

Data and text mining

A framework of integrating gene relations from heterogeneous data sources: an experiment on *Arabidopsis thaliana*

Jiexun Li^{1,*}, Xin Li¹, Hua Su¹, Hsinchun Chen¹ and David W. Galbraith²¹Department of Management Information Systems and ²Department of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

Received on April 19, 2006; revised on June 16, 2006; accepted on June 22, 2006

Advance Access publication July 4, 2006

Associate Editor: Martin Bishop

ABSTRACT

One of the most important goals of biological investigation is to uncover gene functional relations. In this study we propose a framework for extraction and integration of gene functional relations from diverse biological data sources, including gene expression data, biological literature and genomic sequence information. We introduce a two-layered Bayesian network approach to integrate relations from multiple sources into a genome-wide functional network. An experimental study was conducted on a test-bed of *Arabidopsis thaliana*. Evaluation of the integrated network demonstrated that relation integration could improve the reliability of relations by combining evidence from different data sources. Domain expert judgments on the gene functional clusters in the network confirmed the validity of our approach for relation integration and network inference.

Contact: jiexun@eller.arizona.edu**1 INTRODUCTION**

Uncovering gene functional relations is one of the major goals of biological studies. They can be extracted from a variety of sources. Many studies have developed or adopted machine learning techniques to exploit special characteristics of different biological data. However, most studies only focus on one single type of data sources (Friedman *et al.*, 2000; Jenssen *et al.*, 2001). Each single data source often can only reveal a certain perspective of the underlying complex biological mechanism. Furthermore, many single-source-based approaches are criticized for their low reliability as well as limited coverage of genes and relations.

Integrating evidence from multiple data sources is believed to provide a means to overcome these drawbacks, and thereby benefit studies of genomic functions. By combining multiple forms of evidence, we expect to provide a complete genome-wide functional network and more accurate inferences of unknown gene functions.

With specific interests and experiences in plant science research, we focus on relation extraction and integration for *Arabidopsis thaliana* (or *Arabidopsis*). *Arabidopsis* is one of the model organisms for studying plant genetics and development. The genome of *Arabidopsis*, the first to be sequenced in higher plants, is believed to comprise at least 30 700 genes. Of these genes, the function of approximately one-third (9194) remain unknown according to the functional Gene Ontology (GO) category listed by the *Arabidopsis* Information Resource (TAIR). Of the remainder, a large proportion

lack complete or adequate functional annotation. We are aimed at constructing a genome-wide functional network of *Arabidopsis* by integrating relations extracted from diverse data sources.

2 LITERATURE REVIEW

Several approaches have been proposed to integrate gene functional relations extracted from various data sources. We survey the previous related studies of relation integration from three perspectives: data sources, analytical techniques and integration methods.

2.1 Data sources

In general, previous studies have involved examination of three types of genomic data. The first type of data source comprises experimental measurements of genes or proteins, such as microarray-based gene expression data (Jansen *et al.*, 2003), yeast two-hybrid experimental data (Uetz *et al.*, 2000) and *in vivo* pull down experimental data (Gavin *et al.*, 2002). These experimental datasets contain rich biological information but oftentimes also involve noises and errors. The second type of data consists of various genomic and proteomic features, such as gene sequence (Marcotte *et al.*, 1999), gene localization (Yanai and DeLisi, 2002) and metabolic pathway (Marcotte *et al.*, 1999). The third type of data is the body of biological knowledge, which contains mostly known and validated gene/protein relations. Some of these known relations are stored in structured format, e.g. Database of Interacting Proteins (DIP) and Munich Information Center for Protein Sequences (MIPS) (Marcotte *et al.*, 1999). There are also other known relations that are represented in unstructured textual format, e.g. Gene Ontology (GO) annotation (Jansen *et al.*, 2003) and biological literature (von Mering *et al.*, 2005). Automatic extraction of relations from text in semi- or non-structural format is a non-trivial task.

2.2 Analytical techniques

The data sources described above provide diverse insights of gene functions. Some functional interactions can be directly inferred from experimental results, such as yeast two-hybrid interactions (Uetz *et al.*, 2000) and *in vivo* pull down experiments (Gavin *et al.*, 2002). For most biological data sources, various analytical techniques must be applied to extract gene functional relations.

Many studies discover gene relations based on certain criteria of correlation or similarity between genes within a particular data source. For instance, from analysis of microarray data,

*To whom correspondence should be addressed.

comprising simultaneous measurements of the amounts of thousands of transcripts, the observed correlation of the expression levels of different genes can indicate co-expression and regulatory relations (von Mering *et al.*, 2005). Clustering methods, such as *K*-means, self-organizing maps (SOM) and hierarchical clustering, have been devised to identify the similarity of expression of different genes across multiple samples (Gasch and Eisen, 2002).

Co-occurrence analysis can also be applied to various data sources. These methods assume that the co-occurrence of two items in a certain source indicates their relationships. Phylogenetic profiling (PP) (Pellegrini *et al.*, 1999) and Rosetta Stone (RS or domain fusion) method (Marcotte *et al.*, 1999) are co-occurrence-based methods for gene sequence analysis. Some sources may not directly infer protein interactions *per se*, but they also contain information associated with the interactions. Marcotte *et al.* extracted protein pairs that catalyze sequential reactions in metabolic pathways (Marcotte *et al.*, 1999). Jansen *et al.* looked at whether two proteins are either both essential or both non-essential for survival (Jansen *et al.*, 2002, 2003). Co-occurrence analysis have also been applied to extract gene pairs from literature (von Mering *et al.*, 2005).

2.3 Integration methods

The combination of relations from different sources could provide a unified view of gene functional networks with large coverage and high reliability. Previous relation integration methods can be categorized into set-based methods and scoring-based methods.

A set-based method combines different relations using set operations such as intersection or union. Assuming links corroborated by multiple methods are of good quality, intersection can be used to achieve higher reliability (Marcotte *et al.*, 1999). However, intersection often leads to small fractions of relation sets. Alternatively, union can be used to achieve larger coverage of evidence but often lower reliability (Yanai and DeLisi, 2002). Since they treat each relation equally, the difference of relations in ‘quality’ is not taken into account in the integration process.

To overcome these pitfalls, more recent studies have used scoring-based methods for relation integration. These methods are often composed of two steps. First, relations extracted from different resources are scored based on a certain benchmark. Second, based on their quality scores, the relations can be integrated together using different methods. A quality score could be assigned to each relation subset (Jansen *et al.*, 2002; Troyanskaya *et al.*, 2003; von Mering *et al.*, 2005). For example, the subset profiling method (Jansen *et al.*, 2002) partitions relations from different sources into subsets based on their overlapping patterns and assigns a score to each subset. Subsets are sorted by the score in descending order and added into the network one by one until a good balance of coverage and reliability is achieved. Lee *et al.* assigned a score to each relation and integrated relations using a weighted sum scoring method (Lee *et al.*, 2004). Bayesian methods have been shown to be useful in integrating different types of evidence. Naïve Bayes methods assume conditional independence among evidence and are mainly used to integrate relations from unrelated or weakly related data sources (Jansen *et al.*, 2003; Troyanskaya *et al.*, 2003; von Mering *et al.*, 2003, 2005). Fully connected Bayesian networks (BNs) capture the interdependence and therefore could achieve a more accurate accommodation of correlated evidence (Jansen *et al.*,

2003; Troyanskaya *et al.*, 2003). However, full BNs often require higher computational costs.

2.4 Research gaps

In previous studies, various analytical techniques have been applied to integrate relations extracted from multiple data sources to improve the reliability and/or the coverage of gene functional networks. However, few studies have conducted a systematic evaluation of the relation integration method and the integrated network. Therefore, designing a framework for extracting and integrating relations from diverse biological sources and evaluating the integrated network remains a challenge. Furthermore, most related studies focused on yeast genes and proteins. For *Arabidopsis*, there has been no study that constructed a genome-wide network by integrating relations from diverse sources. This research gap motivated us to choose and focus on *Arabidopsis* in this study.

3 RESEARCH QUESTIONS

A successful integration of gene relations from heterogeneous data sources is intended to provide researchers with more insights about biological functions. We focus on five main research questions in this study. Q1. How can gene relations extracted by different techniques from diverse data sources be combined into a genome-wide functional network? Q2. Can relation integration improve the reliability of a gene functional network? Q3. How can we evaluate the reliability of the integrated network? Q4. How does each data source contribute to the integrated gene network in terms of the number of reliable relations? Q5. How can the integrated network help biological researchers identify gene functional clusters?

4 A FRAMEWORK OF GENE RELATION INTEGRATION

In this study, we propose a framework of integrating gene functional relations (Fig. 1).

4.1 Data sources and benchmarks

Our proposed framework involves three types of data sources, i.e. experimental data, biological knowledge and biological features. In particular, we mainly focus on gene expression data, biological literature and gene sequence information as the representatives for the three major data sources, respectively.

In our framework, benchmark sources of trusted relations are important to evaluate the relations and validate the integrated network. Benchmarks for this purpose should be as accurate as possible and are usually human-curated. An ideal benchmark set should be independent from the evidence sources, sufficiently large for reliable statistics, and free of systematic biases (Jansen *et al.*, 2003). Some benchmarks are used to score relations supported by different evidence in the relation integration process. Such benchmarks can be called ‘scoring benchmarks’. The others that are independent of the scoring benchmark are used as a gold standard to validate the correctness of the integrated network. We call these benchmarks as ‘evaluation benchmarks’.

4.2 Relation extraction

From each data source, e.g. gene expression data (*E*), biological literature (*L*) and genome sequence (*S*), multiple analytical techniques can be applied to extract functional relations. In Figure 1,

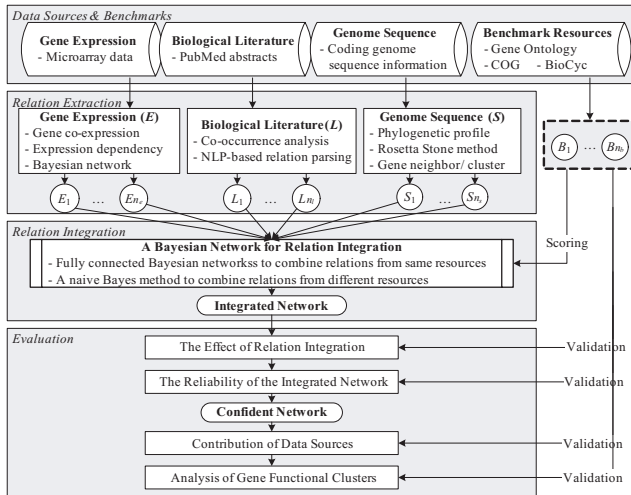


Fig. 1. A framework of gene functional relation integration. The framework contains four major modules: Data sources and benchmarks, Relation extraction, Relation integration and Evaluation. Gene functional relations are extracted from diverse data sources and integrated into a genome-wide network. Benchmark resources are used to score and validate the gene relations in relation integration and evaluation process.

(E_1, \dots, E_{n_e}) , (L_1, \dots, L_{n_l}) , and (S_1, \dots, S_{n_s}) denote the relations extracted by a certain technique from the three data sources, where E_i , L_i or S_i represents relations extracted by a particular technique from the corresponding data source; n_e , n_l and n_s denote the number of techniques applied to each data source, respectively. Each technique applied to a particular data source provides certain evidence of functional relations among genes and proteins.

Many analytical techniques can be used for relation extraction. For example, from gene expression data, correlation coefficient analysis (Jansen *et al.*, 2003; Lee *et al.*, 2004; von Mering *et al.*, 2003, 2005) and BN learning (Friedman *et al.*, 2000; Huang *et al.*, 2006) can be used to extract co-expression relationship between genes. Gene relations in the body of biological literature can be extracted by co-occurrence analysis (Jenssen *et al.*, 2001; Lee *et al.*, 2004; Marcotte *et al.*, 1999) and various NLP-based parsing techniques (McDonald *et al.*, 2004). For genome sequence information, various sequence analysis methods, such as PP, RS, gene cluster (GC) and gene neighbor (GN) analysis, are all options for relation extraction (Bowers *et al.*, 2004; Lee *et al.*, 2004).

4.3 Relation integration

Relations extracted by different analytical techniques from diverse data sources are of different quality and reliability. We propose a two-layered Bayesian network (BN) for integrating these relations into a genome-wide functional network. We chose a BN approach for its following advantages (Jansen *et al.*, 2003). BNs allow for the combining of heterogeneous data by converting them into a common probabilistic framework. They weigh evidence according to its reliability. They are readily interpretable as they present conditional probability relationships among different evidence. The structure of our proposed BN is shown in Figure 2.

In order to integrate relations extracted by different techniques from diverse data sources, we need a unified scoring scheme

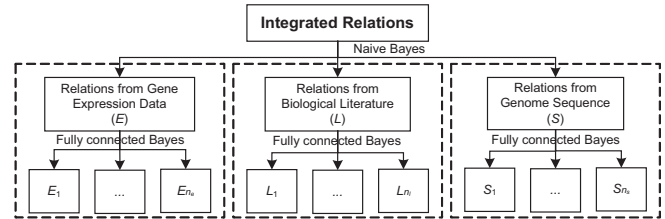


Fig. 2. A two-layered BN for relation integration. Relations from same data sources are integrated using a fully connected BN; and relations from different data sources are integrated using a naïve Bayes method.

to measure their quality. In previous literature, a confidence score has been used for this purpose (von Mering *et al.*, 2003, 2005). We adopt this score as the evaluation criterion for different evidence in this study. Consider evidence f expressed in binary terms (i.e. ‘present’ or ‘absent’). Each evidence f is evaluated for its ability to reconstruct known gene pathways by comparing its predictions to a common benchmark. A relation that matches a known one in the benchmark is called a true positive. This confidence score, $P(f)$, measures the percentage of true positives among the total number of relations support by evidence f . A greater score indicates a higher degree of reliability of the evidence.

As shown in Figure 2, the BN of relation integration is composed of two layers. At the bottom layer of the BN are gene functional relations extracted by different analytical techniques from each data source. Since these relations are from the same resource and provide correlated evidence, these multiple forms of evidence from the same data source are combined using fully connected BNs. Specifically, each relation extracted from a data source D is supported by a particular combination of evidence forms, $f = (f_1, \dots, f_{n_d})$, where n_d is the number of techniques used to extract relations from D . For each combination of evidence form f , we can assign it with a joint confidence score, $P(f) = P(f_1, \dots, f_{n_d})$, by benchmarking the supported relations against the common reference set.

Next, the second layer of the BN combines relations from different data sources into one integrated network. Under the assumption of independence among various data sources, relation integration at this layer is performed in a naïve Bayes fashion. This assumption of independence is valid because relations extracted from the same data source have been joined previously in the first layer. Given N independent forms of evidence (f_1, \dots, f_N) , the joint confidence score is derived as follows (von Mering *et al.*, 2003, 2005):

$$P(f_1, \dots, f_N) = 1 - \prod_{i=1}^N [1 - P(f_i)].$$

The joint confidence score is often higher than the individual sub-scores, indicating higher reliability of relations supported by multiple forms of evidence.

4.4 Evaluation

To examine its validity and to uncover the underlying biological information, we evaluate the integrated network from the following four perspectives.

- *The effect of relation integration.* We examine the effect of relation integration by comparing the integrated network against relations of each evidence form based on an evaluation benchmark. A successful integration should provide a network of relations with higher reliability than relations supported by individual evidence forms.
- *The reliability of the integrated network.* Each relation in the integrated network is assigned with a joint confidence score that measures the reliability of this relation. Based on the evaluation benchmark, we evaluate the reliability of the integrated network by analyzing the correlation between the true positive (TP) rate of top relations and the confidence score. A positive correlation between the TP rate and the confidence score can confirm that a higher score indicates higher reliability.
- *Contribution of data sources.* Furthermore, a threshold for reliable relations can be derived by comparing the evaluation benchmark against the scoring benchmark. Relations with a confidence score greater than this threshold are regarded high-quality relations while the others are regarded as noise and removed. We follow the same terminology used in (Lee et al., 2004) and call such a network of reliable relations a ‘confident network’. We evaluate the confident network by analyzing the distribution of gene functional relations from different data sources. Specifically, we analyze the relations that involve gene of unknown functions. Such relations are interesting to biological researchers because they may infer new functions of genes.
- *Analysis of gene functional clusters.* We analyze the network structure and identify gene clusters (GCs) by grouping genes according to their connectivity. Based on expert judgments, we match each GC to a specific biological function. Analysis of these gene functional clusters can validate the integrated network and provide researchers with more insights about their biological functions.

5 EXPERIMENTAL STUDY

In order to validate our proposed framework for relation integration, we conducted an experimental study on *A.thaliana* and evaluated the integrated functional network.

5.1 Test-bed

Our experimental study involves the three major types of data sources for gene functional relation extraction.

We chose two high-quality microarray series of *Arabidopsis* available at <http://www.weigelworld.org> suggested by a senior plant scientist as the test-bed of gene expression data. These two microarray series include experiments for development (referred as ‘dev’) and abiotic stress (referred as ‘abio’). These two datasets both measure the expression levels of 22 810 *Arabidopsis* genes in 237 and 298 samples, respectively.

The second data source is literature abstracts related to *Arabidopsis* from PubMed, the online portal of MedLine literature database. We used the MeSH (Medical Subject Headings) terms, ‘*Arabidopsis*’ and ‘*Arabidopsis* Proteins’, to create a sub-collection for *Arabidopsis*. By April 2005, we collected 10 548 *Arabidopsis*-related abstracts from PubMed.

For the data source of genome sequence information, we identified and chose the well-known online database named Prolinks

(<http://dip.doe-mbi.ucla.edu/pronav>), constructed by the University of California, Los Angeles (Bowers et al., 2004). This database spans 83 organisms including *Arabidopsis thaliana*. It combines results of four gene sequence analysis techniques in a single database and includes over one million links for *Arabidopsis*.

5.2 Benchmarks

In our proposed framework, benchmarks of known gene functional relations are needed to evaluate the relations and to validate the relation integration approach. Based on the suggestion of domain experts, two reference benchmarks, namely KOG and AraCyc, were selected in our study. The eukaryotic orthologous groups (KOGs) are within the Clusters of Orthologous Groups of proteins (COGs), which define the orthologous proteins among different species with gene functions assigned to 23 broad categories (Tatusov et al., 2003). Of the 25 749 annotated *Arabidopsis* proteins 53% are claimed to be clustered in KOG (Tatusov et al., 2003). The identification of any pair of proteins as belonging to the same KOG clusters serves as a benchmark for the evaluation of relations. This benchmark has also been used in previous studies such as (Bowers et al., 2004). The second benchmark, AraCyc, is an *Arabidopsis* pathway database as a part of the BioCyc projects (Mueller et al., 2003). It is generated from genome annotations, and is verified and corrected by human curators. By January 2005, AraCyc covered 186 genetic and metabolic pathways. About 53% of reactions in these pathways have enzymes/genes annotated to them. For AraCyc, each pair of two genes belonging to the same pathway will be regarded as a benchmark relation.

In total, we identified 13 329 genes and 328 493 relations from KOG, 1141 genes and 25 212 relations from AraCyc. In the relation integration process, we used KOG as the scoring benchmark to score relations supported by different evidence for its larger coverage of genes and relations. The AraCyc benchmark was then used to evaluate the integrated network.

5.3 Selected techniques for relation extraction

In this study, we selected the following techniques to extract gene/protein functional relations from the three data sources.

5.3.1 Gene expression data To extract gene relations from gene expression data, we adopted two different analytical techniques, correlation coefficient and the mutual information (MI), both indicating the strength of gene co-expression.

Pearson correlation coefficient has been commonly used in previous studies for gene co-expression analysis (Lee et al., 2004). A pair of genes with highly correlated expression values are believed to be co-expressed and therefore possibly functionally related. According to this measure, the correlation between a gene x and a gene y can be defined as follows:

$$r(x, y) = \frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_k (x_k - \bar{x})^2} \sqrt{\sum_k (y_k - \bar{y})^2}},$$

where x_k and y_k represent the expression level of x and y in the k -th sample; \bar{x} and \bar{y} are the means of expression of x and y over all samples, respectively.

In BN learning for microarray data analysis, MI can be used to measure the inter-dependency between genes (Huang et al., 2006).

The MI measure between gene x and gene y is defined as follows:

$$I(x, y) = \sum_{i,j} P(X_i, Y_j) \log \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)},$$

where X_i is the i -th expression level of gene x , Y_j is the j -th expression level of gene y ; $P(X_i)$ is the probability that the expression level of x equals X_i ; $P(Y_j)$ is the probability that the expression level of y equals Y_j and $P(X_i, Y_j)$ is the probability that the expression level of x equals X_i and the expression level of y equals Y_j .

5.3.2 Biological literature In this study we used two well-known text mining techniques to extract gene relations from the body of biological literature. Co-occurrence analysis is commonly applied in mining gene relations from literature (Jenssen *et al.*, 2001). It infers the existence of a certain functional relation between a pair of genes when they both appear within the same abstract. We extracted co-occurrence relations by searching genes that are cited in the same literature abstracts. In this study we only focus on binary co-occurrence relations from literature.

Previous studies show that fine-grained text analysis based on linguistic and semantic sentence parsing produces more precise gene relations compared with raw co-occurrence analysis. In this study we adopted the Arizona Relation Parser (McDonald *et al.*, 2004), a text mining tool developed by the University of Arizona, to fully parse meaningful gene relations from textual sentences in literature abstracts. This relation parser was reported to outperform other parsers in its use of a broad coverage syntax-semantic hybrid grammar (McDonald *et al.*, 2004). In order to reduce the ambiguities for biological entities and functional processes in the extracted relations, we created an aggregate lexicon specifically for *Arabidopsis* and used an aggregation module, called BioAggregate Tagger, to aggregate relations and to capture important contextual information (Marshall *et al.*, 2006).

5.3.3 Genome sequence In this study we included *Arabidopsis* protein functional relations extracted by four standard sequence analysis techniques (i.e. PP, RS, GN and GC) available in the Prolinks database (Bowers *et al.*, 2004). The PP method constructs profiles by analyzing the distribution of protein sequences across known genomes. Proteins with similar phylogenetic profiles are likely to participate in the same pathway. The RS method compares genomic sequence information across organisms, by searching for examples of protein coding sequences that represent the fusion of coding sequences that are separated in the genomes of other organisms. The fusion protein is termed the RS protein, in that it allows us to infer the functional linkage between two proteins. The GN method identifies protein pairs encoded in close proximity across multiple genomes. The GC or operon method identifies closely spaced genes, and assigns a probability P of observing a particular gap distance (or smaller), as judged by the collective set of inter-gene distances.

5.4 Relation extraction and integration

Genes are denoted by different symbols or identifiers in different data sources. Relation integration requires that all genes are denoted according to a common naming scheme. We mapped genes from different resources to their Entrez Gene identifiers. Relations involving genes that could not be mapped were eliminated. The numbers

Table 1. Relations extracted from different resources using different analytical techniques

Data resources	Analytical techniques	Number of relations
Gene expression	Correlation (CC)	250 600 078
	Mutual information (MI)	250 600 078
Biological literature	Co-occurrence (CO)	1291
	Relation parsing (RP)	222
Genome sequence	Phylogenetic profiling (PP)	132 637
	Rosetta Stone (RS)	989 795
	Gene neighbor (GN)	18 823
	Gene cluster (GC)	11 586

of extracted relations from different data sources are summarized in Table 1.

Based on our two-layered BN approach, relations supported by different evidence are scored by benchmarking against KOG and combined into an integrated network. The integrated network contains 25 660 genes and 250 762 268 relations.

5.5 Evaluation

Based on our proposed framework, we evaluated the integrated gene functional network.

5.5.1 Evaluation metric Because KOG was used as the scoring benchmark already, we used AraCyc to evaluate the integrated network. We sorted all the relations in the integrated network by the joint confidence score in a descending order. We use TP rate of top relations as the evaluation metric. This metric has been used in previous studies (Bowers *et al.*, 2004; Lee *et al.*, 2004).

5.5.2 The effect of relation integration In order to demonstrate the power of relation integration, we compared the relations in the integrated network with those supported by evidence from each individual resource based on the AraCyc benchmark (Fig. 3).

Relations extracted from gene expression are associated with a correlation or a MI value, corresponding to a different confidence score. We evaluated these relations sorted by the confidence score and drew a curve for each form of evidence from gene expression data. In contrast, relations extracted from literature and genome sequences are binary relations in our study. Each type of these relations is assigned with a single confidence score. Therefore, each of these evidence forms is shown as a dot in Figure 3. In our study, the TP rates for these evidence forms, i.e. CO, RP, PP, RS, GN and GC, are 0.0380, 0.0252, 0.0062, 0.0020, 0.0366 and 0.0044, respectively. We found that among all different evidence, co-occurrence relations from literature had the highest TP rate, while relations extracted by MI analysis from the abiotic stress dataset had the lowest TP rate of all.

In general, the curve of integrated relations (JOINT) is above most of the other curves or dots, except for relations extracted by PP and GN from gene sequence. It demonstrates that the proposed BN-based approach achieved higher TP rate by combining evidence from multiple resources. Relations supported by diverse forms of evidence are more likely to be correct. This highlights the merit of relation integration: weak evidence from multiple resources can be combined to provide strong evidence for a relation.

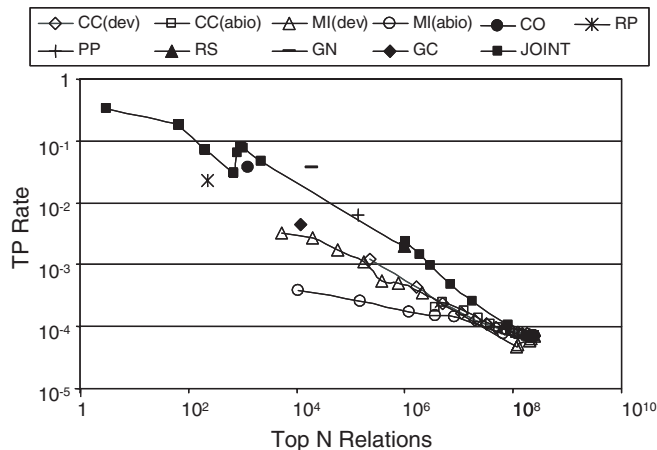


Fig. 3. Comparison between the integrated network (JOINT) and individual evidence form: TP rate of relations versus the number of top relations. CC, MI, CO, RP, PP, RS, GN and GC represent different analytical techniques (Table 1). For CC and MI, dev/abio corresponds to the dataset of development/abiotic stress.

5.5.3 Reliability of the integrated network In the integrated network, relations with a higher joint score are believed to be more reliable. Given different thresholds of joint confidence score, denoted by P_{cut} , we evaluated the TP rate of the relations above P_{cut} based on the AraCyc benchmark. Figure 4 shows the TP rate of top relations in the network against P_{cut} . In general, the TP rate of top relations increases as P_{cut} increases. Regression analysis on the TP rate against P_{cut} shows that R -square between TP rate and P_{cut} is 0.78 and the regression coefficient of the independent variable P_{cut} is 0.28 (P-value < 0.0001). This result demonstrated a significant positive correlation between the TP rate of relations and the confidence score. In other words, gene relations with a higher joint confidence score are more likely to be correct.

5.5.4 Contribution of data sources In the integrated network of over 250 million relations, many of the relations represented noise. We used the AraCyc as a gold standard to eliminate low quality relations and thereby to identify a network with reliable relations. Among the 25 212 relations from AraCyc, 2853 relations can be matched with those from KOG, i.e. the probability of matching is about equal to 0.1. Thus, we used the top relations with a joint confidence score >0.1 to form the 'confident network'. In this network, the number of genes was reduced to 11 082 and the number of relations was reduced to 1 007 883.

The confident network is believed to provide strong evidence about relations between various genes and proteins. We analyzed distribution of relations from different data sources to evaluate the contribution of each data source. The confident network includes many genes of unknown function. Reliable relations involving such genes are interesting to biological researchers in raising new hypotheses about gene biological functions. Among the 11 082 genes in this network, there are 207 genes lacking an associated GO term and 3027 genes associated with the GO term 'biological process unknown', which we collectively regard as genes of unknown function (GUF). Table 2 summarizes the distribution of gene relations from different data sources.

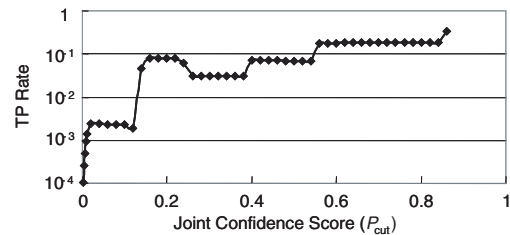


Fig. 4. Evaluating the reliability of integrated network: TP rate of relations versus joint confidence score P_{cut} .

Among the 1 007 883 relations in the confident network, most of them were extracted from genome sequence information and gene expression data. In addition, only 1092 relations (87 + 840 + 14 + 151) were supported by evidence from biological literature. Among all the relations, there are 345 593 relations (34.25%) involving GUF. Again, genome sequence and gene expression data were the major sources of these relations. In general, relations supported by fewer forms of evidence tend to involve genes of unknown functions, and vice versa. Particularly, relations only from literature had the highest percentage of relations involving GUF (74.71%), and relations from all the three data sources had the lowest percentage of relations with GUF (9.27%).

5.5.5 Analysis of gene functional clusters Relation integration provided a genome-wide network for *Arabidopsis*. In order to gain more insight about the network, we analyzed the network structure based on a stepwise creation process. Specifically, in descending order of the joint confidence score, we added one link at a time to construct the network. Thus, the sub-networks at each step are only composed of the functional relations with the highest joint score, i.e. with the strongest evidence to support. At the early stages, links with higher scores were mainly disconnected from each other because of the large coverage of the network. As more links were added, genes were gradually grouped into clusters, which may infer the existence of gene functional groups. After including the top 3000 relations, three major clusters of gene functional relations emerged. Cluster A contains 492 nodes and 937 links, cluster B contains 221 nodes and 668 links and cluster C contains 122 nodes and 145 links. We reviewed these three clusters and found a relatively clear distinction of function among them. Particularly, genes in cluster A are mainly involved in regulation of transcription, genes in cluster B in protein phosphorylation/kinase activity, and those in cluster C in electron transport.

Analyzing the subnet could provide more insight into the gene functions. For example, in a subnet centered on *PHYB* (At2g18790), we found that genes were connected by linkages supported by different evidence. Many genes in this subnet are transcription factors, kinases or photoreceptors, which are typical players in signal transduction pathways. It is also observed that genes involved in the same pathway tend to form sub-clusters. For instance, GA1 (At4g02780), GA3 (At5g25900) and GA4 (At1g15550) are enzymes involved in gibberellic acid biosynthesis and signaling (Cowling *et al.*, 1998; Helliwell *et al.*, 1999; Sun and Kamiya, 1994); FT (At1g65480), LFY (At5g61850), and CO (CONSTANS, At5g15840) are transcription factors that are responsible for flowering control or flower development (Ayre and Turgeon,

Table 2. Distribution of gene relations from different data sources (*E*: gene expression data; *L*: biological literature; *S*: genome sequence) in the confident network. '✓' and '—' indicate whether relations are from the source or not

Data sources			Number of relations	Number of relations with GUF	Percentage of relations with GUF
<i>E</i>	<i>L</i>	<i>S</i>			
✓	—	—	50	30	60.00
—	✓	—	87	65	74.71
—	—	✓	156 052	78 572	50.35
✓	✓	—	840	292	34.76
✓	—	✓	850 689	266 618	31.34
—	✓	✓	14	2	14.29
✓	✓	✓	151	14	9.27
	Sum		1 007 883	345 593	34.25

2004; Huang *et al.*, 2005; Moon *et al.*, 2005). These results provide some evidence for the potential value of our framework to integrate gene relations from various resources.

6 CONCLUSIONS AND FUTURE DIRECTIONS

In this study, we proposed a framework for relation integration that combines evidence from different sources to construct a genome-wide functional network. We employed different analytical techniques to extract gene relations from three data sources: gene expression data, biological literature and genome sequence information, and we constructed a gene network of *Arabidopsis thaliana*. Evaluation on the integrated network confirmed the validity and potential value of the proposed framework for relation integration. To the best of our knowledge, this is the first study to construct such a genome-wide network of *Arabidopsis* by integrating relations from diverse data sources.

In the future, we plan to study the prediction of new gene relations based on topological analysis of the integrated network. The predicted relations can be evaluated and integrated into the network based on the BN. Furthermore, we plan to examine our proposed framework on a larger test-bed of plant genes as well as on other biomedical domains such as cancers.

ACKNOWLEDGEMENTS

The project is supported by grants to H.C.: NIH/NLM, 1 R33 LM07299-01, 2002-2005, 'Genescene: a Toolkit for Gene Pathway Analysis' and to D.W.G.: NSF DBI 0211857, 2002-2005, 'Technology Development: Novel Techniques for Discovery of Patterns of Gene Regulation within Complex Eukaryotic Tissues'.

Conflict of Interest: none declared.

REFERENCES

- Ayre, B.G. and Turgeon, R. (2004) Graft transmission of a floral stimulant derived from *CONSTANS*. *Plant Physiol.*, **135**, 2271–2278.
- Bowers, P.M. *et al.* (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **5**, R35.
- Cowling, R.J. *et al.* (1998) Gibberellin dose-response regulation of GA4 gene transcript levels in *Arabidopsis*. *Plant Physiol.*, **117**, 1195–1203.
- Friedman, N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.
- Gasch, A.P. and Eisen, M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, **3**, RESEARCH0059.
- Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
- Helliwell, C.A. *et al.* (1999) *Arabidopsis* ent-kaurene oxidase catalyzes three steps of gibberellin biosynthesis. *Plant Physiol.*, **119**, 507–510.
- Huang, T. *et al.* (2005) The mRNA of the *Arabidopsis* gene FT moves from leaf to shoot apex and induces flowering. *Science*, **309**, 1694–1696.
- Huang, Z., Jiexun, Li, Hua, Su and George, S. Watts and Hsinchun Chen, Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining, Decision Support Systems, In Press, Corrected Proof, Available online 11 April 2006 (<http://www.sciencedirect.com/science/article/B6V8S-4JP9FN6-2/2/c631841ae222122e215b86b3e5b1f5b6>).
- Jansen, R. *et al.* (2002) Integration of genomic datasets to predict protein complexes in yeast. *J. Struct. Funct. Genom.*, **2**, 71–81.
- Jansen, R. *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
- Jenssen, T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.
- Lee, I. *et al.* (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
- Marcotte, E.M. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
- Marcotte, E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
- Marshall, B. *et al.* (2006) Aggregating automatically extracted regulatory pathway relations. *IEEE Trans. Inform. Technol. Biomed.*, **10**, 100–108.
- McDonald, D. *et al.* (2004) Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics*, **20**, 3370–3378.
- Moon, J. *et al.* (2005) Analysis of flowering pathway integrators in *Arabidopsis*. *Plant Cell Physiol.*, **46**, 292–299.
- Mueller, L.A. *et al.* (2003) AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol.*, **132**, 453–460.
- Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
- Sun, T.P. and Kamiya, Y. (1994) The *Arabidopsis* Gal locus encodes the cyclase ent-kaurene synthetase-a of gibberellin biosynthesis. *Plant Cell*, **6**, 1509–1518.
- Tatusov, R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
- Troyanskaya, O.G. *et al.* (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.
- Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
- von Mering, C. *et al.* (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
- von Mering, C. *et al.* (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.
- Yanai, I. and DeLisi, C. (2002) The society of genes: networks of functional links between genes from comparative genomics. *Genome Biology*, **3**, research0064.0061–research 0064-0012.