Gene Regulatory Networks from Gene Ontology

Wenting Liu[1]*, Kuiyu Chang[1], Jie Zheng[1,2],
Jain Divya[1], Jung-Jae Kim[1], and Jagath C. Rajapakse[1,3,4]

[1]Bioinformatics Research Center, School of Computer Engineering, Nanyang Technological University, Singapore. [2]Genome Institute of Singapore, A*STAR (Agency for Science, Technology, and Research), Biopolis, Singapore.
[3]Singapore-MIT Alliance, Singapore.
[4]Department of Biological Engineering, Massachusetts Institute of Technology, USA.
wliu7@ntu.edu.sg, kuiyu.chang@pmail.ntu.edu.sg, ZhengJie@ntu.edu.sg
divyajain.30@gmail.com, {jungjae.kim, asjagath}@ntu.edu.sg

Abstract. Gene Ontology (GO) provides a controlled vocabulary and hierarchy of terms to facilitate the annotation of gene functions and molecular attributes. Given a set of genes, a Gene Ontology Network (GON) can be constructed from the corresponding GO annotations and semantic relations among GO terms. Transitive rules can be applied to GO semantic relations to infer transitive regulations among genes. Using information content as a measure of functional specificity, a shortest regulatory path detection algorithm is developed to identify transitive regulations in GON. Since direct regulations may be overlooked during the detection of gene regulations, gene functional similarities deduced from GO terms are used to detect direct gene regulations. Both direct and transitive regulations are then used to construct a Gene Regulatory Network (GRN). The proposed approach is evaluated on seven E.coli sub-networks extracted from an existing known GRN. Our approach was able to detect the GRN with 85.77% precision, 55.7% recall, and 66.26% F1-score averaged across all seven networks.

Keywords: gene ontology, gene regulatory network, transitive gene regulation, semantic similarity, functional similarity.

1 Introduction

Gene regulation denotes the cellular activity that arises when a set of genes interact with one another. Gene regulations can be organized into a gene regulatory network (GRN), which provides insights into complex biological processes. However, ground truths of biological regulatory networks are unknown in most cases, so building a GRN that is accurate and biologically plausible remains an open research problem in the field of functional genomics.

Gene Ontology (GO) provides a controlled vocabulary arranged in a hierarchy of terms to facilitate the annotation of gene functions and molecular attributes.

* to whom correspondence should be addressed.

GO has been widely used for validating functional genomics experiments [1], [2]. In this paper, we present a method to build GRN that captures both direct and transitive regulations based on GO. The semantic relations of these gene annotation terms provide some evidence for gene regulations. Applying the transitive rules of these semantic relations, we can also infer transitive gene regulations.

A transitive regulation is a regulation between two genes via one or more transitive genes in the absence of a direct regulation [3]. For example, suppose gene g_1 regulates gene g_2 directly and gene g_2 is related to gene g_3. Then since there is no direct regulation between g_1 and g_3, if we can infer a regulatory relation from g_1 to g_3, we say g_1 transitively regulates g_3 through transitive gene g_2. The vast majority of previous research has focused on finding direct regulations between genes, using co-expressions [4]. In our approach, we use the information content of GO terms to represent the functional specificity and information flow, thereby determining the most probable transitive regulations between genes. Zhou et al. [3] linked genes of the same biological pathway based on the transitive expression similarity among genes. They determine the transitive co-expression genes by applying shortest path analysis on large-scale yeast microarray expression data. Instead of finding the shortest path based on distance, we find the shortest path based not only on distance but also on GO regulatory relations, which gives more reliable transitive regulations, as shown in our experiments.

Semantic Similarity of gene annotations can provide some clues for direct gene regulations. Cheng et al. [5] and Kustra et al. [6] incorporate gene similarity scores from GO semantic similarity into gene expression data to cluster genes. Franke et al. [4] assumed functional interactions among similar genes if they share more GO terms, and incorporate microarray co-expression and protein-protein interaction data to construct a human gene network. Similarly, in order to detect direct regulations, we consider the functional similarity of genes based on their semantic similarity. To better estimate the functional similarity score of gene pairs, we modify the original term probabilities by taking into consideration the chances of a term being annotated to both genes. Since these two GO methods complement each other, we then propose a GO fusion method to combine both direct and transitive regulations to generate the final GRN.

2 Methods

2.1 Gene Ontology Networks

A gene ontology (GO) is a controlled and structured biological vocabulary of various gene terminologies and their inter-related functional characteristics. It describes how gene products behave in a cellular context. The ontology covers three domains: biological process (BP), molecular function (MF), and cellular component (CC). BP is a collection of molecular events, MF defines gene functions in the biological process, and CC describes gene locations within a cell. A gene is associated with GO terms that describe the properties of its products (i.e., proteins). In our approach, only BP and MF terms are used since the cellular component (CC) is not directly related to gene regulation.

GO terms and their semantic relations are represented as a directed acyclic graph (DAG) where vertices represent GO terms and direct edges the relations between GO terms. There are three defined semantic relations between GO terms: $is - a$ is used when one GO term is a subtype of another GO term, $part - of$ is used to represent part-whole relationship in the GO terms, and *regulate* is used when the occurrence of one biological process directly affects the manifestation of another process or quality [7]. Let $\mathcal{R} = \{is - a,\ part - of,\ regulate\}$ denote the set of ontological relations. The GO can thus be represented as a hierarchical directed acyclic graph, where each term is related to one or more terms in the same or different domain. The GO has three roots at the topmost level: BP, MF, and CC. Nodes/terms near the root of the DAG have broader functions and are hence shared by many genes; leaf nodes/terms on the other hand convey more specific biological functions.

GO annotation (GOA) is the process in which GO terms are annotated to gene products. GOA data can be readily obtained from the GO annotation database [8]. The GO hierarchical structure also allows annotators to assign properties to genes or gene products at different levels, depending on the availability of information about an entity. In general, when inferring information from a gene that is annotated by some hierarchical GO terms, biological functions at the lower levels should be chosen as an inference base due to its more specific and richer information content.

As such, a measure to filter the more informative GO terms at the lower levels is clearly needed. A term in the GOA hierarchy that occurs less frequently is considered to be more informative as it has a more specialized function. To capture this frequency-sensitive informativeness of GO terms, the information content of the node was measured with respect to their annotations by [9]. Specifically, the information content $I(t)$ of a GO term $t \in \mathcal{T}$ is given by

$$I(t) = -\log(p(t)) \tag{1}$$

where $p(t) = N(t)/N(root(t))$ and $root(t) \in \{BP, MF\}$ is the GO root of term t, $N(t)$ is the number of occurrences of term t in the given GOA data. The information content strongly correlates with the hierarchical depth of the term with respect to the GO root. If the GO term is less frequent, it is usually located at a deeper/lower level and therefore has a more specific function.

Consider a set $\mathcal{G} = \{i\}_{i=1}^{n}$ of n genes with gene i associated to GO term-set T_i. The total term-set $\mathcal{T}$ of all genes is given by $\mathcal{T} = \bigcup_{i=1}^{I} T_i$. Let $r(t, t') \in \mathcal{R}$ denote a GO relation between terms $t, t' \in \mathcal{T}$. All GO relations between terms in the term-set $\mathcal{T}$ are represented as $\mathcal{E} = \{r(t, t') : t, t' \in \mathcal{T}, \exists\ r(t, t') \in \mathcal{R}\}$. The pair $(\mathcal{T}, \mathcal{E})$ thus constitute the GO network (GON).

2.2 GRNs from GO Regulatory Paths

Recall that in GON, *regulate* denotes the occurrence of one biological process that directly affects the manifestation of another process or quality, e.g., process t regulates process t' means that if both processes occur, t always regulates t'.

Suppose gene i is annotated by GO term t and gene i' is annotated by GO term t'. If GO term t regulates term t', then we can infer that gene i might regulate i'. If this inference comes from more specific terms, gene i should regulate i' with high confidence. In our GRN inference procedure, we therefore choose the most reliable regulate inference path. There are very few direct *regulate* relations in the existing GO database, making it difficult to infer a GRN. As such, we propose to induce transitive *regulate* relations among GO terms, from which we infer a GRN based on both direct and transitive gene regulations derived from the GON.

Consider the transitivity rule:

$$\text{if } t_a \overset{r_1}{\to} t_b \text{ and } t_b \overset{r_2}{\to} t_c, \text{ then } t_a \overset{r_3}{\to} t_c \tag{2}$$

where $t_a, t_b, t_c \in \mathcal{T}$ and $r_1, r_2, r_3 \in \mathcal{R}$. Using rule deduction notation, the above transitive relation can be written as $r_3 = r_1 \wedge r_2$. According to GO database [1], the following transitivity rules exist among the GO terms. For any $r \in \mathcal{R} = \{is - a,\ part - of,\ regulates\}$, the following four transitivity relations are valid.

$$r = r \wedge is - a \tag{3}$$

$$r = is - a \wedge r \tag{4}$$

$$regulates = regulates \wedge part - of \tag{5}$$

$$part - of = part - of \wedge part - of \tag{6}$$

Consider a path $(t_j)_{j=0}^{J}$ in GON where t_0 and t_J denote the source and destination terms, respectively; and $r(t_j, t_{j+1})$ is the parent-child relation between parent term t_j and child term t_{j+1}. Using parent-child relations $\{\ r(t_j, t_{j+1})\ \}_{j=1}^{J-1}$, each term t_j can induce a relation from the source term. Denote path $\pi_J = (t_j)_{j=0}^{J}$, and let $r(\pi_J) = r(t_0, t_J)$ denote the inferred relation along path π_J by applying transitive rules to parent-child relations, we have

$$r(\pi_j) = r(\pi_{j-1}) \wedge r(t_{j-1}, t_j). \tag{7}$$

We then assign a confidence score function $\sigma(\pi)$ for each inferred path π by considering both the number of steps and the information content of the terms along the inferred paths. The confidence score $\sigma(\pi)$ should give preference to paths with fewer inference steps and more informative terms, defined as

$$\sigma(\pi_j) = \sigma(\pi_{j-1}) + \varDelta_{r(t_{j-1}, t_j)}(t_{j-1}, t_j) \tag{8}$$

where $\sigma(\pi_j)$ is the score assigned to the inferred path π_j from source t_0 to term $t_j \in \mathcal{T}$ and $\varDelta_{r(t,t')}(t, t')$ is the score assigned to relation $r(t, t') \in \mathcal{R}$ between terms $t, t' \in \mathcal{T}$.

The cost for deducing a relation between two terms should facilitate the selection of the most informative inferred path. The semantic similarity of GO terms based on their information content, i.e., Lin's semantic similarity measure

[1] `http://www.geneontology.org/GO.ontology.relations.shtml`

[10] and Jiang's semantic distance [11], can be used to define the cost of deducing a relation between two terms. For example, $\mathit{\Delta}_{r(t,t')}(t,t') = 1 - S(t,t')$, where $S(t,t')$ is the semantic similarity of terms t,t'. Jiang's semantic distance can also be directly used as the cost $\mathit{\Delta}_{r(t,t')}(t,t')$.

When there exists no relation $r(t,t')$, the cost is $\mathit{\Delta}_{r(t,t')}(t,t') = \infty$, eliminating empty relations. Thus, the path with the minimum score σ_π is the path inferred collectively using the fewest number of steps and along the most informative terms, i.e., the most reliable inferred path. Dijkstra's shortest path algorithm can be used to find the shortest inferred path. If an inferred path ends with the deduced relation *regulate* at the destination, then there is a regulatory path (RP) between the source and the destination.

We propose an algorithm to detect the most reliable RP between two terms $s,d \in \mathcal{T}$ by using the deduced scores in Dijkstra's shortest path algorithm [12]. For each node $v \in \mathcal{T}$, we use an indicator vector to represent the deduced relations r_v, and $\sigma(r_v,v)$ is a matrix to record the current minimum distance/cost to deduce r_v at term v along path $\pi(r_v,v)$.

Initially, the source term is assigned an $is-a$ relation as it does not change the first transitive relation. Subsequently for each $R \in \mathcal{R}$, S_R denotes the unvisited node set with current relation R, and we iteratively choose $u^* = \arg\min_{u\in S_R}\{\sigma(R,u)\}$ as the starting node of the following iteration. At each inference step, if $\sigma(r_{u^*},u^*) + \mathit{\Delta}_{r(u^*,v)}(u^*,v) < \sigma(r_v,v)$, we update the current deduced score for the three relations by $r_v = r_{u^*} \wedge r(u^*,v)$, $\sigma(r_v,v) = \sigma(r_{u^*},u^*) + \mathit{\Delta}_{r(u^*,v)}(u^*,v)$. The iteration stops when all S_R are empty. Upon termination, if $\sigma(\mathit{regulate},d) < +\infty$, $\pi(\mathit{regulate},d)$ is the most-reliable RP from term s to d, otherwise no RP from term s to d exists. Given a source term and target term, Algorithm 1 finds the most reliable RP.

Algorithm 1 Finding the most-reliable Regulatory Path (RP) between two GO terms

```
Step 0. Given source term $s$, target term $d$, term set $\mathcal{T}$, and inference cost matrix $\mathit{\Delta}$
Step 1. Set $r_s(is-a) = 1$; $\forall u \in \mathcal{T}, R \in \mathcal{R}$. Set $S_R = \mathcal{T}$, $\pi(R,u) = \{u\}$, $\sigma(R,u) = +\infty$ except $\sigma(is-a,s) = 0$
Step 2. For $R \in \mathcal{R}$, If $S_R \neq \{\}$
        Choose $u^* = \arg\min_{u\in S_R}\{\sigma(R,u)\}$, set $S_R = S_R \backslash \{u^*\}$
        If $\sigma(R,u^*) \neq +\infty$: for each $(u^*,v) \in \mathcal{E}$, if $\sigma(r_{u^*},u^*) + \mathit{\Delta}_{r(u^*,v)}(u^*,v) < \sigma(r_v,v)$
            Update $r_v = r_{u^*} \wedge r(u^*,v)$, $\sigma(r_v,v) = \sigma(r_{u^*},u^*) + \mathit{\Delta}_{r(u^*,v)}(u^*,v)$, $\pi(r_v,v) = \{u^*\} \cup \pi(r_v,v)$
Step 3. If $\sigma(\mathit{regulate},d) < +\infty$, return $\pi(\mathit{regulate},d)$ as the most-reliable RP
```

Up till now, we have only determined the most reliable RP between GO terms. For genes i,i' with GO annotations $T_i, T_{i'}$ respectively, if there exists a RP from $t \in T_i$ to $t' \in T_{i'}$, we can infer that gene i regulates gene i'. The confidence score for this inferred path is assigned by the minimum regulatory path score of all RPs (if any): $C(i,i') = \min\{\sigma(\pi(t,t'))|r(\pi(t,t')) = \mathit{regulate}\}$. Then we can construct a GRN with confidence score based on direct and transitive *regulate* relations among the GO terms.

2.3 Complementary GRNs from Functional Similarity

According to the transitive rules, if no *regulate* path exists between two GO terms, then no *regulate* relation can be deduced. As a result, the Regulatory Path method cannot infer a GRN when there exists few *regulate* relations among the GO terms in the GOA data.

Since Gene Ontology (GO) provides a standard vocabulary of functional terms and allows for coherent annotation of gene products, gene products are functionally similar if they have comparable molecular functions and are involved in similar biological processes. Hence, the more similar genes are, the more likely they belong to the same biological pathway, which involves gene interactions/regulations. We can thus assess the functional similarity of gene products by comparing sets of GO terms, and then recover the direct regulations missing from the GO Regulate Paths method based only on genetic functional similarity.

Functional Similarity of Genes based on Semantic Similarity of GO annotations Semantic similarity has been previously proposed to compare concepts within an ontology. It can evaluate the specificity of a GO term's underlying concept in a given GO annotation. There are three popular semantic similarity measures: Resnik similarity [13] measures the semantic similarity of two terms via the information content of their lowest common ancestors (LCA); Lin's similarity [10] assesses how close the terms are to their LCA, but it does not take into account the level of detail of the LCA; simRel [14] combined the semantic similarity of Lin and Resnik, and it takes into account how close terms are to their LCA as well as how detailed the LCA is, i.e., it distinguishes between generic and specific terms. For each term $t \in \mathcal{T}$, let $p(t)$ be the probability of finding t's descendents in the GO annotation database. If t and t' are two terms and $a\,(t,t')$ represents the set of parent terms shared by both t and t', then

$$p(LCA(t,t')) = \min_{t^* \in a(t,t')} p\,(t^*)\,, \tag{9}$$

and the three similarity measures are listed as follows.

$$Sim_{Resnik}\,(t,t') = -log\,(p(LCA(t,t'))) \tag{10}$$

$$Sim_{Lin}\,(t,t') = \frac{2 \times \log\,(p(LCA(t,t')))}{\log\,(p\,(t)) + \log\,(p\,(t'))} \tag{11}$$

$$Sim_{Rel}\,(t,t') = \frac{2 \times \log\,(p(LCA(t,t')))}{\log\,(p\,(t)) + \log\,(p\,(t'))}[1 - p(LCA(t,t'))] \tag{12}$$

In fact, simRel reduces to Lin's measure when $p(LCA(t,t'))$ is very small, i.e., $[1 - p(LCA(t,t'))]$ approaches 1. Thus, in our experiments, we consider both

Lin's and Resnik's measure. Specifically, we consider two genes to be similar if and only if both measures yield high scores.

Gene products annotated with GO terms can be compared using the aforementioned semantic similarity measures. Let GOscore be the measure of functional similarity between two genes with respect to either their biological process (BPscore) or molecular function (MFscore). Each gene pair receives two similarity values, one for each ontology root. [15] defined the functional similarity between two genes i and i', with annotated GO term set T_i and $T_{i'}$, respectively, as the *average* inter-set similarity of terms in T_i and $T_{i'}$, as follows.

$$GOscore_{avg}\left(i,i'\right)=\frac{1}{\left|T_i\right|\left|T_{i'}\right|}\sum_{t\in T_i,t'\in T_{i'}} Sim\left(t,t'\right) \tag{13}$$

The *maximum* similarity measure is also computed as an upper bound, as follows.

$$GOscore_{max}\left(i,i'\right)=\max_{t\in T_i,t'\in T_{i'}} Sim\left(t,t'\right) \tag{14}$$

Finally, the funSim score is calculated from the BPscore and MFscore of a pair of gene products as follows,

$$funSim\left(i,i'\right)=\frac{1}{2}[(\frac{BPscore\left(i,i'\right)}{\max(BPscore)})^2+(\frac{MFscore\left(i,i'\right)}{\max(MFscore)})^2] \tag{15}$$

where max(BPscore) and max(MFscore) denote the maximum score for biological process and molecular function, respectively.

Modified GO Term Probabilities In the previous section, GOscores are defined by treating each term equally in the semantic similarity computation. That is, the semantic similarities are defined based on the term probabilities $p(t) = N(t)/N(root(t))$. However, this ignores the hierarchical structure of GO because $N(root(t))$ is in fact the number of genes assigned by $root(t) \in \{BP, MF\}$, i.e., two GO trees of different sizes are involved. $N(t)$ is the number of genes annotated to term t, thus, the definition of $p(t)$ is in fact the distribution of term t conditioned on a specific GOA data instead of all GOA.

Consider the case of two genes in a GO term list where some terms are commonly assigned to two genes, but some are assigned to only one gene. Clearly, the two terms should have different term probabilities. To account for this imbalance, we model the term probability for three differentoutcomes as follows: 1) term annotates both genes, 2) term annotates only one gene, and 3) term annotates none of the two genes. Given a term t and two genes denoted by $m = N(root(t))$ and $n = N(t)$, with joint probability $p(n, m) = p(t) = n/m$, the probability of term t annotated to (i) both genes is $p(n, m, k = 2) = p(n, m) \times \frac{\binom{m}{2}}{\binom{n}{2}}$; (ii) only one gene is $p(n, m, k = 1) = p(n, m) \times \frac{\binom{m}{1}\binom{n-m}{1}}{\binom{n}{2}}$; (iii) neither of the two genes is the same as the background distribution , i.e., $p(n, m, k = 0) = p(n, m) \times \frac{\binom{n-m}{2}}{\binom{n}{2}}$.

Let us consider an example to illustrate the discriminatory power of the modified term probabilities. Given a term, if it is assigned to both genes, its probability is $p(n, m, k = 2) = \frac{m^2 \times (m-1)}{n^2 \times (n-1)}$; if it is assigned to only one gene, its probability is $p(n, m, k = 1) = \frac{2m^2 \times (n-m)}{n^2 \times (n-1)}$; otherwise, its probability is $p(n, m, k = 0) = \frac{m \times (n-m) \times (n-m-1)}{n^2 \times (n-1)}$. For a specific term, m is always significantly smaller than n, thus, $p(n, m, k = 2) << p(n, m, k = 1) << p(n, m, k = 0)$. Similarly, their information content $-log(p(n, m, k = 2)) >> -log(p(n, m, k = 1)) >> -log(p(n, m, k = 0))$. In other words, the three outcomes represent three distinct levels of information content.

Note that the definition $p(n, m, k)$ also considers the prior probability, hence, our modified term probabilities is consistent with and improves the original definition. With the modified term probabilities, we can then use the semantic similarity definition and GOscore computation method in Section 2.3 to compute the functional similarity of a gene pair.

Deriving GRN from Gene Functional Similarity We next propose a method to build a GRN from the computed gene functional similarity scores of all applicable pairs.

If Lin's average biological similarity between genes exceeds a threshold θ_1, i.e., $funSim_{Lin,GOavg}(i, i') \geq \theta_1$, and Resnik's maximum biological similarity also exceeds some threshold θ_2, i.e., $funSim_{Resnik,GOmax}(i, i') \geq \theta_2$, then we say that gene i and i' are *functionally similar*, and there is a possible regulation between gene i and i'.

The GRN is constructed using the derived gene regulations. In our experiments, we exhaustively evaluated all combinations of semantic and GOscore measures to find the one that gives the best F1-score for GRN.

Fused GRN from Functional Similarity and RP Since the GRN from the RP and functional similarity methods are complementary, we propose a method to fuse the two derived regulations into a GRN as follows.

For all pairs of gene i, i', (i) if there exists transtive gene regulation from gene i to i' detected by GO RP, then gene i regulate i', or(ii) if gene i, i' are *functionally similar*, then there exists a gene regulation between gene i and i'.

3 Results

We evaluate our GO-inferred GRN against benchmark GRNs from GeneNetWeaver (GNW) [16][2]. Specifically, we use the E.coli GRN. Since the complete E.coli network from GNW contains many genes that have no corresponding GO annotations, we extract seven sub-networks from it, which are listed in Table 1. GO terms and relations corresponding to the genes in the networks were obtained from files associated with the GO annotation database [3]. The corresponding informative GO terms related to the target genes were also selected from the gene association files. Only GO terms involved in the molecular function and biological process are considered. To construct a more reliable GRN, we choose only

[2] `http://gnw.sourceforge.net/`

[3] `http://www.geneontology.org/GO.downloads.annotations.shtml`

Table 1. E.coli sub-networks and their GO relations. Each network has 25 genes.

	Net1	Net2	Net3	Net4	Net5	Net6	Net7
No. of edges	18	15	24	11	15	19	29
regulate	2	1	1	1	1	2	4
part-of	3	1	0	2	2	2	2
is-a	26	26	9	28	33	18	36
No. of terms	96	93	44	108	110	82	90

GO terms with information content $I(t) \geq \theta_I$, i.e., above a certain threshold θ_I; a threshold $\theta_I = -\log 0.25$ was used in the experiments. Each GO annotation is classified into one of 5 descending order of quality categories: experimental, computational, author statement, curator statement, and automatic. Annotations derived through direct experiments are deemed higher quality compared to others [17]. We only consider GO terms with the top two quality levels: computational and experimental.

In the following sections, we evaluate the performance of the three GO-inferred methods on the seven networks. The evaluation measurements include accuracy, precision, recall, F1-score, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) numbers. Average results over the seven sub-networks are denoted by "Avg". To compare the three GO methods, namely GO Regulatory Path (denoted by "RegPath"), GO Functional Similarity (denoted by "FunSim"), and GO Fusion (denoted by "Fusion"), we list their GRN prediction performance on the seven target networks in Table 2.

From the "RegPath" results in Table 2, we see that our GO RP approach achieved very high average precision of 87.43%, with corresponding F1-score of 62.93%. In general, very few false positive (FP) edges were extracted, with two networks (Net4 and Net5) consistently having zero FPs, and the remaining five networks registering less than three FPs. To summarize, we have discovered seven new gene regulations via our GO RP method, three self-regulations on gene "argP", "fadR" and "flhC"; four gene pairs: "argP" regulate "gyrA"; "argP" regulate "polA"; "dnaA" regulate "dinB"; and "flhC" regulate "flhD". We tried to look up evidences for these seven gene regulation pairs from the MEDLINE database [4], but was unable to find any evidence. We may eventually need experts in the field to confirm or reject these FPs. Moreover, the FPs could have been generated due to (i) human errors in the GO database: incorrect gene annotations or GO relations; (ii) incompleteness of the target network.

One limitation of the RP method lies in its poor recall, which averaged only 51.41%. This shows that the GO RP method could not detect enough gene regulations in the target GRN, which can be due to (i) incomplete gene GO annotations, and missing GO terms in GON; (ii) lack of annotated *regulate* relations among the GO terms; (iii) incomplete $is - a$ or $part - of$ GO relations; (iv) inaccuracy or incompleteness of the target ground truth network itself. Clearly, the

[4] `http://www.ncbi.nlm.nih.gov/`

incomplete GO information will bound the accuracy of our GO regulatory path method, which originally motivated us to use functional similarity to further improve our GRN inference from GO.

Table 2. Performance of the three proposed GO methods

		TP	FP	FN	TN	Pre.(%)	Rec.(%)	F1(%)	Acc.(%)
RegPath	Net1	10	3	8	604	76.92	55.56	64.52	98.24
	Net2	6	1	9	609	85.71	40.00	54.55	98.40
	Net3	20	1	4	600	**95.24**	**83.33**	**88.89**	99.2
	Net4	5	0	6	614	**100**	45.45	62.5	99.04
	Net5	5	0	10	610	**100**	33.33	50	98.4
	Net6	5	3	14	603	62.5	26.32	37.04	97.28
	Net7	22	2	7	594	**91.67**	**75.86**	**83.02**	98.56
	Avg	10.43	1.43	8.29	604.86	**87.43**	**51.41**	**62.93**	**98.45**
O:FunSim	Net1	7	51	11	556	12.07	38.89	**18.42**	90.08
	Net2	2	2	13	608	**50.00**	13.33	21.05	97.6
	Net3	10	55	14	546	15.38	41.67	22.47	88.96
	Net4	4	35	7	579	10.26	36.36	16	93.28
	Net5	2	4	13	606	33.33	13.33	19.05	97.28
	Net6	16	270	**3**	336	5.594	**84.21**	10.49	56.32
	Net7	19	330	10	266	5.444	**65.52**	10.05	45.6
	Avg	8.57	106.71	10.14	499.57	18.87	**41.90**	16.79	81.30
M:FunSim	Net1	10	3	8	604	76.92	55.56	**64.52**	**98.24**
	Net2	6	1	9	609	85.71	40.00	54.55	**98.4**
	Net3	20	1	4	600	**95.24**	**83.33**	**88.89**	**99.2**
	Net4	5	0	6	614	**100.00**	45.45	62.5	**99.04**
	Net5	5	0	10	610	**100.00**	33.33	50.00	**98.4**
	Net6	5	3	14	603	62.5	26.32	37.04	97.28
	Net7	18	2	11	594	**90.00**	**62.07**	**73.47**	97.92
	Avg	9.86	1.43	8.86	604.86	**87.20**	**49.44**	**61.57**	**98.35**
Fusion	Net1	11	3	7	604	78.57	**61.11**	68.75	**98.4**
	Net2	6	1	9	609	85.71	40.00	54.55	**98.4**
	Net3	20	1	4	600	95.24	**83.33**	**88.89**	**99.2**
	Net4	6	0	5	614	100	54.55	70.59	**99.2**
	Net5	6	1	9	609	85.71	40.00	54.55	**98.4**
	Net6	6	3	13	603	66.67	31.58	42.86	**97.44**
	Net7	23	3	6	593	88.46	**79.31**	**83.64**	**98.56**
	Avg	11.14	1.71	7.57	604.57	**85.77**	**55.70**	**66.26**	**98.51**

The evaluations of GRN predicted from GO Functional Similarity using the original term probabilities (denoted by “O:FunSim”) and modified term probabilities (denoted by and “M:FunSim”) are shown in Table 2. The “O:FunSim” method resembles existing works which extract gene interactions based on gene similarity from GO Semantic Similarity [5, 4, 6]. It can be seen that “M:FunSim” outperforms “O:FunSim” notably on the averaged precision, recall, F1-score, and accuracy. Clearly, the modified term probabilities are extremely effective in capturing the functional similarity of genes.

Due to the complementary strengths of last two methods, they can be fused to capture more information from the GO databases. From the Fusion method in Table 2, we see that the GO Fusion approach gave the best performance in terms of F1 measure and accuracy. The overall low recall of the GO Fusion method is due to the incompleteness of the GO database and the target network. The extremely high TNs (reflected in the accuracy rates of 98% or higher) shows that our GO method was able to filter the vast majority of negative edges. In fact, our approach generally yielded very low false positive rates. As a result, very high precision rates averaging 85.77% were achieved by the Fusion method as shown in Table 2. This conservative behaviour is desirable because gene regulation is hard to validate in general and thus a GRN should have as high precision as possible.

4 Conclusion

We proposed a method to detect both direct and transitive regulations between genes by using their corresponding GO annotations and relations. By developing a novel shortest path detection algorithm, we detected the most likely regulatory paths from GONs. Experiments show that transitive regulations play an important role in GRN and their detection significantly improves the accuracy of the generated GRN. We show that GO can be used effectively to detect transitive regulations.

Due to the incomplete information of the source GO database, the GRN generated from the GO Regulate Path method may overlook some important direct regulations. Inspired by the fact that gene regulations occur between functionally similar genes, we propose the GO FunSim method to detect direct regulations. Gene function similarity scores are computed from the semantic similarities of their corresponding GO terms, using their occurrence probabilities. We then modified the term occurrence probabilities to account for GO term imbalance, e.g., the likelihood of a term being assigned to each, both, or neither of the two genes. Experimental results show that our GO FunSim method based on the modified term probabilities are extremely adept at capturing pairwise gene function similarities.

Lastly, we proposed a simple fusion method to combine the results of the proposed FunSim and Regulate Path methods to generate a fused GRN. Experiments show that our GO Fusion method yielded the best GRN in terms of F-score.

The errors may arise from the following: (i) the incompleteness of the target networks; (ii) the incompleteness of the GO databases; (iii) the erroneous annotations of GO database; (iv) GO allows us to annotate genes and their products with a limited set of attributes, its scope is limited to the three domains, which is not comprehensive. Hence, the GRN we extracted from GO are in fact based upon partial evidence provided by the current GO. The false negatives of the networks could be further reduced by fusing the GO generated GRN with additional data sources such as wet-lab data. One extension to this work is to identify ontology terms that are specific to the pathways under consideration, e.g., terms related to cell-cycle functions in our experiments. The GRN developed by our method could be useful for validation of networks built by other experimental or computational approaches.

Acknowledgement This research was supported in part by Singapore's Ministry of Education Academic Research Tier 2 fund MOE2010-T2-1-056 and MOE AcRF Tier 2 grant of ARC9/11.

References

1. Steuer, R., Humburg, P., Selbig, J.: Validation and functional annotation of expression-based clusters based on gene ontology. BMC Bioinformatics (2006)
2. Mundra, P.A., Rajapakse, J.C.: Svm-rfe with mrmr filter for gene selection. IEEE Transactions on Nanobiosciences **9** (2010)
3. Zhou, X.H., Kao, M.J., Wong, W.H.: Transitive functional annotation by shortest-path analysis of gene expression data. In: PNAS. Volume 99. (2002) 12783–12788
4. Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M., Wijmenga, C.: Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet. **78** (2006) 10111025
5. Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D., Siani-Rose, M.: A knowledge-based clustering algorithm driven by gene ontology. J Biopharm Stat. **14** (2004) 687–700
6. Kustra, R., Zagdaski, A.: Data-fusion in clustering microarray data: Balancing discovery and interpretability. TCBB **7** (2010) 59–63
7. Ashburner, M., Ball, C.A., Blake, J.A.: Gene ontology: tool for the unification of biology. Nat Genet **25** (2000) 25–29
8. Barrell, D., Dimmer, E., Huntley, R.P.: The goa database in 2009 - an integrated gene ontology annotation resource. Nucleic Acids Research **37** (2009) 396 – 403
9. Alterovitz, G., Xiang, M., Mohan, M.: Go pad: the gene ontology partition database. Nucl. Acids Res. **35** (2007) 322–327
10. Lin, D.: An information theoretic definition in similarity. In: ICML. (1998) 266–304
11. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: International Conference on Research in Computational Linguistics. (1998)
12. Johnson, D.B.: A note on dijkstra's shortest path algorithm. JACM (1973)
13. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. In: J Artif Intell Res. Volume 11. (1999) 95–130
14. Schlicker, A., Domingues, F.S., Rahnenfuhrer, J., Lengauer, T.: A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics (2006)
15. Lord, P., Stevens, R., Brass, A., C.A.Goble: Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics **19** (2003) 1275–1283
16. Schaffter, T., Marbach, D., Floreano, D.: GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. Bioinformatics **27** (2011) 2263–2270
17. Rhee, S.Y., Wood, V., Dolinski, K.: Use and misuse of the gene ontology annotations. In: Nat Rev Genet. Volume 9. (2008) 509–515