

# Graphical Models of Residue Coupling in Protein Families

Dartmouth Computer Science Technical Report TR2005-535

John Thomas\*    Naren Ramakrishnan†    Chris Bailey-Kellogg\*

**Abstract:** Identifying residue coupling relationships within a protein family can provide important insights into intrinsic molecular processes, and has significant applications in modeling structure and dynamics, understanding function, and designing new or modified proteins. We present the first algorithm to infer an undirected graphical model representing residue coupling in protein families. Such a model serves as a compact description of the joint amino acid distribution, and can be used for predictive (will this newly designed protein be folded and functional?), diagnostic (why is this protein not stable or functional?), and abductive reasoning (what if I attempt to graft features of one protein family onto another?). Unlike current correlated mutation algorithms that are focused on assessing dependence, which can conflate direct and indirect relationships, our algorithm focuses on assessing independence, which modularizes variation and thus enables efficient reasoning of the types described above. Further, our algorithm can readily incorporate, as priors, hypotheses regarding possible underlying mechanistic/energetic explanations for coupling. The resulting approach constitutes a powerful and discriminatory mechanism to identify residue coupling from protein sequences and structures. Analysis results on the G-protein coupled receptor (GPCR) and PDZ domain families demonstrate the ability of our approach to effectively uncover and exploit models of residue coupling.

**Keywords:** residue coupling, graphical models, evolutionary co-variation.

## 1 Introduction

When studying a family of proteins that have evolved to perform a particular function, a major goal of contemporary biological research is to uncover constraints that appear to be acting on the family, with an eye toward understanding the molecular mechanisms imposing the constraints. For example, amino acid conservation has long been recognized as an important indicator of structural or functional significance [25]. In the 1990s, researchers began generalizing single-position conservation to correlated co-evolution of amino acid pairs, thus revealing cooperativity and coupling constraints (e.g., one early study focused on the HIV-1 envelope protein, with the aim of guiding peptide vaccine design [15]). Such works boosted the discovery of coupled residues, which could previously have been identified only by painstaking *in vitro* approaches such as thermodynamic double mutant analysis [11]. The next step was to summarize information about correlated positions into pathways [14], motifs [1, 18], and structural templates [18] in protein families. Today, projects undertake ambitious large-scale recombination [26] or site-directed and combinatorial mutagenesis studies [21] to identify entire building blocks of proteins important to preserve function.

Knowing which pairs (or sets) of residues are coupled in a protein family aids our understanding of many important processes, e.g., protein folding and conformational change [19, 22], signaling [24], protein-protein interaction, and even protein complex assembly [13]. Since the basis

---

\*Dept. of Computer Science, Dartmouth College, Hanover, NH 03755; {jthomas,cbk}@cs.dartmouth.edu

†Dept. of Computer Science, Virginia Tech, Blacksburg, VA 24061; naren@cs.vt.edu

for coupling can be structural and/or functional, information about coupled residues can be used predictively for assessing protein structure [23], fold classification [9], or even to suggest novel sequences for protein engineering [20].

While there are many computational techniques for studying residue coupling [6], all methods begin by defining a metric to quantify the degree to which two residues co-vary. Global methods then determine pairs of coupled residues by observing correlated mutations in the protein family multiple sequence alignment (MSA) as a whole (e.g., [15]). The state-of-the-art in understanding residue coupling is, however, a local method — so-called ‘perturbation-based’ analysis [4] introduced by Lockless and Ranganathan [16]. The basic idea is to subset the MSA according to some condition (e.g., containing a moderately conserved residue type at a particular position) and observe the effect of the perturbation on residue distributions at other positions. If the subsetting operation significantly alters the proportions of amino acids at some other position, it is inferred to be coupled to the perturbed position, according to the evolutionary record. Even though this approach is purely sequence-based, it has been shown to uncover structural networks of residues underlying important allosteric communication pathways in proteins [24].

A key missing ingredient to date is a formal probabilistic model capturing the constraints inferred from residue coupling studies. Such a model would help assess the feasibility and significance of inferring a network from coupling data, including determining whether such a network is a persistent feature of a protein family or merely a hallucination. The process of inferring such a model would help make explicit the essential constraints underlying the family (e.g., by identifying a small set of correlations that explain the data nearly as well as the complete set). A model would enable the careful combination of multiple information sources (e.g., by integrating priors from structural and functional studies with correlations derived from sequence analysis). Finally, the model would serve as a compact description of the joint amino acid distribution, and could be used for predictive (will this newly designed protein be folded and functional?), diagnostic (why is this protein not stable or functional?), and abductive reasoning (what if I attempt to graft features of one protein family onto another?).

This paper addresses these needs by formulating and elucidating the natural correspondence between a coupling network (qualifying interdependencies among residues) and a probabilistic graphical model (summarizing interrelationships between random variables).

1. We present the *first* algorithm to infer an undirected graphical model underlying residue coupling in protein families. We bring in ideas from the extensive literature on probabilistic models [3] to derive networks that are meaningful as indicators of joint variation of sequence positions and that also explain structural features of protein families.
2. Unlike current correlated mutation algorithms that are focused on assessing dependence (which can conflate direct and indirect relationships) we focus on assessing *independence* (which enables modular reasoning about variation). We thus derive more compact descriptions of underlying networks highlighting the most important relationships.
3. We demonstrate how hypotheses regarding possible underlying mechanistic/energetic explanations for coupling can be used as priors for computational model discovery. For instance, if we have reason to believe that coupling in a given family would be only between nearby

residues, a representative contact graph can be utilized as a valuable prior, benefiting algorithmic complexity and ensuring biological interpretability of the results.

## 2 Background: Correlated Mutations and Residue Coupling

We begin by providing some background about correlated mutations and how they are used as indicators of residue coupling. Typically, we are given a multiple sequence alignment (MSA) whose rows are the members of the family and the columns are the aligned residue positions. Thus the MSA can be thought of a matrix where the value in row  $s$  and column  $j$  refers to the  $j$ th residue according to sequence  $s$ . We ignore columns that are too ‘gapful,’ and ignore in the calculations below the remaining entries that are gaps.

A coupling constraint quantifies the degree to which two positions in the family co-vary. Given positions  $i$  and  $k$ , information about amino acid occurrences contained in the  $i$ th and  $k$ th column vectors of the MSA can be summarized into 20-element vectors of frequencies, or probability distributions  $P(i)$  and  $P(k)$ . Essentially, this allows us to think of residue positions as random variables over a discrete sample space of 20 possibilities (recall that we ignore gaps). Coupling can then be estimated by many information-theoretic and statistical metrics; one example is the (global) *mutual information* between  $P(i)$  and  $P(k)$ , given by:

$$MI(i, k) \equiv \sum_{i=1}^{20} \sum_{k=1}^{20} P(i, k) \log \frac{P(i, k)}{P(i)P(k)}$$

Notice that the mutual information is actually the KL divergence [17] between the distributions  $P(i, k)$  and  $P(i)P(k)$ ; it quantifies the margin of error in assuming that the joint distribution  $P(i, k)$  is decomposable.  $MI(i, k)$  is zero when the underlying distributions are independent and non-zero otherwise. Another way to think of  $MI(i, k)$  is as the difference

$$MI(i, k) \equiv H(i) - H(i|k)$$

where  $H(i)$  is the entropy of the random variable  $i$  and  $H(i|k)$  is the entropy of the probability distribution  $P(i|k)$ . If  $MI(i, k) = 0$ , then knowing the value of  $k$  does not reduce our uncertainty about  $i$ . A high score of  $MI(i, k)$  is typically used as an indicator of coupling [15].

There are other ways to quantify coupling, e.g., using covariances and correlations; see [6]. All metrics, however, suffer from estimation problems under high or low degrees of conservation. For instance, if position  $i$  is always alanine and position  $k$  is always glutamine, then  $MI(i, k)$  would be assigned zero even though we have not observed any variation in either! Similar problems arise with residues that have low frequencies of certain amino acids. It is hence well-recognized that ‘correlated mutation algorithms must favor an intermediate level of conservation’ [6].

The typical use of the above concepts is to posit graphs or networks summarizing the constraints inferred. Traditionally researchers have used coupling constraints as a basis to infer the contact map—since coupled residues are often known to be spatially proximal—and this is still a popular way to validate correlated mutation algorithms (e.g., see [4]). Others compare the constraints to known energetic couplings inferred from double mutant experiments [7]. Still others

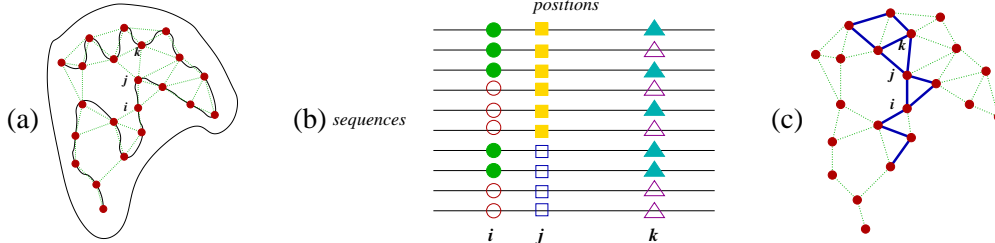


Figure 1: Graphical models of residue coupling. (a) A graph expressing a prior over possible coupling relationships. One source for a prior could be the contact graph representation of a protein’s three-dimensional structure; here, mechanistic explanations for coupling posit either a direct interaction between contacting residues, or an indirect (transitive) propagation of an interaction through networks of contacting residues. (b) The multiple sequence alignment for members of a protein family provides evidence for dependence and independence. In the example, positions *i* and *k* are very correlated — when *i* is a ‘filled in’ residue, *k* tends to be as well; similarly when *i* is ‘empty,’ *k* tends to agree. However, knowing *j* makes the positions rather independent. In the most common case where *j* is filled in, we see the combinations of types at *i* and *k* are more evenly distributed. This suggests that *i* and *k* are conditionally independent, given *j*. (Of course, even in this example, noise obscures the degree of independence.) (c) A graphical model (darkened edges) capture conditional independence. We construct such a model by selecting edges from the prior that best decouple other relationships. For example, we see that the conditional independence of *i* and *k* given *j* can be explained by a transitive propagation of interaction along model edges.

attempt to organize the couplings into pathways of allosteric communication through the protein [14]. The discovery of such pathways has recently been reinvigorated with the work of [24] where the authors perform perturbation-based analysis at numerous positions and subsequently ‘cluster’ the pairs of coupled residues; this procedure has been shown to yield sparse, connected, networks in many protein families.

### 3 Learning Graphical Models of Residue Coupling

If coupled residues indeed capture meaningful relationships, then they must afford a probabilistic interpretation. That is our working hypothesis for this paper and helps highlight where all previous work falls short. All previous approaches to inferring networks from data do so by direct incorporation of couplings as dependencies and, as is well known, such an approach cannot distinguish direct from transitive dependencies. It is also clear that (in)dependence of random variables is a very conditional phenomenon: two random variables may be correlated, become uncorrelated in the presence of new evidence, become correlated again when given further evidence, and so on. This means that we must pay careful attention to conditioning contexts, especially when we employ perturbation-based correlated mutation algorithms.

Our proposed approach is to directly learn an undirected graphical model [3] that encodes the network of residue coupling relationships. Such a model  $N(\mathcal{V}, \mathcal{E})$  encodes a joint distribution over the space of random variables in  $\mathcal{V}$  (residues) as a product of potential functions over  $N$ ’s cliques.

Formally, this joint distribution is given by:

$$P(\{\mathcal{V}\}) = \prod_{c \in \text{cliques}(N)} \phi_c(c) \tag{1}$$

By the Hammersley-Clifford theorem [3], the conditional independence statements represented by such a model are a union of statements of the form ‘a node is independent of its non-neighbors given its neighbors.’ In other words, variation in a residue position is independent of all others when we are given information about neighboring residues. Notice that ‘neighboring’ here indicates adjacency w.r.t. the inferred graphical model, not (necessarily) physical proximity according to the protein structure. For instance, a network with no edges implies that all the variation in the MSA is captured through independent, pointwise variations occurring at the residues. One situation where this would happen is with high degrees of conservation.

Eq. 1 yields a natural likelihood formulation for evaluating a model with respect to a set of input sequences. The likelihood is given by the product of marginals defined over the cliques of  $N$  divided by the product of marginals defined over the clique adjacencies (which could be nodes, edges, or general subgraphs). In this view, each potential of Eq. 1 is either a conditional or a joint marginal distribution (but they cannot all be of the same type).

Uncovering graphical models from datasets is known to be an NP-hard problem in the general case and researchers typically restrict either the topology of the network (e.g., to trees) or adopt heuristics to search the space of possibilities. In this paper, we assume the existence of a candidate set of edges (a graph prior) and propose heuristics that sequentially infer conditional *independencies* among this set (rather than dependencies as followed in prior work). If we know that residues  $i$  and  $k$  become independent given  $j$ , i.e., the conditional mutual information

$$MI(i, k|j) = H(i|j) - H(i|k, j)$$

is zero, then it is easy to see that the removal of  $j$  and its incident edges must separate  $i$  and  $k$  in the unknown network  $N$ . This assessment is made in the context of a prior graph  $G = (V, E)$ , where we assume  $\mathcal{V} = V$  and  $\mathcal{E} \subset E$ . This approach is akin to the ‘sparse candidate’ algorithm [8] for learning (directed) Bayesian networks.

Fig. 1 presents an example of such an inference. In attempting to de-couple position  $i$  from  $k$ , we need only consider neighbors of  $i$  (e.g.,  $j$ ) according to the graph prior. Notice, however, that insisting on complete independence, i.e.,  $MI(i, k|j) = 0$ , is a very stringent criterion especially in the presence of small, finite datasets; instead we assess how much the conditional mutual information is decreased. The above ideas can be used to formulate an approach for scoring a network as well as an algorithm for greedily inferring the network from data. The score for a network is given by:

$$\text{Score}(N(\mathcal{V}, \mathcal{E})) = \sum_{n \in \mathcal{V}} \sum_{m \notin \text{neighbors}(n)} MI(n, m|\text{neighbors}(n))$$

We use this notion of network score to define an edge score as the difference in score between the network without the edge and the network with the edge. Note that the score of an edge can be negative, if adding the edge produces more coupling in the network. This then allows us to incrementally grow a network by, at each step, selecting the edge that scores best with respect to

```

function InferNetwork ( $G = (V, E)$ )
 $\mathcal{V} \leftarrow V; \mathcal{E} \leftarrow \emptyset$ 
 $s \leftarrow \text{Score}(\mathcal{V}, \mathcal{E})$ 
 $C \leftarrow \{(e, s - \text{Score}(\mathcal{V}, \mathcal{E} \cup \{e})) \mid e \in E\}$ 
repeat
   $e \leftarrow \arg \max_{e \in E - \mathcal{E}} C(e)$ 
   $\mathcal{E} \leftarrow \mathcal{E} \cup \{e\}$ 
  for all  $e' \in E - \mathcal{E}$  such that  $e$  and  $e'$  share a vertex do
     $C(e') \leftarrow C(e) - \text{Score}(\mathcal{V}, \mathcal{E} \cup \{e'\})$ 
  end for
until “enough” edges incorporated

```

Figure 2: Algorithm for inferring graphical models of residue coupling.

the current network. Fig. 2 gives this algorithm. The algorithm can be configured to utilize various greedy stopping criteria—stop when the newly added edge’s contribution is not significant enough, stop when a designated number of edges have been added, or stop when the likelihood of the model is within acceptable bounds.

The run-time of our algorithm depends on  $n$ , the number of residues in the protein of interest and  $d$ , the maximum degree of nodes in the prior. With an uninformative prior,  $d$  is  $n$ . For stronger priors (e.g., a contact graph), we can assume a bounded number of neighbors for any residue, so  $d$  is  $O(1)$ . The algorithm scores  $O(dn)$  edges at each iteration. Naive execution of the algorithm requires that the score of the network be computed for each edge at each iteration. Scoring a network requires  $O(n)$  *MI* computations for each residue and there are  $n$  residues, so naive execution requires  $O(dn^3)$  *MI* computations at each iteration. Since conditioning contexts change dynamically during the operation of the algorithm, we cannot perform any *a priori* pre-processing to accumulate sufficient statistics (in contrast to global methods where mutual information between all pairs of residues can be computed in a single pass). However, the cost of making fresh assessments is curtailed since conditioning contexts are merely subsets of neighbors. Thus by caching values efficiently we can improve the runtime by a factor of  $O(n^2)$  at each iteration. First, pre-compute the score of every edge in consideration, requiring  $O(dn^3)$  *MI* computations. At each iteration, rather than recomputing scores, pick the edge in the cache that improves the score of the network the most. This requires  $O(n)$  time, but does not require any *MI* computations. The key observation is that after an edge is added, the only edges whose scores change are those incident to the edge just added. Since there are at most  $O(d)$  of those that need to be updated, we need only  $O(dn)$  *MI* computations, for a speedup of  $O(n^2)$ . Additional constant factor speedups can be achieved by removing at each step edges that produce statistically unsound conditioning contexts.

## 4 Results

We illustrate our algorithm for inferring graphical models of residue coupling with two protein families: GPCRs (G-protein coupled receptors) and PDZ domains. GPCRs are membrane-bound proteins critical in intracellular communication and signaling, and a key target of molecular mod-

eling in drug discovery. Since ligand binding at the extracellular face initiates propagation of structural changes through the transmembrane helices and ultimately to the cytoplasmic domains, GPCRs make an appropriate and compelling study for network identification [24]. PDZ domains are protein-protein interaction domains that occur in many proteins and are involved in a wide variety of biological processes [10]. One role of PDZ domains is assisting in the formation of protein complexes by binding to the C-termini of certain ligands [10]. Traditionally, PDZ domains have been classified into two types according to which type of ligand they bind. The first class of PDZ domains binds to C termini with sequences S/T-X- $\Phi$  ( $\Phi$  is a hydrophobic residue) while the second class targets sequences of the form  $\Phi$ -X- $\Phi$ . Through these two studies we aim to explore many pertinent aspects of our approach, such as how to set priors, studying the progress of the algorithm as new edges are added, using the induced graphical model for classifying protein sequences, and biological interpretation of the results. Due to space considerations, we do not discuss all these issues for both families.

## GPCRs

In the GPCR study, we study the use of protein contact graphs as priors and also explicitly relate the structure of our identified networks with those previously identified [24]. We first retrieved the multiple sequence alignment of 940 members of the class A GPCR family as discussed in [24]. In order to explore contact graph priors, we constructed a contact graph from the three-dimensional structure of one prominent GPCR member, bovine rhodopsin (PDB id 1HZX), identifying 3161 pairs of residues with atoms within 7 Å. We verified that the residues previously identified as belonging to networks [24] form connected subgraphs of this contact graph.

For this study, in testing conditional mutual information, we only considered cases for which at least 15% of the original set of sequences remained after subsetting to a particular residue type. As discussed [16], such a bound is required in order to ensure sufficient fidelity to the original MSA and allow for evolutionary exploration. The above bound of 15% is roughly half that used in [24], since our algorithm subsets according to multiple residues, depending on the number of neighbors available, whereas the previous algorithm subsets according to only one residue. From extensive experiments with this parameter (data not shown), we found that while there is some variation in the edges with changes of this parameter, many ( $> 70\%$ ) of the best edges are insensitive to the exact threshold.

In order to evaluate the implications of restricting dependencies to structural neighbors, we compared the  $MI$  scores for edges in the protein contact map against those for all pairs of residues. For every residue, we identified both the best decoupler *anywhere* in the protein, and the best decoupling contact graph neighbor. Fig. 3 shows the absolute differences between these values. Notice that in most cases, the best neighbor provides nearly as much decoupling as the best residue elsewhere in the graph. However, there are some nodes that incur a large penalty. In general, these nodes are highly conserved and therefore have small scores against all other nodes. However, since the total number of residues is large, the sum of all these small correlations becomes non-trivial. When a node is subsetted, making an originally highly conserved node to become perfectly conserved, the score for that node drops to 0. In this case there is a large difference in improvement between selecting a distant node and a node from the original prior graph. It is important to keep

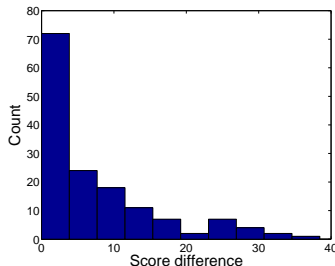


Figure 3: Penalty for decoupling using a contact graph neighbor rather than any residue (frequency distribution). Lower score differences indicate that neighbors perform as well as other residues. The skewing of the frequency distribution toward lower scores lends credibility to the use of the contact graph as a prior for discovering an evolutionarily conserved network.

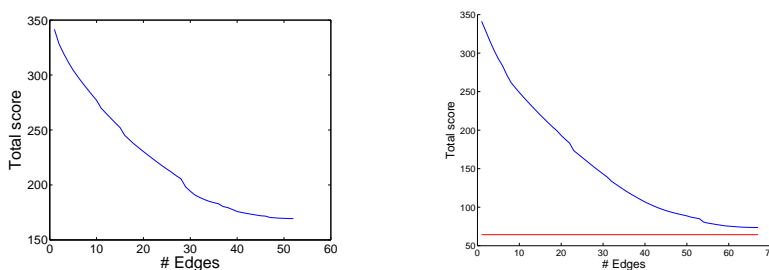


Figure 4: Improvement of  $MI$  score as edges are successively added for the contact graph prior (left) and uninformative prior (right). The red line shows a lower bound for the score for the GPCR MSA.

these caveats in mind in the discussion that follows.

Our first model inference test was to start with the previously identified network [24], use its induced subgraph of the contact graph as input to our algorithm, and see if it recovers the network. The algorithm considers 144 edges and picks 52 of them for inclusion in the model. The top of Fig. 5 (left) illustrates the network identified by our algorithm. Fig. 4 (left) shows the change in score as edges are added to our network. Notice the score decreases as edges are added and levels out toward the end. When the algorithm completes, there are no edges that reduce the score.

To study the influence of the contact graph prior, we re-ran our algorithm using an uninformative prior so that all pairs of residues are now tested for inclusion. This time, the algorithm considers 1080 edges and picks 67 of them for inclusion. As Fig. 4 (right) shows, the algorithm produces a network that has a better score than the one produced by the contact graph prior but, unfortunately, does not have as nice a visualization (Fig. 5 (right)).

Since the score differences between these two runs were substantial, we investigated the best possible score achievable on this protein family. Towards this end, we randomly shuffled the columns of the MSA, yielding a new MSA having the same level of conservation for each residue but with correlation lost due to the independent shuffling. We measured the correlation in 2500 of these MSAs (which consisted of just noise) by computing the score of the empty network (one with no edges) on the MSA. The resulting scores were normally distributed over a small range (63.5 to 65.1) with mean value 64.3. This means that for the GPCR family, if we accounted for all possible

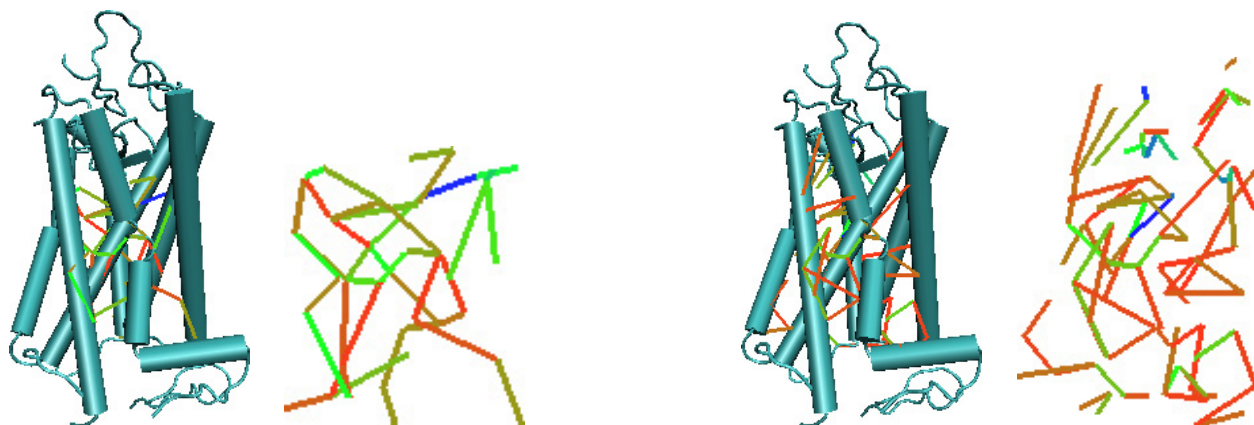


Figure 5: GPCR network identification: three-dimensional structure of bovine rhodopsin with overlaid network, and just the network for model inferred from (left) contact graph induced by the network of [24] and (right) entire contact graph. Edges are colored by score, with red the strongest ‘decouplers’ and blue the weakest.

correlation we would expect a score of about 64.3. The algorithm run with the uninformative prior scores 73.6, well within the margin of error we would expect due to the greedy property of our algorithm or the nature of the conditioning contexts.

While our modeling formulation is different in nature from that of Suel *et al.* (independence vs. dependence), our model identifies many of the same biologically relevant features. For example, Suel *et al.* identify coupling between residues 296 and 265 that form “part of a linked network extending parallel to the plasma membrane from 296 to form the bottom of the ligand-binding pocket.” Our algorithm likewise identifies an edge between residues 296 and 265. Several other identified interactions appear as *indirect* relationships in our model. For example, coupling between residue 296 and 293, identified as a “helical packing interaction” is identified by our model as being indirect. In this case, residue 117 actually makes residues 296 and 293 conditionally independent, lowering their mutual information scores from .3347 to .0259. This is true also of the coupling between residue 296 and residues 298 and 299. These couplings are part of “a sparse but contiguous network of inter-helical interactions linking the ligand-binding pocket with the cytoplasmic surface.” Both 296/298 and 296/299 become conditionally independent in the presence of residue 117.

Although our algorithm does produce many of the relationships as identified by Suel *et al.*, there are several differences between the models. For instance, our network does not identify the coupling between residues 296 and 113 which “makes a salt-bridge interaction with the protonated form of the Schiff base,” as either direct or indirect. Nor does our algorithm find the “inter-helical packing interaction” between residues 296 and 91. Conversely, our algorithm finds a strong direct coupling between residues 296 and 117 as well as between residues 90 and 91. Further investigation into these strong couplings may be of interest to biologists (e.g., by mutagenesis studies). This illustrates the ability of our approach to help formulate testable biological hypotheses.

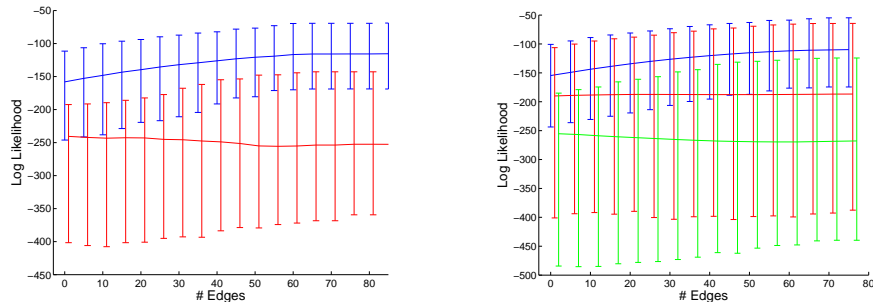


Figure 6: Evolution of likelihood as edges are added to the network. (Left) Sequences from class I (blue) and class II (red) against the class I model. (Right) Sequences from training class I (blue), testing class I (red), and class II (green) against the class I model. Each plot shows the mean, maximum and minimum likelihood.

## PDZs

In this case study, we demonstrate the utility in subsequent analyses of graphical models learned by our algorithm. In particular, we study the ability of our inferred models to capture the ‘essense’ of a protein, namely in classifying PDZ domains. Although the two classes in this protein family may be defined by simple sequence motifs, we show that coupling-based models provide more discriminatory power, and we use this opportunity to subject our approach to a rigorous evaluation in a maximum likelihood framework.

We obtained MSAs for the two classes of PDZ domains from PDZBase [2] by querying according to the ligand and removing duplicate entries, thereby obtaining 95 class I and 14 class II sequences. We ran our algorithm on the sequences in class I and compared the likelihoods from proteins in class I and II against the model. Fig. 6 (left) shows the evolution of likelihood scores as edges are added to our model. On the far left is the likelihood based solely on conservation (i.e., with no edges in the network). As the network grows, so does its power to discriminate classes—conservation alone does not adequately represent the multiple sequence alignment. In the limit, we would derive a clique, with a joint distribution over all residues that would provide a reasonable score only for sequences in the original alignment.

To avoid over-fitting, we adopted a cross-validation approach. We randomly selected 2/3 of the class I multiple sequence alignment data with which to construct a model. We then computed the likelihood of the remaining sequences against the model so inferred and compared it to likelihoods of the sequences used for model building and the sequences from class II. We repeated this experiment 500 times; Fig. 6 (right) shows the average of the results. As is clear, sequences in the model (blue) have the highest likelihoods while the sequences in class I but not in the model (red) have higher likelihoods than those in class II (green).

## 5 Discussion

This work marries research into residue co-variation with probabilistic graphical models, producing a systematic and sound algorithmic approach to inferring networks underlying protein families. Our use of conditional mutual information as a criterion for growing a network means that our algorithm can also be viewed as a perturbation-based approach; however, in contrast to [24] who infer coupling between the perturbed position and another position, we infer independence between residues on either side of the perturbed position. The results indicate that independence of residues can be a good guiding principle for the discovery of evolutionarily conserved structure.

While there are other ways to infer networks from covariation data (e.g., gaussian graphical models [5]) they either require the specification of complete sets (e.g., all pairs) of dependency information or must necessarily make assumptions about the parametric form of interrelationships. In contrast, our approach employs the broader notion of independencies to situate the network. In addition, it models *all* significant couplings and conditional independencies, hence capturing the essence of what it means to belong to a given family. This has tremendous applications in protein fold classification and protein design.

There are several extensions to the work proposed here. First, we would like to scale up our algorithms to work with MSAs involving greater numbers of sequences. There have been systematic approaches proposed to scaling up graphical model inference algorithms [12] and we propose to consider these for inferring coupled residues. Second, we would like to relax our modeling of residues as distributions over amino acids, and instead consider distributions over *classes* of amino acids (e.g., polar, hydrophobic, small). Since there are multiple, overlapping, taxonomies of amino acids [25] we can even assume a hidden variable model (denoting an unknown relabeling of each residue) and attempt to infer the network as well as the relabeling function from a given MSA and contact map. An alternative is to employ a scoring matrix in evaluating extent of co-variation [22]. Yet another form of relaxation involves modeling multiple families simultaneously, perhaps by superposition of Gaussian processes, each with its own graphical model. Finally, we intend to explore the applications to protein analysis and design made possible by our graphical model.

## Acknowledgments

CBK and JT are supported in part by a CAREER award from the National Science Foundation (IIS-0444544). NR is supported in part by NSF grants IBN-0219332 and EIA-0103660.

## References

- [1] W.R. Atchley, W. Terhalle, and A. Dress. Positional Dependence, Cliques, and Predictive Motifs in the bHLH Protein Domain. *Journal of Molecular Evolution*, Vol. 48:501–516, 1999.
- [2] T. Beuming, L. Skrabanek, M.Y. Niv, P. Mukherjee, and H. Weinstein. PDZBase: A Protein-Protein Interaction Database for PDZ-Domains. *Bioinformatics*, Vol. 21(6):827–828, 2005.
- [3] W.L. Buntine. Operations for Learning with Graphical Models. *Journal of Artificial Intelligence Research*, Vol. 2:159–225, 1994.

- [4] J.P. Dekker, A.A. Fodor, R.W. Aldrich, and G. Yellen. A Perturbation-Based Method for Calculating Explicit Likelihood of Evolutionary Co-Variance in Multiple Sequence Alignments. *Bioinformatics*, Vol. 20(10):1565–1572, 2004.
- [5] M. Drton and M.D. Perlman. Model Selection for Gaussian Concentration Graphs. *Biometrika*, Vol. 91(3):591–602, 2004.
- [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*, Vol. 56:211–221, 2004.
- [7] A.A. Fodor and R.W. Aldrich. On Evolutionary Conservation of Thermodynamic Coupling in Proteins. *Journal of Biological Chemistry*, Vol. 279(18):19046–19050, Apr 2004.
- [8] N. Friedman, I. Nachman, and D. Peer. Learning Bayesian Network Structure from Massive Datasets: The “Sparse Candidate” Algorithm. In *Proceedings of the Fifteenth International Conference on Uncertainty in Artificial Intelligence (UAI’99)*, pages 206–215, 1999.
- [9] I.V. Grigoriev and S.-H. Kim. Detection of Protein Fold Similarity Based on Correlation of Amino Acid Properties. *Proceedings of the National Academy of Sciences, USA*, Vol. 96(25):14318–14323, Dec 1999.
- [10] B.Z. Harris and W.A. Lim. Mechanism and Role of PDZ Domains in Signaling Complex Assembly. *Journal of Cell Science*, Vol. 114:3219–3231, 2001.
- [11] A. Horovitz. Double-Mutant Cycles: A Powerful Tool for Analyzing Protein Structure and Function. *Fold. Des.*, Vol. 1:R121–R126, 1996.
- [12] G. Hulten and P. Domingos. Mining Complex Models from Arbitrarily Large Databases in Constant Time. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’02)*, pages 525–531, 2002.
- [13] A.Y. Hung and M. Sheng. PDZ Domains: Structural Modules for Protein Complex Assembly. *Journal of Biological Chemistry*, Vol. 277(8):5699–5702, Feb 2002.
- [14] I. Kass and A. Horovitz. Mapping Pathways of Allosteric Communication in GroEL by Analysis of Correlated Mutations. *Proteins: Structure, Function, and Genetics*, Vol. 48:611–617, 2002.
- [15] B.T.M. Korber, R.M. Farber, D.H. Wolpert, and A.S. Lapedes. Covariation of Mutations in the V3 Loop of HIV Type 1 Envelope Protein: An Information Theoretic Analysis. *Proceedings of the National Academy of Sciences, USA*, Vol. 90:7176–7180, Aug 1993.
- [16] S.W. Lockless and R. Ranganathan. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science*, Vol. 286(5438):295–299, Oct 1999.
- [17] D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [18] M. Milik, S. Szalma, and K.A. Olszewski. Common Structural Cliques: A Tool for Protein Structure and Function Analysis. *Protein Engineering*, Vol. 16(8):542–552, 2003.

- [19] O. Olmea, B. Rost, and A. Valencia. Effective Use of Sequence Correlation and Conservation in Fold Recognition. *Journal of Molecular Biology*, Vol. 295:1221–1239, 1999.
- [20] W.P. Russ and R. Ranganathan. Knowledge-Based Potential Functions in Protein Design. *Current Opinion in Structural Biology*, Vol. 12:447–452, 2002.
- [21] W.S. Sandberg and T.C. Terwilliger. Engineering Multiple Properties of a Protein by Combinatorial Mutagenesis. *Proceedings of the National Academy of Sciences, USA*, Vol. 90(18):8367–8371, Sep 1993.
- [22] M.C. Saraf, G.L. Moore, and C.D. Maranas. Using Multiple Sequence Correlation Analysis to Characterize Functionally Important Protein Regions. *Protein Engineering*, Vol. 16(6):397–406, 2003.
- [23] O. Schueler-Furman and D. Baker. Conserved Residue Clustering and Protein Structure Prediction. *Proteins: Structure, Function, and Genetics*, Vol. 52:225–235, 2003.
- [24] G.S. Suel, S.W. Lockless, M.A. Wall, and R. Ranganathan. Evolutionary Conserved Networks of Residues Mediate Allosteric Communication in Proteins. *Nature Structural Biology*, Vol. 10:59–69, Jan 2003.
- [25] W.S.J. Valdar. Scoring Residue Conservation. *Proteins: Structure, Function, and Genetics*, Vol. 48:227–241, 2002.
- [26] C.A. Voigt, C. Martinez, Z.-G. Wang, S.L. Mayo, and F.H. Arnold. Protein Building Blocks Preserved by Recombination. *Nature Structural Biology*, Vol. 9(7):553–558, Jul 2002.