JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 4, 2015

Pp. 313–323
DOI: 10.1089/cmb.2015.0001

Disease Gene Prioritization Using Network and Feature

BINGQING XIE,[1] GADY AGAM,[1] SANDHYA BALASUBRAMANIAN,[2] JINBO XU,[3]
T. CONRAD GILLIAM,[2] NATALIA MALTSEV,[2] and DANIELA BÖRNIGEN[2,3]

ABSTRACT

Identifying high-confidence candidate genes that are causative for disease phenotypes, from the large lists of variations produced by high-throughput genomics, can be both time-consuming and costly. The development of novel computational approaches, utilizing existing biological knowledge for the prioritization of such candidate genes, can improve the efficiency and accuracy of the biomedical data analysis. It can also reduce the cost of such studies by avoiding experimental validations of irrelevant candidates. In this study, we address this challenge by proposing a novel gene prioritization approach that ranks promising candidate genes that are likely to be involved in a disease or phenotype under study. This algorithm is based on the modified conditional random field (CRF) model that simultaneously makes use of both gene annotations and gene interactions, while preserving their original representation. We validated our approach on two independent disease benchmark studies by ranking candidate genes using network and feature information. Our results showed both high area under the curve (AUC) value (0.86), and more importantly high partial AUC (pAUC) value (0.1296), and revealed higher accuracy and precision at the top predictions as compared with other well-performed gene prioritization tools, such as Endeavour (AUC-0.82, pAUC-0.083) and PINTA (AUC-0.76, pAUC-0.066). We were able to detect more target genes (9/18/19/27) on top positions (1/5/10/20) compared to Endeavour (3/11/14/23) and PINTA (6/10/13/18). To demonstrate its usability, we applied our method to a case study for the prediction of molecular mechanisms contributing to intellectual disability and autism. Our approach was able to correctly recover genes related to both disorders and provide suggestions for possible additional candidates based on their rankings and functional annotations.

Key words: conditional random field, enrichment analysis, gene prioritization.

1. INTRODUCTION

DISCOVERING THE GENETIC AND PATHOPHYSIOLOGICAL mechanisms underlying heritable disorders (e.g., autism, schizophrenia, diabetes, predisposition to cardiovascular diseases, various forms of epilepsy, and brain malformations) is among the greatest challenges for modern biomedical research. Detection of contributing molecular mechanisms associated with the disorders of interest requires the

[1]Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois.
[2]Department of Human Genetics, University of Chicago, Chicago, Illinois.
[3]Toyota Technological Institute of Chicago, Chicago, Illinois.

identification of relevant genes among thousands of candidate genes predicted by high throughput genomics or traditional linkage analysis. However, the experimental validation of a large number of candidate genes proved to be very time- and resource-consuming. Therefore, efficient computational approaches are needed to address this challenge. Conversely, *in silico* gene prioritization of promising candidate genes requires prior knowledge describing genes, their products, functional and structural properties, and molecular interactions (Börnigen et al., 2011). Recent technological advances in genomics (e.g., new generation sequencing technologies and functional genomics) produce this knowledge at unprecedented tera- and petabyte scales (Schadt et al., 2010). These technologies do not only generate high dimensional annotations for individual genes but also provide information describing gene–gene interactions and networks.

The availability of such massive amounts of information, however, poses additional challenges, which include, inter alia, the need for the integration of heterogeneous data from multiple sources and the extraction of the most critical information from the high dimensional feature space. In this study, we address these challenges by introducing a novel approach for predicting new high-confidence genetic factors contributing to disease phenotypes. Our approach, enrichment-based conditional random field (CRF), prioritizes the candidate genes by utilizing different types of information coming from network and annotations and allows us to fully explore the available information. This prioritization of candidate genes was achieved by rank-ordering a list of candidates with respect to their relevance to an input gene list based on current knowledge.

Multidimensional biological information was acquired from our in-house Lynx knowledge base (Sulakhe et al., 2014), which integrates various classes of information from over 35 public databases and private collections (NCBI databases, EMBL, UniProt, TIGR); molecular pathways (e.g., Reactome, BioCarta, KEGG, NCI pathways); phenotypic databases (OMIM, disease ontology, phenotype ontology databases); and ontologies [Gene Ontology (GO)(Ashburner et al., 2000), BioPAX, phenotype ontology, disease ontology, MI- PSI, etc.]. Here, a novel way to prioritize candidate genes is introduced by using both gene annotations and reliable information that describe gene–gene interactions based on natural fusion of an underlying gene interaction network (Szklarczyk et al., 2011), as well as various classes of biological information (Sulakhe et al., 2014). Network information and annotations were retained in their original form without manually converting them into each other.

We validated our approach with independent benchmark studies, which revealed an AUC value of 0.86 and a 22% error reduction rate compared with previous tools, including Endeavour (Tranchevent et al., 2008) and PINTA (Nitsch et al., 2011). Finally, we applied our method to a case study for the identification of genetic factors contributing to autism and intellectual disability, and predicted novel promising candidate genes for these phenotypes.

2. RELATED WORK

Gene prioritization is the process of assigning similarity or confidence scores to genes, and ranking them based on the probability of their association with the disease of interest. In the past, several bioinformatics tools for gene prioritization were developed, including but not limited to Toppgene (Chen et al., 2009), Endeavour (Tranchevent et al., 2008; Aerts et al., 2006), Suspects (Adie et al., 2006), and PROSPECTR (Adie et al., 2005) [for a detailed review see the Gene Prioritization Portal (Tranchevent et al., 2010)]. These tools are based on the assumption that the genes that share functionality or interact with each other would be involved in similar phenotypes. Tools such as ToppFun (Chen et al., 2009), GATHER (Chang and Nevins, 2006), and Endeavor used high-dimensional features to prioritize candidate gene lists; for example, ToppFun uses a statistical enrichment process and summarizes a representative profile of training genes with 17 categories (e.g., GO, human and mouse phenotype, text mining results), while Endeavour applies a statistical approach to build profiles for attribute-based features and term frequency inversed document frequency (TFIDF), and recently implemented a kernel-based version. Suspects, Toppgene, and Endeavour compare candidate genes with the extracted representative features by fuzzy similarity measurement (Chen et al., 2009), Pearl correlation (Aerts et al., 2006), or kernels (Aerts et al., 2009) to produce similarity scores for candidate genes that can be combined through order statistics or kernel-based methods (Aerts et al., 2009). The other tools, including ToppNet (Chen et al., 2009) and PINTA (Nitsch et al., 2011), rather than relying on high-dimensional features instead utilize networks and prioritize candidate genes by

using an underlying global protein interaction network [e.g., String (Szklarczyk et al., 2011; Jensen et al., 2009) or D2I (Brown and Jurisica, 2005].

3. APPROACH OVERVIEW

In this study, we present a novel gene prioritization approach that ranks promising disease candidate genes based on the modified conditional random field (CRF) model that simultaneously uses both gene annotations and gene interactions, while preserving their original representation. The major flow of our method consists of five main steps (Fig. 1). First, a training set is acquired from the user input (i.e., a set of genes known to be associated with a certain disease or phenotype, which is referred to as user input or training/seed genes). Second, these training genes are annotated using various classes of information (e.g., GO terms) from LYNX knowledge base (Sulakhe et al., 2014) to obtain a union of annotations for the user input genes. Here, the union of annotations per source is the feature set for the individual source. Subsequently, an enrichment analysis is performed on each feature in the feature set with respect to the training genes to extract the most important features and to remove irrelevant annotations from the model. This enrichment analysis attaches importance scores to all features, while a gene interaction network is acquired to extract pairwise interactions for the CRF model in order to build edge factors. To combine annotations and network information, the filtered features with importance scores and the underlying network are formulated as factors in the general CRF model. The joint probability is the factorization of all the factors divided by a normalization parameter. The probability of labels on the candidate genes can be inferred from the model. Finally, the probabilities for all candidate genes are combined from multiple LYNX sources into overall scores. To assess the significance of the overall scores on the genes, a permutation is performed to calculate a *p*-value for each candidate gene in the list.

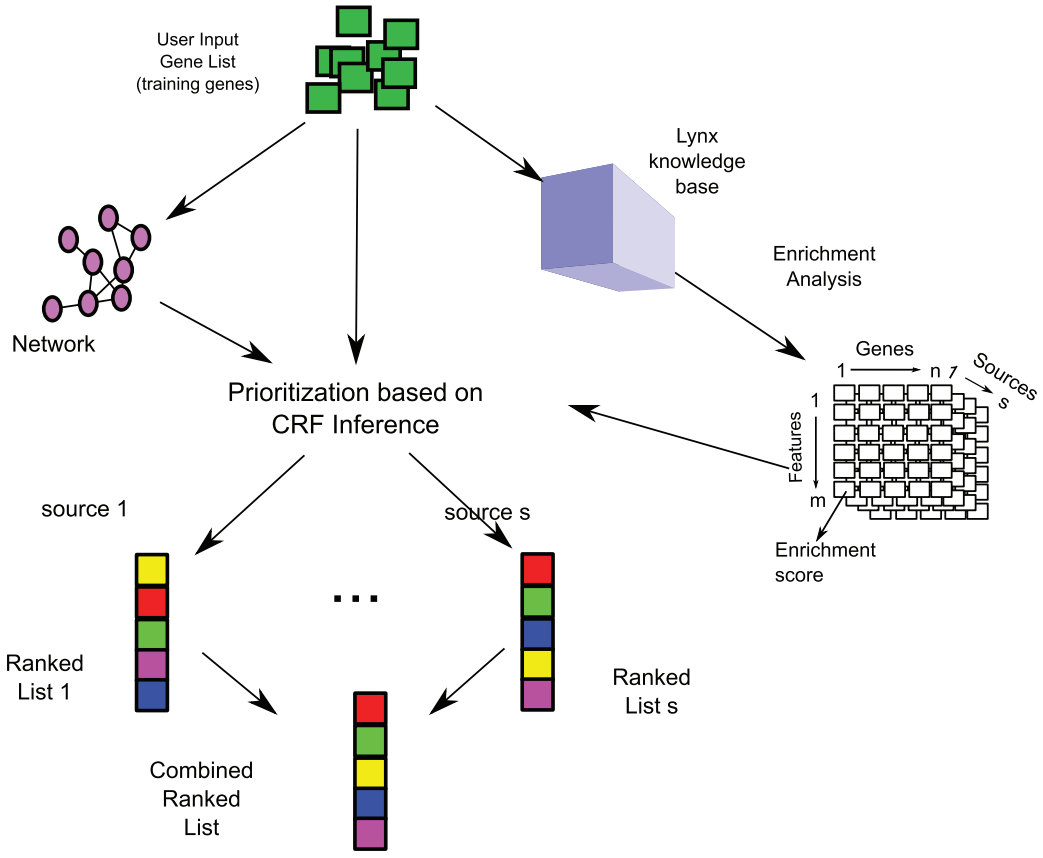

FIG. 1. For a given list of genes, our approach uses both network and multisource feature information from the LYNX knowledge base to build the CRF model and produces the final prediction based on combined scores.

4. ENRICHMENT EMBED CONDITIONAL RANDOM FIELD FOR PRIORITIZATION

In this section, we introduce an enrichment-based CRF method for gene prioritization that ranks a list of candidate genes based upon a list of genes with known disease association, using information including gene–gene interactions and gene annotations. The enrichment analysis is used for feature selection and feature weighting. The CRF is used to integrate gene interaction network information and gene-centric features.

4.1. Conditional random fields

Conditional random field (CRF) is a probabilistic graphical model that can estimate the probability of an unknown label set Y from a given observation set X. In the setting of gene prioritization, X is the set of known disease genes, and Y is the remaining genes for which we would like to predict their disease association. Let $G(V, E)$ denote the gene–gene interaction network where V is the set of genes in the graph and E is the set of edges (i.e., interactions). A factor graph was constructed by integrating gene level annotations over G. Figure 2 shows an example of using CRF modeling a small-scale network with 6 nodes (genes) and 7 edges (interactions). In total there are 13 factors including 6 node factors and 7 edge factors. There are multiple annotations associated with genes, hence the node factors are also associated with those multidimensional features. A global feature is initialized to represent the background probability of random gene associated with any disease. Edge factors share one feature, which represents the scenario that two nodes in the same edge agree or disagree. The nodes with observed labels are colored in red while the labels of the remaining nodes are to be determined. Then the probability of unknown associations between lighter nodes and the disease can be modeled using the CRF. The conditional probability in CRF can be computed

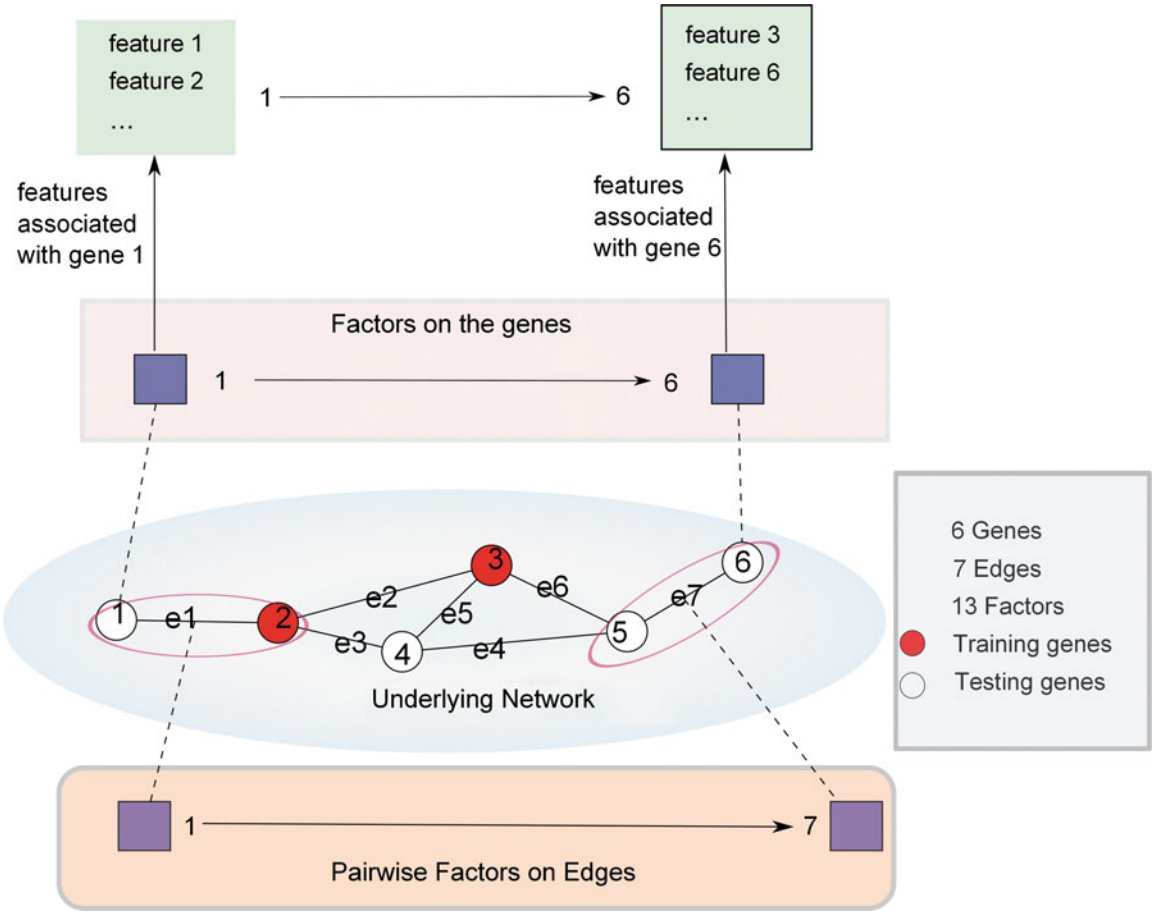

FIG. 2. An example CRF model with 6 nodes (genes) and 7 edges (interactions). There are in total 13 factors including 6 node factors and 7 edge factors. Each node factor is associated with a list of features.

by the summation of all exponential factors on the nodes V and edges E, and the factors can be further parameterized by features on the factors as shown in Equation (1):

$$p(Y|X) = \frac{1}{Z} e^{-\left(\sum_{v \in V,k} \mu_k f_k(Y|_v, X|_v) + \sum_{e \in E} \lambda g(Y|_e, X|_e)\right)} \quad (1)$$

where X is the set of genes known to be associated with a certain disease, Y is the set of genes in the rest of the genome, Z is the normalization parameter, $V|_S$ is the set of nodes in both node set V and subgraph S. Note that in this setting, S is either a single node or a pair of nodes. A node feature function f_k is equal to 1 when gene i is associated with the k^{th} feature and labeled with disease, and 0 otherwise. Edge feature g is equal to 1 when both nodes in the edge have the same label, and 0 otherwise; μ_k is the weight for the k^{th} node feature, and λ is the weight for the edge feature.

After constructing the CRF model, a global inference was performed to determine the marginal probability of each test node. Exact inference is computationally hard for an arbitrary graph structure (Sutton and McCallum, 2010), therefore, approximate inference method, loopy belief propagation (LBP), was used to perform the inference.

4.1.1. Annotation and enrichment. In previous studies (Xie et al., 2013), weights were learned from training examples, which was a time-consuming task and may not converge to the global optimum. To obtain meaningful weights for the factors, features were categorized into two groups: the edge features and the node features. Note that all edge factors shared one feature f_e. A uniform weight w_e was assigned and tuned later. All the node factors on the other hand shared a set of features f_a, $a = 1, \ldots, m$ where m is the total number of annotations. Here, weights of features on the nodes were assigned based on the criterion that they need to reflect the importance of features according to the input training genes. An enrichment analysis was used to characterize sets of target genes with different functional and structural properties (e.g., participation in the same pathway, or biological process, or association with the similar phenotype) (Huang et al., 2009). Here, multiple classes of annotations (e.g., GO terms, associations with particular molecular pathways or phenotypes) from the LYNX knowledge base (Sulakhe et al., 2014) were used for gene annotation and enrichment, while statistical significance scores were assigned to each annotation by estimating differences between the input genes and background genes using the Bayes factor (Chang and Nevins, 2006). Thus, for each factor f_a, a weight w_a was assigned by Bayes factor $B = b_1, \ldots, b_m$ obtained from enrichment analysis.

4.2. Data sources integration

After performing inference on the CRF model, probabilities of disease association on each gene $P_i(Y|X)$ were obtained from data source s_i. Next, the score of one gene associating with a certain label was computed by a weighted sum of all the data sources:

$$\text{Score}(Y|X) = \sum_{i=1}^{m} \theta_i P_i(Y|X) \quad (2)$$

where m is the total number of data sources and θ_i is the weight for the ith source. Currently, the weights were tuned manually by setting to binary values in the experiments and benchmarked in the Results section.

4.3. Statistical validation

Since the probabilities were converted into the overall scores of various data sources, a statistical analysis was performed to estimate the significance of certain gene-label associations. To achieve this, we performed a large-scale randomization of input genes to obtain the score distribution and computed a p-value for each result gene. Standard false discovery rate (FDR)/Bonferroni correction was then applied to adjust p-values.

5. RESULTS

Our method was validated on an unbiased benchmark containing 42 newly discovered disease genes (Börnigen et al., 2011) in two independent settings. First, the backdated validation benchmark was used on a selection of data sources to be able to provide an unbiased and fair performance comparison with other

gene prioritization tools (Börnigen et al., 2011). This analysis revealed a high partial area under the curve (pAUC) values (0.86) for our method. Second, the data sources were extended to a broader collection to find a robust and best-performing combination, again validated on the benchmark but on current data sources, revealing an AUC of 0.89. Finally, our approach was applied to a case study of intellectual disability and autism to predict new causal candidates.

5.1. Validation on a unbiased disease-gene benchmark

First, the performance of our approach was estimated based on an unbiased disease-gene association benchmark (Börnigen et al., 2011). This analysis consists of 42 disease genes that were newly discovered by the time of the benchmark study to mimic new discovery. To enable an unbiased and fair comparison with other tools from this benchmark, the data sources were backdated to the same time as this benchmark study as described below.

5.1.1. Backdated discovery-based cross-validation. To perform an unbiased evaluation and to enable a fair comparison with other existing tools from the published unbiased benchmark study (Börnigen et al., 2011), we needed to backdate our feature data and network data to a time point no later than 2010. Human Gene Ontology (GO) terms (2007) from Uniprot were used as a feature. The String network version 8.2 (Jensen et al., 2009) was used as the underlying functional network. A leave-one-out-cross-validation (LOOCV) was performed and resulted in a receiver of characteristic (ROC) curve (Fig. 3) with an AUC value = 0.86. In this validation, the same candidate set was used as Endeavour and PINTA. For genes without any known annotations or interactions, we initialized all the genes in the

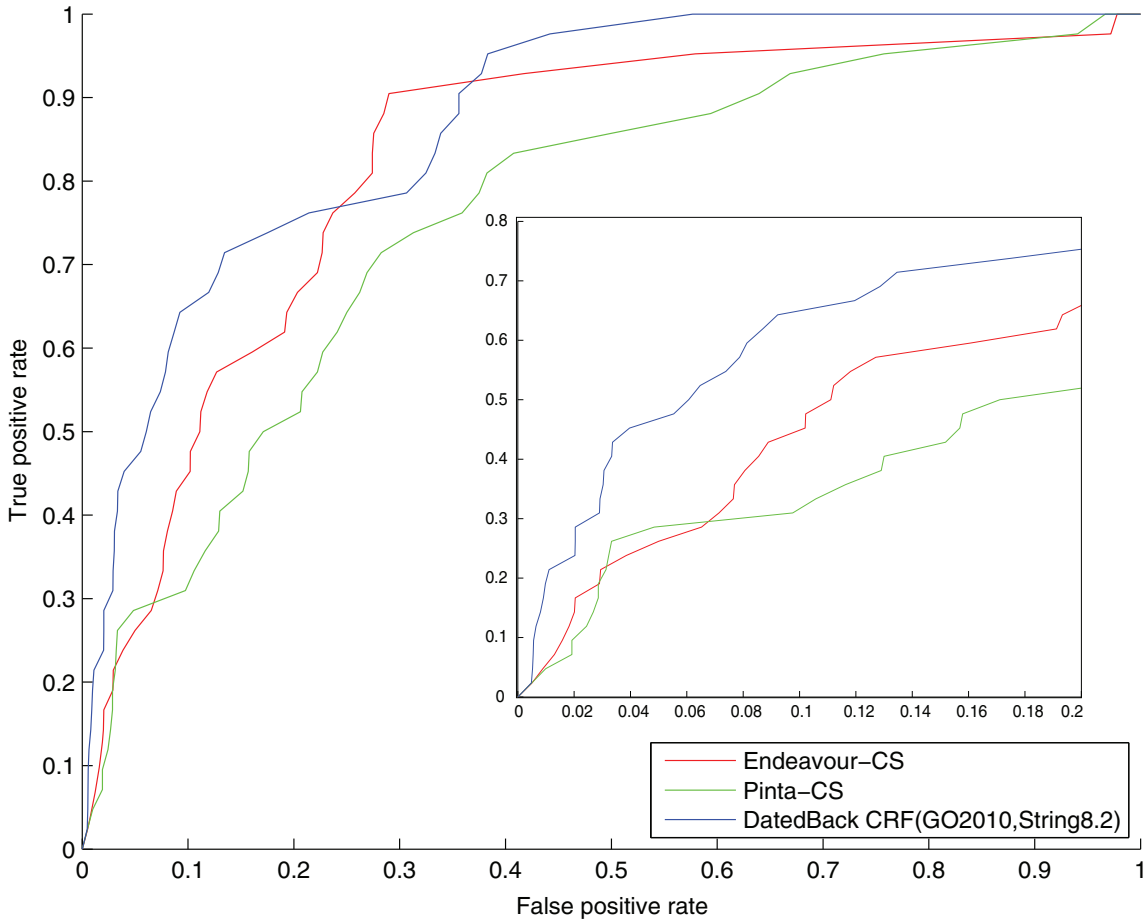

FIG. 3. ROC curves on a disease gene benchmark study using Endeavour, Pinta, and our enrichment-based CRF. The zoomed-in portion is the partial ROC curve within the range of [0, 0.2]

Table 1. Number of True Positives and Partial Area Under Curve Values as Achieved by Enrichment-Based Conditional Random Field Endeavour and PINTA

Method	*TP at top 10%*	*TP at top 15%*	*TP at top 20%*	*pAUC (0–20%)*
E-CRF	27	30	31	0.1296
Endeavour	18	24	27	0.083
PINTA	13	17	21	0.066

pAUC, partial area under curve; TP, true positives.

candidate set with a global feature. This global feature represents the prior probability of a gene associated with diseases, thus was shared by all genes. Since there was no prior distribution of disease genes, the prior probability was set to 0.5.

Next, we compared our results with Endeavour (Tranchevent et al., 2008) and PINTA (Nitsch et al., 2011) from the previous benchmark study (Börnigen et al., 2011). Here, Endeavour used 14 distinct data sources, including annotation (GeneOntology, Swissprot, Interpro, Kegg, EnsemblEst); Interaction (Bind, String); Expression (SonEtAl, SuEtAl); Precalculated (Ouzounis, Prospectr); Motif; Blast; and Text mining, while PINTA only used network information from String version 8.2. The resulting ROC curves of the LOOCV are visualized in Figure 3, showing a higher AUC value for our CRF-based method (0.86) than Endeavour (0.82) and PINTA (0.76) by using only GO term and network information. Compared with Endeavour, our approach reduced the error by 22%, since among the top 20% of gene prioritization results, we observed a much higher pAUC value (0.1296) for our method than for Endeavour (0.083) and PINTA (0.066) (Table 1, Fig. 3), while the achievable maximum pAUC on the top 20% is 0.2; the true positives among the top predictions demonstrate that we outperformed Endeavour and PINTA by identifying more target genes (Table 1).

Absolute ranking positions of target genes. Besides the ranking ratios, the absolute rankings of target genes were also compared among the three methods. As shown in Table 2, more correct target genes were identified as the top ranked gene (9 out of 42) using enrichment-based CRF, while Endeavour ranked 3 and PINTA 6 out of 42 correct target genes as the top gene in the ranking list. Considering the top 5, 10, and 20 genes, more target genes were again identified by our Enrichment-based CRF method than by the other methods. This result demonstrates the high accuracy of our approach for discovering novel target genes.

Correlation and distribution of ranking among the methods. We further determined the Pearson correlation and overall distribution of rankings of target genes among our approach, Endeavour and PINTA (Table 3 and Fig. 4), revealing a low correlation between our method and the other two methods. This implies that our method is complementary to the other two methods, although similar data sources were used. Our method was able to reveal the novel hidden mechanism behind the annotation and network data. The distribution showed in Figure 4 demonstrated that the enrichment-based CRF can generate stable and accurate ranking results. In particular, our method has smaller standard deviation (15 vs. 21/ 24 from

Table 2. Number of Target Genes That Were Ranked at Position x as Determined by Enrichment-Based Conditional Random Field, Endeavour, and PINTA

Methods (data sources)	*No. of target genes ranked in top 1*	*No. of target genes ranked within top 5*	*No. of target genes ranked within top 10*	*No. of target genes ranked within top 20*
Enrichment based-CRF (GO, String)	9	18	19	27
Endeavour (multisources)	3	11	14	23
PINTA (String)	6	10	13	18

CRF, conditional random field.

Table 3. Pearson Pairwise Correlation of Target Gene Rankings Between Enrichment-Based Conditional Random Field and Endeavour/PINTA

Correlation of ranking results for 42 target genes	*Endeavour*	*PINTA*
Enrichment-based CRF	−0.1	0.4

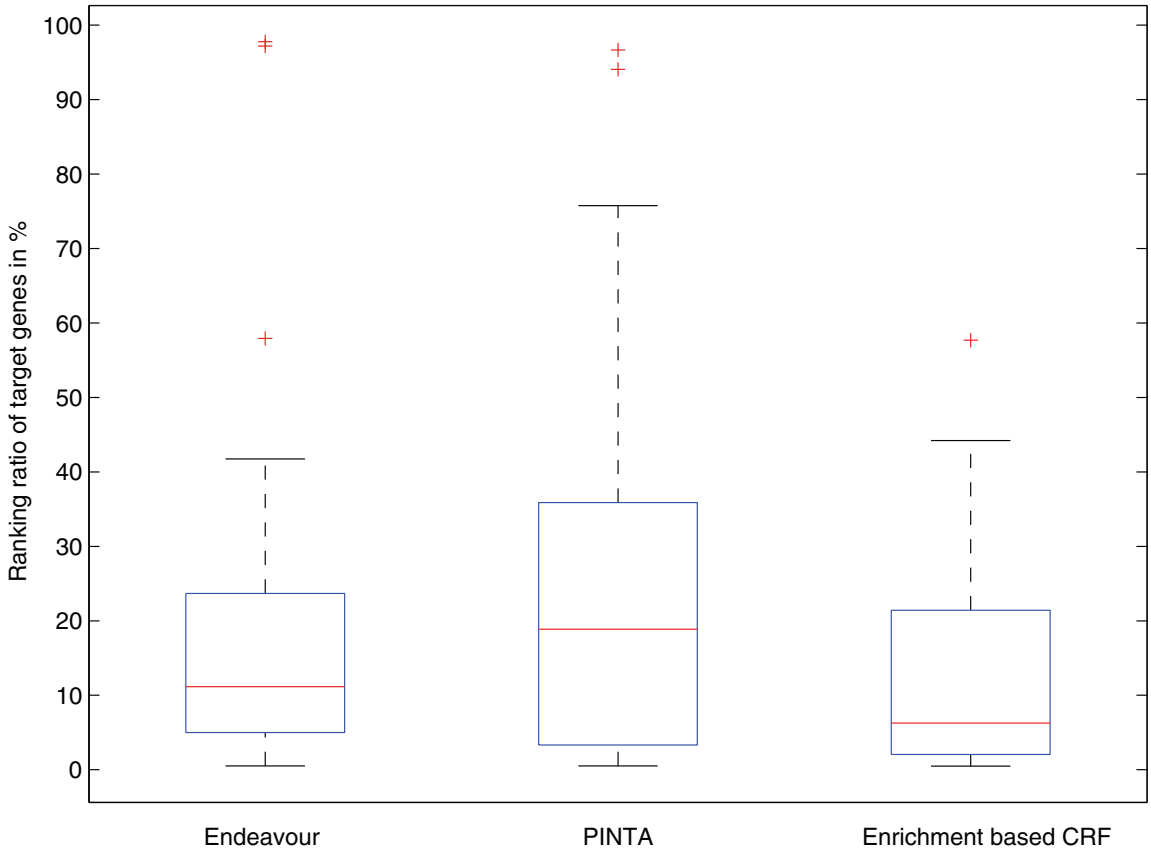

FIG. 4. Distribution of target gene ranking ratio ranked by Endeavour, PINTA, and our enrichment based CRF with medians at 11.0, 18.87, and 6.0, respectively, and standard deviation of 21, 24 and 15, respectively.

Endeavour/PINTA), better median ranking position of target genes (6 vs. 11/18 from Endeavour/PINTA), and a better worst-case scenario (57 vs. 97/96 from Endeavour/PINTA).

5.1.2. Comparison of data sources. In the previous section we compared the performance of our method with Endeavour and PINTA using backdated data source Gene Ontology and network information from STRING. However, there are multiple up-to-date data sources in LYNX database, including the Gene

TABLE 4. AREA UNDER CURVE VALUES AND COVERAGE VALUES OF ROC CURVE FOR LEAVE-ONE-OUT CROSS VALIDATIONS USING ENRICHMENT-BASED CONDITIONAL RANDOM FIELD METHOD AND DATA AVAILABILITY

Table	*AUC*	*Coverage of data source*
DISEASE	0.83	1
GEDI CUSTOM ONTOLOGY	0.78	0.26
GO Biological Process	0.89	1
GO Cellular Components	0.81	0.90
GO Molecular Function	0.89	0.95
GO HIERARCHY Biological Process	0.83	1
GO HIERARCHY Cellular Components	0.77	0.93
GO HIERARCHY Molecular Function	0.84	1
PATH COMMONS	0.85	0.69
PATHWAYS	0.84	0.76
PHENOTYPE	0.84	0.90
TISSUE	0.66	0.55
VISTA ENHANCER	0.71	0.62
VISTA TFBS CLUSTERS	0.75	0.29

AUC, area under curve.

Table 5. Cross-References on Top-Predicted Genes with Known ID and Autism Genes

Top-ranked genes	*No. of genes (out of 190)*	*Total no. of genes in particular category*
Additional ID genes	65	455
Additional known Autism genes (AutDB)	10	219

Ontology source with three main categories (biological process, molecular function, and cellular component) and hierarchical associations between the concepts, Pathway data sources from KEGG, BioCarta, Pathway Commons (PC), Human Phenotype Ontology (HPO), Tissue, Disease database, Vista Enhancer, and Vista TFBS Clusters. To compare these data sources in the LYNX knowledge base, the same benchmark was used. To evaluate the data sources, we introduce the coverage C of a data source:

$$C = \frac{\text{number of target genes annotated}}{\text{total number of target genes}} \tag{3}$$

If the coverage of a data source is close to 1, this source can always produce valid rankings for the target genes, otherwise this data source has missing data on the target genes. Table 4 shows the AUC values for sources separately, as well as the data source coverage, which reflects the proportion of target genes without any annotations in the particular source. Here we can see that Disease, Gene Ontology, Pathway, and Phenotype are the most informative sources with the highest coverage of genes. Since our benchmark covered various types of diseases, we suggest using these sources in the default setting of prioritization if general diseases are studied. However, if a specific brain-related disease is of interest, users should include the Brain Connectivity Ontology source.

5.2. Case study: Intellectual disability (ID) and autism

Genes associated with syndromic and nonsyndromic intellectual disability (ID) were extracted from both OMIM (OMIM, 2012) and literature (Ropers, 2008; van Bokhoven, 2011; Lubs et al., 2012; Inlow and Restifo, 2004). This resulted in a high-quality consolidated list of 475 ID genes that were used in our analysis (see Supplementary Table S1a, available online at www.liebertpub.com/cmb). Genes associated with autism spectrum disorders were obtained from the AutDB database release 3.0 (Basu et al., 2009) (see Supplementary Table S1b).

Forty-eight genes associated both with autism and ID were used as a training gene set for prioritization. The annotation-based feature data sources applied in this prioritization were GO terms, pathways, and phenotypes, while the underlying global network was String 9 (Jensen et al., 2009). Here, we performed a genome-wide ranking. The p-values for the ranked genes were obtained from the distribution generated by random permutations on input genes. The p-value cutoff was set to 0.01 after Bonferroni correction. A total of 190 top-ranked genes with the p-value equal or less than 0.01 were extracted and used for further analysis and annotation.

Table 5 shows the results of the annotation of the top-ranked predicted genes. As it follows from the table, 124 out of 190 of the top-ranked genes were known ID and autism genes, where 94 of them were

Table 6. Cross-References on Top-Predicted Genes with Phenotype Relevant to ID and Autism

Genes associated with phenotype	*No. of genes (out of 160)*	*Total no. of genes in particular category*
Cognitive impairment	80	596
Focal seizures with impairment of consciousness or awareness	1	4
Generalized myoclonic seizures	1	27
Neurological speech impairment	34	162
Delayed speech and language development	30	88
Seizures phenotype	120	605
Additional schizophrenia genes	12	218

predicted, in addition to the 30 genes that were from the original seed set (see Supplementary Table S1 for more details). As it follows, 124 out of 190 of the top-ranked genes were known ID and autism genes, out of which a set of 94 additional genes were correctly predicted to be associated with these phenotypes (see Table 5 for more details).

The prioritized genes were further annotated with the relevant phenotypic information shown in Table 6. The associations with the phenotypes relevant for ID-associated conditions and autism were extracted from the Human Phenotype Ontology (Köhler et al., 2013). A total of 135 out of 160 top-ranked genes were either identified as autism or ID genes from the master list or associated with at least one relevant phenotype with ID/Autism. Meanwhile, 25 genes did not have obvious connections to two conditions. Extensive literature analysis, however, has provided evidence of association of these genes to ID or autism (see Supplementary Table S2). Thirteen of the relevant studies were published after 2011. For example, ATCAY (ranked number 4 out of these 25 genes) may be a promising candidate gene for autism. This gene was previously shown to play a role in the development of neural tissues, particularly the postnatal maturation of the cerebellar cortex (Tan, 2012). It also plays a role in neurotransmission through regulation of glutaminase/GLS, an enzyme responsible for the production in neurons of the glutamate neurotransmitter. Moreover, the ATCAY gene regulates the localization of mitochondria within axons and dendrites.

6. DISCUSSION

In this article we introduced a feature- and network-based gene prioritization method by modeling both types of information on a CRF model. The pairwise relationship on genes is encoded into edge factors, and the annotation-based features are encoded with node factors in the CRF model. The correctness and potential of the method was demonstrated through the validation on a published diseases benchmark by obtaining a higher partial AUC value and better absolute true positives in top-ranked candidate genes than other existing gene prioritization tools. This method is flexible with data sources, allowing for heterogeneous feature types. It outperformed other tools by identifying more true positives on top of the ranked list while using less information. This result shows that using both network and feature information without converting them to each other can enhance the precision on disease genes predictions. The final case study shows a strong association between top-ranked genes and the phenotypes of interest. We were able to identify 25 additional novel genes, potentially related to the phenotypes of interest (ID and autism) and recommend these genes as possible candidates for further testing.

ACKNOWLEDGMENT

The authors are very grateful to Andrew Taylor for his help in preparation of the manuscript.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Adie, E.A., Adams, R.R., Evans, K.L., et al. 2005. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinform.* 6, 55.

Adie, E.A., Adams, R.R., Evans, K.L., et al. 2006. Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics*. 22, 773–774.

Aerts, S., Lambrechts, D., Maity, S., et al. 2006. Gene prioritization through genomic data fusion. *Nat. Biotechnol*. 24, 537–544.

Aerts, S., Vilain, S., Hu, S., et al. 2009. Integrating computational biology and forward genetics in Drosophila. *PLoS Genet*. 5, e1000351.

Ashburner, M., Ball, C.A., Blake, J.A., et al. 2000. Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Basu, S.N., Kollu, R., and Banerjee-Basu, S. 2009. Autdb: a gene reference resource for autism research. *Nucleic Acids Res*. 37, D832–D836.

Börnigen, D., Tranchevent, L.-C., Bonachela-Capdevila, F., et al. 2011. Critical assessment of candidate gene prioritization methods. *Bioinformatics*.

Brown, K.R., and Jurisica, I. 2005. Online predicted human interaction database. *Bioinformatics* 21, 2076–2082.

Chang, J.T., and Nevins, J.R. 2006. Gather: a systems approach to interpreting genomic signatures. *Bioinformatics* 22, 2926–2933.

Chen, J., Aronow, B., and Jegga, A. 2009. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinform*. 10, 73.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 37, 1–13.

Inlow, J.K., and Restifo, L.L. 2004. Molecular and comparative genetics of mental retardation. *Genetics* 166, 835–881.

Jensen, L., Kuhn, M., Stark, M., et al. 2009. String 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*. 37, D412–D416.

Köhler, S., Doelken, S.C., and Mungall, C.J., et al. 2013. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*. 42. D966–D974.

Lubs, H.A., Stevenson, R.E., and Schwartz, C.E. 2012. Fragile X and X-linked intellectual disability: four decades of discovery. *Am. J. Hum. Genet*. 90.4, 579–590.

Nitsch, D., Tranchevent, L.C., Goncalves, J.P., et al. 2011. Pinta: a web server for networkbased gene prioritization from expression data. *Nucleic Acids Res*. 39, W334–W338.

Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Available at: http://omim.org/ Accessed: 2012-10-04.

Ropers, H. 2008. Genetics of intellectual disability. *Curr. Opin. Genet. Dev*. 18, 241–250.

Schadt, E.E., Linderman, M.D., Sorenson, J., et al. 2010. Computational solutions to largescale data management and analysis. *Nat. Rev. Genet*. 11, 647–657.

Sulakhe, D., Balasubramanian, S., Xie, B., et al. 2014. Lynx: a database and knowledge extraction engine for integrative medicine. *Nucleic Acids Res*. 42, D1007–D1012.

Sutton, C., and McCallum, A. 2010. An introduction to conditional random fields. *Found Trends Mach Learn*. 4, 267–373.

Szklarczyk, D., Franceschini, A., Kuhn, M., et al. 2011. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*. 39, D561–D568.

Tan, Ü. 2012. *Latest Findings in Intellectual and Developmental Disabilities Research*. InTech, Rijeka, Croatia.

Tranchevent, C., Capdevila, F.B., Nitsch, D., et al. 2010. A guide to web tools to prioritize candidate genes. *Brief. Bioinform*. 12, 22–32.

Tranchevent, L.C., Barriot, R., Yu, S., et al. 2008. Endeavour update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res*. 36, W377–W384.

van Bokhoven, H. 2011. Genetic and epigenetic networks in intellectual disabilities. *Annu. Rev. Genet*. 45, 81–104.

Xie, B., Agam, G., Maltsev, N., et al. 2013. Conditional random field for candidate gene prioritization. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. ACM, New York.

Address correspondence to:
Ms. Bingqing Xie
Illinois Institute of Technology
Department of Computer Science
3300 South Federal Street
Chicago, IL 60616

E-mail: bxie1@hawk.iit.edu