulge can arise from the insertion of one residue on each β-strand (so that the main diagonal of the pattern is not shifted) or from the insertion of more than one residue on one β-strand (so that the main diagonal of the pattern is shifted by more than one position).

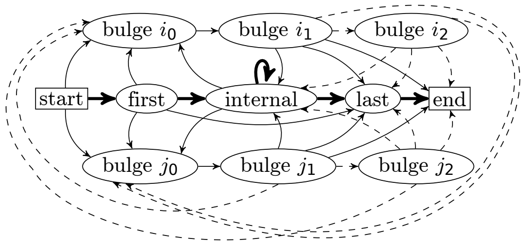

Fig. 2. Hidden Markov model architecture used in this study. Bold arrows represent the most frequently observed transitions and dashed arrows represent rare transitions. For the HMM detecting parallel (resp. antiparallel) contacts, most transitions correspond to a diagonal (resp. antidiagonal) displacement in the coupling matrix, except the transitions between bulge i states (transitions along the j axis with i remaining fixed, such as the one displayed in Figure 1d) and the transitions between bulge j states (transitions along the i axis with j remaining fixed). A β-bulge involves at least one (and at most two) inserted residue(s) on one of the two strands, thus state bulge i_0 can only be followed by bulge i_1 (same for bulge $j_0 \rightarrow$ bulge j_1).

To detect the patterns created by β-β contacts, we designed the hidden Markov model (HMM) architecture shown in Figure 2. In contrast to most HMMs used in bioinformatics applications (e.g. transmembrane helix predictions), our HMM detects patterns in a 2D map and not in a 1D sequence or 1D sequence profile. The nine hidden states circled in Figure 2 represent different types of interactions between pairs of β-residues. The “first” and “last” states represent the first and last non-bulge residue-residue contacts. The “internal” state is the main HMM state and β-β contact predictions can be extended to any length by looping through this state. β-bulges correspond to indels in HMMs used for pairwise sequence alignment. We only take into account β-bulges inducing a shift in the pattern and this shift is achieved through the transition from bulge i_0 to bulge i_1 or from bulge j_0 to bulge j_1. All HMM transitions except those between two bulge states correspond to a diagonal (respectively antidiagonal) displacement for parallel (respectively antiparallel) β-β contacts. Bulge states i_2 and j_2 allow for the insertion of two residues in the same β-bulge; we do not allow for the insertion of more than two residues. The “start” and “end” states are added for modelling both ends of the contact between two extended β-strands. Figures 1c and 1d show coupling patterns with annotated HMM states.

The same HMM architecture is used for parallel and antiparallel β-β contacts, but two different sets of HMM parameters are trained (section 2.5) and the direction of HMM decoding is diagonal for parallel contacts and antidiagonal for antiparallel ones.

The observed variables are the real-valued predicted couplings as well as the predicted secondary structure states (or the true secondary structure states assigned by DSSP (Kabsch and Sander, 1983), when used in order to compare our method with existing ones). The HMM emission probabilities consequently include a product of continuous emissions accounting for the observed couplings (including signals from the main diagonal and the secondary diagonals of the pattern) and discrete emissions accounting for the discrete secondary structure states (H, E and C). These emissions are described in more detail in section 2.5.

A path (sequence of states) traced by the HMM through the coupling matrix corresponds to a contact between two extended β-strands. Each state within the path corresponds to an interaction between two β-residues. We want to detect all patterns in the coupling matrix corresponding to parallel and antiparallel β-strand contacts. To this effect, we use a local version of the Viterbi algorithm, so that we can predict several paths that can start and end anywhere in the coupling matrix. The Viterbi score associated to a path measures the confidence of the corresponding prediction. We rank the paths by decreasing Viterbi score and retain all paths (above a given threshold) which satisfy the topological constraints associated with β-strand pairings.

In the following sections, we describe the training and benchmarking of bbcontacts.

2.2 Datasets

To compare our method with existing ones, two previously published test datasets were used. The BetaSheet916 dataset (Cheng and Baldi, 2005) has been routinely used as a benchmark dataset for β-β contact prediction. It is also our main test dataset. It consists of 916 protein chains containing 31,638 β-residue contacts. Savojardo *et al.* (2013) recently proposed a complementary dataset, built from more recent structures. This new dataset, BetaSheet1452, consists of 1452 protein chains containing 56,552 β-residue contacts. The two test datasets are non-redundant at 20% sequence identity, both internally and with each other.

Our training dataset was built from the CATH database of protein domains v3.5 (Sillitoe *et al.*, 2013). The building process aimed to reduce as much as possible the redundancy between the training dataset and both test datasets at the fold (CATH Topology) level. We extracted all CATH domains that did not belong to any of the fold groups identified by CATH in the test datasets. We filtered the resulting dataset to reduce internal redundancy, using the HH-suite script `pdbfilter.pl` (Remmert *et al.*, 2011). Finally, 943 domains containing β-contacts form our training dataset (Supplementary Dataset S1). Because not all chains from the test datasets were annotated in CATH v3.5, there might be some residual redundancy between the training dataset and the test datasets. We checked that this did not lead to overtraining of bbcontacts, by verifying that the results did not deteriorate when taking the subset of each test dataset that is strictly non-redundant with the training dataset at the fold level (see Supplementary Results and Supplementary Figure S1).

All β-contacts were assigned based on backbone contacts, following the DSSP definition (Kabsch and Sander, 1983). For both training and testing, the DSSP assignment was reduced to three states (B and E were mapped to E; H, I and G were mapped to H; T, S and C were mapped to C).

β-bulges were detected by PROMOTIF (Hutchinson and Thornton, 1996). We then filtered the detected β-bulges to retain only those inducing a shift of the main diagonal in the direct coupling pattern (see Figure 1d).

2.3 Data used for HMM training

For the training and test datasets, we built MSAs by running HHblits (Remmert *et al.*, 2011) against the uniprot20 database dated March 2013. Each MSA was filtered down to 90% sequence identity with HHfilter (Remmert *et al.*, 2011). Supplementary Figure S2 shows the distribution of the number of sequences in the resulting MSAs for each dataset.

Secondary structure predictions were obtained with PSIPRED (Jones, 1999), using the HH-suite script `addss.pl` for improved performance (Remmert *et al.*, 2011). Direct coupling predictions were obtained with CCMpred (Seemayer *et al.*, 2014), a fast implementation of the state-of-the-art methods by Kamisetty *et al.* (2013) and Ekeberg *et al.* (2013).

We observed that the range of predicted couplings varied greatly depending on the number of sequences N in the MSA and the protein length L. We found that $\eta = \sqrt{N}/L$ was a good descriptor for the range of couplings observed in a predicted matrix. For each domain in the training dataset, we filtered the initial MSA in order to build MSAs of reduced diversity, using the qsc parameter of HHfilter (Remmert *et al.*, 2011). We derived 12 datasets for $\eta = 0.05, 0.1, 0.2, ..., 1.0, 1.2$ (see Supplementary Table S1). We ran CCMpred on all diversity-filtered alignments and used the resulting coupling matrices to train the coupling-based part of the HMM emission probabilities.

2.4 Local background correction of the coupling matrices

The coupling matrices sometimes display darker regions that can lead to many false positive predictions (Supplementary Figure S3a). We therefore applied the following local background correction procedure to all coupling matrices in the training and test datasets: from each coupling, we subtracted the average coupling over an area of size $(2S+1)$x$(2S+1)$, extending by S cells in each direction. For single domains with a good alignment coverage (including most domains in the training dataset), this procedure has almost no effect on the coupling values (Supplementary Figure S3b).

2.5 HMM parameters

The HMMs were trained using the labeled data contained in the training dataset. We trained two sets of HMM parameters separately for parallel and antiparallel β-strand pairings. We did not train bbcontacts for the detection of β-bridges, because we do not expect the residues involved in isolated β-bridges to be generally predicted as β-residues by PSIPRED, and because the coupling signals typically do not form patterns for such isolated β-contacts.

The HMM transition probabilities were trained by counting how many times each transition was used in the training dataset. The HMM emission probabilities contain a product of two terms, one based on couplings and one based on secondary structure (from either PSIPRED predictions or DSSP assignments). Each term is expressed as the odds-ratio of the conditional distribution of the observed variables when in one of the HMM states, relative to the background distribution.

2.5.1 Coupling-based emissions. The coupling-based part of the emission probability at position (i, j) was expressed as the product of three odds-ratios relative to the background: one for the central coupling at position (i, j) belonging to the main diagonal of the pattern and one for each of the two couplings at the positions adjacent to (i, j) belonging to the secondary diagonals of the pattern (Supplementary equations 1-4). This is illustrated in Figure 1c: the dotted lines delineate the three couplings entering into the emission probability calculations for cell $(i = 6, j = 2)$, at positions (i, j), $(i-1, j-1)$ and $(i+1, j+1)$. These adjacent cells are chosen rather than $(i, j \pm 2)$ and $(i \pm 2, j)$ to avoid multiple counting.

Because of data scarcity, we did not distinguish between different HMM states and bundled all β-contacts together for this stage of the training. For each decoding direction (parallel and antiparallel), we thus had to describe three coupling distributions: one for the background, one for the main diagonal of the patterns and one for the secondary diagonals of the patterns.

After centering the coupling distributions at zero, we fitted their density using two transformed Gamma distributions, one for positive couplings and one for negative couplings. To describe the density fit for a given value of η, we used 7 parameters: the shift needed to center the coupling distribution, the relative weight of the positive and negative sides, plus two transformed Gamma parameters for negative couplings and three for positive couplings. The shift was fitted as a quadratic function of η and all remaining parameters were expressed as linear functions of η. The optimization was performed by maximum likelihood estimation. The final number of parameters for the coupling-based emissions is 90. Details are given in the Supplementary Methods and the final fits are illustrated in Supplementary Figure S4.

2.5.2 Secondary-structure-based emissions. This part of the emission probabilities was trained separately for the PSIPRED predictions and the DSSP assignments. We used a discrete mapping of the secondary structure observations to three states (E, H and C). In the PSIPRED case, the most probable secondary structure state was used for each position.

We denote by z an HMM state and by (σ_i, σ_j) the pair of secondary structure states at position (i, j). We tested two types of secondary-structure-based emissions. The "non-conditional" emissions were defined as the set of probabilities $p(\sigma_i, \sigma_j|z)$, as expected from the traditional definition of emission probabilities. However, this does not account for the important fact that by definition, a secondary structure element is a segment of residues immediately adjacent in sequence. Therefore, the "conditional" emissions were defined as the set of probabilities $p(\sigma_i, \sigma_j|\sigma_{i_{\text{prev}}}, \sigma_{j_{\text{prev}}}, z)$ of observing a pair of secondary structure states (σ_i, σ_j) in state z, given that we additionally already observed secondary structure states $(\sigma_{i_{\text{prev}}}, \sigma_{j_{\text{prev}}})$ at the previous position $(i_{\text{prev}}, j_{\text{prev}})$.

We also added pseudocounts derived from the non-conditional probability distribution to the conditional probabilities: when calculating the conditional probabilities, we added N_0 counts from the non-conditional frequencies to the observed conditional counts. This effectively interpolates between the conditional and non-conditional distributions: for states with few conditional counts relative to N_0, the conditional probabilities with added pseudocounts will be very similar to the non-conditional probabilities. Different values of N_0 were tested (see section 3.1).

Because we found that the secondary structure states for the coupling matrix cells situated immediately before and immediately after a β-strand interaction contain information about the likelihood to start and end this interaction, the model also contains secondary-structure-based emission terms for the start and end states. The start term is always a non-conditional probability and can also be seen as a prior based on secondary structure.

In total, there are 415 parameters for the DSSP secondary-structure based emissions and 415 parameters for the PSIPRED-based emissions (see Supplementary Methods and Supplementary equations 5-10).

2.5.3 Prior probability distribution depending on sequence separation. The sequence separation between two interacting β-strands has a lower bound due to geometric considerations, especially for the parallel case. In addition, it is strongly biased in practice, especially for antiparallel β-strands, a majority of which are β-hairpins for which the two strands are separated by a short loop.

Thus, we introduced a prior for starting a β-strand interaction depending on the sequence separation between the first pair of interacting residues. This prior is based on 34 fitted parameters for PSIPRED-based predictions and 34 fitted parameters for DSSP-based predictions (Supplementary equation 11 and Supplementary Figure S5). We also introduced constraints to prevent the detection of β-contacts for positions too close to the diagonal of the coupling matrix. More details are given in the Supplementary Methods.

2.6 HMM decoding

Following the ideas introduced by Muckstein *et al.* (2002) for local sequence-sequence alignment and by Biegert and Söding (2008) for local HMM-HMM alignment, we use a local version of the Viterbi algorithm for decoding a matrix of predicted couplings to detect patterns corresponding to β-strand interactions. We dispense with an explicit background state by using odds-ratio emission probabilities: each emission probability corresponding to an HMM state is always divided by a background probability. The HMM paths can start and end anywhere in the coupling matrix. We do not detect a single most likely path corresponding to the best Viterbi score, but instead all paths above a certain Viterbi score threshold.

For each position (i, j) and each state z, the Viterbi variable $V[i, j, z]$ is defined as the probability of ending a path at position (i, j) and in state z. The local Viterbi algorithm consists of four major steps: initialization, recursion, termination and back-tracing (see Supplementary Methods, including Supplementary equations 12-14).

In the initialization step, the Viterbi variables $V[i, j, \text{start}]$ are initialized for all positions (i, j) in the coupling matrix and the priors described above are applied. In the recursion step, all $V[i, j, z]$ for $z \notin \{\text{start}, \text{end}\}$ are calculated using the transition and emission probabilities. In the termination step, the $V[i, j, \text{end}]$ probabilities are calculated. During recursion and termination, pointers are used to keep track of the most likely paths.

The initialization, recursion and termination steps of the Viterbi decoding are performed separately for the parallel and antiparallel directions, but all $V[i, j, \text{end}]$ scores are then merged and sorted in decreasing order for the final back-tracing step. The most likely path, corresponding to the highest $V[i, j, \text{end}]$ probability, is retrieved by back-tracing through the saved pointers. Then, we cross out a region extending by ± 3 residues around this path in the Viterbi matrix corresponding to the path direction (parallel or antiparallel), i.e. we do not take into account any more probabilities for this region. This avoids retrieving many variants of a contact between the same β-strands. The next path that does not contain any crossed-out residue pairs is then saved and a region around this path is crossed-out. We proceed iteratively in this manner until we reach a given Viterbi score threshold.

2.6.1 Prediction-shortening mode (PSM). For PSIPRED-based results, in some cases, the Viterbi paths can be very long because of the spatial architecture of the protein (see Supplementary Figure S6 for an example). We designed a procedure, called "PSM" (prediction-shortening mode), to ensure that the predicted parallel paths stay below 11 residue pairs in length and the predicted antiparallel paths below 15 residue pairs. When PSM is triggered, it shortens the predicted paths by iteratively decreasing the transition probabilities and rerunning the Viterbi algorithm until the predicted paths are below the length threshold. For DSSP-based results, such a procedure is unnecessary because the secondary-structure-based probabilities make it impossible to predict contacts between non-β-residues, so that the length of any path is limited by the length of the longest β-strands.

More details about PSM are given in the Supplementary Methods.

2.7 Topology filtering

When all most likely Viterbi paths have been retrieved, a final post-processing step is applied to filter the incompatible paths given the topological constraints that apply to β-β contacts. We go through the list of retrieved paths sorted by decreasing Viterbi score. Each path can be retained or excluded. A path is excluded if it contains a residue that already has two β-partners in previously retained paths or a residue pair that already belongs to a previous path. Using this residue-based filtering rather than a strand-based filtering means that bbcontacts can handle cases where a β-strand is in contact with more than two strands, as also pointed out for previous methods (Cheng and Baldi, 2005). In addition, when the DSSP assignment is used, we can rely on the exact positions of β-strands, so we exclude any path containing a contact between two residues from the same β-strand or between a pair of β-strands that already belongs to a previous path.

2.8 Evaluation

Performance is measured in terms of precision and recall at the strand level and at the residue level, as was done for previous β-β contact prediction methods. The F1-score (harmonic mean of precision and recall) is also used, as it provides a single value to measure the quality of the β-β contact predictions. bbcontacts is compared with the best methods available so far: BetaPro (Cheng and Baldi, 2005), MLN and MLN-2S (Lippi and Frasconi, 2009), CMM (Burkoff *et al.*, 2013) and BCov (Savojardo *et al.*, 2013).

Residue-level evaluation is straightforward in all cases. Strand-level evaluation is only straightforward for DSSP-based results. For PSIPRED-based results, because the true β-strand positions are unknown, additional conventions need to be adopted for the strand-level evaluation (see Supplementary Methods). Because PSIPRED-based strand-level evaluation is based on these additional criteria, it is provided only in an indicative manner and the residue-level evaluation forms the most solid basis for comparison between different versions of our method.

bbcontacts is also compared with general contact predictors: CCMpred (Seemayer *et al.*, 2014), PhyCMAP (Wang and Xu, 2013) and PconsC2 (Skwark *et al.*, 2014). Baselines for these methods are obtained by restricting predictions to DSSP-defined β-strand regions. To make the comparison fairer to these methods, we exclude all false positives with sequence separation smaller than 6 from their predictions. True positives are unchanged (β-β contacts assigned by DSSP).

3 RESULTS AND DISCUSSION

For the sake of simplicity, all results are shown for the BetaSheet916 test dataset. However, the trends described also hold for the training dataset and the BetaSheet1452 test dataset (see Supplementary Results and Supplementary Figures S21 to S25).

3.1 Contribution of the different terms in bbcontacts

This section illustrates how much the various steps composing the bbcontacts method contribute to its performance. Here, the secondary structure predicted by PSIPRED is used as an input. This section is focused on residue-level performance because this assessment is more stringent.

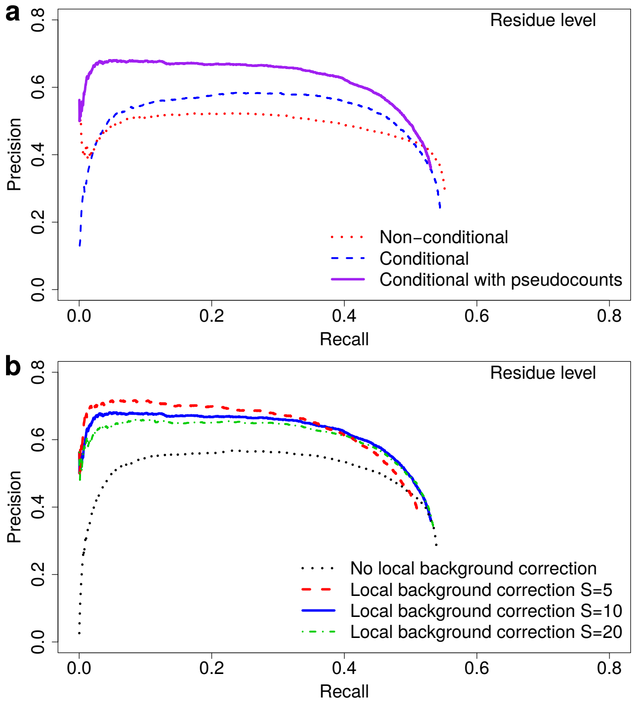

Fig. 3. Influence of different model parameters on the residue-level performance of bbcontacts on the BetaSheet916 dataset, using PSIPRED predictions as an input. (a) Influence of the type of secondary-structure-based emissions: non-conditional (red), conditional (blue), conditional with 10,000 pseudocounts (purple). (b) Influence of local background correction applied to coupling matrices, for different values of S.

We use a reference version of bbcontacts in which local background correction with S=10 is applied to the coupling matrices, the conditional secondary-structure-based emission probabilities with N_0=10,000 pseudocounts from the non-conditional distribution are used, the prior depending on sequence separation is used and prediction-shortening mode (PSM) is turned off. We then modify one element of the model at a time and discuss the impact of each of the corresponding terms.

First of all, when developing bbcontacts, we tested "conditional" and "non-conditional" secondary-structure-based emission probabilities (see Methods). We also tried adding pseudocounts from the non-conditional probability distribution to the conditional probabilities. This was motivated by the observation that the conditional probabilities for some rarely observed states were derived from very low counts; in this case, adding counts from the non-conditional distribution (for which all states are well populated) should make the resulting probabilities more robust. However, we found an optimal number of pseudocounts N_0 of 10,000 on the training dataset (see Supplementary Figure S7), much larger than we would expect if the pseudocounts were just used to avoid overtraining of the conditional probabilities derived from low counts. This large number of pseudocounts actually performs an interpolation between the conditional and non-conditional probability distributions.

This can be understood by noting that the dependency between two consecutive states is already described to some extent by the HMM transitions for the HMM states, but not for the background. Consequently, the conditional probabilities are better suited to the background than to the HMM states. Using 10,000 pseudocounts, the secondary-structure-based probabilities are very close to the conditional ones for the background (which has several million counts in the training data), but intermediate between conditional and non-conditional for the HMM states (which have a few hundred to a few thousand counts).

Figure 3a shows the prediction results using the different types of secondary-structure-based emissions: purely non-conditional, purely conditional, or conditional with 10,000 pseudocounts taken from the non-conditional distribution. Adding pseudocounts from the non-conditional distribution to the conditional probabilities significantly improves the performance of bbcontacts.

Another major contribution to the performance of bbcontacts is brought by correcting the local background of the coupling matrices in order to remove false positives, by avoiding the presence of dark regions concentrating strong couplings in the predicted coupling matrices (see Methods). The results in Figure 3b show a notable performance improvement when using local background correction, for the three displayed values of S. We chose S=10 as the default parameter for bbcontacts, because it displays the best improvement in precision without loss in recall. The results for S=5 show a slightly higher precision for high-confidence predictions, but the final recall is also lower.

Other terms in the bbcontacts model have a smaller influence on the final performance. The prior depending on sequence separation has a small, but consistently positive effect on the performance of bbcontacts (Supplementary Figure S8a). The influence of including signal from the secondary diagonals of the patterns as well as signal from the main diagonal is analyzed in Supplementary Figure S8b. Inclusion of secondary diagonal signal notably increases the final recall reached by bbcontacts. However, taking signal only from the main diagonal slightly increases the precision for high-scoring predictions. This effect is dampened when prediction-shortening mode (PSM) is turned on. Finally, the red traces in Figure 4b show that PSM slightly increases the precision for high-confidence residue-level predictions, because it removes false positives belonging to long paths with large Viterbi scores. The strand-level performance is almost unaffected by PSM (Figure 4a).

Although all methodological choices for bbcontacts illustrated in this section relate to PSIPRED-based β-β contact predictions, the performance of bbcontacts when using the DSSP assignments as an input is either improved or unchanged by these choices (see Supplementary Figures S9 and S10).

3.2 Comparison with previous methods

Figure 4 presents the precision-recall results on the BetaSheet916 dataset for the final version of bbcontacts compared to previous β-β contact prediction methods, at the strand level (Figure 4a) and at the residue level (Figure 4b). Supplementary Tables S2 and S3 present a summary of this comparison. Supplementary Figure S11 shows the strand-level results when testing for correct orientation of predicted β-strands as well as correct pairing.

As a complementary view, Figure 1b and Supplementary Figure S12 show examples of contact maps predicted by bbcontacts using

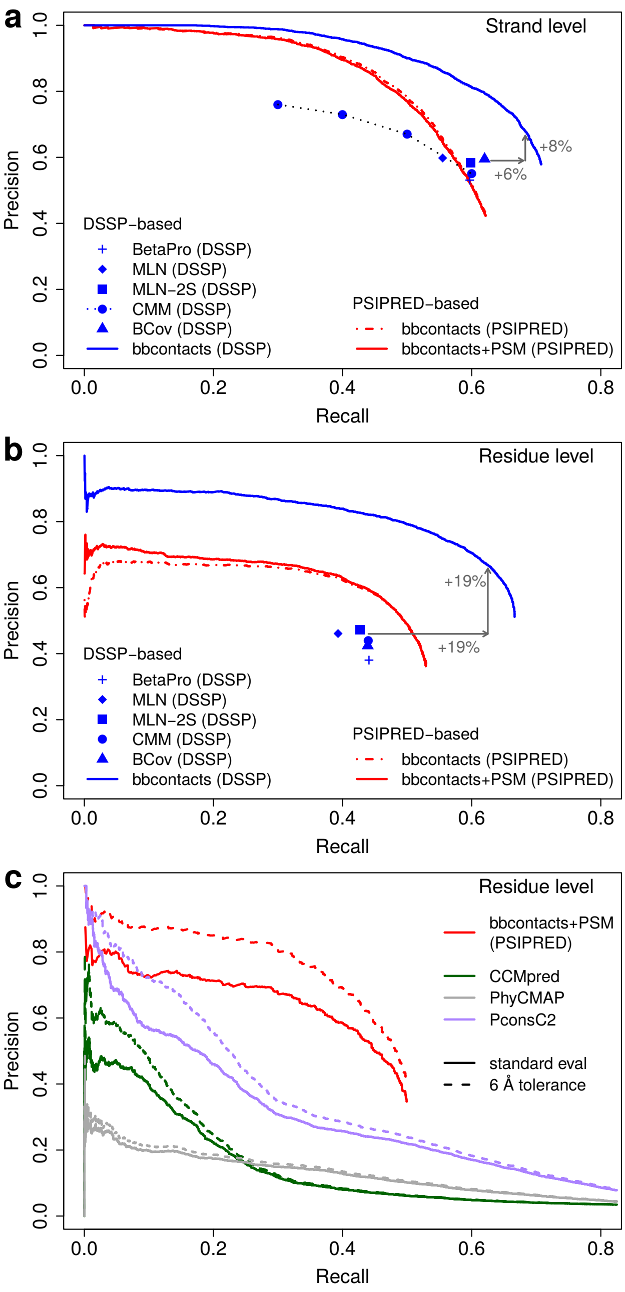

Fig. 4. (a-b) Performance of bbcontacts compared to previous β-β contact prediction methods on the BetaSheet916 dataset: (a) strand-level performance for correct β-strand pairing and (b) residue-level performance. Results for BetaPro, MLN, MLN-2S and BCov are taken from Savojardo *et al.* (2013), results for CMM from Burkoff *et al.* (2013). Note that methods in blue use known secondary structure assigned by DSSP, which is unavailable in practice. (c) Comparison of residue-level performance of bbcontacts with CCMpred, PhyCMAP and PconsC2 baselines (obtained by restricting the predictions to β-strand regions), in both the default and 6 Å tolerance evaluation frameworks. The solid lines use DSSP assignments for evaluation. The dashed lines use the 6 Å tolerance evaluation framework in which false positive residue pairs within a Cβ distance of 6 Å are ignored (i.e. excluded from the set of false positives), while true positives are unchanged.

predicted secondary structure. These illustrate the capacity of bbcontacts to remove noise from CCMpred coupling matrices.

Figures 4a, 4b and S11 clearly show that for DSSP-based predictions (blue traces, blue symbols), bbcontacts outperforms all previous methods, even though it was not specifically designed to perform well for DSSP-assigned secondary structure. This is particularly striking at the residue level. The effect is not as strong at the strand level because bbcontacts was not trained to detect isolated β-bridges, as we cannot expect secondary structure prediction methods to detect isolated β-residues. However, β-bridges are included in the performance assessment and the strand-level recall particularly suffers from this as β-bridges represent 17.6% of the β-strand pairs in the BetaSheet916 dataset.

As expected, PSIPRED-based predictions (red traces in Figures 4 and S11) are less confident than DSSP-based predictions, but they still exhibit remarkable precision and recall. In particular, residue-level PSIPRED-based predictions display better precision and recall than previous methods when they use the true secondary structure. The gap between DSSP-based and PSIPRED-based performances (i.e. between the blue and red lines in Figure 4) is explained by errors in the PSIPRED predictions, but also by the major advantage of knowing the exact DSSP strands compared to strand predictions (however accurate), since using true secondary structure we can rule out entirely any β-β contact occurring outside of the strand regions.

The precision-recall curves show that bbcontacts not only has good overall performance, but also provides (through the Viterbi score associated with each path) a measure of the confidence we can place in each prediction. The results of bbcontacts are quite robust: the precision drops very slowly with recall, up to around 50% recall for DSSP-based and 40% recall for PSIPRED-based predictions.

To investigate how much of the performance improvement obtained with bbcontacts is due to using better contact predictions as input, we introduced additional reference points BCov* and CMM*, corresponding to results obtained when couplings predicted with CCMpred are used as an input to the β-contact prediction algorithms from BCov and CMM. Better input couplings strongly improve β-β contact predictions, but this does not explain the full extent of the bbcontacts DSSP-based performance (see Supplementary Results and Supplementary Figure S13).

In Supplementary Figures S14 and S15, we analyze the performance of bbcontacts for each of the 916 test cases depending on the alignment size N and on the CCMpred precision for $L/5$ predicted contacts (where L is the length of the protein). For each test case, the F1-score is calculated for all β-β contact predictions above a Viterbi score threshold chosen to maximize the residue-level F1-score on the training dataset (Supplementary Figure S16). As expected, the trend for bbcontacts performance is to increase with N and with CCMpred precision. However, even for alignments containing a few hundred sequences, a number of cases display similar performance compared to cases with many homologs.

Note that the precision and recall reported in this work are rather conservative estimates of how useful the predicted contacts might be for structural modeling. Indeed, in compliance with previously published approaches, the evaluation adopted in this paper relies on the DSSP definition of β-β contacts, based on backbone hydrogen bonds. Because of this rigid definition, close side-chain contacts between pairs of residues which do not form backbone contacts (such as pairs of bulge residues or pairs of residues immediately adjacent to β-strands, but not assigned as β-residues by DSSP) can be detected as false positives. A few such false positives are shown in Figure 1b: several residue pairs predicted by bbcontacts (triangles) but not assigned by DSSP (i.e. not in open circles) display distances below 6 Å. Comparison of the solid and dashed red traces in Figure 4c and Supplementary Figure S17 shows that if we exclude all residue pairs within 6 Å Cβ distance from the false positives during evaluation, there is a large increase in bbcontacts precision for PSIPRED-based predictions, reaching $\sim$80% precision at 40% recall on the full BetaSheet916 dataset (Figure S17).

Finally, bbcontacts was compared with general contact prediction methods: CCMpred (Seemayer *et al.*, 2014), PhyCMAP (Wang and Xu, 2013), and PconsC2 (Skwark *et al.*, 2014). The results are shown in Figure 4c, using both the DSSP-based evaluation framework (solid lines) and the 6 Å distance tolerance (dashed lines). The latter is a middle ground for evaluation, since bbcontacts was trained to detect only β-β contacts strictly assigned by DSSP (mostly distributed around 5 Å), while the general contact predictors are designed to detect all contacts up to 8 Å. Due to the computational cost of PconsC2, this evaluation is performed on a subset of BetaSheet916 containing 186 protein chains (Supplementary Dataset S2). However, the evaluation of bbcontacts, CCMpred and PhyCMAP on the full BetaSheet916 dataset reported in Supplementary Figure S17 shows very similar trends. As expected, PhyCMAP, which does not use direct coupling analysis, displays low precision even at low recall. PconsC2 improves largely over PhyCMAP and CCMpred, but its precision drops much earlier with recall than the bbcontacts precision, even when relaxing the evaluation criteria to 6 Å tolerance.

Supplementary Figures S17-S19 show the results when using 8 Å tolerance in the evaluation: in this case, we also remove many false positives corresponding to predicted β-β contacts with a shifted register. Even so, bbcontacts improves greatly over CCMpred and PhyCMAP, and the PconsC2 precision-recall curve drops earlier than the bbcontacts curve.

4 CONCLUSION

bbcontacts is the first predictor of β-β contacts that does not require known secondary structure from DSSP and that therefore can be used in practice. Having to use predicted instead of true secondary structure makes β-β contact prediction a much more challenging problem. In particular, we can no longer rely on the knowledge of the exact β-strand positions.

Analyzing the contributions of different terms in bbcontacts shows that the choices that have a strong impact on PSIPRED-based predictions do not affect DSSP-based results to the same extent (compare for instance Figure 3 with Supplementary Figure S9). This underlines the importance of specifically designing methods to deal with predicted secondary structure.

bbcontacts also illustrates that HMMs are an attractive approach for statistical modelling of variable-length β-β contacts based on the detection of specific patterns in the coupling matrices. The use of a local Viterbi algorithm enables the detection of local patterns.

In addition, bbcontacts can pick up signal in a number of alignments with relatively few homologous sequences. Another interesting feature is that it provides a score for each predicted β-strand contact, which expresses the reliability of the prediction. Finally, bbcontacts is provided as free and open-source software with few dependencies. The runtimes on a single core of an Intel Xeon E5-2650 processor are typically under one minute for proteins up to 500 residues, and up to a few minutes when PSM gets triggered (Supplementary Figure S20). This makes bbcontacts easily applicable to a variety of situations, including predictions for large proteins and large-scale β-contact prediction. Another possible application for bbcontacts would be the case of inter-chain β-sheets, provided enough homologs are available for both chains.

bbcontacts could be further improved by including additional information in the HMM emissions, such as amino acid profiles and solvent accessibility predictions, or other input sources complementing CCMpred, for instance PSICOV (Jones *et al.*, 2012). Further planned work involves incorporating input from predictors such as CMAPpro (Di Lena *et al.*, 2012) or PhyCMAP (Wang and Xu, 2013) which do not use direct coupling analysis and perform better than CCMpred or PconsC2 when few homologous sequences are available (Skwark *et al.*, 2014). One advantage of our approach in this respect is that it can handle new sources of input in a probabilistic manner, through the addition of new terms in the emission probabilities. Also, one could improve the use of topological constraints by calculating posterior probabilities for all β-strand topologies that can be built from the list of best Viterbi paths. Finally, the approach developed in this study can be generalised to detect patterns of other interacting secondary structure elements such as helix-helix and helix-strand interactions and thus points to a promising avenue for future research.

ACKNOWLEDGEMENT

We thank Stefan Seemayer and Susann Vorberg for stimulating discussions and Stefan Seemayer and Armin Meier for feedback on the paper draft.

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (grants GRK1721 and SFB646 to JS); by the Bavarian Center for Molecular Biosystems (BioSysNet); and by the Alexander von Humboldt Foundation (Research Fellowship for Postdoctoral Researchers to JA). We thank NVIDIA Corporation for donating two Tesla K40 GPUs used in this work.

REFERENCES

Baldi,P. *et al.* (2000). Matching protein beta-sheet partners by feedforward and recurrent neural networks. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 25–36.

Biegert,A. and Söding,J. (2008). De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, **24**, 807–814.

Burkoff,N.S. *et al.* (2013). Predicting protein β-sheet contacts using a maximum entropy-based correlated mutation measure. *Bioinformatics*, **29**, 580–587.

Chan,A.W. *et al.* (1993). Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci.*, **2**, 1574–1590.

Cheng,J. and Baldi,P. (2005). Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, **21 Suppl 1**, i75–84.

Craveur,P. *et al.* (2013). β-Bulges: extensive structural analyses of β-sheets irregularities. *Protein Sci.*, **22**, 1366–1378.

Di Lena,P. *et al.* (2012). Deep architectures for protein contact map prediction. *Bioinformatics*, **28**(19), 2449–2457.

Ekeberg,M. *et al.* (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E*, **87**, 012707.

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

Holm,L. and Sander,C. (1996). Mapping the Protein Universe. *Science*, **273**, 595–602.

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

Hubbard,T. (1994). Use of beta-strand interaction pseudo-potentials in protein structure prediction and modelling. In *Proc. Twenty-Seventh Hawaii Int. Conf. Syst. Sci. HICSS-94*, volume 5, pages 336–344. IEEE Comput. Soc. Press.

Hutchinson,E.G. and Thornton,J.M. (1996). PROMOTIF–a program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220.

Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones,D.T. *et al.* (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

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

Kamat,A.P. and Lesk,A.M. (2007). Contact patterns between helices and strands of sheet define protein folding patterns. *Proteins*, **66**, 869–876.

Kamisetty,H. *et al.* (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 1–6.

Kim,D.E. *et al.* (2014). One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins*, **82 Suppl 2**, 208–218.

Klepeis,J.L. and Floudas,C.A. (2003). ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.*, **85**, 2119–2146.

Kortemme,T. (1998). Design of a 20-Amino Acid, Three-Stranded -Sheet Protein. *Science*, **281**, 253–256.

Leaver-Fay,A. *et al.* (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.*, **487**, 545–574.

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

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 fluorescent protein. *J. Biol. Chem.*, **275**, 29200–29206.

Michel,M. *et al.* (2014). PconsFold: improved contact predictions improve protein models. *Bioinformatics*, **30**, i482–i488.

Muckstein,U. *et al.* (2002). Stochastic pairwise alignments. *Bioinformatics*, **18**, S153–S160.

Nugent,T. and Jones,D.T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, E1540–E1547.

Remmert,M. *et al.* (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Richardson,J.S. *et al.* (1978). The beta bulge: a common small unit of nonrepetitive protein structure. *Proc. Natl. Acad. Sci. U. S. A.*, **75**, 2574–2578.

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

Savojardo,C. *et al.* (2013). BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. *Bioinformatics*, **29**, 3151–3157.

Seemayer,S. *et al.* (2014). CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, page 10.1093/bioinformatics/btu500.

Sillitoe,I. *et al.* (2013). New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.*, **41**, D490–D498.

Skolnick,J. *et al.* (1997). MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.*, **265**, 217–241.

Skwark,M.J. *et al.* (2014). Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS Comput. Biol.*, **10**(11), e1003889.

Smith,C.K. and Regan,L. (1995). Guidelines for Protein Design: The Energetics of beta Sheet Side Chain Interactions. *Science*, **270**, 980–982.

Tai,C.H. *et al.* (2014). Assessment of template-free modeling in CASP10 and ROLL. *Proteins*, **82 Suppl 2**, 57–83.

Wang,Z. and Xu,J. (2013). Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics*, **29**(13), i266–273.

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