Modeling and inference of sequence-structure specificity

Hetunandan Kamisetty* Bornika Ghosh† Chris Bailey-Kellogg†§ Christopher James Langmead*‡§

Abstract: In order to evaluate protein sequences for simultaneous satisfaction of evolutionary and physical constraints, this paper develops a graphical model approach integrating sequence information from the evolutionary record of a protein family with structural information based on a molecular mechanics force field. Nodes in the graphical model represent choices for the backbone (native vs. decoys), amino acids (conservation analysis), and side-chain conformations (rotamer library). Edges capture dependence relationships, in both the sequence (correlated mutations) and the structure (direct physical interactions). The sequence and structure components of the model are complementary, in that the structure component may support choices that were not present in the sequence record due to bias and artifacts, while the sequence component may capture other constraints on protein viability, such as permitting an efficient folding pathway. Inferential procedures enable computation of the joint probability of a sequence-structure pair, thereby assessing the quality of the sequence with respect to both the protein family and the specificity of its energetic preference for the native structure against decoy structures. In a case study of WW domains, we show that by using the joint model and evaluating specificity, we obtain better prediction of foldedness of designed proteins (AUC of 0.85) than either a sequence-only or a structure-only model, and gain insights into how, where, and why the sequence and structure components complement each other.

Keywords : Protein Design, Probabilistic Graphical Models, Markov Random Fields

1 Introduction

Understanding the interrelationship among protein sequence, structure, and function is a central challenge in protein science. On the one hand, we can adopt a physical perspective, modeling the energetics determining the structure (and thereby function) of an amino acid sequence. On the other hand, we can adopt an evolutionary perspective, modeling the constraints on sequence modifications that are acceptable in maintaining structure and function. We seek here to unify these two perspectives.

The sequence-structure-function interrelationship also has implications in significant applications such as rational protein engineering, in which a variant or novel protein is designed for desired properties such as structure [4, 15, 7] or catalytic activity [18, 16]. Some design methods focus on sequence and function, selecting amino acids based on homology to existing proteins (e.g., [12, 17]). Other design methods focus on the structure, employing physical models to identify sequences with low internal energies for prescribed backbone structure (e.g., [4, 16]).

By focusing on just sequence or just structure, existing design methods lose important information. Sequence-based methods may be myopic, not considering mutations that are energetically acceptable but are not part of the evolutionary record (as currently sampled). At the same time,

*Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.
†Department of Computer Science, Dartmouth College, Hanover, NH.
‡Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA
§Corresponding Authors: cjl@cs.cmu.edu, cbk@cs.dartmouth.edu

structure-based methods, may not account for interactions that aren't evident in the native structure, such as those relevant to protein folding. Interestingly, conservation-only sequence statistics have been used as the metric by which to evaluate the plausibility of structure-based designs [14].

Another disadvantage of sequence or structure-only design algorithms is that they cannot account for sequence-structure *specificity*, whether a sequence prefers the target structure over alternatives. A few algorithms have been proposed to address the issue of specificity (e.g., [15, 2, 8]). But these have been tried for only small peptides [2], or have been shown to produce sequences that exhibit non-natural folding pathways [25], indicating the need for better approaches. By modeling the joint probability over sequence and structure, our approach naturally accounts for sequence-structure specificity.

This paper presents a graphical model approach (Fig. 1) in which we **integrate** sequence and structure information into a joint model, and use the joint model to **evaluate** sequence-structure pairs for simultaneous satisfaction of evolutionary and physical constraints, as captured in the joint probability under the model. Our model can thus assess whether a possible sequence preserves amino acid combinations that are favorable in the sequence record *and* that lead to energetically favorable interactions. It can trade off between these two aspects, accepting amino acids that didn't appear in the sequence record but have good energies, or those that don't appear to have good energies, but are prominent in the sequence record.

We demonstrate our method in a retrospective analysis of designed WW domain proteins [22], classifying 77 new sequences against a model integrating sequence information from a family of 42 wild-type WWs with structure information for the WW domain fold and near-native decoys. We show that the joint model performs better at this task than either a sequence-only model or a structure-only model. That is, the two sources of information are complementary. In particular, we show that sequence-based aspects of the model decrease false positives, while structure-based ones decrease false negatives.

In summary, the primary contribution of this paper is the first graphical model for sequence-structure specificity. In contrast to [7], our model considers both native and decoy structures. We use the graphical model to classify putative WW domain sequences, demonstrating that our model is more accurate (AUC of 0.85) than a sequence-only or a structure-only model for the same task. While we focus on modeling sequence-structure specificity, our approach directly supports the design of proteins to *maximize* this specificity using methods similar to [7] or [24].

2 Methods

We describe how to construct and utilize a probabilistic model integrating sequence and structure information for a specific protein family. The model has the following state space and variables, detailed in the rest of this section.

Sequence Let $\mathbf{S} = \langle S_1, S_2, \ldots, S_n \rangle$ be a vector of random variables for the amino acid content of a protein with n residues. The set of all possible protein sequences is $\mathcal{S}$, and $s \in \mathcal{S}$ is one such sequence composed of amino acids $\langle s_1, s_2, \ldots, s_n \rangle$.

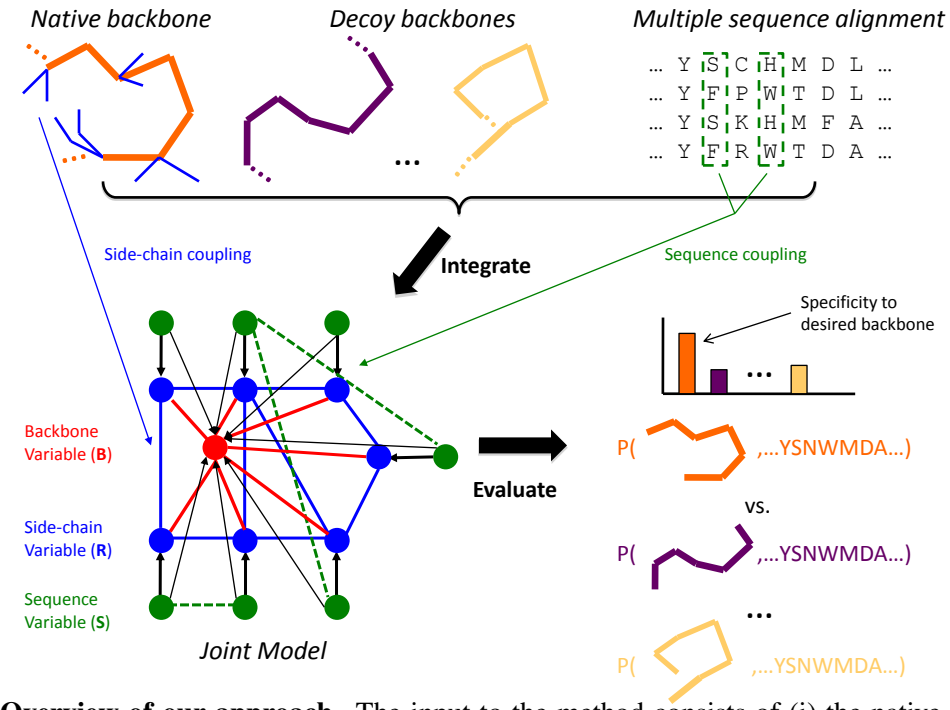


Figure 1: **Overview of our approach**. The input to the method consists of (i) the native backbone, (ii) a set of alternative (i.e., decoy) backbones, and (iii) a multiple sequence alignment (MSA). These data are integrated into a Probabilistic Graphical Model (PGM) with a single node (shown in red) encoding the choice of backbone, and a sequence (green) and side-chain (blue) node encoding the choices for each residue position. Edges in the PGM capture residue couplings based on the MSA (dashed edges) and structure (solid edges). The resulting joint model can be used to evaluate the joint probability over a given sequence and any backbone. A sequence is said to be *specific* to a backbone if their joint probability is highest.

Backbone Let $\mathbf{B}$ be the random variable representing the backbone conformation (i.e., 3D coordinates for all backbone atoms).

Side-chain conformations Let $\mathbf{R} = \langle R_1, R_2, \ldots, R_n \rangle$ represent the side-chain conformations (rotamers), $\mathcal{R}$ the set of all possible rotamers, and $r = \langle r_1, r_2, \ldots, r_n \rangle$ the side-chain conformations of the protein. Since allowed rotamers at a position depend on the amino acid, let $\mathcal{R}^s$ be the set of rotamers that are consistent with the choice of the sequence s (i.e., the states that have non-zero measure when conditioned on the event $S = s$.)

The model then provides a joint distribution over the sequence and structure variables, defining probabilities of interest in evaluating proteins (Sec. 2.1). It can be efficiently constructed and stored (Secs. 2.2, 2.3, and 2.4), and supports inference of the probabilities of interest (Sec. 2.5).

2.1 An Integrated Probabilistic Model of Sequence and Structure

We model the conditional distribution over backbone and side-chains for a given sequence as a Boltzmann distribution of the form:

$$P(\mathbf{B} = \mathbf{b}, \mathbf{R} = \mathbf{r} | \mathbf{S} = \mathbf{s}) = \frac{1}{\mathcal{Z}_\mathbf{s}} e^{-E(\mathbf{r},\mathbf{b})} \tag{1}$$

where $E_{\mathbf{r},\mathbf{b}}$ is the energy of all atoms(backbone and side-chain) of the protein as computed by a molecular mechanics force-field; and $\mathcal{Z}_{\mathbf{s}}$ the sequence specific partition function $\sum_{r\in\mathcal{R}^s,b\in\mathcal{B}} e^{-E(\mathbf{r},\mathbf{b})}$ over backbone and side-chain configurational space ensures that we have a probability measure. ¶

The joint distribution therefore has the form

$$P(\mathbf{B}=\mathbf{b},\mathbf{R}=\mathbf{r},\mathbf{S}=\mathbf{s}) = \frac{1}{\mathcal{Z}} e^{-E(\mathbf{r},\mathbf{b})} P(\mathbf{s}), \tag{2}$$

where $P(S)$ is a prior distribution of amino acids for this protein family.

Given this joint distribution, we can determine various probabilities of interest by marginalizing out the remaining variables. This leads to a set of different objective functions by which to evaluate a designed protein sequence. Most energy-based approaches to protein design for example assume a fixed backbone, b^0, and find the sequence with the best (i.e., lowest) energy for this backbone. Under a Boltzmann distribution, negative energies are related to probabilities by an exponential factor. Thus, finding the best sequence for a given backbone is the same as finding the most likely sequence for that backbone, i.e., $\arg\max_S P(\mathbf{S}|\mathbf{b}^0)$. Similarly, given a sequence, $\mathbf{s}$, finding the best backbone is the same as finding the most likely structure for that sequence, i.e., $\arg\max_B P(\mathbf{B}|\mathbf{s})$. These are different objective functions, and the backbone that maximizes $\arg\max_B P(\mathbf{B}|\mathbf{s})$ may not be $\mathbf{b}^0$. In other words, the conformation adopted by a sequence need not be that for which it was designed.

We take advantage of our probabilistic model to develop an objective function overcoming this pitfall by accounting for sequence-structure specificity. To do so, note that the joint probability $P(S, b^0)$ incorporates both how good the sequence is for the backbone, and how good the backbone is for the sequence.

$$P(\mathbf{B}=\mathbf{b},\mathbf{S}=\mathbf{s}) = \sum_R P(\mathbf{B}=\mathbf{b},\mathbf{R}=\mathbf{r},\mathbf{S}=\mathbf{s}) = \frac{1}{\mathcal{Z}_{\mathbf{s}}} Z_{\mathbf{b},\mathbf{s}} P(\mathbf{s}) \tag{3}$$

where $Z_{\mathbf{b},\mathbf{s}} = \sum_{\mathbf{r}\in\mathcal{R}^s} e^{-E(\mathbf{b},\mathbf{r})}$ is the sequence and backbone specific normalizing constant and $Z_{\mathbf{s}} = \sum_{\mathbf{b}\in\mathcal{B}} Z_{\mathbf{b},\mathbf{s}}$ is the sequence specific normalizing constant Notice that the joint probability automatically incorporates the propensity for competing backbones via the partition function $Z_{\mathbf{s}}$.

In the results, we compare predictive power of the joint probability $P(\mathbf{B}=\mathbf{b},\mathbf{S}=\mathbf{s})$ with the power of the conditional probability

$$P(\mathbf{B}=\mathbf{b}|S=\mathbf{s}) = P(\mathbf{B}=\mathbf{b},\mathbf{S}=\mathbf{s})/P(\mathbf{S}=\mathbf{s}) = \frac{1}{\mathcal{Z}_s} Z_{b,s} \tag{4}$$

and the power of the sequence prior $P(\mathbf{S}=\mathbf{s})$ in determining the foldability of a set of sequences.

Notice that the conditional probability evaluates the quality of the backbone for the sequence, but does not assess the quality of the sequence itself ($P(\mathbf{S}=\mathbf{s})$) thus using only structural information while $P(\mathbf{S}=\mathbf{s})$ ignores any structural information. The joint probability, however, incorporates both sources of information.

Unfortunately, these distributions are over extremely large spaces. For example, the joint distribution is defined over a state space of size $|\mathcal{R}\times\mathcal{B}\times\mathcal{S}|$, which is exponential in the number

¶For ease of presentation, all energies shown here are in k_BT units

of residues in the protein. In order to efficiently reason about the joint distribution, we use a Probabilistic Graphical Model (PGM) to compactly encode the entire probability distribution.

A PGM is defined as a graph over a set of random variables $\mathbf{X}$ and a set of functions F called *factors*; given an instantiation $\mathbf{x}$ of these variables, the probability $P(\mathbf{x})$ according to the PGM is given by a normalized product over the factors:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \prod_{f_a \in F} f_a(\mathbf{x_a}) \tag{5}$$

where $\mathbf{X_a}$ is the set of variables connected to factor f_a in the factor graph. While the factor graph is usually depicted as a bi-partite graph, it can be equivalently depicted as a chain-graph [27]; a form we use for its clearer physical interpretability. In this form, an edge between vertices in the graph encodes for a direct dependence quantified by a factor over the edge, while the lack of an edge encodes a conditional independency. Factors over variables quantify the prior distribution over the variables.

A PGM can represent nearly any probability distribution but it is particularly useful in providing a compact encoding of multivariate probability distributions which have conditional independencies between their variables. In the following, we will construct a PGM over $\mathbf{R}, \mathbf{B}, \mathbf{S}$ variables one variable at a time. By accounting for the conditional independencies at each stage, we can compactly encode the entire distribution. allowing us to efficiently and accurately compute the probabilities of interest as described in Sec. 2.5.

2.2 Side-Chains: Probabilistic Graphical Model for $P(\mathbf{R}|\mathbf{B}, \mathbf{S})$

Let us first assume that we already know the amino acid sequence and backbone conformation, and want to evaluate a set of side-chain conformations. Due to the nature of physical interactions, the energy of interaction between atoms falls off rapidly with distance; in many practical implementations of molecular mechanics force-fields, it is set to zero beyond a threshold distance. For example, in the toy peptide in the top left of Fig. 1, while we may consider the interaction energies between the side-chains of the first and last residue, we may ignore as negligible those between these side-chains and the side-chain that would be placed in the middle of the turn. However, there may be indirect influences, where one side-chain interacts with another which interacts with a third, although the first and third do not directly interact. If we use random variables R_1, R_2, R_3 for the choice of three such side-chain conformations, in statistical terms, we say that $R_1 \perp R_2 | R_3$, i.e., R_1 and R_2 are conditionally independent of each other when conditioned on R_3.

We [10] and others [26] have previously developed probabilistic graphical models to encode all such conditional independencies between side-chain variables, given the sequence and backbone conformation (blue subgraph in Fig. 1). By exploiting the large number of conditional independencies present in the distribution, the model dramatically reduces the space-complexity needed to store the distribution from exponential space to polynomial space (typically O(n)). We refer to the previous publications for details of the model; in summary a PGM is constructed among variables $\mathbf{R} =< R_1, \ldots, R_n >$ representing a side-chain configuration among a discrete set of *rotamers*. A backbone dependent rotamer library [3] is used to perform this discretization. Variables that

are situated at a distance lower than a threshold according to the backbone are connected by an edge. A factor $\psi_{i,j}(r_i, r_j) = \exp(-E_{\mathbf{b}}(r_i, r_j))$ is computed for each pair of variables connected by an edge using the ROSETTA force-field to compute the energy $E_{\mathbf{b}}(r_i, r_j)$ between rotamers r_i, r_j according to backbone $\mathbf{b}$. The backbone dependent library also defines a prior distribution for each rotamer; this is incorporated by a factor ψ_i for each R_i.

2.3 Backbone: Probabilistic Graphical Model for $P(\mathbf{R}, \mathbf{B}|\mathbf{S})$

Now let us consider accounting for variability in the backbone conformation, computing $P(\mathbf{B}, \mathbf{R}|\mathbf{S})$. Sec. 2.2 describes the PGM for $P(\mathbf{B}|\mathbf{R}, \mathbf{S})$; the extended PGM for $P(\mathbf{B}, \mathbf{R}|\mathbf{S})$ essentially imposes a mixture model over backbones by using the fact that $P(\mathbf{B}, \mathbf{R}|\mathbf{S}) = P(\mathbf{R}|\mathbf{S}, \mathbf{B})P(\mathbf{B}|\mathbf{S})$. In the red+blue subgraph in Fig. 1, energetic interactions between backbone atoms and side-chain atoms ($E(r_i, \mathbf{b})$) are shown with red edges, while interactions between side-chain atoms ($E_{\mathbf{b}}(r_i, r_j)$) are shown with blue edges. The energetic interactions within backbone ($E(\mathbf{b})$) atoms is stored as a vertex factor in the backbone variable. In all cases, the factor is defined as $\exp -E$ where E is the energy of the specific interaction.

Notice that the choice of the factors, along with Eq. 5 and the fact that $E(\mathbf{r}, \mathbf{b}) = E(\mathbf{b}) + E_{\mathbf{b}}(\mathbf{r})$ leads to the probability of a specific configuration $\mathbf{r}, \mathbf{b}$, $P(\mathbf{R} = \mathbf{r}, \mathbf{B} = \mathbf{b})$ being exactly the value according to the Boltzmann distribution. Thus, the red+blue subgraph exactly encodes the Boltzmann distribution over a discrete set of backbone and side-chain configurations. Due to space constraints, we refer to [9] for more details.

2.4 Sequence: Probabilistic Graphical Model for Prior $P(\mathbf{S})$

For the sequence prior, we utilize our previous work in developing undirected graphical models of residue coupling [23, 24]. While hidden Markov models (HMMs) provide a probabilistic basis for representing and reasoning with conservation,in order to best predict whether a new protein will be folded and functional, it is also necessary to account for residue coupling [22, 20]. Thus our models generalize HMMs to account for coupling in addition to conservation.

To construct this portion of our graphical model (green part in Fig. 1), we start with a multiple sequence alignment (MSA) $\mathcal{F}$ of the protein family being studied. Standard conservation statistics provide the statistical energy for amino acid type a at residue position i:

$$\phi_i(a) = -\lambda * \log(|\{s \in \mathcal{F} : s[i] = a\}| \,/\, |\mathcal{F}|) \tag{6}$$

The corresponding factor of the PGM is then defined as $\exp -\phi_i(a)$ as earlier.

In practice, we employ "pseudo-counts" to ensure that there is non-zero probability of any amino acid at any position; this acknowledges the incompleteness of the extant sequences. λ is a hyper-parameter of our model which defines the strength of the statistical energy. In our experiments, we choose a λ such that the statistical energies have the same importance as the structural energies. This amounts to giving equal importance to both sources of information.

Likewise, for a pair of coupled positions, the statistical energy for a pair of amino acids a, b at a pair of positions i, j is based on the number of occurrences in the family:

$$\phi_{i,j}(a, b) = -\lambda * \log(|\{s \in \mathcal{F} : s[i] = a, s[j] = b\}| \,/\, |\mathcal{F}|) \tag{7}$$

again in practice adding pseudo-counts (now for each possible amino acid pair).

The remaining question is which edges to include in the model. Traditional residue coupling studies employ one of a number of possible statistical and information-theoretic metrics [6] to identify highly covarying residue pairs, and simply list all such pairs. However, including all such edges in a model may lead to incorrectly evaluating the joint probability. For example, in the MSA in Fig. 1, including 2-4, 2-5, and 4-5 is in some sense "double counting" the 2-4-5 relationship, since once we have evaluated a sequence for consistency against two of the three edges (say 2-4 and 2-5), there's no additional information provided by testing the third edge. Thus in a graphical model, we "factorize" the relationships into direct and indirect dependencies, as well as independencies. The semantics are such that a residue is conditionally independent of all others, given its immediate neighbors. Edges capture direct relationships, multi-edge paths capture indirect relationships, and residues with no path between them are independent. Note that there is not necessarily a unique way to separate direct vs. indirect relationships, but that several factorizations may be equally valid (except for noise and effects of small counts), capturing the same information about the constraints and expressing the same probability model.

We have developed an algorithm for learning the edges of a graphical model of residue coupling, and demonstrated its effectiveness in a variety of proteins, including GPCRs and PDZs [23]. The algorithm considers a set of residues pairs deemed to have significantly significant covariation according to a χ^2 test (i.e., the number of pairs of some amino acid types is much more than would be expected if the positions were independent). It then incrementally adds edges to a growing model. When choosing the next potential edge to add, it evaluates the impact on the residual coupling across the protein, as determined by a conditional mutual information score. That is, it seeks edges that not only connect coupled residues, but are also good "decouplers", rendering other residues conditionally independent (in keeping with the underlying semantics of the model).

2.5 Inference: Computing $Z_{\mathbf{b},\mathbf{s}}, Z_{\mathbf{s}}$

Given a PGM as described in the previous sections, our task is to now compute the joint probability of a sequence $\mathbf{s}$ in a backbone $\mathbf{b}$, $P(\mathbf{b}, \mathbf{s})$, the prior probability of $\mathbf{s}$ for this protein family $P(\mathbf{s})$ and the conditional probability of $\mathbf{s}$ for backbone $\mathbf{b}$, $P(\mathbf{b}|\mathbf{s})$. These probabilities, according to Eq. 3, Eq. 4 depend on $Z_{\mathbf{b},\mathbf{s}}, Z_{\mathbf{s}}$.

We use a statistical inference algorithm called loopy *Belief Propagation* [19] (LBP or simply BP) to compute $Z_{\mathbf{b},\mathbf{s}}$. Statistical inference algorithms like BP compute marginal distributions over the random variables in the graph. BP takes $O(|\mathcal{E}|)$ time per iteration, where $|\mathcal{E}|$ is the number of edges in the graph; in acyclic factor graphs BP takes two iterations. The PGMs considered here however are not trees; in such cases BP is run until the iterations converge. While convergence is not guaranteed, is has always done so in our experiments. Significantly, it has been proved, that using BP [19] is equivalent to using the Bethe approximation [1] of the log of the partition function. This approximation is quick and has been shown to accurately predict experimentally verifiable properties of the protein (changes in free energies) [10, 11].

We can thus use BP on the PGM encoding $P(\mathbf{R}|\mathbf{B}, \mathbf{S})$ to efficiently and accurately approximate $Z_{b,s}$, the partition function specific to a sequence and backbone, i.e., over all rotamers. Given this approximation to $Z_{b,s}$ to each backbone b, the Z_s can be easily computed using its definition

according to Eq. 3.

Finally, in our experiments instead of using the probabilities listed above, it will be convenient to use their logarithms instead. Due to the monotonicity of the logarithm, this does not affect our results in any way and has the benefit of being numerically more stable. Additionally, we will ignore any universally common normalizing factors (that do not depend on instantiation of $\mathbf{s}, \mathbf{R}$ or $\mathbf{B}$) since they are constant additive factors in the log probabilities that do not affect the classification results. We will refer to the log probabilities $\log(P(\mathbf{s}))$, $\log(P(\mathbf{b}^0|\mathbf{s}))$ and $\log(P(b^0, \mathbf{s}))$ as sequence, structure and joint scores respectively to reflect the source of information that they incorporate.

3 Results

We performed a detailed case study of our approach in analysis of WW domains, taking advantage of the existence of a relatively large set of artificial proteins generated and experimentally evaluated by the Ranganathan lab [22, 20].

3.1 Joint Model Construction

We used the backbone structure from the NMR structure of the WW domain of the ubiquitin ligase NEDD4 (PDB ID 1I5H) as our native backbone. Decoy backbones were generated using ROSETTA's backrub generating mode, which generates backbones by manipulating the backbone dihedrals and pruning away energetically unfavorable conformations [21, 5]. We generated 20 alternate backbones for each sequence in the dataset. Our results are not sensitive to the choice of number of backbones, and are nearly identical even if we use as few as 4 backbones or as many as 40 backbones. We stress the fact that our choice of generating backbones is not binding; indeed we expect that for larger proteins, alternate strategies of generating backbones will have to be used in order to efficiently sample the equilibrium backbone conformational space.

For the sequence portion of our model, we used an MSA of 42 natural WW domains provided by the Ranganathan lab [22]; our results are similar when incorporating additional sequences obtained by PSI-BLAST. We selected the 34 columns represented in the backbone structure.

The sequence coupling model factorizes the statistically significant residue couplings (p value > 0.005; $> 8.91e^{-6}$ after a Bonferroni adjustment) into a set of 25 edges. The structure coupling model puts an edge between variables that are when the corresponding C_α atoms are closer than 10 **r**A. in the backbone, consists of 201 edges. This threshold is sufficient to ensure that all non-zero energies of interaction according to the ROSETTA force-field [13] are incorporated. The joint model, which is a union of the two models, thus consists of 214 edges. The potentials under the joint model are a linear combination of the potential scores under the individual models. That is, joint score = structure model score + sequence model score. We chose $\lambda = 0.26$ as this was the ratio of the score ranges from the two model components. This choice equalizes the contributions of the two components as explained in Sec. 2.4.

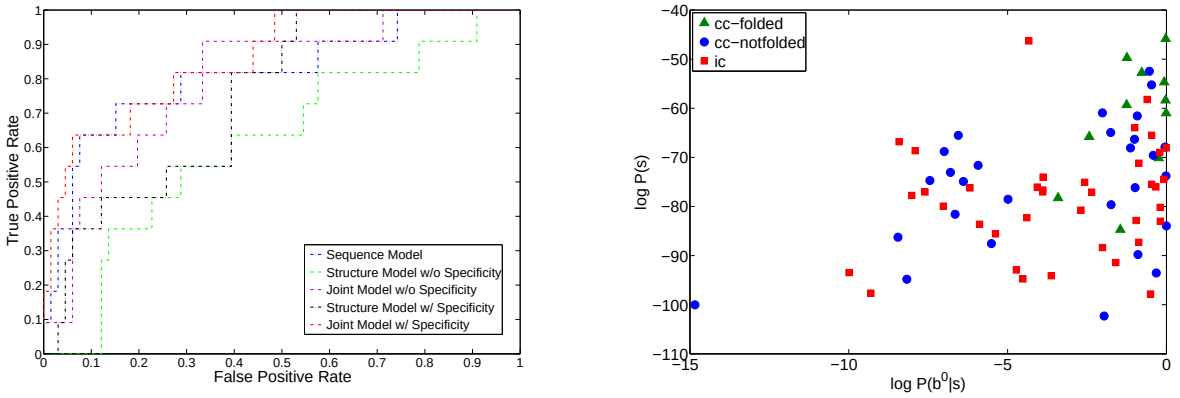
3.2 Predictive power

We evaluate our model against the collection of 77 artificial sequences evaluated for foldedness by the Ranganathan lab [22]. We compare and contrast the predictive power of a sequence-only model$P(\mathbf{S})$, a structure-only model($P(\mathbf{B}|\mathbf{S})$), and our joint model ($P(\mathbf{B},\mathbf{S})$) in terms of the area under the curve (AUC) of the receiver operating characteristic (ROC) curve (Fig. 2-A). The AUC of the sequence-only model is 0.82, while that of the structure-only model is 0.75. The joint model provides an AUC of 0.86, an improvement over both the individual models. The figure also shows the AUC for the joint model when no competing backbones are allowed("w/o specificity"). As can be seen, the performance of the models when no structural specificity is encoded is significantly lower (AUC falls from 0.75 to 0.62 in the structural model and from 0.86 to 0.80 in the joint model).

The sequence-only model has high TPR for low values of FPR but doesn't retrieve all true positives until the FPR is nearly 0.7. The structure model with backbone specificity, on the other hand, has lower TPR than the sequence only model, but is able to retrieve all true positives by the time FPR=0.5. The joint model is able to capture the best behavior of both models by obtaining high TPR for low values of FPR and retrieving all TPRs by the time FPR=0.5.

The primary parameter in the joint model is our choice of scaling factor, λ, for weighting the potential scores under the two models. Our results are not sensitive to the choice of λ: the AUC of the joint model is similar for values of λ between 0.2 and 1.0.

Fig. 2(B) shows the scores of the two sources of information – the sequence prior(on the y-axis, in log scale) and the structural specificity (on the x-axis, in log scale). This figure clearly shows the complementary nature of the two sources of information: each model perfoms reasonably well by themselves, but more importantly the sequences that they misclassify (low/high scores for folded/unfolded sequences respectively). Thus, the joint model which combines both scores is able to outperform both scores.



(A) (B)

Figure 2: (A) ROC curves for the three different objectives with and without backbone specificity. (B) Scatter plot of the log probabilities of the structural score ($\log P(\mathbf{b}^0|\mathbf{S})$, on the x-axis) and of the prior distribution ($\log P(\mathbf{S})$, on the y-axis) for the sequences in the test set. The true positives(cc-folded) are shown in green triangles while the false positives are shown in blue circles and red squares

3.3 Model analysis

There are a few interesting observations which explain the increased AUC of the joint model. The sequence model has a lower TPR, i.e. it has more false negatives than the structure model, and hence the joint model does well on this part due to addition of structure knowledge. On the other hand, the FPR is high under the structure only model, whereas the sequence only model has a much lower FPR and hence the joint model does well on this part because of the additional sequence information.

To obtain a more detailed understanding of the underlying interactions producing these behaviors, we computed a score for each edge with respect to each protein in the test set. The score for an edge i, j according to the sequence model is simply its contribution to the total score, $\lambda\phi_{i,j}(a, b)$. For an edge in the structure model, it is harder to assign an individual score since the score for a sequence is a difference in the log partition functions, a term that isn't decomposable into a sum of edge-wise scores. We therefore approximate this term by its first order moment, a value that physically corresponds to the difference in the average energy of interaction of this edge between b^0 and the most favorable backrub structure. The score between a pair of residues that were connected by both a sequence edge and a structure edge was simply the sum of the scores of the individual components. Given this score, we ranked the edges based on decreasing hinge loss in classifying the sequence correctly (top rank implies more predictive edge)

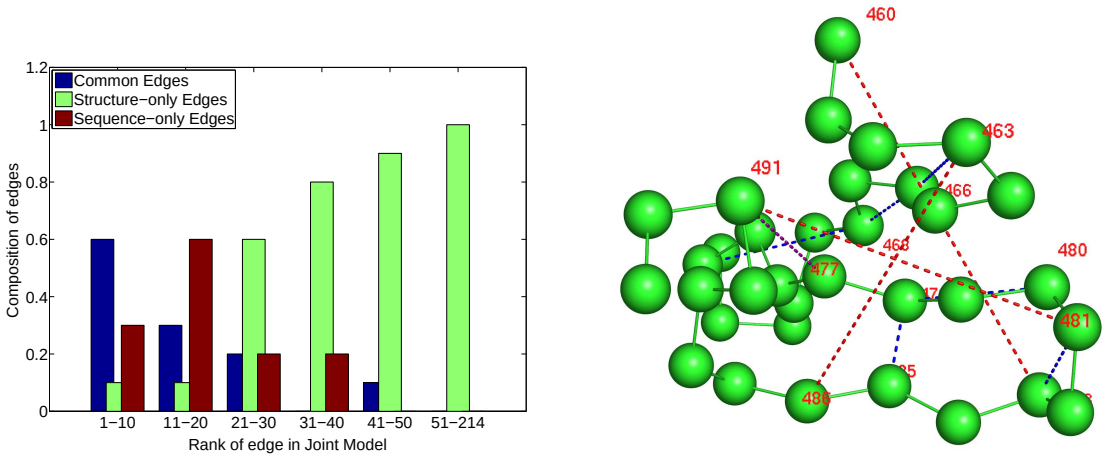


Figure 3: (A) Edge composition (sequence coupling/structure coupling/joint coupling) of the edges in the joint model ranked according to their individual classification performance. (B) Top-ranked edges, color-coded according to model (red:sequence only, purple:structure only, blue:structure and sequence), visualized in the backbone trace of the WW structure used as native in our study (PDB ID 1I5H).

Fig. 3.3(A) shows the composition of the edges (sequence only,structure only,common) in the ranked list. In the top ten, there are 3 exclusive to the sequence model, 1 exclusive to structure coupling model, and 6 common to both. These top ten edges are shown as dashed edges in the pdb structure in Fig. 3.3(B). The common edge between residue positions 478 and 485(both in a

$\beta - sheet$) is low ranked in the sequence coupling model, but is the highest ranked in the structure coupling model. The lack of having enough data in the training set for this interaction results in a low score for this edge (average score of this interaction in the test set was nearly half the score of interactions in the training set). This is however compensated by the structure information which captures the importance of this interaction.

The edge between residue positions 468 and 474 is high ranked in the sequence coupling model, but rather lowly ranked in the structure coupling model. This is due to these two positions are distally located in the protein leading to a negligible energy of interaction. This is an instance of information being captured by the sequence coupling model, that was not captured well by the structure model, and hence this improves the performance of the joint model. Amongst the top ranked joint model edges, edges which are exclusively in the sequence coupling model, are the edges that explain important long range interactions between residues, and which go amiss under the structure coupling model. Again the joint model utilizes the information encapsulated in the two separate models, which are complementary to each other and hence when used together builds a more powerful model.

In conclusion, we have developed and demonstrated a model that integrates information from both structure and sequence to accurately and efficiently model sequence-structure specificity. Our results indicate that by integrating these complementary sources of information, this model is able to outperform each of its individual components.

References

[1] H. A. Bethe. Statistical theory of superlattices. *Proc. Roy. Soc. London A*, 150:552–575, 1935.

[2] Parbati Biswas, Jinming Zou, and Jeffery G. Saven. Statistical theory for protein ensembles with designed energy landscapes. *The Journal of Chemical Physics*, 123(15), 2005.

[3] A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr. A graph theory algorithm for protein side-chain prediction. *Protein Sci.*, 12:2001–2014, 2003.

[4] B.I. Dahiyat and S.L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, Oct 1997.

[5] I.W. Davis, W.B. Arendall, D.C. Richardson, and J.S. Richardson. The backrub motion: How protein backbone shrugs when a sidechain dances. *Structure*, 14(2):265–274, 2006.

[6] A.A. Fodor and R.W. Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56:211–221, 2004.

[7] M. Fromer and C. Yanover. Accurate prediction for atomic-level protein design and its application in diversifying the near-optimal sequence space. *Proteins: Structure, Function, and Bioinformatics*, page In Press, 2008.

[8] J. J. Havranek and P. B. Harbury. Automated design of specificity in molecular recognition. *Nat Struct Biol*, 10(1):45–52, January 2003.

[9] H. Kamisetty, C. Bailey-Kellogg, and C.J. Langmead. A graphical model approach for predicting free energies of association for protein-protein interactions under backbone and side-chain flexibility. Technical Report CMU-CS-08-162, Carnegie Mellon University, 2008.

[10] H. Kamisetty, E.P. Xing, and C.J. Langmead. Free Energy Estimates of All-atom Protein Structures Using Generalized Belief Propagation. In *Proc. 7th Ann. Intl. Conf. on Research in Comput. Biol. (RECOMB)*, pages 366–380, 2007.

[11] H. Kamisetty, E.P. Xing, and C.J. Langmead. Free Energy Estimates of All-atom Protein Structures Using Generalized Belief Propagation. *J. Comp. Bio.*, 15(7):755–766, 2008.

[12] S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik, and M.H. Hecht. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262(5140):1680–1685, Dec 1993.

[13] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *PNAS*, 99(22):14116–14121, October 2002.

[14] B. Kuhlman and D. Baker. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*, 97(19):10383–10388, 2000.

[15] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, Nov 2003.

[16] R. Lilien, B. Stevens, A. Anderson, and B. R. Donald. A novel ensemble-based scoring and search algorithm for protein redesign, and its application to modify the substrate specificity of the gramicidin synthetase A phenylalanine adenylation enzyme. *J. Comp. Biol.*, 12(6-7):740–761, 2005.

[17] C. Loose, K. Jensen, I. Rigoutsos, and G. Stephanopoulos. A linguistic model for the rational design of antimicrobial peptides. *Nature*, 443(7113):867–869, Oct 2006.

[18] C.R. Otey, J.J. Silberg, C.A. Voigt, J.B. Endelman, G. Bandara, and F.H. Arnold. Functional evolution and structural conservation in chimeric cytochromes P450: Calibrating a structure-guided approach. *Chem. Biol.*, 11(3):309–318, Mar 2004.

[19] J. Pearl. Fusion, propagation, and structuring in belief networks. *Artif. Intell.*, 29(3):241–288, 1986.

[20] W.P. Russ, D.M. Lowery, P. Mishra, M.B. Yaffee, and R. Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437(7058):579–583, Sep 2005.

[21] C. A. Smith and T. Kortemme. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.*, 380(4):742–756, July 2008.

[22] M. Socolich, S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, Sep 2005.

[23] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Graphical models of residue coupling in protein families. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 5(2):183–197, 2008.

[24] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Protein design by sampling an undirected graphical model of residue constraints. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2009. In press.

[25] A. L. Watters, P. Deka, C. Corrent, D. Callender, G. Varani, T. Sosnick, and D. Baker. The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell*, 128(3):613–624, February 2007.

[26] C. Yanover and Y. Weiss. Approximate inference and protein folding. *Proc. NIPS*, pages 84–86, 2002.

[27] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51:2282–2312, 2005.