BMC Bioinformatics

SOFTWARE Open Access

GOAL: A software tool for assessing biological significance of genes groups

Alain B Tchagang*, Alexander Gawronski, Hugo Bérubé, Sieu Phan, Fazel Famili and Youlian Pan*

Abstract

Background: Modern high throughput experimental techniques such as DNA microarrays often result in large lists of genes. Computational biology tools such as clustering are then used to group together genes based on their similarity in expression profiles. Genes in each group are probably functionally related. The functional relevance among the genes in each group is usually characterized by utilizing available biological knowledge in public databases such as Gene Ontology (GO), KEGG pathways, association between a transcription factor (TF) and its target genes, and/or gene networks.

Results: We developed *GOAL*: **G**ene **O**ntology **A**na**L**yzer, a software tool specifically designed for the functional evaluation of gene groups. *GOAL* implements and supports efficient and statistically rigorous functional interpretations of gene groups through its integration with available GO, TF-gene association data, and association with KEGG pathways. In order to facilitate more specific functional characterization of a gene group, we implement three GO-tree search strategies rather than one as in most existing GO analysis tools. Furthermore, *GOAL* offers flexibility in deployment. It can be used as a standalone tool, a plug-in to other computational biology tools, or a web server application.

Conclusion: We developed a functional evaluation software tool, *GOAL*, to perform functional characterization of a gene group. *GOAL* offers three GO-tree search strategies and combines its strength in function integration, portability and visualization, and its flexibility in deployment. Furthermore, *GOAL* can be used to evaluate and compare gene groups as the output from computational biology tools such as clustering algorithms.

Background

In the midst of the functional genomic era, DNA microarray technology is widely used to study the expression level of thousands of genes under different experimental conditions or time points. Grouping of genes is a norm in functional genomics. Statistical criteria and biological hypotheses, combined with well defined computational techniques such as clustering [1-5], are then used to group genes that could be functionally related. Post clustering processes, such as functional characterization that relate the gene expression profile with the functional significance must be conducted. A prevailing procedure in functional characterization is to search the annotation of genes in respective group through Gene Ontology (GO) [6-8], KEGG pathways [9], REACTOME pathways [10], transcription factor (TF) gene association data [11], and/or gene networks. These post processing steps allow identifying biological functions which are highly represented in the given group of genes.

The Gene Ontology provides controlled vocabularies for the description of the molecular function, biological process, and cellular component of gene products [6-8]. TF-gene association data describes with a given probability whether or not a given gene is regulated by a specific transcription factor. One of the goals of identifying clusters of functionally related genes is to associate the genes and their interconnections with known biological pathways. In this regard, the KEGG pathways database provides a very rich resource to test this hypothesis. Its aim is to link individual level information such as genes, proteins, enzymes, with system level information such as interactions, enzymatic reactions, and pathways.

* Correspondence: alain.tchagang@nrc-cnrc.gc.ca, youlian.pan@nrc-cnrc.gc.ca
Knowledge Discovery Group, Institute for Information Technology, National Research Council, Canada, 1200 Montreal Road, Ottawa, ON K1A 0R6 Canada
Full list of author information is available at the end of the article

There are several packages available for functional characterization of gene groups: FuncAssociate [12], GOstat [13], BINGO [14], DAVID [15,16], CLENCH [17], FATIGO [18], GOSt [19], US FDA's ArrayTrack [20], EXPANDER [21], and many others that are referenced in the microarray tools' section of the GO website [22-24]. However, each of these has its limitations. Among these limitations has been the failure to consider which genes were actually assayed in the experiment that produced the list of interesting genes [12]. Several of these pioneering functional evaluation tools have mainly been implemented using only the GO annotations [12-14]. Some of these tools, such as BINGO [14] and CLENCH [17], are available as a part of some broader software packages or dedicated to specific applications, and do not offer flexibility in terms of deployment and biological applications. BiNGO [14] only offers GO evaluation and it is designed as a plug-in to Cytoscape [25]. CLENCH [17] is specifically designed to allow *Arabidopsis thaliana* researchers to perform automated retrieval of GO annotations from TAIR database [26]. DAVID [15,16], FATIGO [18], GOSt [19], ARRAYTRACKTM [20], and EXPANDER [21] offer the use of more advanced functions, such as protein-protein interactions and/or KEGG pathways, than the others. However, DAVID does not offer the possibility for user to incorporate available TF-gene association data in their analysis. Although FATIGO and EXPANDER make use of TF binding data, the user is limited to use the one provided by the software. Furthermore, packages such as FATIGO and several others that are only available as web server application make the users completely dependent on the availability of the server and unable to update associated data, such as GO data files.

In this paper we introduce a **G**ene **O**ntology **A**na**L**yzer, *GOAL*, which is a software tool specifically designed for the biological evaluation of groups of genes. *GOAL* implements and supports efficient and statistically rigorous biological interpretations of groups of genes through its integration with the Gene Ontology, available TF-gene association data, and KEGG pathways. Unlike other GO analysis tools, *GOAL* provides three GO-tree search strategies to detect the GO terms that are specific to a gene group and enables users to update source data, such as GO files, at anytime. Additionally, *GOAL* offers flexibility in terms of integration and deployment.

The hypotheses behind the use of Gene Ontology, TF-gene associations, the KEGG pathways in the implementation of *GOAL* are described as follows. Ideally, a good clustering algorithm identifies genes with similar expression patterns, which are probably co-regulated by the same transcription factors [1-5]. In addition, genes that are co-expressed frequently participate in the same or related biological pathways [1-5]. In both cases, if the genes in co-expressed groups are functionally related, they should be enriched with and annotated to related GO terms, involved in related biological pathways and regulated by related transcription factors [3].

The rest of this paper is described as follows. First, we describe the implementation. Then we illustrate the use of *GOAL* through application examples, and show how the *GOAL* package can be deployed. Finally, we discuss and conclude.

Implementation

GOAL is a Java module designed specifically to provide the functional characterization of groups of genes. Within the input list of genes, *GOAL* identifies statistically significant GO terms, TF-gene associations, and KEGG pathways. The framework relies on the *p*-value computed using a *Fisher's exact test* on a *2* × *2* contingency table derived either from the GO enrichment level, TF-gene associations, or KEGG pathways. The output of *GOAL* is a list of significant GO terms, TF-gene associations or KEGG pathways together with their *p-values* that state how significant the associations are for the list of input genes.

Statistical Significance

For a given set of input genes (gene id types are provided in the user manual) and a reference gene list, *GOAL* determines all annotated GO terms and their parental GO terms (defined by the input parameter *max path length from leaf* or *min path length from root*), all TF-gene associations, and all associations with KEGG pathway for the set of input genes. It then counts the number of appearances of each GO term, each TF-gene association, and each association with a KEGG pathway for the genes within the group and for the reference genes. *Fisher's exact test* is performed to determine whether or not the observed difference is significant. This will result in a *p-value* for each GO term, TF-gene association, or the association with a KEGG pathway that the observed counts could have been achieved from the background list by chance. The *p-value* is evaluated using the hypergeometric distribution for each GO term, TF-gene association, or KEGG pathway:

$$p = \sum_{i=n}^{\min(K,I)} \frac{\binom{K}{i}\binom{N-K}{I-i}}{\binom{N}{I}} \qquad (1)$$

where *n* is the number of genes in the intersection between the input gene list and those involved in the GO term, TF-gene association, or KEGG pathway of interest; *N* is the total number of unique genes in the reference gene list, *K* is the total number of genes that are annotated to the GO term, TF-gene association, or KEGG pathway of interest, and *I* is the number of genes in the input dataset. To give more statistical power to *GOAL*, the computed *p-value* is further corrected using one of the three multiple-testing corrections (Bonferroni correction, Bonferroni step-down, and Benjamini false discovery rate) [27]. These multiple-testing correction approaches are listed in the order of their stringency, with the Bonferroni being the most stringent, and the Benjamini FDR being the least stringent. The more stringent a multiple-testing correction, the less false positive observations are allowed.

Gene Ontology Search Strategies

Three GO-tree search strategies are implemented in *GOAL*: I) *max path length from leaf*, II) *min path length from root*, and III) the combination of I and II which we call *combined search strategy* in this article. Note that Strategies I and III are specific in *GOAL*, whereas, Strategy II is used in most other Gene Ontology tools mentioned above.

Max path length from leaf

here, we define *leaf* as the most specific GO term of a gene. This GO term may not be a leaf node along the GO-tree hierarchy. But there is no GO-term at the current stage that is more specific in defining the gene. *GOAL* first locates the most specific GO term of the gene along the hierarchy (leaf), users are allowed to define the maximum number of steps along its parental path to backtrack to a parent, grand-parent or great grand-parent, etc., of the GO term. This parameter defines the GO-tree search space to the k^{th} parental GO term from the most specific GO term. For example, if a user wants to search the most specific GO term only, the value is 0.

Min path length from root

GOAL first locates the most general GO term of the gene along the hierarchy (root, e.g. Biological Process); users are allowed to define a threshold number of steps from the root along the offspring path. This threshold number of steps is the cut off level in the GO-tree. Most existing tools implement only this strategy.

Combined search

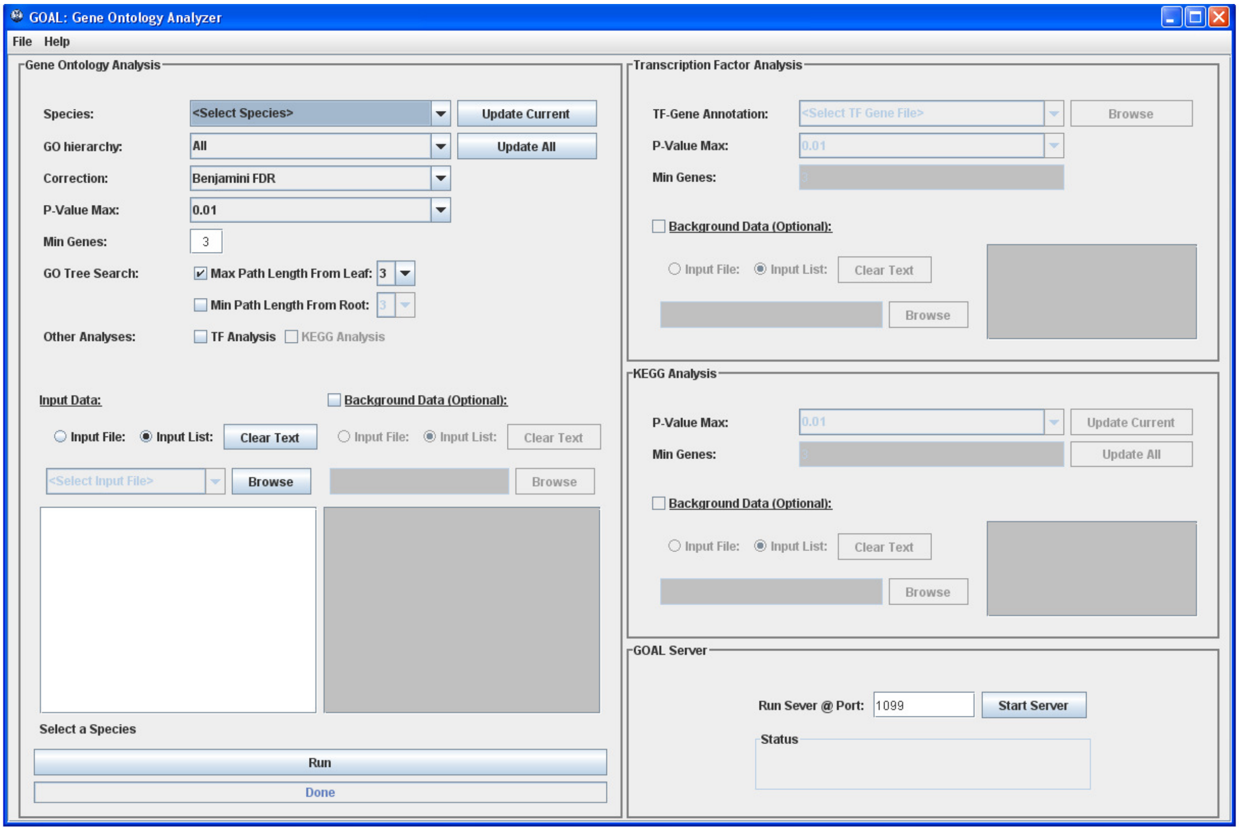
This is a combination of Strategies I and II. When this approach is used, *GOAL* first locates the most specific GO term as described in strategy I. Then, the *p-value* of each one of them is computed as in strategy II. By doing this, *GOAL* considers the most specific GO terms, while discarding the ones that are more general than the threshold set by Strategy II. Another advantage is that it also considers genes that are associated with parental GO terms of the most specific GO terms located within strategy II.

Development

GOAL is implemented entirely in Java. It is available as an executable jar file and technically works with all operating system supporting Java SDK 1.6 or later. We tested under Microsoft Windows (XP, Vista) and Linux systems (Ubuntu 9.10) and provided a start-up script for these two operating systems. The GUI is written using the Java Swing and AWT libraries which are part of the Java SDK. For network communication between clients and the server, the Java RMI library is used, which is also part of the SDK. Other packages included are Jargs (command line interface) and CSVreader (File IO). *GOAL* also makes use of external Gene Ontology and gene annotation files [6-8]. The GO files are accessible for analysis by *GOAL* using a partial, modified version of the BioJava 1.6.1 API. *GOAL* can download these files directly from the Gene Ontology [22] or European Bioinformatics Institutes [28]. All GO file can be updated at any time by clicking an update button.

The GUI provides user convenience of inputting the list of genes and selecting various parameters. Users first specify a list of gene identifiers of interest in the text area of the input interface (Figure. 1). The list of gene identifiers can also be uploaded as a delimited data file specified in the User Manual (accessible through the help menu of *GOAL*). It is the same for reference gene list if users desire to input their own references. By default, the system provides the whole genome of the selected species as the reference. Next, the user specifies the gene annotation source (species) from a drop-down list, which specifies the GO file that *GOAL* will use to seek for annotation and to calculate the *p*-values. A list of default parameters (GO hierarchy, *p*-value correction, maximum *p*-value, maximum path length from leaf and/or minimum path length from root, minimum number of genes in a term) is also available in the GUI (Figure 1). The user may change these default values if they desire. If the TF-gene association data for the corresponding study and species are available, users can upload them as tab delimited data files. Finally, users can also choose the KEGG pathway evaluation if the selected species is supported.

The TF-gene association file is compiled by the user. As we mentioned earlier, the TF-gene associations data can be derived from Chromatin Immunoprecipitation (ChIP) [11] experiments. They can also be compiled from literature and/or TF databases of well known and well characterized biological interactions such as TRANSFAC [29], JASPAR [30], and RegulonDB [31], etc. For example, TRANSFAC is a database on eukaryotic transcriptional regulation. The database contains data on transcription factors, their target genes and their experimentally-proven binding motifs in genes' promoter. The TF-gene



GOAL: Gene Ontology Analyzer

Gene Ontology Analysis

Species: <Select Species>

GO hierarchy: All

Correction: Benjamini FDR

P-Value Max: 0.01

Min Genes: 3

GO Tree Search: ☑ Max Path Length From Leaf: 3

□ Min Path Length From Root: 3

Other Analyses: □ TF Analysis □ KEGG Analysis

Input Data: ○ Input File: ● Input List:

<Select Input File>

□ Background Data (Optional): ○ Input File: ● Input List:

Select a Species

Transcription Factor Analysis

TF-Gene Annotation: <Select TF Gene File>

P-Value Max: 0.01

Min Genes:

□ Background Data (Optional): ○ Input File: ● Input List:

KEGG Analysis

P-Value Max: 0.01

Min Genes:

□ Background Data (Optional): ○ Input File: ● Input List:

GOAL Server

Run Sever @ Port: 1099

Status

Figure 1 ***GOAL* input interface**. The GUI is divided into four sections: GO analysis, TF-gene analysis, KEGG pathways analysis, and the *GOAL* server. The first three sections are further divided into subsections, where the user can specify the input parameters, the list of gene identifiers, which specify the group of genes of interest and the background data. Users have an option to input their own list of reference. The executable button is located at left bottom of the interface.

association data is represented as a matrix, where the rows correspond to genes and the columns to transcription factors. The entries of the matrix are either 0 or 1, with 1 for a known interaction between a TF and its target gene and 0 for no interaction.

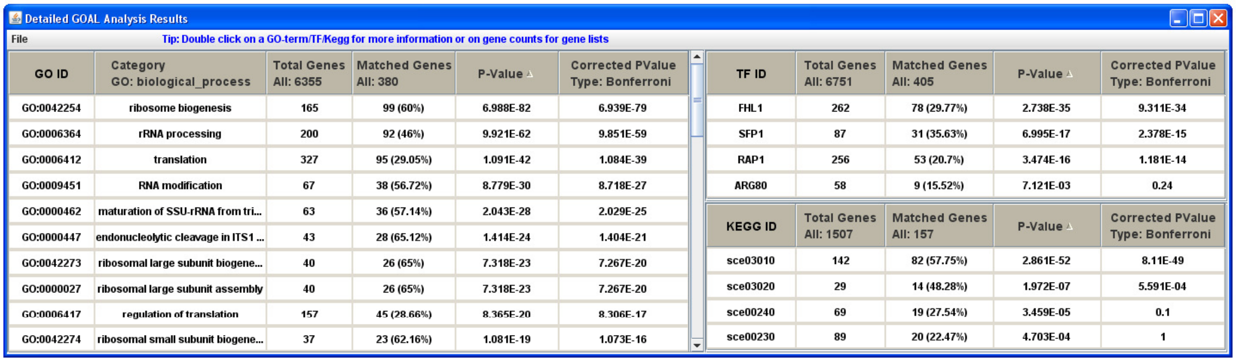
Following the input phase, the *GOAL* algorithm executes as described above, and new windows will appear displaying the functional evaluation results. Few examples are shown in Figures 2, 3 and 4. From these new windows, a user has the option to explore or save the results as a comma separated values (CSV) file. Furthermore, the GO terms in the output contain hypertext links to a visualization tool for the GO hierarchy; genes are linked to NCBI [32], and KEGG pathways ids are linked to KEGG database [9].

GOAL is available as Additional File 1. A quick video tutorial is also available as Additional File 2. The most recent version of *GOAL* can be downloaded directly from the *GOAL* project website: http://bioinfo.iit.nrc.ca/GOAL/.

Results

Application Examples

We used three different datasets to test *GOAL*. Groups of co-expressed genes were obtained from each set of data using a well defined clustering algorithm [3,4]. The first dataset is a group of co-expressed gene from the time-series gene expression data of the *Saccharomyces cerevisiae* amino acid (AA) starvation dataset [33]. For the TF-gene association data, we used the published AA starvation ChIP-chip dataset on 34 transcription factors [34]. Figure 2 shows the combined results of the GO analysis, TF-gene association, and KEGG pathway association of an identified gene group using *GOAL*. The results shown in Figure 2 are consistent with experimental result; the transcription factors FHL1 (p-value = 9.31e-34) and SFP1 (p-value = 1.18e-15) regulate many genes involved in ribosome biosynthesis [35]. This is the most significant GO biological process in this cluster. This is also consistent with the KEGG pathway analysis where the ribosome pathway (sce03010, p-value = 8.11e-49) is the most significant.



Tip: Double click on a GO-term/TF/Kegg for more information or on gene counts for gene lists

GO ID	Category GO: biological_process	Total Genes All: 6355	Matched Genes All: 380	P-Value	Corrected PValue Type: Bonferroni
GO:0042254	ribosome biogenesis	165	99 (60%)	6.988E-82	6.939E-79
GO:0006364	rRNA processing	200	92 (46%)	9.921E-62	9.851E-59
GO:0006412	translation	327	95 (29.05%)	1.091E-42	1.084E-39
GO:0009451	RNA modification	67	38 (56.72%)	8.779E-30	8.718E-27
GO:0000462	maturation of SSU-rRNA from tri...	63	36 (57.14%)	2.043E-28	2.029E-25
GO:0000447	endonucleolytic cleavage in ITS1 ...	43	28 (65.12%)	1.414E-24	1.404E-21
GO:0042273	ribosomal large subunit biogene...	40	26 (65%)	7.318E-23	7.267E-20
GO:0000027	ribosomal large subunit assembly	40	26 (65%)	7.318E-23	7.267E-20
GO:0006417	regulation of translation	157	45 (28.66%)	8.365E-20	8.306E-17
GO:0042274	ribosomal small subunit biogene...	37	23 (62.16%)	1.081E-19	1.073E-16

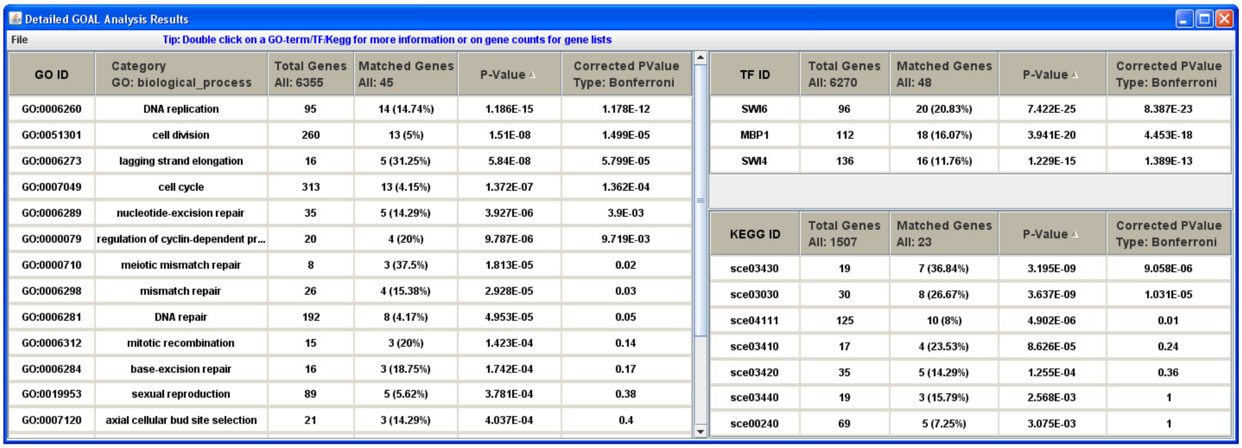
TF ID	Total Genes All: 6751	Matched Genes All: 405	P-Value	Corrected PValue Type: Bonferroni
FHL1	262	78 (29.77%)	2.738E-35	9.311E-34
SFP1	87	31 (35.63%)	6.995E-17	2.378E-15
RAP1	256	53 (20.7%)	3.474E-16	1.181E-14
ARG80	58	9 (15.52%)	7.121E-03	0.24

KEGG ID	Total Genes All: 1507	Matched Genes All: 157	P-Value	Corrected PValue Type: Bonferroni
sce03010	142	82 (57.75%)	2.861E-52	8.11E-49
sce03020	29	14 (48.28%)	1.972E-07	5.591E-04
sce00240	69	19 (27.54%)	3.459E-05	0.1
sce00230	89	20 (22.47%)	4.703E-04	1

Figure 2 ***GOAL*** **output of Gene Ontology, transcription factor and target gene association, and KEGG metabolic pathways gene association evaluation of the** ***Saccharomyces cerevisiae*** **amino acid starvation data**.

The second dataset is a group of co-expressed genes from the *Saccharomyces cerevisiae* cell cycle dataset [36]. The TF-gene association data we used here are derived from [37]. This dataset contains 113 yeast TFs and their target genes. Figure 3 shows the combined results of the GO analysis, TF-gene association, and KEGG pathway association of an identified cluster using *GOAL*. The most significant GO biological processes identified for this cluster is "cell cycle", with a *p*-value = 8.6e-09 (Figure 3). KEGG pathways and TF-gene results are consistent with the GO analysis. TF-gene analysis shows that the transcription factors MBP1 (P-value = 4.4e-18), SWI4 (p-value = 1.3e-13) and SWI6 (p-value = 8.3e-23) have significantly overrepresented number of target genes in this cluster. These three TFs are known to participate in the two major transcription complexes regulating G1/S transition in cell cycle process [38]. KEGG pathways analysis also revealed the cell cycle pathway (sce04111, p-value = 4.6e-03) and DNA replication pathway (sce03030, p-value = 5.1e-06). This indicates that the three parts of analysis are consistent.

The third dataset is a co-expressed gene group from the time series gene expression data of *Brassica napus* during seed development [39]. Since the *B. napus* genome is not yet complete, we use the orthologs in *A. thaliana*, a close relative of *B. napus*, for computational analysis. The search for orthologs in *A. thaliana* is done by blasting the sequences of interest *B. napus* genes against the TAIR [40] database (version 8) and keeping only the matches with scores better than 1e-20. The TF-gene interactions matrix was inferred through a carefully selected literature search, bias towards experimental evidence [41-43]. More precisely, we compiled from the literature a set of *A. thaliana* TFs that have been shown by microarray experiments and/or quantitative RT-PCR to promote, suppress, or induce genes that are related to seed development, FA metabolism and other related biological processes such as lipid metabolism and biosynthetic process, with a *p-value*



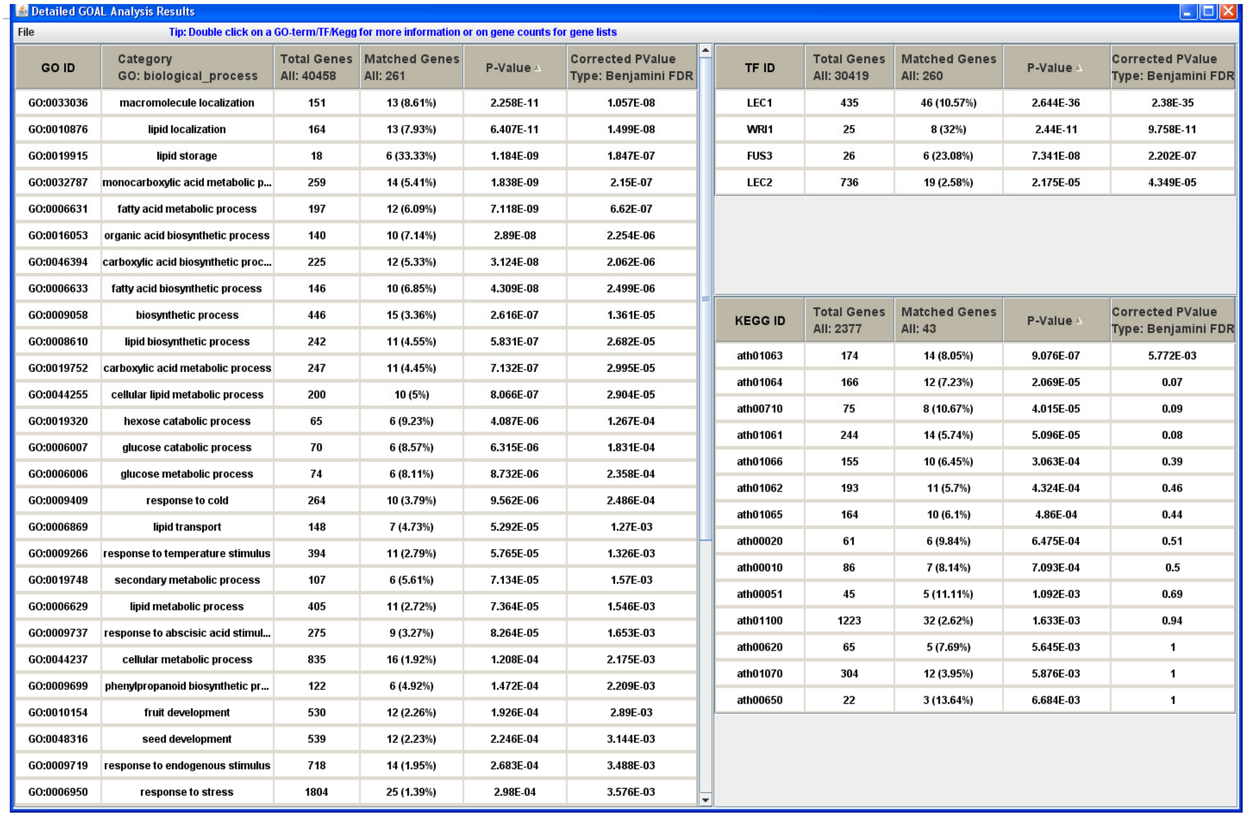
Tip: Double click on a GO-term/TF/Kegg for more information or on gene counts for gene lists

GO ID	Category GO: biological_process	Total Genes All: 6355	Matched Genes All: 45	P-Value	Corrected PValue Type: Bonferroni
GO:0006260	DNA replication	95	14 (14.74%)	1.186E-15	1.178E-12
GO:0051301	cell division	260	13 (5%)	1.51E-08	1.499E-05
GO:0006273	lagging strand elongation	16	5 (31.25%)	5.84E-08	5.799E-05
GO:0007049	cell cycle	313	13 (4.15%)	1.372E-07	1.362E-04
GO:0006289	nucleotide-excision repair	35	5 (14.29%)	3.927E-06	3.9E-03
GO:0000079	regulation of cyclin-dependent pr...	20	4 (20%)	9.787E-06	9.719E-03
GO:0000710	meiotic mismatch repair	8	3 (37.5%)	1.813E-05	0.02
GO:0006298	mismatch repair	26	4 (15.38%)	2.928E-05	0.03
GO:0006281	DNA repair	192	8 (4.17%)	4.953E-05	0.05
GO:0006312	mitotic recombination	15	3 (20%)	1.423E-04	0.14
GO:0006284	base-excision repair	16	3 (18.75%)	1.742E-04	0.17
GO:0019953	sexual reproduction	89	5 (5.62%)	3.781E-04	0.38
GO:0007120	axial cellular bud site selection	21	3 (14.29%)	4.037E-04	0.4

TF ID	Total Genes All: 6270	Matched Genes All: 48	P-Value	Corrected PValue Type: Bonferroni
SWI6	96	20 (20.83%)	7.422E-25	8.387E-23
MBP1	112	18 (16.07%)	3.941E-20	4.453E-18
SWI4	136	16 (11.76%)	1.229E-15	1.389E-13

KEGG ID	Total Genes All: 1507	Matched Genes All: 23	P-Value	Corrected PValue Type: Bonferroni
sce03430	19	7 (36.84%)	3.195E-09	9.058E-06
sce03030	30	8 (26.67%)	3.637E-09	1.031E-05
sce04111	125	10 (8%)	4.902E-06	0.01
sce03410	17	4 (23.53%)	8.626E-05	0.24
sce03420	35	5 (14.29%)	1.255E-04	0.36
sce03440	19	3 (15.79%)	2.568E-03	1
sce00240	69	5 (7.25%)	3.075E-03	1

Figure 3 ***GOAL*** **output of Gene Ontology, transcription factor and target gene association, and KEGG metabolic pathways gene association evaluation of the** ***Saccharomyces cerevisiae*** **cell cycle data**.



Tip: Double click on a GO-term/TF/Kegg for more information or on gene counts for gene lists

GO ID	Category GO: biological_process	Total Genes All: 40458	Matched Genes All: 261	P-Value △	Corrected PValue Type: Benjamini FDR
GO:0033036	macromolecule localization	151	13 (8.61%)	2.258E-11	1.057E-08
GO:0010876	lipid localization	164	13 (7.93%)	6.407E-11	1.499E-08
GO:0019915	lipid storage	18	6 (33.33%)	1.184E-09	1.847E-07
GO:0032787	monocarboxylic acid metabolic p...	259	14 (5.41%)	1.838E-09	2.15E-07
GO:0006631	fatty acid metabolic process	197	12 (6.09%)	7.118E-09	6.62E-07
GO:0016053	organic acid biosynthetic process	140	10 (7.14%)	2.89E-08	2.254E-06
GO:0046394	carboxylic acid biosynthetic proc...	225	12 (5.33%)	3.124E-08	2.062E-06
GO:0006633	fatty acid biosynthetic process	146	10 (6.85%)	4.309E-08	2.499E-06
GO:0009058	biosynthetic process	446	15 (3.36%)	2.616E-07	1.361E-05
GO:0008610	lipid biosynthetic process	242	11 (4.55%)	5.831E-07	2.682E-05
GO:0019752	carboxylic acid metabolic process	247	11 (4.45%)	7.132E-07	2.995E-05
GO:0044255	cellular lipid metabolic process	200	10 (5%)	8.066E-07	2.904E-05
GO:0019320	hexose catabolic process	65	6 (9.23%)	4.087E-06	1.267E-04
GO:0006007	glucose catabolic process	70	6 (8.57%)	6.315E-06	1.831E-04
GO:0006006	glucose metabolic process	74	6 (8.11%)	8.732E-06	2.358E-04
GO:0009409	response to cold	264	10 (3.79%)	9.562E-06	2.486E-04
GO:0006869	lipid transport	148	7 (4.73%)	5.292E-05	1.27E-03
GO:0009266	response to temperature stimulus	394	11 (2.79%)	5.765E-05	1.326E-03
GO:0019748	secondary metabolic process	107	6 (5.61%)	7.134E-05	1.57E-03
GO:0006629	lipid metabolic process	405	11 (2.72%)	7.364E-05	1.546E-03
GO:0009737	response to abscisic acid stimul...	275	9 (3.27%)	8.264E-05	1.653E-03
GO:0044237	cellular metabolic process	835	16 (1.92%)	1.208E-04	2.175E-03
GO:0009699	phenylpropanoid biosynthetic pr...	122	6 (4.92%)	1.472E-04	2.209E-03
GO:0010154	fruit development	530	12 (2.26%)	1.926E-04	2.89E-03
GO:0048316	seed development	539	12 (2.23%)	2.246E-04	3.144E-03
GO:0009719	response to endogenous stimulus	718	14 (1.95%)	2.683E-04	3.488E-03
GO:0006950	response to stress	1804	25 (1.39%)	2.98E-04	3.576E-03

TF ID	Total Genes All: 30419	Matched Genes All: 260	P-Value △	Corrected PValue Type: Benjamini FDR
LEC1	435	46 (10.57%)	2.644E-36	2.38E-35
WRI1	25	8 (32%)	2.44E-11	9.758E-11
FUS3	26	6 (23.08%)	7.341E-08	2.202E-07
LEC2	736	19 (2.58%)	2.175E-05	4.349E-05

KEGG ID	Total Genes All: 2377	Matched Genes All: 43	P-Value △	Corrected PValue Type: Benjamini FDR
ath01063	174	14 (8.05%)	9.076E-07	5.772E-03
ath01064	166	12 (7.23%)	2.069E-05	0.07
ath00710	75	8 (10.67%)	4.015E-05	0.09
ath01061	244	14 (5.74%)	5.096E-05	0.08
ath01066	155	10 (6.45%)	3.063E-04	0.39
ath01062	193	11 (5.7%)	4.324E-04	0.46
ath01065	164	10 (6.1%)	4.86E-04	0.44
ath00020	61	6 (9.84%)	6.475E-04	0.51
ath00010	86	7 (8.14%)	7.093E-04	0.5
ath00051	45	5 (11.11%)	1.092E-03	0.69
ath01100	1223	32 (2.62%)	1.633E-03	0.94
ath00620	65	5 (7.69%)	5.645E-03	1
ath01070	304	12 (3.95%)	5.876E-03	1
ath00650	22	3 (13.64%)	6.684E-03	1

Figure 4 ***GOAL*** **output of Gene Ontology, transcription factor and target gene association, and KEGG metabolic pathways gene association evaluation of the** ***Brassica napus*** **data**.

< 0.01 [36-38]. Figure 4 shows the results of the GO analysis, TF-gene association, and the involvement in the metabolic pathways. GO biological process shows that most of the genes in this group are highly enriched under seed development, lipid and fatty acid biosynthetic and metabolism. This is consistent with the TF-gene association results and KEGG pathway results. This group of genes is significantly co-regulated by the transcription factors LEC1 (p-value = 2.3e-35), LEC2 (P-value = 4.3e-05), FUS3 (p-value = 2.2e-07), and WRI1 (p-value = 9.7e-11). Significant number of genes in this group are involved in the biosynthesis of alkaloids derived from shikimate pathway (ath01063, 14 genes), phenylpropanoid (ath01061, 14 genes), which are related with fatty acid biosynthesis (ath00061) and metabolism (ath00071), and citrate cycle (ath00020). These results are consistent with the fact that the LEC1 and LEC2 have been shown to play major roles during seed development, and also with the fact that LEC1 function is partially dependent on FUS3 and WRI1 in the regulation of FA and FA-derived complex lipid [41].

Comparative analysis of GO-tree search Strategies I and II (Figures 5 and 6) shows that Strategy I usually outperforms Strategy II in terms of percentage of genes associated to a GO term (Figure 5). Here, the percentage of genes associated to a GO term corresponds to the number of genes in the intersection of a cluster with a GO term divided by the total number of genes in the background that are annotated to that GO term. Nevertheless, the *p-values* of GO terms in Strategy II are in some cases slightly more significant as compared to the ones in Strategy I (Figure 6). Further analysis showed that several GO terms identified using Strategy II are more generic compared to the ones identified by Strategy I which are more specific. Thus groups of genes in Strategy I are biologically more parsimonious than the ones in Strategy II. Note that, there might be some values for which the two GO-tree search strategies yield the same output. Furthermore, from a statistical perspective, a search strategy that is related to more genes is preferable. But, in biological applications, this assertion is biologically relevant if the GO terms depicted by the corresponding search strategy are more specific. In other words, a search strategy that is related to more genes and that depicts more generic GO terms is less biologically relevant compared to a search strategy that depicts more specific GO terms as strategy I.

The data used for these analyses are provides in the *GOAL* tool as default data and can be exploited by the

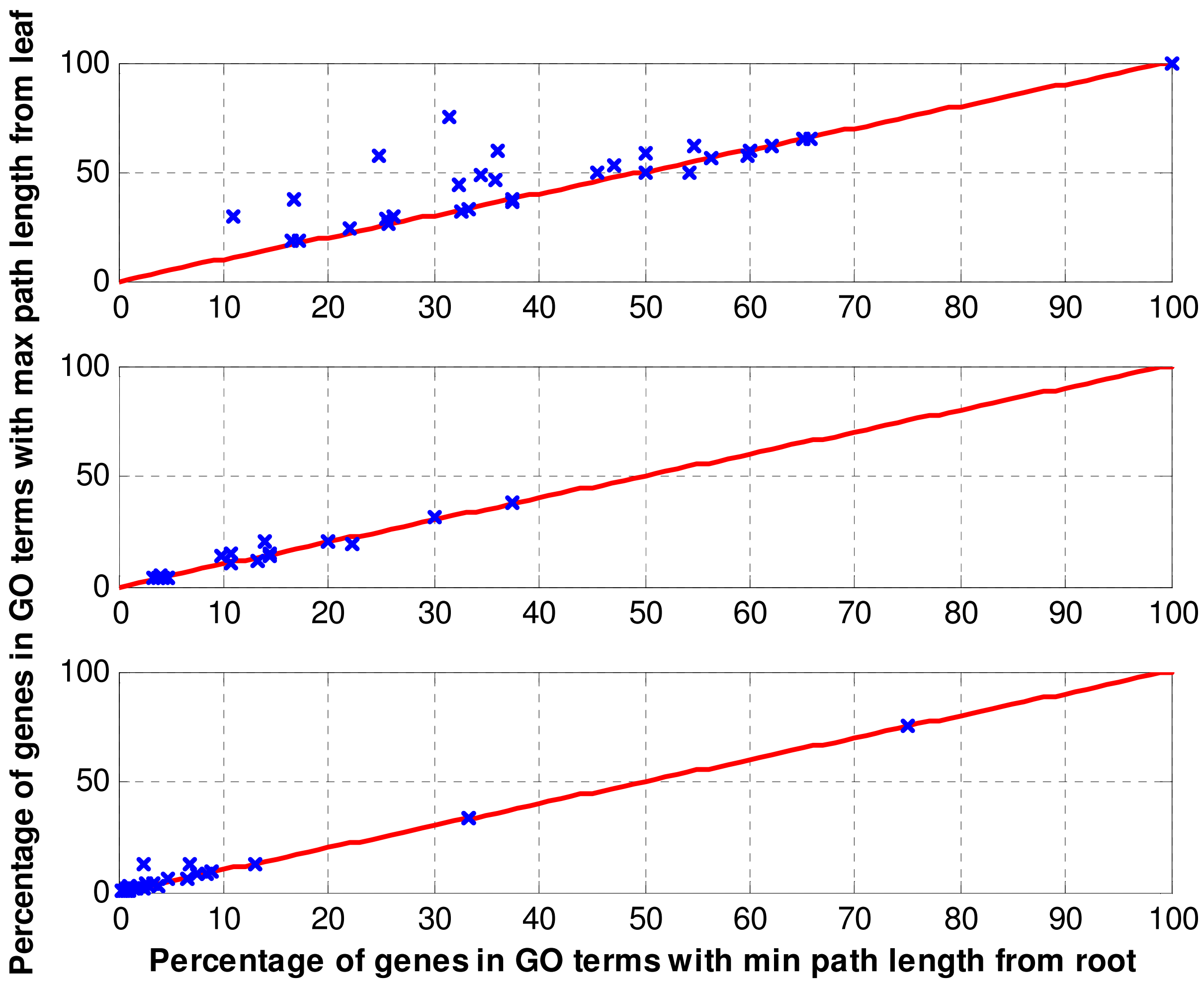


Figure 5 Comparative analysis of the *max path length from leaf* (Strategy I) and the *min path length from root* (Strategy II) approaches in terms of percentage of genes in each GO category. Each row represents one of the three datasets: *Saccharomyces cerevisiae* amino acid starvation data, *Saccharomyces cerevisiae* cell cycle data, and *Brassica napus* data respectively. The *y*-axis represents the percentage of gene in a GO terms depicted using Strategy I (0, the most specific term by itself; no parent is considered). Likewise, the x-axis represents the percentages of genes in the same GO terms identified using Strategy II (3, a default in several tools). Points above the central diagonal line represent the GO terms with higher percentage of genes in Strategy I than Strategy II, whereas points below the line, the GO terms with higher percentage of genes identified using Strategy II. As can be seen, Strategy I has the largest number of GO terms with higher percentage of genes.

users in their own research. *GOAL* has also been successfully used to characterize the following six human genes (IFNA4 IL12B IL2RB STAT1 STAT2 IRF9) provided to us by one of the reviewers (Additional File 3), to characterize groups of correlated genes in DNA microarray experimental data of *Arabidopsis thaliana* under pathogens attack [44], and several in house (private) DNA microarray datasets of *Brassica napus* during seed development.

Deployment

GOAL can be used as a standalone tool and also offers flexibility to be integrated with other applications. For example, it can be a plug-in to other computational biology tools, or deployed as a web server application.

The standalone application

the tool is packaged into a single executable jar file and only requires Java to be installed. A configuration file is automatically generated on the first run and can be modified to suit the user's need. The installer also creates necessary file structure and a shortcut to a batch file to run the tool. GO files are downloaded from GO database [22] on the first run.

Plug-in application

GOAL can be easily integrated into other tools with minimal coding by the user. The user can create an instance of the *GOALPlugin* class which gives access to all components of *GOAL*. It can be used to retrieve or display graphical user interface (GUI) components for integra-

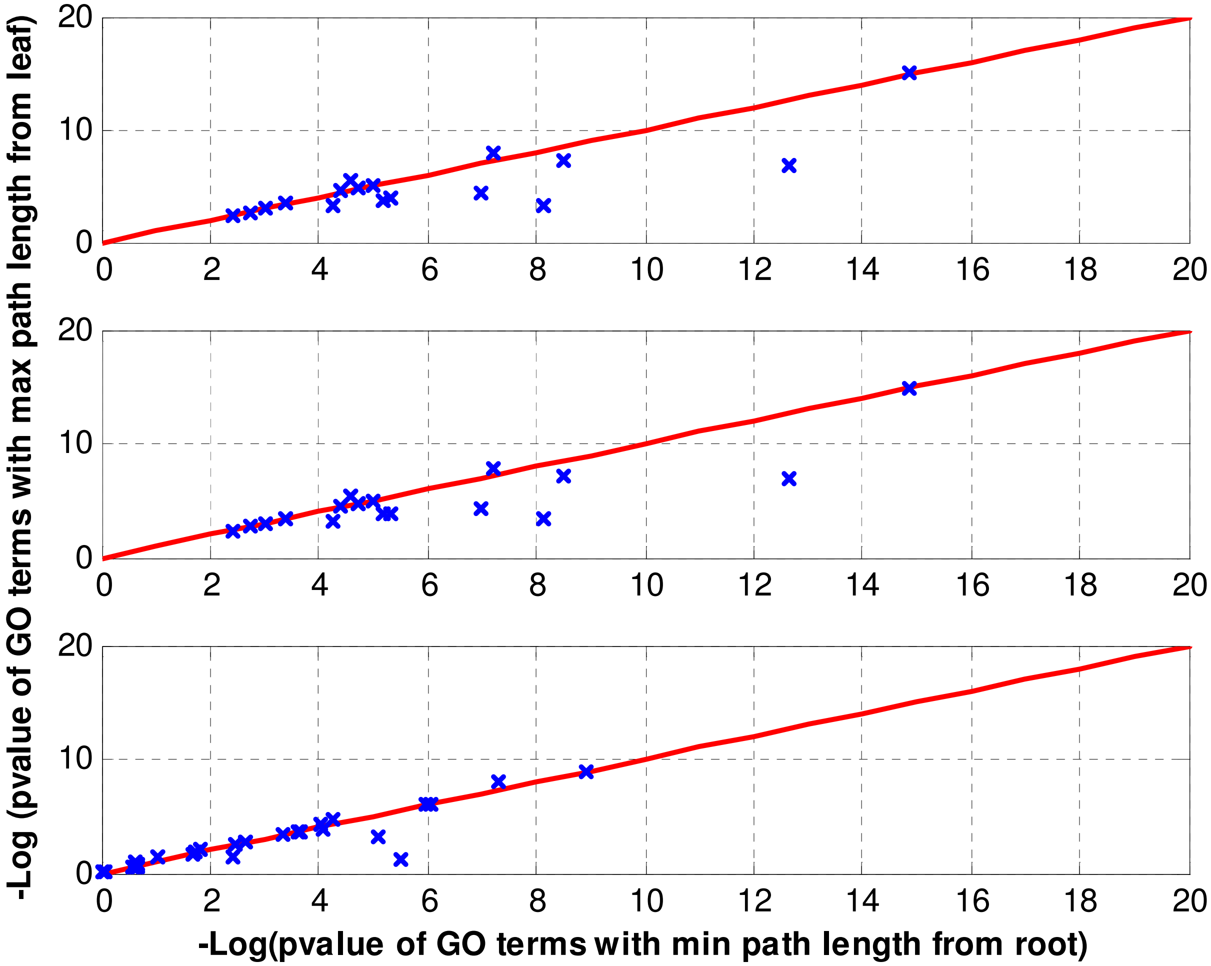


Figure 6 Comparative analysis of the *max path length from leaf* (Strategy I) and the *min path length from root* (Strategy II) approaches in terms of *p*-value of each GO term. Each row represents one of the three datasets: *Saccharomyces cerevisiae* amino acid starvation data, *Saccharomyces cerevisiae* cell cycle data, and *Brassica napus* data respectively. The *y*-axis represents the minus log *P*-value for GO enrichment using the Strategy I (0, the most specific term by itself; no parent is considered). The *x*-axis represents the minus log *P*-value for GO enrichment using Strategy II (3, a default in several tools). Points above the central diagonal line represent the most significant GO-terms in Strategy I, whereas points below the line represent categories with most significant terms in Strategy II. As can be seen, Strategy II has slightly larger number of significant GO terms.

tion into existing GUIs or display in a separate window. The class also allows for running functional components independently of the GUI in a separate or currently used thread. Furthermore the user can execute the jar file through a provided command line interface.

Web server application

Another functionality of *GOAL* is the ability to be used remotely through a java Remote Method Invocation (RMI) server. Along with the standard package, a client package is also available. The client can be installed on any remote computer and communicate with the server machine that host a web server. All the computation initiated by various clients is done by the same server. This creates an organized environment where only the server needs to keep annotation files up-to-date. Furthermore only the server would require suitable hardware to run the computations. This way, the server version can also be a middleware to a database that host gene data. More details are available in the User Manual.

Discussion

The recent proliferation of high-throughput functional genomics methods has brought with it a variety of ways to produce large lists of genes based on well defined statistical assumptions and biological hypotheses. For example, clustering algorithms are used to collect sets of co-expressed genes. Genes in each group that behave similarly in the expression profile across experimental conditions or time points most likely share similar biological functions and processes, genetic pathways, and probably co-regulated. Several solutions have been offered to address these issues. However, each of these has its limitations, of which, the lack of appropriate corrections for multiple hypothesis testing and the failure to consider whether a gene was actually validated in wet lab experiment are the two most obvious ones [12]. The *GOAL* package addresses this issue by implementing several multiple correction approaches and by allowing users to incorporate ChIP-chip data.

Another important aspect of ontological analysis regardless of the tool or statistical method is the choice of the reference gene list against which the list of differentially regulated genes is compared. Inappropriate choice of reference genes may lead to false functional characterization of a gene list. Khatri and Dr№ghici [24] pointed out that only the genes represented on the array, although quite incomplete, should be used as reference list instead of the whole genome as it is a common practice. The *GOAL* package addresses this issue by allowing the users to either use the entire genome of the selected species as the reference, or upload their own reference genes.

In addition, GO allows for the annotation of genes at different levels of abstraction due to its directed acyclic graph (DAG) structure. In this particular hierarchical structure, each GO term can have one or more child terms as well as one or more parent terms. For instance, the same gene list is annotated with a more general GO term such as "metabolic process" at a higher level of abstraction, whereas the lowest level provides a more specific ontology term such as "lipid metabolic process." It is important as in the *GOAL* package, to integrate the hierarchical structure of the GO in the analysis since various levels of abstraction usually give different p-values. In this regard, *GOAL* first locate the most specific GO term of the gene (we define it as a leaf) along the hierarchy, users are allowed to define the maximum number of steps along parental path to backtrack to a parent or grand-parent of the GO term. *GOAL* also implements a GO-tree search strategy (min path length from root) common to most existing Gene Ontology packages. Additionally, users may want to combine both strategies to do more specific functional characterization of a gene group. *GOAL* provides the users with three options to determine significant function of a gene group and to better address their research problem.

Other than some GO evaluation tool as implemented in several existing packages, *GOAL* takes a step further by also including the association between transcription factors and their target genes, and gene's involvement in various metabolic pathways in its evaluation process. The co-expressed genes identified through clustering processes are probably co-regulated by common transcription factors and frequently participate in the same or related biological pathways [1-5]. Thus, they should be enriched with certain GO terms, related with certain TFs, and associated with certain biological pathways. *GOAL* is not only capable of evaluating whether or not a given group of genes may be involved in the same genetic pathways, but also capable of evaluating whether or not they might be co-regulated at the transcriptional level.

Conclusion

We have introduced a new software package for functional characterization of a gene list and to evaluate the functional domain of these genes. *GOAL* integrates Gene Ontology with available TF-gene associations and KEGG pathways information. *GOAL* provides users with three different GO-tree search strategies. *GOAL* is highly portable, can be used as a standalone tool, a plug-in to other computational biology tools, or deployed as a web server application. This way, users have an option to keep privacy of their datasets without sending over their data through internet to the server provided by the software owner. Also, users may analyze their data using most updated information by a single click that updates the data source files (e.g. GO files). *GOAL* presents its result of analysis in a highly visual and interactive manner. The unique automated analysis capabilities of *GOAL* combined with its visualization and integration capabilities should merit *GOAL* to be the software tool of choice for functional evaluation of gene groups. *GOAL* is available in Additional File 1. Subsequent update will be available at *GOAL* home page: http://bioinfo.iit.nrc.ca/GOAL/.

Availability And Requirements

Software name: GOAL: **G**ene **O**ntology **A**na**L**yzer

Project home page: http://bioinfo.iit.nrc.ca/GOAL/

Operating system(s): Tested in Microsoft Windows (XP, Vista) and Linux (Ubuntu 9.10); technically, it should work with all operating systems supporting Java 1.6 or higher.

Programming language: Java

Other requirements: Java 1.6 or higher, internet connection

License: non-commercial research use license

Any restrictions to use by non-academics: license needed from authors for commercial use.

Additional material

Additional file 1 *GOAL* jar file and *GOAL* user manual (*GOAL*-1.0.zip).

Additional file 2 Quick video tutorial of GOAL (GOAL_Quick_Video_Tutorial.zip) (HTML, JavaScript & Shockwave Flash files). Require web browser to visualize. This tutorial is also accessible via the GOAL project website at the following URL: http://bioinfo.iit.nrc.ca/GOAL/demo/GOAL-Demo.htm

Additional file 3 *GOAL* (GO biological process and KEGG pathway association analysis) and GOSt [19] detail results of six human genes: IFNA4 IL12B IL2RB STAT1 STAT2 IRF9. They were provided to us by one of the reviewers.

Authors' contributions

ABT initiated the *GOAL* project. ABT and YP both contributed to the design of *GOAL*. AG implemented *GOAL*. All participated in the drafting and revising of the manuscript.

Acknowledgements

We are thankful to three anonymous reviewers for their constructive criticisms. This work is supported by Genomics and Health Initiative and Institute for Information Technology of the National Research Council Canada.

Author Details

Knowledge Discovery Group, Institute for Information Technology, National Research Council, Canada, 1200 Montreal Road, Ottawa, ON K1A 0R6 Canada

Received: 8 September 2009 Accepted: 6 May 2010
Published: 6 May 2010

References

1. Androulakis IP, Yang E, Almon RR: **Analysis of Time-Series Gene Expression Data: Methods, Challenges, and Opportunities.** *Annu Rev Biomed Eng* 2007, **9:**205-228.
2. Madeira SC, Oliveira AL: **Biclustering algorithms for biological data analysis: a survey.** *IEEE Transactions on Computational Biology and Bioinformatics* 2004, **1:**24-45.
3. Tchagang AB, Tewfik AH, Benos PV: **Biological Evaluation of Biclustering Algorithms using Gene Ontology and ChIP-chip data.** *Proceedings of IEEE, International Conference on Acoustics, Speech and Signal Processing: 30 March - 4 April 2008; Las Vegas, Nevada* .
4. Tchagang AB, Bui KV, McGinnis T, Benos PV: **Extracting biologically significant patterns from short time series gene expression data.** *BMC Bioinformatics* 2009, **10:**255.
5. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data.** *Bioinformatics* 2006, **22:**1122-9.
6. The Gene Ontology Consortium.: **Gene ontology: tool for the unification of biology.** *Nat Genet* 2000, **25(1):**25-9.
7. Day-Richter J, Harris MA, Haendel M: **Gene Ontology OBO-Edit Working Group, OBO-Edit--an ontology editor for biologists.** *Bioinformatics* 2007, **23(16):**2198-200.
8. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO Hub, Web Presence Working Group. AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25(2):**288-9.
9. Okuda Shujiro, Yamada Takuji, Hamajima Masami, Itoh Masumi, Katayama Toshiaki, Bork Peer, Goto Susumu, Kanehisa Minoru: **KEGG Atlas mapping for global analysis of metabolic pathways.** *Nucleic Acids Research* 2008, **36:**.
10. **REACTOME pathways** [http://www.reactome.org/]
11. Buck MJ, Lieb JD: **ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.** *Genomics* 2004, **83:**349-60.
12. Berriz GF, King OD, Bryant B, Sander C, Roth FP: **Characterizing gene sets with FuncAssociate.** *Bioinformatics Applications Note* 2003, **19(18):**2502-2504.
13. Beißbarth T, Speed TP: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics Applications Note* 2004, **20(9):**1464-1465.
14. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks.** *Bioinformatics* 2005, **21:**3448-3449.
15. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources.** *Nature Protoc* 2009, **4(1):**44-57.
16. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4(5):**P3.
17. Shah NH, Fedoroff NV: **CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology.** *Bioinformatics* 2004, **20(7):**1196-7.
18. Al-Shahrour F, Minguez P, Tárraga J, Montaner D, Alloza E, Vaquerizas JMM, Conde L, Blaschke C, Vera J, Dopazo J: **BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments.** *Nucleic Acids Research* 2006, **34:**W472-W476.
19. **GOSt** [http://biit.cs.ut.ee/gprofiler/]
20. **ARRAYTRACK™** [http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/]
21. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R: **EXPANDER-an integrative program suite for microarray data analysis.** *BMC Bioinformatics* 2005, **21:**6:232.
22. **Gene Ontology** [http://www.geneontology.org]
23. Coulibaly I, Page GP: **Bioinformatic Tools for Inferring Functional Information from Plant Microarray Data II: Analysis Beyond Single Gene.** *International Journal of Plant Genomics* **2008:**.
24. Khatri P, Dr№ghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, **21(18):**3587-3595.
25. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
26. Swarbreck D, Wilks C, Lamesch P: **The Arabidopsis Information Resource (TAIR): gene structure and function annotation.** *Nucleic Acids Research* 2008, **36:**D1009-D1014.
27. Kanji GK: **100 Statistical Tests.** Thousand Oaks, CA: SAGE Publications; 1999:110.
28. **European Bioinformatics Institutes** [http://www.ebi.ac.uk/]
29. Wingender E, Dietze P, Karas H, Knüppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Research* 1995, **24(1):**238-241.
30. Sandelin A, Alkema Engström WP, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Research* 2004, **32:**D91-D94.
31. Huerta AM, Salgado H, Thieffry D, Collado-Vides J: **RegulonDB: a database on transcriptional regulation in *Escherichia coli*.** *Nucleic Acids Research* 1997, **26(1):**55-59.
32. **NCBI** [http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene]
33. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11(12):**4241-4257.
34. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431(7004):**99-104.
35. Jorgensen P, Rupes I, Sharom JR, Schneper L, Broach JR, Tyers M: **A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size.** *Genes Dev* 2004, **18(20):**2491-2505.
36. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2:**65-73.
37. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in Saccharomyces cerevisiae.** *Science* 2002, **298:**799-804.
38. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409:**533-538.
39. Wu GZ, Shi QM, Niu Y, Xing MQ, Xue HW: **Shanghai RAPESEED Database: a resource for functional genomics studies of seed development and fatty acid metabolism of Brassica.** *Nucleic Acids Res* 2008, **36:**D1044-147.
40. **TAIR** [http://www.arabidopsis.org/]
41. Mu J, Tan H, Zheng Q, Fu F, Liang Y, Zhang J, Yang X, Wang T, Chong K, Wang XJ, Zuo J: **LEAFY COTYLEDON1 is a key regulator of fatty acid biosynthesis in Arabidopsis.** *Plant Physiology* 2008, **148:**1042-1054.

42. Braybrook SA, Stone SL, Park S, Bui AQ, Le BH, Fischer RL, Goldberg RB, Harada JJ: **Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis.** *PNAS* 2006, **9(103):**3468-3473.
43. Baud S, Mendoza MS, To A, Harscoët E, Lepiniec L, Dubreucq B: **WRINKLED1 specifies the regulation action of LEAFY COTYLEDON2 towards fatty acid metabolism during seed maturation in Arabidopsis.** *The Plant Journal* 2007, **50:**825-838.
44. Tchagang AB, Shearer H, Phan S, Bérubé H, Famili FA, Fobert P, Pan Y: **Towards a temporal modeling of the genetic network controlling systemic acquired resistance in *Arabidopsis thaliana*.** *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2010), 2-5 May 2010; Montreal, Canada* .

doi: 10.1186/1471-2105-11-229
Cite this article as: Tchagang *et al.*, GOAL: A software tool for assessing biological significance of genes groups *BMC Bioinformatics* 2010, **11**:229