bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

.

Application of Global Transcriptome Data in Gene Ontology Classification and Construction of a Gene Ontology Interaction Network

Mario Fruzangohar[1], Esmaeil Ebrahimie[1], David L. Adelson[1,2]*

[1]School of Molecular and Biomedical Science, The University of Adelaide, South Australia 5005, Australia.

[2] Zhendong Australia – China Centre for Molecular Chinese Medicine, The University of Adelaide, South Australia 5005, Australia

mario.fruzangohar@adelaide.edu.au

esmaeil.ebrahimie@adelaide.edu.au

*david.adelson@adelaide.edu.au

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

.

Abstract

Background

Gene Ontology (GO) classification of statistically significant over/under expressed genes is a common method for interpreting transcriptomics data as a first step in functional genomic analysis. In this approach, all significant genes contribute equally to the final GO classification regardless of their actual expression levels. However, the original level of gene expression can significantly affect protein production and consequently GO term enrichment. Furthermore, even genes with low expression levels can participate in the final GO enrichment through cumulative effects.

GO terms have regulatory relationships allowing the construction of a regulatory directed network combined with gene expression levels to study biological mechanisms and select important genes for functional studies.

Results

In this report, we have used gene expression levels in bacteria to determine GO term enrichments. This approach provided the opportunity to enrich GO terms in across the entire transcriptome (instead of a subset of differentially expressed genes) and enabled us to compare transcriptomes across multiple biological conditions. As a case study for whole transcriptome GO analysis, we have shown that during the infection course of different host tissues by *streptococcus pneumonia*, Biological Process and Molecular Functions' GO term protein enrichment proportions changed significantly as opposed to those for Cellular Components. In the second case study, we compared *Salmonella Enteritidis* transcriptomes between low and high pathogenic strains and showed that GO protein enrichment proportions remained unchanged in contrast to a previous case study.

In the second part of this study we show for the first time a dynamically developed enriched interaction network between Biological Process GO terms for any gene samples. This type of network presents regulatory relationships between GO terms and their genes. Furthermore, the network topology highlights the centrally located genes in the network which can be used for network based gene selection. As a case study, GO regulatory networks of *streptococcus pneumonia* and *Salmonella enteritidis* were constructed and studied.

Conclusions

In both *Streptococcus pneumonia* and *Salmonella enteritidis*, the pathways related to GO terms "Environmental Information Processing", "Signal transduction" and "two-component system" were associated with increasing pathogenicity, breaching host barriers and the generation of new strains.

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

This study demonstrates a comprehensive GO enrichment based on whole transcriptome data, along with a novel method for developing a GO regulatory network showing overview of central and marginal GOs that can contribute to efficient gene selection.

Keywords

Biological pathway, Expression level, Gene ontology regulatory network, GO, Central gene, Eigen gene, Transcriptomics, RNA-Seq, PRKM, Directed acyclic graph, DAG, Microarray, System Biology, PHP, PostgreSQL, Cytoscape, *Salmonella Enteritidis*, *streptococcus pneumonia*, pathogen, pathogenesis, Kolmogorov–Smirnov, chi-square

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

.

Background

The functional genomic changes in bacterial pathogens during disease progression or in emerging highly pathogenic strains are poorly understood. Classifying genes into distinct functional groups through Gene Ontology (GO) is a commonly used and powerful tool for understanding the functional genomics and underlying molecular pathways. However, GO protein enrichment is related to the amount and number of proteins described in that GO, and in eukaryotes mRNA levels are often poorly correlated with protein expression. Bacteria are attractive organisms for GO analysis since they have less Post-transcriptional gene silencing compared to animals and plants [1] making gene expression an accurate proxy for protein levels.

Because of the lack of specific resources for GO analysis in bacteria, we recently developed Comparative GO, a PHP based web application for statistical comparative GO and GO-based gene selection in bacteria [2]. Comparative GO has the potential to provide a comprehensive view of bacterial functional genomics by categorizing genes into a limited number of annotated GO groups [2, 3].

Another major advantage of GO analysis is developing quality-based gene selection strategies compared to the common approach of gene selection in bacteria which is solely based on the level of gene expression (quantity based gene selection) [2, 3]. It should be noted that expression level alone cannot be used as a sole index of gene significance because some genes with lower expression levels (such as transcription factors) play a prominent role in bacterial systems biology [2, 3]. An integrative approach, combining quality-based metrics such as GO classification, promoter analysis, and network construction in conjunction with quantity-based gene selection criteria provides a more robust approach for identifying key bacterial genes and describing bacterial systems biology. Such an approach can contribute to the discovery of genes associated with specific function(s) for investigation as novel vaccine candidates or pathways for pharmacological targeting.

Biological process GO terms are analogous to genes because they have regulatory relationships with each other that can be used to construct a directed acyclic network. Compared to common gene networks, GO regulatory networks can identify key functional genomics based interactions in a broader sense. Classifying a large number of genes in a small number of GO classes and visualising the GO networks can significantly decrease the network complexity and, more importantly, offers a new approach for gene selection by considering the genes which contribute to central nodes in GO networks. To our knowledge there is no tool and methodology currently available to dynamically construct GO regulatory networks.

The common approach in transcriptome experiments is that GO analysis is carried out on a short list of genes with statistically significant differential expression (up/down regulated) [4-6]. In this approach, all selected genes contribute equally in the final GO classification regardless of their actual expression levels.

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The major drawback to this approach is that the original levels of gene expression can significantly affect protein production and consequently actual GO term enrichment. In addition, even genes with low or statistically non-significant expression levels can participate in final GO enrichment through cumulative effects.

In this report we show for the first time how gene expression levels in bacteria can be used to determine GO term enrichments. By using gene expression levels as coefficients, we also took into account the impact of non-significantly expressed genes in GO enrichment. This approach provided the opportunity to enrich GO terms in the entire transcriptome genome (instead of samples of a short list of genes) and enabled us to compare GO terms of transcriptomes across multiple biological conditions. In order to achieve this, we enhanced our recently developed web server, Comparative GO[2, 7]. To enable analysis of very large gene sets such as from a whole genome, we implemented cache technology to improve web server performance. We also integrated robust non-parametric chi-square based tests into our web application to enable genome scale GO based comparison of gene expression.

We applied our new methods to two important bacterial pathogens, *streptococcus pneumonia* and *Salmonella Enteritidis* in order to unravel the global, transcriptome based, GO pattern of *streptococcus pneumonia* during infection of host tissues and breaching of tissue barriers as well as the comparison of low and highly pathogenic *Salmonella Enteritidis* strains [8].

In the second part of this study we describe the implementation of GO based gene selection and GO network discovery. We show for the first time a dynamically constructed interaction network between Biological Process GO terms for any given bacterial gene sample. To this end, GO relationships were extracted from Gene Ontology database [9-11], and used to build a directed acyclic graph (DAG). To visualise the final DAG, we used the Cytoscape web browser plug-in [12]. We used our *streptococcus pneumonia* and *Salmonella Enteritidis* data sets as case studies for this method.

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

.

Material and Methods

Incorporation of gene expression levels into GO analysis

The system accepts any type of expression level such as microarray fold-change data and RPKM counts of RNA-Seq data. In all cases, for each gene, one coefficient is estimated based on its expression level within the sample or within the genome. If we want to perform comparative GO analysis on n sample of genes, and the expression level of gene i in list j is e_{ij} and also given that the smallest expression level across n samples is denoted by e_{min}, then the coefficient of gene i in list j (C_{ij}) is estimated as :

$$C_{ij} = e_{ij} / e_{\min} \qquad \text{where } e_{\min} > 0$$

Furthermore, if GO term t in sample r is associated with genes $G_{1r}...G_{mr}$, then the protein enrichment level of GO term t in sample r (PE_{tr}) is estimated as:

$$PE_{tr} = \sum_{i=1}^{m} C_{ir}$$

We have implemented these methods on our server with PHP.

Hypothesis Testing Tool

We implemented a tool to test the hypothesis of a significant difference between 2 genomes/samples GO term distributions. Specifically, we implemented a Chi-Square test for 2 samples and we compared it with the Kolmogorov–Smirnov test using the R-statistical package[13]. These two methods are both non-parametric and are suitable for comparing 2 lists of paired numbers like GO term enrichment values for 2 samples.

GO regulatory Network Construction

Regulatory relationships (up/down regulation) were extracted between Biological Process terms from the Gene Ontology database [9, 10]. We stored these relationships in our internal database [2]. For any given gene sample, our application builds a GO DAG (Directed Acyclic Graph) network, based on regulatory relationships.

In order to infer new relationships from available relationships we expanded initial GO network to include parental nodes; then, new relationships were inferred from relationships

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

between parental GO nodes to the nodes in the network. Figure 1 depicts a simple GO regulatory network, where grey nodes represent the GO terms related to the sample, and the relationships between GO terms are depicted by green arrows. As we can see at the top of the graph, there is a relationship between parental GO terms 2 and 3. Accordingly, we inferred 3 new relationships between nodes 4, 5, 6 and node 7, depicted as green dotted arrows. The final enriched network can describe novel regulatory relationships between GO terms and consequently between their associated genes.

Figure 1:

Web Application Enhancements

Methods and algorithms were implemented in our web application [2] using PHP 5 and PostgreSQL. Because of the additional functionality to analyse the GO distribution of all expressed genes within a genome (global transcriptomics), significant memory and processing resources were required by the Apache web server. To enhance performance and husband system resources we implemented file based caching technology to cache the whole genome GO graphs. When a GO graph was built for the first time, subsequent references to that GO graph, even by other users, was instantaneous. For a better user experience in web application pages where long running tasks were performed, we used Ajax technology to implement task progress bars.

Visualising the GO interaction network

In order to visualize the enriched GO regulatory network, we used the Cytoscape [12] flash player plug-in for web. We initialized and used this component through JavaScript. Cytoscape contains advanced dynamic network customization options such as zooming, network filtering, node re-locating, node and edge re-sizing, and colour scheming. These user-friendly options allow developers and users to dynamically change the look and feel of the network.

Case study data

To demonstrate the biological application of our new methods in global transcriptome GO analysis and GO network analysis, data from 2 previous gene expression experiments were

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

used. *Streptococcus pneumonia* and *Salmonella enteritidis* were selected since they are responsible for high morbidity, mortality, and infection worldwide and have been well studied.

The first data set [3] was two colour microarray data from *Streptococcus pneumonia in vivo* derived RNA samples, where the relative expression of each gene in one niche was calculated in comparison to expression in the previous niche when bacteria moved from nose ⟶ lungs ⟶ blood ⟶ brain. The relative expression of all 2236 genes of *Streptococcus pneumonia* during the course of infection are presented in Additional File 1 (lung versus nose), Additional File 2 (blood versus lung), and Additional File 3 (brain versus blood). Additional files are in MS Excell worksheet format.

The second case study [8] was RNA-Seq global transcriptome data from 6 strains of *Salmonella enteritidis*, where 3 highly pathogenic strains and 3 low pathogenic strains were compared. The average whole genome expression of (4402) genes of the 3 highly pathogenic strains is presented in Additional File 4. While Additional File 5 contains the average expression of the 3 low pathogenic strains. The goal of this analysis was to unravel significantly different GO terms between highly and low pathogenic strains of bacteria using *Salmonella enteritidis* as a model.

For GO network analysis, in case study 1, the 30 highest over expressed genes in *Streptococcus pneumonia* during infection in lung versus, blood versus lung and brain versus blood were used (Additional File 6). Also, in case study 2, 18 genes with the highest fold change in expression levels between highly pathogenic strains versus low pathogenic strains are presented in Additional File 6.

Results

Introduction of gene expression levels into GO analysis

Addition of expression level data with GO term data provided the opportunity of (1) quantifying exact GO enrichments, (2) extending analysis coverage from sample-wide to genome-wide, and (3) developing statistical tests for comparison of GO distributions across transcriptomes. Considering the influences of all expressed genes in functional genomics, even those with low levels of expression, could possibly increase the accuracy of the analysis in prokaryotes.

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

.

GO regulatory network

GO regulatory networks for a sample of genes initially present three types of information: regulatory relationships between GO terms and their associated genes depicted by directed edges of the graph, enrichment levels of GO terms that are proportional to nodes' diameter of graph and finally, the genes associated with each GO term.
Furthermore, network topology revealed GO groups and their genes that had the highest number of interactions with other groups. Specifically, genes located in centre of the network were selected as good candidates for further experiments and gene discovery. In addition, the enrichment levels of GO terms that were proportional to the size of the nodes in the graph were in accordance with the regulatory relationships between GO terms.

Case studies

As case studies, we used publicly available two colour microarray and global transcriptomics data of two important bacterial pathogens, *Streptococcus pneumonia* and *Salmonella enteritidis* respectively. For each bacterium, 2 types of analysis were carried out: transcriptome based GO enrichment and GO network discovery. In *Streptococcus pneumonia*, all expressed genes were subjected to GO analysis in order to characterise functional changes in *Streptococcus pneumonia* during the course of infection. Then, using a selection of significantly up-regulated genes during infection in each tissue, GO networks were constructed to identify the central GO node and the key genes associated with the central GO node. In the *Salmonella enteritidis* case study, we first compared transcriptome GO enrichment levels between highly pathogenic and low pathogenic strains to highlight GO functional groups correlated with pathogenicity. We then constructed the GO network using the genes which were significantly more highly expressed in pathogenic strains

Case Study 1: Changes in the transcriptome GO during *Streptococcus pneumonia* from nose ⟶ lungs ⟶ blood ⟶ brain

After downloading microarray data [3] from the NCBI GEO database for *Streptococcus pneumonia* , we selected data of strain WCH43 after 72 hours infection across 4 different tissues. We estimated the geometric means of the fold-change for each gene in the genome.

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

.

The result was 3 genome-wide lists (Nose vs. Lung, Lung vs. Blood and Blood vs. Brain) each containing 2236 genes along with their mean fold-changes (Additional File 1, 2 and 3). These lists were submitted to the web server.

First, we used the pie chart visualisation to determine GO term proportions (protein enrichment distribution percentage) at different levels of the GO tree. GO term proportions of some GO groups didn't change across multiple tissues. Hence, the GO term proportions of 3 genome-wide lists were mutually compared by Kolmogorov–Smirnov test and the calculated p-values are presented in Table 1.

Table 1:

Table 1 suggests that Cellular Components GO enrichment proportions did not change during the course of infection at all. Interestingly, when bacteria moved from blood to its final destination (brain), the overall proportions of Biological Process GO terms did not change.

We then produced a tabular report of the last level (most detailed) of the GO tree. From a large list of GO terms, this report highlighted GO terms that were consistently up/down regulated. Surprisingly, in this study only identified a few such GO terms (Figure 2). GO terms with upward or downward arrows had consistent up/down expression patterns. The continuously up regulated GOs were "barrier septum assembly" and tryptophan synthase activity which are involved in propagation of *Streptococcus pneumonia*. This result confirmed a known, experimentally verified mechanism in this organism [3]. The list of genes in each GO is also presented to assist with GO based gene selection. GOs such as "histidine biosynthesis process" and "amidase activity" were down regulated. This report also highlights GO terms with more than 4 fold average fold-change.

Figure 2:

The GO regulatory network during *Streptococcus pneumonia* infection from nose ⟶ lungs ⟶ blood ⟶ brain

The GO network during movement of *Streptococcus pneumonia* from nose to lung is presented in Figure 3A. Upon inspection, regulation of transcription (Gene Ontology ID: 6355) is a central node in the network. SP_0798 is the only component of this GO network. Interestingly, the GO group (regulation of transcription) governed by SP_0798 plays a key

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

role in breaching the brain-blood barrier and infection of brain tissue. We previously demonstrated that the SP_0798 transcription factor positively regulates the Sp-0927 transcription factor and activates a sub network through interaction with proteins such as SP_0797, SP_0084, SP_2083, SP_1226, and SP_0799 [3]. The SP_0798 sub network is one of the key sub networks conferring high virulence to *Streptococcus pneumonia* [3].

When comparing lung-nose niche expression patterns, the SP_0798 governed GO has interactions with \ GOs such as: "phosphorylation", "fatty acid biosynthesis process", "establishment of competence for transformation" and "oxidation-reduction process". The "establishment of competence for transformation" GO (SP_0798 gene) can play a significant role in the translocation of *Streptococcus pneumonia* from nose to lung.

Figure 3C showed that the SP_0798 governed GO (Gene Ontology ID: 6355) had a considerable number of regulatory effects in the brain-blood comparison. The brain is the final destination of *Streptococcus pneumonia* WCH43 where it causes meningitis. SP_0798 activated different GO groups such as "metabolic process", "establishment of competence for transformation", "phosphorylation" and "antibiotic transport" while reaching and infecting the brain. Activation of "antibiotic transport process" helps *Streptococcus pneumonia* resist antibiotics.

It was previously [3] known that in meningitis-inducing strains of *Streptococcus pneumonia* such as WCH43, relative global gene expression significantly decreased in blood compared to the previous niche (lung) or the subsequent niche (brain). Interestingly, the GO network shown in Figure 3B helps illustrate the underlying mechanism of this global down regulation and shows that Gene Ontology ID 45892 ("negative regulation of transcription, DNA-dependent") governed by SP_1713 transcriptional repressor NrdR is central to this relative decrease in expression. Gene Ontology ID 45892 has interactions with "CTP/GTP biosynthesis process", "barrier septum assembly" (involved in propagation), "cytokinesis binary fission", and "tryptophan biosynthesis process" (Figure 3B). The SP_1664 protein is involved in barrier septum assembly. SP_1813, SP_1814 and SP_1815 proteins participate in tryptophan biosynthesis process.

Discovery of the Gene Ontology ID 45892 ("negative regulation of transcription, DNA-dependent") governed by SP_1713 and its considerable influence in suppression of genes opens a new avenue for the treatment of blood stream-based diseases such as Bacteremia and Sepsis.

Figure 3:

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

.

Case Study 2: Comparison of whole transcriptome based GO enrichment between low and highly pathogenic *Salmonella enteritidis*

We collected RNA-Seq data for 6 strains of low and high pathogenic *Salmonella enteritidis* [8] including 3 low pathogenic strains and 3 highly pathogenic ones. We averaged the RPKM counts for each gene of the 3 low pathogenic strains and created a single list of genome expression levels. We did the same for the 3 highly pathogenic strains (Additional File 4 and 5). After submission of both gene lists (4402 genes for each one) to the web server, we used the pie chart to visualise the GO term proportions and navigate the GO term tree. The comparison revealed very similar GO proportions at nearly all levels of the GO tree. This encouraged us to perform hypothesis tests to compare the GO enrichment proportions between low and highly pathogenic strains. Table 2 shows the result of this comparison for Biological Process, Molecular Function, and Cellular Components.

Table 2:

Based on a 0.05 level of significance for our tests, Table 2 indicates that there is probably no significant difference in GO protein enrichment proportions between low and highly pathogenic strains of *Salmonella enteritidis* bacteria. This is a significant finding that demonstrates that the change from low pathogenic strain to highly pathogenic strain is not associated with a global shift in GO terms. However, as seen below, a shift in a subset of GO terms can be associated with higher pathogenicity.

GO regulatory network changes between high and low pathogenic strains of *Salmonella enteritidis*

A list of the 18 most differentially expressed genes were submitted to the Web server (Additional File 6), including fljB, SEN1084, motA, flgK, cheA, invF, invA, invG, , fliD, prgH, osmY, , ipB, sipC, yeaG, sipA, dps, yjbJ, and bfr. The resulting GO network is presented in Figure 4.

Interestingly, the GO term “signal transduction by phosphorylation” (Gene Ontology ID: 23014) is central in the overrepresented GO expression network of highly pathogenic

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Salmonella enteritidis strains. The protein kinase encoded by cheA is the sole component of "signal transduction by phosphorylation process". This shows that higher pathogenicity in *Salmonella enteritidis* appears to be associated with increased signal transduction and phosphorylation. We speculate that up regulating GO "Signal transduction by phosphorylation" may allow *Salmonella enteritidis* to more rapidly sense environmental changes and activate more genes through stronger phosphorylation activity. "Response to stress", "iron ion transport" (bfr gene), "pathogenesis", "transcription DNA dependent", "protein phosphorylation" (yeaG gene) and "chemotaxis" are the other GO terms which are differentially expressed in highly pathogenic strain.

Figure 4:

Commonality between GO Regulatory Networks of Case Studies

Selection of *Streptococcus pneumonia* during the course of infection in nose, blood, and brain of host allowed us to apply whole genome based GO enrichment and GO in study of tissue-based pathogenesis and breaking host barriers by pathogen. In addition, comparative study of GO enrichment and GO network between highly pathogenic and low pathogenic strains of *Salmonella* provided to investigate mechanisms involved in generation of highly pathogenic strains using GO concept.

Go network analysis in *Streptococcus pneumonia* and *Salmonella enteritidis* resulted in detection of new biological results and genes that were not reported in original works. Furthermore, central roles of GO classes of "regulation of transcription" and "signal transduction by phosphorylation" governed by SP_0798 and cheA in induction of pathogenesis were unravelled. Phosphorylation, performed by kinases, is one of the main pathways of rapid signal response and gene activation. Interestingly, even in plants, protein kinases are the central compartment of inducing high stress resistance and evolution [14].

cheA (chemotaxis protein CheA) is a sensor histidine kinase and a member of two-component system. cheA is majorly involves in "Environmental Information Processing" and "Signal transduction" (KEGG database [15]). According to Pfam database [16], cheA contains the following domains: PF01584 (CheW-like domain), PF01627 (Hpt domain), PF02518 (Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase), PF02895 (Signal

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

transducing histidine kinase, homodimeric domain), PF09078 (CheY binding), and PF13589 (Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase).

SP_0798 is a DNA-binding response regulator CiaR and a member of two-component system. According to Pfam database [16], SP_0978 contains PF0072 (response regulator receiver domain) and PF00486 (Transcriptional regulatory protein, C terminal). Similar to cheA, SP_0798 is also involved in "Environmental Information Processing", "Signal transduction" and two-component system (KEGG database [15]). It can be concluded that SP_0798 and cheA are ortholog to each other.

Based on the above discussion and the observed similar observed mechanism between *Streptococcus pneumonia* and *Salmonella enteritidis*, it can be suggested that "Environmental Information Processing" which carries ON by "Signal transduction" and two-component system pathways are methods of choice by bacterial pathogens in increasing pathogenicity, host barrier breaking and generation of new strains. In fact, successful pathogens such as *Streptococcus pneumonia* and *Salmonella enteritidis* are developing expert systems to recognise faster external environment and also react more promptly by a more efficient signal transduction system. Two-component system is a head-tail pathway which one member sits outside the cell and other member inside the cell and informs the bacteria about environmental signals/changes. Rapid recognition of environmental alterations such as antibiotic stress and nutrient change allows bacteria to act more rapidly and increase the chance of surviving. Two-component system has a confirmed role in bacterial virulence [17, 18].

On the other hand, SP_1713 is the major player of negative regulation in blood infection of *Streptococcus pneumonia*. The fact that SP_1713 has the ability to regulate a large number of other gene ontology terms and dramatically decreases the global transcriptome expression levels in blood, offers a new possibility for treatment of blood-based infections such as Bacteremia and Sepsis. This example shows how GO network construction can be employed for discovery of key GO groups and GO based gene selection.

Discussion

GO analysis provides a new avenue for a deeper understanding of gene expression and function, which can be exploited in the context of quality-based gene selection strategies [2,

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

3]. While other GO web servers [6, 19] support gene annotation in model eukaryotes via user submitted gene lists that must match the single source of annotation used by the server, our web server supports all sequenced prokaryotes and viruses and automatically recognizes gene names from all annotation sources.

In contrast to other web servers, our web server provides interactive visual navigation of the hierarchical tree structure of GO groups weighted according to gene expression values at all levels. Furthermore, our server provides dynamic visual reports (using AJAX technology) such as pie charts (to visualize GO group proportions) and bar charts (to compare GO term enrichments versus reference genome based on hyper-geometric distribution), whereas other web servers present this information in text format or rely on visualization capacity provided by other websites [20].

The most significant analytical advantage provided by our web server is the ability to compare GO terms across multiple gene samples (or whole genomes) from multiple biological conditions. At present other web servers [6, 19] can only compare one sample against a reference genome. Comparative GO analysis is particularly important as a means to identify the underlying biological pathways recruited under different biological conditions. This is an essential method if one wishes to identify important genes for perturbation experiments.

Unlike other GO web servers that compare one GO term compared to a reference genome at a time (using the Fisher Exact test), our web server can compare all the GO term enrichments from two or more samples (or whole genomes) simultaneously by using robust non-parametric statistical tests. This enables detection any shift in GO distribution as a function of experimental conditions.

Finally, our comparative table report takes into account protein enrichment to detect GO terms with special enrichment patterns or with specific enrichment fold-change across multiple samples. This helps identify key GO terms and their associated genes because their expression prevalence. At present, this is a unique analytical approach that is not found elsewhere.

Global transcriptome based GO analysis was achieved by integrating gene expression levels with GO classifications. This allowed us to compare GO enrichment that better reflected the biological reality of the experiments across multiple samples by taking into account the abundance of gene products. This type of comparison was not previously possible, most likely because the prevalence of eukaryotic GO databases and web servers [6] would not have benefited from such an analysis. Current GO web applications are mostly developed in

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

eukaryotic genomes [4-6] where protein abundance levels are poorly correlated with gene expression levels, making the need for transcript abundance weighting less relevant.

In this report we have presented a method to build GO regulatory networks using public GeneOntology data [9]. GO regulatory networks from differentially expressed genes can reveal underlying biological pathways [21]. In particular the topology of such networks can highlight highly connected/central GO terms and their associated genes, supporting the discovery of candidate genes.

Furthermore, by looking at networks from different bacterial species we can elucidate common biological pathways. Even though we have only implemented GO regulatory networks for bacteria, this type of network could be very effective for eukaryotes as well, particularly for proteomics data. To our knowledge, no current GO web server provides this capability.

We have also demonstrated how to combine a GO regulatory network with gene expression data. The resultant network can be used to study regulatory effects of genes and GOs on each other. For example, by comparing and overlapping multiple GO regulatory networks for the same genes across multiple biological conditions, we can detect areas of the network that confirm or contradict expected regulatory relationships. This can be used as a mean to support or question the validity of original transcriptomic data or indicate the existence of any unknown environmental effects in the experiment. Moreover, by replacing the GO regulatory network's nodes with their associated genes one can generate a GO-based gene regulatory network (GRN).

Finally, combining GO-based gene regulatory networks with other types of gene regulatory networks [21] (those that are reverse engineered from transcriptome data) such as co-expression networks [22, 23] can lead to the discovery of unknown biological entities or biological mechanisms, particularly where such results contradict one another.

Together, the global transcriptomics based GO enrichment and GO regulatory network, developed in the present investigation and implemented in Comparative GO Web application [2, 7] can significantly increase the knowledge of bacterial regulatory mechanisms of pathogenesis as well as functional genomics arrangements which result in emerging new highly pathogenic strains.

Conclusion

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

.

We applied whole transcriptome data and gene expression levels to GO classification analysis leading to new meaningful biological reports. We have also developed a method to dynamically construct GO regulatory networks for any given sample. Finally, we have demonstrated the efficiency of our developed methods and tools through case studies on two types of bacteria. The results of these analyses either identified new candidate genes and GO terms that were not reported in the original work or confirmed the functionality of known genes.

Availability of supporting data

The data sets supporting the results of this article are included within the article and its additional files

List of abbreviations

GO: Gene Ontology; DAG: Directed acyclic Graph; GRN: Gene Regulatory Network; RPKM: Reads Per Kilo Base Per Million;

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceived and designed the methodology and experiments: MF, EE, DLA

Performed the experiments: MF, EE

Data Analysis: MF, EE, DLA

Wrote the paper: MF, EE, DLA

All authors read and approved the final manuscript.

Acknowledgments

- We would like to greatly thank Dr. Abiodun Ogunniyi, Dr. Layla Mahdi and Prof. James Paton from the Research Centre for Infectious Diseases of The University of Adelaide for their comments and help. We would also like thank Dr. Dan Kortschak for his helpful comments.

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Cogoni C, Macino G: **Post-transcriptional gene silencing across kingdoms**. *Current opinion in genetics & development* 2000, **10**(6):638-643.
2. Fruzangohar M, Ebrahimie E, Ogunniyi AD, Mahdi LK, Paton JC, Adelson DL: **Comparative GO: A Web Application for Comparative Gene Ontology and Gene Ontology-Based Gene Selection in Bacteria**. *PloS one* 2013, **8**(3):e58759.
3. Mahdi LK, Ebrahimie E, Adelson DL, Paton JC, Ogunniyi AD: **A transcription factor contributes to pathogenesis and virulence in Streptococcus pneumoniae**. *PloS one* 2013, **8**(8):e70862.
4. Conesa A, Götz S: **Blast2GO: A comprehensive suite for functional analysis in plant genomics**. *International journal of plant genomics* 2008, **2008**.
5. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.
6. Da Wei Huang BTS, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2008, **4**(1):44-57.
7. **Comparative GO: A Web Application for Comparative Gene Ontology Analysis** [genomes.ersa.edu.au/BacteriaGO/]
8. Shah DH: **RNA-Seq reveals differences in the global transcriptome between high-and low-pathogenic Salmonella Enteritidis strains**. *Applied and environmental microbiology* 2013:AEM. 02740-02713.
9. **Gene Ontology Database** [http://www.geneontology.org/]
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: **Gene Ontology: tool for the unification of biology**. *Nature genetics* 2000, **25**(1):25-29.
11. Consortium GO: **The gene ontology: enhancements for 2011**. *Nucleic acids research* 2012, **40**(D1):D559-D564.
12. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang P-L, Lotia S, Pico AR, Bader GD, Ideker T: **A travel guide to Cytoscape plugins**. *Nature methods* 2012, **9**(11):1069-1076.
13. Team RC: **R: A language and environment for statistical computing**. *R foundation for Statistical Computing* 2005.
14. Alimohammadi A, Shiran B, Martínez-Gómez P, Ebrahimie E: **Identification of water-deficit resistance genes in wild almond< i> Prunus scoparia</i> using cDNA-AFLP**. *Scientia Horticulturae* 2013, **159**:19-28.
15. **Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.jp/kegg/]
16. **Pfam Protein Database** [http://pfam.sanger.ac.uk/]
17. Miller SI, Kukral AM, Mekalanos JJ: **A two-component regulatory system (phoP phoQ) controls Salmonella typhimurium virulence**. *Proceedings of the National Academy of Sciences* 1989, **86**(13):5054-5058.
18. Stibftz S, Aaronson W, Monackt D, Falkowt S: **Phase variation in Bordetella pertussis by frameshift mutation in a gene for a novel two-component system**. 1989.
19. Al-Shahrour F, Díaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes**. *Bioinformatics* 2004, **20**(4):578-580.
20. **The European Bioinformatics Institute** [http://www.ebi.ac.uk/]

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

21. Zinman GE, Zhong S, Bar-Joseph Z: **Biological interaction networks are conserved at the module level**. *BMC systems biology* 2011, **5**(1):134.
22. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC bioinformatics* 2008, **9**(1):559.
23. Liu LZ, Wu FX, Zhang WJ: **Reverse engineering of gene regulatory networks from biological data**. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2012, **2**(5):365-385.

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

.

Figures and Legends

Figure 1: Schematic presentation of a simple GO regulatory network. Grey nodes represent GO terms related to the sample and the relationships between GO terms are depicted as green arrows. Parental GO nodes 2 and 3 have a relationship that can be extended to children GO nodes 4, 5, 6 and node 7, depicted as green dotted arrows.

Figure 2: Amended "Table report" which lists consistently up and down regulated GO terms and also GO terms with more than 4 times change in protein enrichment.

Figure 3: GO regulatory network constructed based on differentially expressed *Streptococcus pneumonia* genes in (A) Lung versus Nose (B) Blood versus Lung (C) Brain versus Blood.

Figure 4: GO regulatory network based on 18 genes with significant differential expression levels in highly pathogenic versus low pathogenic *Salmonella enteritidis* strains.

Tables and Captions

Table 1: Comparison of genome-wide GO enrichment levels by Kolmogorov–Smirnov test during the infection course of *Streptococcus pneumonia* from nose ⟶ lungs ⟶ blood ⟶ brain

	Biological Process	**Molecular Function**	**Cellular Components**
(Lung vs. Nose ~ Blood vs. Lung)	P=0.01	P=0.01	Not significant
(Blood vs. Lung ~ Brain vs. Blood)	Not Significant	P=0.01	Not significant

Table 2: Comparison of genome wide GO enrichment levels of low pathogenic strains of *Salmonella enteritidis* versus high pathogenic strains by Kolmogorov–Smirnov test

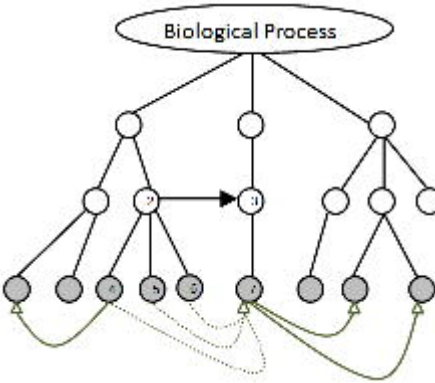
	Biological Process	***Molecular Function***	***Cellular Components***
Low Pathogenic strains Vs. High Pathogenic strains	P value =0.86	P value = 0.34	0.7590978

bioRxiv preprint doi: https://doi.org/10.1101/004911; this version posted May 8, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

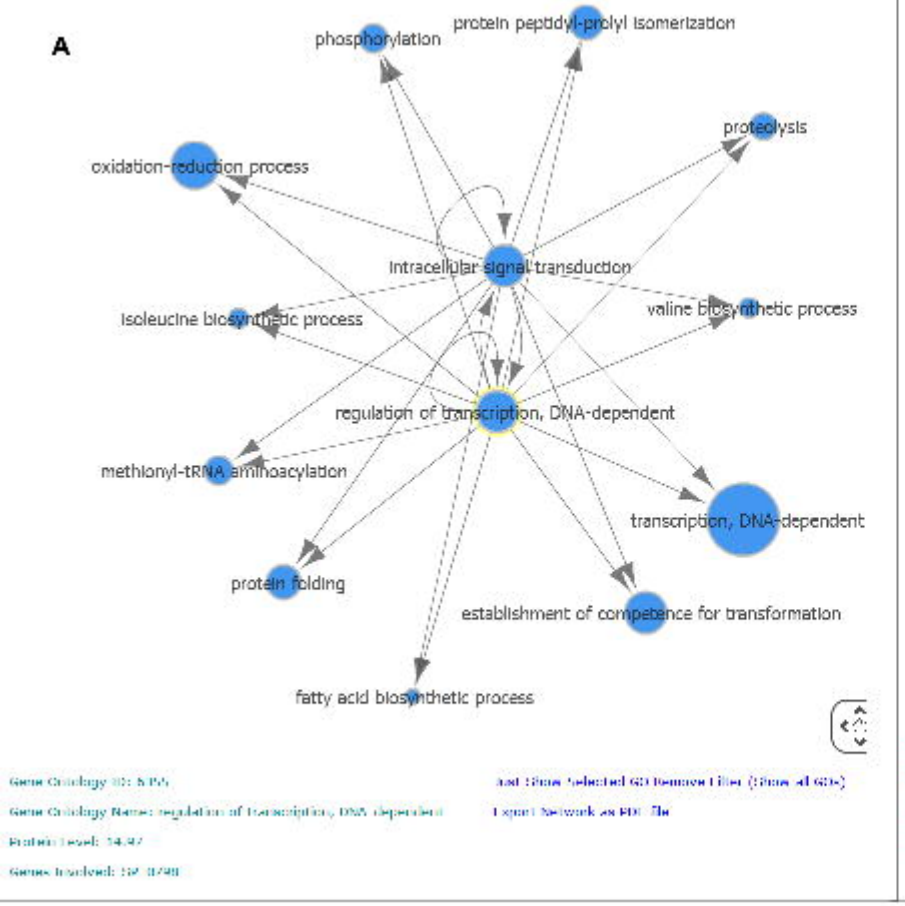
.

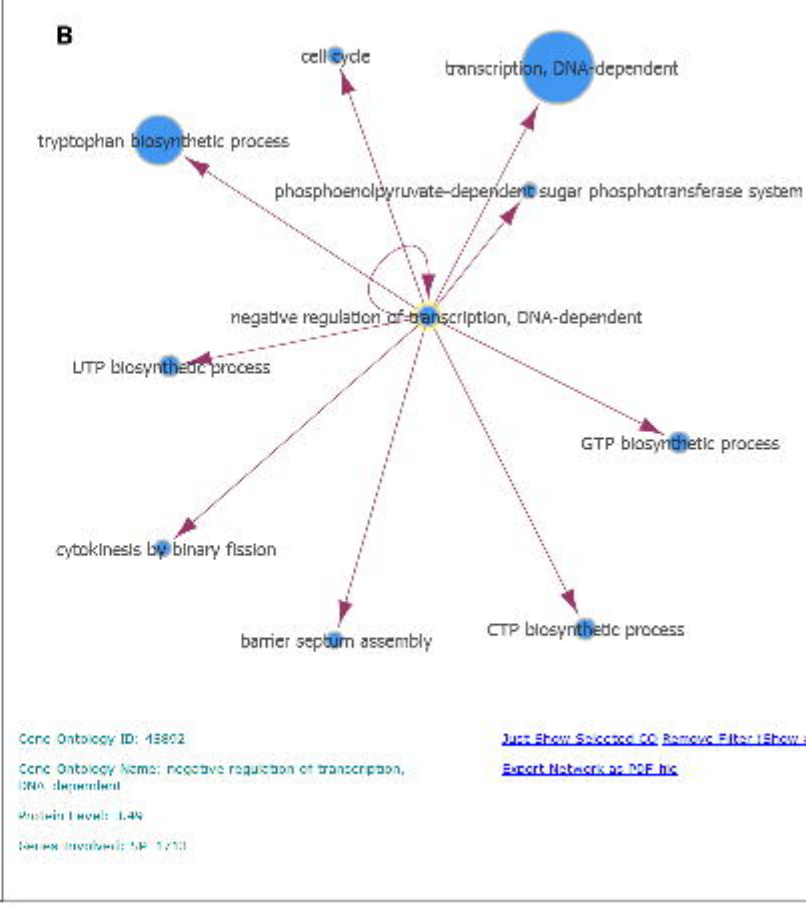
Additional Files

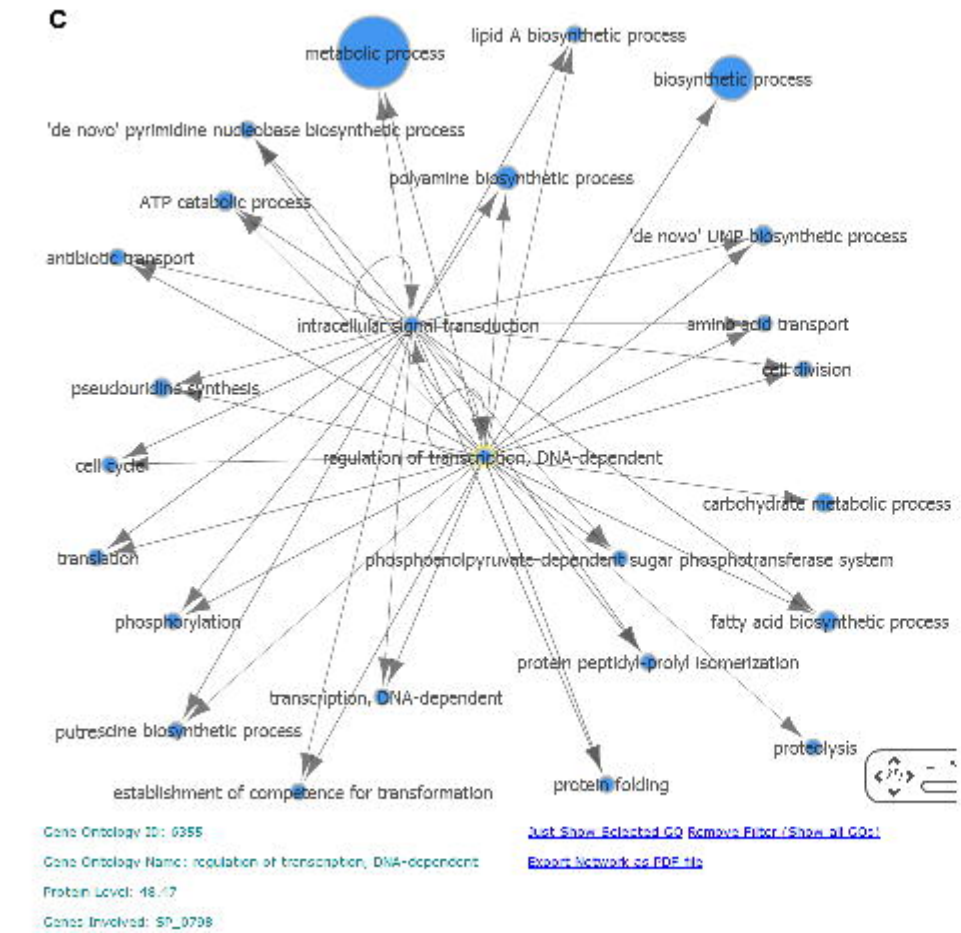
File Name	File Format	Title of Data	Description of Data
Additional_File1.xlsx	xlsx	transcriptome	*streptococcus pneumonia* Whole transcriptome Relative Expression Levels Lung vs. Nose (2 Colour Microarray)
Additional_File2.xlsx	xlsx	transcriptome	*streptococcus pneumonia* Whole transcriptome Relative Expression Levels Blood vs. Lung (2 Colour Microarray)
Additional_File3.xlsx	xlsx	transcriptome	*streptococcus pneumonia* Whole transcriptome Relative Expression Brain vs. Blood (2 Colour Microarray)
Additional_File4.xlsx	xlsx	transcriptome	*Salmonella enteritidis* global transcriptome high pathogenic (average of 3 strains) RPKM Counts
Additional_File5.xlsx	xlsx	transcriptome	*Salmonella enteritidis* global transcriptome low pathogenic (average of 3 strains) RPKM Counts
Additional_File6.xlsx	xlsx	samples	List of Differentially Expressed Genes in Case Study 1 and Study 2 for GO Network Analysis

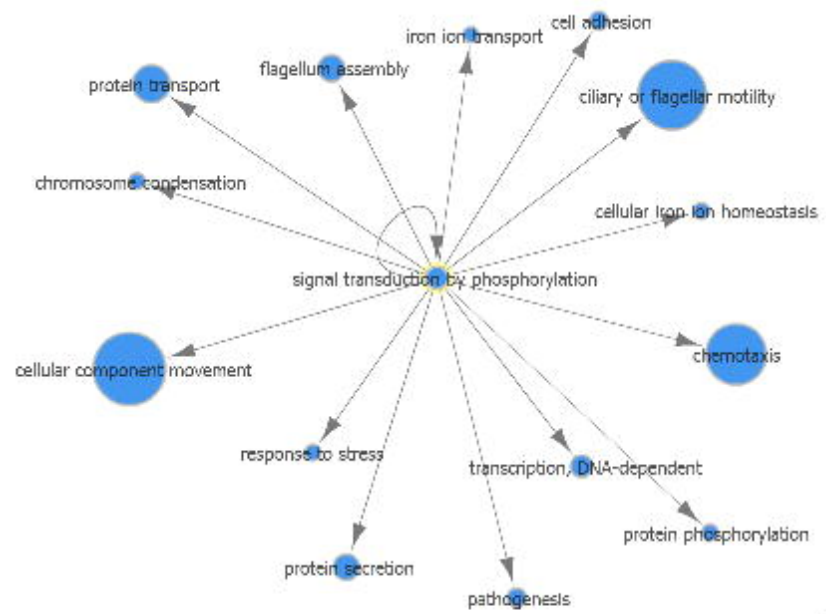

Biological Process

[illegible] binding	2.67	4.44	11.28		2.573	SP_0410,SP_0879,SP_0971,SP_0110,SP_0574,SP_0770,SP_0861,SP_0221,SP_0226,SP_0561,SP_0768,SP_1059,SP_0250,SP_0271,SP_0274,SP_0372,SP_1373,SP_0412,	SP_0410,SP_0879,SP_0971,SP_0110,SP_0574,SP_0770,SP_0861,SP_0221,SP_0226,SP_0561,SP_0768,SP_1059,SP_0250,SP_0271,SP_0274,SP_0372,SP_1373,SP_0412,
[illegible]ne biosynthetic process	1.2	0.91	0.67	↓	0.730	SP_0035,	SP_0035,
[illegible]-dimethyl-8-ribityllumazine synthase activity	1.69	0.1	17.05		2.273	SP_0175,	SP_0175,
[illegible]ase complex assembly	0.90	0.17	11.21	↑	1.374	SP_1361,SP_1362,SP_0367,SP_1362,SP_1733,	SP_1361,SP_1362,SP_0367,SP_1362,SP_1733,
[illegible]ase activity	1.12	0.67	0.77	↓	0.773	SP_0967,	SP_0967,
[illegible]dihydroorotase activity	4.70	5.2	0.34		4.324✓	SP_0761,SP_0961,	SP_0761,SP_0961,
[illegible]isopentenyl-diphosphate delta-isomerase activity	1.43	0.17	22.24		4.011✓	SP_0701,	SP_0701,
[illegible]methylene-tetrahydrofolate cyclohydrolase activity	1.2	0.91	0.65	↓	0.736	SP_0525,	SP_0525,
[illegible]methionine adenosyltransferase activity	1.11	0.17	15.03		4.132✓	SP_0762,	SP_0762,
[illegible]methionyl-tRNA formyltransferase activity	1.41	0.25	7.35		1.440	SP_1737,	SP_1737,
[illegible]methylenetetrahydrofolate dehydrogenase (NADP+) activity	1.3	0.91	0.75	↓	0.776	SP_0727,	SP_0727,
[illegible] bond-forming dihydroxybutanone kinase activity, [illegible] reduction of flavin or flavin as acceptor	3.67	1.81	1.01	↓	0.552	SP_1170,SP_1180,	SP_1170,SP_1180,
[illegible]ase complex activity	2.17	0.12	47.09		4.670✓	SP_0767,	SP_0767,
[illegible]thymidylate synthase activity	1.64	0.11	25.87		4.019✓	SP_0560,	SP_0560,
[illegible]biotin synthase activity	1.29	1.05	4.29	↑	1.846	SP_1611,SP_1612,	SP_1611,SP_1612,
[illegible]-succinyl-coenzyme A synthetase complex	0.67	1.71	1.01	↓	0.703	SP_1179,SP_1180,	SP_1179,SP_1180,
[illegible]alcohol metabolic process	1.12	1.13	1.05	↓	0.953	SP_2026,	SP_2026,
[illegible]purine ribonucleotide biosynthetic process	1.2	0.91	0.67	↓	0.730	SP_0035,	SP_0035,
[illegible]-NAD biosynthetic process	1.64	0.11	25.87		4.019✓	SP_0560,	SP_0560,
[illegible]-acetylglucosamine biosynthetic process	1.11	0.13	18.56		4.193✓	SP_0712,	SP_0712,
[illegible]polyamine biosynthetic process	5.74	0.05	140.49		5.092✓	SP_0437,	SP_0437,
[illegible]succinate metabolic process	2.17	0.12	47.09		4.670✓	SP_0765,	SP_0765,
[illegible]amino acid transport	2.63	0.07	52.94		4.724✓	SP_0740,	SP_0740,
[illegible]-succinyltransferase (succinylating) activity	1.12	1.13	1.05	↓	0.953	SP_2026,	SP_2026,
[illegible]phosphoenolpyruvate protein phosphotransferase activity	1.5	0.12	24.73		4.26✓	SP_1175,	SP_1175,
[illegible]-type D-Ala D-Ala carboxypeptidase activity	1.13	0.15	20.46		4.257✓	SP_0572,	SP_0572,
[illegible]-methionine-methionine aminotransferase activity	0.81	0.14	14.94		4.117✓	SP_1150,	SP_1150,
[illegible]-sugar bond-forming dihydroxybutanone metabolic process	1.1	0.86	0.95	↓	0.872	SP_1150,	SP_1150,
[illegible]-deoxyribonucleotide biosynthetic process	1.1	0.85	0.96	↓	0.872	SP_1180,	SP_1180,
[illegible]-pteridine biosynthetic process	4.26	0.07	65.95		4.062✓	SP_0121,	SP_0121,
[illegible]carbon dioxide	1.14	1.17	1.03	↓	0.903	SP_2000,	SP_2000,
[illegible]-formyltetrahydrofolate cyclo-ligase activity	0.22	0.15	17.35		4.346✓	SP_2095,	SP_2095,
[illegible]-metabolic process	1.51	0.13	25.00		4.097✓	SP_1274,	SP_1274,
[illegible]carbonate transport	2.64	0.06	72.49		5.24✓	SP_1955,	SP_1955,
[illegible]cellular nucleobase metabolic process	1.22	0.66	0.67	↓	0.773	SP_0165,	SP_0165,


A
phosphorylation
protein peptidyl-prolyl isomerization
proteolysis
oxidation-reduction process
intracellular signal transduction
isoleucine biosynthetic process
valine biosynthetic process
regulation of transcription, DNA-dependent
methionyl-tRNA aminoacylation
transcription, DNA-dependent
protein folding
establishment of competence for transformation
fatty acid biosynthetic process


B
cell cycle
transcription, DNA-dependent
tryptophan biosynthetic process
phosphoenolpyruvate-dependent sugar phosphotransferase system
negative regulation of transcription, DNA-dependent
UTP biosynthetic process
GTP biosynthetic process
cytokinesis by binary fission
barrier septum assembly
CTP biosynthetic process


C
metabolic process
lipid A biosynthetic process
biosynthetic process
'de novo' pyrimidine nucleobase biosynthetic process
polyamine biosynthetic process
ATP catabolic process
'de novo' UMP biosynthetic process
antibiotic transport
intracellular signal transduction
amino acid transport
cell division
pseudouridine synthesis
cell cycle
regulation of transcription, DNA-dependent
carbohydrate metabolic process
translation
phosphoenolpyruvate-dependent sugar phosphotransferase system
phosphorylation
fatty acid biosynthetic process
protein peptidyl-prolyl isomerization
transcription, DNA-dependent
putrescine biosynthetic process
proteolysis
establishment of competence for transformation
protein folding
Gene Ontology ID: 6355
Gene Ontology Name: regulation of transcription, DNA-dependent
Protein Level: 48.47
Genes Involved: SP_0798
Just Show Selected GO Remove Filter (Show all GOs)
Export Network as PDF file



Gene Ontology ID: 23014

Gene Ontology Name: signal transduction by phosphorylation

Protein Level: 100.23

Genes Involved: chsA

Just Show Selected GO Remove Filter | Show all GO

Export Network as PDF file