Metric Labeling and Semi-metric Embedding for Protein Annotation Prediction

Emre Sefer and Carl Kingsford
Department of Computer Science, University of Maryland, College Park
{esefer,carlk}@cs.umd.edu

Abstract

Computational techniques have been successful at predicting protein function from relational data (functional or physical interactions). These prediction techniques have been used to generate hypotheses and to direct experimental validation. With few exceptions, these predictive tasks are modeled as multi-label classification problems where the labels (functions) are treated independently or semi-independently. However, databases such as the Gene Ontology provide more information about the similarities between functions. It is a largely open question how much the use of relationships between functions can improve the quality of function prediction techniques. In this paper, we explore the use of the Metric Labeling combinatorial optimization problem to make use of heuristically computed distances between functions to make more accurate predictions of protein function in networks derived from both physical interactions and a combination of other data types. To do this, we give a new technique (based on convex optimization) for converting heuristic semimetric distances (from, e.g. Gene Ontology) into a metric that finds an embedding of the semimetric into a metric with minimum least-squares distortion (LSD). The Metric Labeling approach is shown to outperform 5 existing techniques for inferring function from networks. These results suggest Metric Labeling is useful for protein function prediction, and that our LSD minimization approach can help solve the problem of converting heuristic distances to a metric.

1. Introduction

Networks encoding pairwise relationships between proteins have been widely used for protein function prediction and for data aggregation, display, and visualization. Sometimes these networks are derived from a single data source such as protein-protein interactions [19, 25, 37, 41, 47]. In other instances, they are constructed from the integration a large collection of experiments involving different data types, such as gene expression [e.g. 17], protein localization [e.g. 24], protein complex dataset [e.g. 18, 23] etc. The precise meaning of an edge in these can differ, but their common feature is that two proteins connected by an edge often have similar functions. By extension, these networks globally generally have the property that two proteins that are "close" in the network are more likely to have closely related functions. This correlation has given rise to a number of computational approaches to extract hypotheses for protein function from relational data [13, 22, 26, 40, 44, 45].

Nearly all of these computational methods treat the function prediction problem as a labeling problem, where the labels are drawn from a hand-curated vocabulary of biological functions or processes. Further, they typically ignore any relationships between the functions, treating them as independent labels. However, there are usually relationships among functions that ought to be useful to make more accurate predictions of protein function. For example, the Gene Ontology (GO) [1] is a manually curated database of biological functions and processes that represents the hierarchical relationships among different functions as a DAG. However, most prediction methods have ignored such a structure.

In the few cases it has been done, integrating Gene Ontology knowledge into protein function prediction methods [4, 13] and clustering [11] has resulted in improved predictions. For example, Barutcuoglu et al. [4] developed a Bayesian framework for combining multiple SVM classifiers

based on the GO constraints to obtain the most probable, consistent set of predictions. Their approach used a hierarchy of support vector machine (SVM) classifiers trained on multiple data types. By GO constraints, this method also exploits the relationship between functions in GO but does not exploit distances between functions directly. Taking another approach, Deng et al. [13] uses the correlations between which proteins are labeled with each functions but they estimate these correlations from training data and don't consider GO structure.

1.1 Metric Labeling for Function Prediction

Here, we propose to integrate Gene Ontology relationships with relational data by modeling the protein function prediction problem as an instance of Metric Labeling [29] which is a special case of MRF in which the distance function among labels X is a metric. The Metric Labeling problem seeks to assign labels (here, protein functions) to nodes in a graph (here, proteins or genes) to minimize the distance (in the metric) between labels assigned to adjacent nodes. The advantage of this formulation is that rather than treating function labels as independent, unrelated entities, their similarities can be directly incorporated into the objective function. A more detailed description of the Metric Labeling problem is given in Section 2.1.

The Metric Labeling formulation can be seen as an generalization of minimum multiway cut. For example, the minimum multiway cut formulation [48] implicitly assigns distance 0 between two identical functions and distance 1 between any pair of distinct functions. Metric Labeling softens this to account for varied levels of similarities between the functions. Metric Labeling can also be seen as special case of Markov Random Field approach. Markov Random Fields encode the same combinatorial problem, but the distance function is not restricted to metrics or semimetrics [30]. However, MRF optimization with such arbitrary distance functions is NP-Hard [12, 30], and there is no approximation algorithm that can approximate the global optimum within a non-trivial bound when the distance is arbitrary. In that case, only local optimum can be computed [30]. However, in practice, the distance function is either a metric or close to a metric most of the time, and Metric Labeling becomes a reasonable approach because there are practical approximation algorithms for Metric Labeling with logarithmic approximation guarantees [9, 10, 29]. In this paper, we will use the integer programming formulation by Chekuri et al. [9] which yields an $O(\log k)$ approximation algorithm for Metric Labeling where k is number of labels.

1.2 Constructing a Metric Distance Between GO Functions

Metric Labeling (and MRF models) have been typically been used in applications related to computer vision [5, 6, 30, 36, 49] where often the distance between the labels naturally can be expressed by metrics or a special metrics which can be approximated better. However, in the case of function prediction from relational data, while heuristic relationships between functions can be readily computed from the structure of the Gene Ontology graph, it is more difficult to make these distances obey the requirements of a metric. Recall that a **metric** $d(\cdot, \cdot)$ over items X satisfies the following 4 properties for all x, y, z in X:

$$d(x, y) \geq 0 \qquad \text{(Nonnegativity)} \tag{1}$$

$$d(x, y) = 0 \text{ if and only if } x = y \tag{2}$$

$$d(x, y) = d(y, x) \qquad \text{(Symmetry)} \tag{3}$$

$$d(x, z) \leq d(x, y) + d(y, z) \qquad \text{(Triangle Inequality)} \tag{4}$$

Typically, properties (1)–(3) can be easily satisfied, but often natural distance measures do not satisfy the triangle inequality (4). When d satisfies all of them except triangle inequality (4), it becomes a **semimetric**.

To apply Metric Labeling when the distance function on the labels is merely a semimetric, we will first convert the semimetric into a metric that is as similar to the semimetric as possible. Approximating a semimetric by a close metric and MRF optimization when the distances are semimetric are topics of recent interest and Kumar and Koller [34] have recently suggested an algorithm based on minimizing the distortion. If $\mathcal{S}$ is a semimetric, and $\mathcal{M}$ is a metric approximating $\mathcal{S}$, *contraction* of this mapping is the maximum factor by which distances are shrunk and *expansion* or stretch of this mapping is the maximum factor by which distances are stretched. *Distortion* of this approximation is the product of the contraction and the expansion. Although distortion minimization has traditionally been used in metric embeddings, distortion by its definition tells us about the maximum expansion and contraction and doesn't tell us about the distribution of the error over all the distances. In other words, in low-distortion approaches, only the error introduced in the largest outlier is considered. However, this doesn't take the distribution of errors into account. For noisy, imperfect data that is far from a metric, intuition indicates that minimizing the error introduced in the other distances would yield a better metric.

To design metric approximations to semimetrics that better preserve all distances, we propose a least-squared minimization algorithm that tries to minimize the total squared error among all distances. To contrast it with traditional distortion, we call this approach *least squared distortion* (LSD). This problem can easily be solved in polynomial time [28, 32] because it is a convex case of quadratic programming. Thus, to apply Metric Labeling in cases when the distances among the labels are not metric, we first map the semimetric to a close metric using the LSD algorithm and then run Metric Labeling on the new metric. Experiments on protein function prediction suggest this is a good metric approximation method. The issue of converting a set of heuristically estimated distances arises in many practical contexts and the LSD approach may also be useful for other applications.

1.3 Improvement in Function Prediction

We test the LSD algorithm and the Metric Labeling approach for function prediction on relational data for 7 species: *S. cerevisae*, *A. thaliana*, *D. melanogaster*, *M. musculus*, *C. elegans*, *S. pombe* and Human. For *S. cerevisae*, we apply the algorithms to an integrated data set that derives pairwise relationships between proteins from several lines of evidence such as gene expression, protein localization data, and known protein complexes. For all 7 species, we also test the approaches on networks derived from high-throughput protein-protein interaction experiments.

The algorithms are tested in a variety of settings. The set of functional labels are drawn from the Gene Ontology. The number of considered GO terms is varied between 90 and 300 in order to evaluate the effect of the size and specificity of the label set on performance. The specific GO terms are selected to match sets of terms used in previous publications. Various metrics and semimetrics relating the GO terms are also tested. A simple shortest-path metric is compared with two other semimetrics derived from lowest common ancestor in the Gene Ontology DAG, semimetrics computed from a training set of labels and semimetrics computed from both training set and GO. See Section 2.4 for more details.

1.4 Our Contributions

We introduce the use of Metric Labeling for protein function prediction from relational data and show that under many reasonable metrics it outperforms the approaches based on Markov Random Fields [35], Functional Flow [40], minimum multiwaycut [27, 48], neighborhood enrichment [22], and simple majority rule [44]. We extensively test on 7 species in both protein-protein and integrated networks using several different collections of GO terms. The results indicate that the clean, simple Metric Labeling formulation is useful for automated function prediction.

In addition, we introduce the LSD objective function for finding a metric that approximates a semimetric with the goal of preserving many distances rather than just limiting the maximum distortion. The convex optimization approach for this problem may be useful in other contexts where reasonable heuristic distances do not satisfy the triangle inequality. We compare the performance of running first our LSD metric approximation algorithm and then running METRIC LABELING on the LSD's output metrics with a recent algorithm by Kumar and Koller [34] and METRIC LABELING with LSD metric approximation seems to result in better predictions.

2. Methods

2.1 The Metric Labeling Problem

The METRIC LABELING problem has been extensively investigated from a theoretical point of view Chekuri et al. [9, 10], Kleinberg and Tardos [29]. Formally, we have a graph $G = (P, E)$ over a set P of n nodes (here, proteins) and a set L of k possible labels (here, functions) that we want to assign to objects. We have a metric $d(\cdot, \cdot)$ satisfying properties (1)–(4) defined between any labels in L. We are also given a function $c(p, \ell)$ that provides the cost of assigning label $\ell \in L$ to $p \in P$. METRIC LABELING seeks an assignment $f : P \to L$ of labels to proteins that minimizes the objective function:

$$Q(f) = \sum_{p \in P} c(p, f(p)) + \sum_{(p,q) \in E} w_{pq} d(f(p), f(q)) \tag{5}$$

where $w_{pq} = w_{qp}$ is the weight of the edge between proteins p and q in the graph. The first summation is called the *assignment costs* and depends only on individual choice of label we make for each protein and second summation is called the *separation costs* and is based on the pair of choices we make for two interacting proteins.

The intuition is that pairs of proteins that are highly related (w_{pq} is high) ought to be assigned labels that are highly similar ($d(f(p), f(q))$ is low). The assignment costs prevent the problem from becoming trivial by forbidding the assignment of the same label to every protein. For a protein p with a known function b, typically $c(p, b)$ will be 0 and $c(p, \ell) = \infty$ for all $\ell \in L$ except b.

2.2 Integer Programming Formulation of Metric Labeling

The METRIC LABELING problem defined above can be written as an ILP as in Chekuri et al. [9]. In this formulation, $x(u, i)$ is binary variable indicating that vertex u is labeled with i and $x(u, i, v, j)$ is binary variable indicating that vertex u is labeled with i and vertex v is labeled with j for edge $(u, v) \in E$. The objective is then to

$$\text{minimize} \sum_{v \in V} \sum_{i \in L} c(u, i) x(u, i) + \sum_{(u,v) \in E} \sum_{i \in L} \sum_{i \in L} w(u, v) d(i, j) x(u, i, v, j) \tag{6}$$

The variables are subject to the following constraints:

$$\textstyle\sum_{i \in L} x(u, i) = 1 \qquad \forall u \in V \tag{7}$$

$$\textstyle\sum_{j \in L} x(u, i, v, j) = x(u, i) \qquad \forall u \in V, v \in N(u), i \in L \tag{8}$$

$$x(u, i, v, j) = x(v, j, u, i) \qquad \forall u, v \in V, i, j \in L \tag{9}$$

$$x(u, i) \in \{0, 1\} \qquad \forall u \in V, i \in L \tag{10}$$

$$x(u, i, v, j) \in \{0, 1\} \qquad \forall (u, v) \in E, i, j \in L \tag{11}$$

Constraints (7) mean each vertex must receive some label. Constraints (8) force consistency in the edge variables: if $x(u, i) = 1$ and $x(v, j) = 1$, they force $x(u, i, v, j)$ to be 1. Constraints (9) express the fact that (u, i, v, j) and (v, j, u, i) refer to the same edge.

Solving this integer programming instance optimally is NP-Complete. Since we are dealing with large networks, we use the $O(\log k)$ approximation algorithm given by Chekuri et al. [9] that is based on rounding the linear programming relaxation to identify a deterministic HST metric [2] approximation of the given metric such that the cost of solution of this HST metric is at most $O(\log k)$ times LP cost on original metric. We implemented and ran the LP formulation in GLPK [20].

We used the rounding scheme described in GenMultiCut algorithm in Section 2.6 to convert fractional assignment costs returned by the LP relaxation of the above ILP. We use the fraction of times that rounding scheme chose a given function for a protein a the probability of annotating this protein with that function.

2.3 Metric Approximation Via Least Square Distortion Minimization

The algorithms suggested above have guaranteed performance bounds when the distance d is a metric. However, finding metric distance in practical contexts can be quite difficult. Ideally, the distance encodes and summarizes a large amount of existing knowledge about the relationship between protein functions. It is likely that such as distance will not satisfy the triangle inequality (and several of the distances we consider below do not).

We define a novel metric approximation algorithm, called LSD, based on minimizing the total least square error between a given semimetric set of distances and the computed metric distances. Least square error approximation is intuitive because the error of every distance contributes to the total error of the metric approximation instead of only the maximum expansion and contraction as in distortion case.

The LSD algorithm is defined as a quadratic program below, where $S = \{s_1, \ldots, s_{\binom{n}{2}}\}$ is the given set of semimetric distances between each pair of n items, and $M = \{m_1, \ldots, m_{\binom{n}{2}}\}$ is corresponding set of metric distances, where for all i, s_i and d_i represent distances between the same pair of proteins. Let $I = \{1, \ldots, \binom{n}{2}\}$ be the set of indices of distances.

To find a good approximation to the distances in S we seek values for the $\{m_i\}$ variables to

$$\text{minimize} \sum_{i \in I} (s_i - m_i)^2.$$

We require that the m_i values satisfy the following constraints for all $i, j, k \in I$ that should be related by the triangle inequality:

$$m_i + m_j - m_k \geq 0 \tag{12}$$

$$m_i + m_k - m_j \geq 0 \tag{13}$$

$$m_k + m_j - m_i \geq 0 \tag{14}$$

The objective function can be written as $1/2x^T Qx + c^T x$ where $n \times n$ matrix Q is symmetric, and c is any $n \times 1$ vector. In our case, the matrix Q is positive definite and if the problem has a feasible solution then the global minimizer is unique. In this case, the problem can be solved by interior point methods in polynomial time [28] and the problem is not NP-hard as in the general case [32]. This is important since it turns the problem into a practical one. We implemented and ran the problem in CGAL [8].

2.4 Metrics and Semimetrics

We test 4 different distance measures between protein functions:

1. $d_{\text{SP}}(x, y)$ = the shortest path distance in the GO DAG between x and y divided by diameter of GO. This is a metric and intuitively simple.

2. $d_{\mathrm{LCA}}(x,y) = (b+c)/(2a+b+c)$, where a is shortest path distance from the root of Ontology to the lowest common ancestor u of x and y and b is the shortest distance from x to the u and c is the shortest distance from y to u. The LCA distance measure does not satisfy triangle inequality and is only a semimetric.
3. $d_{\mathrm{Lin}}(x,y) = (\log \Pr(x) + \log \Pr(y))/(2 \log \Pr(\mathrm{lca}(x,y)))$, where $\Pr(x)$ is the empirical probability that a protein is annotated with x, and lca(x,y) is the LCA of x and y. This is defined in Lin [38] as a similarity measure and we take its reciprocal as a distance. It is similar to the LCA distance above but uses the probabilities in each annotation instead of distances. It has mostly been used in NLP applications Budanitsky and Hirst [7], Lin [39]. However, it has recently been used in other applications of Gene Ontology distances [14, 43] and became useful. It is a semimetric.
4. $d_{\mathrm{KB}}(x,y) = \sum_{p_1 \in P_x} \sum_{p_2 in P_y} \mathrm{sp}(x,y)/(\mathrm{diameter} \cdot |P_x| \cdot |P_y|)$, where P_x and P_y are sets of proteins in the training set annotated with x and y respectively, $\mathrm{sp}(x,y)$ is the shortest path distance between x and y, *diameter* is the diameter of network.

We also consider the combination of the structure-based d_{SP}, d_{LCA}, and d_{Lin} with the knowledge-based d_{KB} using the formula:

$$d_{\mathrm{comb}}(x,y) = (1-\alpha) d(x,y) + \alpha d_{\mathrm{KB}}(x,y)\,, \tag{15}$$

where α is a weight of contribution of training set estimations. For $\alpha < 1$, none of the combined distances are metric (but are semimetric).

When the distance is not a metric (and in almost none of the tested cases is it a metric), we first run the LSD metric approximation algorithm (Section 2.3) to obtain a closer metric and then run Metric Labeling on those metric distances. When it is a metric, we just run Metric Labeling.

In addition, the assignment costs $c(u,i)$ of assigning label i to node u is chosen either to be uniformly 1 or according to the density of a label in a particular region of the graph as follows: We estimated for each protein p and label i cost $c(p,i) = n_p/n_{pi} n_p = 1/n_{pi}$ where n_p and n_{pi} are number of neighbors of p and number of neighbors of p in the training set that have function i respectively. In the case where p has no neighbors with function i, $c(p,i) = 2$.

2.5 Network Data

We tested our algorithm on the protein-protein interaction (PPI) networks of 7 species obtained from BIOGRID [46]: *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *A. thaliana*,*M. musculus*, *H. sapiens*, and *S. pombe*. We used all physical interactions in BIOGRID. Duplicate edges were counted as single edges. We consider only the largest connected component. We used GO annotations downloaded from the Gene Ontology as our true annotations. For *S. cerevisiae* there were 5757 proteins total and 5748 proteins in the largest connected component, of which 5512 had some annotation. There were 54717 interactions in the largest connected component.

For *S. cerevisiae*, we also considered an integrated network derived from several data sources, including gene expression [17], protein localization [24], protein complexes [18, 23], and protein interaction [46]. We used protein complex dataset by assigning binary interactions between any two proteins participating in the same complex, yielding 49313 interactions. For gene expression data, we assigned binary interactions between genes whose correlation in Gasch et al. [17] is greater than 0.8 or smaller than -0.8. We assigned binary interactions between any proteins that are annotated to the same location in Huh et al. [24].

We combined these data sources into one network by using Noisy-Or with their reliability scores, where the interaction score between A and B is taken to be $\mathrm{Score}(A,B) = \prod_{i \in E}(1-r_i)$.

The reliability r_i of each source i was estimated by $r_i = \log \frac{\Pr(L|i)/(1-\Pr(L|i))}{\Pr(L)/(1-\Pr(L))}$, where $\Pr(L \mid i)$ is the probability that the annotation set links the proteins, given they are linked by experiment i and $\Pr(L)$ is the probability of edges being linked in the annotation set.

2.6 Comparison to Other Methods

We run algorithms on Mac which has 2 GHz Intel Core 2 Duo processor and 2 Gb memory. The Metric Labeling algorithm took approximately 15 minutes to run. We compared Metric Labeling predictions with several well-known direct function prediction methods such as:

Majority: Each protein is annotated with the function that occurs most often among its neighbors as described in Schwikowski et al. [44]. The main disadvantage of this method is that full topology of network is not considered. This runs in < 5 seconds on yeast network.

Neighborhood: For each protein, we consider all other proteins within a radius $r = 2$ as described in Hishigaki et al. [22] and a χ^2-test is used determine if each function is overrepresented. We implemented this in Python, and it runs less than 5 seconds on yeast network.

GenMultiCut: This algorithm is described in Vazquez et al. [48] and Karaoz et al. [27]. It tries to cluster the network by minimizing the number of edges between clusters. This algorithm is a simpler version of our algorithm in which distance between two functions are either 1 (if they are not same) or 0 (if they are equal). Hence, it cannot take the relations among functions into account. We followed the same approach by [40] and run the ILP formulation 50 times each time perturbing the weights by a very small offset drawing from uniform distribution on $(-w_{\max}10^{-5},\ w_{\max}10^{-5})$ where w_{max} is the maximum edge weight in the graph. Then probability of assigning a function to a protein will be the fraction of number of annotations of this protein with that function. We implemented this by using MathProg and GLPK, taking less than 1 minute on yeast.

FunctionalFlow: Each function is independently flowed through the whole network according to an update rule and each node is assigned to functions depending on the amount of flow it gets [40]. We re-implemented this method, and it runs less than 2 minutes on yeast network.

MRF: This method is from Lee et al. [35]. It is based on kernel logistic regression which is the improvement over previous MRF models [13, 33]. This method also tries to exploit the relation between different functions by identifying a set of functions that are correlated with the function of interest. However, it doesn't use the structure of GO when estimating the correlation. This approach takes less than 5 minutes to run on yeast network. We modified code provided by the authors to work on our data sets.

We also compare LSD with a recent approach for MAP Estimation under a semimetric:

Semimetric MAP Estimation Algorithm: This algorithm from Kumar and Koller [34] tries to approximate a given semimetric distance function using a mixture of r-hierarchically well-separated tree (r-HST) metrics [3, 15]. Then, it solves each resulting r-HST metric labeling problem. We followed the same approach as in GenMultiCut, run the formulation 50 times by perturbing the edges and assign the fraction of number of annotations of this protein with that function as probability of annotating this protein with that function. We modified code provided by the authors to work on our data sets. It ran in less than a 1 minute on yeast.

2.7 Evaluating Performance

We use fivefold cross-validation to compare the predictive performance of the algorithms. We divided the set of annotations into 5 equal parts and then tried to predict functions for one part

by using GO and remaining 80% of the total annotations as our training set. The d_{KB} distance is computed using only the remaining 80% of annotated proteins each time. All performance measurements are the average of the 5 runs. Each method described in Section 2.6 yields a score, and we assess performance at different false positive rates by varying the score thresholds from 0 to 1 by 0.05 increments.

We varied the number of considered functions from 90 to 300. We used the functions in Kourmpetis et al. [31] as our 90 functions. We extended this set to 150 and 300 elements by randomly choosing from among the functions seen in the network annotations. For each of these collections of functions, we report standard performance measures for classification problems: sensitivity (#TP/(#TP + #FP)) and false-positive rate (#FP/(#FP + #TN)), counting each annotation separately as a prediction. We show the prediction results by ROC Curve Analysis [21].

3. Results

3.1 Function Prediction in Yeast Using a PPI Network

Predictive performance on the yeast PPI network is shown in Figure 1. The curves show that Metric Labeling combined with our LSD metric approximation algorithm performs better than the other tested algorithms with various number of elements in ontology. Metric Labeling is more accurate than GenMultiCut in every case since GenMultiCut ignores the effect of distances between functions. FunctionalFlow also does not perform as well as Metric Labeling, which again may be due to its independence assumption between functions. When the number of terms of considered in the ontology increases (Figures 1(b) and (c)) and the functions become more detailed, and the performance of all algorithms decreases but other algorithms are affected more than the Metric Labeling-based approaches.

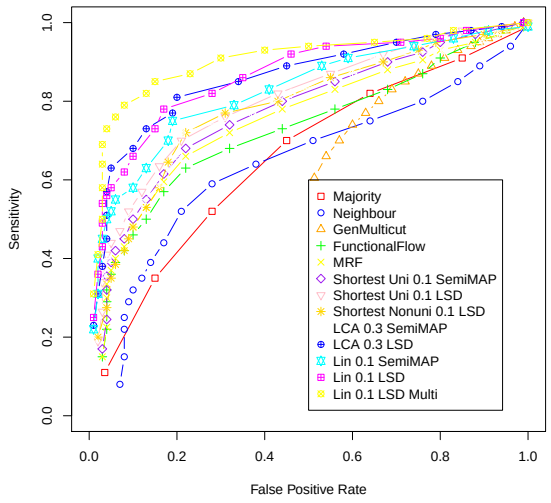
Metric Labeling also outperforms the MRF-based algorithm [13]. This may be because the correlation estimations between functions using in that approach depend solely on training data whereas our distances are estimated from both the training set and the structure of the GO DAG. This indicates that, while the Gene Ontology is an imperfect, incomplete, manually edited resource, the distances between annotations in the ontology do contain useful information that can be exploited to make more accurate predictions.

Among various distance heuristics we used, the LCA and Lin distances are better in general since they take the lowest common ancestor into account. Typically, d_{Lin} performs slightly better than the d_{LCA} distance but they both perform better than the d_{SP} metric. This further indicates that lowest common ancestor is a good distance estimator when there are hierarchical relations among points as shown previously in WordNet [16]. This also echos results in several other papers [7, 39, 42] in terms of showing effectiveness of lowest common ancestor as a measure between ontology terms. In addition, in almost all cases the nonuniform assignment costs performs slightly better than uniform assignment costs, although the effect is not large, and if nonuniform assignment costs are not available, uniform assignment costs can be nearly as effective.

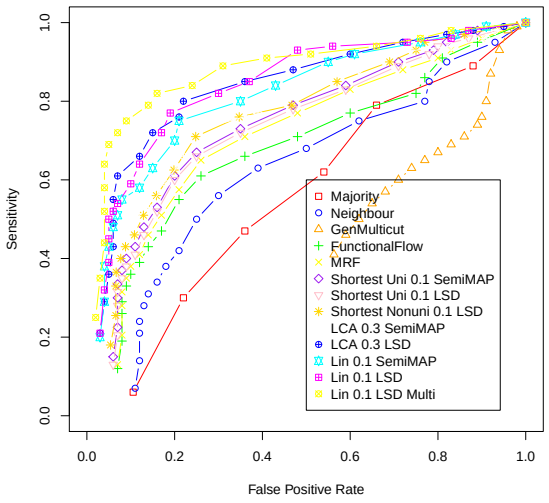
Running the LSD minimization for semimetrics and then running Metric Labeling performs better than Semimetric MAP Estimation algorithm [34] on most of the cases which shows optimizing least square error, rather than the classical distortion, for metric approximation is effective in the protein function prediction application.

3.2 Tradeoff Between GO-distances and Network Distances

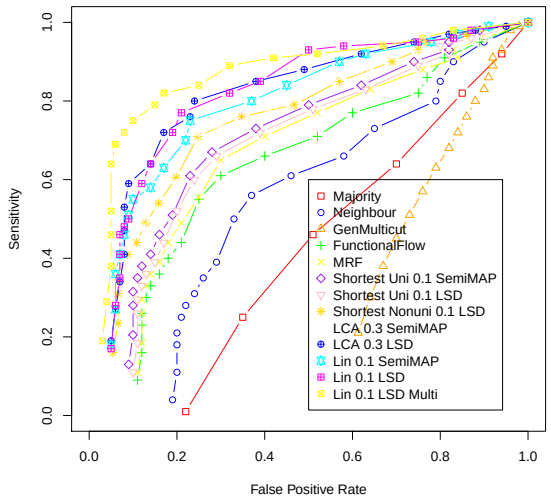
We also investigate how performance varies as the tradeoff between a distance computed from the GO structure ($d_{\mathrm{SP}}, d_{\mathrm{LCA}}, d_{\mathrm{Lin}}$) and a distance computed from proximity in the network (d_{KB}) is varied. Figure 1(d) shows the performance of Metric Labeling with LSD metric approximation



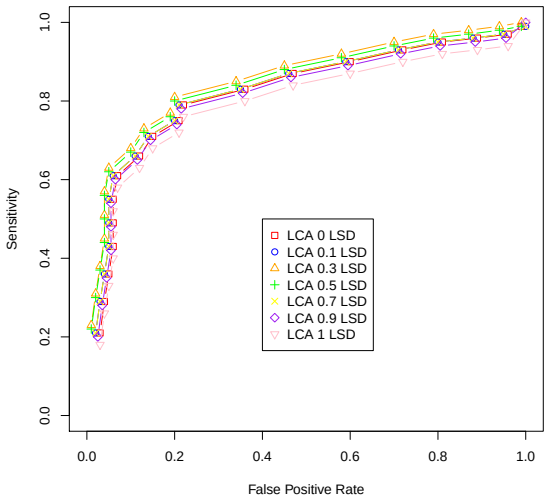
(a) 90 Terms



(b) 150 Terms



(c) 300 Terms



(d) Varied α

Figure 1: Performance of various algorithms on yeast considering (a) 90 (b) 150, or (c) 300 ontology terms. **SemiMap** indicates the semimetric-to-metric conversation algorithm by Kumar and Koller [34] is run; **LSD** means we first run our LSD minimization algorithm and then Metric Labeling. **Uni** indicates the assignment costs of Metric Labeling are uniformly 1 except for known annotations and **Nonuni** means assignment costs are nonuniform as described in Section 2.4. **Shortest**, **LCA**, **Lin** indicate the different distances functions in Section 2.4. The tradeoff α between the GO-based distance and the training distance (Equation (15)) is either 0.1 or 0.3 as indicated. (d) Performance of the d_{LCA} distance combined with the d_{KB} distance with various α using the LSD algorithm.

and the LCA distance for different tradeoffs α between the GO-based structural distance (d_{LCA}) and the trained distances d_{KB} as described in Equation (15). In almost all cases, distances based solely on GO performs better than using only d_{KB} but using estimations both from training set and Gene Ontology structure performs better than using either one alone.

Combining the Gene Ontology knowledge with training set estimations using low values of α ($\alpha = 0.1$ or $\alpha = 0.3$) achieves the best performance by a slight margin. After the initial benefit of using some of the d_{KB} distances, the performance starts to decrease as the weight α is increased. This may mean that d_{KB} is most effective when it operates as a tie-breaker between terms that have the same GO distances. (The dependence on α of the performance of the other GO-based distances d_{SP}, d_{Lin} is similar.) Hence, we show only the results for $\alpha = 0.1$ and $\alpha = 0.3$ in Figures 1(a-c).

3.3 Robustness on the Yeast PPI Network

Metric Labeling combined with LSD metric approximation is more robust to noise in both misannotations and edge removal. We tested for robustness of the predicted results in two ways. First, we removed various percentages of edges randomly from the PPI network and re-run our algorithm. Performance clearly decreased but even when 50% and 40% of the PPI edges are removed on 90 and 150 element ontologies respectively a Metric Labeling approach performs as well as other algorithms run on the true PPI network. The fewer elements the ontology has, the more robust it is in terms of edge removal. The Lin and LCA distance measures again outperform shortest path distance and running LSD minimization for semimetrics and then Metric Labeling does better than using the Semi-Metric MAP Estimation algorithm [34]. We believe LSD minimization algorithm may handle the noise in the data better than other methods during its error minimization.

Secondly, we also tested robustness by misannotating various percentages of protein annotations and then running our algorithm. Performance even when 30% of the proteins are misannotated on both 90 and 150 element ontologies is still comparable with its performance with the true labels, and it is not worse than other algorithms on the true labels. However, in the case of misannotations, combining GO knowledge with slight training set estimations ($\alpha = 0.1$ or $\alpha = 0.3$) does not perform the best anymore. Rather, the GO structure-based distances in isolation perform the best.

3.4 Performance on Other Networks

When we created an integrated networks from multiple sources as described in Section 2.5, the performance increases (last curve in Figure 1). This shows that the Metric Labeling approach is also useful on relational data other than PPI networks. We also tested our algorithm on several species. Among those species, performance strongly depends on how complete PPI network is, with sparser networks generally exhibiting worse performance. Again, the Metric Labeling approach performs competitively with existing methods. Due to space limitations, the complete results for the 7 considered species are available at `http://www.cs.umd.edu/~esefer/metriclabeling`.

4. Conclusions

We show that GO structural information can be exploited to achieve better protein function prediction. We also show that the clean, combinatorial problem of Metric Labeling can exploit these distances and produce accurate predictions in a reasonable amount of computational time.

Our novel LSD metric approximation algorithm combined with Metric Labeling performs better than the semimetric MAP estimation algorithm in most cases. This is interesting since distortion defined as in Section 1 has nearly always been used as the performance measure for metric embeddings. However, as mentioned, distortion doesn't consider the distribution of the error on all points. Its minimization considers just the minimization of the boundary cases (of maximum contraction and expansion). LSD minimization instead tries to minimize the total least square error which makes sense both intuitively and experimentally as we have seen on protein function prediction. Its effectiveness on different application domains is an open question, but the LSD approach is likely to be useful for the common problem of converting a set of heuristic distances into a metric for subsequent processing with an algorithm (such as that for Metric Labeling) that assume a metric.

In addition, the LSD metric approximation is completely independent of Metric Labeling. Either of these algorithms can be changed without affecting the other. However, this is not the case for Semimetric MAP Estimation algorithm, for which the two phases of metric estimation and prediction are not independent and not easy to modify.

References

[1] 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. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556. URL `http://dx.doi.org/10.1038/75556`.

[2] Yair Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. In *In 37th Annual Symposium on Foundations of Computer Science*, pages 184–193, 1996.

[3] Yair Bartal. On approximating arbitrary metrics by tree metrics. In *In Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, pages 161–168, 1998.

[4] Zafer Barutcuoglu, Robert E. Schapire, and Olga G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, January 2006. ISSN 1367-4803. URL `http://view.ncbi.nlm.nih.gov/pubmed/16410319`.

[5] Yuri Boykov, Olga Veksler, and Ramin Zabih. Markov random fields with efficient approximations. In *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–655, 1998.

[6] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 2001. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.396`.

[7] Alexander Budanitsky and Graeme Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *In Workshop On WordNet and Other Lexical Resources, Second Meeting Of The North American Chapter Of The Association For Computational Linguistics*, 2001.

[8] CGAL. Computational Geometry Algorithms Library, 2010. http://www.cgal.org.

[9] C. Chekuri, S. Khanna, J. Naor, and L. Zosin. A linear programming formulation and approximation algorithms for the metric labeling problem. *SIAM J. Discret. Math.*, 18(3):608–625, 2005. ISSN 0895-4801. doi: http://dx.doi.org/10.1137/S0895480101396937.

[10] Chandra Chekuri, Sanjeev Khanna, Joseph (Seffi) Naor, and Leonid Zosin. Approximation algorithms for the metric labeling problem via a new linear programming formulation. In *Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2001)*, pages 109–118, 2001.

[11] Jill Cheng, Melissa Cline, John Martin, David Finkelstein, Tarif Awad, David Kulp, and Michael A Siani-Rose. A knowledge-based clustering algorithm driven by Gene Ontology. *J Biopharm Stat*, 14(3):687–700, 2004. ISSN 1054-3406. URL `http://www.biomedsearch.com/nih/knowledge-based-clustering-algorithm-driven/15468759.html`.

[12] Julia Chuzhoy and Joseph (Seffi) Naor. The hardness of metric labeling. *Foundations of Computer Science, Annual IEEE Symposium on*, 0:108–114, 2004. ISSN 0272-5428. doi: http://doi.ieeecomputersociety.org/10.1109/FOCS.2004.67.

[13] Minghua Deng, Zhidong Tu, Fengzhu Sun, and Ting Chen. Mapping gene ontology to proteins based on protein–protein interaction data. *Bioinformatics*, 20(6):895–902, 2004. ISSN 1367-4803. doi: http://dx.doi.org/10.1093/bioinformatics/btg500.

[14] Dikla Dotan-Cohen, Simon Kasif, and Avraham A. Melkman. Seeing the forest for the trees: using the Gene Ontology to restructure hierarchical clustering. *Bioinformatics*, 25 (14):1789–1795, 2009. doi: 10.1093/bioinformatics/btp327. URL `http://bioinformatics.oxfordjournals.org/content/25/14/1789.abstract`.

[15] Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *In Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, pages 448–455, 2003.

[16] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, illustrated edition edition, May 1998. ISBN 026206197X. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/026206197X`.

[17] Audrey P. Gasch, Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, Patrick O. Brown, and Pamela A. Silver. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11:4241–4257, 2000.

[18] Anne-Claude Gavin, Markus Bosche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jorg Schultz, Jens M. Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Hofert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck, Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R. Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, January 2002. ISSN 0028-0836. doi: 10.1038/415141a. URL `http://dx.doi.org/10.1038/415141a`.

[19] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of drosophila melanogaster. *Science*, 302(5651):1727–1736, December 2003. ISSN 1095-9203. doi: 10.1126/science.1090289. URL `http://dx.doi.org/10.1126/science.1090289`.

[20] GLPK. GNU Linear Programming Kit, 2010. http://www.gnu.org/software/glpk/.

[21] J. A. Hanley and B. J. Mcneil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, April 1982. ISSN 0033-8419. URL `http://radiology.rsnajnls.org/content/143/1/29.abstract`.

[22] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, 18(6):523–531, April 2001. ISSN 0749-503X. doi: http://dx.doi.org/10.1002/yea.706. URL `http://dx.doi.org/10.1002/yea.706`.

[23] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D. Bader, Lynda Moore, Sally-Lin L. Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, Lingyun Yang, Cheryl Wolting, Ian Donaldson, Søren Schandorff, Juanita Shewnarane, Mai Vo, Joanne Taggart, Marilyn Goudreault, Brenda Muskat, Cris Alfarano, Danielle Dewar, Zhen Lin, Katerina Michalickova, Andrew R. Willems, Holly Sassi, Peter A. Nielsen, Karina J. Rasmussen, Jens R. Andersen, Lene E. Johansen, Lykke H. Hansen, Hans Jespersen, Alexandre Podtelejnikov, Eva Nielsen, Janne Crawford, Vibeke Poulsen, Birgitte D. Sørensen, Jesper Matthiesen, Ronald C. Hendrickson, Frank Gleeson, Tony Pawson, Michael F. Moran, Daniel Durocher, Matthias Mann, Christopher W. Hogue, Daniel Figeys, and Mike Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, January 2002. ISSN 0028-0836. doi: 10.1038/415180a. URL `http://dx.doi.org/10.1038/415180a`.

[24] Won-Ki Huh, James V. Falvo, Luke C. Gerke, Adam S. Carroll, Russell W. Howson, Jonathan S. Weissman, and Erin K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691, October 2003. ISSN 0028-0836. doi: 10.1038/nature02026. URL `http://dx.doi.org/10.1038/nature02026`.

[25] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–4574, April 2001. ISSN 0027-8424. doi: 10.1073/pnas.061034498. URL `http://dx.doi.org/10.1073/pnas.061034498`.

[26] L. J. Jensen, R. Gupta, H.-H. Strfeldt, and S. Brunak. Prediction of human protein function according to Gene Ontology categories. *Bioinformatics*, 19(5):635–642, 2003. doi: 10.1093/bioinformatics/btg036. URL `http://bioinformatics.oxfordjournals.org/content/19/5/635.abstract`.

[27] Ulas Karaoz, T. M. Murali, Stan Letovsky, Yu Zheng, Chunming Ding, Charles R. Cantor, and Simon Kasif. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2888–2893, 2004. doi: 10.1073/pnas.0307326101. URL `http://www.pnas.org/content/101/9/2888.abstract`.

[28] N. Karmarkar. A new polynomial-time algorithm for linear programming. In *STOC '84: Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, New York, NY, USA, 1984. ACM. ISBN 0-89791-133-4. doi: http://doi.acm.org/10.1145/800057.808695.

[29] Jon Kleinberg and Eva Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. In *In IEEE Symposium on Foundations of Computer Science*, pages 14–23, 1999.

[30] Nikos Komodakis and Georgios Tziritas. Approximate labeling via graph-cuts based on linear programming. In *In Pattern Analysis and Machine Intelligence*, page 2007, 2007.

[31] Yiannis A. Kourmpetis, Aalt D. van Dijk, Marco C. Bink, Roeland C. van Ham, and Cajo J. Ter Braak. Bayesian markov random field analysis for protein function prediction based on network data. *PloS one*, 5(2):e9293+, February 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009293. URL http://dx.doi.org/10.1371/journal.pone.0009293.

[32] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan. Polynomial solvability of convex quadratic programming. *Doklady Akademiia Nauk SSSR*, 248, 1979.

[33] Minghua Deng Kui, Kui Zhang, Shipra Mehta, Ting Chen, and Fengzhu Sun. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, 10: 947–960, 2002.

[34] M. Pawan Kumar and Daphne Koller. Map estimation of semi-metric mrfs via hierarchical graph cuts. In *UAI '09: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 313–320, Arlington, Virginia, United States, 2009. AUAI Press. ISBN 978-0-9749039-5-8.

[35] H. Lee, Z. Tu, M. Deng, F. Sun, and T. Chen. Diffusion kernel-based logistic regression models for protein function prediction. *OMICS*, 10(1):40–55, 2006. ISSN 1536-2310. doi: 10.1089/omi.2006.10.40. URL http://dx.doi.org/10.1089/omi.2006.10.40.

[36] S. Z. Li. *Markov random field modeling in computer vision*. Springer-Verlag, London, UK, 1995. ISBN 4-431-70145-1.

[37] Siming Li, Christopher M. Armstrong, Nicolas Bertin, Hui Ge, Stuart Milstein, Mike Boxem, Pierre-Olivier Vidalain, Jing-Dong J. Han, Alban Chesneau, Tong Hao, Debra S. Goldberg, Ning Li, Monica Martinez, Jean-Francois Rual, Philippe Lamesch, Lai Xu, Muneesh Tewari, Sharyl L. Wong, Lan V. Zhang, Gabriel F. Berriz, Laurent Jacotot, Philippe Vaglio, Jerome Reboul, Tomoko Hirozane-Kishikawa, Qianru Li, Harrison W. Gabel, Ahmed Elewa, Bridget Baumgartner, Debra J. Rose, Haiyuan Yu, Stephanie Bosak, Reynaldo Sequerra, Andrew Fraser, Susan E. Mango, William M. Saxton, Susan Strome, Sander van den Heuvel, Fabio Piano, Jean Vandenhaute, Claude Sardet, Mark Gerstein, Lynn Doucette-Stamm, Kristin C. Gunsalus, J. Wade Harper, Michael E. Cusick, Frederick P. Roth, David E. Hill, and Marc Vidal. A Map of the Interactome Network of the Metazoan C. elegans. *Science*, 303(5657):540–543, 2004. doi: 10.1126/science.1091403. URL http://www.sciencemag.org/cgi/content/abstract/303/5657/540.

[38] Dekang Lin. An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.

[39] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA, 1998. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/980691.980696. URL http://portal.acm.org/citation.cfm?id=980696.

[40] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1, June 2005. ISSN 1367-4803. URL http://view.ncbi.nlm.nih.gov/pubmed/15961472.

[41] Jean-Christophe Rain, Luc Selig, Hilde De Reuse, Veronique Battaglia, Celine Reverdy, Stephane Simon, Gerlinde Lenzen, Fabien Petel, Jerome Wojcik, Vincent Schachter,

Y. Chemama, Agnes Labigne, and Pierre Legrain. The protein-protein interaction map of helicobacter pylori. *Nature*, 409(6817):211–215, 01 2001. URL http://dx.doi.org/10.1038/35051615.

[42] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, 1999.

[43] Andreas Schlicker, Francisco Domingues, Jorg Rahnenfuhrer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7(1):302, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-302. URL http://www.biomedcentral.com/1471-2105/7/302.

[44] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature biotechnology*, 18(12):1257–1261, December 2000. ISSN 1087-0156. doi: 10.1038/82360. URL http://dx.doi.org/10.1038/82360.

[45] Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Molecular systems biology*, 3, March 2007. ISSN 1744-4292. doi: 10.1038/msb4100129. URL http://dx.doi.org/10.1038/msb4100129.

[46] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1):D535–D539, 2005. doi: 10.1093/nar/gkj109. URL http://nar.oxfordjournals.org/content/34/suppl_1/D535.abstract.

[47] Peter Uetz, Loic Giot, Gerard Cagney, Traci A. Mansfield, Richard S. Judson, James R. Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, Alia Qureshi-Emili, Ying Li, Brian Godwin, Diana Conover, Theodore Kalbfleisch, Govindan Vijayadamodar, Meijia Yang, Mark Johnston, Stanley Fields, and Jonathan M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 02 2000. URL http://dx.doi.org/10.1038/35001009.

[48] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nat Biotech*, 21(6): 697–700, 06 2003. URL http://dx.doi.org/10.1038/nbt825.

[49] Olga Veksler. *Efficient graph-based energy minimization methods in computer vision*. PhD thesis, Cornell University, Ithaca, NY, USA, 1999. Adviser-Zabih, Ramin.