GMR
Genetics and Molecular Research
ISSN 16765680

Using the Gene Ontology tool to produce *de novo* protein-protein interaction networks with IS_A relationship

G.S. Oliveira[1] and A.R. Santos[2]

[1]Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais,
Belo Horizonte, MG, Brasil
[2]Faculdade de Computação, Universidade Federal de Uberlândia,
Uberlândia, MG, Brasil

Corresponding author: A.R. Santos
E-mail: santosardr@ufu.br

Genet. Mol. Res. 15 (4): gmr15049273
Received September 13, 2016
Accepted November 11, 2016
Published December 19, 2016
DOI http://dx.doi.org/10.4238/gmr15049273

ABSTRACT. Since the first assembled genomes, gene sequences alone have not been sufficient to understand complex metabolic processes involving several genes, each playing distinct roles. To identify their roles, a network of interactions, wherein each gene is a node, should be created. Edges connecting nodes are evidence of interaction, for instance, of gene products coexisting in the same cellular component. Such interaction networks are called protein-protein interactions (PPIs). After genome assembling, PPI mapping is used to predict the possibility of proteins interacting with other proteins based on literature evidence and several databases, thus enriching genome annotations. Identifying PPIs involves analyzing each possible protein pair for a set of features, for instance, participation in the same biological process and having the same function and status in a cellular component. Here, we investigated using the three categories of the Gene Ontology (GO) database for efficient PPI prediction, because it provides data about the three features exemplified here. For a broader conclusion, we investigated the genomes of ten different human pathogens, looking for commonality regarding the GO hierarchical relationship-denominated IS_A. The plasmids were examined separately from their main genomes. Protein pairs sharing at least one IS_A value were considered as interacting proteins. STRING results certified the probed interactions as sensitivity (score >0.75) and specificity (score <0.25) analysis. The average areas under the receiver operating characteristic curve for all organisms were 0.66 and 0.53 for their genomes and plasmids, respectively. Thus, GO categories alone could not potentially provide reliable PPI prediction. However, using additional features can improve predictions.

Key words: UniprotKB; Gene ontology; String-DB; CpDB;
Protein interaction; Interlog-free

INTRODUCTION

Proteins are the building blocks of life. They play multiple roles in the structural and enzymatic functions of a cell by interacting with each other. Determining the role of proteins is a difficult task. An example of this fact is that almost half of the prokaryote genes have unknown function (Hanson et al., 2009). Despite the lack of information about the role of novel genes, we can glimpse possible interactions between the unknown proteins and the known ones. Such annotation could help us to focus on a particular set of genes when trying to understand biological processes. To achieve this objective, we used certain possible evidence features of the gene, for instance, evolution, conserved neighborhood, expression, a biological process involved, function, and cellular component (Snel et al., 2000). Each feature creates an edge between the nodes representing genes in a network called Protein-Protein Interaction (PPI) network. It should be noted that such a network is just an automated prediction because only by looking at the genome of an organism we cannot guarantee the expression of its predicted genes. Even so, it is a valuable prediction for biologists. PPI networks have been developed in various organisms, allowing the comprehension of numerous biological processes. The assembled set of associations allows the investigation of the structural and functional protein singularities within the pathways of interest. PPI maps have helped to investigate new disease-related proteins (Stelzl et al., 2005).

The search tool for the retrieval of interacting genes/proteins (STRING) database comprises millions of predicted protein associations for more than 2000 organisms. These protein associations are predicted based on the literature and various databases, as well as genomic context (Szklarczyk et al., 2015). The STRING PPI prediction tool considers the Gene Ontology (GO) database (String-DB.org, 2016). GO systematically tags different attributes within the three master domains: biological process, molecular function, and cellular component (Gene Ontology Consortium, 2015). Each GO domain is crucial for predicting the gene functionality, adding novel insights about gene clustering, and better comprehending the complex biological mechanisms (Chen et al., 2007). Even though the STRING prediction tool considers GO, it solely imports the GO protein complexes to infer interaction, not the GO terms by themselves (String-DB.org, 2016).

Since the GO term categories are supposed to contain accurate biological information about different proteins, we hypothesized that by estimating the number of shared-IS_A GO terms among protein pairs, we can predict the positive PPIs. To test this hypothesis, we compared STRING protein association predictions with a GO-based semantic measurement approach in different microorganisms using the area under the receiver operating characteristic curve (AUC) (Sing et al., 2005). Despite performing this investigation in several organisms, we got a similar result confirming the inability of GO-IS_A relationship for creating strong PPI predictions.

MATERIAL AND METHODS

Genomes

The complete genome sequences of the following microorganisms were downloaded from the National Center for Biotechnology Information (NCBI) GenBank database: *Bacillus anthracis* (NC_003997.3), *Clostridium botulinum A2 Kyoto* (NC_012563.1), *Clostridium botulinum F Langeland* (Chromosome NC_009495.1, plasmid NC_009496.1), *Clostridium perfringens* (NC_003366.1), *Clostridium tetani E88* (Chromosome NC_004557.1, Plasmid NC_004565.1), *Corynebacterium diphtheriae* (NC_002935.2), *Escherichia coli ED1a* (NC_011745.1), *Escherichia coli S88* (Chromosome CU928161.2, plasmid CU928146.1) *Mycobacterium tuberculosis* (NC_000962.3), *Streptococcus pneumoniae Taiwan 19F14* (NC_012469).

Pipeline

Given the studied genomes, we linked the GO terms for protein identifiers before inferring PPI probabilities. We opted to link the protein identifiers to their appropriate GO terms using solely the local processing, avoiding the utilization of the interolog method commonly provided by current PPI annotation services, including STRING, which transfers predictions according to the sequence similarity of protein pairs. As shown in Figure 1, the *Corynebacterium pseudotuberculosis* DataBase (CpDB) genome tools (Santos, 2012) were employed to store all data, including genes, GO, and link tables, in the form of BLAST (basic local alignment search tool) results for genes and their respective GO terms. A STRING database dump was filtered, using scripting tools. True positive and true negative interactions for scores above 0.750 and below 0.250, respectively, were considered per studied organism. The filtered set of STRING proteins was analyzed as a golden standard. The annotated protein sequences of the genome samples were exported for performing GO tag association. The data from the UniProtKB database (UniProt Consortium, 2015) were dumped to provide the GO terms for the proteins of the genomes under study. Since the protein identifiers listed in the STRING database dump were different from those presented in the UniProtKB database, a similarity analysis, using the BLASTp program for filtering the higher scores, was performed to access the predicted GO term values. Considering the low chances of two proteins having the same GO term, we used a GO hierarchical relationship, which encompasses the closely related GO terms and the IS_A relationship. The IS_A was employed as the main point to infer protein interaction using the three GO term categories. Besides, the categories obtained from the GO database (molecular function, biological process, and cellular component), the gene neighborhood was considered in the second round of analyses. The neighboring genes were empirical, and were admitted as a positive feature when the distance between them was equal or lower than five genes way in the DNA strands.

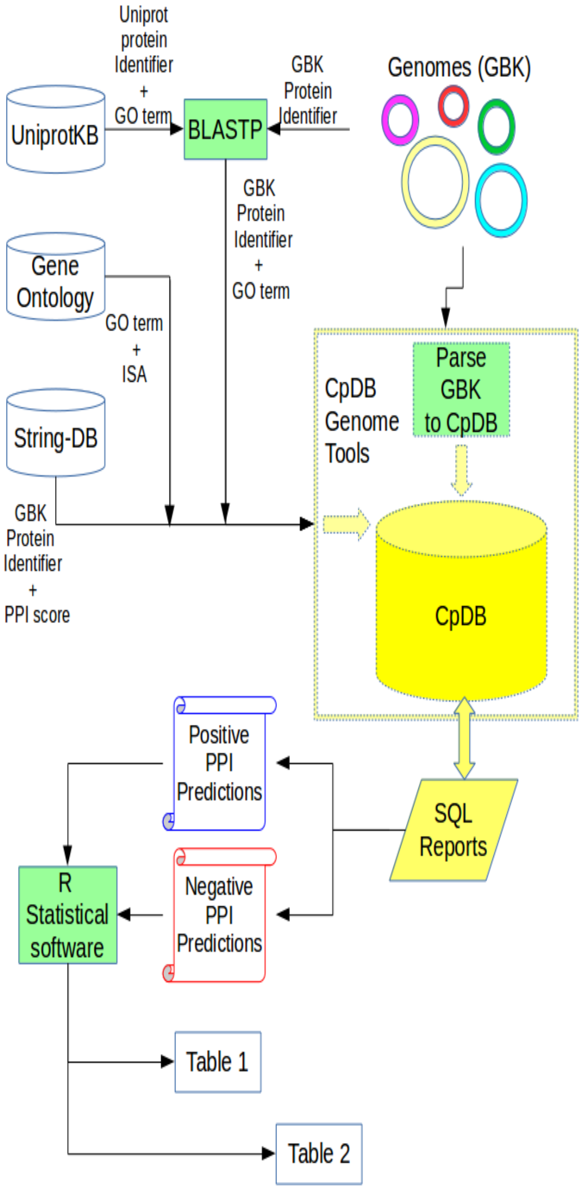

Figure 1. Workflow pipeline and the tools utilized in this study. The green forms represent the compiled programs. The CpDB genome tools comprise a database schema and a parser written in C language. The yellow arrows represent the input and output in the form of structured query language (SQL) commands. A local Postgres server maintains the CpDB schema.

RESULTS AND DISCUSSION

Using STRING-based protein interactions as a golden standard, we attempted to validate the small plasmid and whole genome PPIs using a *de novo* approach for identifying the GO relationship with IS_A commonality. The average AUC of all organisms was 0.62 and 0.53 for genomes and their plasmids, respectively. These values were not significant because the AUC values of 0.50 were random chance, and the obtained values were marginally above the random chances. Due to the low AUC achievements associated with the use of only the GO terms categories, the distance between the neighboring genes led further predictions culminating into AUC averaged values summed up to 0.79 for genomes and 0.80 for plasmids; these values were considered significantly different from the random chance values. Furthermore, the contemplation of the four categories of PPI predictions caused an average increment of 29 and 51% in the AUC of the whole genomes (Table 1) and plasmids (Table 2), respectively. These average increments in the AUC were significant, and demonstrated a notable PPI prediction ability using only those four features. Our result opens the possibility of creating a standalone PPI predictor, based on a few features and benefits of the high-level result, independently of the external servers.

Table 1. Area under the receiver operating characteristic (ROC) curve (AUC) values of bacterial genomes created by a *de novo* PPI prediction comprising the common Gene Ontology (GO)-IS_A relationships of protein-protein pairs.

Microorganism	N	GO	GO + Distance	Improvement (%)
Bacillus anthracis	167,072	0.6688	0.7396	10.58
Clostridium botulinum A2 Kyoto	133,380	0.5122	0.8749	70.82
Clostridium botulinum F Langeland	124,911	0.5603	0.8671	54.75
Clostridium perfringens	94,145	0.6595	0.7564	14.68
Clostridium tetani E88	73,828	0.6619	0.7539	13.9
Corynebacterium diphtheriae	59,717	0.5625	0.7266	29.17
Escherichia coli ED1a	197,455	0.7045	0.7911	12.3
Escherichia coli S88	73,828	0.6598	0.7681	16.41
Mycobacterium tuberculosis	148,893	0.5282	0.8090	53.16
Streptococcus pneumoniae Taiwan 19F14	67,159	0.6693	0.7935	18.57

N = sample size; GO = AUC values for the three GO categories; GO + Distance = AUC values for GO categories plus the neighborhood gene distance; Improvement = percentage increase after packing the distance feature within the GO categories for an extra round of predictions.

Table 2. Area under the receiver operating characteristic (ROC) curve (AUC) values of bacterial plasmids, listed in Table 1 as whole genomes, generated by a *de novo* PPI prediction comprising the common GO-IS_A relationships of protein-protein pairs.

Microorganism	N	GO	GO + Distance	Improvement (%)
Bacillus anthrax p1	80	0.5122	0.8749	70.82
Bacillus anthrax p2	107	0.5603	0.8671	54.75
Clostridium botulinum F Langeland p1	16	0.5625	0.7266	29.17
Clostridium tetani E88 p1	76	0.5282	0.8090	53.16
Escherichia coli S88 p1	389	0.4999	0.7448	48.99

N = sample size; GO = AUC values for the three Gene Ontology (GO) categories; GO + Distance = AUC values for GO categories plus the neighborhood gene distance; Improvement = percentage increase after packing the distance feature within the GO categories for an extra round of predictions.

Furthermore, no significant difference was observed between the average AUC values for genomes and their respective plasmids. However, significant improvements were observed in the average AUC of the specific plasmids of *B. anthracis* (70%) and *C. tetani* (54%). A relatively higher improvement in the average AUC observed for these two plasmids sequences, as compared to their whole genomes, was likely due to their evolutionary function. As a small and adopted DNA sequence for the synthesis of specific products, for instance, antibiotic-resistance and heavy-metal-resistance genes, plasmids can increase the chances for bacterial fitness under toxic conditions. Such evolutionary advantage could also explain a conservative gene neighborhood in these plasmids as compared to their whole genomes (MacLean and San Millan, 2015).

The overall AUC increments provided by the gene neighborhood compensated for the lack of common IS_A terms. Even so, considering the AUC values, we can assume the augmented PPI prediction ability supported by the GO categories plus gene neighboring. We hypothesize that such outcomes are because GO does not tag some products of the known biological pathways as belonging to the same biological process. For instance, consider the thiamine pyrophosphate (TPP) synthesis pathway. TPP is also known as vitamin B1, an essential coenzyme for the catabolism of sugars and amino acids in aerobes. Within the TPP synthesis pathway, the enzyme thiamine biosynthesis oxidoreductase (*thiO*) participates in the production of dehydroglycine from glycine. In the next step, another enzyme, thiazole synthase (*thiG*), is required for the production of thiazole phosphate carboxylate tautomer from dehydroglycine (catalyzed from *thiO*) and 1-deoxyxylulose 5-phosphate (Broderick et al., 2014). Both these enzymes are known to interact with each other in their biological functions and processes, for instance, in the TPP metabolic process of *B. subtilis* (Settembre et al., 2003; Du et al., 2011). However, GO tags *thiO* with five IS_A values and *thiG* with 11 IS_A values, with none of them in commonality (Table 3). Besides, by using the GO enrichment analysis tool, it is possible to generate *thiG* as thiazole synthase and *thiO* as a member of a subfamily not named.

Table 3. A sample of GO terms and their corresponding IS_A relationships for two genes known to interact. In this sample, no common GO-IS_A value was obtained, configuring a pair of non-interacting proteins in accordance with the three GO categories.

DIP0031 IS A *thiO* product	DIP0033 IS A *thiG* product
(GO:0016491 IS A GO:0003824)	(GO:0005737 IS A GO:0044424)
(GO:0050660 IS A GO:0000166)	(GO:0009228 IS A GO:0006772)
(GO:0050660 IS A GO:0043168)	(GO:0009228 IS A GO:0042724)
(GO:0050660 IS A GO:0050662)	(GO:0009229 IS A GO:0009108)
(GO:0055114 IS A GO:0044710)	(GO:0009229 IS A GO:0019438)
	(GO:0009229 IS A GO:0042357)
	(GO:0009229 IS A GO:0044272)
	(GO:0009229 IS A GO:0072528)
	(GO:0009229 IS A GO:0090407)
	(GO:0016783 IS A GO:0016782)
	(GO:0036355 IS A GO:0016830)

Our results do not intend to reduce the significance of GO, but point out its limitations for a *de novo* PPI prediction strategy.

As a limitation of the method used here to predict PPI, it entirely depends on the UniProtKB predictions, mostly automatically made by the Interproscan application. If the UniProtKB database fails to provide the GO terms for the studied proteins, the PPI predictions based on the three GO characteristics sharing IS_A terms will also fail. However, the extensive use of the Interproscan tool by the scientific community makes it a reliable tool. Another possible limitation of this study is the chosen organisms. In our study, we used four distinct classes of organisms, namely, four *Clostridia*, two *Bacilli*, two *Actinobacteria*, and two *Gammaproteobacteria*. We made these choices based on the organisms present in both database dumps, STRING and UniprotKB. It was a practical decision because we could not apply additional efforts for mapping the protein identifiers to the GO terms. Due to this limitation, we selected the organisms causing human diseases. Our study demonstrated the potential of a *de novo* PPI prediction in microorganisms important in the context of public health.

CONCLUSION

A *de novo* PPI prediction employing only three GO categories was proven unable to generate a reliable result according to the golden standard of the STRING database. The addition of a fourth PPI prediction feature, the gene neighborhood, significantly improved the GO PPI predictions, but still with fairly significant AUC values. Our study points out the need for several complementary features to achieve accurate PPI predictions.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by Pró-Reitoria de Pesquisa e Pós-Graduação (PROPP) da Universidade Federal de Uberlândia (UFU), Fundação Amparo à Pesquisa de Minas Gerais (FAPEMIG), and Genetics and Molecular Research (GMR) journal.

REFERENCES

Broderick JB, Duffus BR, Duschene KS and Shepard EM (2014). Radical S-adenosylmethionine enzymes. *Chem. Rev.* 114: 4229-4317. http://dx.doi.org/10.1021/cr4004709

Chen JL, Liu Y, Sam LT, Li J, et al. (2007). Evaluation of high-throughput functional categorization of human disease genes. *BMC Bioinformatics* 8 (Suppl 3): S7. http://dx.doi.org/10.1186/1471-2105-8-S3-S7

Du Q, Wang H and Xie J (2011). Thiamin (vitamin B1) biosynthesis and regulation: a rich source of antimicrobial drug targets? *Int. J. Biol. Sci.* 7: 41-52. http://dx.doi.org/10.7150/ijbs.7.41

Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43: D1049-D1056. http://dx.doi.org/10.1093/nar/gku1179

Hanson AD, Pribat A, Waller JC and de Crécy-Lagard V (2009). 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list - and how to find it. *Biochem. J.* 425: 1-11. http://dx.doi.org/10.1042/BJ20091328

MacLean RC and San Millan A (2015). Microbial evolution: Towards resolving the plasmid paradox. *Curr. Biol.* 25: R764-R767. http://dx.doi.org/10.1016/j.cub.2015.07.006

Santos AR (2012). CpDB. Uberlândia: Sourceforge.net. Available at [https://sourceforge.net/projects/cpdb] Accessed on October 1, 2016.

Settembre EC, Dorrestein PC, Park JH, Augustine AM, et al. (2003). Structural and mechanistic studies on ThiO, a glycine oxidase essential for thiamin biosynthesis in *Bacillus subtilis*. *Biochemistry* 42: 2971-2981. http://dx.doi.org/10.1021/bi026916v

Sing T, Sander O, Beerenwinkel N and Lengauer T (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* 21: 3940-3941. http://dx.doi.org/10.1093/bioinformatics/bti623

Snel B, Lehmann G, Bork P and Huynen MA (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* 28: 3442-3444. http://dx.doi.org/10.1093/nar/28.18.3442

Stelzl U, Worm U, Lalowski M, Haenig C, et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122: 957-968. http://dx.doi.org/10.1016/j.cell.2005.08.029

String-DB.org (2016). Does STRING contain any Gene Ontology information? String Frequent Answer Questions. Available at [http://http://string-db.org/cgi/info.pl]. Accessed September 9, 2016.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43: D447-D452. http://dx.doi.org/10.1093/nar/gku1003

UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43: D204-D212. http://dx.doi.org/10.1093/nar/gku989