**Protein Expression Data Improves Gene Function Prediction**

**Abstract**

**Background**: The progress of Biological experiment technology improves the understanding of genes and their products, produces a lot of biological information. But many genes function information are still lacking. Detection of gene function in experiment is time and vigour. These biological information is fully used with the development of bioinformatics and there are a growing number of function prediction methods of genes. And the gene functions predicted by researchers can be applied to calculate the relevance between genes and disease by employing the directed acyclic graph structure of Gene Ontology(GO).

**Method**: Human protein expression obtained from mass spectrometry was applied to predict genes functions. We build Pearson correlation coefficient matrix and function probability matrix of GO term annotations. The genes related with unknown genes are filtered by gene-gene interactions data and the genes which rarely interacts with other genes would be removed. The GO terms score were generated by multiplying the two matrix and the GO terms were annotated according to sorting the scores. Our method also was integrated into a framework containing a wide variety of biological information sources, include protein sequences, RNASeq and gene interaction network data.

**Results:** We proposed a novel Pearson correlation coefficient matrix based on protein expression and the genes interactions was employed to filter the matrix. Predicted GO terms were ranked on the basis of their relative association sores with the unannotated gene and the approach was evaluated quantitatively by plotting the precision-recall curves. Comparison with function prediction based on protein sequences, RNASeq and gene interaction network data, Our method performs best for three functional domains. It is found that the protein expression obtained from MS(mass spectrometry) has a competitive edge on gene function prediction. After integrating the four types of biological information, the integrative approach added MS data provides clearly increased accuracy over other methods.

**Conclusion**: Protein expression data has the different proteins` expression in 30 tissues and can improve gene function prediction accuracy. Genes interactions that filter the the genes which rarely interacts with other genes are good for the final prediction precision. When we integrate different types of data in a framework, the result is better than prediction based single biological information.

Keywords: mass spectrometry, rotein sequences, RNASeq, gene interaction network.

**1 INTRODUCTION**

In many areas of biological research, ontologies plays an important role for unification of gene function description. Among widely used ontology is Gene Ontology(GO), which describes genes and gene products function using a hierarchical structure with three distinct annotation categories, molecular function(MF), cellular component(CC) and biological process(BP). GO team also constructs directed acyclic graph(DAG) to describe relationship of GO terms and many terms may have a lot of branches to relate with other terms, even some terms included different categories frequently occur together to annotate the same gene. GO has been growing in size and currently includes a total of 34765 terms and 64635 term-term relations annotating genes and it is very convenient for researchers to search functions of a query sequence.

Researchers found out closely-related genes with unknown gene and associated unknown gene with closely-related genes’ function. This principle called ‘guilt-by association’ which transfers function from one gene to another via biological relationships derives many methods based on gene sequence similarity, gene structure similarity and gene interaction network.

Conventional protein function prediction methods based on sequences such as BLAST[[1]](#endnote-0), FASTA[[2]](#endnote-1) and SSEARCH[[3]](#endnote-2) rely on the concept of homology. BLAST produces the E-values of sequence hits and the hits include many sequences that are homologous with the query sequence. But when the sequences hits include less homologous sequences of the query sequence, the prediction accuracy is low and sequence similarity does not necessary imply functional similarity. ESG[[4]](#endnote-3) method iterates the search hits from PSI-BLAST to capture functions of a protein, even from weakly similar sequences and reduces the dependence of accuracy for homology.

Furthermore, protein sequences do not provide enough information on the biological functions which involved in metabolic pathway and biological process. And protein-protein interaction(PPI) and protein structure can provide additional function information. The proteins that lie closer to one another in the PPI network are more likely to have similar function. Schwikowski[[5]](#endnote-4) annotated a protein three functions that are most frequent among its neighbors. Although it is effective, the approach does not make full use of the topology of the network. The clustering algorithms are applied to gene interaction network to predict genes functions by more methods. Vladimir[[6]](#endnote-5) used penalized non-negative matrix tri-factorization(PNMTF) and it takes all interaction network in a matrix to perform simultaneous clustering of genes and GO terms. Some methods predict function for fold structure searching, the best known being DALI[[7]](#endnote-6), SSM[[8]](#endnote-7) and GRATH[[9]](#endnote-8).

The methods, such as PEDANT[[10]](#endnote-9) and GeneQuiz, mining the sequence information from free text annotation. GOtcha[[11]](#endnote-10) reconstructs the structure of Gene Ontology vocabulary and search relevant terms for scoring relevant genes.

Some researchers have integrated data from multiple sources for the gene annotation. MRF[[12]](#endnote-11) approach reconstructs multiple networks in the annotating process and Lee[[13]](#endnote-12) applied the method to predict yeast proteins function. MNet[[14]](#endnote-13) optimize the composite network and unify the prediction problem in a objective function. So optimization of the model of prediction must produce positive influence and can effectively integrate multiple networks.

In 2014, Mathias Wilhelm and coworkers present a draft map of the human proteome using high-resolution Fourier-transform mass spectrometry. The proteomic profiling includes 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells.This article uses human protein expression obtained from mass spectrometry to predict genes functions and compares with the methods based on protein sequences, RNASeq and gene interaction network data.

**2 Method**

**2.1 Pearson correlation coefficient Matrix Building**

It is assumed that two gene have a closed functions relationship if they have similarity expression profiles. Based on the assumption it is inferred that a unknown gene` functions can be predicted from the genes which are annotated by GO and associated with the unknown gene. So we transfer the problem from gene functions prediction to searching the genes having strong association with unknown gene.

A item of gene  expression  from MS data was fetched out as unknown gene` expression and then the whole protein expressions ,,…, were traversed to calculate the Pearson correlation coefficient as follows:



Where  is the correlation of the two genes,  and ,  is the average and  denotes the standard deviation of gene ` expression in 30 tissues.

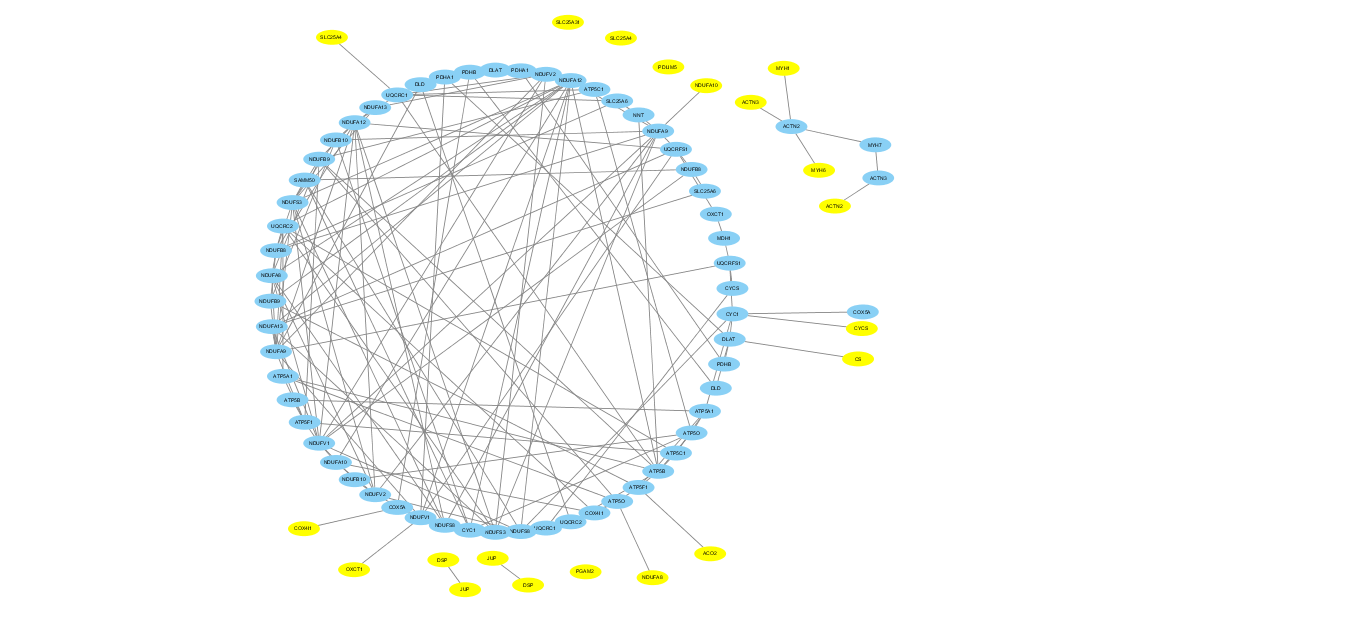
If  is more than 0.6, we think that the gene  is associate with the unknown gene  and add its expression to Pearson correlation coefficient matrix. If  is less than 0.6, we think that the correlation is too low to have function intersection between the two genes. If there are m genes added in matrix, the dimension of the matrix is 1\*m for the unknown gene as follow:



Where  is the Pearson correlation coefficient matrix and there are m genes is associate with unknown genes on biological function. The genes strongly associated with the gene unannotated in GO are extracted from MS data. And these genes had potential for continuing being filtered.

**2.2 filtered by interaction network**

Next, we filter the genes from the first matrix by building their interaction network. We construct network of genes in  from genes interactions data and if genes rarely interact withe others, they are considered false positive results and would be removed.



**Figure1 Interaction network of genes from Pearson correlation coefficient Matrix associated with TTN as the gene unknown function.** The genes painted with yellow interact with other genes in the network less than twice.

In Figure 1, isolate genes that rarely interact with other genes are considered to not involved in the biological process that involves the unknown gene TTN and have little shared function with TTN. So these genes should be remove to let the genes that really associate with TTN join the next step.

**2.3 function prediction building**

The second matrix is function probability matrix and the element is calculated as follow:





Where  denotes the probability of  occurring under gene  being considered in first matrix,  is the count of GO terms of gene ,  denotes the count of genes annotated by  and ,  is a priori probability denoting the frequency of GO term  in Gene Ontology DAG,  is probability of  annotating a gene when  annotates the gene. Then the function matrix  is built as follow:



The final score matrix assigned to all GO term is determined by multiplying the Pearson correlation coefficient matrix and function probability matrix.



 is the final score for the GO term , dimension of  is the number of GO terms. The unannotated gene  is assigned GO terms ranked based on the their value in matrix .

**2.4 Recall-precision calculate**

Precision is calculated the number of genes correctly classified as having a given GO term divided by the total number of genes classified as having that GO term. Recall is defined as the percentage of genes annotated with a given GO term that were classified as having that GO term.





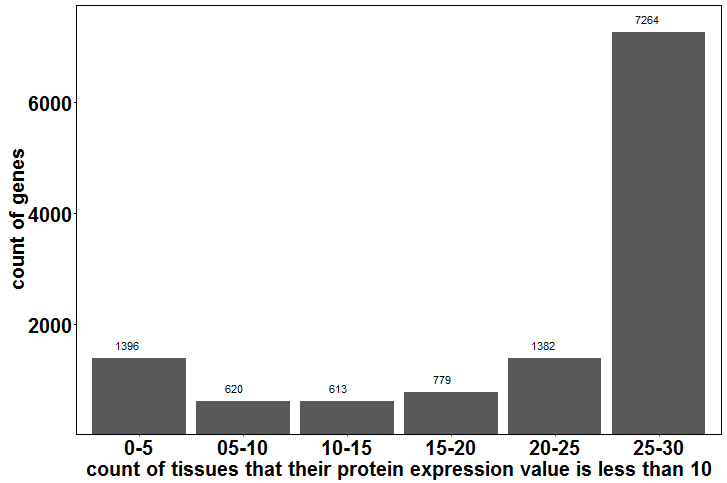
Where TP is the number of the genes predicted the GO term when the GO term annotates the genes in GO, FP is the number of genes predicted the GO term but the GO term is not included in the genes` annotation and FN is the number of genes which are not predicted the GO term.

**3 Result**

**3.1 MS data filtering**

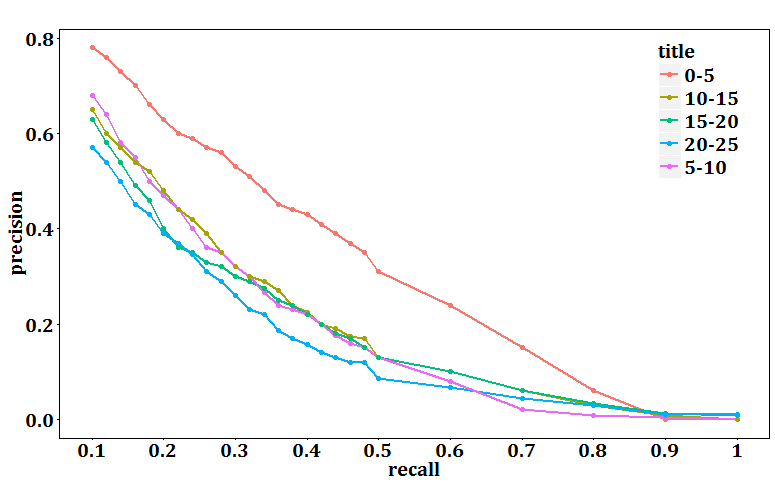
Genes standard functions are downloaded from Gene Ontology and its DAG works for search the relations among the GO terms . There are 17294 protein-coding genes expression in 30 tissues includes 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells from MS data. RNASeq downloaded from HPA database(www.proteinatlas.org) includes 11745 genes at RNA levels in 45 cell lines and 32 tissues. Gene interaction network downloaded from Human Interactome Database includes 13944 pairs genes. Proteins sequences are downloaded from Uniprot database.

But the MS data includes 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells is inadequacy to describe the protein expression profile in human and protein expressions are unknown in many other tissues. The gene expressions are unknown in human tissues except 30 tissues from MS data. If the protein expressions in many tissues from MS data are too low, Pearson correlation coefficient between two genes will be not trustworthy. We divide proteins expression data into five data sets according to the count of tissues that their protein expression values are less than 10 in 30 tissues and the six data sets are applied to predict gene function.



**Figure 2 distribution of six data sets divided according to the count of tissues that their protein expression values are less than 10 in 30 tissues in MS data.**

In Figure 1, the count of x-axis between 15 and 30 accounts for 60 percent of total proteins expression and it implies that expressions of 60 percent of proteins in MS data are too low to provide enough information about performance of proteins in 30 tissues. We draw precision-recall curve to show the result and the tissues count less than 5 performs better than others.



**Figure 3 Precision-recall of the method based on five data sets.**

Figure 1 obviously shows that the red curve represents the precision of method based on the data that having the least count of tissues that their protein expression value is less than 10 in 30 tissues. And the red curve is above other curves across all the sessions on the recall coordinate system. It implies prediction result is better as the the count of tissues that their protein expression less than 10 decreasing and can be infered the correlation coefficient is inaccurate if there are low protein expression values occurring in many tissues.

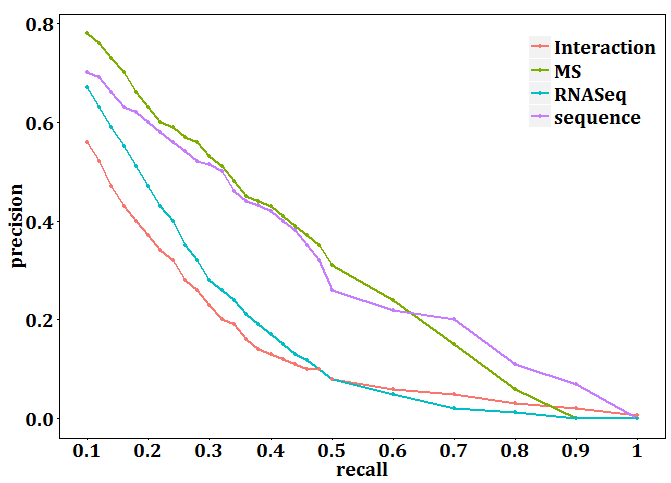
Incomplete MS data has some influence on the final function prediction result and the result will be better if the proteins expression is known in all human tissues. So the MS data has potential for clearly improving the precision of genes functions prediction. It is unfair if we compare our method based on the whole MS data with other methods and we need filter the whole MS data set. So the data set having the best performance was chosen as our training set.

**3.2 comparison of four types of data sets**

The MS(mass spectrometry) data was divided into two data set, training set and test set according to whether the gene`s function is known. We construct two matrix, the expression similarity matrix and the function association matrix. The expression similarity matrix is built by calculating the value of Pearson correlation coefficient and describes the relevance of two genes. The genes related with unknown genes are filtered by gene-gene interactions data and the genes which rarely interacts with other genes would be removed. Next, the function association matrix is built by constructing the bayesian network to calculate the probability that two GO terms are associated to the same gene based on frequency at which they co-occur in all genes functions. We can obtain score matrix by multiplying the two matrices that indicates the relationship between unknown gene and GO terms. So we put the high score GO terms to unknown gene.

RNASeq data, gene interaction data and protein sequences are also applied to predict gene function. We extends the original PNMTF algorithm[[15]](#endnote-14) to handle the gene interaction network. It takes network data in a matrix form and performs simultaneous clustering of genes and GO terms and clustering process is under the guidance of prior knowledge given in the form of intra-type pairwise constraints. Protein sequences can be the input of ESG method which performs iterative protein sequence database searches. ESG uses two level neighbors by performing PSI-BLAST software which produces E-values of query sequences` hits. And function association matrix is used by ESG to view correlation between GO terms in annotation files. The RNASeq data includes 11745 gene` expression on RNA level in 32 tissues and cell lines and its data format is similar with proteins expression from MS analysis. So we adopt the same methods that predicts gene function based on MS data applying to RNASeq.

We perform gene function prediction method described from second part based on MS data and compared it with the different methods based on other data and plot the average precision and recall curve across the whole set of gene from MS data after filtering.



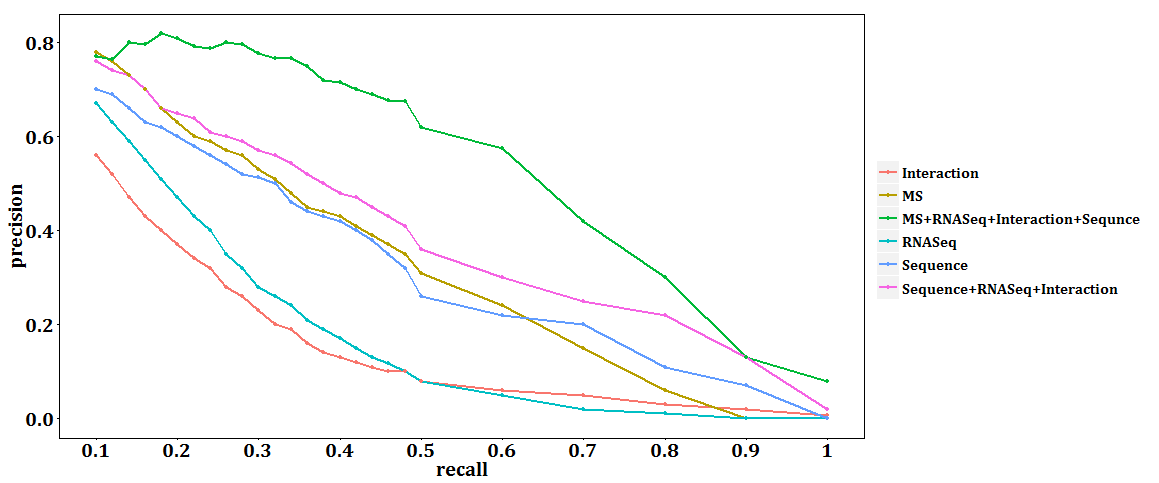
**Figure 4 Precision-recall curve of four methods based on four types of biological information.**

The comparison results show that the method based on protein sequences and based on MS data both can reach a similar high precision and our method achieves the higher prediction accuracy. The accuracy gradually reduces as the recall growth and the performance on molecular function and cellular component is better than the performance on biological process. With the recall between 0. and 0.2, it is obvious that the curve of MS data is above others.

We can draw a conclusion that proteins expression data has a advantage on gene function prediction and we will watch its performance in integrating MS data with others.

**3.3 integration of data sets**

Then we integrate MS data, RNASeq, gene sequences and gene interaction network data to predict gene function and compare with the method only using MS data. We develop three different probabilistic scores to combine RNASeq, Interaction network and protein sequence data and four probabilistic scores to combine RNASeq, Interaction network, protein sequence data and MS data. We compare the prediction accuracy among integrating data sets and single data, such as interaction network data and protein expression data.

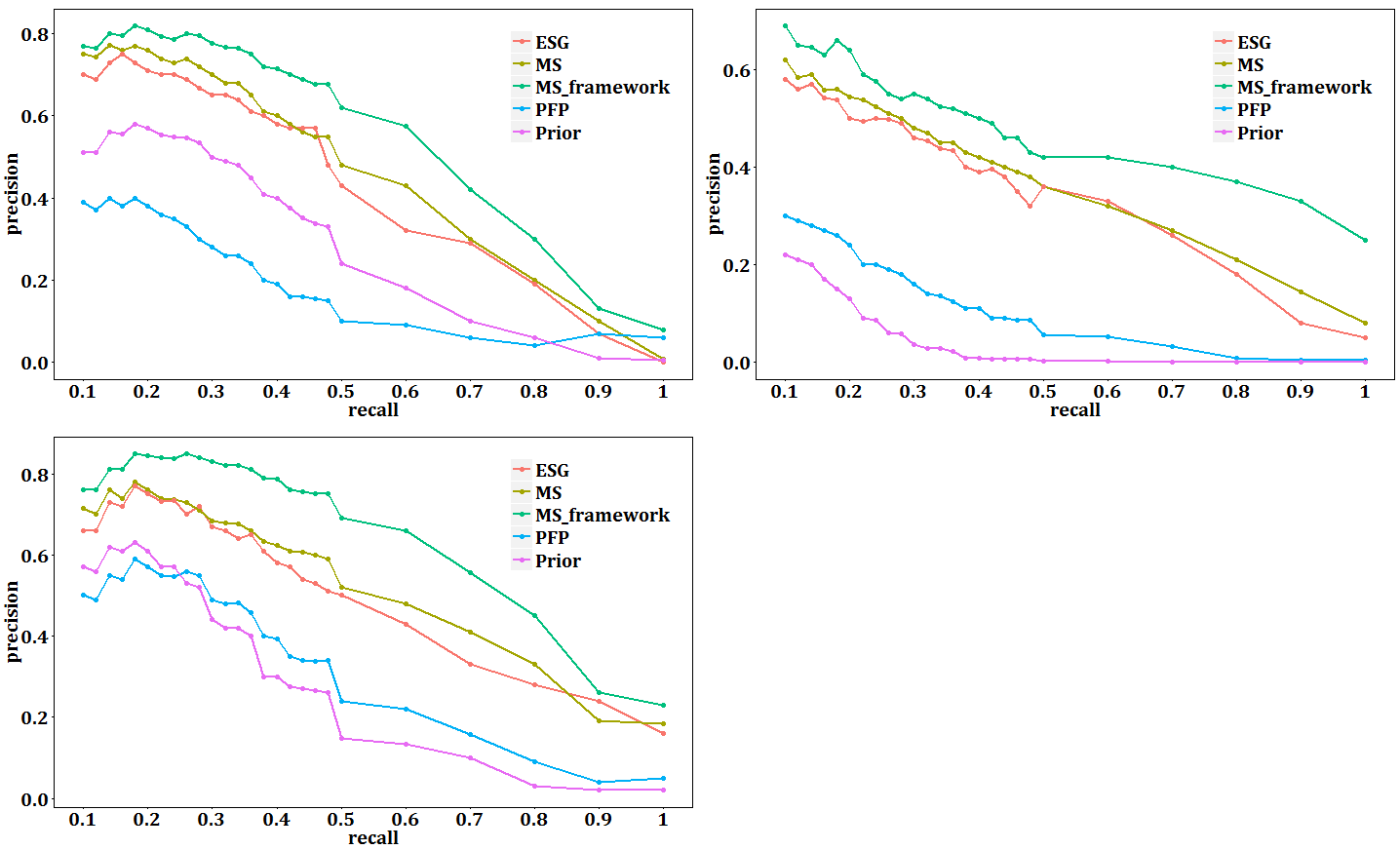


**Figure 5 Precision-recall curve of methods based on integration of different types of biological information.**

The comparison results show that integrating data makes the accuracy higher than only using MS data. The curve of combination of three data sets is slightly above the curve of protein sequence data and it shows that it helps to integrate sequence with RNASeq and genes interactions for improving prediction precision. The result of integration of four data which adds MS data based on the combination of three data is clearly improved across the whole recall-coordinate. The Figure5 implies that the description of integrating data for biological activity is more comprehensive.

**3.4 Comparison with three classical gene function prediction methods**

We compare the performance of our method and framework with the three predictors ESG, PFP and Prior.



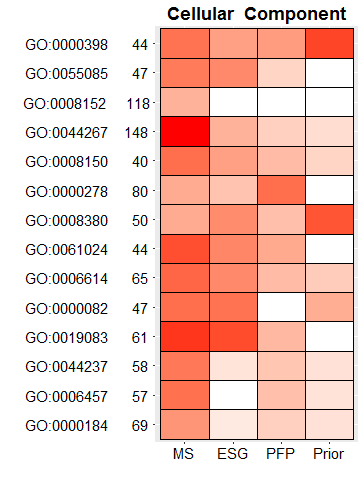
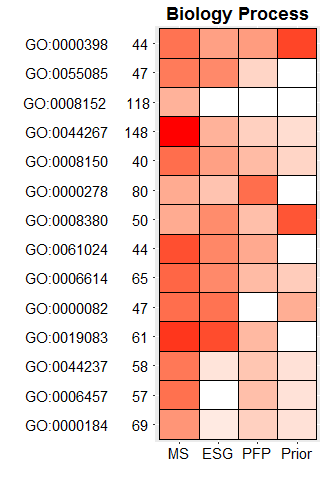
**Figure 6 Precision-recall curve of five prediction methods.** MS\_framework represents the method based on integration MS data with other three data sets.

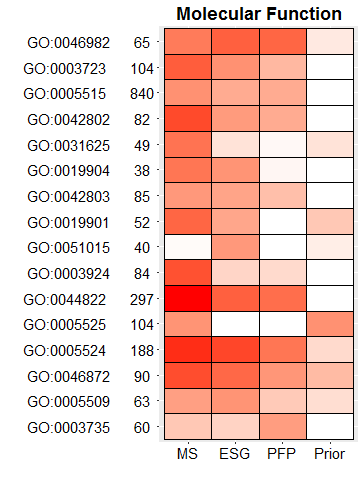
The results clearly show that the framework that integrates MS with other three data sets significantly outperforms the three methods on molecular function, cellular component and biological process when evaluated by the precision-recall plot.

**3.5 Prediction accuracy for different GO terms**

For four methods we analyze the prediction accuracy for different GO terms for each functional category. Only GO terms that are used for annotating 40 or more targets are considered. The F1 measure is used for evaluation. The average F1 scores across the terms for our method, PFP[[[16]](#endnote-15)], ESG[[[17]](#endnote-16)] and Prior represents the effect of the gene function prediction. Our method performs best for three functional domains.

In Figure 4, we analyze the effect of function prediction for different FO terms. Only GO terms that are annotated 40 or more genes are chosen. This results include 144 BP terms, 87 MF terms and 64 CC terms.





**Figure 7 Prediction accuracy evaluated for each functional category.** Each row represents a GO term category and each column represents a function prediction method. The number behind the GO term id is the count of target genes that were annotated by the term. The intensity of the color represents the F1 measure that was used for evaluation and calculated in equation(6).

For a GO term under consideration, we counted how many target genes that were annotated by the GO term and the number of genes that were correctly predicted to have the same term for each prediction methods. For example, there were 44 out of 1396 targets that were annotated by BP term GO:0061024. The true positive(TP) was considered as the number of targets that were correctly predicted and the number of targets that were not annotated by the GO term but predicted to have the term was considered as false positive(FP). The number of targets that were annotated by GO:0061024 but were not predicted was consider as false negatives(FN). For each term, the precision and recall were calculated in equation(4)(5) for each method. Further, F1 measures for the term was given by equation(6) and represents the effect of prediction for each method.

 (4)

 (5)

 (6)

We compared method based on MS data with other prediction methods, such as ESG, PFP, Prior.

14 out of the 144 BP terms, our method showed the highest F1 measure among the four methods for 8 terms. PFP, ESG, Prior showed highest F1measure for 3, 2 and 1 terms, respectively. The average F1 measures across the 44 BP terms by MS, ESG, PFP, Prior were respectively 0.384, 0.359, 0.27, 0.22. In Figure 4A, there is a sample of 14 BP terms out of 44. For the term GO:0000184 nuclear-transcribed mRNA catabolic process that is used to annotate 69 BP targets, method based on MS data performed the best F1 measure value of 0.43 followed by ESG(0.412), PFP(0), Prior(0). On the other hand, for GO:0044267 cellular protein metabolic process that shares annotations with 148 genes, all four methods showed comparable performance with F1 measure around 0.36 and for GO:0000278 mitotic cell cycle that annotate 40 targets, the F1 measure value of all four methods is very low. Overall, in most cases, our method performed better than the others for BP terms.

There are 16 MF terms showed in Figure 4B. In seventeen out of twenty-seven terms annotating more than 40 genes, our method performed the best F1 measure. The average F1 measures across the 16 MF terms by MS, ESG, PFP, Prior were respectively 0.362, 0.358, 0.231 and 0.167, respectively. As the case with CC domains, overall our method performed best for all three domains.

**4 function prediction for genes unannotated in Gene Ontology from MS data**

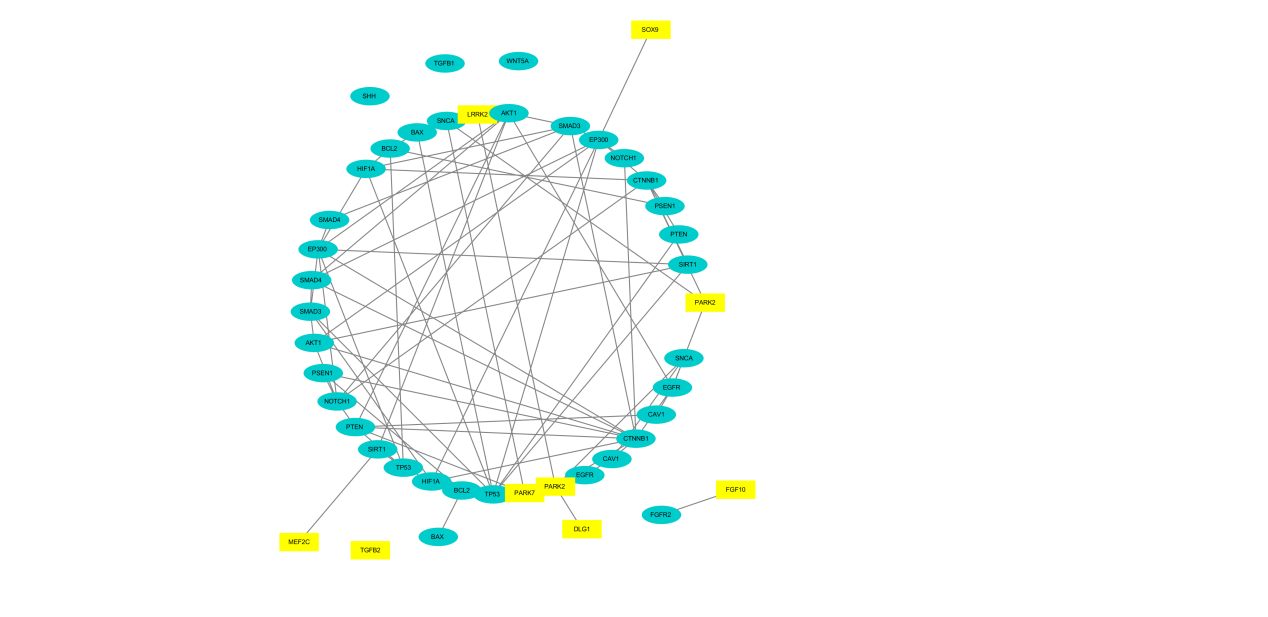
In the mass spectrometry experimental process, a uniqe and comprehensive strategy for proteogenomic analysis was enabled researchers to discover 2535 protein-coding genes that have not been previously observed by large community-based proteomic data sets. These genes were a focus on their performance in biological process and finally, we discuss the prediction example that our method succeeds in function prediction of genes which express highly in 30 human tissues and is unannotated in Gene Ontology. Since the number of actual and predicted GO terms for a gene may be very large when low scores that was assigned to, Table 1 includes top 10 terms that are strongly associate with the unannotated gene.

**Table 1 three unannotated genes and their top ten predicted GO terms ranked by score.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Genes** | **GO Term** | **Definition** | **Score** |
| LOC390956 | GO:0070062 | extracellular exosome | 0.68 |
| GO:0005634 | nucleus | 0.66 |
| GO:0005515 | protein binding | 0.54 |
| GO:0005737 | cytoplasm | 0.54 |
| GO:0005829 | cytosol | 0.53 |
| GO:0044822 | poly(A) RNA binding | 0.42 |
| GO:0000786 | nucleosome | 0.3 |
| GO:0046982 | protein heterodimerization activity | 0.29 |
| GO:0003677 | DNA binding | 0.18 |
| GO:0005925 | focal adhesion | 0.11 |
| HIST1H4K | GO:0008380 | RNA splicing | 0.98 |
| GO:0005515 | protein binding | 0.91 |
| GO:0005654 | nucleoplasm | 0.87 |
| GO:0003723 | RNA binding | 0.74 |
| GO:0000786 | nucleosome | 0.65 |
| GO:0016020 | membrane | 0.63 |
| GO:0010467 | gene expression | 0.59 |
| GO:0046982 | protein heterodimerization activity | 0.56 |
| GO:0008152 | metabolic process | 0.42 |
| GO:0006412 | translation | 0.28 |
| COX2 | GO:0005739 | mitochondrion | 0.88 |
| GO:0044281 | small molecule metabolic process | 0.75 |
| GO:0005743 | mitochondrial inner membrane | 0.72 |
| GO:0070062 | extracellular exosome | 0.72 |
| GO:0044237 | cellular metabolic process | 0.69 |
| GO:0005515 | protein binding | 0.67 |
| GO:0022904 | respiratory electron transport chain | 0.52 |
| GO:0005759 | mitochondrial matrix | 0.44 |
| GO:0005634 | nucleus | 0.39 |
| GO:0043209 | myelin sheath | 0.27 |

**5 Genes associated with disease prediction**

After these test, our method is proved that it has a competitive edge on gene function prediction and we predict disease related genes based on semantics similarity using the shortest distance in GO directed acyclic graph. In this work, we introduce a approach[[18]](#endnote-17) for ranking disease candidate genes on functional comparisons involving Gene Ontology and extend the methodology to introduce additional relevance between the GO term and their annotated genes. Firstly, the functional annotations of known disease genes was assigned to scores with the disease based on the correlation coefficient between the known disease genes and the disease. Then we exploits the similarity between the gene functional annotations of diseases and candidate genes. The relevance of them is determined to search for the shortest path between GO terms from annotations of disease and candidate genes annotation in Gene Ontology DAG construction. The annotations of disease was traversed to calculate the correlation with candidates and the final score was assigned to averaging the results of traversing.



**Figure 8 the results of interaction network of genes predicted to associate with stomach cancer.** The blue color represents the true positive results that were known as disease genes and the yellow color represents the genes considered irrelevant with stomach cancer.

In Figure 8, the genes painted with yellow are less interacting with other genes in comparison with genes known as disease genes.

So if we get a gene`s expression at protein level, we can predict its GO terms and calculate its correlation coefficient with genetic diseases.

**Conclusion**

A novel method based MS data was proposed to establish relation between genes unannotated in GO and GO terms. The performance of the proposed method was evaluated and compared with three other methods based on different types of biological information, protein sequence, RNASeq, interaction network. The proposed method outperforms above mentioned methods. At the same time, a integration framework was built to integrate protein expression data and other data sets. The framework was evaluated and compared with the classical gene function prediction methods, PFP, ESG, Prior. With the results it was clear that our framework based on MS data and other data performed better than any other methods. The improving of function prediction also benefits the prediction of new disease-gene associations for a particular disease.

The current prediction results indicate that there is considerable room for improvement in the field. And as the growth of biological information, the need for accurate automatic function prediction methods will be more strong. We expect that improved, more accurate methods make readily accessible to the biological community and there will be more breakthroughs on functional annotation.

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402. doi: 10.1093/nar/25.17.3389. [↑](#endnote-ref-0)
2. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 1990;183:63–98. doi: 10.1016/0076-6879(90)83007-V. [↑](#endnote-ref-1)
3. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85:2444–8. doi: 10.1073/pnas.85.8.2444. [↑](#endnote-ref-2)
4. Chitale M, Hawkins T, Park C, Kihara D: ESG: extended similarity group method for automated protein function prediction. Bioinformatics. 2009, 25: 1739-1745.10.1093/bioinformatics/btp309. [↑](#endnote-ref-3)
5. A network of protein-protein interactions in yeast. Schwikowski B, Uetz P, Fields S. Nat Biotechnol. 2000 Dec; 18(12):1257-61. [↑](#endnote-ref-4)
6. Vladimir Gligorigevic, Vuk Janjic, Natasa Przulj. Integration of molecular network data reconstructs Gene Ontology. Bioinformatics, 2014, 30(17): i594-i600. [↑](#endnote-ref-5)
7. L. Holm, C. Sander. Protein structure comparison by alignment of distance matrices. J Mol Biol, 233 (1993), pp. 123–138. [↑](#endnote-ref-6)
8. E. Krissinel, K. Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr, D60 (2004), pp. 2256–2268. [↑](#endnote-ref-7)
9. A. Harrison, F. Pearl, I. Sillitoe, T. Slidel, R. Mott, J.M. Thornton, C. Orengo. Recognising the fold of a protein structure. Bioinformatics, 19 (2003), pp. 1748–1759. [↑](#endnote-ref-8)
10. Functional and structural genomics using PEDANT.

    Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW

    Bioinformatics. 2001 Jan; 17(1):44-57. [↑](#endnote-ref-9)
11. Martin D, Berriman M, Barton G. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics. 2004;5:178–94. doi: 10.1186/1471-2105-5-178. [↑](#endnote-ref-10)
12. Deng M, Tu Z, Sun F, Chen T. Mapping Gene Ontology to proteins based on protein-protein interaction data. Bioinformatics. 2004, 20(6):895-902. [↑](#endnote-ref-11)
13. Lee H, Tu Z, Deng M, Sun F, Chen T. Diffusion kernel-based logistic regression models for protein function prediction. OMICS. 2006 Spring; 10(1):40-55. [↑](#endnote-ref-12)
14. Guoxian Yu, Hailong Zhu, Carlotta Domeniconi, at al. Integrating multiple networks for protein function prediction[J]. BMC Syst Biol, 2015, 10:1186-1752. [↑](#endnote-ref-13)
15. Wang D, Gao X and Wang X. Semi-Supervised nonnegative matrix factorization via constraint propagation. IEEE Trans Cybem, 2016, 46(1): 233-244. [↑](#endnote-ref-14)
16. [] Hawkin T, et al. Enhanced automated function prediction using distantly related sequences and contextual association by PFP[J]. Protein Sci, 2006, 15:1550-1556. [↑](#endnote-ref-15)
17. [] Meghana Chitale, Troy Hawkins, Changsoon Park, Daisuke Kihara. ESG: extend similarity group method for automated protein function prediction[J]. Bioinformatics, 2009, 25(14):1739-1745. [↑](#endnote-ref-16)
18. Andreas Schlicker, Thomas Lengauer and Mario Albrecht. Improving disease gene prioritization using the semantic similarity of Gene Ontology term[J]. Bioinformatics, 2010, 26(18): i561-i567. [↑](#endnote-ref-17)