Inferring Protein-Protein Interactions Based on Sequences and Interologs in Mycobacterium Tuberculosis

Zhi-Ping Liu[1], Jiguang Wang[2], Yu-Qing Qiu[3], Ross K.K. Leung[4],
Xiang-Sun Zhang[3], Stephen K.W. Tsui[4], and Luonan Chen[1]

[1] Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences, Shanghai 200031, China
[2] Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China
[3] Academy of Mathematics and Systems Science, Chinese Academy of Sciences,
Beijing 100190, China
[4] Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong,
Shatin N. T., Hong Kong, China
lnchen@sibs.ac.cn

Abstract. *Mycobacterium tuberculosis* is a pathogenic bacterium that poses serious threat to human health. Inference of the protein interactions of *M. tuberculosis* will provide cues to understand the biological processes in this pathogen. In this paper, we constructed an integrated *M. tuberculosis* H37Rv protein interaction network by machine learning and ortholog-based methods. Firstly, we developed a support vector machine (SVM) method to infer the protein interactions by gene sequence information. We tested our predictors in *Escherichia coli* and mapped the genetic codon features underlying protein interactions to *M. tuberculosis*. Moreover, the documented interactions of other 14 species were mapped to the proteome of *M. tuberculosis* by the interolog method. The ensemble protein interactions were then validated by various functional linkages i.e., gene coexpression, evolutionary relationship and functional similarity, extracted from heterogeneous data sources.

1 Introduction

M. tuberculosis is the causative agent that causes tuberculosis and leads to lesions in lungs and other organs. Tuberculosis is the second leading cause of death in infectious diseases. An extensive protein-protein interaction (PPI) network of *M. tuberculosis* can lead to more comprehensive screens of its cellular operations. To date, genome-wide experimental and computational systems for studying PPIs in *M. tuberculosis* is not available [1]. It is urgently necessary to develop approaches capable of converting available genomic data into functional information for *M. tuberculosis*. *E. coli* is one of the best model systems to study bacterial physiology, with well-characterized interactome, genome and transcriptome [2]. Interaction features can be learned by machine learning methods, such as support

D.-S. Huang et al. (Eds.): ICIC 2011, LNBI 6840, pp. 91–96, 2012.

vector machines (SVMs) [3], and also it is common to predict protein interactions from known interactions of other organisms by interolog method [4].

Genetic information in the form of codons, i.e. tri-nucleotide sequences, specifies amino acid sequence in the polypeptide during the synthesis of proteins. It is well known that codon usage is correlated with expression level [5]. Genetic codons will be selected as the sequence features in the learning of interaction patterns. Moreover, the corresponding orthologs of interacting proteins in other organisms will provide more information about the potential interaction mappings by comparative genomics.

In this work, we developed a systematic method combining heterogeneous data sources to infer a comprehensive protein interaction network in *M. tuberculosis.* The codon features of interacting protein pairs are detected and used to train an SVM classifier. Moreover, the interactions from other 14 species are mapped to *M. tuberculosis* by the interolog method. The available data from multiple levels including gene coexpression and evolutionary relationship to functional similarity are implemented to assess these predicted interactions by confidence significance. The predicted protein interaction network as well as the proposed method provide a framework for the functional specificities study of *M. tuberculosis.*

2 Methods

2.1 Framework of Prediction

The protein interactions were predicted by two main pipelines. Firstly, we built the protein interaction network of *M. tuberculosis* from codon features of interacting proteins in *E. coli* by machine learning approach. The integrated interaction maps and gene sequences of *E. coli* were retrieved from EcID [2]. The ORFs of *M. tuberculosis* were derived from the laboratory strain H37Rv. We used the information of protein interaction network of *E. coli* to train an SVM classifier to get the genetic codon features underlying the interacting pairs. The interactions in *M. tuberculosis* were then predicted by the trained SVM predictor with the genetic codons of ORFs in gene sequences of *M. tuberculosis.* Secondly, we inferred the protein interactions of *M. tuberculosis* by interolog method from the documented protein interactions in 14 species. We collected these interactions from IntAct [6] and DIP [7] and the *M. tuberculosis* orthologs of these interologs were identified by BLAST [8]. The homologs of two interacting proteins will be identified as the predicted interactors. As for the validation of predicted results, we tested our method in *E. coli.* Three pieces of available information of *M. tuberculosis*, i.e., gene expression profiling, evolutionary relationship from ortholog database and functional similarity, were used to evaluate the prediction results.

2.2 Validation from Multiple Resources

We implemented multiple available resources to access the constructed PPI network in *M. tuberculosis.* The confidence of interactions was evaluated by three extra data sources, namely, gene expression, evolutionary relationship and functional similarity. Firstly, we identified the Pearson correlation coefficients (PCC) of gene coexpression of pairwise proteins in the predicted network. We downloaded the gene expression data of *M. tuberculosis* from NCBI GEO (ID: GSE9776). Secondly, we presented the evaluation of evolutionary relationship between the predicted interacting proteins. Clusters of orthologous groups (COGs) were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain. The maximum of COG value between two groups in which the interacting proteins located were regarded as the value representing their evolutionary relationship. Thirdly, Gene Ontology (GO) similarity between the predicted pairs were identified to evaluate their functional relationship. We used semantic similarity measures [9] to evaluate the similarity of GO term lists corresponding to the interacting proteins.

3 Results

3.1 Performance of Predictor

E. coli is one of the best characterized organisms and has been served as a model system to study many aspects of bacterial physiology [2]. The positive and negative sets of protein interactions in *E. coli* were designed to test the performance of our codon-based prediction methods. The genome and proteome of *E. coli* were downloaded and prepared for the interacting sets as well as all known opening reading frames (ORFs). The distance of two ORFs in terms of usage of codon c is defined as $d_{ij}(c) = |f_i(c) - f_j(c)|$, where $f_i(c)$ and $f_j(c)$ are relative frequencies of codon c in ORF i and ORF j. By codon definition, $\sum_k f_i(c_k) = 1$ and $\sum_k f_j(c_k) = 1$ for $k = 1, 2, ..., 64$ in all codons. There are 14058 pairs of interactions and 27882 pairs of non-interactions in 4227 proteins of *E. coli*. A five-fold cross validation process is implemented in these pairs. Figure 1 shows the performance of prediction results by the SVM predictor using genetic codon features. There are several codons corresponding to the same amino acid in genetic code. The prediction performance of merging the frequency of these degenerate codons ('codon-mer') is also shown in Figure 1. The details of prediction precision and accuracy are listed in Table 1. The results provide evidences for the effectiveness and efficiency of predicting protein interactions from the genetic codons by machine learning method.

Table 1. Prediction performances of the codon-based SVM predictor in *E. coli.*

Feature	ACC	SN	SP	PRE	AUC
Codon	0.9003	0.7576	0.9486	0.8327	0.9507
Codon-mer	0.9595	0.8986	0.9801	0.9386	0.9835

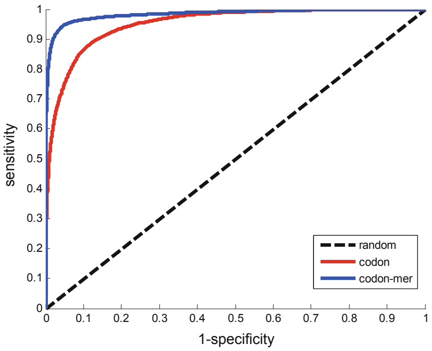

Fig. 1. ROC curves of the five-fold cross validation predictions in *E. coli.*

3.2 Protein Interactions in *M. tuberculosis*

To explore protein interactions in *M. tuberculosis*, we used the formerly trained SVM classifier to infer the interactions of *M. tuberculosis* by the codon message of ORFs in gene sequence level. Based on the genetic codons of *M. tuberculosis* H37Rv, we predicted 12,899 interactions in 3,266 proteins. Furthermore, the known protein interactions of other species were mapped to the proteome of *M. tuberculosis* by interolog method. We collected the documented interactions of 14 species from PPI databases, IntAct and DIP, and the sequence features of interacting proteins were transferred into the *M. tuberculosis* proteome by ortholog detection. Table 2 lists the detailed prediction results by interolog method. The known protein interactions were also included in our inferred interactome of *M. tuberculosis.* So far, we also found 530 pairs of protein interactions of *M. tuberculosis* from various databases, such as BIND [10] and Reactome [11]. Combining with these known interactions, we built a comprehensive protein interaction map totally with 46,119 interactions of 3,465 proteins in *M. tuberculosis.*

3.3 Validation Results

Protein interacting pairs are identified with close relationship with gene coexpression, coevolution, similar GO annotations. To every predicted interacting pairs of *M. tuberculosis*, we collected these available heterogeneous data sources to annotate them. Firstly, we annotated the predicted pairs by their corresponding PCC of gene coexpression. For comparison, we calculated the corresponding correlation values of these same-size random selected protein pairs. Every prediction was then annotated by a coexpression value in gene expression profiling. Figure 2 (a) shows the boxplot of coexpression values in the predictions. From Figure 2, we identified that the coexpression values in the predicted interacting pairs tend to be more correlated when compared to that of random selected ones. Secondly, we identified the evolutionary relationship of the interacting proteins by COG information. The interacting proteins were detected in their own COG individually. Figure 2 (b) shows the boxplot of evolutionary relationship values in

Table 2. Details of predicted protein interactions in *M. tuberculosis*

Species	Database	Original PPI	Predicted PPI	Percentage (%)
By machine Learning				
E. coli	ECID	14,058(positive)+ 27,882(negative)	12,899	27.97
By interolog				
Escherichia coli	IntAct	14,158	16,468	35.71
Campylobacter jejuni	IntAct	11,870	7,674	16.64
Treponema pallidum	IntAct	3,744	324	0.70
Synechocystis	IntAct	2,625	2,481	5.38
Myxococcus xanthus	IntAct	384	253	0.55
Synechocystis sp.	IntAct	219	220	0.48
Rickettsia sibirica	IntAct	282	24	0.05
Streptococcus pneumoniae	IntAct	193	47	0.10
Drosophila melanogaster	DIP	22,650	1,558	3.38
Saccharomyces cerevisiae	DIP	21,769	2,701	5.86
Caenorhabditis elegans	DIP	3,979	229	0.50
Homo sapiens	DIP	1,485	84	0.18
Mus musculus	DIP	287	36	0.06
Rattus norvegicus	DIP	69	2	0.15
Total: 46,119 interactions in 3,465 proteins (with 530 known PPIs)				

the predicted interacting pairs and that of the same-size random selected protein pairs. Every predicted interaction gets a confidence of evolutionary relationship. Thirdly, we calculated the functional similarities underlying these predicted interactions. We detected the semantic similarity between the GO term pairs of interacting proteins. The boxplots of the three values of GO similarities, i.e., cellular component ('CC'), molecular function ('MF') and biological process ('BP'), in random pairwise proteins and that in predicted pairs are shown in Figure 2 (c), (d) and (e), respectively. The predicted interactions have higher functional similarity than random ones, which further provides evidence for the accuracy of our results.

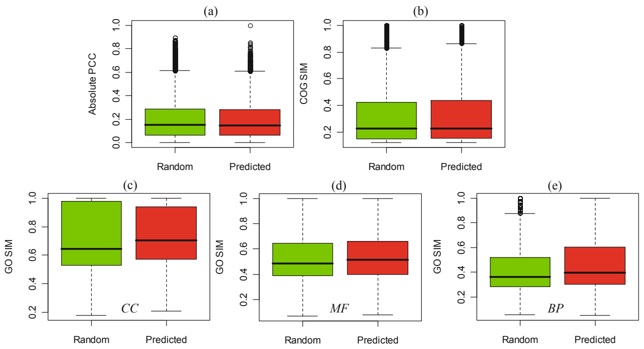

Fig. 2. Boxplot of coexpression (a), coevoluation (b) and cofunction values (c)–(e) of the predicted interactions and that of the same-size random selected protein pairs

4 Conclusion

In conclusion, we established a novel framework to integrate genomic data to infer PPIs in *M. tuberculosis*. We predicted the protein interactions in *M. tuberculosis* by an SVM based classifier by genetic codons. And the documented protein interactions from various species were also mapped to the proteome of *M. tuberculosis* by interolog method. The information from gene expression, evolutionary and functional relationship provided reliable measures of evaluation of our predictions. Our framework can easily be extended to infer the large-scale protein interactions in other species. These predicted interactions provide a valuable reference of interactome for *M. tuberculosis* research. The PPIs are available at: `http://www.aporc.org/doc/wiki/MTBPPI`.

Acknowledgements. This work was supported by Shanghai NSF under Grants No. 11ZR1443100, and by the Chief Scientist Program of SIBS, CAS under Grant No. 2009CSP002. Part of the authors were also supported by NSFC under Grant No. 61072149 and 91029301.

References

1. Singh, A., Mai, D., Kumar, A., et al.: Dissecting Virulence Pathways of Mycobacterium Tuberculosis Through Protein-Protein Association. Proc. Natl. Acad. Sci. USA 103, 11346–11351 (2006)
2. Andres, L.E., Ezkurdia, I., et al.: EcID. A Database for the Inference of Functional Interactions. E. coli. Nucleic Acids Res. 37, D629–D635 (2009)
3. Shen, J., Zhang, J., et al.: Predicting protein-protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA 104, 4337–4341 (2007)
4. Yu, H., Luscombe, N.M., et al.: Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs. Genome Res. 14, 1107–1118 (2004)
5. Najafabadi, H.S., Salavati, R.: Sequence-based Prediction of Protein-Protein Interactions by Means of Codon Usage. Genome Biol. 9, R87 (2008)
6. Kerrien, S., Alam-Faruque, et al.: IntAct–open Source Resource for Molecular Interaction Data. Nucleic Acids Res. 35, D561–D565 (2007)
7. Xenarios, I., Salwinski, L., et al.: DIP, the Database of Interacting Proteins: A Research Tool for Studying Cellular Networks of Protein Interactions. Nucleic Acids Res. 30, 303–305 (2002)
8. Altschul, S.F., Madden, T.L., et al.: Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Res. 25, 3389–3402 (1997)
9. Lord, P.W., Stevens, R.D., et al.: Investigating Semantic Similarity Measures Across the Gene Ontology: The Relationship Between Sequence and Annotation. Bioinformatics, 1275–1283 (2003)
10. Alfarano, C., Andrade, C.E., et al.: The Biomolecular Interaction Network Database and Related Tools. Nucleic Acids Res. 33, D418–D424 (2005)
11. Vastrik, I., D'Eustachio, et al.: Reactome: A Knowledge Base of Biologic Pathways and Processes. Genome Biol. 8, R39 (2007)