Inferring Regulatory Networks through Orthologous Gene Mapping

Guoyi Zhao*, Li Guo†, Lixin Gao* and Li-Jun Ma†
*Department of Electrical and Computer Engineering
University of Massachusetts Amherst, USA
†Department of Biochemistry & Molecular Biology
University of Massachusetts Amherst, USA

***Abstract*—In recent years, constructing regulatory networks using gene expression data has received extensive attentions. From Boolean network, Bayesian network to Module network, a number of models have been applied in order to learn the regulatory networks more accurately. The statistical power of network modeling is directly affected by sample size of available expression data used as training data. However, training data are not always abundantly available, except a few well-studied model organisms. It is also infeasible to perform a large number of experiments which require a lot of resources and labor. How to learn a reliable network using minimal training data making use of well-characterized model organisms becomes an important problem with pressing needs. In this paper, we developed a method that infers regulatory sub-networks for a species with limited expression data by learning from a known reference network through orthologous gene mapping. Inspection of three predicted sub-networks confirms biological relevance of our predictions and demonstrates the ability of the method in extracting core regulatory relationships.**

I. Introduction

Cellular organisms maintain proper biological functions and respond to environmental stimuli through fine-tuning their sophisticated and complex cell circuitries including transcriptional regulation, signal transduction and metabolic pathway networks. Studying the structures and regulatory mechanisms of these cellular networks is not only important to understand the fundamental processes of cellular organisms, but also has significant impact on medical practices. In recent years, researches have blossomed on reconstruction of local and global cellular networks applying computational models and whole genome gene expression profiling in many organisms such as bacteria (*Escherichia coli*) [1, 2], baker's yeast (*Saccharomyce cerevisea*) [3, 4] and even some mammals [5]. Many of these studies have proved successfully reconstructed transcriptional regulatory networks from transcriptome data using probabilistic models such as Bayesian network model [6] and Boolean network model.

Fusarium is a genus of filamentous fungi that have diverse ecological roles in nature and impact on human life [7]. Many *Fusarium* species cause devastating diseases on agricultural crops and some are human pathogens able to cause infection on immune-compromised individuals such as HIV patient and those receiving organ transplants. In this study we focused on three *Fusarium* species: a cereal pathogen *F. graminearum* (*Fg*), a pathogen with wide host range *F. oxysporum* (*Fo*) and a *F. verticillioides* (*Fv*) and studied their gene regulatory networks. Currently, knowledge on the transcriptional regulation networks in Fusarium species is very limited, despite the importance of these fungi in human health and food safety. Learning such networks is crucial for understanding their biology and making sound strategies controlling the problems caused by them.

Through genome sequencing and bioinformatics analysis, it has been shown that the three genomes shared high degree of synteny (>79%) and high average sequence identity (>80%) [8–10]. In addition, based on syntenic mapping, over 8000 genes are conserved among the three species and identified as orthologous genes [8]. However, it has previously been shown that genome conservation does not necessarily reflect its functional conservation as the gene regulatory networks are often dissimilar even among closely related organisms [11]. It is unknown how much of this sequence conservation is consistent with the functional conservation of these genes.

Our study reveals that the orthologous genes in the three *Fusarium* species are highly conserved in functions and regulatory mechanism. Taking advantage of this fact, we propose a new method to infer the regulatory network of one species based on the regulatory network of a closely related species, and demonstrate the feasibility for inferring a large portion of a regulatory network using limited experimental samples. To our knowledge, this is the first study of global gene regulatory network inference in filamentous fungi and we believe our network construction has captured true biological meanings and provide important framework for experimentally verification of the reconstructed networks in the future. The major contributions in this paper are the following:

- We reconstructed the gene regulatory networks in *F. graminearum* from combined transcriptome data using Bayesian network model. Using the algorithm, we successfully predicted the top regulators of *F. graminearum* gene expressions and correlated these regulator genes with their target genes. With the predicted *F. graminearum* regulatory networks, we obtained gene regulatory networks of core genome in three *Fusarium* species by mapping the ortholog data on the *Fg* networks.
- Based on the *Fusarium* orthologous genes, we determined the conservation level of the gene regulatory networks among the three *Fusarium* species using an additional

transcriptomics dataset. With a small set of transcriptomic data containing just 4 biological conditions (less than 10% of *Fg* network data), the orthologous networks have shown high consistency (51.8% to 77.5%) in all three species in cross validation test.

- Biological validation of several sub-networks proved that the consistent orthologous networks we predicted have correctly captured the regulatory relationships between the regulators and target genes. Using current biological knowledge, we vigorously validated the consistent regulatory networks and showed that the correlation of the top regulators and their target genes has high credibility. And many of our predictions have given novel annotation to the regulators that previously do not have known biological functions, which can be validated by future biological experiments.

The rest of the paper is organized as follows. In Section 2, we give the *Fusarium* gene expression data description and biology experiments setting. Section 3 briefly describes how we learn the *Fg* regulatory network. Two levels of consistency check are provided in Section 4. Section 5 describes the new method of inferring the regulatory network through the orhologous gene mapping. Validation results are illustrated in Section 6. We review the related work in Section 7 and conclude the paper in Section 8.

II. Data Description

The transcriptomic data used for gene regulatory network inference are a collection of *F. graminearum* microarray data on 198 samples in 55 experimental conditions such as signaling pathway perturbations, nutrient starvation, sexual and asexual development, toxin production inducing conditions and time-series of plant infection. Majority of these data were obtained from a publicly accessible collection of microarray data in Plexdb (www.plexdb.org). Additional *F. graminearum* gene expression datasets were generated in this study using the Affymetrix Fungal Genechip, covering experimental conditions such as nitrogen and carbon starvation, cAMP-PKA signaling mutants and several chromatin modification mutants. Expression levels of total 13,331 genes from 198 samples were calculated from raw microarray data using robust multiple array algorithm.

To conduct the comparative analysis of gene regulatory networks in three *Fusarium* species, using RNA sequencing approach, we measured genome-wide gene expression of three species under 4 biological conditions: normal temperature (28°C), high temperature (37°C), normal pH (pH=7) and low pH (pH=5). Total RNAs were extracted from the 3 species culturing in the above conditions, purified and sequenced using Illumia Hiseq technology. The data was processed and analyzed in CLC Genomics Workbench and gene expression values were calculated and presented as RPKM (Reads Per Kilobase of transcript per Million mapped reads).

The gene and orthologs information for the 3 *Fusarium* species is shown in Figure 1. Based on the orthologous gene mapping, around 50% to 60% regulators are conserved

	Total genes	orthologs	TF genes	Orthologs TF genes
F. verticillioides	14,169	12,218	629	425
F. oxysporum	17,708	12,291	811	430
F. graminearum	13,331	9,722	661	470

Fig. 1. Gene and orthologs information for 3 *Fusarium* species.

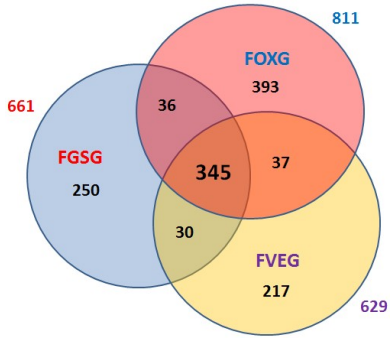

Fig. 2. Candidate TF overlap in 3 *Fusarium* species .

in both genes and candidate regulators. Figure 2 shows the exact candidate regulators overlap among 3 species. The high conservation of the genes is the foundation of the regulatory networks consistency among these species. In section 4, we will further analysis the consistency of these 3 species in both gene expression level and regulatory logic level.

III. Regulatory Network Learning

In order to infer regulatory networks from known networks, good reference networks are quite necessary for the practical inferring. In this section, we describe how to learn the regulatory networks of *F. graminearum* from gene expression data using the MinReg [6] algorithm. The idea of the MinReg algorithm is to construct a constrained structure of Bayesian networks to depict the regulatory influences in gene networks. Learning Bayesian network model [12] has been applied widely in studying biology networks [6, 13, 14]. With the reasonable assumption that there is only a certain number d of regulators will regulate a target gene, the MinReg algorithm can scale up to learn a complex network with tens of thousands of genes in a short period of time.

A. Learning Bayesian Networks

Bayesian networks are directed acyclic graphs whose nodes represent random variables under the Bayesian theory. Since they reflect the probabilistic relationships among variables, it is a good model to learn the Biology networks such as the regulatory networks we want to learn. However, the traditional learning procedure is infeasible because there are tens of thousands of the nodes and millions of variables need to be modeled in this problem. With the biology background knowledge, some restriction of this complex network can be

made to simplify the learning. Therefore, we used the greedy scheme algorithm MinReg to learn our regulatory network.

The objective in constructing a Bayesian network is to find a structure that gives the highest probability with the observed dataset. The Bayesian score, based on the likelihood function, is the common measure of the fitness of a given Bayesian network. The local score function is defined based on Bayesian paradigm and uses the Bayesian BDe scoring function [12] as shown in Equation 1.

$$Score_B(\mathcal{G} : \mathcal{D}) = logP(\mathcal{D}|\mathcal{G}) + logP(\mathcal{G}) \quad (1)$$

The BDe score is calculated based on a certain class of priors with several desirable properties. In this work, the BDe score of the entire regulatory network G can be decomposed into the sum over the local scores for each variable as in Equation 2.

$$Score(\mathcal{G} : \mathcal{D}) = \sum_i Score_B(X_i, Pa_i : \mathcal{D}) \quad (2)$$

B. Constructing Regulatory Networks

To learn Bayesian Network structure, we greedily increase the regulator set by adding the candidate regulator that gives the highest Bayesian score. For the first d regulators, we pick the candidate regulators that give the top-d scores in the entire gene set. After that, any new regulator added to the network will compete with previously assigned regulator parents. A dynamic process will fit network structure for a higher score. To bound the number of key regulators, we terminate this greedy process of picking regulators when k regulators have been selected.

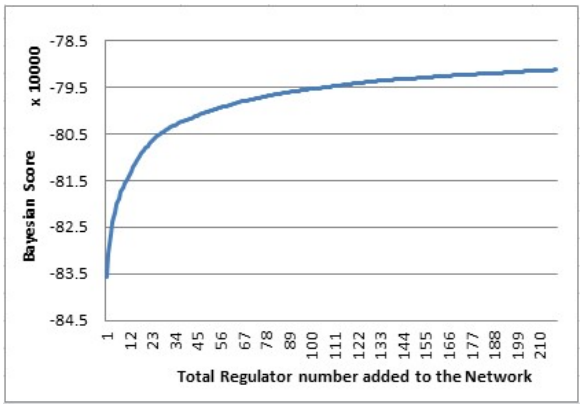

Fig. 3. The Bayesian Score and regulators children genes number changes towards the variation of key regulators number.

In each iteration, the Bayesian score will increase but the growth rate will decrease. As shown in the Figure 3, after 100 regulators have been picked, the change of Bayesian score is quite small. We can finish the learning process by setting a threshold as a significant difference in Bayesian score. At last, 120 regulators have been selected as the key regulators. For each target gene, we choose the 3 parents regulators with the highest local BDe score as their parents.

IV. Transcription Regulatory Network Consistency Analysis

In this section, the feasibility of inferring regulatory networks is shown with the consistency check in 2 aspects. The first analysis is performed on the expression data level. Since the RNA-seq data contains the same experiment settings for the *Fg*, *Fv* and *Fo* species, we check whether the expression states of the orthologous genes are consistent across the species in the same expression. Another analysis is performed on the regulatory network expression logic level. After learning the *Fg* regulatory network, each target gene has 3 parents as the regulators. Under this regulatory relationship, how target genes will express can be predicted based on a certain pattern of the parents' expression.

A. Data Normalization

Data normalization is the process of discretizing the input gene expression value to the 3 states which are standard reference (–), up-regulated (↑) and down-regulated (↓). Usually, two standard methods are used to normalize the expression data. One is to use fold change cutoff and the other one is based on the statistical hypothesis T-test. Using fold change cutoff, correlations among the genes can be maintained in each duplicate directly and maximally. However, noise in the data can greatly affect the network inferring. The T-test based normalization can keep the most reliable relationship, but it may filter too many data instances. In this paper, since the data reliability is the primary concern, T-test normalization is a better choice.

$$t = \frac{\overline{X}_{sample} - \overline{X}_{cm}}{\sqrt{\frac{Var_{sample}}{N_{sample}} + \frac{Var_{cm}}{N_{cm}}}} \quad (3)$$

The formula of T-test is shown in Equation 3. The value t is the score which quantifies the distribution difference between experiment $sample$ and the reference cm (terms of complete medium). $\overline{X}$ is the arithmetic mean of the experiment samples. Var is the variance. N is the numbers of duplicate samples. For each gene under one experiment setting, all the duplicates are treated as a whole to compare with standard reference. If the distributions of the experiment and the reference are significantly different from each other, this gene is either up-regulated or down-regulated.

B. Gene Expression Data Consistency

The consistency of gene expression data is reflected by how the orthologous genes express under the same condition. Instead of using numerical gene expression value, 3 gene expression states are discretized based on how the expression is different from the normal condition. With the help of the data normalization process, genes expression states can be compared through the one-to-one orthologs mapping.

Table I shows the summary of the gene expression states analysis on the RNA-seq raw data. Before doing the gene orthologs mapping, the overall state distributions of the 3 *Fusarium* species are quite consistent. More than 60% of the

gene samples are in normal condition, 15% to 20% of the genes are up-regulated while the down-regulated genes have a similar ratio. Those genes which are up-regulated or down-regulated will reveal the connection among the regulators and the genes that take part in a particular biology process.

TABLE I
SAMPLE STATES ANALYSIS FOR RAW DATA

	Fg	Fv	Fo
total samples	119,889	127,521	141,664
reference (ratio)	84,647 (71%)	79,874 (63%)	85,106 (60%)
up-regulated (ratio)	19,263 (16%)	23,162 (18%)	25,633 (18%)
down-regulated (ratio)	15,979 (13%)	24,485 (19%)	30,925 (22%)

Figure 4 shows the expression data consistency across the 3 species. From a total number of 24,243 sample points, 56.3% of the genes share the common state. As illustrated in Figure 1, *Fv* and *Fo* have the closest relationship. The pairwise overlap of these 2 species is over 80%, while *Fg* have an overlap ratio around 65% with the *Fo* and *Fo* species. With additional data consistency check shown in Table I, the high consistent ratios in expression level data make the gene connection among 3 species really strong. Therefore, great confidence can be built on these data to perform further network inferring.

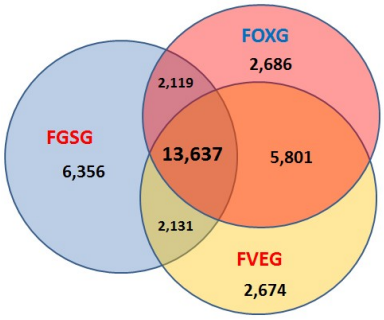

Fig. 4. Pairwise gene expression consistency among 3 species.

C. Network Regulatory Expression Logic Consistency

From the learned *Fg* regulatory network, besides the three regulator parents for each target gene, the regulatory relationship between the regulators and the genes is the most concern in this study. In this section we first describe how to derive the regulatory expression logic from the regulatory network and then show the expression logic consistency results.

1) Derive Regulatory Logic from Regulatory Network: The regulatory expression logic specifies how the target gene will express under a particular configuration of the parent regulators. Based on the learned regulatory network, the experiment sample numbers for different regulator parents and gene configurations are shown in Table II. Only the configurations that contain at least one sample are listed.

Based on the statistics of input experiment sample, gene regulatory expression logic can be derived from the majority samples of each parent configuration. For the 3 possible gene expression states, the state with the largest number of samples should be the expression logic for that configuration. For example, like the up-regulated for row 1 and down-regulated for row 8, 9, 10 in Table II.

TABLE II
SAMPLE NUMBERS UNDER DIFFERENT REGULATOR PARENTS AND GENE CONFIGURATION

No.	parents' configuration pa_1	pa_2	pa_3	sample numbers ↓	–	↑
1	↑	↑	↑	0	0	13
2	↑	↑	–	0	1	0
3	↑	–	–	0	1	1
4	–	↑	–	0	1	0
5	–	–	↑	0	1	0
6	–	–	–	0	24	0
7	↓	–	–	1	2	0
8	↓	↓	↑	1	0	0
9	↓	↓	–	1	0	0
10	↓	↓	↓	5	0	0

TABLE III
DERIVED GENE EXPRESSION LOGIC FROM DIFFERENT PARENTS' CONFIGURATION

No.	parents' configuration pa_1	pa_2	pa_3	gene expression states basic	integrated
1	↑	↑	↑	↑	↑
2	↑	↑	–	–	– / ↑
3	↑	↑	↓	unknown	↑
4	↑	–	↑	unknown	↑
5	↑	–	–	noMajor	–
6	↑	–	↓	unknown	impossible
7	↑	↓	↑	unknown	impossible
8	↑	↓	–	unknown	impossible
9	↑	↓	↓	unknown	impossible
10	–	↑	↑	unknown	↑
11	–	↑	–	–	–
12	–	↑	↓	unknown	impossible
13	–	–	↑	–	↑
14	–	–	–	–	–
15	–	–	↓	unknown	↓
16	–	↓	↑	unknown	impossible
17	–	↓	–	unknown	–
18	–	↓	↓	unknown	↑
19	↓	↑	↑	unknown	impossible
20	↓	↑	–	unknown	impossible
21	↓	↑	↓	unknown	impossible
22	↓	–	↑	unknown	impossible
23	↓	–	–	–	–
24	↓	–	↓	unknown	↓
25	↓	↓	↑	↓	↓
26	↓	↓	–	↓	↓
27	↓	↓	↓	↓	↓

Sometimes there are no majority samples in the original data, e.g., row 3 in Table II. Initially, they be will treated as *unknown* or *noMajor*. To help determining the majority states, we include additional samples from the opposite configuration and related configurations. The opposite configuration is the one with the states fo each parent being changed from up to down and vice versa. The related configurations are defined to be those that maintain the ↑ and ↓ but replace the – with 3 possible states. If the sample number is not sufficient we

will also include the opposite related configurations. Instead of considering the configuration itself, all related configuration samples are combined to form an integral group to derive the expression logic. Sometimes parent regulators cannot have opposite states; there will be no data points under these configurations. We treated these cases as *impossible* configurations. The derived regulatory expression logic is shown in Table III.

Algorithm 1 Gene expression logic derivation algorithm

Input: Gene normalized expression data
Output: Express gene states for each parents' configuration

for parents' configuration $parents_j$ of each gene gn_i **do**
  if there is a largest sample number (majority, short for mj) $Stat_{mj}$ in the 3 gene states **then**
    Output gene state $Stat_{mj}$
  else if at least two of the 3 gene states have the same sample number but not 0 **then**
    if the opposite configuration contains a $Stat_{mj}$ **then**
      Output gene state $-Stat_{mj}$
    else if the integrated sample number which combined the relative configurations contains a $Stat_{mj}$ **then**
      Output $Stat_{mj}$
    else
      Output $unknown$ gene state
    end if
  else if all the same sample number are 0 **then**
    if the opposite configuration contains a $Stat_{mj}$ **then**
      Output gene state $-Stat_{mj}$
    else if the integrated sample number table contains a $Stat_{mj}$ **then**
      Output gene state $Stat_{mj}$
    else if no samples in all relative configurations **then**
      Output $Impossible$ gene state
    end if
  end if
end for

Algorithm 1 describes how to derive the gene expression logic for each situation. For the configurations with no majority samples, the worst case is that all the related combinations also do not contain a major state. Since the data is noisy and this predicted network is not perfect, an unknown state could exist for these configurations. But those cases are less than 0.2% in the total gene configurations in our data.

2) Regulatory Logic Consistency Check Results: With the derived regulatory expression logic, the consistency of the RNA-seq data with the reference *Fg* network is shown in this section. Under each regulatory parents' combination, if the target gene expresses as specified by the expression logic, the RNA-seq data is considered to be consistent with the reference network.

Since the orthologs only cover a subset of the *Fusarium* genes, after orthologs mapping, sometimes only 1 or 2 parents of a gene will remain. For this case, simpler expression logic table can be generated, similar to Table II. Although this is inconsistent with the assumption that there are 3 parents for each target gene, it is still a good derived sub-network maintained from the reference network with a high consistent ratio shown in consistency analysis. Results of expression logic with any parents number and 3 parents only are shown in Figure 5 and Figure 6.

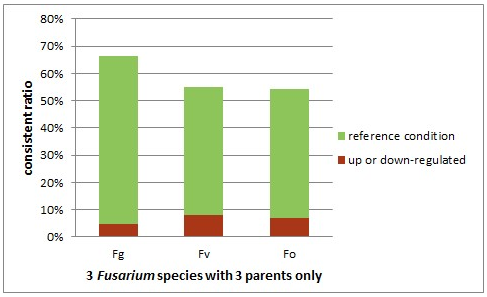

Fig. 5. Regulatory logic consistency of orthologous genes with 3 parents only.

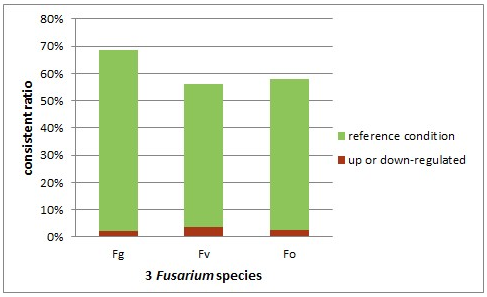

Fig. 6. Regulatory logic Consistency of RNA-seq data after orthologous genes mapping.

In Figure 5, the results show the consistency of all the RNA-seq data, including 1 or 2 parents' cases. All the 3 *Fusarium* species data have a high consistent ratio in comparison to the inconsistent case and the unknown cases. Figure 6 is the consistency results for genes which still have 3 parents after orthologs mapping. The consistent ratio is 54% to 64%. Plus the expression data consistent ratio from 60% to 70%, the regulatory network of one specie can be a good reference to infer networks for the other two species.

V. Inferring Regulatory Networks through Orthologous Gene Mapping

In Section 3, the *F. graminearum* regulatory network is inferred using the MinReg algorithm. In theory, regulatory networks of any species can be learned from the gene expression data under the Bayesian network model. However, learning a sound regulatory network requires a large amount of experiment data. It is not only labor intensive but also resource consumption. In this paper, we propose a new method to infer the regulatory network from a reference network and orthologous genes. Even though there are only 10% of the experiment samples a good sub-network can be inferred with great confidence. After analyzing the infeasibility of learning regulatory networks with small data, the new method and results will be presented.

A. Learning Regulatory Networks from Small Samples

Using the MinReg algorithm, 3 regulatory networks are generated based on the small RNA-seq expression data. On average, there are 110 regulators learned for each species. The learned networks contain more than 40 orthologous regulators, but their overlaps are quite small as shown in Figure 7. The regulator and gene relationship overlaps are even smaller, which is only 1% to 2%.

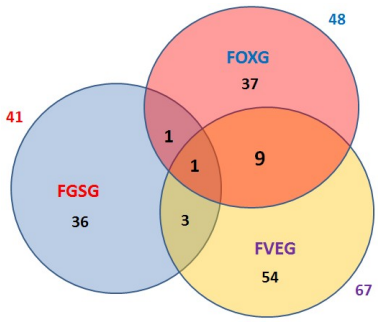

Fig. 7. 3 species regulatory network consistency.

The small overlap among the networks of the 3 *Fusarium* species contradicted with the biology knowledge that the core regulatory relationships should be conserved. With a small number of experiments, the noisiness of the data and the bias of the experiments chosen will usually make the learned network quite different from the real network. We propose a new method to infer the regulatory network with high confidence even there are only 3 experiment samples.

B. Inferring Regulatory Networks of Fv and Fo

Since the *Fg* regulatory network learned from a sufficient number of samples is proved to be sufficient [6], using the gene orthologs information, the network can be mapped to the *Fv* and *Fo* regulatory network. Although these are only sub-networks, they maintain the key relationships of *Fusarium* species based on the homologous genes. With additional experiments samples, the sub-network can be further refined.

1) Infer Fv and Fo Regulatory Networks through Orthologs: To infer regulatory networks of *Fv* and *Fo* with the *Fg* reference network, using the one-to-one mapping in the orthologs is the basic step. As shown in the first two rows of Table IV, after the orthologs mapping, 104 out of 120 (86.7%) regulators are retained; from all the regulators and genes relationships, approximately 18,000 from 39,459 are retained (at rate 45.6%). The 104 regulators are the same from all the three species. The small difference in the numbers of edges is due to the fact that there are some genes from *Fv* and *Fo* species which do not have the expression data.

Although the mapped sub-network is a good starting point for analyzing and understanding the networks, without having much information from *Fv* and *Fo* experiments, to assume that the orthologous genes will have exactly the same biology function seems to be arbitrary. As will be shown by biology analysis in the next section, this network make a good connection between regulators and target genes with similar functions. The diversity among species will be concealed by the pure orthologs mapping.

TABLE IV
3 SPECIES ORTHOLOGOUS NETWORKS REGULATORS AND EDGES RETAINED

	Fg	Fv	Fo
orthologous Regulator	104	104	104
orthologs edges (ratio)	17,989 (46%)	18,000 (46%)	17,951 (46%)
consistent edges (ratio)	13,933 (35%)	11,102 (28%)	9,294 (24%)
consistent in orthologs	78%	62%	52%

2) Refine Fv and Fo Regulatory Networks with Expression Data: After inferring the regulatory network from the reference network, whether it actually reflects the regulatory relationship of different species still remains unclear. Further refinement of the sub-networks using the expression data is a way to integrate the useful information together. Because the gene expression data is the key evidence for the interaction between regulators and target genes, small samples can be used to verify the inferred relationships.

To refine the inferred relationships only the edges that are consistent with the new RNA-seq data should be retained. The consistent edge here is defined based on the number of samples in the RNA-seq data that act the same as the expression logic predicted from the expression data. If there is at least one sample in the expression data that contains the same pattern, this data sample point is consistent. After evaluating the consistency of the 9 samples, if the consistent sample number is larger than a certain number, this regulatory relationship should be kept in the cross species sub-network. Then these 3 regulators and the target gene can be added as the nodes and the 3 relationships as edges.

From the previous analysis of the expression data consistency and the regulatory logic consistency, around 60% to 70% of the data is consistent; therefore, a threshold number of 5 ($5/9 = 55.6\%$) is chosen to keep the edges in the mapped network for further biology case study. As shown in the 4 and 5 rows of Table IV, after the orthologs mapping and expression data consistency refinement, around 23% to 28% of the edges are retained in *Fv* and *Fo* from the total 39,459 edges in the original *Fg* regulatory network. Those edges are the most confident sub-networks we can infer from the reference network. Since the orthologous edges are only 45% of the total edges, more than 50% of the edges are retained after the refinement.

The refined sub-networks integrate the knowledge for a good *Fg* reference regulatory network, orthologs information and the new RNA-seq expression data. For *Fv* and *Fo* species, with only 3 experiment conditions (which is only 5.8% samples comparing to the *Fg* expression data), we can learn a robust sub-network containing more than 50% of the regulatory relationships through orthologous genes. Further network analysis with biology background knowledge will be shown in the next section.

3) Overview of the Inferred Sub-network: The orthologous gene networks inferred from our work contain 104 regulators. The number of target genes for the 104 regulators ranges from 5 to 321. Instead of validating the local regulatory relations such as the one between an individual regulator and the target gene, a global biological validation of the regulatory networks was conducted on the whole target gene set. Assuming that these orthologous genes are functionally conserved, each regulator uses its existing GO (Gene Ontology) annotation knowledge in *F. graminearum* genome database (http://mips.helmholtz-muenchen.de/genre/proj/FGDB/). We first conducted GO term enrichment of all target genes for each regulator and compared the top enriched GO terms of the target genes with the GO annotations of the regulator. The results were summarized by the functional capture ratio (FCR) defined as the total overlapped GO functions between target genes and the regulator divided by total GO functions of the regulator. Among the 104 regulators, about 66% regulators have over 0.5 function capture ratio (FCR) and 34% regulators have less than 0.5 FCR in Figure 8. About 45.5% regulators have an FCR of more than 0.7 and 25% regulators have an FCR of 1. This result shows that the majority of the top regulators predicted have high correlation with their target genes in terms of molecular functions and biological processes, suggesting the high relevance and robustness of our network prediction results.

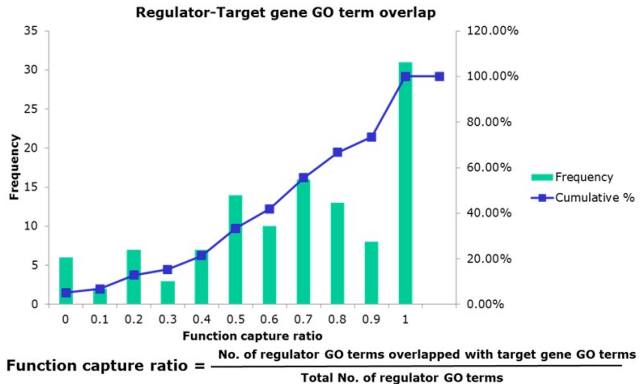

Fig. 8. Histogram of GO term analysis of top regulators and target genes in the orthologous gene regulatory networks.

VI. Biology Analysis on the Orthologs Networks

Close inspection of inferred sub-networks using three examples further illustrated the biological relevance of our predictions. Biological validation based on GO annotation and *F. graminearum* transcription factor phenotype database (FgTFPD) shows strong association of biological processes between regulator and target gene in the inferred networks.

In *Fg* networks, FGSG_01350 is a Zinc finger transcription factor and regulates 92 target genes as predicted. FgTFPD shows that FGSG_01350 mutant completely lost its sexual reproduction and pathogenicity and grows poorly on axenic media. These defects can be explained in our network prediction results. Biological processes enriched in the predicted 92 target genes include metabolism (16 genes), energy (6 genes), cell defense (5 genes) and fungal specific cell type differentiation (5 genes) etc. The 5 genes controlling fungal specific cell type differentiation are FGSG_04998, FGSG_07946, FGSG_08621, FGSG_07418 and FGSG_10048 that are either pheromone response, mating-type determination, sex-specific proteins (FGSG_08621 and FGSG_07418) or development of asco- basidio- or zygospore (FGSG_04998, FGSG_07946 and FGSG_10048). In addition, genes involved in cell defense are important for fungal pathogen to resist host antifungal processes and defective transcription of these genes caused by mutation of their regulator can lead to reduction or loss of virulence. Our prediction results show that FGSG_01350 regulates variety of biological processes that involve transcription of 92 target genes, consistent with previous finding that deletion of FGSG_01350 can lead to disruption of these processes and pleiotropic effects on the fungus.

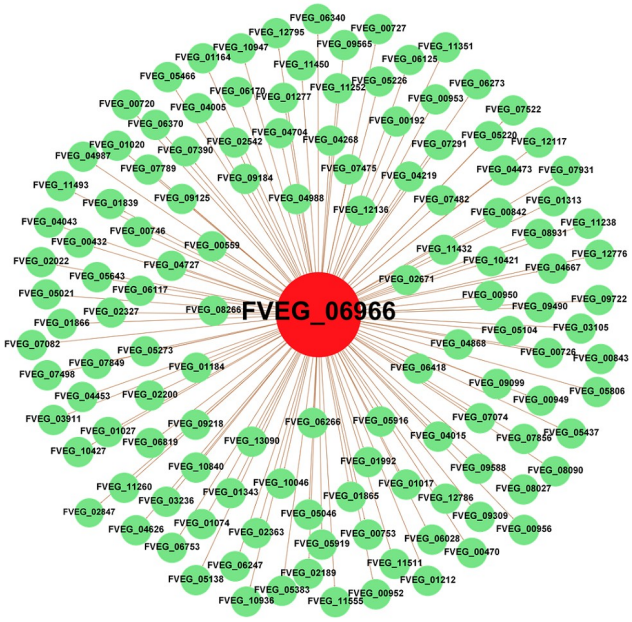

Fig. 9. Regulator networks of three Fusarium regulators FVEG_06966. Red nodes are regulators and green nodes are target genes.

Take FVEG_06966 in *Fv* networks for another example, it encodes an endoplasmic reticulum ATPase homologous to CDC48 in *S. cerevisea*, and regulates 127 target genes that are significantly enriched for biological processes in protein fate (folding, degradation), cellular transport and unfolded protein response (ER quality control). This is interesting because CDC48 is a key component of ERAD (endoplasmic reticulum associated degradation) pathway that controls the degradation of misfolded proteins via ubiquitination. Our network prediction has accurately captured the target genes of FVEG_06966 that are involved in protein degradation pathways : FVEG_11252 (probable 20S proteasome subunit Y7), FVEG_11432 (26S proteasome regulatory particle chain RPT3), FVEG_06117

(20S proteasome subunit alpha 5), FVEG_07482 (RPT2 - 26S proteasome regulatory subunit), FVEG_00753 (26s proteasome p44.5 protein), FVEG_01017 (26S proteasome regulatory subunit YTA3) and FVEG_01343 (nuclear protein localization factor and ER translocation component NPL4). These genes encode proteins that are essential components of proteasomes where the protein degradation takes place. Fv network prediction results show that FVEG_06966, a key part of ERAD, regulate these genes. This discovery provides guidance to future functional analysis of the ERAD pathway in *Fv* using experimental approaches.

In *Fo* networks, FOXG_13852 encoding a serine/threonine protein kinase is regulated by FOXG_03523, FOXG_01068 and FOXG_03041 and it regulates 274 target genes. GO annotation shows FOXG_13852 is involved in phosphate metabolism, cell growth, morphogenesis and pheromone response and mating-type determination. GO term analysis of the 274 target showed that all 3 processes were captured by the target genes. For example, 10 genes are involved in cell growth/morphogenesis and 4 genes are involved in pheromone response and mating-type determination. Unlike *F. graminearum*, studies on *F. oxysporum* mutagenesis have been limited. Currently, it is unknown what phenotypes the FOXG_13852 mutant may possess. It will be interesting to see whether FOXG_13852 mutant has defects in these biological processes. Our results showed that FOXG_13852 regulated target genes controlling these processes, which can be experimentally verified in the future.

VII. Related Works

Accompany with the popularity of regulatory network [6, 12–14], the *Fusarium* comparative genomic study [8–10] has made great progress in analyzing the horizontal gene transfer among different *Fusarium* species. With the help of orthologous genes [8], we have a better understanding of the core functional connection among those closely related species. However, the study of the common regulatory networks among the species and how they can infer the network of one another still remain unclear.

Regulatory network learning is a key problem in biology regulatory interactions research. Many works have been done to learn the regulatory networks with gene expression data and other information. The MinReg [6] algorithm proposed a greedy scheme method to learn the key regulators of the network. The Module network [14] defined modules for each cluster of genes and derives the regulatory interaction using decision tree. However, most of these works need a sufficient number of experiment data to make the prediction more accurate. In our work, we present the method to use small samples to infer the core regulatory network from reference networks.

VIII. Conclusion

In this paper, we proposed a new method to infer a regulatory network by extracting a sub-network from a reference network. We demonstrated that it is feasible to infer good sub-network using orthologs information by applying a series of data consistency and regulatory logic consistency analysis. We inferred different levels of sub regulatory networks and refined these networks using only a small number of data samples (less than 10%). The validation of the 3 examples from predicted *Fg*, *Fv* and *Fo* sub-networks gave us great confidence on the method we use in the sub-networks inferring.

Acknowledgment

This work is partially supported by NSF grants CNS-1217284 and CCF-1018114 and the United States Department of Agriculture, National Institute of Food and Agriculture Grant awards MASR-2009-04374, MAS00441.

References

[1] Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., et al. (2004). Integrating high-throughput and computational data elucidates bacterial networks. Nature 429, 92-96.

[2] Zare, H., Sangurdekar, D., Srivastava, P., Kaveh, M., and Khodursky, A. (2009). Reconstruction of Escherichia coli transcriptional regulatory networks via regulon-based associations. BMC systems biology 3, 39.

[3] Guelzim, N., Bottani, S., Bourgine, P., and Kepes, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. Nature genetics 31, 60-63.

[4] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298, 799-804.

[5] Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122, 947-956.

[6] Pe'er, D., Tanay, A. and Regev, A. (2006). MinReg: A Scalable Algorithm for Learning Parsimonious Regulatory Networks in Yeast and Mammals.. Journal of Machine Learning Research, 7

[7] Leslie, J.F., and Summerell, B.A. (2006). Fusarium laboratory manual (Blackwell Publishing).

[8] Ma, L.J., van der Does, H.C., Borkovich, K.A., Coleman, J.J., et al. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 464, 367-373.

[9] Ma, L.-J., H.C. Kistler, and M. Rep, Evolution of plant pathogenicity in Fusarium species. in Evolution of Virulence in Eukaryotic Microbes. L. D. Sibley, B. J. Howlett, J. Heitman (eds), 2012: 486 - 498.

[10] Ma, L.-J., et al., Fusarium Pathogenomics. Annu Rev Microbiol, 2013. 67: 399-416.

[11] Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nature genetics 42, 631-634.

[12] Heckerman D., Geiger D., and Chickering D. M. Learning Bayesian networks: The combination of knowledge and statistical data. In Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence, (1994). 293-301.

[13] Friedman N., Linial M., Nachman I., Peer D. Using Bayesian networks to analyze expression data, J. Comput. Biol. 7 (2000).

[14] Segal, E., Shapira, M., Regev, A., Peer, D., Botstein, D., Koller, D., Friedman, N.: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, Nat. Genet. 34, 166-176 (2003).

[15] Son, H. et al. (2011). A phenome-based functional analysis of transcription factors in the cereal head blight fungus, Fusarium graminearum. PLoS Pathog 7, e1002310.