Gatsby Computational Neuroscience Unit
University College London
http://www.gatsby.ucl.ac.uk

17 Queen Square, London
WC1N 3AR, United Kingdom
+44 20 7679 1176

Funded in part by the Gatsby Charitable Foundation.

February 7, 2007

GCNU TR 2007–001

Small sets of interacting proteins suggest latent linkage mechanisms through analogical reasoning

Ricardo Silva
rbas@gatsby.ucl.ac.uk
Gatsby Unit

Edoardo Airoldi
eairoldi@princeton.edu
Carl Icahn Laboratory, Princeton University

Katherine A. Heller
heller@gatsby.ucl.ac.uk
Gatsby Unit

Abstract

We consider the problem of automatically discovering analogies between pairs of proteins that are known to interact. For example, given a pair of interacting proteins, P_1:P_2, which of two other interacting pairs, P_3:P_4 or P_5:P_6, interacts in a way that is most "similar" to P_1:P_2? In other words, is the interaction P_3:P_4 more analogous to P_1:P_2 than P_5:P_6 is? The goal of such exploratory analysis is to find new subclasses of interactions that might be relevant for further study: e.g., P_1:P_2 might belong to a class of interactions that is not yet fully formalized, and scientists exploring the interaction between P_1:P_2 might want to find other interactions which behave in an analogous way. This can lead to novel ways of categorizing proteins based on functional similarity. We present a Bayesian formulation of this question and illustrate its application to exploring new taxonomies of protein-protein interactions. In order to objectively evaluate the performance of our method against state-of-the-art and popular methods for information retrieval, we adopt an evaluation measure based on the Munich Institute for Protein Sequencing taxonomy. The results indicate a significant advantage for our approach.

Small sets of interacting proteins suggest latent linkage mechanisms through analogical reasoning

Ricardo Silva
rbas@gatsby.ucl.ac.uk
Gatsby Unit

Edoardo Airoldi
eairoldi@princeton.edu
Carl Icahn Laboratory, Princeton University

Katherine A. Heller
heller@gatsby.ucl.ac.uk
Gatsby Unit

1 Contribution

Many university admission exams, like the American Scholastic Assessment Test (SAT) and Graduate Record Exam (GRE), used to include a section on analogical reasoning. A prototypical analogical reasoning question is as follows:

```
   DOCTOR:HOSPITAL ::
A) sports fan :  stadium
B) cow :  farm
C) professor :  college
D) criminal :  jail
E) food :  grocery store
```

The examinee has to answer which of the 5 pairs best matches the relation implicit in `DOCTOR:HOSPITAL`. Although all candidate pairs interact in some way, pair `professor:college` seems to best capture the notion of (*object*, *place of work*) implicit in the relation between doctor and hospital.

Performing this type of analogical reasoning could be extremely useful in less mundane domains. Consider the following question, composed solely of pairs of interacting proteins according to the MIPS classification system (Mewes and et. al, 2004):

```
   YDL061C : YLR167W ::
A) YBR084CA : YJL189W
B) YBL092W : YKR094C
C) YDL083C : YGL189C
D) YBL027W : YJL189W
E) YDR178W : YKL148C
```

In the given question, `YDL061C` is a protein of type *cytoplasm* (40.03), while `YLR167W` is of type *cytoplasmic and nuclear degradation* (6.13.01). Such is also the case of B) `YBL092W : YKR094C`. Other pairs contain members which are both in the 40.03 class but none in 6.13.01, and are in this sense not as close to the question pair as option B.

Dividing the population of protein interactions into *subpopulations with similar mechanisms of linkage* is therefore a way of categorizing proteins and their functional roles. The population of interacting pairs of proteins is not uniform. The biological mechanism by which protein pair P1:P2 is linked might not be the same mechanism behind the linkage of P3:P4, as illustrated above. The result of this effort is that taxonomies such as the Gene Ontology (Ashburner et al., 2000), or the Munich Institute for Protein Sequencing (MIPS) database (Mewes and et. al, 2004) can be enriched by suggesting analogical similarities of protein interactions.

As an example, consider the following thought experiment. Suppose, for the sake of illustration, that one was not aware of the above mentioned MIPS categories 40.03 and 6.13.01. Still, the scientist has observed a set **S** composed of a few pairs of proteins interacting in the cytoplasm, where one protein in each pair appears to have a functional role related to cytoplasmic degradation. The biological relevance of this regularity could be strengthened if other pairs with similar behavior were detected. This detection can be difficult in the early stages of investigation, where no rule that classifies pairs according to this class of interaction is known, and no labeled data for building classifiers is available. However, if one could query a database of protein-protein interactions to identify pairs similar to those in **S**, this information could be used to further clarify functional roles.

In this article we propose a novel way of determining the relational similarity of protein-protein interactions. Like some other approaches (Valencia and Pazos, 2002; Bader, 2003; Nabieva et al., 2005), this method analyses data containing: 1. information about which pairs of protein interact; 2. features of proteins, such as the respective gene expression levels corresponding to a given protein. Unlike some other approaches, our goal is to retrieve protein pairs that would belong to the same subpopulation ("cluster") of interactions represented by **S**. This is not a classification problem; our goal is *not* to predict if two proteins interact or not.

We emphasize that analogies might be ambiguous.[1] As a consequence, our methodology (and available software) is meant to provide an exploratory data analysis tool that is able to formulate and rank plausible analogies between protein-protein relations. The solution we propose for making less ambiguous queries is to allow for *sets* of pairs to be used. In this setup, the implicit qualitative relation is the same for all members in "question set" **S**, and the goal is to rank all other relations by how similar they are to this common, implicit, **S**-interaction[2]. In information retrieval terminology (Manning et al., 2007), our set **S** is a *query set* for which a retrieval system has to provide similar items.

To summarize, the core ideas that form our contribution are:

1. motivation: finding similar protein-protein interactions is a relevant and important problem, because it allows for novel taxonomies of proteins;

2. good methods for finding similar protein-protein interactions must be able to identify and select the relevant (predictive) physical attributes, or features, for any given interaction.

3. using only a single example to describe a subpopulation of protein-protein interactions is arguably a very ambiguous description. Larger query sets provide more consistent descriptions;

2 Approach

The basic approach we take in this paper was originally introduced by Silva et al. (2007). We detail its fundamental principles, and describe the particular way by which this approach is applicable to scalable protein-protein exploratory data analysis.

Let $\mathcal{A}$ and $\mathcal{B}$ represent object spaces. To say that an interaction $A{:}B$ is analogous to $\mathbf{S} = \{A^1{:}B^1, A^2{:}B^2, \ldots, A^N{:}B^N\}$ is to define a measure of similarity between the pair and the set of pairs. However, this similarity is not (directly) given by the information contained in the distribution of objects $\{A^i\} \subset \mathcal{A}$, $\{B^i\} \subset \mathcal{B}$, but by the *mappings* classifying such pairs as being linked:

> **Bayesian analogical reasoning formulation**: Consider a space of latent functions in $\mathcal{A} \times \mathcal{B} \rightarrow \{0, 1\}$. Assume that A and B are two objects classified as linked by some unknown function $f(A, B)$, i.e., $f(A, B) = 1$. We want to quantify how similar the function $f(A, B)$ is to the function $g(\cdot, \cdot)$, which classifies all pairs $(A^i, B^i) \in \mathbf{S}$ as being linked, i.e., $g(A^i, B^i) = 1$. The similarity should be a function of the observations $\{\mathbf{S}, A, B\}$ and our prior distribution over $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$.

Such a similarity will be defined through a Bayes factor, as explained next. For simplicity, we will consider a family of latent functions that is parameterized by a finite-dimensional vector: the logistic regression function with multivariate Gaussian priors for its parameters.

[1] One of the reasons why ETS has dropped analogical reasoning questions from its college admission exams in the United States.

[2] To borrow an example from a clustering task (Ghahramani and Heller, 2005), words such "republican" and "US president" seem to express the concept implicit in the set {"Bush", "Nixon", "Reagan"}, while the set {"Bush", "Putin", "Blair"} suggests concepts such as "current world leader".

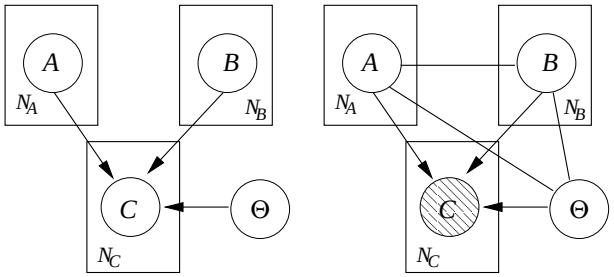

Figure 1: (a) Graphical plate representation for the relational Bayesian logistic regression, where N_A, N_B and N_C are the number of objects of each class, where N_C is equal to the number of elements in L_{AB}. (b) Extra dependencies induced by further conditioning on C are represented by undirected edges.

For a particular pair ($A^i \in \mathcal{A}$, $B^j \in \mathcal{B}$), let $X^{ij} = [\Phi_1(A^i, B^j)\ \Phi_2(A^i, B^j) \dots \Phi_K(A^i, B^j)]^T$ be a point on a feature space defined by the mapping $\Phi : \mathcal{A} \times \mathcal{B} \to \Re^K$. Let $C^{ij} \in \{0, 1\}$ be an indicator of the existence of a link between A^i and B^j in the database. Let $\Theta = [\theta_1, \dots, \theta_K]^T$ be the parameter vector for our logistic regression model

$$P(C^{ij} = 1 | X^{ij}, \Theta) = \text{logistic}(\Theta^T X^{ij}) \tag{1}$$

where $\text{logistic}(x) = (1 + e^{-x})^{-1}$. Our measure of similarity for a pair (A^i, B^j) with respect to a query set $\mathbf{S}$ is the probabilistic similarity measure of "Bayesian sets" (Ghahramani and Heller, 2005) on a log-scale:

$$\begin{array}{rcl} score(A^i, B^j) & = & \log P(C^{ij} = 1 | X^{ij}, \mathbf{S}, \mathbf{C}^{\mathbf{S}} = 1) \\ & - & \log P(C^{ij} = 1 | X^{ij}) \end{array} \tag{2}$$

where $\mathbf{C}^{\mathbf{S}}$ is the vector of link indicators for $\mathbf{S}$: i.e., $C^1 = 1, C^2 = 1, \dots, C^N = 1$ indicates that all pairs in $\mathbf{S}$ are linked.

The general framework is as follows. We are given a relational database $(\mathbf{D}_A, \mathbf{D}_B, \mathbf{L}_{AB})$, where the first two components of this database are sampled respectively from $\mathcal{A}$ and $\mathcal{B}$. Relationship table $\mathbf{L}_{AB}$ is a binary matrix assumed to be generated by a logistic regression model of link existence. A query proceeds according to the steps below:

1. the user selects a set of pairs $\mathbf{S}$ that are linked in the database;
2. the system performs Bayesian inference to obtain the corresponding posterior distribution for Θ, $P(\Theta | \mathbf{S}, \mathbf{C}^{\mathbf{S}})$, given a Gaussian prior $P(\Theta)$;
3. the system iterates through all linked pairs, computing for each pair the integrals

$$\begin{array}{l} P(C^{ij} = 1 | X^{ij}, \mathbf{S}, \mathbf{C}^{\mathbf{S}} = 1) = \\ \int P(C^{ij} = 1 | X^{ij}, \Theta) P(\Theta | \mathbf{S}, \mathbf{C}^{\mathbf{S}} = 1)\, d\Theta \end{array} \tag{3}$$

$$P(C^{ij} = 1 | X^{ij}) = \int P(C^{ij} = 1 | X^{ij}, \Theta) P(\Theta)\, d\Theta \tag{4}$$

4. given the value of such integrals, the system sorts pairs according to the score in Equation (2). This sorted list is the output.

Since the integral used in the Bayesian logistic function (3) does not have a closed formula, in all of these expressions we use the Bayesian variational approximation by Jaakkola and Jordan (2000). For completeness, we briefly review this approach in the Appendix.

The corresponding plate model is illustrated in Figure 1(a). Latent parameter vector $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}^T$ and objects A and B are ancestors of link indicator C. By conditioning on $C = 1$, elements of Θ will be connected to and share information from input data $\{A, B\}$, as in Figure 1(b). This information can be passed forward to evaluate other points. The suggested setup scales as $O(K^3)$ due to the matrix inversions necessary for (variational) Bayesian logistic regression (Jaakkola and Jordan, 2000). If necessary, a further approximation for $P(\Theta | \mathbf{S}, \mathbf{C}^{\mathbf{S}})$ might be imposed if the dimensionality of Θ is too high.

2.1 Modeling protein-protein interactions

In our molecular biology experiments, we use data that indicates if two pairs of proteins interact or not. This reduces to a binary class problem with two classes: *interaction exists* and *interaction does not exist*. The type of interaction will depend on the data. For instance, the MIPS data indicate co-complex protein pairs. It is hand curated data and does not include information from high-throughput datasets. Since the population of protein-protein interactions depend on the experimental conditions under which they were measured, the subpopulations that our method is meant to find also depend on such conditions.

The approach cannot succeed if the underlying model is not a good classifier of protein-protein interactions. Results from Qi et al. (2006) indicate that logistic regression can provide reasonable predictions for the data we use in our experiments. It is always important to check this condition before using our method, since it assumes from the outset that the data can be reasonably separated into the interaction/no interaction classes.

2.2 Priors and the role of "negative data"

Two important parts of the algorithm still need to be clarified. First, the algorithm described seemingly never uses the "negative" data, i.e., the pairs that do not interact. Second, one needs to decide on a prior for the model, which plays an important role in the proposed approach.

The answers for both questions are interwined. As discussed by Ghahramani and Heller (2005) and Silva et al. (2007), empirical priors are used. To be more precise, in our setup we initially fit a logistic regression classifier using a maximum likelihood estimation (MLE) and our data, obtaining the estimate $\widehat{\Theta}$.

A covariance matrix $\widehat{\Sigma}$ for Θ is set proportional to the MLE estimated covariance:

$$(\widehat{\Sigma})^{-1} = c \cdot \left(\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X}\right) / N \tag{5}$$

where N is the total number of (positive and negative) pairs in our data, $\mathbf{X}$ is the $N \times K$ matrix containing the protein-protein features that were used in the MLE computation. Matrix $\mathbf{M}$ is a diagonal matrix. Entry $(\mathbf{M})_{rr}$ is given by $\widehat{p}(r) \cdot (1 - \widehat{p}(r))$, where $\widehat{p}(r)$ is the MLE predicted probability of positive interaction for the t-th row of $\mathbf{X}$.

The constant c is a user-defined parameter. In our experiments, we set c to be equal to the number of elements in $\mathbf{S}$. The interpretation is that we are using as a prior a sample of size equivalent to our query. Results are fairly insensitive to small variantions of c. In general, for larger queries one might want to limit the size of c, depending on the domain.

The prior for Θ is then set to be the normal $\mathcal{N}(\widehat{\Theta}, \widehat{\Sigma})$. As previously mentioned in Silva et al. (2007), empirical priors are a sensible choice: this is a retrieval, not a predictive, task. Basically, the entire data set is the "population." A data-dependent prior based on the population is quite important for an approach like Bayesian sets, since deviances from the "average" behaviour in the data are useful to discriminate between subpopulations.

Notice this data for prior fitting includes both positively (interacting) and negatively (non-interacting) classified pairs. Therefore, the negative data plays an important role through the prior. Moreover, this indirect role of negative data is a computational asset: the computational bottleneck lies on the Bayesian inference steps of our approach, which do not use the negative data (and which usually are in much larger numbers than the positive pairs). This contrasts with other common approaches (Silva et al., 2007).

A visual intuition on the role of the prior, negative data, and resulting score function is given in Section 4.1.

3 Results

Our approach measures similarity through a function space instead of a feature space. To evaluate the merits of our approach, we designed experiments in which we perform protein-protein retrieval.

In this section, we evaluate our approach on data collected from yeast cells. Our gold standard for protein-protein interactions is the Munich Information Center for Protein Sequences (MIPS) catalog (Mewes and et. al, 2004). Details on the particular data we use are given in Appendix B. Moreover, we make use of the MIPS classification system for proteins in the evaluation criteria described shortly. We also describe competing approaches against which we compare our algorithm, and perform a sensitivity analysis on the importance of the prior distribution and feature space.

3.1 Evaluation metric

Evaluating the relevance of proposed analogical similarities is not a straightforward process, and is prone to subjective assessments and extended discussions, as typically happens in the development of a new taxonomy. In our studies, we propose an objective measure of evaluation that is used to rank different algorithms.

Conside a query set $\mathbf{S}$, and a ranked response list $\mathbf{R} = \{R^1, R^2, R^3, \ldots, R^N\}$ of protein-protein pairs. Every element of $\mathbf{S}$ is a pair of proteins P_i:P_j such that P_i is of class M_1 and P_j is of class M_2, where M_1 and M_2 are classes lying on the bottom of the MIPS classification system. For instance, $M_1 = 67.04.01.02$ (*other cation transporters (Na, K, Ca , NH4, etc.))* and $M_2 = 67.5$ (*transport mechanism*).

The retrieval algorithm that generates $\mathbf{R}$ does not receive any information concerning the MIPS taxonomy. It ranks $\mathbf{R}$ starting from the protein pair that is judged most similar to $\mathbf{S}$, followed by the other protein pairs in the population in decreasing order of similarity. Each algorithm has its own measure of similarity.

The evaluation criterion for each algorithm is as follows: we generate a precision-recall curve (Manning et al., 2007) and calculate the area under the curve (AUC). For more details concerning precision-recall curves, see Appendix C.

Notice that in the experiments that follow, we want to focus on very specific MIPS subclasses. This is solely for evaluation purposes, since such classes are known to the experimenter *but not known to the algorithm*. Our criterion is rather stringent, in the sense it requires a perfect match of each R^I with the MIPS categorization, which is not an unique gold standard for analogical similarity. AUC scores will be lower than in typical information retrieval applications. However, we will adopt this score in our study due to its several advantages:

1. it is readily available and does not rely on further subjective assessments;
2. there are several ways by which a pair R^I might be analogous to the relation implicit in $\mathbf{S}$, and they do not need to agree with MIPS. Still, we postulate that there is a correlation between the MIPS categorization given to $\mathbf{S}$ and relevant subpopulations of protein-protein interactions similar to $\mathbf{S}$. Therefore, the corresponding AUC is a tool for comparing the usefulness of different algorithms;

3.2 Algorithms

We compare our approach against two widely used similarity metrics for information retrieval, and one state-of-the-art method:

- the nearest neighbor measure (NN) with Euclidean distances: for a given query set $\mathbf{S}$ and a given candidate point R^I, the distance between the point and the set is given by the minimum distance between R^I and any individual point in $\mathbf{S}$;
- the cosine distance metric (COS): the distance between any two vectors is taken as the inner product of the normalized vectors, where the normalized vector is obtained by dividing it by its Euclidean norm. To measure the distance between a point and a set, we take the average of the distances;
- the Gaussian Bayesian sets metric (GBSets): Bayesian sets (Ghahramani and Heller, 2005) give state-of-the-art performance for tasks such as retrieval of word concepts and images. Since our features are continuous, we used a variation based on Gaussian models. Details are given in Appendix D.

Because our variation of Bayesian sets is motivated by relational data, we call our approach the relational Bayesian sets method (RBSets), to contrast it with the Gaussian Bayesian set (GBSets) described above.

None of these approaches can be interpreted as measures of analogical similarity, since they do not take into account how the protein pair features (gene expression, in our case) contribute to their interaction[3]. We are not aware of any other measure which does. It is true that a direct measure of analogical similarity is not theoretically required to perform well according to our evaluation metric. However, we will see that in this task our measure can often perform an order of magnitude better than other approaches by reasoning analogically.

[3]As a consequence, none uses negative data. Another consequence is the necessity of modeling the input space, a difficult task given the dimensionality and the continuous nature of the features.

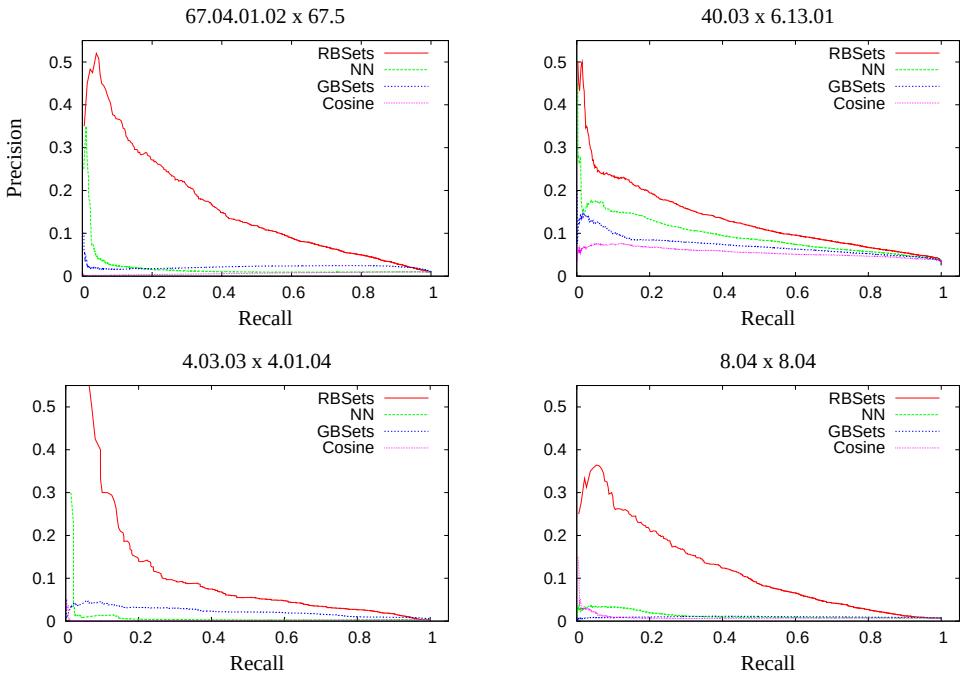

Figure 2: Average precision/recall curves for four different types of queries of size 15. Each curve is an average over 20 random trials.

3.3 Studies

We chose 4 different protein-protein combinations for our benchmark. They were chosen according to the MIPS categorization and shown below, along with the percentage of interacting pairs they represent after we remove the query elements:

- **Query 1:** 67.04.01.02 × 67.5 (i.e., *other cation transporters (Na, K, Ca , NH4, etc.)* × *transport mechanism*), 1% of the interacting pairs;
- **Query 2:** 40.03 × 06.13.01 (i.e., *cytoplasm* × *cytoplasmic and nuclear degradation*), 2.5% of the interacting pairs;
- **Query 3:** 04.03.03 × 04.01.04 (i.e., *tRNA processing* × *rRNA processing*), 0.3% of the interacting pairs;
- **Query 4:** 8.04 × 8.04 (i.e., *mitochondrial transport* × *mitochondrial transport*), 0.7% of the interacting pairs;

For each query evaluation, we randomly choose 15 elements of the given class of pairs and run the 4 algorithms with the selected input. This is repeated 20 times. Figure 2 shows the average precision-recall curves for each query, with the coordinates of each point in the curve being the average of the 20 query results.

As expected, such curves are lower than typical precision-recall curves for the binary classification problem of predicting protein-protein interactions, such as the ones depicted in Qi et al. (2006). A direct comparison between the classification curves and the retrieval curves of Figure 2 is not appropriate, since the classification curves have a well-defined loss function (0/1, for wrong and correct predictions)[4]. The loss function employed in the retrieval problem is not optimal, but chosen for the reasons pointed in Section 3.1.

We can see how much better RBSets performs when compared against different approaches. Table 3.3 summarizes the difference in the area under curve between our approach and the others. All differences are significant under a sign test at a 0.01 level (all differences are all positive).

[4]It is also clear that we are dealing with many fewer "positive examples" (i.e., the selected query size) than in a common binary classification setup. In Qi et al.'s setup, there are thousands of "positive examples" against 15 of ours.

Query	RBSets - NN	RBSets - GBSets	RBSets - Cosine
1	0.14 (0.05)	0.14 (0.05)	0.16 (0.05)
2	0.04 (0.03)	0.06 (0.02)	0.08 (0.02)
3	0.10 (0.03)	0.08 (0.03)	0.10 (0.04)
4	0.11 (0.04)	0.12 (0.04)	0.12 (0.04)

Table 1: Differences in the area under the curve between our algorithm and each of the other three algorithms. See Section 3.3 for details. Each entry contains the average (over 20 trials) and the respective standard deviations in parenthesis. The areas under the curve for our algorithm are $(0.17, 0.14, 0.10, 0.13)$, with respective standard deviations of $(0.05, 0.02, 0.03, 0.04)$.

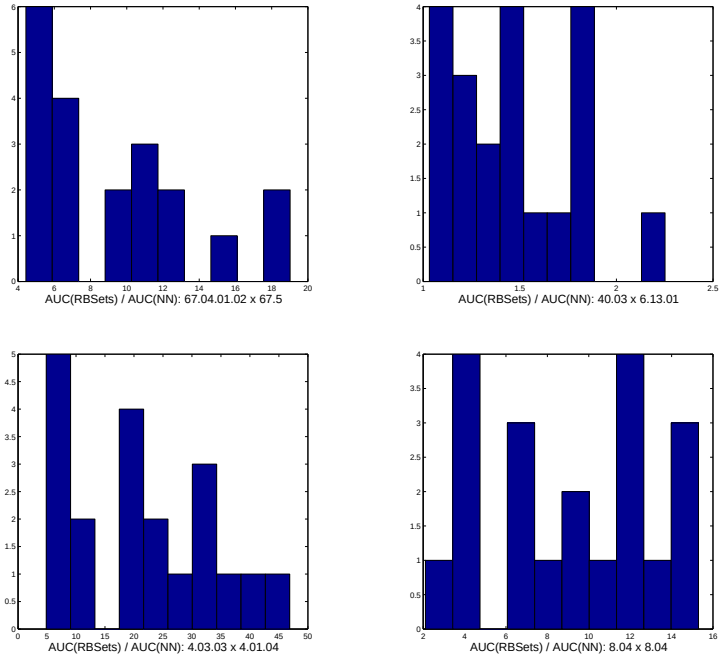

Figure 3: Histograms of the ratio AUC(RBSets) / AUC(NN) for the four different types of queries.

The limited performance of Gaussian Bayesian sets can be attributed not only to the relational nature of the data, which this model was not designed for, but also to the multimodality of the distribution of a few variables. However, removing these variables did not alter significantly the behavior the algorithm, which now might be due to the loss of information given by these variables.

It is also interesting to visualize the distribution of the ratio statistics. For the nearest neighbor algorithm, we computed the ratio between the area under curve of RBSets and the area for NN in each of the 20 trials. The distribution can be visualized in Figure 3. Gains over an order of magnitude are common.

There are some explanations for the rapid degradation of precision in Query 2 ($40.03 \times 06.13.01$). The particular difficulty in this case is the fact that every protein in all queries is also of MIPS types $(5.01, 40.03)$ (with the ribosome biogenesis type (5.01) being one of the most numerous categories in the MIPS data). This is a good illustration of the strictness of our evaluation criterion, since in some sense pairs of type $40.03 \times 06.13.01$ are also of type 40.03×5.01, 40.03×4.03, 5.01×5.01. Had we counted pairs of type $(5.01, 40.03) \times (5.01, 40.03)$ as valid hits, we would have achieved very high precision-recall curves (e.g., close to 100% precision in the top 50 hits), but the query would then be uninteresting due to the high concentration of pairs of this subpopulation. The restriction of perfect matching makes results appear less dramatic, but implicitly it might be ranking relevant subpopulations within $(5.01, 40.03) \times (5.01, 40.03)$. Only further studies on determining the significance of such pairs might provide a more fair evaluation.

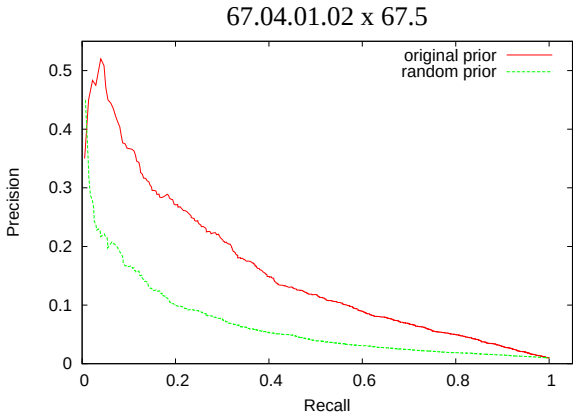

Figure 4: Degradation of performance by using a random prior instead of the empirically derived one.

3.3.1 Sensitivity analysis: prior

In order to evaluate the importance of the empirical prior, we re-run the 20 experiments of Query 1 using a random prior. This prior uses the same covariance matrix $\widehat{\Sigma}$ from our original definition, but the mean parameter $\widehat{\Theta}$ is now a random vector, sampled from a standard Gaussian with independent variables.

The result is illustrated in Figure 4, where we replicate the curve obtained with the original prior (average over 20 trials) with the curve obtained with the same queries for the random priors. The average area under the curve is 0.07, contrasted with the 0.17 area for the original prior.

3.3.2 Sensitivity analysis: feature set

Ultimately, the success of our approach will depend on a good predictive model of interaction existence. We perform some preliminary studies on how a more flexible classifier could improve the results presented here. In order to achieve a more flexible discriminant with logistic regression, we expand our feature set to include other non-linear transformations of the data. We used all second-order monomials that can be obtained from our original features (e.g., for a feature set $\{x, y\}$, the resulting expanded feature set is $\{x^2, y^2, xy\}$). Unfortunately, this did not increase the evaluation statistic compared to our simpler model. This might be a limitation of the information contained in the data. However, further research on integrating more flexible discriminants than logistic regression are likely to improve results in general. How to incorporate nonparametric classifiers in a computationally efficient way for our task is an interesting open question.

4 Discussion

In biology, there is a historical division between analogies and homologies (Griffiths et al., 2002, p. 622). Homologies are similarities due to physical structure. One example of physical structure is the genotype. In this case, the common genetic ancestry between the wing of a bat and the arm of a human make them genetically homologous structures. In constrast, analogies are similarities due to functionality, such as the wings of bats and wings of birds, which do not share common structural genetic ancestry. In the light of such definitions and that the interaction of proteins follows from their function, we postulate that functional (interaction) similarities of interest should be analogical in nature, and derive measures of similarity accordingly.

It is fair to say that structural features ultimately underly functional similarity, depending on how such features are defined. However, given a set of structural features that are fixed a priori, the importance of each feature will vary with respect to the target interaction. For instance, consider a SAT-like exam where for a given pair (say, *water*:*river*) we have to choose (out of 5 pairs) the one that best matches the type of relation implicit in such a query. In this case, it is reasonable to say $car{:}traffic$ would be a better match than (the somewhat nonsensical) *soda*:*ocean*, since cars flow through traffic,

and so does water through a river. The feature that water, river, soda and ocean are either liquids or liquid bodies is not relevant in this relation.

In this sense, one can say that weights given to features of objects (proteins) in our similarity measure can only be assigned based on the extent to which such features are useful to predict the existence of the target relation (protein-protein interactions). We formalized this idea with our approach in Section 2. We provided evidence of its usefulness in Section 3.

4.1 An illustration

A clearer understanding of the algorithm and the role of negative data can be obtained by visualizing a particular example.

We generated data from two classes. The data is represented in Figure 5(a), where the x's are the positive data points, and the triangles the negative points. Positive points were generated over two independent circles of radius 0.15. To facilitate visualization, positive and negative regions do not overlap.

Let $\{X_1, X_2\}$ be the original space. We use the Bayesian logistic classifier with feature space $\Phi(X) = \{X_1, X_2, X_1^2, X_2^2, X_1X_2\}$. A possible empirical prior, which uses the MLE estimated covariance instead of Equation (5), is shown in (b). The role of the negative data should be clear from Figure (c): while the positive datapoints are the same as in (a), a different distribution of negative points induces a completely different empirical prior. In Figure 5(d), we show the prior suggested in Section 2.2, with $c = 1$. Increasing c brings this prior closer to the one in (b), with equality attained at $c = N$.

When conditioned on a query, the probability mass moves around, as exemplified by Figure 5(e) and (f). A query set $\mathbf{S}$ of 5 points points is depicted within small circles in Figure 5(e). The contours in (e) correspond to different scores (Equation 2) for points in this input space, when conditioned in $\mathbf{S}$. Notice that this score is a function of the prior boundary and the query set. The top 20 data points according to our score are show within diamonds in Figure (e). Figure (f) shows the posterior distribution given $\mathbf{S}$ (corresponding to Equation 3).

Figures 5(g) and (h) depict the results for a different query. Again the query points are depicted as small circles in (g), with the top 20 points within small diamonds[5]. Notice that in both (e) and (g), the posterior space defines a distance function that is very different from what could be expected from Euclidean distance.

This can be visualized in Figure 6. This distance, measured with respect to the same query set of Figure 5(g), uses a type of nearest neighbor measure. The result consists on concentric ellipses. This measure is more favorable to the southmost points than ours. This is not in accord to our prior decision boundary (Figure 5), which treates the curvature at the south points rather differently from the curvature at the southwest points, where our query lies.

A geometric intuition of our analogical reasoning formulation can be summarized as follows: the query points define a new curvature on the decision surface (the probability of interaction surface). This new surface results from the initial prior surface, with the space being bent by the introduction of the query points. The regions that change most towards this new curvature are those whose relationship best matches the one defined by the query.

4.2 Related work

To define an analogy is to define a measure of similarity between structures of related objects (pairs of proteins, in our case). A key aspect of this is that, typically, we are not interested in how each individual object (protein) in a candidate pair (protein-protein pair) is similar to individual objects in the query pairs. This concept of assessing analogical similarity through a measurement of relational similarity instead of object similarity was formalized by the Structure Mapping Theory (SMT) of Gentner (1983). SMT eventually became an influential work on analogical reasoning in the artificial intelligence and cognitive science literature (French, 2002).

There is a large literature on analogical reasoning in artificial intelligence and psychology. We refer to French (2002) for a survey. Other recent references are discussed in Silva et al. (2007). The graphical model formulation of Getoor et al. (2002) incorporates models of link existence in relational databases,

[5]For this example, we illustrate a query that does not correspond to points in the original dataset.

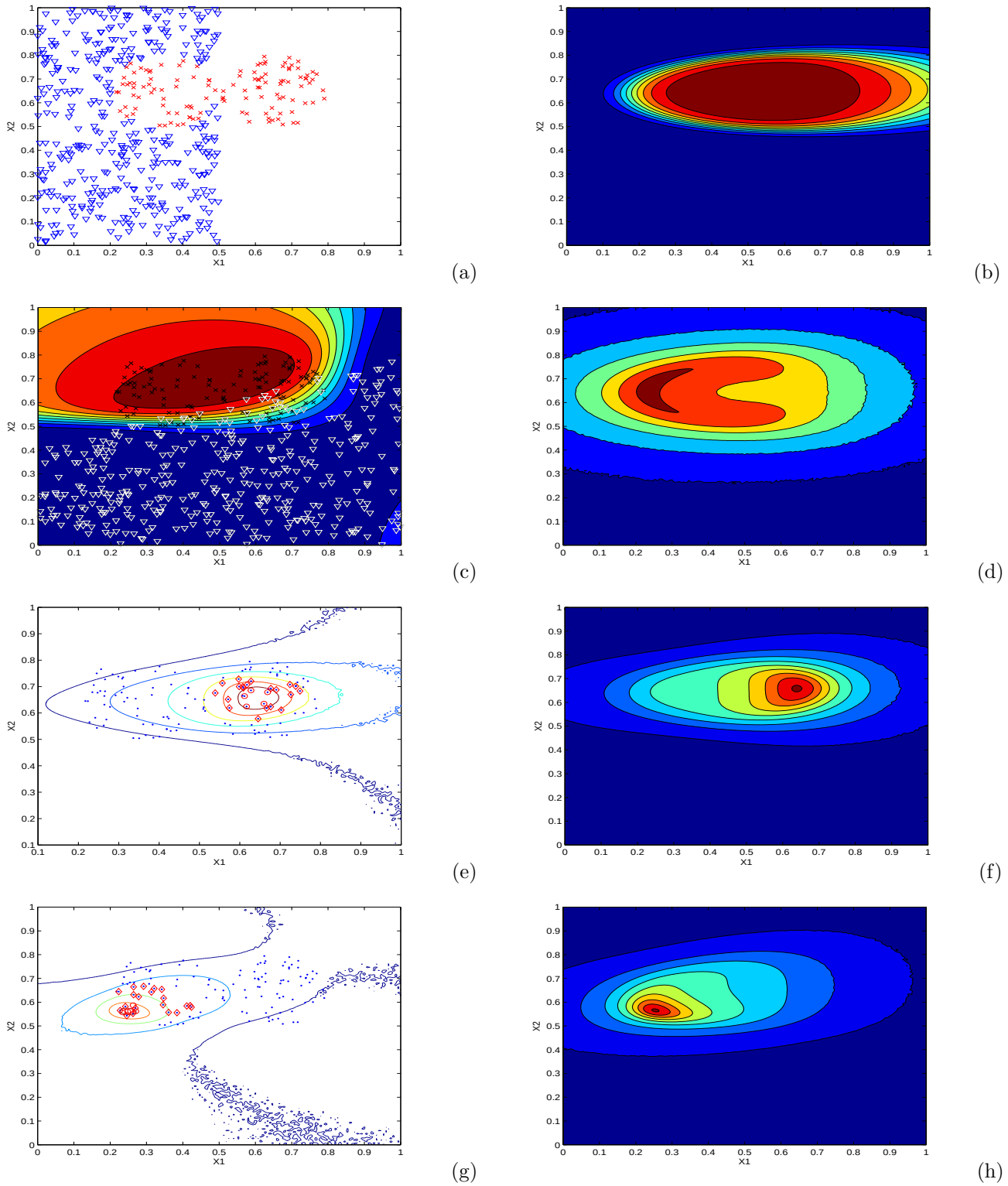

Figure 5: Contour plots for different priors and query-dependent posteriors in a two-dimensional input space. Color varies from blue to red as the probability of positive class increases. In (a), positive points are indicated by x's, and negative ones by triangles. In (b), the prior that is obtained using the maximum likelihood estimates (MLE) as parameters. In (c), a dataset that has a different distribution for negative data, but the same for positive data, induces a completely different empirical prior using the same MLE criterion. In (d), the prior obtained when we used the smoothed MLE parameters described in Section 2.2 (we use $c = 1$ in this case). In (e), a contour of the scores (Equation 2) obtained by a particular query set and the prior in (d). The query points are depicted as circles, and the top 20 positive data points are depicted within diamonds. In (f), the posterior distribution (Equation 3) induced by such a query. Another query is depicted in the last row: (g) shows the score contours, and (h) shows the posterior contours. All values obtained with the variational approximation.

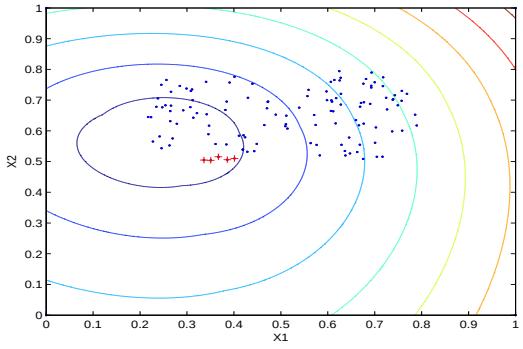

Figure 6: The same query of Figure 5(g) results in very different score contours if an Euclidean nearest neighbor criterion is used (described in Section 3.2). Points indicated with crosses are some of the points whose score are relatively inflated when compared to our metric.

an idea used explicitly in this work. In the clustering literature, the probabilistic approach of Kemp et al. (2006) is motivated by principles similar to those in our formulation: the idea is that there is an infinite mixture of subpopulations that generates the observed relations. See (Silva et al., 2007) for a detailed comparison.

4.3 What this approach is not about

To emphasize once more, our focus here is not on predicting the presence or absence of links, as in, e.g., (Qi et al., 2006) but rather on retrieving similar links from among those already assumed to exist in the relational database. Neither is our focus to provide a fully unsupervised clustering of the whole database of pairs (as in, e.g., Airoldi et al., 2006; Kemp et al., 2006).

Although the problem of predicting protein-protein interactions is popular in bioinformatics literature, it has little to do with the task we describe in this paper. Qi et al. (2006), for instance, use machine learning techniques to predict if a particular pair interacts or not. We are not solving this problem. Our very different setup assumes from the outset that the protein-protein interactions are given. The problem we tackle in this paper is a type of information retrieval or *clustering on demand* problem, where we search over the space of already linked proteins. Our goal is to discover, given a set of interacting pairs (a "cluster"), which other interacting pairs are plausible elements of this set.

5 Conclusion

We presented a novel measure of similarity between biological structures based on the principe of analogical comparison. It provides a way of clustering biological data that is considerably different from other methods, due to its focus on analysing the space of functions that map object features to their relations, instead of the feature space itself. For small size queries, our method find analogies that are functionally relevant among the top matches

This work can be expanded in many ways, including but not limited to: allowing for extra dependencies between interactions that are not due to input features X; scaling up the algorithm to allow for higher-dimensional data; apply it to other domains such as evaluating analogies between cells from different species. We believe several useful variations of our approach can be designed in the future.

Acknowledgement

We thank Zoubin Ghahramani and Olga Troyanskaya for helpful comments. EA is supported by NIH grant R01 GM071966 and NSF grant IIS-0513552 to Olga Troyanskaya at Princeton.

References

E. Airoldi, D. Blei, E. Xing, and S. Fienberg. Mixed membership stochastic block models for relational data with application to protein-protein interactions. *Proceedings of the International Biometrics Society Annual Meetings*, 2006.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubinand, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:155–170, 2000.

J. S. Bader. Greedily building protein networks with confidence. *Bioinformatics*, 19:1869–1874, 2003.

R. French. The computational modeling of analogy-making. *Trends in Cognitive Sciences*, 6:200–205, 2002.

D. Gentner. Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, 7:155–170, 1983.

L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of link structure. *Journal of Machine Learning Research*, 3:679–707, 2002.

Z. Ghahramani and K. Heller. Bayesian sets. *18th NIPS*, 2005.

A. Griffiths, W. Gelbart, R. Lewontin, and J. Miler. *Modern Genetic Analysis*. Freeman & Co, 2nd edition, 2002.

K. Heller and Z. Ghahramani. Bayesian hierarchical clustering. *Technical Report 2005-002, Gatsby Computational Neuroscience Unit*, 2005.

T. Jaakkola and M. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

C. Kemp, J. Tenenbaum, T. Griffths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. *Proceedings of AAAI'06*, 2006.

C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. In press, 2007.

H. Mewes and C. Amid et. al. MIPS: analysis and annotation of proteins from whole genome. *Nucleic Acids Research*, 32, 2004.

E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and Mona Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(Suppl 1):i302–i310, 2005.

Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Bioinformatics*, 63:490–500, 2006.

R. Silva, K. Heller, and Z. Ghahramani. Analogical reasoning with relational Bayesian sets. *11th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2007.

A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Current Opinions in Structural Biology*, 12(3):368–373, 2002.

Appendix

A. Variational method

In this appendix we summarize the variational approximation for Bayesian logistic regression, as described by Jaakkola and Jordan (2000). The task is to compute the integrals of the type shown by Equations (3) and (4).

Let $g(y)$ be the logistic function, $g(y) = (1 + e^{-y})^{-1}$, and consider the case for the single data point evaluation, Equation (4). The method lower-bounds the integrand as follows

$$P(C|X, \Theta) = g(\Theta^T X) \geq g(\xi) \exp\{(H_C - \xi)/2 - \lambda(\xi)(H_C^2 - \xi^2)\} \tag{6}$$

where $H_C = (2C - 1)\Theta^T X$ and $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$, $\tanh(\cdot)$ being the hyperbolic tangent function.

Consider first approximating $P(\Theta|X, C)$, which is obtained by normalizing

$$P(C|X, \Theta)P(\Theta) \geq Q(C|X, \Theta)P(\Theta)$$

where $Q(C|X, \Theta)$ is the expression on the righ-hand side of (6).

Since this bound assumes a quadratic form as a function of Θ and our priors are Gaussian, the approximate posterior will be Gaussian, which we denote by $\mathcal{N}(\mu_{pos}, \Sigma_{pos})$. However, this bound can be loose unless a suitable value for the free parameter ξ is chosen. The key step in the approximation is then to optimize the bound with respect to ξ.

Let the Gaussian prior $P(\Theta)$ be denoted as $\mathcal{N}(\mu, \Sigma)$. The procedure reduces to an iterative optimization algorithm where for each step the following updates are made:

$$\begin{aligned} \Sigma_{pos}^{-1} &= \Sigma^{-1} + 2\lambda(\xi) X X^T \\ \mu_{pos} &= \Sigma_{pos}^{-1}\left[\Sigma^{-1}\mu + (C - 1/2)X\right] \\ \xi &= (X^T \Sigma_{pos} X + (X^T \mu_{pos})^2)^{1/2} \end{aligned} \tag{7}$$

To approximate an integral such as (3), a sequential update is performed: starting from the prior $P(\Theta)$ for the first data point (X, C) in $(\mathbf{S}, \mathbf{C^S})$, the resulting posterior $\mathcal{N}(\mu_{pos}, \Sigma_{pos})$ is treated as the new prior for the next point. The ordering is chosen from an uniform distribution in our implementation.

Finally, given the optimized approximate posterior, the predictive integrals (3) (and, analogously, (4) can be approximated as

$$\begin{aligned} \log Q(C^{ij}|X^{ij}, \mathbf{S}, \mathbf{C}) &= \log g(\xi_{ij}) - \xi_{ij}/2 + \lambda(\xi_{ij})\xi_{ij}^2 \\ &\quad -\tfrac{1}{2}\mu_{\mathbf{S}}^T \Sigma_{\mathbf{S}}^{-1} \mu_{\mathbf{S}}^T + \tfrac{1}{2}\mu_{ij}^T \Sigma_{ij}^{-1} \mu_{ij}^T \\ &\quad +\tfrac{1}{2}\log(|\Sigma_{ij}^{-1}|/|\Sigma_{\mathbf{S}}^{-1}|) \end{aligned} \tag{8}$$

where parameters $(\mu_{\mathbf{S}}, \Sigma_{\mathbf{S}})$ are the ones in the approximate posterior $\Theta|(\mathbf{S}, \mathbf{C^S}) \sim \mathcal{N}(\mu_{\mathbf{S}}, \Sigma_{\mathbf{S}})$, and (μ_{ij}, Σ_{ij}) come from the approximate posterior $\Theta|(\mathbf{S}, \mathbf{C^S}, X^{ij}, C^{ij})$. Parameters $(\xi_{ij}, \xi_{\mathbf{S}})$ come from the respective approximate posteriors.

B. Features and data

We use the data collected by Qi et al. (2006). Their MIPS data consists of 8236 positive protein-protein interactions, and approximately 230,000 pairs that are labeled as negative examples. There are 162 attributes, 20 of which are derived from gene expression data. Details on how this data was collected is given in the reference.

In our studies, we make use of the gene expression data only: most of the remaining attributes contained a sizeable proportion of missing data[6]. Although there are standard ways of dealing with missing data in logistic regression, we wanted to avoid adding another source of variance to our first results. Better ways of dealing with missing data will be treated in a future extension of this paper.

Given the 20 attributes of a given pair P_i:P_j, we non-linearly transform them as follows: we first calculate the Euclidean norm of the vector of 20 attributes, and normalize the vector by dividing each entry by this norm. This normalization is reminiscent of the cosine distance metric (Manning et al., 2007), a common distance metric in information retrieval applications. On top of the normalized 20 attributes, we also included the squared value of each of the 20 original variables as another source of non-linear transformation.

C. Precision-recall curves

The curve is generated by scanning $\mathbf{R}$ in a sequential manner. For each element $R^I \in \mathbf{R}$, we calculate

- its corresponding *precision*, that is, the number of elements in $\{R_1, R_2, \ldots, R_I\}$ that are either of M_1:M_2 or M_2:M_1 class[7] (we assume assymetric interactions are being measured), divided by I;
- its corresponding *recall*, that is, the number of elements in $\{R^1, R^2, \ldots, R^I\}$ that are either of M_1:M_2 or M_2:M_1 class, divided by the total number of elements in $\mathbf{R}$ that are of such classes;

The precision/recall of each R^I defines a point in a space with *recall* in the x-axis and *precision* in the y-axis.

D. Gaussian Bayesian sets

The most common implementation of Bayesian sets is based on binary data. Given the relatively small dimensionality of our space (20 features), binarization is unlikely to produce any meaningful model. Instead, we will treat the data as continuous, with a multivariate Gaussian model. The prior for the parameters is a normal-Wishart prior with a mean and covariance hyperparameters. We set those

[6] Those attributes that are mostly complete are either poor predictors − as analyzed by Qi et al. − or are Gene Ontology (GO) features. We did not want to add GO features to our study, since this would bias our results towards a particular human-designed taxonomic system.

[7] Each protein in MIPS belongs to several classes. We consider P_i to be of class M_j if M_j is one of its classes.

hyperparameters proportional to the maximum likelihood estimator using the positive pairs. Two user-defined parameters, r and υ are needed: r is a multiplication factor for the mean hyperparameter, and υ is a multiplication factor for the inverse covariance matrix hyperparameter (Heller and Ghahramani, 2005). We set $r = 20$ and $\upsilon = 1$ to achieve broad priors.