BIOINFORMATICS

Vol. 30 ISMB 2014, pages i60–i68
doi:10.1093/bioinformatics/btu269

Inductive matrix completion for predicting gene–disease associations

Nagarajan Natarajan* and Inderjit S. Dhillon

Department of Computer Science, University of Texas at Austin, Austin, TX 78712, USA

ABSTRACT

Motivation: Most existing methods for predicting causal disease genes rely on specific type of *evidence*, and are therefore limited in terms of applicability. More often than not, the type of evidence available for diseases varies—for example, we may know linked genes, keywords associated with the disease obtained by mining text, or co-occurrence of disease symptoms in patients. Similarly, the type of evidence available for genes varies—for example, specific microarray probes convey information only for certain sets of genes. In this article, we apply a novel matrix-completion method called Inductive Matrix Completion to the problem of predicting gene-disease associations; it combines multiple types of evidence (features) for diseases and genes to learn latent factors that explain the observed gene–disease associations. We construct features from different biological sources such as microarray expression data and disease-related textual data. A crucial advantage of the method is that it is *inductive*; it can be applied to diseases *not* seen at training time, unlike traditional matrix-completion approaches and network-based inference methods that are *transductive*.

Results: Comparison with state-of-the-art methods on diseases from the Online Mendelian Inheritance in Man (OMIM) database shows that the proposed approach is substantially better—it has close to one-in-four chance of recovering a true association in the top 100 predictions, compared to the recently proposed CATAPULT method (second best) that has <15% chance. We demonstrate that the inductive method is particularly effective for a query disease with *no* previously known gene associations, and for predicting novel genes, i.e. genes that are previously not linked to diseases. Thus the method is capable of predicting novel genes even for well-characterized diseases. We also validate the novelty of predictions by evaluating the method on recently reported OMIM associations and on associations recently reported in the literature.

Availability: Source code and datasets can be downloaded from http://bigdata.ices.utexas.edu/project/gene-disease.

Contact: naga86@cs.utexas.edu

1 INTRODUCTION

In silico prioritization of disease genes is an important step towards discovering causal genes and understanding genetic disorders. Many disease–gene prioritization tools have been developed in the last decade, some generic and some disease-class specific. Due to the inherent difficulty and latency in human gene–disease studies, very few reliable associations are reported to public databases such as the Online Mendelian Inheritance in Man (OMIM) and the Genetic Association Database (Becker *et al.*, 2004). Therefore, exploiting multiple auxiliary sources of data is essential for predicting genes related to polygenic traits, and many existing methods have been developed for this purpose. For example, a popular family of network-based methods include CIPHER (Wu *et al.*, 2008), GeneWalker (Köhler *et al.*, 2008), Prince (Vanunu *et al.*, 2010), RWRH (Li and Patra, 2010) and CATAPULT (Singh-Blom *et al.*, 2013). These methods exploit biological networks such as the functional gene interactions network and disease similarity network; they infer gene–disease connections by using random walk procedures on different biological networks or computing a similarity measure between nodes.

The problem of predicting gene–disease associations can be thought of as analogous to designing a *recommender system* where the goal is to predict the 'preference' that a user (gene) would give to an item (disease). An important formulation used in recommender systems such as the Netflix movie recommendations (Bennett and Lanning, 2007) is matrix completion, where the problem is to 'complete' the user-item preference matrix given a sample of observed preferences. The standard matrix completion techniques for recovering the user-item preference matrix assume that the true underlying matrix is low-rank. To the best of our knowledge, there is no existing successful application of the matrix completion approach to recovering the gene–disease associations matrix. Two reasons are the extreme sparsity of the associations matrix and the lack of 'negative' associations. Also, all matrix completion approaches suffer from the *cold-start* problem, that of making predictions for a new user (see Section 2). Our approach in this article is based on matrix completion and is best motivated by the limitations of the existing methods discussed next.

Most of the aforementioned methods typically rely on a seed or candidate set of genes already linked to the query disease and therefore fail to make predictions for a new disease of interest, for which there are no gene linkage studies yet; a few make reasonable predictions *if* we could compute some similarity measure with existing diseases on which the methods were trained. However, more often than not, the type of evidence available for diseases of interest varies—for example, we may know already linked genes, keywords associated with the disease obtained by mining text, or co-occurrence of disease symptoms in patients. Methods relying on a specific type of evidence (such as disease similarities) cannot be applied to a query disease with a different type of evidence (say keywords associated with the disease). The same is true for the type of evidence available for genes. Network-based methods cannot predict a gene that is not connected to any other node in the network. On the other hand, methods that exploit gene-expression profiles, functional annotations and signaling pathways exist but have primarily

*To whom correspondence should be addressed.

been developed for specific disease-classes, and therefore fall short in generalizing to new diseases.

It is imperative that complementary types of evidence be merged to provide better coverage and generalization than any single data source. The survey article by Piro and Di Cunto (2012) discusses the following different types of evidence used by prioritization tools: text-mining of biomedical literature, functional annotations, pathways and ontologies, phenotype relationships, intrinsic gene properties, sequence data, protein–protein interactions, regulatory information, orthologous relationships and gene expression information. In this article, we propose a framework that can seamlessly integrate features from the aforementioned data sources. Our approach involves two steps. First, we derive features for diseases and genes from multiple sources. Next, we incorporate the features while trying to learn gene–disease associations in a novel inductive matrix completion (IMC) approach [recently developed and theoretically analyzed by Jain and Dhillon (2013)]. The entries of the associations matrix are assumed to be generated by applying the corresponding gene and disease feature vectors on an unknown low-rank matrix Z. The parameter matrix Z is learnt using a training set of OMIM gene–disease associations, and predictions for a disease are obtained as a function of the features of all genes and the feature vector for the disease. We evaluate our proposed approach through comprehensive experiments and demonstrate substantial increase in the quality of predictions compared to state-of-the-art methods. Our findings and contributions are summarized below.

(1) Integrating diverse feature sets of genes and diseases obtained through a wealth of publicly available data overcomes extreme sparsity in the gene–disease associations data.

(2) Our approach is a novel application of the inductive matrix completion method; it can be applied to diseases not seen at training time, unlike traditional matrix completion approaches and other network-based inference methods that are transductive.

(3) The approach is particularly effective for a query disease with no previously known gene associations, and for predicting novel genes, i.e. genes that are previously not linked to diseases, thus capable of making novel predictions even for well-characterized diseases.

(4) Comparison with state-of-the-art methods on OMIM diseases shows the superiority of the inductive method. We also validate the novelty of predictions by evaluating the method on recently discovered gene associations recorded in the OMIM database, as well as on associations recently reported in the literature curated by Börnigen *et al.* (2012).

We begin by discussing the limitations of traditional matrix completion techniques, motivating our approach and describing the inductive method in Section 2. In Section 3, we describe our experimental datasets and construction of gene and disease features. Extensive quantitative analysis of the new approach is presented in Section 4, and conclusions are presented in Section 5.

1.1 Related work

In the past two decades, a number of tools have been developed for prioritizing disease genes, leveraging the advances in statistical and machine learning techniques. We refer the reader to the excellent survey articles by Moreau and Tranchevent (2012) and by Piro and Di Cunto (2012) for a near-comprehensive treatment of different classes of methods, contexts in which they are best-suited and what sources of data they integrate.

Recently, predicting gene–disease links based on network analysis has become popular (Lee *et al.*, 2011; Li and Patra, 2010; Linghu *et al.*, 2009; Singh-Blom *et al.*, 2013; Vanunu *et al.*, 2010; Wu *et al.*, 2008). These methods work by determining similarity between candidate gene and disease nodes in heterogeneous networks composed of different biological networks [see Barabási *et al.* (2011) for a detailed review of network-based approaches]. In particular, the recently proposed Catapult framework and Katz on the heterogeneous network (Singh-Blom *et al.*, 2013) integrate different biological networks and phenotypes from multiple other species such as mouse and fly. The main drawbacks of network-based methods are that they are limited to the genes that belong to the network and often are not capable of making predictions for new diseases. In contrast, our proposed approach is inductive, and can integrate multiple diverse sources of data in the form of features, including the different biological networks.

Sequence-based features have been employed for prioritizing disease genes by López-Bigas and Ouzounis (2004) and by Adie *et al.* (2005). However, these methods do not use any prior knowledge about the disease and prioritize genes *a priori*. Miozzi *et al.* (2008) show that high-throughput gene-expression data can predict gene function through the 'guilt by association' principle and exploit it to find new candidate genes for many OMIM diseases. On the other hand, methods like Catapult, ToppGene (Chen *et al.*, 2009) and Xu and Li (2006) use topological network features. Our proposed approach seamlessly integrates different types of features and therefore provides better generalization to new diseases and new types of evidence.

2 INDUCTIVE METHOD

Our goal is to predict potential genes for a given disease of interest. We form the gene–disease associations matrix $P \in \mathbb{R}^{N_g \times N_d}$, where each row corresponds to a gene (total number of genes is N_g), and each column corresponds to a disease (total number of training diseases is N_d), such that $P_{ij} = 1$ if gene i is linked to disease j and 0 if the relationship is *unobserved*. Our approach is based on matrix completion, which is one of the most successful and well-studied techniques for recommender systems. Given a sample of observed entries Ω from a true underlying matrix $M \in \mathbb{R}^{m \times n}$, the goal is to estimate missing entries under additional assumptions on the structure of the matrix. The most common assumption is that the matrix is low-rank, i.e. $M = WH^T$, where $W \in \mathbb{R}^{N_g \times k}$ and $H \in \mathbb{R}^{N_d \times k}$ are of rank $k \ll m, n$. Applying the standard low-rank model on the gene–disease associations matrix $P \approx WH^T$, we could solve the following optimization problem:

$$\min_{W \in \mathbb{R}^{N_g \times k}, H \in \mathbb{R}^{N_d \times k}} \sum_{(i,j) \in \Omega} (P_{ij} - W_i^T H_j)^2 + \frac{1}{2}\lambda(\|W\|_F^2 + \|H\|_F^2), \qquad (1)$$

where λ is a regularization parameter, W_i and H_j denote the latent factor for the i-th gene and the j-th disease, respectively. We want to learn factors $W \in \mathbb{R}^{N_g \times k}$ and $H \in \mathbb{R}^{N_d \times k}$ such that the estimated values are

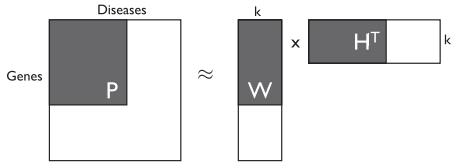

Fig. 1. Low-rank modeling of gene–disease associations matrix. The shaded region in the P matrix corresponds to genes or diseases with at least one known association. Traditional matrix completion would fail to make predictions for genes and diseases with no known associations

close to the observed entries, and the rank of WH^T is small. The gene–disease association matrix P is typically very sparse. For example, in our dataset consisting of diseases from the OMIM database, most columns (diseases) have just one known entry, and many rows (genes) have no known entries. Figure 1 illustrates why using traditional matrix completion Equation (1) on P is not a good idea—it fails to predict on rows and columns with no known entries. Of course, to make meaningful predictions, we would need more information about genes and diseases that have no associations data.

Different data sources provide evidence for genes and diseases: text-mining of biomedical literature, functional annotations, phenotype relationships, protein–protein interactions, regulatory information, orthologous phenotypes in other species and gene-expression information. The question we ask here is if we can directly use the rich set of features for genes and diseases, for the prioritization task. One naïve way is to solve a regression problem associated with each disease independently, where the gene features form the covariates and associations for the disease are the responses. This is called *single-task* learning. The fundamental problem here is that most diseases do not have enough training examples. In contrast, we need a *multi-task* learning approach, as we would expect closely related diseases to have similar predictions. The idea is to learn gene associations for multiple diseases jointly. We formulate a multi-label learning problem, where each gene is an example and each disease is a label or a task, and the goal is to jointly learn associations for all diseases. The recently developed framework (Yu *et al.*, 2014) for multi-label learning formulates the problem as that of learning a low-rank linear model $Z \in \mathbb{R}^{d \times L}$, where each example (gene) is represented by d features and has up to L labels (diseases). If $\mathbf{x} \in \mathbb{R}^d$ denotes the feature vector for a gene, then the corresponding prediction for disease j is given by $\mathbf{x}^T Z_j$, where Z_j is the j-th column of Z. Two key observations given below are in order.

(1) In typical multi-label problems arising in machine learning applications [considered, for example, by Yu *et al.* (2014)], the set of labels is usually fixed and when presented a new example we would want to predict which of the labels are most relevant. In the case of gene–disease associations, as discussed earlier, it would be desirable to make predictions for a new disease—for example, one that was not previously known to be a polygenic disorder. But this is not possible in the standard multi-label formulation because it is transductive—the labels are fixed during the training phase, and predictions on new labels are not possible.

(2) On the other hand, it would be helpful to construct features from other auxiliary sources such as text articles on diseases, studies on patients, symptoms, etc. Relationships (such as co-occurrences) with other existing polygenic traits also make viable biological features. We would want to be able to exploit available information to make informed predictions on diseases.

To this end, we adopt the recently developed IMC method (Jain and Dhillon, 2013) for the task of learning gene–disease associations. The method can be interpreted as a generalization of the transductive multi-label learning formulation. IMC assumes that the associations matrix is generated by applying feature vectors associated with its row as well as column entities to a low-rank matrix Z. The goal is to recover Z using observations from P. The idea is illustrated in Figure 2.

Let $\mathbf{x}_i \in \mathbb{R}^{f_g}$ denote the feature vector for gene i, and $\mathbf{y}_j \in \mathbb{R}^{f_d}$ denote the feature vector for disease j. Let $X \in \mathbb{R}^{N_g \times f_g}$ denote the training feature matrix of N_g genes, where the i-th row is the gene feature vector $\mathbf{x}_i$, and let $Y \in \mathbb{R}^{N_d \times f_d}$ denote the training feature matrix of N_d diseases, where the i-th row is the disease feature vector $\mathbf{y}_i$. The IMC problem is to recover a low-rank matrix $Z \in \mathbb{R}^{f_g \times f_d}$ using the observed entries from the gene–disease association matrix P. Denote the set of observed entries (i.e. training gene–disease associations) by Ω. The entry P_{ij} of the matrix is modeled as $P_{ij} = \mathbf{x}_i^T Z \mathbf{y}_j$ and the goal is to learn Z using the observed entries Ω. Z is of the form $Z = WH^T$, where $W \in \mathbb{R}^{f_g \times k}$ and $H \in \mathbb{R}^{f_d \times k}$, and k is small. The low-rank constraint on Z is NP-hard to solve. The standard relaxation of the rank constraint is the trace norm, i.e. sum of singular values. Minimizing the trace-norm of $Z = WH^T$ is equivalent to minimizing $\frac{1}{2}(\|W\|_F^2 + \|H\|_F^2)$. The factors W and H are obtained as solutions to the following optimization problem:

$$\min_{W \in \mathbb{R}^{f_g \times k}, H \in \mathbb{R}^{f_d \times k}} \sum_{(i,j) \in \Omega} \ell(P_{ij}, \mathbf{x}_i^T W H^T \mathbf{y}_j) + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2). \quad (2)$$

The loss function ℓ penalizes the deviation of estimated entries from the observations. A common choice for loss function is the squared loss function given by $\ell_{sq}(a, b) = (a - b)^2$. The regularization parameter λ trades off accrued losses on observed entries and the trace-norm constraint. Given a new disease j' that was not a part of the training data, the predictions $P_{ij'}$ can be computed for all genes i as long as we have feature vector $\mathbf{y}_{j'}$. Typically, when the number of features is very large, a small value of k implies that the number of parameters to be learnt is *much smaller* than $f_g \times f_d$. Note that in the standard matrix completion, we would learn $(N_g + N_d) \times k$ parameters, but in IMC the number of parameters is independent of the number of genes or diseases, but depends only on the number of gene and disease features.

2.1 Principal components as features

We perform dimensionality reduction on different types of data sources to obtain robust gene and disease features. Most of our data sources are in the form of networks represented by adjacency matrices. One way to obtain real-valued features for nodes is to look at the principal components of the adjacency matrices. In particular, we use the leading eigenvectors of the adjacency matrix as *latent* features. For example, consider the gene-interactions network G of size $N_g \times N_g$. Let $U \in \mathbb{R}^{N_g \times m}$ denote the matrix of eigenvectors corresponding to the top m eigenvalues of G. Now, the i-th row of U gives m latent features for gene i. We perform PCA (principal components analysis) on the microarray expression and word-count data to obtain low-dimensional informative features for genes and diseases, respectively. We discuss data sources and feature extraction in detail in Section 3.

2.2 Optimization

The objective function in Equation (2) is non-convex. We adapt the LEML solver provided by Yu *et al.* (2014) for solving Equation (2) since the traditional multi-label learning problem can be thought of a special case of Equation (2) where the disease feature matrix Y is set to the identity matrix of size N_d. The solver uses alternating minimization (fix W and solve for H and vice versa) to optimize Equation (2). If the loss function ℓ is convex, then the objective function becomes convex when W or H is fixed. The resulting convex problem in one variable (W or H) is solved using the Conjugate Gradient iterative procedure. Common choices of loss function include squared loss, logistic loss $\ell_{\log}(a, b) = \log(1 + e^{-ab})$ and squared-hinge loss $\ell_{\text{sqhinge}}(a, b) = (\max(0, 1 - ab))^2$. Note

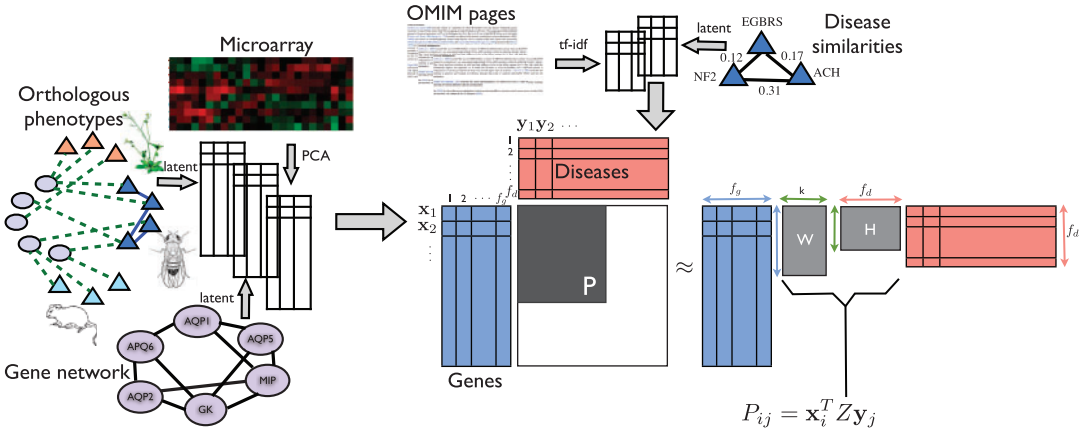

Fig. 2. Schematic of the proposed approach. First, we construct gene and disease features using different sources. Then, we perform IMC using row and column features. The shaded region in the P matrix corresponds to genes or diseases with at least one known association

that the set Ω contains only positive (known gene–disease) associations. However, the number of positive associations typically is very small. In many machine-learning applications, we also have access to negative examples. Unfortunately, we do not have any negative examples (i.e. absence of a gene–disease connection) for our task. Common strategies to cope with the situation are treating all unknowns as negative associations, randomly sampling negative associations from unknowns, or using label-dependent costs (Kshirsagar *et al.*, 2013; Natarajan *et al.*, 2013; Singh-Blom *et al.*, 2013).

2.3 Computational efficiency

Computational cost of solving the optimization problem potentially differs with the choice of the loss function. In our experiments, we use squared loss in the objective, and treat missing values as zeros. For squared loss with fully observed labels, we can essentially use the Algorithm 2 in Yu *et al.* (2014), which yields a fast procedure for solving Equation (2). In particular, the time taken per alternating minimization step is $O((nnz(P) + N_g f_g + N_d f_d)k^2 T)$, where $nnz(P)$ denotes the number of non-zeros in P, and T is a small constant. In our experiments, f_d, f_g and k are very small (few hundreds), and the alternating minimization procedure converges in <10 iterations (takes under 2 min on average on a 2.8 GHz, 8-core machine).

3 DATASET AND FEATURES

3.1 OMIM associations

We obtained human gene–disease associations data from the OMIM project. OMIM phenotypes have become the standard data set for the evaluation of prediction of gene–disease associations (Karni *et al.*, 2009; Köhler *et al.* 2008; Li and Patra, 2010; Mordelet and Vert, 2011; Singh-Blom *et al.*, 2013; Vanunu *et al.*, 2010; Wu *et al.*, 2008). For quantitative evaluation (using 3-fold cross-validation discussed in Section 4.1), we use the OMIM data used in Singh-Blom *et al.* (2013). There are 3209 diseases with at least one known gene association and 3954 gene–disease associations (i.e. the total number of non-zeros in the gene–disease associations matrix). The matrix is extremely sparse, with >90% of the columns with exactly one non-zero entry and ~75% of the rows with no non-zero entries. To compare different gene prioritization methods on the novelty of predictions (see Section 4.7), we use more recently reported associations in the OMIM database (reported between August 2011 and November 2013).

3.2 Gene features

Microarray measurements of gene-expression levels in different tissue samples, obtained from BioGPS (www.biogps.org) and Connectivity Map (www.broadinstitute.org/cmap), serve as the first source of the gene features. In particular, each feature corresponds to a gene-expression level in a sample of a given cell type. Typically, microarray measurements are given for 'probes' (that encode possibly multiple genes). If a probe involves more than a single gene, then the probe is discarded—thus favoring gene-specific probes. If a gene is part of many such probes, then the measurements are averaged across probes. There are some genes for which we do not have any measurements, and hence no features are available for those genes. In total, there are 4536 features for each of the 8755 genes. We observe that the features are highly correlated. This is understandable as samples of same cell type from two different individuals (as in the case of BioGPS features) tend to have similar gene-expression profiles. In our experiments, we project the data to a lower dimensional space. In particular, we use PCA that performs a linear mapping of the data onto the lower dimensional space spanned by the leading 100 eigenvectors of the covariance matrix, maximizing the variance of the data in the new representation.

The second source of gene features is the functional interaction data between genes. HumanNet (Lee *et al.*, 2011) is a large-scale functional gene network which incorporates multiple datasets, including mRNA expression, protein–protein interactions, protein complex data and comparative genomics (but not disease or phenotype data). HumanNet contains 21 different data sources, which are combined into one integrated network using a regularized regression scheme trained on GO pathways. HumanNet has been shown to be very useful for the gene-prioritization task (Singh-Blom *et al.*, 2013). We obtain latent graph features for genes from HumanNet given by the leading 100 eigenvectors of the network.

The third source of gene features arises from gene orthology—gene–phenotype associations of other species that are relatively richer compared to gene–disease studies in humans. We use phenotypes of eight different species (Singh-Blom *et al*., 2013), namely, plant [*Arabidopsis thaliana*, from TAIR Swarbreck *et al*. (2008)], worm [*Caenorhabditis elegans* from WormBase Chen *et al*. (2005) and Green *et al*. (2011)], fruit fly [*Drosophila melanogaster* from FlyBase Tweedie *et al*. (2009)], mouse [*Mus musculus* from MGD Eppig *et al*. (2007)], yeast [*Saccharomyces cerevisiae* from Dwight *et al*. (2002), Saito *et al*. (2004), McGary *et al*. (2007) and Hillenmeyer *et al*. (2008)] *Escherichia coli* [Nichols *et al*. (2011)], zebrafish [*Danio rerio* from ZFIN Sprague *et al*. (2006)] and chicken [*Gallus gallus* from GEISHA Bell *et al*. (2004)]. We form a large gene–phenotype associations matrix (similar to the gene–disease associations matrix), whose columns correspond to phenotypes of the aforementioned organisms. We use the leading 100 singular vectors of the matrix as features for genes.

3.3 Disease features

Analogous to the use of HumanNet for obtaining latent features for genes, we extract 100 latent disease features from the disease similarity network MimMiner (Van Driel *et al*., 2006). The MimMiner network has been previously used for prioritizing disease genes (Li and Patra, 2010; Singh-Blom *et al*., 2013; Vanunu *et al*., 2010). Another source of disease features we incorporate comes from the web pages for the OMIM diseases. In particular, we look at the 'Clinical Features' and 'Clinical Management' sections of the web pages that document the symptoms, medication and responses by patients, and related studies of effects of different courses of therapies. We want a representation for diseases such that two diseases that are biologically close (such as variants of the same disease) are also close in the feature space. To this end, we form the so-called term-document matrix M, where M_{ij} gives the frequency of the term i in the web page corresponding to disease j. The term-document matrix is commonly used in text mining such as topic extraction from a corpus of documents. Often it is better to use a re-weighting scheme called tf–idf (term frequency–inverse document frequency): M_{ij} is offset by the frequency of the word i in the entire collection of documents. This helps filtering out common words. After applying the tf–idf scheme, we trim the feature space by purging the most common (like 'dose', 'day', 'regimen' that are not informative) and very rare words (specific to a disease such as 'vicriviroc' that appears only in the OMIM page for *Susceptibility to Human Immunodeficiency Virus Type I* phenotype). The resulting feature space for diseases is still high dimensional (~20 000 words). We reduce the dimensionality of the feature space using PCA as in the case of microarray gene features, retaining the top 100 principal components.

4 RESULTS AND DISCUSSION

4.1 Evaluation methods

To quantitatively evaluate our approach and to compare to the state-of-the-art disease–gene-prioritization methods, we measure the recovery of genes using a cross-validation strategy similar to the one used by Mordelet and Vert (2011) and by Singh-Blom *et al*. (2013) on OMIM data. We split the known gene–disease pairs into three equally sized groups. We hide the associations in one group and run the methods on the remaining associations, repeating three times to ensure that each group is hidden exactly once. For each disease in our dataset, we order all the genes by how strongly the method predicts them to be associated with the disease. Finally, for every gene–disease pair (g, d) in the hidden group we record the rank of the gene g in the list associated with disease d. We use the cumulative distribution of the ranks as a measure for comparing the performances of different methods, i.e. the probability that the rank (at which hidden gene–disease pair is retrieved) is less than a threshold r. The motivation for using this performance measure is to distinguish methods based on the probability of recovering a true association in the top-r predictions for a given disease. A small value of r is desired by biologists; Here, we report results for $r \leq 100$. Recent methods including ProDiGe (Mordelet and Vert, 2011) and CATAPULT (Singh-Blom *et al*., 2013) have adopted this performance measure for evaluation.

We are also interested in studying the ability of our method to correctly identify associations between diseases and genes that are less well studied. Singh-Blom *et al*. (2013) propose validation on *singleton* genes, i.e. genes with only one known association in the dataset but none in the training, for highlighting methods that discover novel genes. We employ this validation strategy in Section 4.5. In Section 4.6, we study how different methods perform on the task of predicting genes for a new disease, i.e. diseases for which there are no known associations at the training time.

Assessing the novelty of predictions is often challenging. It is common to hand-pick biologically relevant genes from the top few predictions and corroborate with findings in the existing literature, which is then sometimes followed by wet-lab experiments. Here, we choose to use the unbiased evaluation scheme adopted by Börnigen *et al*. (2012) for assessing the novelty of gene-prioritization methods. In particular, we train all the competitive methods using all the available OMIM data collected until August 2011. Then, we evaluate the methods on the associations recently reported in the literature, collected by Börnigen *et al*. (2012), and the associations recorded in the OMIM database between August 2011 and November 2013 (Section 4.7).

4.2 Baselines and competitive methods

A natural baseline for our proposed approach is the standard matrix completion on the gene–disease associations matrix P given in Equation (1), i.e. IMC with the gene feature matrix X and the disease feature matrix Y set to identity matrices of appropriate sizes. We compare to two recently proposed methods that rely on combining the different biological networks we use to derive latent features from: the gene-interactions network, the disease similarity network and the bipartite gene–phenotype associations networks of multiple species. We also consider a stronger baseline method—matrix completion on a heterogeneous network that is composed of the aforementioned biological networks. Finally, we compare to the LEML method (Yu *et al*., 2014), which is the transductive equivalent of IMC, when there are no disease features. Below, we describe the methods in more detail:

1. CATAPULT (Singh-Blom *et al*., 2013): Train a bagging support vector machine classifier over gene–disease pairs;

each gene–disease pair is represented by features corresponding to the number of paths of increasing lengths in the combined network given by:

$$C = \begin{bmatrix} G & \bar{P} \\ \bar{P}^T & \bar{Q} \end{bmatrix}, \quad (3)$$

where $\bar{P}$ includes the gene–disease associations matrix P and phenotypes of other species discussed in Section 3.2, and $\bar{Q}$ is set to the disease similarity matrix Q corresponding to human disease nodes and 0 elsewhere. This method was shown to outperform a number of graph-based inference methods such as PRINCE (Vanunu *et al.*, 2010), RWRH (Li and Patra, 2010) and ProDiGe (Mordelet and Vert, 2011).

2. Katz on the combined network (Singh-Blom *et al.*, 2013): The method computes similarities between a pair of nodes based on how many paths of different lengths connect the pair. The similarity between nodes i and j in the combined network Equation (3) is given by:

$$S^{Katz}(C)_{ij} = \sum_{l=1}^{k} \beta^l \left(C^l\right)_{ij},$$

where C^l_{ij} gives the number of paths of length l connecting nodes i and j in the network C. Typically, a small value of k is used (node similarity is better captured by shorter paths). Letting $k = 3$, we can write the matrix of scores between gene and disease nodes as:

$$S^{Katz}_{ij} = \beta P + \beta^2 (GP + PQ) + \beta^3 \left(\bar{P}\bar{P}^T P + G^2 P + GPQ + PQ^2\right).$$

The parameter β (typically set to a small value like 0.01) dampens the contribution from paths of higher lengths. This method is closely related to RWRH (Li and Patra, 2010).

3. Matrix completion on the combined network: We consider matrix completion on the combined network C instead of the bipartite network P; the low-rank model $C \approx WH^T$ suggests that the factors W and H should explain not only the gene–disease associations, but also the observed gene interactions, other species phenotypes and disease similarities. Letting the size of the matrix C to be $N \times N$, the optimization problem we solve is:

$$\min_{W,H \in \mathbb{R}^{N \times k}} \sum_{(i,j) \in \Omega} (C_{ij} - W_i^T H_j)^2 + \alpha \sum_{(i,j) \notin \Omega} (C_{ij} - W_i^T H_j)^2 + \lambda(\|W\|_F^2 + \|H\|_F^2).$$

A large value of α *biases* the estimation of the unobserved entries in the different biological networks comprising C toward 0 (reflecting the assumption that the true underlying biological networks are sparse). Of course, setting $\alpha = 0$ corresponds to applying the standard matrix completion Equation (1) on C. In our experiments, we set $\alpha = 0.2$ (the best value chosen by cross-validation).

4. LEML (Yu *et al.*, 2014) Implemented identical to IMC, but does not use disease features.

Note that the first three methods do not use gene- or disease-specific features such as microarray or tf–idf, respectively. For all the matrix-completion-based methods (including LEML and IMC), we rank the predictions using the estimated values of the matrix (higher the estimated value P_{ij}, more relevant is the gene i for disease j). For the IMC method (i) we construct the gene and disease features ($f_g = 300$, $f_d = 200$) as described in Section 3, (ii) we use squared loss in the optimization problem Equation (2), which we find to be the best-performing loss function in our experiments, and missing entries are treated as zeros in the objective function, which together with squared loss results in a fast procedure for solving Equation (2) as discussed in Section 2.3. We set $\lambda = 0.2$ in the optimization problem Equation (2) for both IMC and LEML. We use the best value for parameters obtained by cross-validation for all the competitive methods.

4.3 Overall performance

The 3-fold cross-validation results on 3209 OMIM diseases are presented in Figure 3a. The vertical axis in the plots gives the probability that a true gene association is recovered in the top-r predictions for various r values in the horizontal axis. We observe that the proposed method IMC significantly dominates every other competitive method consistently over all r values. Our method has close to 25% chance of retrieving a true gene in the top-100 predictions for a disease, whereas even the second best performing method CATAPULT has only ~15%. The three competitive methods Katz, CATAPULT and matrix completion on the combined network which use the same information, albeit in different ways, perform very similarly within the top-100 predictions. As expected, matrix completion on C performs significantly better than the baseline matrix completion. The importance of using disease features cannot be emphasized more—LEML performs significantly worse. In Figure 3b, we present precision–recall curves for different methods. Precision is the fraction of true positives (genes) recovered in the top-r predictions for a disease. Recall is the ratio of true positives recovered in the top-r predictions to the total number of true positives for the disease in the test set. We observe a consistent ordering of curves with respect to the standard precision and recall measures.

Integration of multiple informative features is important for the success of the method. However, we do see that disease features play a dominant role in the predictive power, as most of the OMIM diseases have a single known gene. We find that (data not shown in plots) features obtained from gene-interactions network are important particularly for genes that are well-connected. Gene-expression data is highly noisy and correlated, and therefore we rely on using a few PCA features. Not all genes and diseases have all sets of features; genes that are not connected in the gene network have some microarray expression based features, and similarly diseases not connected in the disease similarity network have tf–idf features. We would like to emphasize that integrating features from different sources that bring

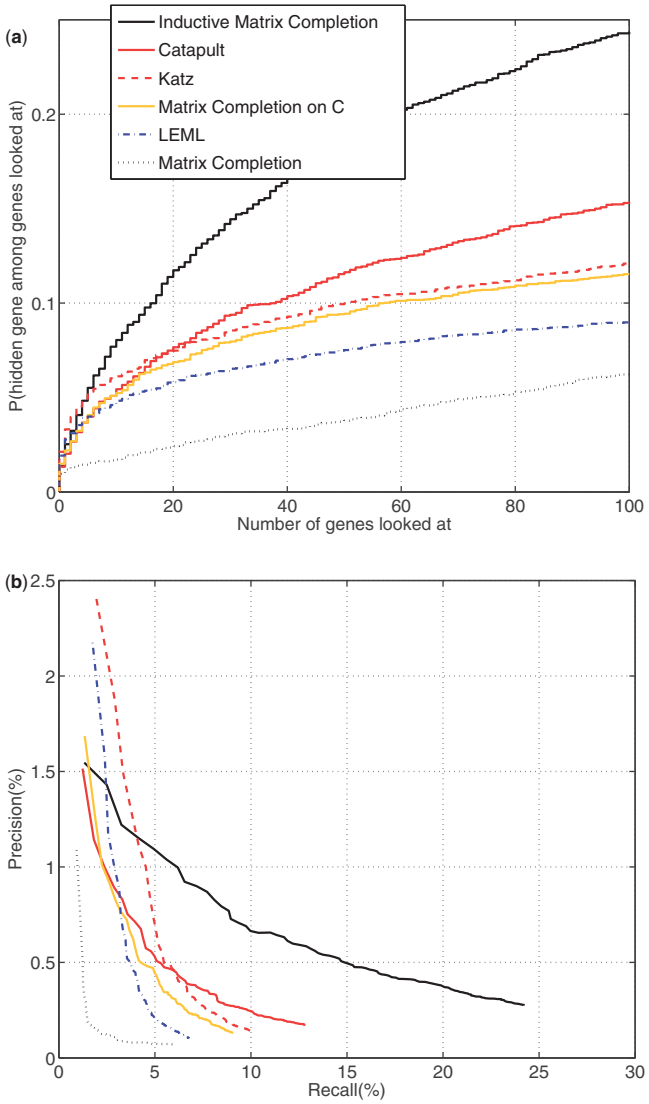

Fig. 3. Comparison of disease gene prioritization methods. The top panel shows the empirical cumulative distribution function for the rank of the withheld gene under cross validation. The vertical axis shows the probability that a true gene association is retrieved in the top-r (shown on the horizontal axis) predictions for a disease. The proposed method IMC is trained with 300 gene features, 200 disease features and $k = 200$. Katz, CATAPULT and Matrix completion on C all use the same combined network Equation (3). IMC (solid black) consistently and significantly outperforms competitive methods by a large margin. The significance of using disease features is apparent by comparing to LEML (dash-dotted blue)

in complementary information, helps improve the predictive performance.

4.4 Effect of rank k

The key parameter in IMC is k, the rank of the model matrix $Z \in \mathbb{R}^{300\times200}$. The effect of rank k on the performance of the method is shown in Figure 4. In general, we observe that performance increases with k. More importantly, a small rank $k = 30$ yields a competitive performance—much better than CATAPULT as compared with Figure 3a. This indicates the success

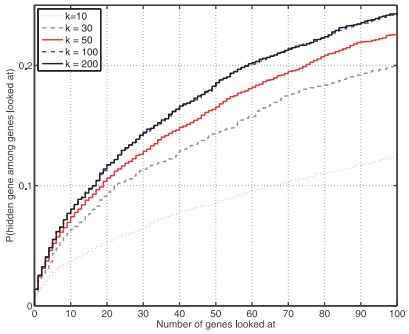

Fig. 4. Performance of IMC ($f_g = 300, f_d = 200$) for different values of k. The performance increases with the rank k of the parameter matrix Z. Even when $k = 30$, IMC starts performing much better than the competitive methods [compare with Figure 3a]. Curves for $k = 100$ and $k = 200$ almost coincide

of multi-task learning approach that exploits correlations between multiple diseases and between multiple genes. At $k = 100$ IMC performs as good as the effective full rank $k = 200$. For comparison with competitive methods, we use $k = 200$ in our experiments.

4.5 Singleton genes

One problem that plagues the evaluation of prioritization methods are the 'popular' genes. Genes that are well-connected tend to be predicted more often and therefore tend to yield inflated recall rates. To this end, we adopt the following strategy used in Singh-Blom *et al.* (2013): We look at *singleton* genes, i.e. genes that have only one association in the data, and evaluate the methods on how highly the corresponding gene associations are ranked. Note that the genes have no known associations at training time, and therefore the ability of a method to predict singleton gene associations also attests to the novelty of predictions. The results are shown in Figure 5a. We see that using additional sources like the biological networks directly are helpful in the case of Katz and matrix completion on C and to lesser extent in case of CATAPULT in the top 1–20. IMC attains a significant increase in the performance around top 50–100 predictions because it uses additional microarray features.

4.6 Induction on new diseases

Next, we turn to an important aspect of evaluation of the proposed method—ability to make predictions for a new disease. By new disease, we mean a disease for which there are no existing gene associations. However, the features for the new disease may be available. For evaluation purposes, we look at the CDF of ranks of genes associated to diseases for which no gene associations were available at the training time (corresponding to columns with no known entries). Figure 5b shows how different methods perform on the task. The significance of using disease features cannot be emphasized more—IMC is substantially better than all other methods. Note that the baseline matrix completion is missing from the plot, as it cannot make predictions for diseases with no known entries.

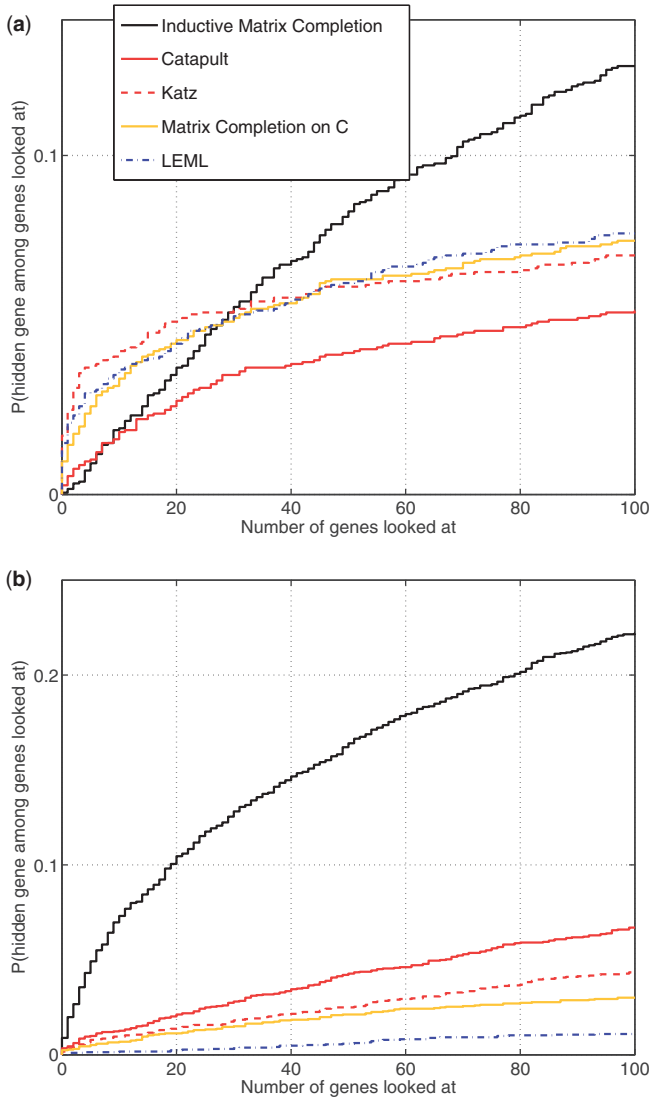

Fig. 5. Evaluating novelty with singleton genes and induction on new diseases. The methods are evaluated in two extreme settings: rows with no known entries (top panel) and columns with no known entries (bottom panel). (A) While using additional sources like the biological networks directly are helpful in the case of Katz and Matrix completion on C, we see a significant increase in the performance around top 50–100 predictions by also using microarray features (solid black). (B) The significance of using disease features is distinct. Note that the baseline matrix completion is missing from both the panels as it cannot make predictions for rows or columns with no known entries (see Section 2)

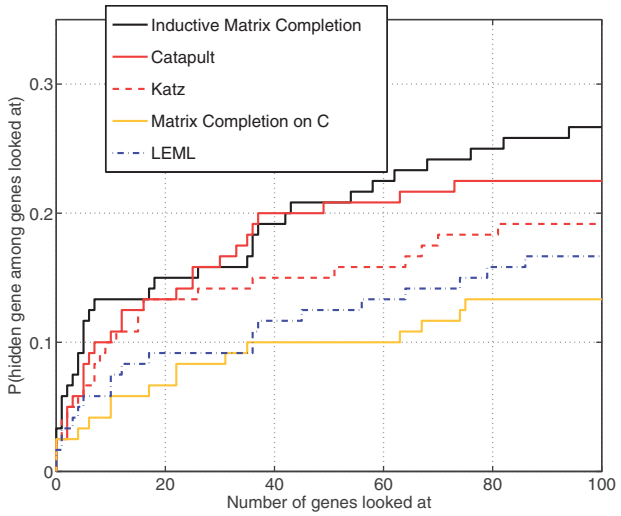

Fig. 6. Evaluating methods on 120 newly discovered associations. The plot shows empirical cumulative distribution function of the ranks of recently discovered associations for a few diseases (see Section 4.7). We see that IMC outperforms all the methods for most values of r, and CATAPULT is competitive. Note that the training data consists of many more associations than that of Figure 3, and partly explains the noticeable increase in performance of all methods

4.7 Evaluation on newly discovered associations

Börnigen *et al.* (2012) evaluate gene prioritization tools using 42 then recently reported associations. The evaluation is unbiased in the following sense: cross-validation on retrospective data is likely to yield overoptimistic performance estimate, as some of the data sources are contaminated with knowledge from gene–disease associations. It is important to understand the subtlety here—the contamination can be by more indirect ways. For example, certain gene interactions may have been discovered precisely because of the associations with the particular disease under evaluation. Though the associations themselves are hidden, the other features are 'contaminated' with this information. The approach therefore mimics novel discovery more closely.

We train all the methods using all gene associations for 3209 OMIM diseases collected until August 2011. For evaluation, we use 36 of the 42 associations curated by Börnigen *et al.* (2012) (we use the expanded set of OMIM phenotypes in our experiments; six of the 42 associations correspond to collapsed phenotypes such as 'Complex heart defect' which can potentially be associated with many phenotypes in our dataset, and therefore we choose to exclude them from evaluation). Of the 36 associations, six associations correspond to six new diseases that are not a part of our training data. It has been a year since the paper was published, so we supplement the data by including 84 recent associations added to the OMIM database between August 2011 and November 2013. The total 120 associations involve 115 unique genes, of which 56 genes did not have any known associations before. Thus, evaluation on the new associations also helps characterize the ability of the methods to recommend novel genes. The ranking performances of competitive methods on the 120 new associations are shown in Figure 6. We see that IMC outperforms all the methods for most values of r and CATAPULT is competitive. We observe a noticeable increase in performance of all methods compared to the cross-validation results in Figure 3. Note that the training data for this experiment consists of many more associations than previous experiments, and in particular all the diseases [except six new diseases from Börnigen *et al.* (2012)] have at least one known association.

Gene MUTYH for Gastric Cancer: As a concrete example, consider the case of the OMIM disease *Gastric Cancer*. We

observe that our method recovers the hidden gene MUTYH in the top-100 predictions but no other method does. Notably, MUTYH is also associated with the OMIM phenotype *Familial Adenomatous Polyposis 2*. The tf–idf features for the two diseases have a high similarity (normalized inner product) which helps our method recover MUTYH for gastric cancer.

5 CONCLUSIONS

In this article, we have proposed a novel approach based on inductive matrix completion for prioritizing disease genes. Our approach combines complementary types of evidence which is essential for generalizing to new diseases, as no single source of data can potentially capture all relevant relations. Comprehensive quantitative analysis of the proposed approach substantiates the claim, as we observe that approaches that rely on particular sources or features (such as biological networks in case of CATAPULT or only gene features as in the case of LEML) perform significantly worse in many cases. The inductive method is not restricted to the types of features used in our experiments—new sources of information can be incorporated easily. Typically, prioritization strategies for finding a novel gene related to an already well-characterized disease would differ from those for which limited or no prior knowledge is available. In our experiments, we find that our approach consistently performs the best on almost all types of diseases and genes, well-characterized or new, which makes our approach a suitable prioritization tool to use for biologists.

Funding: This research was supported by DOD Army grant W911NF-10-1-0529 to ID.

Conflict of Interest: none declared.

REFERENCES

Adie,E.A. *et al.* (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinform.*, **6**, 55.

Barabási,A.-L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Becker,K.G. *et al.* (2004) The Genetic Association Database. *Nat Genet*, **36**, 431–432.

Bell,G.W. *et al.* (2004) GEISHA, a whole-mount in situ hybridization gene expression screen in chicken embryos. *Dev. Dynam.*, **229**, 677–687.

Bennett,J. and Lanning,S. (2007) The netflix prize. In: *Proceedings of KDD Cup and Workshop*. Vol. 2007, p. 35.

Börnigen,D. *et al.* (2012) An unbiased evaluation of gene prioritization tools. *Bioinformatics*, **28**, 3081–3088.

Chen,J. *et al.* (2009) Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37** (**Suppl 2**), W305–W311.

Chen,N. *et al.* (2005) WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res.*, **33**, D383–D389.

Dwight,S.S. *et al.* (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.

Eppig,J.T. *et al.* (2007) The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res.*, **35**, D630–D637.

Green,R.A. *et al.* (2011) A high-resolution *C. elegans* essential gene network based on phenotypic profiling of a complex tissue. *Cell*, **145**, 470–482.

Hillenmeyer,M.E. *et al.* (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (New York, N.Y.)*, **320**, 362–365.

Jain,P. and Dhillon,I.S. (2013) Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*.

Karni,S. *et al.* (2009) A network-based method for predicting disease-causing genes. *J. Comput. Biol.*, **16**, 181–189.

Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

Kshirsagar,M. *et al.* (2013) Multitask learning for host–pathogen protein interactions. *Bioinformatics*, **29**, i217–i226.

Lee,I. *et al.* (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.

Li,Y. and Patra,J.C. (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics/Comput. Appl. Biosci.*, **26**, 1219–1224.

Linghu,B. *et al.* (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol.*, **10**, R91.

López-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, **32**, 3108–3114.

McGary,K.L. *et al.* (2007) Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. *Genome Biol.*, **8**, R258.

Miozzi,L. *et al.* (2008) Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data. *PLoS One*, **3**, e2439.

Mordelet,F. and Vert,J.-P. (2011) Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinform.*, **12**.

Moreau,Y. and Tranchevent,L.-C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Gen.*, **13**, 523–536.

Natarajan,N. *et al.* (2013) Learning with noisy labels. *Adv. Neural Inf. Process. Syst.*, 1196–1204.

Nichols,R.J. *et al.* (2011) Phenotypic landscape of a bacterial cell. *Cell*, **144**, 143–156.

OMIM. Online Mendelian Inheritance in Man, OMIM (2011), Aug. URL http://omim.org/.

Piro,R.M. and Di Cunto,F. (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.*, **279**, 678–696.

Saito,T.L. *et al.* (2004) SCMD: *Saccharomyces cerevisiae* morphological database. *Nucleic Acids Res.*, **32**, D319–D322.

Singh-Blom,U.M. *et al.* (2013) Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS One*, **8**, e58977.

Sprague,J. *et al.* (2006) The zebrafish information network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**, D581–D585.

Swarbreck,D. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.

Tweedie,S. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.

Van Driel,M.A. *et al.* (2006) A text-mining analysis of the human phenome. *European J. Hum. Genet.*, **14**, 535–542.

Vanunu,O. *et al.* (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.

Wu,X. *et al.* (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.

Xu,J. and Li,Y. (2006) Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, **22**, 2800–2805.

Yu,H.-F. *et al.* (2014) Large-scale multi-label learning with missing labels. In: *Proceedings of the 31st International Conference on Machine Learning (ICML), 2014*.