Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Data and text mining

Supplementary file of 'Isoform function prediction based on bi-random walks on a heterogeneous network'

Guoxian Yu [1], Keyao Wang [1], Carlotta Domeniconi [2], Maozu Guo [3, 4]* and Jun Wang [1]*

[1]College of Computer and Information Science, Southwest University, Chongqing, China.
[2]Department of Computer Science, George Mason University, Fairfax 22030, USA.
[3]School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China.
[4]Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing, China.

*guomaozu@bucea.edu.cn (M. Guo);kingjun@swu.edu.cn(J. Wang)

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

1 Construction of $\mathbf{W}$

To construct the isoform functional association network, we downloaded 588 RNA-seq runs (of total 311 samples) of Human from the ENCODE project (access date: 2017-12-15). These 311 samples are obtained from different tissues and conditions. We process and control the quality of original data as follows:

(1) HISAT2(v.2-2.1.0) (Kim *et al*., 2015; Pertea *et al*., 2016) is firstly used to align the short-reads of each RNA-seq dataset of the Human genome (build GRCh38.90) from Ensemble(Zerbino *et al*., 2017).

(2) A GTF annotation file of the same build is used with an option of no-novel-junction. Then, we use Stringtie(v.1.3.3b) (Pertea *et al*., 2015) to calculate the relative abundance of the transcript as Fragments Per Kilobase of exon per Million fragments mapped fragments (FPKM). We separately compute the FPKM values of a total of 60675 genes with 199184 isoforms for each sample.

(3) The FPKM values of very short isoforms are exceptionally higher. Therefore, these isoforms with less than 100 nucleotides are discarded. For example, the tRNA usually has less than 100 nucleotides. It is important to have sufficient nonzero values in the expression vector during the network building step. Therefore, we only keep isoforms with FPKM value larger than 1 in at least half of all samples.

(4) To further control the quality of isoforms, we used known protein coding gene names to map those genes obtained in step (3). Finally, we obtained a total of 8417 genes with 84519 isoforms.

$\mathbf{W}_{gg}$ encodes the hierarchical dependency between GO terms. To construct the GO term subnetwork, we downloaded the ontology file from the GO website[1]. We then directly used the hierarchical relationship between GO terms stored in this file to construct the subnetwork $\mathbf{W}_{gg}$ as follows:

$$\mathbf{W}_{gg}(s,t) = \begin{cases} 1, \text{ if } t \in child(s) \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

where $child(s)$ includes all the direct child terms of s in the GO hierarchy. $\mathbf{W}_{gg}$ is asymmetric, and it stores the hierarchical relationship between l GO terms.

For the gene-level network, we collected the data from BioGrid[2], which is a publicly accessible database of physical and genetic interactions of genes/proteins with an exhaustive curation (Chatr-Aryamontri *et al*., 2017). If an interaction between genes i and j exists, $\mathbf{W}_{pp}(i,j) > 0$; otherwise $\mathbf{W}_{pp}(i,j) = 0$.

2 Evaluation metrics and Comparing methods

The performance of gene function prediction can be evaluated by different evaluation metrics. To reach a comprehensive comparison, we adopt three evaluation metrics recommended by CAFA (Jiang *et al*., 2016), namely *AUROC*, *Fmax* and *Smin*. Beside, we also use *AUPRC* and *RankLoss* , which is a representative metric in multi-label learning, since isoform function prediction can also be handled as a multi-label learning problem(Zhang and Zhou, 2014).

[1] http://geneontology.org/page/download-ontology

[2] https://thebiogrid.org/download.php

AUROC firstly calculates the area under Receiver Operating Characteristic (ROC) curve of each term and then takes the average value of these areas as a whole to measure the performance. ROC curve plots the true positive rate (sensitivity) as a function of the false positive rate (1-specificity) under different classification thresholds. It measures the overall quality of the ranking induced by the classifier, instead of the quality of a single value of the threshold in that ranking.

AUPRC calculates the average value of the area under the precision-recall curve of each term, and then measure the performance with the average value. The precision of PRC is the percentage of correct associations among the predicted ones, while recall is the same as sensitivity in the ROC.

Fmax is the overall maximum harmonic mean of precision and recall across all possible thresholds on the aggregated gene-term association matrix. The precision (pr), recall (rc) and the resulting *Fmax* are defined as:

$$pr(\eta) = \frac{1}{m(\eta)} \sum_{i=1}^{m(\eta)} \frac{\sum_t \mathbb{1}(t \in p_i(\eta)) \wedge t \in \mathcal{T}_i}{\sum_t \mathbb{1}(t \in p_i(\eta))}, \tag{2}$$

$$rc(\eta) = \frac{1}{n_e} \sum_{i=1}^{n_e} \frac{\sum_t \mathbb{1}(t \in p_i(\eta)) \wedge t \in \mathcal{T}_i}{\sum_t \mathbb{1}(t \in T_i)}, \tag{3}$$

$$F_{max} = \max_{\eta \in [0,1]} \{\frac{2 \cdot pr(\eta) \cdot rc(\eta)}{pr(\eta) + rc(\eta)}\}, \tag{4}$$

where $p_i(\eta)$ denotes the set of terms that have predicted scores greater than or equal to η for protein i, $\mathcal{T}_i$ denotes the corresponding ground-truth set of terms for that protein, $m(\eta)$ is the number of proteins with at least one predicted score greater than or equal to η, $\mathbb{1}(\cdot)$ is an indicator function, and n_e is the number of targets used in a particular mode of evaluation.

Smin uses information theoretic analogs of remaining uncertainty (ru) and misinformation (mi) by referring to GO hierarchy to compute the minimum semantic distance between the predictions and ground-truths across all possible thresholds. The ru, mi and the resulting *Smin* are defined as:

$$ru(\eta) = \frac{1}{n_e} \sum_{i=1}^{n_e} \sum_t ic(t) \cdot \mathbb{1}(t \in p_i(\eta) \wedge t \in \mathcal{T}_i), \tag{5}$$

$$mi(\eta) = \frac{1}{n_e} \sum_{i=1}^{n_e} \sum_t ic(t) \cdot \mathbb{1}(t \in p_i(\eta) \wedge t \notin \mathcal{T}_i), \tag{6}$$

$$S_{min} = \min_{\eta \in [0,1]} \{\sqrt{ru(\eta)^2 + mi(\eta^2)}\}, \tag{7}$$

where $ic(t)$ is the information content of the ontology term t (Clark and Radivojac, 2013).

RankLoss computes the average fraction of wrongly predicted annotations ranking ahead of ground-truth annotations of genes, it is defined as:

$$RankLoss = \frac{1}{N} \sum_{i=l}^{N} \frac{1}{|\mathcal{T}_i||\bar{\mathcal{T}}_i|} |\{(t_1, t_2) \in \mathcal{T}_i \times \bar{\mathcal{T}}_i | R(i, t_1) \leq R(i, t_2)\}| \tag{8}$$

where $\bar{\mathcal{T}}_i$ is the complement set of $\mathcal{T}_i$; $R(i,t)$ is the predicted likelihood for the i-th protein annotated with t.

AUROC and *AUPRC* are GO term-centric metrics, *Fmax* and *Smin*, and *RankLoss* are gene-centric metrics. We want to remark that, a small *Smin* and *RankLoss* means a better performance, and the other four metrics are opposite.

To comparatively and quantitatively study the performance of IsoFun, we take mi-SVM, MI-SVM (Eksi *et al*., 2013), iMILP (Li *et al*., 2014), miFV and miVLAD (Wei *et al*., 2017) as comparing methods. MI-SVM, mi-SVM and iMILP were introduced in the Introduction Section. In identifying the 'responsible' isoform(s), mi-SVM selects the top 25% isoforms as 'responsible' isoforms of a gene for a function, whereas MI-SVM only selects the maximum score isoform as the 'responsible' isoform of a gene for a function. miFV and miVLAD were introduced to solve the large scale MIL problem, miVLAD is based on the Vector of Locally Aggregated Descriptors (VLAD) representation, and miFV is based on the Fisher Vector (FV) representation to map the original MIL bags into new vector representations. To avoid the misled effect of overwhelming negative examples, only five times number of negative samples to that of positive samples were used to generate the new feature vectors for miFV and miVLAD, for training mi-SVM and MI-SVM.

3 Sensitivity analysis of k

To analyze the sensitivity of k, we use different input values of k to construct the isoform functional association network, and then follow the experimental setup in Section 3.2 of the main text to test the performance of IsoFun. The Fmax values and Smin values of IsoFun under different values of k are provided in Fig. S1. From Fig. S1, we can see that the performance of IsoFun increases as k increasing, and remains stable when $k > 30$. Therefore, an effective k can be easily selected from a wide range of values. From this analysis, we adopted $k = 100$ for experiments.

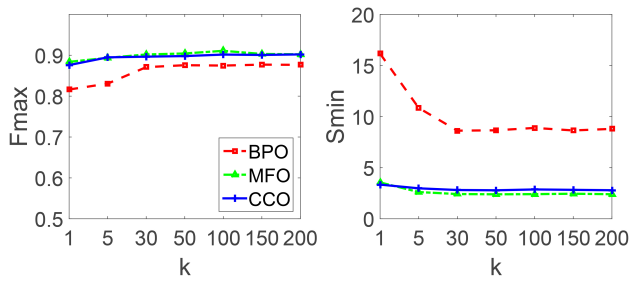

Fig. S1. Fmax and Smin vs. k.

4 Comparison with iMILP

In this section, we adopted the filtering protocol used by iMILP to filter the data. For each RNA-seq data set, an isoform is retained for further analysis if and only if the coefficient of variation (the ratio of standard deviation to mean) of its expression profile is ≥ 0.3, and it is significantly expressed with the expression value ≥ 10 FPKM in at least two experiments. After that, we obtained 7069 genes with 15826 isoforms. The experimental results of IsoFun and iMILP on this dataset is reported in Fig. S2.

From Fig. S2., we can see that IsoFun still has a better performance than iMILP across different evaluation metrics. These experiments prove the effectiveness of IsoFun under different data filtering protocols.

5 Runtime analysis

We also record the runtime costs of these comparing methods and report the cost in Table S1. All the comparing methods are run on a server with CentOS 6.5, Intel Xeon E5-2678v3 and 256GB RAM. Both IsoFun and iMILP run faster than other comparing methods, and IsoFun runs even faster than iMILP, although they both apply label propagation on sparse networks to infer GO annotations of isoforms. The cause is that IsoFun separately propagate label information on subnetworks,

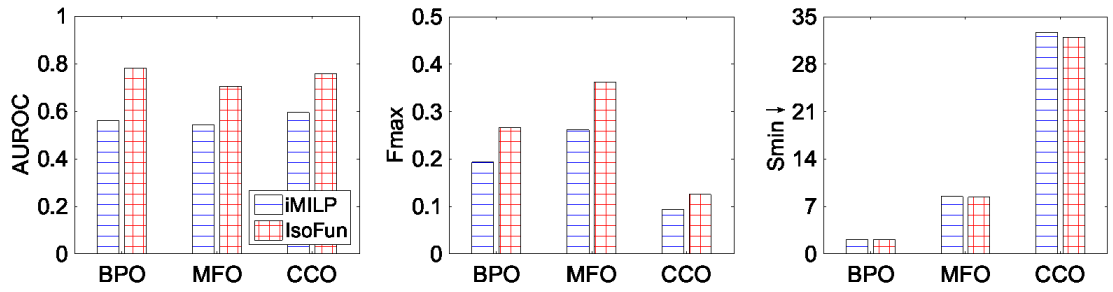

Fig. S2. IsoFun vs. iMILP on the dataset filtered similar as iMILP did.

simultaneously handle multiple labels, whereas iMILP includes a time-consuming normalization process and converges slower than IsoFun. Furthermore, similar as other comparing methods, iMILP also separately handles each label. For miFV, miVLAD, mi-SVM and MI-SVM, we use five times of negative genes to positive ones to accelerate training. The experimental results of reduced negative examples are similar as those of using all negative genes. miFV and miVLAD take a large portion of time to learn the vector of Locally Aggregated Descriptors (VLAD) and Fisher Vector (FV) representation, respectively, so they have larger runtime cost than iMILP and IsoFun. mi-SVM and MI-SVM separately handles each label, and thus have much larger runtime cost than others.

Table S1. Statistics of runtime (seconds).

	miFV	miVLAD	mi-SVM	MI-SVM	iMILP	IsoFun
BPO	8178	6643	210703	154576	15518	661
MFO	1474	1393	42463	37844	2675	160
CCO	1508	1378	51653	40162	3200	124

References

Chatr-Aryamontri, A. *et al*. (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, **45**(D1), D369-D379.

Clark, W.T. and Radivojac, P. (2013) Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, **29**(13), i53-i61.

Eksi, R. *et al*. (2013) Systematically Differentiating Functions for Alternatively Spliced Isoforms through Integrating RNA-seq Data. *PLoS Computational Biology*, **9**(11), e1003314.

Jiang, Y. *et al*. (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, **17**(1), 184.

Kim, D. *et al*. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**(4), 357-360.

Li, W. *et al*. (2014) High-resolution functional annotation of human transcriptome: predictingisoform functions by a novel multiple instance-based label propagationmethod. *Nucleic Acids Research*, **42**(6), e39.

Pertea, M. *et al*. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.. *Nature Protocols*, **11**(9), 1650.

Pertea, M. *et al*. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**(3), 290-295.

Wei, X.S. *et al*. (2017) Scalable Multi-instance Learning. *IEEE Transactions on Neural Networks and Learning Systems*, **28**(4), 975–987.

Zerbino D.R. *et al*. (2017) Ensembl 2018. *Nucleic acids research*, **46**(D1), D754-D761.

Zhang, M.L. and Zhou, Z.H. *et al*. (2014) A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge & Data Engineering*, **26**(8), 1819-1837.