Bioinformatics, 32(24), 2016, 3774–3781
doi: 10.1093/bioinformatics/btw553
Original Paper

OXFORD

Structural bioinformatics

KScons: a Bayesian approach for protein residue contact prediction using the knob-socket model of protein tertiary structure

Qiwei Li[1,*], David B. Dahl[2], Marina Vannucci[1], Hyun Joo[3] and Jerry W. Tsai[3]

[1]Department of Statistics, Rice University, Houston, TX, USA, [2]Department of Statistics, Brigham Young University, Provo, UT, USA and [3]Department of Chemistry, University of the Pacific, Stockton, CA, USA

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on March 1, 2016; revised on July 15, 2016; accepted on August 18, 2016

Abstract

Motivation: By simplifying the many-bodied complexity of residue packing into patterns of simple pairwise secondary structure interactions between a single knob residue with a three-residue socket, the knob-socket construct allows a more direct incorporation of structural information into the prediction of residue contacts. By modeling the preferences between the amino acid composition of a socket and knob, we undertake an investigation of the knob-socket construct's ability to improve the prediction of residue contacts. The statistical model considers three priors and two posterior estimations to better understand how the input data affects predictions. This produces six implementations of KScons that are tested on three sets: PSICOV, CASP10 and CASP11. We compare against the current leading contact prediction methods.

Results: The results demonstrate the usefulness as well as the limits of knob-socket based structural modeling of protein contacts. The construct is able to extract good predictions from known structural homologs, while its performance degrades when no homologs exist. Among our six implementations, KScons MST-MP (which uses the multiple structure alignment prior and marginal posterior incorporating structural homolog information) performs the best in all three prediction sets. An analysis of recall and precision finds that KScons MST-MP improves accuracy not only by improving identification of true positives, but also by decreasing the number of false positives. Over the CASP10 and CASP11 sets, KScons MST-MP performs better than the leading methods using only evolutionary coupling data, but not quite as well as the supervised learning methods of MetaPSICOV and CoinDCA-NN that incorporate a large set of structural features.

Contact: qiwei.li@rice.edu

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

1 Introduction

Due to the resource intensive nature of experimental protein structure determination, the amount of information about protein structure lags far behind the amount of known protein sequence data. Yet, a protein structure yields a deeper understanding of how a protein functions as well as how it is controlled in a biological system. Since Anfinesen's insightful experiments on protein folding (Anfinesen, 1973), it is generally accepted that all the information on a protein's structure is coded in its amino acid sequence. Therefore, automatic methods that predict protein structure from its sequence are a particular focus of computational biochemistry. As a step towards developing a protein structure prediction method, this

work describes the statistical model for the prediction of protein residue contacts using the novel knob-socket model of protein residue packing structure.

There are four distinct levels of protein structure. The sequential chemical bonded nature of a polypeptide chain of amino acids allows the primary (1°) structure of a protein to be simply represented as the linear sequence of the amino acids. At the secondary (2°) level of protein structure, the patterns of chemical interactions called hydrogen bonds determine the state assigned at that position in the sequence. The set of interactions that bring these 2° structure elements together within a polypeptide chain is called the tertiary (3°) level of protein structure. The last quaternary (4°) level of protein structure occurs between two separate polypeptide chains. In prediction of protein structure, the goal is to predict a protein's 3° structure from the amino acid 1° sequence. An important part of predicting protein 3° structure is to correctly identify the residues close in three-dimensional space that defines the topology of the folded backbone. As a result, residue contact prediction has become its own category in computational structural biology.

As discussed in the most recent Critical Assessment of Structure Prediction (CASP) review of contact prediction methods (Monastyrskyy *et al.*, 2015), accurate predictions of residue contacts have been successfully used to model soluble (Kamisetty *et al.*, 2013; Marks *et al.*, 2011) as well as transmembrane (Hopf *et al.*, 2012; Nugent and Jones, 2012) protein structures. Recently, it has been estimated that 1 long range contact is needed per every 12 residues for an accurate prediction of protein 3° structure (Kim *et al.*, 2014). A thorough review comparison of the major contact prediction methods has been performed and serves as a good benchmark (Ma *et al.*, 2015). Generally, these approaches to contact prediction are based on the concept of evolutionary coupling (EC), where residues packing near in space will most likely mutate with each other (Gobel *et al.*, 1994; Shindyalov *et al.*, 1994). This approach is most successful when there are many sequence homologs to provide a deep enough multiple sequence alignment to identify covariation between positions. To address cases with low sequence information, characteristics from protein structure usually on the order of hundreds have been added to EC. To properly integrate these large number of multiple inputs, machine learning methods like support vector machines (Cheng and Baldi, 2007; Wu and Zhang, 2008) or neural nets (Ma *et al.*, 2015; Tegge *et al.*, 2009) have been used. At the most recent 11th CASP meeting, the CONSIP2 server of the MetaPSICOV method (Kosciolek and Jones, 2015) demonstrated quantifiable progress even on targets with little to no sequence homologs (Monastyrskyy *et al.*, 2015) by implementing a supervised machine learning approach that combined several EC protocols with over 600 protein features to address the lack of sequence information.

By identifying exact residue contacts from 3° structure packing, the knob-socket model provides a simpler and more direct approach to incorporate structural data that is complementary to the aforementioned methods for predicting protein contacts. In a series of papers (Fraga *et al.*, 2015; Joo and Tsai, 2014; Joo *et al.*, 2012, 2015), a novel description of protein 3° packing has been developed that uses a new construct to predict protein contacts. The knob-socket model represents a paradigm shift in the characterization of protein structure by directly relating protein sequence and the arrangement of residues in 3° structure. As shown in Figure 1, the model decomposes the complexity of the multi-body residue interactions into discrete patterns of a four-residue packing unit called the knob-socket. Essentially, the entire 2° structure simply maps into 3-residue sockets (which include irregular coil structure) that exist in either a filled or free state to define 3° structure. A filled socket packs with a knob residue to form 3° structure and reduces

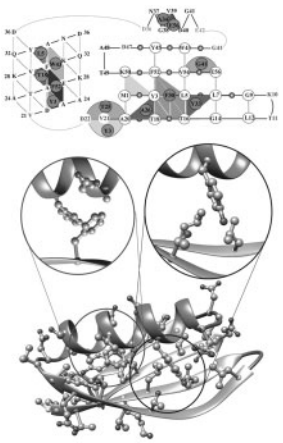

Fig. 1. The structure of Protein G 1pgb (Gallagher *et al.*, 1994) is shown with the backbone in ribbon and residues as ball and stick. For clarity, the sheet is shown in blue, the helix in orange. Because of the many-bodied contacts between residue, the 3° interactions are difficult to identify even when isolated in the 2 circles. In contrast, the knob-socket representation of 3° packing provides a clear decomposition of the packing surfaces between the helix and sheet, such that the interpretation of packing are simple pairings of knobs packing into sockets (Fraga *et al.*, 2015). To be consistent with the color theme, the knob-socket packing surface topology map shows knobs from sheets in blue, helix in orange and coil in green. As examples, the interactions in the circles reference the sheet knob F52 packing into a helix socket made up of residues A26,E27:F30 and also the helix knob Y33 packing into the sheet socket made up of residues L5,L7:T16, where the sockets are given in the standard notation (Joo *et al.*, 2015). Therefore, instead of pairwise residue contacts, the 3° residue packing contacts are classified as knob-socket interactions, which allow a direct mapping of structure to sequence patterns

the many-bodied residue packing interactions into a simple two-body mapping of knob-socket patterns. A free socket indicates those regions of the structure that are not involved in 3° structure. It has been shown that the types of amino acids that make up the socket determine whether a socket is filled or free. As a result, filled sockets provide a scoped definition for predicting residue contacts in 3°. In other words, protein contacts can be modeled based on the propensity of the filled knob-socket interactions.

In this work, we investigate the usefulness of a knob-socket based approach to protein contact prediction. One challenge in identifying the residue contacts is projecting the 1° and 2° one dimensional structures into a three dimensional arrangements of amino acids. The characterization of packing as a knob residue contacting three residues in a socket allows the direct mapping of three dimensional residue arrangements to one dimensional contact pairs. Unlike the supervised learning approaches that use hundreds of inputs to include structural information, our approach uses only the knob-socket construct. We incorporate this protein structural information defined by the knob-socket construct into a principled statistical approach. The 3° packing information provided by the knob-socket model is captured by our likelihood and the evolutionary data from structural homologs provides prior information in our Bayesian approach. We name our proposed residue contact prediction program KScons. To properly assess predictions, KScons is run against three structure sets. The first is the set of 150 structural families used to comprehensively compare a number of current contact prediction routines (Ma *et al.*, 2015) that was originally used to characterize PSICOV (Jones *et al.*, 2012). The last two are the more challenging sets of structures from CASP10 (Monastyrskyy *et al.*, 2014) and CASP11 (Kryshtafovych *et al.*, 2015). From the standard scores used in these comparisons, KScons produces an improvement

in the prediction of protein contacts and therefore demonstrates the usefulness of a structure-based modeling of protein contacts in prediction.

2 Approach

2.1 Datasets

The SCOPe: Structural Classification of Proteins—extended family Astral 2.05 dataset (Fox *et al.*, 2014) with sequence identity cutoff of 70% was used to calculate the knob-socket frequencies. After cleaning up the fragmented structures, 20 158 domains were used to collect knob-socket frequencies as described in previous work (Fraga *et al.*, 2015; Joo and Tsai, 2014; Joo *et al.*, 2012, 2015). The training data (ALN) set was created from the same Astral 2.05 dataset by including only sequences with sequence identities between 70% and 90%, and families with more than three domains were used, which left 212 families including 2008 domains. For each family, all the sequences were aligned using MUSCLE multiple sequence alignment (Edgar, 2004a,b) and the structures were aligned using MUSTANG multiple structure alignment(Konagurthu *et al.*, 2006). The knob-socket data and residue contact matrices were calculated using aligned structures following the methods used in previous studies (Fraga *et al.*, 2015; Joo and Tsai, 2014; Joo *et al.*, 2012, 2015).

As test sets, 150 PSICOV (Jones *et al.*, 2012), 124 CASP10 (Moult *et al.*, 2014) and 110 CASP11 (Kinch *et al.*, 2016) targets were used. As it is common, 21 of 131 of the CASP11 target domains have not yet been released to the public. Among 110 CASP11 target domains, no good templates were found for 22 domains and we used templates for only 88 domains. For each target sequence, homologous sequences were searched in non-redundant sequence database using Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) (Altschul *et al.*, 1997). Templates were selected from the list of RCSB protein data bank (Berman *et al.*, 2000) structures released before December 2012, excluding target proteins. Multiple sequence alignments and multiple structural alignments were performed for each target/template as for the training dataset. The knob-socket analyses were carried out and the residue contact matrices were calculated on the aligned structures.

3 Model

3.1 Socket triplets and filled knob-socket quartets

The amino acid sequence or 1° structure of a protein consisting of L amino acids is represented by a linear string as $\boldsymbol{a} = (a_1, \ldots, a_L)$, where a_i is one of the 20 naturally occurring amino acids. Each position in the 1° sequence can be assigned a 2° structure state $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_L)$. Kabsch and Sander (1983) proposed the Dictionary of Protein Secondary Structure (DSSP) for protein 2° structure with single letter codes. The original eight structure types (in parentheses) are commonly collapsed into the following four types (in italics):

- Helix '*H*': 310 helices (G), α-helices (H), or π-helices (I);
- Strand '*E*': extended strands in parallel or anti-parallel β-sheets (E);
- Turn '*T*': hydrogen bonded turns of length 3 or more amino acids (T);
- Coil '*C*': β-bridge residues (B), bends (S), or random coils (C).

These four types of 2° structure appear in contiguous stretches of the same type giving the general form of local segments and referred to as 'block types.'

The 3° structure of a protein is its geometric shape and is defined by the multi-body packing of residues distant in sequence but close in three-dimensional space. As introduced in Figure 1, the knob-socket model of protein packing structure provides a simple construct to characterize protein 3° structure and decomposes the complexity of many-body residue interactions into a simpler two-body problem between elements of 2° structure. For a knob-socket pair, the socket made up of 3 residues on one element of 2° structure presents a surface that favors packing a knob residue from another 2° structure. Therefore, the socket represents short range protein contacts, while the knob-socket interactions are a representation of long-range protein contacts. Contact prediction then reduces down to identifying which amino acids form these 3-residue sockets. In this paper, the 3-residue sockets are modeled based the propensity to be filled with a knob or free, given the 2° structure type and the amino acid sequence (Joo and Tsai, 2014; Joo *et al.*, 2012, 2015). Notationally, we write a three-residue socket as $(u, v, w)|(\rho_u, \rho_v, \rho_w)$, where u, v, w are positions such that $1 \leq u < v < w \leq L$ and ρ_u, ρ_v, ρ_w are the secondary structure types of the corresponding positions. Each socket may be filled with a knob (i.e. a residue whose position is far in sequence from those forming the socket) or 'free' of the 'knob' (i.e. when the socket is not involved in 3° structure). If all of the free sockets and filled knob-sockets are known, then these can be used as simple constraints to rebuild the entire protein fold or 3° structure (i.e. its geometric shape).

The inferential goal of the statistical model presented in this paper is as follows. Given the 1° structure $\boldsymbol{a}$ and the 2° structure $\boldsymbol{\rho}$ of the target protein with length L, we seek to provide a list of all the three-residue free sockets and $(1+3)-$filled knob-sockets of the protein. Because of sampling variability, our list of predicted sockets may exclude some true sockets and include sockets that are not actually present in the protein.

All of the sockets in the 2° and 3° structure of a protein are a small fraction of all possible combinations of $(u, v, w)|(\rho_u, \rho_v, \rho_w)$ that could be formed from the integers $1, \ldots, L$ and the associated 2° structure types. We let a *triplet* be one possible combination of $(u, v, w)|(\rho_u, \rho_v, \rho_w)$ that may or may not form a socket in the 3° structure. For a target protein which has L residues in total, the total number of such triplets, denoted by K, is $C(L, 3) = L(L-1)(L-2)/6$. Only a small fraction of those K triplets actually form sockets. We postulate the existence of a latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$, with $\gamma_k = 1$ if combination k, denoted $(u_k, v_k, w_k)|(\rho_{u_k}, \rho_{v_k}, \rho_{w_k})$, is actually a socket and $\gamma_k = 0$ otherwise.

Searching through all possible K combinations is computationally prohibitive. As such, we reduce the space to those triplets of amino acids that could possibly form sockets based on known propensities calculated in the characterization of the knob-socket model (Joo and Tsai, 2014; Joo *et al.*, 2012, 2015). As discussed in those analyses, the entire 2° structure forms sockets, and the composition of sockets from different 2° structure exhibits preferences for certain amino acids. Furthermore, the amino acid preferences also identify if the socket tends to be free or filled with a knob residue. First, we divide all the triplets into two groups according to their ranges, i.e. $w_k - u_k$. If $w_k - u_k > 5$, those triplets are non-local and only the triplets with secondary structure $(\rho_u, \rho_v, \rho_w) = (\mathrm{E}, \mathrm{E}, \mathrm{E})$ and the following four structures could form sockets: $(u_k, v_k = u_k + 1, w_k)$, $(u_k, v_k = u_k + 2, w_k)$, $(u_k, v_k = w_k - 1, w_k)$ and $(u_k, v_k = w_k - 2, w_k)$. We can also denote the above four structures by their gaps 1_x, 2_x, x_1 and x_2, respectively, where the first number indicates the gap between position u_k and v_k, the second number indicates the gap between position v_k and w_k, and x means any other possible value. On the other hand, if $w_k - u_k \leq 5$,

those triplets are local and there are 10 possible structures: $(u_k, v_k = u_k + 1,\ w_k = v_k + 1)$, $(u_k, v_k = u_k + 1, w_k = v_k + 2)$, $(u_k, v_k = u_k + 1, w_k = v_k + 3)$, $(u_k, v_k = u_k + 1, w_k = v_k + 4)$, $(u_k, v_k = u_k + 2, w_k = v_k + 1)$, $(u_k, v_k = u_k + 2,\ \ w_k = v_k + 2)$, $(u_k, v_k = u_k + 2, w_k = v_k + 3)$, $(u_k, v_k = u_k + 3, w_k = v_k + 1)$, $(u_k, v_k = u_k + 3, w_k = v_k + 2)$, and $(u_k, v_k = u_k + 4, w_k = v_k + 1)$, denoted by their gap notations 1_1, 1_2, 1_3, 1_4, 2_1, 2_2, 2_3, 3_1, 3_2 and 4_1, respectively. Finally, some triplets must be form sockets and we therefore do not need to predict them. For example, it is well known that the triplets $(u_k, v_k = u_k + 1, w_k = u_k + 4)|(\mathrm{H,H,H})$ and $(u_k, v_k = u_k + 3, w_k = u_k + 4)|(\mathrm{H,H,H})$ always form sockets. Table 1 summarizes those triplets that must be, might be, and must not be true sockets based on the biochemistry.

We introduce another binary latent vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)$ to indicate whether a triplet is a filled socket. We let $\delta_k = 1$ if and only if the triplet k forms a socket (i.e. $\gamma_k = 1$) and it is filled with a knob. We let $z_k \in \{1, \ldots, L\}$ be the position of the knob in socket k when $\delta_k = 1$ and let $z_k = 0$ otherwise.

3.2 Sampling model

Given the secondary structure $\boldsymbol{\rho}$ and the constraints listed in Table 1, we can enumerate all the triplets, indicated by their positions $\mathbf{Tri} = \{(u_1, v_1, w_1), \ldots, (u_K, v_K, w_K)\}$. We then denote the corresponding set of amino acids and secondary structure types as $\boldsymbol{a}(\mathbf{Tri}) = \{(a_{u_1}, a_{v_1}, a_{w_1}), \ldots, (a_{u_K}, a_{v_K}, a_{w_K})\}$ and $\boldsymbol{\rho}(\mathbf{Tri}) = \{(\rho_{u_1}, \rho_{v_1}, \rho_{w_1}), \ldots, (\rho_{u_K}, \rho_{v_K}, \rho_{w_K})\}$, respectively. Our aim is to make inference on $\boldsymbol{\gamma}$ (i.e. which triplets are sockets), $\boldsymbol{\delta}$ (i.e. which sockets are filled) and $\boldsymbol{z}$ (i.e. where are the knobs filling the sockets), based on the primary structure $\boldsymbol{a}$, secondary structure $\boldsymbol{\rho}$, and the valid triplets list $\mathbf{Tri}$. We take a Bayesian approach and sample from a posterior distribution in the following form:

$$\begin{aligned} &p(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z} | \boldsymbol{a}(\mathbf{Tri}), \boldsymbol{\rho}(\mathbf{Tri})) \\ &\quad \propto \mathrm{p}(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z}) \mathrm{p}(\boldsymbol{a}(\mathbf{Tri}), \boldsymbol{\rho}(\mathbf{Tri}) | \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z}) \\ &\quad = \mathrm{p}(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z}) \textstyle\prod_{\mathrm{k}} \mathrm{p}(\mathrm{a}_{\mathrm{u_k}}, \mathrm{a}_{\mathrm{v_k}}, \mathrm{a}_{\mathrm{w_k}}, \mathrm{a}_{\mathrm{z_k}}, \rho_{\mathrm{u_k}}, \rho_{\mathrm{v_k}}, \rho_{\mathrm{w_k}} | \gamma_{\mathrm{k}}, \delta_{\mathrm{k}}), \end{aligned} \tag{1}$$

with the assumption that knobs are independent of sockets. We also assume that sockets are independent of each others. These independence assumptions are taken for mathematical tractability with the hope that the model performs well in practice even in the presence of dependence. We take the sampling model in (1) to be

$$\begin{aligned} &p(a_{u_k}, a_{v_k}, a_{w_k}, \rho_{u_k}, \rho_{v_k}, \rho_{w_k} | \gamma_k, \delta_k) \\ &\quad = p_{\rho_{u_k}, \rho_{v_k}, \rho_{w_k}}(a_{u_k}, a_{v_k}, a_{w_k} | \gamma_k, \delta_k). \end{aligned} \tag{2}$$

According to Table 1, there are 27 different types of predictable local triplets and only one predictable non-local triplet (EEE), each having a sampling model component like (2). Each of them has 3 different conditions, i.e. $(\gamma_k = 1, \delta_k = 1)$, $(\gamma_k = 1, \delta_k = 0)$ and $(\gamma_k = 0, \delta_k = 0)$, corresponding to filled sockets, free sockets and non-sockets. The 28 sampling models can be evaluated by using Dirichlet-multinomial distributions based on the known true free and filled knob-sockets, and false sockets from training dataset. In this model, conditional upon (ρ_u, ρ_v, ρ_w) and (γ, δ), the vector of (a_u, a_v, a_w) is of length $20^3 = 8000$, with all zeros except a single 1, and $p(a_u, a_v, a_w | \theta)$ is a multinomial distribution with one trial. Given the observed counts $\boldsymbol{X} = (X_1, \ldots, X_{8000})$ for each possible combination of (a_u, a_v, a_w) in the training dataset, we assume the following Bayesian model: $\boldsymbol{X}|\theta \sim \mathrm{Multinomial}(n, \theta)$ and $\theta \sim \mathrm{Dirichlet}(1, \ldots, 1)$, where $n = \sum_{k=1}^{8000} X_k$. Due to conjugacy, the posterior distribution is $\theta|\boldsymbol{X} \sim \mathrm{Dirichlet}(X_1 + 1, \ldots, X_{8000} + 1)$. Integrating θ out from the product of $p(a_u, a_v, a_w|\theta)$ and $p(\theta|\boldsymbol{X})$ results in a Dirichlet-multinomial distribution. As the number of trials is simply 1, evaluating $p(a_u, a_v, a_w|\boldsymbol{X})$ requires only that we add one to the number of times the combination (a_u, a_v, a_w) is present in the condition of interest in the training dataset, and then divide this number by $n + 8,000$.

3.3 Priors

3.3.1 Independent Bernoulli prior

The model is completed by specifying the prior distribution $p(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z})$. One possible choice is the binomial prior, $p(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z}) \propto \prod_{\mathrm{k=1}}^{\mathrm{K}} \omega^{\gamma_k} (1 - \omega)^{1 - \gamma_k}$. This arises by considering $\boldsymbol{\gamma}$ and $(\boldsymbol{\delta}, \boldsymbol{z})$ to be *a priori* independent and $p(\boldsymbol{\delta}, \boldsymbol{z}) \propto 1$ to be a flat and non-informative prior. This is equivalent to assuming independent Bernoulli distributions on each individual $\gamma_k \sim \mathrm{Bern}(\omega)$. We recommend to choose $\omega = 0.05$, based on the counts obtained from the training dataset. The drawback of choosing this independent Bernoulli prior is that we are unable to predict the knob.

3.3.2 Multiple structure/sequence alignment prior

We also consider an informative prior distribution which incorporates information from the sequence alignment of homologous structures, i.e. those sequences in the same structural family as the target sequence. Suppose we know the multiple structure alignment (MST) with length $L' \geq L$ between the target sequence and those N homologous sequences. Also, we have known the L'-by-L' 3° residue contact matrices $\boldsymbol{C}^{\mathrm{MST}(1)}, \ldots, \boldsymbol{C}^{\mathrm{MST}(N)}$ of those homologous structures as defined by the knob-socket model. For the symmetric contact matrix $\mathbf{C}$, we define $\mathbf{C}_{ij} = 1$ if residue i and j belong to the same socket and $\mathbf{C}_{ij} = 0$ otherwise.

Given any proposed $(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z})$ of the target protein and the MST with its homologous sequences, we generate a L'-by-L' contact matrix $\mathbf{C}$ from a Beta-binomial model as follows: $\sum_{n=1}^{N} \mathrm{C}_{ij}^{\mathrm{MST}(n)} | \psi_{ij}$

Table 1. Summary of triplets that must be ('✓'), might be ('?') or must not be ('×') sockets by secondary structure and gap patterns. We reduce the sample space for γ by only considering those triplets that might be sockets

	Local gap patterns										non-local gap patterns			
Structure	1_1	1_2	1_3	1_4	2_1	2_2	2_3	3_1	3_2	4_1	1_x	2_x	x_1	x_2
HHH	×	×	✓	×	×	×	×	✓	×	×	×	×	×	×
EEE	?	?	?	?	?	?	?	?	?	?	?	?	?	?
CTC	×	×	?	?	×	?	?	?	?	?	×	×	×	×
TCE	?	×	×	×	×	×	×	×	×	×	×	×	×	×
*Set	?	?	?	?	?	?	?	?	?	?	×	×	×	×
Others	×	×	×	×	×	×	×	×	×	×	×	×	×	×

*Set = {CCC, TTT, HHC, HCC, CCH, CHH, HHT, HTT, TTH, THH, EEC, ECC, CCE, CEE, EET, ETT, TTE, TEE, CCT, CTT, TTC, TCC, ECT}.

$\sim$ Binomial(N, ψ_{ij}) and $\psi_{ij} \sim$ Beta(α, β), where α and β can be set to 1 for convenience. Therefore we have conjugate posterior $\psi_{ij}|\cdot \sim$ Beta$(\alpha + \sum_{n=1}^{N} C_{ij}^{\mathrm{MST}(n)}, \beta + N - \sum_{n=1}^{N} C_{ij}^{\mathrm{MST}(n)})$. Then we assume that the contact matrix follows a product of $L'(L'-1)/2$ p.m.f.'s, i.e. $p(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z}) = \mathrm{p}(\boldsymbol{C}) = \prod_{\mathrm{i}=1}^{L'-1} \prod_{\mathrm{j}=\mathrm{i}+1}^{L'} \psi_{\mathrm{ij}}^{C_{\mathrm{ij}}} (1-\psi_{\mathrm{ij}})^{(1-C_{\mathrm{ij}})}$. The more similar $\boldsymbol{C}$ is to the average of $\boldsymbol{C}^{\mathrm{MST}(1)}, \ldots, \boldsymbol{C}^{\mathrm{MST}(N)}$, the higher the prior probability will be.

Analogously, using the multiple sequence alignment (MSQ) information $\boldsymbol{C}^{\mathrm{MSQ}(1)}, \ldots, \boldsymbol{C}^{\mathrm{MSQ}(N)}$, we can estimate another set of hyperparameters ψ_{ij} to reshape the prior distribution.

4 Model fitting

4.1 MCMC algorithm

We use Markov chain Monte Carlo (MCMC) methods to sample from the posterior distribution $p(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z}|\boldsymbol{a}, \boldsymbol{\rho})$. We update the three parameters using a Metropolis algorithm. We note that this algorithm is sufficient to guarantee ergodicity for our model. See the details in the supplement.

4.2 Posterior estimation

The goal is to infer the socket list $(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z})$. However, for a proper assessment of performance, we map predicted sockets and knobs into the corresponding contact matrix $\boldsymbol{C}$. We considered two ways to summarize the posterior distribution to yield point estimators: (1) Choosing a particular $\boldsymbol{C}$ that maximizes the posterior probability $p(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z}|\boldsymbol{a}, \boldsymbol{\rho})$, i.e. $\widehat{\boldsymbol{C}}^{\mathrm{MAP}} = (\widehat{\boldsymbol{\gamma}}^{\mathrm{MAP}}, \widehat{\boldsymbol{\delta}}^{\mathrm{MAP}}, \widehat{\boldsymbol{z}}^{\mathrm{MAP}}) = \mathrm{argmax}_{\gamma} p(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{z}|\boldsymbol{a}, \boldsymbol{\rho})$ and (2) Selecting the most likely C_{ij} from its marginal probability, which can be calculated by $\sum_{b=1}^{B} (C_{ij}^{(b)}|\boldsymbol{a}, \boldsymbol{\rho})/\mathrm{B}$, where B is the number of iterations after burn-in. A point estimate of $\boldsymbol{C}$ is then obtained by identifying those marginal probabilities that exceed a given threshold t, i.e. $\widehat{C}_{ij}^{\mathrm{MP}} = \mathrm{I}\left(\sum_{b=1}^{B} (C_{ij}^{(b)}|\boldsymbol{a}, \frac{\boldsymbol{\rho}}{B} \geq t\right)$. We recommend $t = 0.5$, resulting in the median model. We name these 2 posterior estimators as the maximum *a posteriori* method (MAP) and the marginal probability method (MP), respectively.

5 Results and discussion

5.1 Results

We first trained the KScons sampling model, i.e. Equation (2), with the ALN dataset. The training set does not contain any target from any test sets, and so is properly jack-knifed with respect to the test datasets. The statistics of these four datasets are summarized in the supplement. Then, we evaluated the performance of our method with different prior distributions and posterior estimates against the 150 targets in the PSICOV test set, the 124 targets in the CASP10 test set, and the 110 available targets in the CASP11 test set. Using common accuracy metrics, KScon's pairwise contact prediction was compared consistently based on test set with leading methods of CoinDCA (Ma *et al.*, 2015) and PSICOV (Jones *et al.*, 2012), including the new implementations MetaPSICOV (Jones *et al.*, 2015) and CONSIP2 (Kosciolek and Jones, 2015).

For KScons with an independent Bernoulli prior, we set the hyperparameter $\omega = 0.05$. When using the MST or MSQ priors, we set the hyperparameters $\alpha = \beta = 1$, which leads to a non-informative hyperprior. Comparisons of different variants and

Table 2. Means and standard deviations (in parenthesis) of contact prediction accuracy achieved by PSICOV, CoinDCA and our method (KScons) under different prior distributions and posterior estimates in the PSICOV set

	Short range			Medium range			Long range		
	L/10	*L*/5	*L*/2	*L*/10	*L*/5	*L*/2	*L*/10	*L*/5	*L*/2
PSICOV	0.369	0.299	0.205	0.375	0.312	0.213	0.446	0.400	0.311
CoinDCA	0.528	0.446	0.316	0.496	0.435	0.312	0.561	0.502	0.391
KScons w/	0.195	0.136	0.090	0.171	0.131	0.086	0.163	0.145	0.118
Bern-MAP	(0.19)	(0.13)	(0.08)	(0.19)	(0.13)	(0.08)	(0.19)	(0.14)	(0.09)
KScons w/	0.335	0.281	0.198	0.249	0.218	0.162	0.203	0.186	0.162
Bern-MP	(0.24)	(0.20)	(0.15)	(0.22)	(0.17)	(0.12)	(0.18)	(0.14)	(0.11)
KScons w/	0.533	0.485	0.326	0.497	0.460	0.337	0.485	0.477	0.440
MSQ-MAP	(0.26)	(0.23)	(0.19)	(0.26)	(0.23)	(0.19)	(0.28)	(0.25)	(0.22)
KScons w/	0.792	0.701	0.478	0.731	0.676	0.485	0.764	0.747	0.660
MSQ-MP	(0.18)	(0.20)	(0.16)	(0.25)	(0.22)	(0.18)	(0.23)	(0.20)	(0.21)
KScons w/	0.581	0.516	0.336	0.537	0.490	0.354	0.532	0.520	0.470
MST-MAP	(0.24)	(0.22)	(0.19)	(0.27)	(0.24)	(0.20)	(0.29)	(0.25)	(0.22)
KScons w/	**0.823**	**0.717**	**0.493**	**0.782**	**0.703**	0.499	0.787	**0.767**	**0.686**
MST-MP	(0.17)	(0.18)	(0.16)	(0.21)	(0.21)	(0.18)	(0.23)	(0.21)	(0.21)

The numbers which achieve the highest value are shown in boldface.

Table 3. Means of contact prediction accuracy achieved by our method with MST prior and MP posterior estimation, PSICOV, CoinDCA, MetaPSICOV and CoinDCA-NN in the CASP10 set

	Short range			Medium range			Long range		
	L/10	*L*/5	*L*/2	*L*/10	*L*/5	*L*/2	*L*/10	*L*/5	*L*/2
KScons	0.536	0.481	0.341	0.467	0.434	0.341	0.414	0.394	0.352
PSICOV	0.234	0.191	0.140	0.310	0.259	0.192	0.276	0.225	0.168
CoinDCA	0.517	0.435	0.311	0.500	0.440	0.340	0.412	0.351	0.279
MetaPSICOV				0.700	0.615	0.458	0.637	0.592	0.488
CoinDCA-NN				0.725	0.640	0.471	0.665	0.615	0.509

Table 4. Means of contact prediction accuracy achieved by our method with MST prior and MP posterior estimation, PSICOV, CoinDCA, MetaPSICOV, CoinDCA-NN and CONSIP2 in the CASP11 set

	Short range			Medium range			Long range		
	$L/10$	$L/5$	$L/2$	$L/10$	$L/5$	$L/2$	$L/10$	btw553	$L/2$
KScons	0.462	0.425	0.305	0.395	0.350	0.268	0.350	0.319	0.275
PSICOV	0.190	0.144	0.112	0.196	0.163	0.115	0.198	0.172	0.127
CoinDCA	0.452	0.391	0.286	0.430	0.365	0.254	0.279	0.240	0.186
MetaPSICOV				0.680	0.582	0.419	0.555	0.492	0.407
CoinDCA-NN				0.685	0.584	0.423	0.585	0.526	0.432
CONSIP2	0.598	0.533		0.548	0.458		0.313	0.282	

estimation methods are summarized in Tables 2–4 with complete tables in the supplement. The prediction accuracy is defined as the percentage of native contacts among the top predicted $L/10$, $L/5$ and $L/2$ predicted contacts, where L is the sequence length. Contacts are short-, medium- and long-range when the sequence distance between the two residues in a contact falls into three intervals (6, 12), (12, 24) and (24, L), respectively.

In Table 2, only the performance of all 6 different combinations of the 3 prior distributions and 2 posterior estimates on the PSICOV test set is compared to identify the best implementation of KScons. Because the Bernoulli prior was limited to only modeling sockets, no information about long range knob residue contacts are modeled. As a result, this prior performed the poorest at predicting contacts for long range, but also was not much more accurate for medium and short range contacts. The multiple structure/sequence alignment (MST/MSQ) priors that incorporate evolutionary information improved the prediction accuracy. However, these priors need to be paired with an appropriate posterior estimate to take advantage of the evolutionary relationships and produce significantly improved accuracy. Of the 2 posterior estimates, the maximum *a posteriori* method (MAP) estimates did the worst with both priors. This approach favored the socket triplets and knob-socket quartets with the highest probabilities in the ALN dataset at the cost of identifying the unique and correct contacts for each protein. In contrast, using the marginal probability (MP) that selected pairs based on homology over a threshold included more correct contacts. Of the 3 priors, the Bernoulli exhibited the worst improvement with the MP posterior estimate. Between the MSQ and MST priors, the MST using a structural alignment prior performed just about 10% better than the MSQ using sequence alone. This similarity in the performance of alignment priors is expected as both are one-dimensional decompositions of the three-dimensional structure. Overall, the best results for KScons were obtained using the MST structural evolutionary prior with the MP posterior estimate, although the MSQ-MP prior performs reasonably well. Compared to the EC methods of PSICOV and COINDCA, KScons MST-MP exhibits marked improvement on this dataset. In each of the categories from the PSICOV test set (Table 2), the KScons MST-MP predictions improve prediction accuracy by 0.3 or 30% over EC methods like PSICOV (Jones *et al.*, 2012) and CoinDCA (Ma *et al.*, 2015) that leverage significantly more sequence data than KScons. Generally, the performance trends shown in Table 2 were consistent across each test set (see supplement). For these reasons, further comparisons to the CASP10 and CASP11 test sets in Tables 3 and 4, respectively, are from the MST-MP KScons models, although complete tables are shown in the supplement.

Regarding contact predictions in the more challenging CASP10 set (Table 3), KScons MST-MP is again comparable to the best EC approaches of PSICOV and CoinDCA. KScons performs markedly better than PSICOV. In this instance however, KScons MST-MP contact predictions are just 7% on average better than CoinDCA, which is within the equivalence of the standard deviation. The supervised learning methods of MetaPSICOV and CoinDCA-NN are both about 30% better than the KScons MST-MP. For the most recent CASP11, KScons MST-MP contact prediction accuracies are lower by 6%, but show similar improvement over the strictly EC methods of PSICOV and CoinDCA. Again, the supervised learning MetaPSICOV and CoinDCA-NN programs are about 45% better. The CONSIP2 server using MetaPSICOV (Kosciolek and Jones, 2015) had the strongest showing in CASP11 for contact prediction (Monastyrskyy *et al.*, 2015). CONSIP2 is an implementation of MetaPSICOV, a supervised machine learning approach combining a number of EC methods. Only the L/10 and L/5 values were reported, so our discussion is limited to these prediction accuracies. Also, KScons was tested on all 89 structures, while the CONSIP2 server results are only for 36. KScons MST-MP generally did not perform as well as the CONSIP2 server results, especially for short and medium range contacts. Additionally, the CASP11 set contained many targets that did not have many homologs (Monastyrskyy *et al.*, 2015). The present implementation of KScons MST-MP needs homologs to make accurate predictions. In general, current supervised learning methods are more accurate than our implementation of the knob-socket approach due to this requirement for structural homologs. Even so, results from all three datasets are impressive considering that the KScons approach is based primarily on modeling structural data using the simple knob-socket construct. The supervised learning methods are more complex in there using hundreds of inputs and furthermore, do not provide insight on their successes or failures. On the other hand, the knob-socket construct allows us to investigate the performance of this approach.

To better understand the current strengths and limitations of the knob-socket implementation in KScons, the classification precision and recall for γ were calculated. Complete tables for all 6 variations are given in the supplement, while Table 5 shows the values over the 3 tests sets for the MST-MAP and MST-MP implementations of KScons. Precision is defined as the percentage of actual sockets that are correctly estimated over all predicted sockets. Consistent with previous methods, the precision results showed that many false positives are predicted, which has been a recognized problem in contact prediction (Monastyrskyy *et al.*, 2015). The improvement of the MP over the MAP was clear in the precision and can be seen especially for the CASP11 set, where the MP identified more true contacts than the MAP. Recall is defined as the percentage of sockets that are correctly estimated over all actual sockets. In comparison, the MSQ and MST priors improved the recall of true contacts, but more false positives were found. Comparing the posterior estimates in Table 5, the MAP predicted fewer true contacts and more contacts that were false than the MP. These results indicate the need for homologs for accurate predictions in this implementation of KScons. Because the

Table 5. Percent socket prediction accuracies for CASP10 and CASP11 target datasets for helix, sheet and coil secondary structures (SS)

	CASP10			CASP11		
	TP	FP	FN	TP	FP	FN
Helix	60	40	1	56	44	0
Sheet	20	26	54	20	22	58
Coil	19	21	60	18	19	63
All	34	28	37	34	30	36

The true positives (TP) columns give the percents of sockets predicted correctly as free or filled, false positives (FP) columns are percents of filled sockets predicted as free or free sockets predicted as filled, and false negatives (FN) columns are percents of sockets not predicted.

MAP predicts contacts only using the isolated preference of knobs with sockets without context, the predictions produce many false positives and fewer true negatives. Therefore, certain knob-socket combinations that have high frequency in the ALN library of structures will always be predicted. The MP uses the homolog and structural topology to provide context and attenuate the highly preferred knob-sockets of the MAP to make more correct predictions.

More specifically, an analysis of contact prediction accuracy based on a three state description of secondary structure between CASP10 and CASP11 sets further confirms the strengths and limitations of KScons MST-MP. Table 5 shows the accuracy of prediction using a 3 state description of secondary structure. The improved performance of KScons MST-MP in the CASP10 dataset over the CASP11 dataset involves not just identifying fewer correct contacts, but also more false positives and false negatives. For KScons MST-MP, the strength is that the knob-sockets are able to pull significant information from the packing of the structural homologs. For both datasets, the trend is that helices are the most accurately predicted contacts above 59%. With a drop to about 20% for true positives, a significant fall off is seen in accuracy of predicting sheets. Coil is just a little worse with just over 18% true positives. The lack of homologs in CASP11 compared to CASP10 shows that helical predictions suffer from more false positives, while the sheets and coil have more false negatives. Helices are well represented in the training ALN set with a local socket and non-local knob-socket. For helices, the decrease in accuracy between CASP10 and CASP11 is mainly due to a 4% increase in the false positives. A closer look at the more complete analysis in the supplement shows the helical contact predictions in the CASP11 dataset make many more false positives. Therefore, KScon's difficulty with helical contact prediction is inclusion of more incorrect long-range contacts without homologs or poor knob-socket predictions. Sheet sockets and knob-sockets are both non-local and problematic for the knob-socket construct. For sheets, the decrease in accuracy is due to a 4% increase in false negatives. Looking at the more detailed breakdown, this increase is primarily due to more false negatives from long-range contacts in the CASP11 dataset. The coil contacts are local, but random sockets, while knob-sockets are long-range. Like sheet predictions, the coil contacts are difficult to model with very low accuracies. The 2% increase in false positives seen in coil predictions also mirrors the difficulties with long-range interactions.

6 Conclusion

As a complement to current EC and supervised learning approaches to residue contact prediction, the knob-socket construct allows the direct statistical modeling of structure into sequence space (Fig. 1). Therefore, the purpose of this work has been to develop a Bayesian statistical model using knob-socket information that maximizes contact prediction accuracy from a combination of priors and posteriors. The resulting program was then compared over three different and difficult test sets to gauge the overall performance on contact predictions against current leading methods. This article is a natural extension of our preliminary work (Li *et al.*, 2014), where we proposed a Bayesian model to predict 2° structure from 1° structure based on the knob-socket model of protein packing in 2° structure.

The combinations and comparisons revealed that a simple implementation of the knob-socket preferences requires known homologs to guide correct predictions. Testing 3 priors and 2 posterior estimates revealed that the multiple structure alignment prior and a marginal probability posterior (MST-MP) combination maximized KScons prediction accuracy. Without the MP, the knob-socket preferences from the ALN training set are too dominant and lead to decreased contact prediction accuracy because of the inclusion of more false positives and false negatives. In other words, KScons MST-MP improved the prediction true positives and also true negatives. While KScons MST-MP demonstrated greater contact prediction accuracy in comparison to EC methods over the PSICOV dataset, the more challenging CASP10 and CASP11 revealed that KScons MST-MP performs better than EC methods without the need for deep sequence alignments, but also has the same need for homologs. For proteins with no homologs, the supervised learning methods are the most accurate. Even so, this initial implementation of the knob-socket model performs well in extracting correct contact predictions using a simple and intuitive construct, as opposed to the complicated amalgamation of hundreds of inputs needed in the supervised learning approaches.

Because the knob-socket construct maps a direct correspondence between sequence and structure space, the strengths and weaknesses of the current implementation in contact prediction can be identified. This analysis provides a clear path to method improvement. Further work will be to understand how to improve the MAP that uses only the preferences from a protein structure training set. In particular, the goal is to identify how to decrease the false positives for helical contacts and the false negatives for sheet and coil contacts. Our objective is to provide a structural explanation based on knob-socket patterns for prediction of residue contacts in the cases of no sequence data.

Acknowledgements

The authors thank Keith Fraga for helpful discussions of the manuscript.

Funding

This work was supported by a NIH-NIGMS grant R01-GM104972.

Conflict of Interest: none declared.

References

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

Anfinesen,C. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

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

Cheng,J. and Baldi,P. (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.

Edgar,R.C. (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

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

Fox,N.K. *et al.* (2014) SCOPE: structural classification of proteins extended, integrating scop and astral data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.

Fraga,K.J. *et al.* (2015) An amino acid code to define a protein's tertiary packing surface. *Proteins Struct. Funct. Bioinf.*, **84**, 201–216.

Gallagher,T. *et al.* (1994) Two crystal structures of the b1 immunoglobulin-binding domain of streptococcal protein g and comparison with NMR. *Biochemistry*, **33**, 4721–4729.

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

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

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.

Jones,D.T. *et al.* (2015) METAPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, **31**, 999–1006.

Joo,H. and Tsai,J. (2014) An amino acid code for β-sheet packing structure. *Proteins Struct. Funct. Bioinf.*, **82**, 2128–2140.

Joo,H. *et al.* (2012) An amino acid packing code for α-helical structure and protein design. *J. Mol. Biol.*, **419**, 234–254.

Joo,H. *et al.* (2015) An amino acid code for irregular and mixed protein packing. *Proteins Struct. Funct. Bioinf.*, **83**, 2147–2161.

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

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**, 15674–15679.

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

Kinch,L.N. *et al.* (2016) Casp 11 target classification. *Proteins Struct. Funct. Bioinf.*, doi: 10.1002/prot.24982.

Konagurthu,A.S. *et al.* (2006) Mustang: a multiple structural alignment algorithm. *Proteins Struct. Funct. Bioinf.*, **64**, 559–574.

Kosciolek,T. and Jones,D.T. (2015) Accurate contact predictions using covariation techniques and machine learning. *Proteins Struct. Funct. Bioinf.*, doi: 10.1002/prot.24863.

Kryshtafovych,A. *et al.* (2015) Some of the most interesting casp11 targets through the eyes of their authors. *Proteins.*, doi: 10.1002/prot.24942.

Li,Q. *et al.* (2014) Bayesian model of protein primary sequence for secondary structure prediction. *PLoS One*, **9**, e109832.

Ma,J. *et al.* (2015) Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*, **31**, 3506–3513.

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

Monastyrskyy,B. *et al.* (2014) Evaluation of residue–residue contact prediction in casp10. *Proteins Struct. Funct. Bioinf.*, **82**, 138–153.

Monastyrskyy,B. *et al.* (2015) New encouraging developments in contact prediction: assessment of the casp11 results. *Proteins Struct. Funct. Bioinf.*

Moult,J. *et al.* (2014) Critical assessment of methods of protein structure prediction (casp)round x. *Proteins Struct. Funct. Bioinf.*, **82**, 1–6.

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.

Shindyalov,I. *et al.* (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.*, **7**, 349–358.

Tegge,A.N. *et al.* (2009) Nncon: improved protein contact map prediction using 2d-recursive neural networks. *Nucleic Acids Res.*, **37**, W515–W518.

Wu,S. and Zhang,Y. (2008) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, **24**, 924–931.