Graphical Models of Residue Coupling in Protein Families

John Thomas, Naren Ramakrishnan, and Chris Bailey-Kellogg

Abstract—Many statistical measures and algorithmic techniques have been proposed for studying residue coupling in protein families. Generally speaking, two residue positions are considered coupled if, in the sequence record, some of their amino acid type combinations are significantly more common than others. While the proposed approaches have proven useful in finding and describing coupling, a significant missing component is a formal probabilistic model that explicates and compactly represents the coupling, integrates information about sequence, structure, and function, and supports inferential procedures for analysis, diagnosis, and prediction. We present an approach to learning and using probabilistic graphical models of residue coupling (GMRCs). These models capture significant conservation and coupling constraints observable in a multiply aligned set of sequences. Our approach can place a structural prior on considered couplings, so that all identified relationships have direct mechanistic explanations. It can also incorporate information about functional classes, and thereby learn a differential graphical model that distinguishes constraints common to all classes from those unique to individual classes. Such differential models separately account for class-specific conservation and family-wide coupling, two different sources of sequence covariation. They are then able to perform interpretable functional classification of new sequences, explaining classification decisions in terms of the underlying conservation and coupling constraints. We apply our approach in studying both G protein-coupled receptors and PDZ domains, identifying and analyzing family-wide and class-specific constraints, and performing functional classification. The results demonstrate that GMRCs provide a powerful tool for uncovering, representing, and utilizing significant sequence-structure-function relationships in protein families.

Index Terms—Correlated mutations, graphical models, evolutionary covariation, sequence-structure-function relationships, functional classification.

1 Introduction

Functional pressures on proteins constrain their sequences and three-dimensional (3D) structures. Constraints thus manifested in sequence-structure-function relationships can be inferred from the evolutionary record, along with information from available structural studies and functional assays. Identified relationships can then be employed in all different "directions," e.g., to predict function from the sequence of a newly discovered protein, discriminate predicted structures for a sequence according to functional tests, and design variant (homologous) protein sequences with related functions.

This paper develops a new approach, based on undirected graphical models, for modeling and exploiting a particular type of sequence-structure-function relationship: residue coupling.

1.1 Residue Coupling

While amino acid conservation has long been recognized as an important indicator of structural or functional significance, many recent studies have generalized this notion to include *coupling* of amino acid pairs in a protein family. Typically, two residue positions are considered to be coupled if the amino acids occurring in these positions show concerted variation, e.g., one residue occurs as either A or S, the other occurs as either L or C, but taken together, they do not occur in all of the four possible combinations. A chosen metric (e.g., mutual information or correlation, among many others [8]) is used to quantify the degree to which two residues in the family covary, presumably due to compensating mutations in the face of cooperativity. Couplings thus identified have been used in a variety of applications, e.g., to identify protein-specific motifs [2], [24], map allosteric pathways [16], [36], predict protein structures [10], [27], [34] and interactions [13], [29], and design new peptide vaccines [17]. Ranganathan et al. demonstrated that coupling information enables the design of new, stably folded [35], and functional [32] WW domains.

Our goal is to represent coupling in a formal probabilistic model, in order to better support investigation, characterization, and design. Just as a hidden Markov model (HMM) provides a probabilistic basis for reasoning about sequence *conservation*, we aim to provide a probabilistic basis for reasoning about sequence *covariation*. Just as an HMM makes explicit the structure of position-specific amino acid distributions [7], we aim to make explicit the factorization of covariation within a family into a small set of dependencies. Just as the Markov property exposes the direct relationship of a residue on only the preceding one, we aim to distinguish direct couplings from indirect ones. Just as key conservation and variation within a protein family can be identified by examining amino acid distributions and transition probabilities in an HMM, we seek to construct models from which essential coupling constraints can

- *J. Thomas and C. Bailey-Kellogg are with the Department of Computer Science, Dartmouth College, 6211 Sudikoff Laboratory, Hanover, NH 03755. E-mail: {jthomas, cbk}@cs.dartmouth.edu.*
- *N. Ramakrishnan is with the Department of Computer Science, Virginia Tech, 660 McBryde Hall, Blacksburg, VA 24061. E-mail: naren@cs.vt.edu.*

Manuscript received 3 Nov. 2006; revised 1 Feb. 2007; accepted 3 Apr. 2007; published online 21 June 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0198-1106.
Digital Object Identifier no. 10.1109/TCBB.2007.70225.

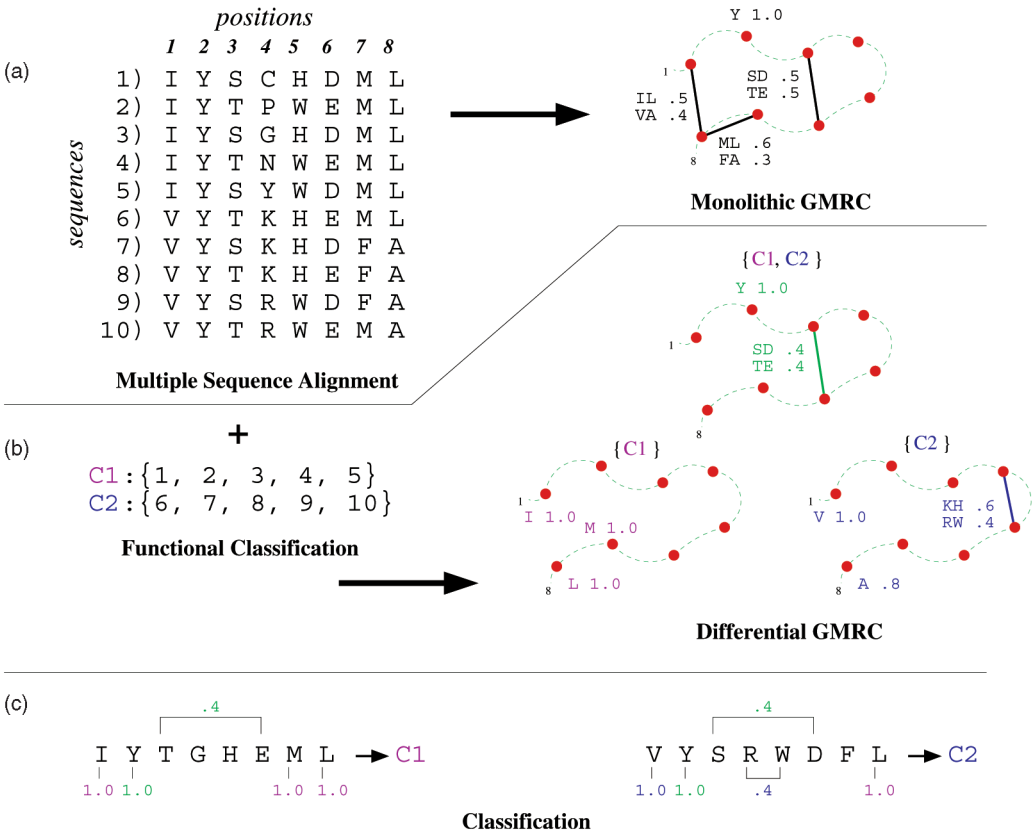


Fig. 1. Schematic illustration of GMRCs. (a) Given an MSA for a protein family, a monolithic GMRC captures the underlying constraints (conservation and coupling) acting on the family. Here, the dashed lines form the backbone of a hypothetical structure, from 1 (N terminus) to 8 (C terminus), and the solid lines are the edges between coupled residues. (b) After assigning each sequence in the alignment to its functional class (sets of indices in the MSA), a differential GMRC separately identifies constraints that are common to all functional classes and constraints that are unique to individual functional classes (or particular subsets of classes). This allows it to, for example, separately account for family-wide coupling (e.g., between positions 3 and 6 in the example) and class-specific conservation (e.g., at positions 1 and 8), both of which appear as a covariation in the MSA. (c) A differential GMRC provides a transparent mechanism for classifying sequences whose functional class is unknown.

readily be extracted. Such significant constraints captured by a model are particularly useful in explaining where and how well another protein fits or does not fit the model. They also enable the formulation of hypotheses for further experimental study, e.g., by site-directed mutation (conservation) and double mutation (coupling).

We develop here the first approach that meets all of these goals: *graphical models of residue coupling (GMRCs)*.[1] Fig. 1a illustrates how a GMRC captures residue coupling from a multiple sequence alignment (MSA). (For reasons that will be clearer later, we call this GMRC *monolithic*, as it treats all sequences in the MSA the same.) The GMRC has nodes representing residues and edges (solid lines) representing the conditional dependence structure. One can "read" such a model as saying "a node is conditionally independent of all other nodes given its neighbors." Thus, residue 3 is independent of all other residues when we know the amino acid type at residue 6. We have labeled the model with some of the conservation constraints observed for the individual residues and some of the coupling constraints observed for the pairs of residues. For example, we see in the MSA that residue 2 is a completely conserved Y (100 percent of the sequences have a Y at position 2) and residues 3 and 6 covary (when 3 is S, 6 is D and when 3 is T, 6 is E). The frequency of the covarying pairs is shown next to the corresponding edge in the graph. Some covariation is captured in the edges, but some is explained indirectly. For example, the covariation between residues 1 and 7 can be explained by the covariation between residues 1 and 8 and that between residues 7 and 8. This model thus has 1-8 and 7-8 edges but leaves the 1-7 relationship as indirect. Our approach explicitly represents a compact set of direct couplings and implicitly (transitively) represents the remaining residue relationships.

1. Our preliminary work, presented at the 2005 BioKDD workshop [37] first formulated some of the basic ideas that we build on here.

1.2 Sequence-Structure Function

While "coupling" connotes a mechanistic relationship, observed residue covariation can have a variety of sources. Many studies have tested the extent of correlation between covariation and structural contact [8], [9], [27]. Our graphical model approach can incorporate structural information, if desired, to limit the edges considered for addition to a model (e.g., only between residues that are in contact).

Phylogeny can play a significant role in observed sequence covariation [28], [30]. Consider two scenarios. In the first, a pair of positions has one pair of amino acid types

in one subtree and another pair of amino acid types in another subtree (of about the same size as the first). In the second, the covariation in the two pairs of amino acid types is dispersed throughout the tree. Considering just the extant sequences, both scenarios yield the same amount of covariation, although it could be due just to two mutation events in the first scenario and multiple ones in the second. Techniques have thus been developed to account for such diverse origins of covariation using a phylogenetic tree, e.g., by performing maximum likelihood inference of an independent evolutionary model versus a coevolutionary one [31]. One can then be more confident that couplings identified as coevolutionary are due to a compensatory process. Since these methods are affected by the choice of phylogenetic tree and these trees are hard to construct, other methods stochastically build trees to detect correlated mutations [6], [25].

While the above discussion deals with the confounding role of phylogeny, techniques such as Evolutionary Trace [20] and ConSurf [1] exploit phylogenetic information in order to identify possible functional sites. The idea is that those residues conferring functional variation will be differentially conserved in different subtrees (e.g., a residue might be an invariant Arg in one subtree and an invariant Lys in another). A spatial relationship of such residues with respect to a known 3D structure is then inferred to be indicative of an active site. An alternative to explicit phylogeny is demonstrated by CASTOR [21], [22], which seeks to identify important functional regions by discovering hierarchical motifs. The approach is unsupervised, in that it does not require a tree or functional information as input but rather uses the recursive identification of motifs to determine a hierarchy and infer functional clusters and their defining sequence patterns.

We focus here on the incorporation within probabilistic graphical models of limited coarse-grain functional information, represented as class labels for the sequences. A *differential* GMRC distinguishes constraints (coupling and conservation) common to all sequences from those that are specific to particular classes. In the example in Fig. 1b, residue 2 is a conserved Y in both functional classes, while the conservation of M at residue 7 is unique to functional class C_1. Further, the coupling between residues 3 and 6 is common to both functional classes, while the coupling between residues 4 and 5 is unique to functional class C_2. Note that class-specific conservation and family-wide coupling have similar appearances in the sequence record; with functional class labels, we can see that residues 1 and 7 are highly conserved in each functional class, but without the labels, it appears that residues 1 and 7 are coupled. The differential model separates these two forms of covariation.

Thus, in contrast to the other work discussed above, our approach does not make explicit use of phylogenetic information and does not attempt to explain the evolutionary history of how residue covariation manifests in the sequence record, is supervised in its ability to profile functional classes using patterns among residues but unsupervised in the sense of not requiring specification of the "true" underlying coupling, and focuses on capturing information about coupling rather than conservation alone. It remains interesting in future work to pursue various combinations of these approaches (e.g., restricting our modeled coupling according to phylogenetic models or hierarchically uncovering coupling models).

1.3 Application

Our graphical models serve as compact descriptions of joint amino acid distributions, and support a variety of applications, including 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?). Here, we develop the use of differential GMRCs for *transparent* classification of sequences (illustrated in Fig. 1c). Given a new sequence, a GMRC computes a score that indicates how likely it is that a sequence belongs to a functional class. The score is based on transparent properties of the model for the functional class: we can examine the model to explain the classification decisions in terms of conservation and coupling constraints. For example, we assign the left sequence to functional class C_1 because it has many of the important features of that class, as identified by the differential GMRC. The right sequence has features that are important to both functional classes but has more features important only for functional class C_2. We note that the differential construction is necessary for this type of interpretation. There is not a unique encoding of independencies; multiple different monolithic models can represent exactly the same information. Thus, if we simply compared separately learned monolithic models, the differences in their constraints would not necessarily indicate class-specific features.

We demonstrate the power of our GMRC approach in the analysis of the G protein-coupled receptor family (GPCRs). Since GPCRs are vital in many signaling processes such as vision, smell, and mood regulation, they are a significant target for molecular modeling and drug discovery. Our graphical models make explicit the essential constraints underlying such a family, identifying and modeling a small set of couplings that explain the observed sequence covariation nearly as well as a complete list of covarying residues and that are deemed statistically significant. We illustrate some of the essential constraints learned by our models of GPCRs. Our algorithm can combine multiple information sources, e.g., by integrating priors from structural and functional studies with sequence analysis, so as to predispose a search toward couplings that conform to a priori background knowledge. We show that a differential model of GPCRs is able to use functional information to identify and model coupling and conservation common to various GPCR functional classes. We also show that the models are able to perform interpretable classification of GPCR sequences, assigning credit (or blame) for the classification to significant coupling and conservation constraints.

As a further validation, we demonstrate the ability of our models to learn and utilize significant constraints for two functional classes of PDZ domains. Many PDZs are involved in protein recognition for complex formation and are classified by characteristics of their recognized ligands. We show that GMRCs can uncover commonalities and

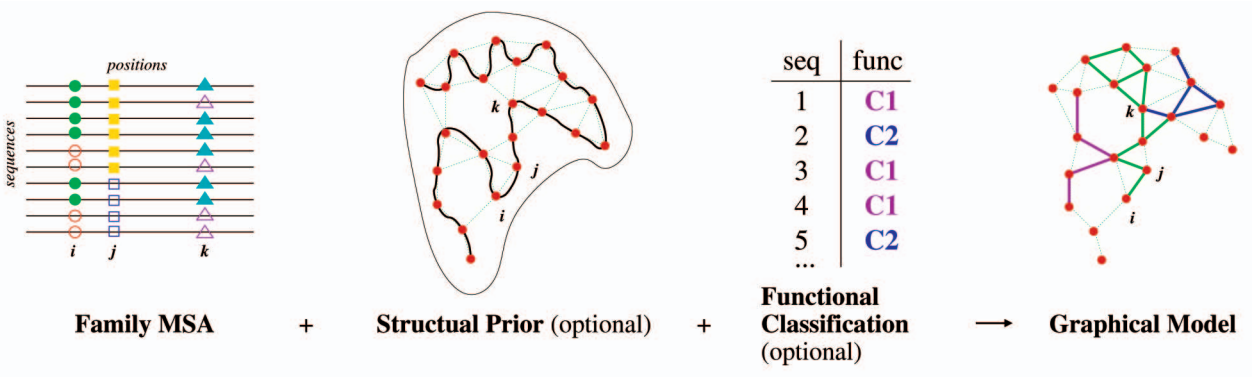


Fig. 2. Learning GMRCs. Evidence for evolutionary constraints is found in an MSA for members of the protein family. This evidence from sequence information is the basis for learning a graphical model (far right; thick solid edges), which captures conditional independence among residues. If desired, the graphical model may also take both structural and functional information into account. Structural constraints are expressed in terms of a prior over possible coupling relationships, shown here in terms of the contact graph (pairs of "close enough" residues). Functional information is represented in a table classifying the functionality of the sequences (here, simply class C1 or class C2). In this case, a base model common to both C1 and C2 (green edges) is extended by differential models for the separate functional classes (magenta and blue edges), capturing class-specific relationships.

differences in both conservation and coupling for members of two such classes, in a manner enabling sequence-based functional classification.

2 Methods

Fig. 2 overviews how our method uses information about sequence, structure, and function, in order to learn a graphical model of residue coupling (GMRC). Central to our approach is information about conservation and covariation within a protein family, captured in an MSA. The inferred graphical model captures conditional dependence and independence among residues, as revealed by the MSA. If desired, mechanistic constraints according to a representative 3D structure (presumed conserved across the family by homology) can be incorporated as a prior on considered couplings; otherwise, an uninformative prior can be employed. A model can be learned for a family as a whole (monolithic) or can be used to analyze distinguishing characteristics of functional classes (differential). In the latter case, we construct the differential graphical model by first building a model of conservation and coupling common to an entire family and then extending it with separate models for subsets of the functional classes, thereby identifying what distinguishes them in terms of conservation and coupling. Once learned, a differential graphical model enables functional classification of other family members according to their likelihood under the model.

2.1 Detecting Coupled Residues

An MSA $\mathcal{S}$ allows us to summarize each residue position in terms of the probabilities of encountering each of the 20 amino acids in that position. Let $V = \{v_1, \ldots, v_n\}$ be a set of random variables, one for each residue position. The MSA then gives a distribution of amino acid types for each. Coupling between these random variables can be quantified by many statistical and information-theoretic metrics [8]. While our methods could use any such metric, the presented results are based on a "perturbation" variant of conditional mutual information, since it has many desirable qualities, and we find that it produces high-quality models.

The traditional definition of *mutual information* $MI(v_i, v_j)$ between residues i and j follows:

$$MI(v_i, v_j) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} P(v_i = a, v_j = b) \cdot \log \frac{P(v_i = a, v_j = b)}{P(v_i = a)\ P(v_j = b)}, \quad (1)$$

where $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{D}, \ldots\}$ is the set of amino acid types, and the probabilities are all assessed from the given sequences $\mathcal{S}$ (see below).

In their SCA analysis, on the other hand, Lockless and Ranganathan [23] introduced a perturbation-based estimator that first subsets the MSA according to some condition (here, containing a moderately conserved residue type at one position); the effects of this perturbation on the residue distribution at another position is observed. If the perturbation significantly alters the proportions of amino acids at the observed position, the latter is inferred to be coupled to the perturbed position, according to the evolutionary record. However, note the asymmetry in this definition wherein we could detect coupling between v_i and v_j when v_i is the perturbed and v_j is the observed but not necessarily the other way around.

We adopt the basic idea of Lockless and Ranganathan but preserve the symmetry of the traditional $MI(v_i, v_j)$ definition by employing the notion of conditional mutual information:

$$MI(v_i, v_j | v_k) = \sum_{c \in \mathcal{A}^*} P(v_k = c) \left[\sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} P(v_i = a, v_j = b | v_k = c) \cdot \log \frac{P(v_i = a, v_j = b | v_k = c)}{P(v_i = a | v_k = c)\ P(v_j = b | v_k = c)} \right], \quad (2)$$

where we estimate the conditionals by subsetting residue k to its most frequently occurring amino acid types ($\mathcal{A}^* \subset \mathcal{A}$), defined as those that appear in at least 15 percent of the original sequences in the subset. As discussed [23], such a bound is required in order to maintain fidelity to the original

MSA and allow for evolutionary exploration. We also ensure that $P(v_k = c)$ distributes a probability mass of 1 among just these indices, in proportion to the number of sequences in each subset, so that $\sum_{c \in \mathcal{A}^*} P(v_k = c) = 1$.

Note that the perturbation limits (and the key difference from straight mutual information) imply both a necessary and sufficient degree of conservation for the detected coupling to be meaningful. As with traditional mutual information, this measure quantifies the error in assuming that a joint distribution is decomposable and is zero when the underlying distributions are independent and nonzero otherwise.

2.2 Summarizing Coupling and Conservation in Graphical Models

The traditional way to summarize coupling data is to seek out relationships between residue pairs with high MI (and, hence, high covariation). The problem with this approach is that it does not explicitly separate direct from indirect relationships. In contrast, we look for relationships with low or near zero MI and encode the independencies they represent in an undirected *GMRC*, $G = (V, E)$. The vertices V are the residue random variables, as above. The edges E encode probabilistic independence constraints, such that a vertex is conditionally independent of all other vertices, given its immediate neighbors. Thus, a GMRC represents a factorization of the joint probability distribution function (pdf) of the residue random variables [5].

More formally, a GMRC $G = (V, E)$ defines a pdf $P_G(R)$ on residue types R for its vertices V by the way of the conditional relationships in its edges E. To compute the pdf, we only need to combine the scores ("potentials") for all the cliques in the graph:

$$P_G(R) = \frac{1}{Z} \prod_{C \in \text{cliques}(G)} \phi_C(R_C). \tag{3}$$

Here, R is a set of amino acid types for V, and R_C denotes values for only those vertices in clique C. The ϕ_C are potential functions for the cliques, and Z normalizes their product into a probability measure. The structure of the potential functions satisfies:

$$\prod_{C \in \text{cliques}(G)} \phi_C(R_C) = \frac{\prod_C P_C(R_C)}{\prod_{A \in \text{cliqueadj}(G)} P_A(R_A)}. \tag{4}$$

Notice that the potentials are given by the product of marginals defined over the cliques divided by the product of marginals defined over the clique adjacencies A, which could be nodes, edges, or general subgraphs. In this view, each potential of (3) is either a conditional or a joint marginal distribution.

For the models to faithfully capture both the conditional independencies and the joint factorization in terms of potentials, the Hammersley-Clifford theorem [18] necessitates positivity of the pdf everywhere. The positivity assumption is likely to be violated in real contexts since it is unlikely that the MSA is sufficiently representative of every possible clique value in the model. This is a well-studied problem in statistical inference, and several approaches exist to apportion the nonzero probability mass among nonoccurring instance values [15]. We adopt the following estimator:

$$P_C(R_C) = \frac{f_C(R_C) + \frac{\rho |\mathcal{S}|}{21^{|C|}}}{|\mathcal{S}|(1+\rho)}. \tag{5}$$

R_C is again a set of residue types for a clique C, $f_C(R_C)$ is the number of times R_C is encountered for the clique vertices in the MSA, $|\mathcal{S}|$ is the total number of sequences in the MSA, $|C|$ is the cardinality of C, and ρ is a parameter that weights the importance of missing data. Notice that even when a particular clique value does not appear in the MSA, it still has a positive (but small) probability, thereby enabling the factorization according to the Hammersley-Clifford theorem [18].

Note that a graphical model captures both conservation and coupling constraints, since it uses pointwise, as well as joint probabilities to factorize distributions. In this sense, our models generalize traditional motif-based approaches to characterizing protein sequences.

2.3 Learning a Graphical Model

To infer a model that affords the above interpretation, we sequentially find *decouplers*, sets of residues that help make other residues independent. For example, in the MSA in Fig. 2, positions i and k are very correlated—when i is a "filled-in" residue, k tends to be as well (5 out of 6 times); similarly, when i is "empty," k tends to agree (3 out of 4 times). However, knowing j makes the positions rather independent. In the most common case, where j is filled in, we see that the combinations of types at i and k are more evenly distributed—when i is "empty," k is "empty" once and "filled in" once; similarly, when i is "filled in," k is "empty" once and "filled in" three times. This suggests that i and k are conditionally independent given j; j decouples i and k. (Of course, even in this example, noise obscures the degree of independence.) We thus construct a model by selecting the edges that best decouple other residue relationships. In our example, i-j and j-k edges decouple i and k (any coupling between i and k is explained transitively by the direct i-j and j-k edges).

Further, we assume the availability of a prior that could, for instance, be based on a contact graph for a representative member of the family (center of Fig. 2). Such a prior places edges between all pairs of residues that are interacting (e.g., because some atoms are within a distance threshold) in the 3D structure of the protein. Alternatively, the prior could be a graph accounting for coupling via an intermediate (ligand binding), long-range electrostatics, or simply an uninformative one (assumes all edges are equally likely). Given a prior, we sequentially assess conditional independencies from this set and incrementally combine the decoupling edges into a graphical model.

Our algorithm (Fig. 3) greedily grows a graph by, at each step, selecting the edge from the list of possible edges (defined by the prior) that scores best with respect to the current graph. The score is given by

function *monolithicGMRC*
input D: possible edges
input $\mathcal{S}$: multiple sequence alignment

$V = \{v_1, \ldots, v_n\}$; $E \leftarrow \emptyset$
$s \leftarrow \text{Score}(G = (V, E))$
$C \leftarrow \{(e, s - \text{Score}(G = (V, \{e\}))|e \in D)\}$
repeat
  $e \leftarrow \arg\max_{e \in D-E} C(e)$
  if e is significant **then**
    $E \leftarrow E \cup \{e\}$
    $s \leftarrow s - C(e)$
    for all $e' \in D - E$ s.t. e and e' share a vertex **do**
      $C(e') \leftarrow s - \text{Score}(G = (V, E))$
    end for
  end if
until stopping criterion satisfied

return $G = (V, E)$

Fig. 3. Algorithm for inferring a GMRC.

$$\text{Score}(G = (V, E)) = \sum_{v \in V} \sum_{u \notin \text{neighbors}(v)} MI(u, v|\text{neighbors}(v)). \quad (6)$$

The algorithm can be configured to utilize various 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. In this paper, we select a candidate edge that both improves the score and retains a significant portion of the sequences in subsetting for MI calculations, so that we may be confident about independence assessments. Further, to ensure that our model does not overfit the coupling relationships, we require that each edge be statistically significant (see below). Edges that are significant are added to the graph, those that are not are rejected. We terminate when no statistically significant edge remains that reduces the score further.

At each iteration, a naïve implementation of Fig. 3 would require $O(dn^2)$ MI computations, where n is the number of residues in the family, and d is the maximum degree of nodes in the prior. With an uninformative prior, d is $O(n)$, resulting in $O(n^3)$ MI computations for each iteration. By caching the assessments of conditioning contexts, as described in [37], and careful preprocessing, we can bring down the complexity to just $O(dn)$ MI computations per iteration, yielding a speedup of $O(n)$.

2.4 Statistical Significance

When learning a graphical model, a common approach is to compute how significant it is given the data. However, in the context of a GMRC, we are more interested in individual edges (for both biologically interesting couplings and for transparent classification) than in the model as a whole. Thus, it is more appropriate to focus on the significance of individual edges. Since our algorithm for learning a GMRC adds a single edge at a time, we can compute the statistical significance of each edge at runtime and reject edges that are not significant. This allows us to search for alternate significant decouplers and to avoid overfitting the data. Fig. 3 incorporates the statistical significance test as a filter on edges before adding them to the growing network. The idea is that the score function proposes possible edges that account for a large part of the remaining coupling in the network, and the statistical significance test only passes those that are significant. Those that are not significant are discarded and not reconsidered.

An edge in a GMRC indicates a direct relationship between two residues. To compute the significance of an edge, we use a p-value. Intuitively, the p-value gives us the probability that the null hypothesis is true—that two edges are truly independent rather than coupled. Smaller p-values indicate stronger confidence in dependent relationships, while larger p-values mean that the residues are most likely independent. For our algorithm, we use the p-value from a χ-squared test. For two residues i and j, we compute the χ^2 as

$$\chi^2 = \sum_{a \in \mathcal{A}_i} \sum_{b \in \mathcal{A}_j} \frac{\left(f_{\{i,j\}}(\{a, b\}) - \frac{f_{\{i\}}(\{a\}) \cdot f_{\{j\}}(\{b\})}{|\mathcal{S}|^2}\right)^2}{\frac{f_{\{i\}}(\{a\}) \cdot f_{\{j\}}(\{b\})}{|\mathcal{S}|^2}}. \quad (7)$$

Here, $f_C(R_C)$ is, as in (5), the number of occurrences of residue types R_C at positions C. The first term in the numerator is the actual number of pairs observed; the second term is the expected number, if the two residues were independent.

A potential concern for such a χ-squared test is that it can be unreliable when there are small counts of the terms it is evaluating. In such a case, the χ^2 can become artificially inflated, making relationships appear more significant than they are. However, this is not the case here, since our algorithm only considers edges between residues that have sufficient representation. As we discussed, we require perturbations to maintain sufficient fidelity to the original MSA, thereby ensuring that only edges with large counts are considered by the algorithm.

2.5 Differential Models

When we have more than one functional class, we learn a *differential* GMRC. Instead of the graphical model being only "family-wide" or for a single "class-specific" functional class, a differential model captures constraints over all subsets of functional classes. A differential graphical model is arranged in a lattice. For instance, if we have three functional classes, C_1, C_2, and C_3, the graphical model at the top level of the lattice corresponds to constraints that act on all of the functional classes. We denote this graphical model by G_{C_1,C_2,C_3}. As we move down the lattice, we capture constraints that act on some functional classes but not on others. For instance, G_{C_1,C_2} captures the constraints that act on functional classes C_1 and C_2 but not on C_3, while G_{C_1} captures the constraints that are unique to C_1.

Fig. 4 gives the algorithm for learning a differential GMRC. It is similar to the algorithm in Fig. 3 for monolithic GMRCs but has some key differences. The main extension for the differential algorithm is that the model for a child in the lattice extends the model for its parent (which extends the model for its parent). Thus, the algorithm takes an

function *differentialGMRC*
input D: possible edges
input $\mathcal{S}_1, \ldots, \mathcal{S}_m$: the m functional classes
input F: ancestor edges

$V = \{v_1, \ldots, v_n\}$; $E \leftarrow \emptyset$
for all $i = 1$ to m **do**
  $s_i \leftarrow \text{Score}(G = (V, F))$ using sequences $\mathcal{S}_i$
  $C_i \leftarrow \{(e, \text{NormalizedScore}(G=(V,F),\{e\}) | e \in D)\}$
end for
repeat
  $e \leftarrow \arg\max_{e \in D-E} \sum_{i=1}^{n} C_i(e)$
  if e is significant for all $\mathcal{S}_j$ **then**
    $E \leftarrow E \cup \{e\}$
    for all $i = 1$ to m **do**
      $s_i \leftarrow s_i - (s_i \cdot C_i(e))$
      for all $e' \in D-E$ s.t. e and e' share a vertex **do**
        $C_i(e') \leftarrow \text{NormalizedScore}(G=(V_i, E \cup F), e')$
      end for
    end for
  end if
until stopping criterion satisfied

return $G = (V, E)$

Fig. 4. Algorithm for inferring a differential GMRC.

additional parameter giving the edges in the ancestor models.

The other differences in the differential algorithm arise because a differential model deals with multiple functional classes instead of just one. The algorithm must ensure that the learned conservation and coupling constraints are representative of *all* the classes. We first extend the score function from (6) to combine information from the classes. Since the score may be quite different for different classes, we normalize the reduction in score caused by edge e for a class:

$$\text{NormalizedScore}(G = (V, E), e) = 1 - \frac{\text{Score}(G = (V, E + e))}{\text{Score}(G = (V, E))}. \quad (8)$$

We select the edge that maximizes the sum of the normalized scores over all functional classes. We require that an edge have a $\text{NormalizedScore} > 0$ in each class, be statistically significant in each class, and be represented by at least 15 percent of the sequences in each class remain upon subsetting (to ensure fidelity to each class alignment). Enforcing these conditions for each class ensures that the constraints are indeed representative.

2.6 Assessing Likelihood

The pdf $P_G(R)$ (3) for residue types R according to a GMRC G can be used as a likelihood in order to evaluate the probability that a sequence R belongs to a model. The likelihood depends on the estimator employed for $P_C(R_C)$ (5) used in the potential function factors in the pdf. For a monolithic GMRC, the estimator simply uses the frequencies of amino acid types in the given set of sequences. At the leaves of a differential GMRC, the estimator uses the frequencies in the sequences of the particular functional class. At the interior nodes, the estimator uses the frequencies in the sequences in the union of the involved functional classes. We denote by $\mathcal{L}_{G_{C_1,C_2}}(s)$ the likelihood of s under a model of classes C_1 and C_2 and similarly for other models.

2.7 Classification

One advantage of using a formal probabilistic method for scoring is that likelihoods from different models are comparable. Given two graphical models for two different functional classes, G_{C_1} and G_{C_2}, we can classify a new sequence s to either functional class C_1 or C_2 by computing the log likelihood ratio LLR:

$$LLR = \log \frac{\mathcal{L}_{G_{C_1}}(s)}{\mathcal{L}_{G_{C_2}}(s)} \quad (9)$$

and assigning s to C_1 if LLR is greater than 0 and to C_2 otherwise. Furthermore, LLR provides a measure of confidence in our assignment. The larger LLR is, the more confident we are that the sequence belongs to C_1 and not C_2. We can make predictions only when we are significantly more confident in one model than another. For instance, we can make classifications when the likelihood ratios indicate that one model is 10 or 100 times more likely than any other.

The likelihood scores in differential graphical models can likewise be used in classifying sequences. Since the likelihoods are directly comparable at each level of the lattice, we perform a hierarchal classification by assigning a sequence to the most likely subset model at each level of the lattice. Each step down, the lattice removes from consideration one functional class; at the bottom, the sequence is classified to a single functional class.

3 Results and Discussion

We demonstrate the power of our GMRC approach in analysis of the GPCRs. GPCRs are membrane-bound proteins essential in cellular signaling: ligand binding at the extracellular face initiates the propagation of structural changes through the transmembrane helices and ultimately to the cytoplasmic domains where a G protein is activated. Fig. 5 shows the 3D structure of a representative GPCR and an unrolled 2D schematic. In each structure, the extracellular portion is at the top, while the cytoplasmic portion is represented at the bottom. The seven transmembrane helices (numbered 1-7) form a barrel through which the structural changes are propagated upon ligand binding.

GPCRs have been the object of previous residue coupling studies (e.g., [26] and [36]). We obtained the set of 940 aligned GPCR sequences used in the coupling study by Ranganathan et al. [36], which they had manually adjusted from sequences in the GPCRDB [11]. The alignment contains 940 sequences, each with 348 residues. GPCRs can be divided into five major classes, labeled classes A though E. Class A GPCRs (Rhodopsin and andreenergiclike receptors) are the most studied GPCRs, and the sequences chosen by Ranganathan and colleagues were selected to be representative of this class.

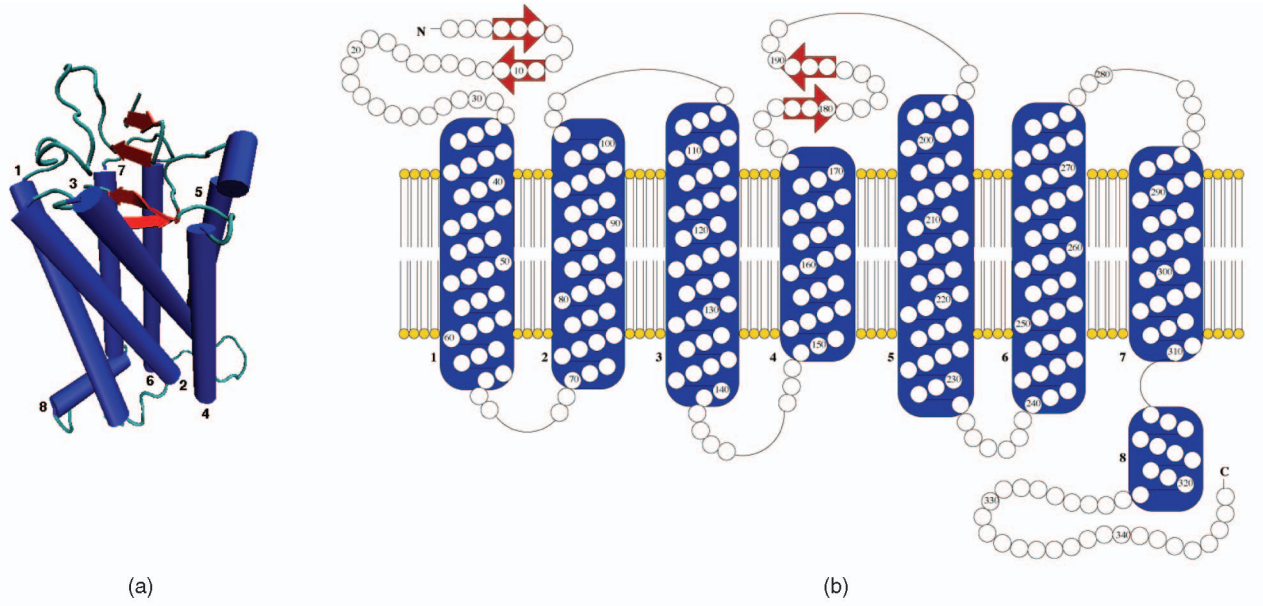


Fig. 5. The structure of a representative GPCR, bovine Rhodopsin (PDB id 1GZM). (a) A 3D structure, which consists of seven transmembrane helices, numbered 1 to 7, and one helix that resides inside the cell, numbered 8 (figure prepared using VMD [12]). (b) A two-dimensional schematic with the residue numbers from the MSA. The top seven helices (1-7, from left to right) are the transmembrane helices, and the lower helix is helix 8. In each structure, the top is the extracellular portion of the GPCR, where ligand binding occurs, while the bottom represents the inside of the cell where the G protein is activated.

Using the GPCRDB [11], we further annotated the alignment with functional class information according to the type of ligand each GPCR binds. The ligand binding data in the GPCRDB is manually culled from the literature and updated frequently, ensuring high-quality annotations. Using these classifications, each GPCR from the data set is assigned to one of 16 classes; see Fig. 6. For the purposes of our results, we refer to the model of the entire set as "family wide" (treating 940 members of class A as the family) and the models of the classes as "class specific" (treating each class as a family in its own right).

Using their SCA approach (see Section 2.1) on the family-wide sequences, Ranganathan et al. found that many interesting coupling relationships that they conjecture form a network of allosteric communication [36]. Our GMRC approach identifies many of the same constraints [37] but, as we discussed above, provides a compact representation with a probabilistic semantics, supporting a variety of reasoning techniques (as discussed in Section 1.3). We will illustrate here the effect of functional annotations and some of the interpretations and applications supported by our approach.

3.1 Coupling as Differential Conservation

As discussed in the introduction and illustrated in Fig. 1, functional class information enables differentiation of conservation and coupling constraints that are specific to the individual functional classes from those that are common to all classes (or even particular sets of classes). To evaluate the necessity for such distinctions in GPCRs, we considered the 10 "perturbations" used as the basis for coupling identification in the SCA analysis of GPCRs [36]. A perturbation selects those sequences with a particular (relatively common) amino acid type at a particular position. If, for the selected sequences, the amino acid frequencies at a second position are very different from their frequencies in the full set of sequences, then that second position is deemed to be coupled to the perturbed residue. With functional class information at hand, we can now also consider whether the frequencies of functional classes in the sequence subset are similar to those in the full set. When covariation is due to family-wide coupling, the perturbation will result in a subset of sequences that is representative of the various functional classes in the original set. On the other hand, when covariation is due to class-specific conservation, the perturbation will select sequences primarily from that class.

Fig. 7 shows the effect of the perturbations on the three largest functional classes, Amine, Peptide, and Rhodopsin (individually), as well as all other functional classes (as a group) from the data set. Notice that each of the 10 perturbations involves a residue that is highly conserved in the Amine functional class, and each thus leaves a large number of Amine sequences. In fact, each perturbation is for an amino acid type that is at least 71 percent conserved in the Amine functional class; most are more than 85 percent conserved. This high level of conservation does not occur in the other large functional classes. Some perturbations for the Peptide functional class do leave a moderate fraction—four perturbations leave more than 30 percent of the sequences—however, others leave less than 10 percent of the sequences. The Rhodopsin functional class is not well represented under the perturbations; only three of the 10 identified perturbations capture more than 10 percent of the sequences, while the remaining seven maintain less than 3 percent of the sequences. The other functional classes are also not as well represented under the perturbations as the Amine class is. Furthermore, the different functional classes have different behaviors for different perturbations, e.g., 40

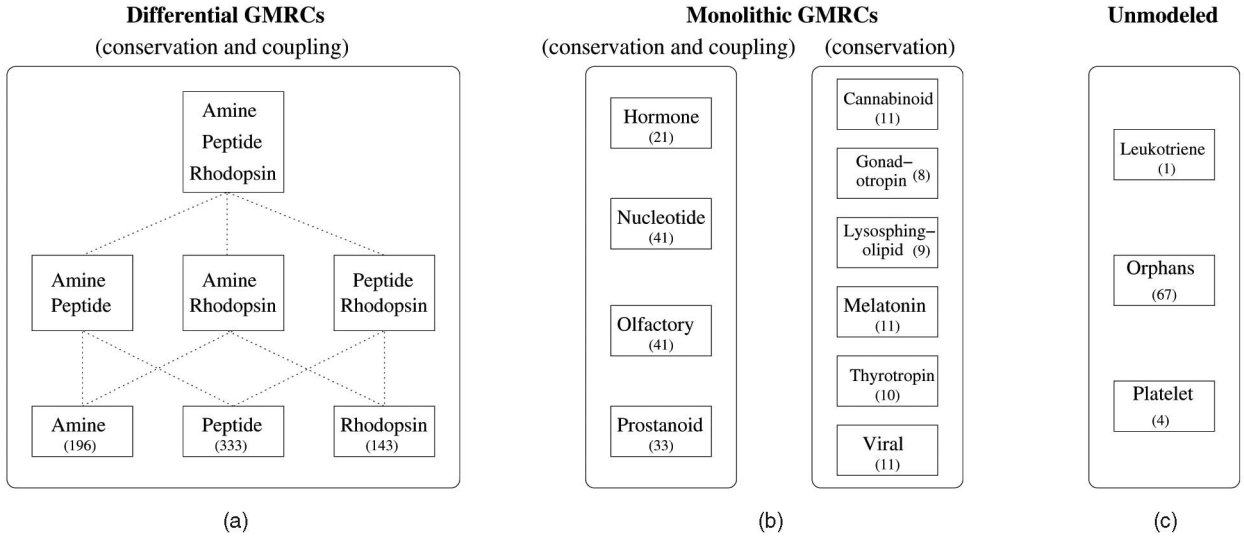


Fig. 6. Schematic of graphical models for GPCRs. For the three largest functional classes, we use a (a) differential GMRC. The top level of the lattice contains constraints that act on all the functional classes. The middle level adds constraints that act on two of the classes, while the bottom level has those constraints that are unique to the individual classes. For functional classes with 5-50 sequences, we learn (b) monolithic GMRCs. For functional classes with more than 20 sequences, we learn graphical models using both conservation and coupling information, while for functional classes with fewer than 20 sequences, we learn graphical models using only conservation information. (c) We do not learn GMRCs for functional classes with fewer than five sequence or for sequences whose functional class is unknown.

of the 41 Olfactory sequences have a Y at position 144, while none of them have an F at 268.

We conclude in Fig. 7 that the previously observed coupling is due primarily to the selection of the Amine functional class by the perturbations, due to the high level of conservation in Amines of the chosen amino acid types.

3.2 Differential Model

By "factoring out" specific constraints, our differential GMRC allows us to further uncover additional constraints that are common to the entire family or specific to other classes. Fig. 6 summarizes the structure of the differential model. Due to the variation in the number of sequences in the various functional class, we focused the differential modeling effort on the three best represented classes (Amine, Peptide, and Rhodopsin). We also constructed individual monolithic models for the four functional classes with 20-50 sequences (Hormone, Nucleotide, Olfactory, and Prostanoid). Although these monolithic models do not directly capture the similarities and differences, they still lend insight into the constraints acting on the classes and allow us to avoid the combinatorial explosion in considering all subsets of classes. For the six functional classes with between 5-19 sequences (Cannabinoid, Gonadotropin, Lysosphingolipid, Melatonin, Thyrotropin, and Viral), we created graphical models using only conservation information, since coupling becomes unreliable with such a small number of sequences. We ignored the two functional classes (Leukotriene and Platelet) with fewer than five sequences. We used the 67 sequences from the data set annotated as Class A orphans (sequences believed to be members of class A GPCRs but whose functional class is unknown) to show the utility of our models to perform classification.

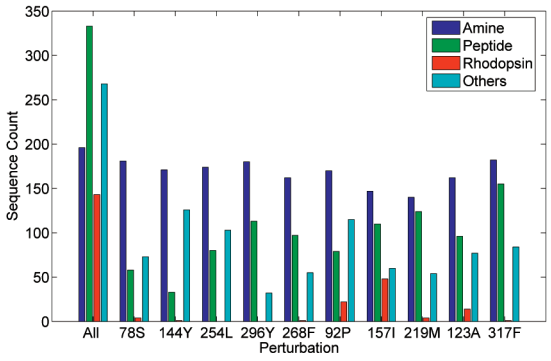


Fig. 7. The effect of each perturbation identified by SCA analysis on the functional classes. The x-axis shows the perturbation (the selected moderately conserved amino acid), while the y-axis shows the number of sequences from each functional class that remain after the perturbation. The first perturbation, All, shows the number of sequences from each functional class in the original MSA. Notice that for each perturbation, the Amine class is highly conserved, while the remaining classes are not. This is an indication that the observed covariation is due to class-specific conservation, rather than family-wide coupling.

We learn models using either an uninformative prior (allowing an edge between all pairs of residues) or a contact graph prior (allowing an edge only between residue pairs that are within a certain distance in the 3D structure). For the contact graph prior, we used the structure determined by Li et al. [19] (PDB id 1GZM); see Fig. 5a. Two residues are considered to be in contact if any atom pair is separated by at most seven Å. Since structure is more conserved than sequence, we assume that all members adopt essentially the same contact graph. The uninformative prior considers all possible 60,378 edges (i.e., between all pairs of the 348 residues), whereas the contact graph prior only considers 3,161 of these edges.

For the results that follow, we set $\rho = 0.1$ as the weight for missing data in (5). Previously, we have found that our results are relatively insensitive to a range of ρ values [37]. As discussed, we impose a 15 percent threshold for the

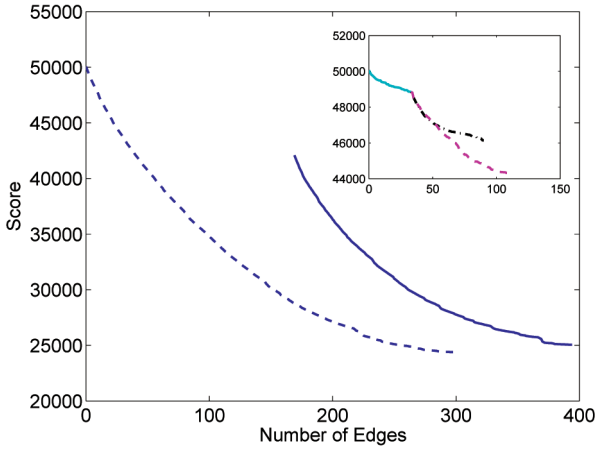


Fig. 8. Evolution of scores for the Amine class as edges are added to both a monolithic GMRC (dashed blue) and differential GMRC (solid blue). The x-axis shows the number of edges in the model, while the y-axis shows the score of the model. The differential model is learned in several steps (inset). Initially, a model is learned that accounts for constraints shared with the Peptide and Rhodopsin functional classes (solid cyan). The model is then augmented with constraints shared with just the Peptide (dot-dashed black) and Rhodopsin (dashed magenta) functional classes. Finally, a model is learned for constraints that are unique to the Amine functional class (blue).

calculation of conditional mutual information (2). We found negligible impact from small deviations in the threshold; values of 10 percent, 15 percent, and 20 percent all produce classification accuracy greater than 97 percent. We use a p-value threshold of 0.005 as the significance level for each edge in (7). Note that this selection is somewhat arbitrary and could be adjusted as desired to account for multiple hypothesis testing. For example, a simple Bonferroni correction for the 60,378 edges in the uninformative prior would yield a threshold of approximately 10^{-7}. Most of the edges identified by our algorithm meet even this far more stringent threshold.

The next section describes important constraints in the models thus constructed and demonstrates the ability of our method to identify important similarities and differences among the classes. A potential concern, however, is that the overhead of the differential graphical model results in models that are of lower quality than if we had simply constructed individual monolithic models for the different functional classes. (As discussed in the introduction, however, multiple monolithic models would not be directly comparable, since the same independencies can be factored in several different ways.) We compared the models learned by these two approaches for the Amine class, i.e., a differential model starting with Amine, Peptide, and Rhodopsin versus a monolithic model for just Amine. Using an uninformative prior, the differential graphical model identified 394 statistically significant couplings for the Amine functional class, while the monolithic model identified 298 statistically significant couplings. The score of a model (6) measures how much coupling remains; the lower the score, the better the model explains the observed coupling. Fig. 8 shows the evolution of the scores as edges are added to each model. The dashed and solid lines represent the score as edges are added in the monolithic and differential models, respectively. The differential model is learned in multiple parts, and the inset represents the evolution of score over those different parts. The solid cyan line shows how much the score decreases when learning the constraints that act on the Amine, Peptide, and Rhodopsin functional classes. Next, the score reduction is shown for the constraints that act on both the Amine and Peptide classes (dot-dashed black), as well as on both the Amine and Rhodopsin classes (dashed magenta). Finally, the score reduction is shown for the constraints unique to the Amine functional class (solid blue). The final differential model and the final monolithic model have similar scores (within 3 percent), and no remaining edges that reduce the score were found to be significant. We found the same behavior when we used a contact graph prior (not shown), leading us to conclude that the differential model explains the coupling about, as well as the monolithic model, but with the added benefit of uncovering similarities and differences among the functional classes.

3.3 Important Constraints

The models learned by our algorithms explicate the constraints (coupling and conservation) acting on the family and the functional classes. It is a significant feature of our approach that we can examine a model and identify biologically significant constraints suitable for experimental evaluation (e.g., by standard double mutant cycles or by combinatorial recombination [33], [38]). We illustrate here some of the strongest constraints identified for some of the models.

We annotate edges by whether they were found in the model learned under the uninformative prior (U) or the contact graph prior (C), or common to both (U + C). The two models have relatively few couplings in common because the contact graph prior significantly limits the edges being considered. For example, for the Amine functional class, the uninformative prior considers 50,684 statistically significant edges, while the contact graph prior considers only 2,539. The problem is further exacerbated by the nonuniqueness of factorization—the same independencies can be encoded multiple ways. Ultimately, the two models are developed under alternative hypotheses about the general relationship between coupling and direct contact, and as discussed above, the resulting constraints may suggest experiments to test particular relationships.

3.3.1 *Amine + Peptide + Rhodopsin*

Our differential GMRC finds common constraints acting on multiple functional classes. For the Amine, Peptide, and Rhodopsin functional classes, our model identifies 33 statistically significant couplings with an uninformative prior and 14 with the contact graph prior. Of the 14 couplings identified under the contact graph prior, seven are also identified under the uninformative prior.

Table 1 highlights several of the more revealing constraints in the model. The probabilities given in the table are the minimums across the three functional classes. In other words, for each functional class, the probability of the constraint is at least that value. For example, residue 267 is a P in at least 99 percent of the sequences in each functional class. In addition, residues 264 and 299 are CS in at least 29 percent of the sequences for each functional class, i.e., when residue 264 is a C, residue 299 tends to be an S, and when residue 299 is an S, residue 264 tends to be a C in each

TABLE 1
Important Constraints in the
$\mathrm{Amine} + \mathrm{Peptide} + \mathrm{Rhodopsin}$ Model

Conservation (5 of the 12 that are $> .9$)			
Model*	Residue	Value	Probability
U + C	55	N	.99
U + C	79	L	.91
U + C	187	C	.94
U + C	267	P	.99
U + C	303	P	.98
Coupling (5 of the 33U and 14C)			
Model*	Edge	Values	Probability
U	57–82	LA	.43
U + C	313–314	FR	.47
U + C	305–306	IY	.45
U + C	302–304	NI	.33
C	264–299	CS	.29

* U: uninformative prior; C: contact graph prior

of the functional classes. Notice that in this table, we only identify one residue pair for the couplings. This is because different functional classes may have differing pairs of amino acid types coupled, and we only point out the pairs that are common across all functional classes.

While several of the identified couplings are sequentially close (313-314, 305-306, and 302-304), others are long range. For example, our model identifies residues 57 (on the first helix) and 82 (on the second helix) as significantly coupled. Fig. 5 shows they are in close proximity; in fact, they may come in contact upon ligand binding in the Amine, Peptide, and Rhodopsin functional classes. Furthermore, our algorithm identifies several highly conserved residues across the Amine, Peptide, and Rhodopsin functional classes. Residues 55N, 79L, and 303P are all highly conserved in the Amine, Peptide, and Rhodopsin functional classes and are all in close proximity to each other in the 3D structure, possibly indicating a key structural motif for GPCRs.

3.3.2 $Amine + Peptide$

As we move down the differential graphical model (Fig. 6a), we identify constraints that act on only two of the three functional classes. For the Amine and Peptide functional classes, our algorithm identifies further 57 statistically significant couplings with the uninformative prior and further 31 with the contact graph prior. Of the 31 couplings identified by the model using the contact graph prior, eight are also identified by the model using the uninformative prior. Note that each model consists of the union of couplings from the current level and couplings from its ancestors' models. Thus, under the uninformative prior, both the Amine and Peptide functional classes have 90 couplings—the 33 that they share with the Rhodopsin functional class (from above) and 57 that are unique and common to the Amine and Peptide functional classes.

Table 2 highlights several of the stronger constraints common to the Amine and Peptide functional classes. Again, the probabilities are minimums over the two functional classes, and the constraints reveal interesting biological insights into the function of GPCRs. For example, residue 103 is highly conserved in both the Amine and Peptide functional classes (more than 70 percent) but not the Rhodopsin function class (less than 30 percent). Residue

TABLE 2
Important Constraints in the $\mathrm{Amine} + \mathrm{Peptide}$ Model

Conservation (all 5 that are $> .7$)			
Model*	Residue	Value	Probability
U + C	54	G	.81
U + C	103	W	.78
U + C	124	S	.76
U + C	134	D	.85
U + C	294	L	.72
Coupling (3 of the 57U and 31C)			
Model*	Edge	Values	Probability
U	72–138	TA	.42
C	80–85	AL	.55
C	70–153	TA	.28

* U: uninformative prior; C: contact graph prior

103 resides at the extracellular face of the membrane, where ligand binding occurs, and may indicate a difference between the type of ligand recognized by Amine and Peptide functional classes versus the Rhodopsin functional class. Another interesting biologically phenomenon identified by our model is the significant coupling between residues 72 and 138 in both the Amine and Peptide functional classes but not the Rhodopsin functional class. Notice that residue 72 (in helix 2) and residue 138 (in helix 3) are close to the active site for the G protein. The fact that they are coupled in the Amine and Peptide functional classes but not the Rhodopsin functional class could be because the Amines and Peptides interact differently with their associated G proteins than do the Rhodopsins.

3.3.3 Amine and Peptide Separately

At the lowest level of the differential graphical model, our algorithm identifies constraints that are unique to individual functional classes. For the Amine functional class, our differential graphical model identifies 225 statistically significant couplings with an uninformative prior and 200 with a contact graph prior. Of the 200 couplings identified by the model using a contact graph prior, nine are also identified by the model using an uninformative prior. The graphical model for the Peptide functional class identifies 140 statistically significant couplings with an uninformative prior and 96 with the contact graph prior. Of the 140 couplings identified by the model using a contact graph prior, six are also identified by the model using an informative prior.

Tables 3 and 4 show some of the stronger constraints that act on the Amine and Peptide functional classes, respectively. Note that now, the probabilities are those just for the individual classes, so the probabilities are exact (no longer minimums), and we include the most common amino acid pairs for each coupling. At the bottom level our differential GMRC provides significant insights into what makes each functional class unique. For example, residues 78 and 296 are uniquely conserved in the Amine functional class (more than 90 percent while less than 40 percent in both the Peptide and Rhodopsin functional classes), and both residues are involved in ligand interaction. Recall that both residues 78 and 296 were residues identified as coupled by SCA [36]; see Fig. 7. The Amine functional class also has unique couplings between residues, which may come in contact upon ligand binding. For instance, residues 90 (helix 2) and 122 (helix 3),

TABLE 3
Important Constraints in the Amine Model

Conservation (all 5 that are > .9)			
Model*	Residue	Value	Probability
U + C	78	S	.92
U + C	117	D	1
U + C	196	F	1
U + C	293	W	.98
U + C	296	Y	.91
Coupling (5 of the 225U and 200C)			
Model*	Edge	Values	Probability
U	90–122	VT	.83
		SN	.12
U	207–268	SF	.69
		TY	.14
U	147–226	KI	.39
		RI	.15
		LV	.17
C	168–171	SP	.63
		WA	.11
C	259–264	GC	.76
		LT	.11

* U: uninformative prior; C: contact graph prior

TABLE 4
Important Constraints in the Peptide Model

Conservation (all 4 of those > .6)			
Model*	Residue	Value	Probability
U + C	44	Y	.67
U + C	230	L	.67
U + C	300	C	.75
U + C	308	F	.69
Coupling (4 of the 140U and 96C)			
Model*	Edge	Values	Probability
U	93–292	FH	.11
		WE	.18
U	143–246	LA	.17
		VK	.12
U + C	127–157	FS	.14
		LC	.18
		SI	.14
C	90–92	CP	.18
		LF	.28

* U: uninformative prior; C: contact graph prior

as well as residues 207 (helix 5) and 268 (helix 6) are significantly coupled, and both pairs are in close physical proximity in the 3D structure. Similar insights can be found from the differential GMRCs for the Peptide functional class. For example, residues 230 (helix 5) and 308 (helix 7), both of which are located near the activation site of the G protein, display conservation that is unique to the Peptide functional class (more than 60 percent is conserved in the Peptide functional class, while less than 30 percent is conserved in both the Amine and Rhodopsin functional classes). Furthermore, residues 127 (helix 3) and 156 (helix 4) are uniquely coupled in the Peptide functional class and are in contact in the 3D structure.

3.3.4 Other Functional Classes

In addition to the differential GMRC for the Amine, Peptide, and Rhodopsin functional classes, we learn monolithic GMRCs for the Hormone, Nucleotide, Olfactory, and Prostanoid functional classes. Since the models were not learned in a differential framework, their constraints cannot be directly compared. For completeness, however, we note that the models identify 143, 271, 308, and 283 couplings with the uninformative prior and 141, 310, 292, and 334 couples with the contact graph prior for the Hormone, Nucleotide, Olfactory, and Prostanoid functional classes, respectively. We use these models below in classification tests.

3.4 Classification

The graphical models learned by our algorithm can be used to assign functional classes to family members of unknown class membership. A GMRC allows us to compute a likelihood score that gives the probability that a sequence belongs to a particular functional class. Since probabilities are comparable, we classify a sequence by computing its likelihood under each GMRC and assigning it to the model with the highest likelihood score. When we have a differential GMRC, as is the case for the Amines, Peptides, and Rhodopsins, we perform a hierarchical classification. At each level in the lattice, we assign the sequence to the model for the subset of classes with the highest likelihood score. We proceed to the next level, considering only the subsets of the selected subset. The process continues until a leaf is reached and the sequence assigned to a particular functional class. In the case of GPCRs, recall that we have a differential GMRC for Amine, Peptide, and Rhodopsin and monolithic GMRCs for the remaining functional classes (Fig. 6). Thus, the first step in GPCR classification is to decide upon one of the monolithic GMRCs or the top level of the differential GMRC. If the most likely model is a monolithic GMRC, the sequence is assigned to that class. If, on the other hand, the most likely model is the top level of the differential GMRC, we hierarchically classify the sequence as just described. Note that this is only one of many possible ways to classify sequences using a differential graphical model. Another possibility is to adopt a Bayesian viewpoint and cast the likelihood for a model as a marginalized combination of the likelihoods of it and its ancestors. Such a weighting scheme allows one to place more emphasis on the constraints that make each functional class unique from all others.

To test the quality of this classification mechanism, we performed a fivefold cross-validation test. We divided each functional class into five parts and performed five training/testing runs. For each run, four of the five parts were used to learn models, and the remaining sequences were classified to their most likely models. When classifying, we did not set a likelihood ratio threshold (i.e., we always made a classification decision), but we always observed the ratios to be significant. Table 5 shows the cumulative results over all five runs. As the table shows, for most classes, the classification is perfect. Of the 16 total errors, seven come from classes with 11 or fewer sequences, where we rely entirely on sparse conservation data. The overall classification is highly accurate, assigning the functional class correctly for 98.3 percent of the sequences. Although some SVM approaches have been shown to obtain similar accuracy for GPCRs [4], [14], as we illustrate below, our graphical models are transparent, making apparent the key

TABLE 5
Cross-Validation Results

Functional Class	Total	Correct	Accuracy (%)
Amine	196	195	99.5
Peptide	333	333	100
Rhodopsin	143	141	98.6
Nucleotide	41	36	87.8
Olfactory	41	41	100
Prostanoid	33	33	100
Hormone	21	21	100
Cannabinoid	11	11	100
Melatonin	11	11	100
Viral	11	6	54.6
Thyrotropin	10	9	90
Lysosphingolipid	9	9	100
Gonadotropin	8	7	87.5
Overall	868	853	98.3

differential aspects (conservation and coupling) in a manner readily permitting biological study.

As a further utility of our GMRCs, recall that the data set contains 67 orphan sequences that are assumed to be class A GPCRs but whose functional class is unknown. Since our models proved to be accurate in cross-validation tests, we used them to classify these orphans to their functional classes. We used the models learned from all the sequences (as in the previous section, rather than with the $4/5$ used in cross-validation tests). Of the 67 orphan sequences, three were classified as Amine and the remaining 64 as Peptide. Table 6 shows several of the sequences and their classifications. The log likelihood ratio (between the most likely model and the second most likely one) provides a level of confidence for our classification (higher is more confident). Since all of the classifications were to Amine, Peptide, or Rhodopsin, we also include in the table information about the subsets considered during the hierarchical classification. At both the top level ($\mathrm{Amine} + \mathrm{Peptide} + \mathrm{Rhodopsin}$ versus other classes) and bottom level (one of the three classes), we are very confident in the classification. Our confidence is not as high at the middle level, because each sequence appears in two different subsets, which capture its commonalities with the other classes. For example, an Amine class GPCR would have a high likelihood under the $\mathrm{Amine} + \mathrm{Peptide}$ model, as well as the $\mathrm{Amine} + \mathrm{Rhodopsin}$ model. We have found that the ultimate classification results do not depend on which of the two such models it is assigned to during the hierarchical process (data not shown).

As we have discussed, the classification decisions are transparent, traceable to the underlying conservation and coupling constraints. Table 6 highlights several of the strong constraints behind the orphan sequence classifications. We only show the constraints listed in the previous section. For instance, YQNJ_CAEEL was classified as an Amine because it has 78S and 296Y, which are both conserved residues unique to the Amine class. Furthermore, it has 90V-122T and 207S-268F, two unique significant couplings in that class. Such interpretable classification justifications are a significant advantage of our approach over "black box" classifiers.

3.5 PDZ Domains

As a further test of our method, we applied it to a different protein family, namely, PDZ domains. PDZ domains occur in many different proteins and often assist in the formation of complexes by binding to the C-termini of other proteins, their ligands. Previous coupling studies of PDZ domains have found many interesting biological relationships among coupled residues [23]. Here, we focus on the ability of our models to use these statistically significant residue couplings to perform functional classification of PDZ domains. Traditionally, PDZ domains have been classified into two functional classes, depending on the type of ligands they recognize. The first class of PDZ domains recognizes ligands of the form S/T-X-Φ, where Φ is a hydrophobic residue, while the second class recognized ligands of the form Φ-X-Φ. We obtained MSAs for the two classes of PDZ domains by querying the PDZBase [3] by ligand type and removing duplicate entries. The

TABLE 6
Orphan Classification

Identifier	Class	LLR	Justifications
GPR1_HUMAN	APR	223.75	55N, 79L, 187C, 267P, 303P, 302N–304I, 264C–299S
	AP	3.40	54G, 103W, 124S, 134D, 294L
	Peptide	378.80	44Y, 300C
GPRA_HUMAN	APR	321.39	55N, 79L, 187C, 267P, 303P, 313F–314R, 305I–306Y
	AP	28.92	54G, 103W, 124S, 134D, 294L, 80A–85L
	Peptide	267.42	44Y, 230L, 300C, 90C–92P
YDBM_CAEEL	APR	388.25	55N, 79L, 187C, 267P, 303P, 305I–306Y, 302N–304I, 264C–299S
	AP	4.29	103W, 124S, 134D, 294L, 70T–153A
	Amine	131.90	117D, 293W, 296Y, 90V–122T, 207S–268F, 168S–171P, 259G–264C
YQNJ_CAEEL	APR	414.96	55N, 79L, 187C, 267P, 303P, 313F–314R, 264C–299S
	AP	32.89	54G, 103W, 124S, 134D, 294L
	Amine	138.85	78S, 117D, 293W, 296Y, 90V–122T, 207S–268F, 147K–226I, 168S–171P
YXX5_CAEEL	APR	272.86	55N, 79L, 187C, 267P, 303P, 313F–314R, 305I–306Y
	AP	6.85	54G, 103W, 124S, 134D, 294L
	Peptide	220.08	44Y, 300C, 308F
YYI3_CAEEL	APR	375.42	55N, 79L, 187C, 267P, 303P, 57L–82A, 313F–314R, 305I–306Y, 264C–299S
	AP	28.07	103W, 124S, 134D
	Amine	93.56	78S, 117D, 293W, 90V–122T, 207S–268F, 147K–226I, 259G–264C

TABLE 7
PDZ Domain Classification

min LLR	Functional Class	Number Classified	Number Correct	Accuracy (%)
0	Class I	80	63	78.7
	Class II	13	13	100
	Total	93	76	81.7
2	Class I	76	62	81.6
	Class II	13	13	100
	Total	89	75	84.3
5	Class I	72	61	84.7
	Class II	13	13	100
	Total	85	74	87.1
10	Class I	68	60	88.2
	Class II	12	12	100
	Total	80	72	90.0

class I alignment consists of 80 sequences, and the class II alignment consists of 13 sequences; each sequence is 80 residues long.

It is important to note that the PDZ domains are an especially difficult test case. First, detecting coupling is hard with a small number of sequences (13 in class II). Second, the active site is relatively small, providing relatively little "direct" information about functional class. Finally, there is high homology between the class I and class II PDZ domains. Thus, conservation is unlikely to provide much information about classification. In fact, with models containing only conservation constraints, 90 of the 93 sequences are classified as class I PDZ domains. Nonetheless, we will see how, even with this small number of sequences, statistically significant couplings incorporated in a GMRC can improve classification accuracy.

Since our focus is on classification, we learned a differential GMRC for the PDZ domains using an uninformative prior. The model identifies 11 statistically significant couplings common to both classes. It also identifies statistically significant differential couplings, 85 class I-specific and 10 class II-specific. The large difference in the number of couplings identified in the class-specific models is due to the smaller number of class II sequences.

To test the power of our classification on PDZ domains, we performed a 13-fold cross-validation test for PDZ domains. (This is a leave-one-out cross validation for the class II PDZ domains.) In each fold, we selected 12 of the 13 parts for training and used the other for testing. By classifying each testing sequences to its most likely functional class, our models achieve an overall classification accuracy of 81.7 percent. However, we can improve the accuracy of the classification by only classifying sequences that have a high log likelihood ratio LLR (9). Table 7 shows the classification results for various minimum LLR values. Notice that as the minimum LLR value increases (meaning our confidence in one class over the other increases), the overall accuracy of our classification increases, while the number of sequences classified decreases. For a minimum LLR of 10, we classify the sequences with an accuracy of 90 percent while still classifying 86 percent of the total sequences.

4 Conclusions

Our GMRCs have a number of advantages over traditional approaches to representing coupling: rather than assessing dependence (which can conflate direct and indirect relationships), they capture independence (which enables modular reasoning about variation), they make explicit the essential constraints underlying the family (e.g., by identifying a small set of couplings that explain the data nearly as well as the complete set), they enable the integration of structural and functional information, and their rigorous probabilistic semantics enables prediction and inference. The differential approach further helps distinguish residue conservation and coupling constraints within a protein family as a whole from those specific to particular functional classes.

In ongoing research, we intend to connect sequence-structure-function relationships of the type learned here with evolutionary models that aim to explain divergence of classes within a family. We also hope to probe our models by protein engineering experiments [33], [38] that test the importance of identified constraints. We intend to develop sampling approaches that use a model to design novel sequences that satisfy user-specified constraints but also remain faithful to the evolutionarily compensatory behavior encoded in the model. The ability to interpolate between class specific models is potentially quite significant in that context. These and similar ideas have been given a significant boost with the promising results by Ranganathan et al. [32], [35]. Finally, we are working to explore constraints acting on a protein family not just among the residues of its members but also through their interaction with other proteins (e.g., ligands). Specifically, we aim to create "coupled" graphical models that capture residue covariation both within their respective families and across them.

Acknowledgments

J. Thomas and C. Bailey-Kellogg are supported in part by a CAREER Award from the US National Science Foundation (NSF) Grant IIS-0444544. N. Ramakrishnan is supported in part by NSF Grants IBN-0219332 and EIA-0103660. This work was inspired by conversations with Dr. Rama Ranganathan (UT Southwestern) and Dr. Alan Friedman (Purdue).

References

[1] A. Armon, D. Graur, and N. Ben-Tal, "ConSurf: An Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Phylogenetic Information," *J. Molecular Biology*, vol. 307, pp. 447-463, 2001.

[2] W.R. Atchley, W. Terhalle, and A. Dress, "Positional Dependence, Cliques, and Predictive Motifs in the bHLH Protein Domain," *J. Molecular Evolution*, vol. 48, pp. 501-516, 1999.

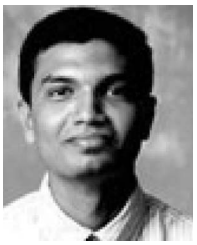
[3] T. Beuming, L. Skrabanek, M.Y. Niv, P. Mukherjee, and H. Weinstein, "PDZBase: A Protein-Protein Interaction Database for PDZ-Domains," *Bioinformatics*, vol. 21, no. 6, pp. 827-828, 2005.

[4] M. Bhasin and G.P.S. Raghava, "GPCRpred: An SVM-Based Method for Prediction of Families and Subfamilies of G-Protein Coupled Receptors," *Nucleic Acids Research*, vol. 32, pp. 383-389, 2004.

[5] W.L. Buntine, "Operations for Learning with Graphical Models," *J. Artificial Intelligence Research*, vol. 2, pp. 159-225, 1994.

[6] M.W. Dimmic, M.J. Hubisz, C.D. Bustamante, and R. Nielsen, "Detecting Coevolving Amino Acid Sites Using Bayesian Mutational Mapping," *Bioinformatics,* vol. 21, no. S1, pp. i126-i135, 2005.
[7] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge Univ. Press, 1998.
[8] 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, pp. 211-221, 2004.
[9] U. Göbel, C. Sander, R. Schneider, and A. Valencia, "Correlated Mutations and Residue Contacts in Proteins," *Proteins: Structure, Function, and Genetics,* vol. 18, no. 4, pp. 309-317, 1994.
[10] I.V. Grigoriev and S.-H. Kim, "Detection of Protein Fold Similarity Based on Correlation of Amino Acid Properties," *Proc. Nat'l Academy of Sciences,* vol. 96, no. 25, pp. 14318-14323, Dec. 1999.
[11] F. Horn, G. Vriend, and F.E. Cohen, "Collecting and Harvesting Biological Data: The GPCRDB and NucleaRDB Databases," *Nucleic Acids Research,* vol. 29, no. 1, pp. 346-349, 2001.
[12] W. Humphrey, A. Dalke, and K. Schulten, "VMD—Visual Molecular Dynamics," *J. Molecular Graphics,* vol. 14, pp. 33-38, 1996.
[13] A.Y. Hung and M. Sheng, "PDZ Domains: Structural Modules for Protein Complex Assembly," *J. Biological Chemistry,* vol. 277, no. 8, pp. 5699-5702, Feb. 2002.
[14] R. Karchin, K. Karplus, and D. Haussler, "Classifying G-Protein Coupled Receptors with Support Vector Machines," *Bioinformatics,* vol. 18, no. 1, pp. 147-159, 2002.
[15] K. Karplus, "Regularizers for Estimating Distributions of Amino Acids from Small Samples," technical report, Computer Eng. and Information Sciences, Univ. of California, Mar. 1995.
[16] I. Kass and A. Horovitz, "Mapping Pathways of Allosteric Communication in GroEL by Analysis of Correlated Mutations," *Proteins: Structure, Function, and Genetics,* vol. 48, pp. 611-617, 2002.
[17] 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," *Proc. Nat'l Academy of Sciences,* vol. 90, pp. 7176-7180, Aug. 1993.
[18] S. Lauritzen, *Graphical Models.* Oxford Univ. Press, 1996.
[19] J. Li, P.C. Edwards, M. Burghammer, C. Villa, and G.F. Schertler, "Structure of Bovine Rhodopsin in a Trigonal Crystal Form," *J. Molecular Biology,* vol. 343, no. 5, pp. 1409-1438, Nov. 2004.
[20] O. Lichtarge, H.R. Bourne, and F.E. Cohen, "An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families," *J. Molecular Biology,* vol. 257, pp. 342-358, 1996.
[21] A.H. Liu and A. Califano, "CASTOR: Clustering Algorithm for Sequence Taxonomical Organization and Relationships," *J. Computational Biology,* vol. 10, no. 1, pp. 21-45, 2003.
[22] A.H. Liu, X. Zhang, G.A. Stolovitzky, A. Califano, and S.J. Firestein, "Motif-Based Construction of a Functional Map for Mammalian Olfactory Receptors," *Genomics,* vol. 81, no. 5, pp. 443-456, 2003.
[23] S.W. Lockless and R. Ranganathan, "Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families," *Science,* vol. 286, no. 5438, pp. 295-299, Oct. 1999.
[24] M. Milik, S. Szalma, and K.A. Olszewski, "Common Structural Cliques: A Tool for Protein Structure and Function Analysis," *Protein Eng.,* vol. 16, no. 8, pp. 542-552, 2003.
[25] O. Noivirt, M. Eisenstein, and A. Horovitz, "Detection and Reduction of Evolutionary Noise in Correlated Mutation Analysis," *Protein Eng.,* vol. 18, no. 5, pp. 247-253, 2005.
[26] L. Oliveira, A.C.M. Paiva, and G. Vriend, "Correlated Mutation Analyses on Very Large Sequence Families," *Chembiochem,* vol. 3, pp. 1010-1017, 2002.
[27] O. Olmea, B. Rost, and A. Valencia, "Effective Use of Sequence Correlation and Conservation in Fold Recognition," *J. Molecular Biology,* vol. 293, pp. 1221-1239, 1999.
[28] M. Pagel, "Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters," *Proc. Biological Sciences,* vol. 255, no. 1342, pp. 37-45, 1994.
[29] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia, "Correlated Mutations Contain Information about Protein-Protein Interaction," *J. Molecular Biology,* vol. 271, pp. 511-523, 1997.
[30] D.D. Pollock and W.R. Taylor, "Effectiveness of Correlation Analysis in Identifying Protein Residues Undergoing Correlated Evolution," *Protein Eng.,* vol. 10, pp. 647-657, 1997.
[31] D.D. Pollock, W.R. Taylor, and N. Goldman, "Coevolving Protein Residues: Maximum Likelihood Identification and Relationship to Structure," *J. Molecular Biology,* vol. 287, pp. 187-198, 1999.
[32] W.P. Russ, D.M. Lowery, P. Mishra, M.B. Yaffee, and R. Ranganathan, "Natural-Like Function in Artificial WW Domains," *Nature,* vol. 437, pp. 579-583, 2005.
[33] L. Saftalov, P.A. Smith, A.M. Friedman, and C. Bailey-Kellogg, "Site-Directed Combinatorial Construction of Chimaeric Genes: General Method for Optimizing Assembly of Gene Fragments," *Proteins: Structure, Function, and Bioinformatics,* vol. 64, no. 3, pp. 629-642, Aug. 2006.
[34] O. Schueler-Furman and D. Baker, "Conserved Residue Clustering and Protein Structure Prediction," *Proteins: Structure, Function, and Genetics,* vol. 52, pp. 225-235, 2003.
[35] M. Socolich, S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner, and R. Ranganathan, "Evolutionary Information for Specifying a Protein Fold," *Nature,* vol. 437, pp. 512-518, 2005.
[36] 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, no. 1, pp. 59-69, Jan. 2003.
[37] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg, "Graphical Models of Residue Coupling in Protein Families," *Proc. Fifth ACM SIGKDD Workshop Data Mining in Bioinformatics (BIOKDD '05),* pp. 12-20, 2005.
[38] X. Ye, A.M. Friedman, and C. Bailey-Kellogg, "Hypergraph Model of Multi-Residue Interactions in Proteins: Sequentially-Constrained Partitioning Algorithms for Optimization of Site-Directed Protein Recombination," *Proc. Int'l Conf. Research in Computational Molecular Biology (RECOMB '06),* pp. 15-29, 2006.

John Thomas received the BS degree in mathematics and the BA degree in computer science from Gettysburg College and the MS degree in computer science from Dartmouth College, where he worked on scheduling theory, proving the existence of schedules that are simultaneously near optimal for two optimality criteria. He is currently a PhD candidate in computer science at Dartmouth College. His current research interest includes bioinformatics, in particular identifying and modeling evolutionary constraints to assist in protein analysis, classification, and design.



Naren Ramakrishnan received the PhD degree in computer sciences from Purdue University in August 1997. He is an associate professor of computer science and a faculty fellow at Virginia Tech. He also serves as an adjunct professor at the Institute of Bioinformatics and Applied Biotechnology, Bangalore, India. His research interests include computational science, problem solving environments, mining scientific data, and information personalization. He currently serves on the editorial board of *Computer*. He is a member of the IEEE Computer Society.

Chris Bailey-Kellogg received the BS and MS degrees from MIT and the PhD degree, with Feng Zhao, from Ohio State University and Xerox PARC. He conducted postdoctoral research in computational biology with Bruce Donald at Dartmouth. He is an associate professor of computer science at Dartmouth. His current research focus is on integrated experiment planning and data analysis for protein structure determination and protein engineering. He has received an US National Science Foundation Career award and an Alfred P. Sloan Foundation Fellowship.