GENOME SCALE META ANALYSIS OF MICROARRAYS FOR BIOLOGICAL INFERENCES

A Dissertation Presented to the Faculty of the Graduate School
University of Missouri-Columbia

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

By
GYAN PRAKASH SRIVASTAVA

Dr. Dong Xu, PhD
Dissertation Supervisor

DECEMBER 2009

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

GENOME SCALE META ANALYSIS OF MICROARRAYS FOR BIOLOGICAL INFERENCES

Presented by Gyan Prakash Srivastava a candidate for the degree of Doctor of Philosophy and hereby certify that in their opinion it is worthy of acceptance.

_____

Dr. Dong Xu

_____

Dr. Youssef G. Saab

_____

Dr. Jianlin (Jack) Cheng

_____

Dr. Michael X. Wang

_____

Dr. John Fresen

Dedicated to:

My parents, who gave me this wonderful life...

My teachers, who taught me the essence of life…

My son, with whom I am learning with each passing day of my life…

Acknowledgement

I would like to express my gratitude to Dong Xu, Ph.D. for giving me an opportunity to work with him. I would also like to thank him for his expert tutelage, patience, guidance, and confidence entrusted upon me to complete the research in this topic area, without which this project would not have been successful.

I acknowledge Dr. Jing Qiu, Ph.D. for her support. Her patience, guidance and insight have been invaluable assets throughout the course of my study. Michael Wang, M.D., Ph.D., Jianlin (Jack) Cheng, Ph.D. and John Fresen, Ph.D., have been great source of inspiration and encouragement to me throughout our collaborative effort. I would like to covey a special thanks to them. Youssef Saab, Ph.D., made excellent suggestions many times that prompted me to think deeply and look at my research problem critically. For his support, I express my gratitude.

I would express my heartfelt gratitude and thanks to my wife, my parents, my parents-in-law, my brothers and sisters, and other family members for their encouragement and support during my entire course of study. A special thanks to Larry Marsh, Ph.D., Prawal Sinha, Ph.D., whose guidance motivated me in my difficult days of life and without whose inspiration, I would have never come into research.

Finally, I would like to convey a special thanks to my colleagues from Digital Biology Laboratory, former colleagues from the lab and friends for their support and constant encouragement throughout my research.

Table of Contents

ACKNOWLEDGEMENT ---------- **II**

LIST OF TABLES ---------- **V**

LIST OF FIGURES ---------- **VI**

1. INTRODUCTION ---------- **1**

1.1. Functional Genomics ---------- **1**

1.2. Systems Biology ---------- **6**

1.3. Biological inferences from high-throughput data ---------- **10**

2. META-ANALYSIS OF MICROARRAYS ---------- **13**

2.1. Background ---------- **13**

2.2. Significance tests for correlation ---------- **14**

2.3. Test of combing results ---------- **15**

2.4. Implementation ---------- **18**

2.4.1. Dynamic meta-analysis and significance level ---------- 18

2.4.2. Dataset level meta-analysis ---------- 20

2.4.3. Network level meta-analysis ---------- 21

3. META-ANALYSIS APPLICATION: FUNCTIONAL GENOMICS ---------- **23**

3.1. Gene function prediction ---------- **23**

3.2. Break-down analysis ---------- **25**

3.3. Hypothesis testing ---------- **27**

3.4. Prototype development ---------- **29**

3.4.1. Microarray data processing ---------- 30

3.4.2. Statistical co-expression linkage graph ---------- 32

3.4.3. Function Prediction algorithm ---------- 39

3.4.4. Performance Evaluation ---------- 41

3.5. Implementation ---------- **43**

3.5.1. Yeast datasets ---------- 44

3.5.2. Human dataset ---------- 47

3.5.3. Case study: Sin1 & PCBP2 interaction ----- 52

3.6. Discussion and Conclusions ----- 54

4. META-ANALYSIS APPLICATION: SYSTEMS BIOLOGY ----- 57

4.1. Regulatory networks ----- 58

4.2. Gene Regulation Model ----- 59
4.2.1. Kinetic Model for Time Lag in Regulation ----- 60
4.2.2. Learning Model Parameters ----- 65

4.3. Implementation ----- 66

4.4. Evaluation and Comparative Analysis ----- 69

4.5. Case study of E2F transcription factor ----- 74
4.5.1. E2F-target identification and evaluation ----- 74
4.5.2. GO-enrichment analysis ----- 77

4.6. Network feature analysis ----- 79

4.7. Discussion and Conclusion ----- 81

4.8. Future work ----- 84

5. HIGH-THROUGHPUT OLIGO DESIGN ----- 85

5.1. Introduction ----- 85

5.2. PRIMEGENSv2 Framework ----- 86
5.2.1. PRIMEGENSv2 Input ----- 88
5.2.2. PRIMEGENSv2 Output ----- 90

5.3. PRIMEGENSv2 Graphical User Interface ----- 91
5.3.1. Data input panel ----- 92
5.3.2. Execution option panel ----- 93
5.3.3. Execution display panel ----- 94
5.3.4. Result visualization panel ----- 95

5.4. Application of PRIMEGENS: Cancer Epigenetics ----- 95
5.4.1. Introduction ----- 96
5.4.2. Implementation ----- 98
5.4.3. Experimental Validation ----- 104
5.4.4. Conclusion ----- 106

6. REFERENCES ----- 109

7. VITA ----- 123

List of Tables

Table 1: High-throughput data and measurement of gene relationship ---------- 10

Table 2: Decision table for function prediction ---------- 42

Table 3: Selection of microarray datasets for the yeast study. ---------- 44

Table 4: Selection of microarray datasets for the human study. ---------- 49

Table 5: List of all tissue groups used for meta-analysis. ---------- 70

Table 6: Comparative analysis of Arabidopsis networks of ~40K (A) and ~70K (B) sizes. ---------- 73

Table 7: Predicted E2F-targets evaluation from ~12K-size network with other studies. - 77

Table 8: GO term enrichment analysis of 178 predicted E2F-target genes. ---------- 79

Table 9: Global regulators from ~12K network having most target genes. ---------- 82

Table 10: Summary of comparison of primer design software---------- 107

List of Figures

Figure 1: Coding high-throughput biological data into a functional-linkage network ---- 2

Figure 2: A graph of query gene and its neighbours based on sequence similarities------- 3

Figure 3: Reconstruction of regulatory network using microarray data ----------10

Figure 4: Combining results from multiple sources to estimate pair-wise correlation----14

Figure 5: Gene pair-specific meta analysis and significance level calculation----------19

Figure 6: Dataset level meta analysis----------21

Figure 7: Network level meta analysis ----------22

Figure 8: Top down analysis of problem of gene function prediction ----------26

Figure 9: kernel density of observed $\hat{t}$ statistics (dashed/blue) with theoretical density (solid/red).----------28

Figure 10: Modular framework for gene function prediction model ----------30

Figure 11: Local function prediction algorithm----------40

Figure 12: Conditional probability of functional similarity given an individual p-value (on log scale) for a single dataset or given the meta p-value for the multiple datasets for yeast. ----------46

Figure 13: Performance comparison between single dataset versus meta-analysis in yeast. ----------47

Figure 14: Conditional probability of functional similarity given an individual p-value (on log scale) for a single dataset or given the meta p-value for the 13 sets for human study.----------50

Figure 15: Prediction performance of single dataset (in blue) versus meta-analysis (in red) in human.----------51

Figure 16: Classes of annotated genes that demonstrate expression similar to both SIN1 and PCBP2. ----------53

Figure 17:. Kinetic model for time lag in TF-target regulation.----------61

Figure 18: Global regulatory network with 4968 nodes (genes) and 12,300 edges for Arabidopsis.----------71

Figure 19: A cluster identified using MCODE. Black node is TF and red node is target gene. ----------80

Figure 20: Basic PRIMEGENS model. ----------87

Figure 21: Input database format. ----------90

Figure 22: PRIMEGENS software user operation flow chart----------91

Figure 23: PRIMEGENS main input window----------92

Figure 24: Primer design specification/option window ----------93

Figure 25: Primer design run time display----------94

Figure 26: Alternate Primer design results display ----------95

Figure 27: PRIMEGENSv2 algorithm flow chart ---------- 101

Figure 28: A CpG island can be located near the TSS in three different ways. --------- 102

Figure 29: Partitioning method to cover the CpG island region located far from the TSS.
---------- 103

Figure 30: Experimental validation of the PCR primers designed by PRIMEGEN. ---- 105

1. Introduction

1.1. Functional Genomics

Determining function of a gene is the central problem in the field of functional genomics. According to Wikipedia Functional genomics is a field of molecular biology that attempts to make use of the vast wealth of data produced by genomic projects (such as genome sequencing projects) to describe gene (and protein) functions and interactions. Genomics experiments have been proven very valuable in making hypotheses, which can be tested by biologist using follow-up experimentations. Computational predictions and annotations of gene functions can serve as a statistically sound form of triage, focusing experimental resources on predictions, which are likely to be true. Among strong predictions, the most interesting can be chosen by individual biologist with intuition and domain specific knowledge.

Model organism databases in the Gene Ontology (GO) Consortium (e.g., SGD, FlyBase, and MGI) track the types of evidence that support functional annotations. According to these evidences from GO database, a substantial fraction of annotated genes are annotated solely by virtue of predictions. When Biologists browse functional annotation database for any gene, they react negatively to annotations that are presented as in silico prediction but not confidently known to be true. On the other hand, when a prediction of gene function is placed alongside conclusions derived from direct experimentation, it is assigned higher confidence or labeled clearly as prediction to avoid misleading. To address this issue, model organism databases have developed evidence codes to label

annotations based on prediction or uncritical transfer from other sources. Furthermore, the tolerance of researchers for false positives depends on a complex tradeoff between the importance of the biological question and the cost of follow-up experiments. Thus, to achieve their full potential value predictions should be provided with interpretable levels of confidence, e.g., an estimated probability that the prediction is correct.

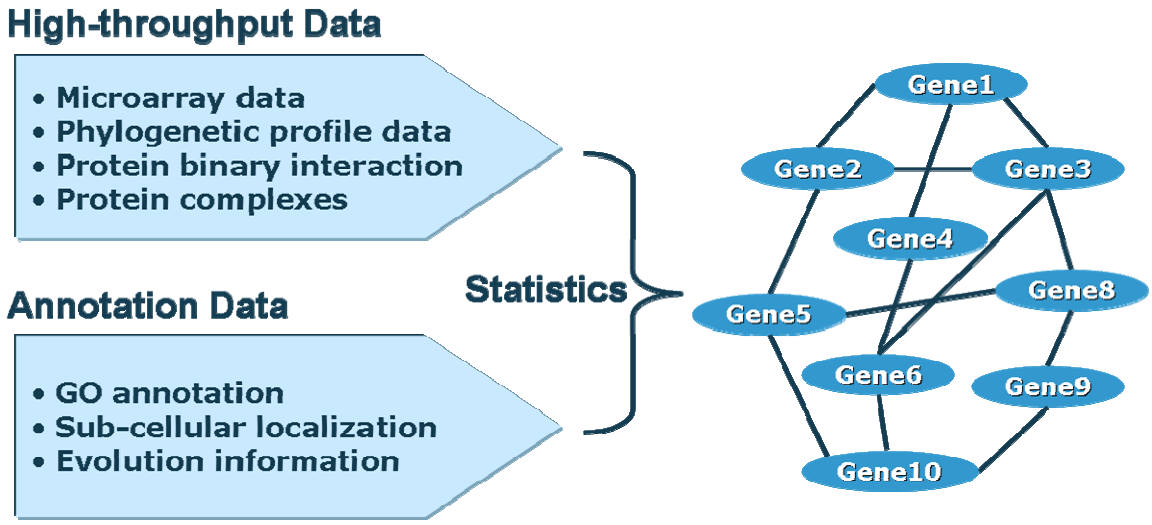

Figure 1: Coding high-throughput biological data into a functional-linkage network

Characterizing gene function is one of the major challenging tasks in the post genomic era. With a myriad of high-throughput data like microarray gene expression profiles, Phylogenetic profiles, protein–protein interactions, protein complexes, and annotations of known proteins, integration of different types of data to draw functional inferences has become the method of choice today to understand the biological mechanism. Amongst these different types of quantitative data, microarray profiles are well known for inferring functions for unknown genes. These profiles are systematic measurements of gene

expression across the whole cells or tissues under different experimental conditions or over a time course. In the past few years, hundreds of research institutes have collected a huge number of microarray data. Many microarray data are available publicly. Gene Expression Omnibus (GEO) from NCBI is a public repository for gene expression data. It supports retrieval and some analyses of gene expression data from any organism. Stanford Microarray Database (SMD) is also a database for microarray gene expression data. It provides various basic analysis capacities for the data. These microarray data for various organisms have provided a rich opportunity for computational analyses of gene products. Multiple microarray datasets can be useful for functional inferences in terms of reducing noise and resulting in a significant addition of sensitivity to extract information from these data.

Figure 2: A graph of query gene and its neighbours based on sequence similarities

Although microarray gene expression data become fundamentally important resources in today's biology, mining microarray gene expression data for discovering new biological knowledge is still challenging. In many cases, the information-rich microarray data are heterogeneous in nature, noisy and incomplete, and often contain misleading outliers, which creates a high likelihood of producing "false positives" in biological inference (Piatetsky-Shapiro G, 2003). In particular, the number of genes far exceeds the number of time points (or conditions) in a dataset, making the problem of elucidating biological information an ill-posed one. Multiple microarray datasets can be useful for functional inferences in terms of reducing noise and resulting in a significant addition of sensitivity to extract information from these data.

A number of studies on meta-analyses of integrating multiple sets of microarray data are reported in literature. Park et al. (Park, et al., 2006) extended the ANOVA model to account for an additional variability resulting from many confounding variables from different microarrays. Rhodes et al. (Rhodes, et al., 2002; Rhodes, et al., 2004) demonstrated a statistical model for performing meta-analysis of independent microarray datasets. Culhane et al. (Culhane, et al., 2005) developed a co-inertia analysis method, which is a multivariate method that identifies trends or co-relationships in multiple datasets that contain the same samples. Warnat et al. (Warnat, et al., 2005) performed cross-platform classification of multiple microarray datasets to identify discriminative gene expression signatures and predict disease associated genes. Stevens and Doerge (Stevens and Doerge, 2005) proposed a statistics-based meta-analytic approach for understanding genes' relationships to specific conditions of interest. Huttenhower et al.

(Huttenhower, et al., 2006) proposed a scalable Bayesian framework for predicting gene functional relationship using multiple microarray datasets. Other related studies include Choi et al. (Choi, et al., 2003) and Jiang et al. (Jiang, et al., 2004). Furthermore, gene expression patterns across different species have been compared (Grigoryev, et al., 2004; Schlicht, et al., 2004). While these studies focused on finding differentially expressed genes, some other studies focused on functional prediction and classification. Alter et al. (Alter, et al., 2000) used singular value decomposition in transforming genome-wide expression data to classify genes into groups of similar functions. Zhou et al. (Zhou, et al., 2005) introduced a 2^{nd}-order expression analysis approach to identify functionally related genes using the cross-platform integration of yeast microarray dataset. Stuart et al. (Stuart, et al., 2003) developed a computational method to analyze cross-species microarray data to identify gene interactions that are evolutionary conserved and inferred biological functions for meta-genes amongst yeast, worm, fly, and human. Reverter et al. (Reverter, et al., 2004) proposed an approach for combining multiple studies using multivariate mixed models, with the assumption of a nonzero correlation among gene across experiments, while imposing a null residual covariance.

Despite these studies of integrating multiple sets of microarray data for meta-analysis, most of them focused on finding differentially expressed genes. Some of the studies address gene function, but mostly characterizing the functional relationship between genes, instead of explicitly predicting each gene's function. In addition, these studies typically used yeast as the model organism, which is simple and the data are often clean. Generalization to more complex organism like human is often challenging. Clearly, more

developments are needed to take advantage of tremendous amounts of public microarray data for biological discovery.

1.2. Systems Biology

Plants often respond and adapt to different environmental stresses, such as drought, cold and chemicals through various transcriptional regulatory systems (Shinozaki, et al., 2003). Identification of these regulations not only enhances our knowledge of biological processes in plants, but also helps a great deal in developing bio-engineered crops that can better sustain challenging environments (Kasuga, et al., 1999). Typically, a handful of key transcription factors (TFs) control various biological pathways by regulating downstream target genes. In many cases, these target genes share functions or pathways. While basic ideas of these TFs and their target genes' general functions may be known, lack of knowing explicit target genes often limits the experimental design for validating intuitive hypotheses or developing new crop traits. A comprehensive list of putative targets of a TF could be used to provide more insight of a key TF through functional enrichment analysis or mapping these target genes into different biological pathways.

High-throughput expression profiling experiments (Schena, et al., 1995) have generated a large amount of data that makes it possible to develop computational approaches for predicting regulatory relations. Public repositories like NCBI Gene Expression Omnibus (GEO) (Barrett, et al., 2005; Barrett, et al., 2007), SMD (Stanford Microarray Database) (Marinelli, et al., 2008), TAIR (Poole, 2007), etc. contain extensive microarray data from time series, developmental stages, genetic interventions or manipulative treatments for

Arabidopsis thaliana, a model organism for plants (Kilian, et al., 2007; Schmid, et al., 2005). These data as well as ChIP-chip data have been used to study interactions of TFs to their downstream genes (Buck and Lieb, 2004; de la Fuente, et al., 2002; Lee, et al., 2002; Yugi, et al., 2005) . However, mining microarray data for discovering complicated regulatory relationships is still challenging partially due to the fact that these data are often incomplete, noisy, and contain misleading outliers, all of which likely produce false positives in biological inferences.

Many computational approaches for predicting genome-wide targets of a TF are based on finding co-occurrence of TFs and their targets. These include Standard Pearson correlation technique to measure statistical significance of synchronous co-regulation of genes and order of regulation (Markowetz and Spang, 2007). However, correlation coefficient is a weak criterion for measuring dependence and can lead to many false positives in predicting TF targets (Brazhnik, et al., 2002). Another approach is Graphical Gaussian Model (GGM) based on the concept of partial correlation for learning high-dimensional dependency networks from genomic data (Toh and Horimoto, 2002; Wichert, et al., 2004), which is valid when number of genes is comparable to number of samples in the microarray data (Schafer and Strimmer, 2005; Wille and Buhlmann, 2006). One way to avoid this limitation is to use GGM with regularization and moderation, which is implemented as an R package *GeneNet* (Opgen-Rhein and Strimmer, 2007; Schafer and Strimmer, 2005). This method has been used to infer genome-scale regulatory network for *A. thaliana* transcriptome (Ma, et al., 2007). Some other methods are based on probabilistic models, such as the Bayesian network

(Friedman, et al., 2000) and regression tree method (Segal, et al., 2003). Such methods cannot be directly applied to many time series expression profiling data, because the apparent time lag between initiation of a TF and activation of its targets is not accounted in these models. For example, a study suggests a clear time lag between the mRNA levels of a TF, CBF and its known targets (Seki, et al., 2002). In part, the time lag is used to translate the mRNAs of a TF into proteins before the proteins can act on activating/repressing TF's targets. To address this issue, it is important to adjust time-series transcription profiling data for detection of TF-target relationship (Spellman, et al., 1998).

Another group of methods to identify TF-target genes are specifically designed for time-series expression profiling data, including a method based on Needleman-Wunsch algorithm (Filkov V., 2001) and a dynamic probabilistic model based on chemical kinetics and linear differential equations (Chen, et al., 1999). The dynamic probabilistic model, introduced by Friedman et al. (Friedman, et al., 2000), is able to learn the kinetic parameters of TFs binding to their target promoters and the structure of gene regulation network simultaneously. However, it requires estimation of a large number of parameters, and it does not provide an explicit way of identifying TFs' targets from predicted active regulator's protein profiles. The linear differential equation model by Chen et al. (Chen, et al., 1999) describes the production and degradation of all mRNAs and their corresponding proteins with equations of chemical kinetics. While it is an interesting and promising theoretical approach, it tends to be very complex and requires concentration measurements of both mRNA and protein, at least at the initial state.

Many existing studies for retrieving regulatory information use a large collection of microarray data. A potential problem in using microarray data this way is ignoring the heterogeneity in topology of regulatory network due to biological/experimental factors, which could be different tissues, developmental stages or artificial treatments. A specific tissue type often has its own set of genes expressed to keep its identity. This may lead to different sets of target genes regulated by the same TF. In our approach, we addressed these issues by performing tissue-wide meta-analysis of expression pattern in at least certain number of tissue types out of all tissue types as shown in Figure 1. Such an approach allows us to identify recurring and stable regulatory relationships under multiple biological conditions. To avoid the risk of biasing towards housekeeping genes, which are expressed in all tissues all the times, we consider only those genes whose expression profiles are differentially expressed in at least one tissue. The novelty of this approach lies in combining the meta-analysis technique to find consensus regulatory interactions with the kinetic model to estimate the time lag between a TF and its associated targets. The scope of our work is smaller than general regulatory network construction, as we are only interested in recurring targets of known TFs. The reduced scope is practically useful and makes the problem more tractable. We chose the model plant *Arabidopsis Thaliana* for this work given its rich availability of biological data and knowledge.

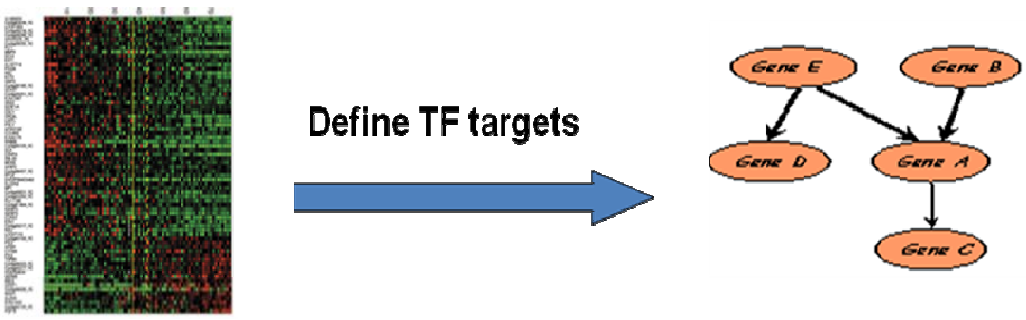

Figure 3: Reconstruction of regulatory network using microarray data

1.3. Biological inferences from high-throughput data

There is rich information contained in high-throughput data to characterize functional relationships among genes and to predict gene function. In order to explore relation between high-throughput data and biological knowledge, we used four different types of high-throughput data and quantified the pair-wise distance between two genes by using different distance measures for different types of data as shown in Table 1.

Date Type	Method
Gene Expression	Pearson correlation coefficient
Protein-Protein Interaction	Binary (0,1)
Phylogenetic Profile	Pearson correlation coefficient
Protein Domain	Maryland-Bridge coefficient

Table 1: High-throughput data and measurement of gene relationship

In the following, we will discuss the method to characterize the gene-gene relationship from each data type.

- **Gene Expression**

  Gene expression is recorded as an n x m matrix with n genes, each of which has m experimental conditions or time points. We used the Pearson correlation coefficient, r, as the pair-wise measure of the linear relationship between two gene profiles. The following equation measures the Pearson correlation between profiles X and Y:

$$r = \frac{m\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{m\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{m\left(\sum y^2\right) - \left(\sum y\right)^2}} \tag{1.1}$$

- **Protein-Protein Interaction**

  Protein-protein interaction data is recorded as an n x n matrix I for n genes. If two proteins i and j have an interaction, $I_{ij} = 1$, otherwise $I_{ij} = 0$.

- **Phylogenetic Profiles**

  A phylogenetic profile is a string that encodes the presence or absence of a homologous gene in a set of genomes. It is represented by an n x p matrix, where n is the number of homologous genes (orthologs) considered and p is the number of organisms used to generate the profile. The Pearson correlation coefficient is used as a distance measure for phylogenetic profiles.

- **Protein Domains (Pfam and InterPro)**

  The protein domain data are represented by an *n x d* binary matrix, where *n* is the number of genes and *d* is the number of domains. We calculated the Maryland Bridge distance (Glazko, et al., 2005) to characterize the relationship between domain profiles as follows,

$$S_{ab} = \frac{X_{ab}}{2}\left(\frac{1}{X_{aa}} + \frac{1}{X_{bb}}\right) \tag{1.2}$$

  where X_i represents the binary vectors of gene *i* corresponding to the *i*-th row of the matrix and $X_{ij}=X_iX_j$ is the dot product of two vectors.

2. Meta-analysis of Microarrays

2.1. Background

In recent years, a plethora of meta-analyses have emerged in biological science specifically with high-throughput data like microarrays. The need of inferring useful biological information affecting basic biological processes has fostered the momentum towards combining research from multiple studies. Meta-Analysis refers to the statistical analyses that are used to synthesize summary data from a series of studies. If the effect size (or treatment effect) is consistent across all the studies in the synthesis, then the meta-analysis yields a combined effect that is more precise than any of the separate estimates, and also allows us to conclude that the effect is robust across the kinds of studies sampled. By contrast, if the effect size (or treatment effect) varies from one study to the next, the meta-analysis may allow us to identify the reason for the variation and report (for example) that the treatment is more effective in a particular kind of patient, or in a particular dose range.

In other non-parametric approach for meta-analysis includes combining results from different studies. That is based on combining p-values obtained independently from each source. Such omnibus methods can be used to summarize biological data when very little information is available on each study, the extreme non-parametric nature of these tests allows to infer useful information from biological experiments where no hypothesis is

known. In this research work, we used meta-analysis on microarrays data to estimate pair-wise gene functional relationship in an empirical way as shown in figure 4.

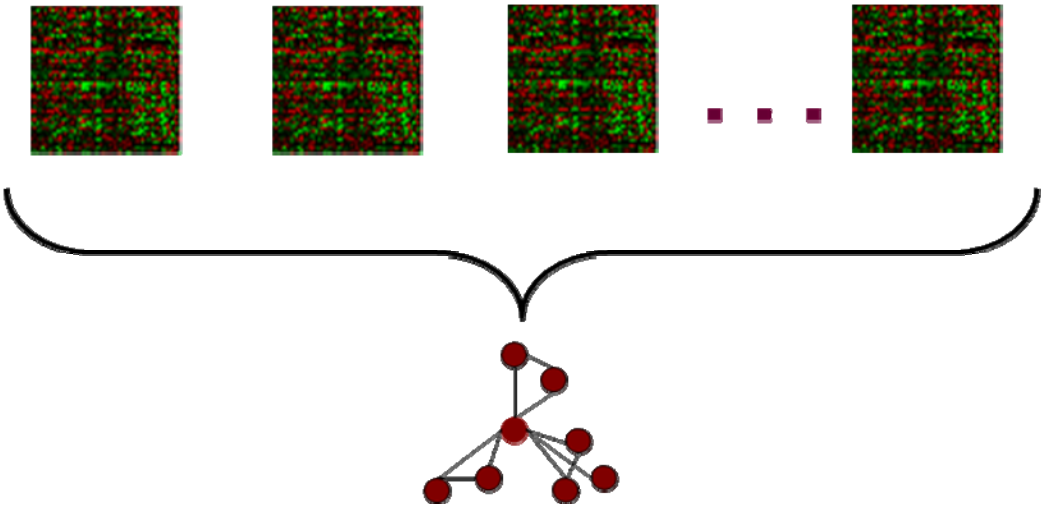

Figure 4: Combining results from multiple sources to estimate pair-wise correlation

Such method assumes that microarray datasets are obtained independently that is pair-wise relationship between a gene pair from an experiment doesn't depend upon the relationship of the same pair from another experimental source.

2.2. Significance tests for correlation

After calculating a correlation coefficient, it is usually reasonable to check its significance. Even if the variables have no correlation, for samples of finite size the correlation coefficient will be non-zero. Zero correlation coefficient is even more improbable than exactly 500 heads from 1000 coin tosses.

A correlation coefficient of zero indicates that there is no linear relationship between two variables. In order to test the significance of a correlation coefficient we can use a test statistic t:

$$\hat{T} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim T(n-2) \tag{2.1}$$

This test statistic is distributed according to a t-distribution. The correlation coefficient is considered to be statistically significant if the computed t value is greater than the critical value of a t-distribution with a level of significance of a.

Another test, which can be used, is Z-test, which assumes that gene profile (X, Y) have jointly bivariate normal distribution. In this case, the pearson correlation coefficient can be transformed such that the transformed statistics follows the normal distribution as follows.

$$W = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \quad \sim N\left(0, \frac{1}{n-3}\right) \tag{2.2}$$

These two are the most common test used for testing the significance of correlation.

2.3. Test of combing results

Here we mention the most widely used tests of significance for combining results. Each of these methods satisfies the monotonicity principle and is therefore optimal for some

testing situation. These methods can be classified into two general categories depending on the nature of the combined test statistics. Some procedures use selected ordered values p-values as the test-statistics, whereas other use linear combination of monotone functions of a p-value.

- **Inverse Chi-square method**

Perhaps the most widely used combination procedure is that of Fisher. Given k independent studies and the p-values $p_1,\ldots, p_k$

According to inverse chi-square method,

$$\left\langle \chi^2 \right\rangle = -2\sum_{i=1}^{k} \log p_i \sim \chi^2_{2k} \tag{2.3}$$

The statistics has chi-square distribution with 2k degree of freedom.

- **The Logit method**

Yet another method for combining k independent p-values is transforming p-values into a logit function, log[p/(1-p)], and then combine the logits via the statistics,

$$\left\langle L \right\rangle = \sum_{i=1}^{k} \log\left(\frac{p_i}{1-p_i} \right) \tag{2.4}$$

The exact distribution of L is not simple, but when null hypothesis is true, the distribution of L can be closely approximated by student t-distribution with 5k+4 degrees of freedom as shown in the equation (2.5) below.

$$\sqrt{3(5k+4)/\pi^2(5k+2)}\ L \sim Student-t(dof = 5k+4) \tag{2.5}$$

- **Weighted Logit method**

A modification of the logit method permits to assign different weights to all studies in the combined test statistics as shown below.

$$\langle L \rangle = \sum_{i=1}^{k} w_i \log\left(\frac{p_i}{1-p_i}\right) \tag{2.6}$$

The statistics has a distribution that can be approximated by student t-distribution. More specifially certain transform of the statistics like

$$L^* = \frac{L}{\sqrt{c}} \sim Student-t(dof = m) \tag{2.7}$$

has t-distribution with m degrees of freedom, where

$$c = \frac{3m}{(m-2)\pi^2 \sum_{i=1}^{k} w_i^2}; m = 4 + 5\frac{\left(\sum_{i=1}^{k} w_i^2\right)^2}{\sum_{i=1}^{k} w_i^4} \tag{2.8}$$

2.4. Implementation

We used non-parametric approach such that we first calculate p-value of a gene pair from each individual microarray datasets and then using the combine test procedures, we calculate the meta-statistics. Since during microarray data pre-processing, the datasets become heterogeneous, it is required to do meta-analysis dynamically for individual pair as explained in next section. We also calculate the significance level of a gene pair as explained below.

2.4.1. Dynamic meta-analysis and significance level

The pre-processed microarray dataset is heterogeneous in nature. That is the set of genes for which expression profile exist, is not same across different dataset. It is mostly due to 30% removal of genes, which are counted as noise in each dataset. This problem leads to the solution of doing meta analysis independently for each gene pair based on its presence across different datasets out of all datasets. Following figure shows an example of pair-wise dynamic meta analysis for two a gene pairs, which appear in different datasets differently (as shown in figure). The figure shows (as shown in the table in the figure) that due to heterogeneity of the microarray datasets gene pair (G2, G5) appear only in 3 out of 6 datasets simultaneously. So meta analysis for this gene pair means combining evidences from these 3 datasets only on the other hand meta analysis of gene pair (G3, G4) includes only 2 out of 6 datasets.

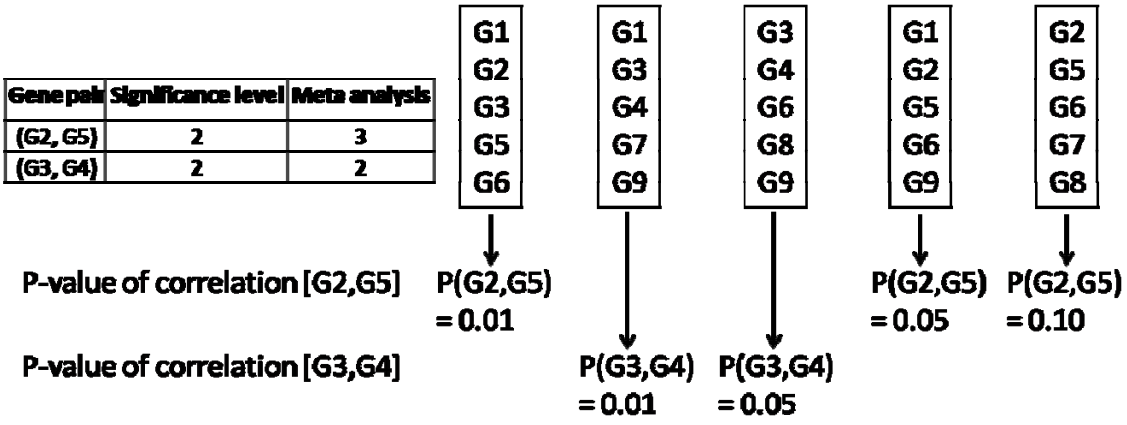

Figure 5: Gene pair-specific meta analysis and significance level calculation

Significance level of a gene pair is defined as count of datasets, which shows significant evidence of correlation for the gene pair based on associate p-value. For example as shown in the figure, gene pair (G2, G5) shows significant correlation (p-value < 0.05) only in 2 out 3 datasets in which the gene pair is present simultaneously. Same is true for other gene pair that is (G3, G4), which appear in 2 datasets only (out of total 6).

The p-value approach only uses the binary information about the relationship of a gene pair, namely, it only tells whether a gene pair is significantly correlated or not and doesn't measure the strength of the relationship. Hence just combining the individual p-values will lose information in regard to the strength of the correlation. To keep the information about the strength of the relationship, one can combine the individual Pearson correlation to obtain a meta correlation coefficient. The individual Pearson correlations need to be converted to a standard normal metric using Fisher's z transformation

$$Z_{ri} = \frac{1}{2}\log\left(\frac{1+r_i}{1-r_i}\right) \tag{2.9}$$

and then a weighted average of these transformed scores is calculated as follows:

$$\overline{Z}_r = \frac{\sum_{i=1}^{k}(n_i - 3)Z_{ri}}{\sum_{i=1}^{k}(n_i - 3)}, \tag{2.10}$$

where the weights are the inverse variances of each Pearson correlation and are the optimal weights that minimize the variance. To obtain the meta correlation, the above weighted average z score need to be transformed back to the original scale of correlation by

$$meta - r = \frac{e^{2\overline{Z}_r} - 1}{e^{2\overline{Z}_r} + 1} \tag{2.11}$$

The meta correlation coefficient could also be used as co-expression statistics to calculate the conditional probability that two genes have the same function.

2.4.2. Dataset level meta-analysis

There are two types of meta-analysis, which could be performed with microarray for any biological inference like network construction. The first one is dataset level meta analysis and second is network level meta analysis. As shown in figure 6 dataset level meta-analysis combine p-values of same gene pair correlation to estimate final correlation of the same gene pair.

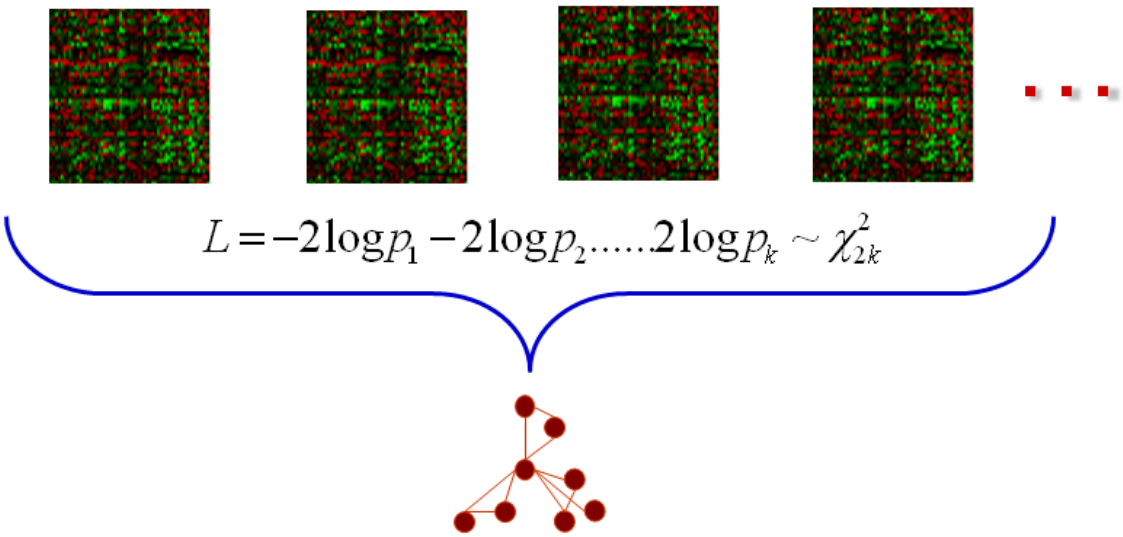

Figure 6: Dataset level meta analysis

The limitation of this method is that same data type should be chosen for meta analysis. Not only that microarray data should be from platform and addressing the same biological hypothesis in order to accurately apply the statistics procedure.

2.4.3. Network level meta-analysis

Network level meta-analysis is more flexible that dataset level meta-analysis as it can allow using different methods with different data type or microarray from different platforms. As shown in figure 7 network level meta analysis can use different methods for inferring biological network from different data sources. If used properly this type of meta analysis could result incorporating different types of high-throughput data to combine for any biological inference.

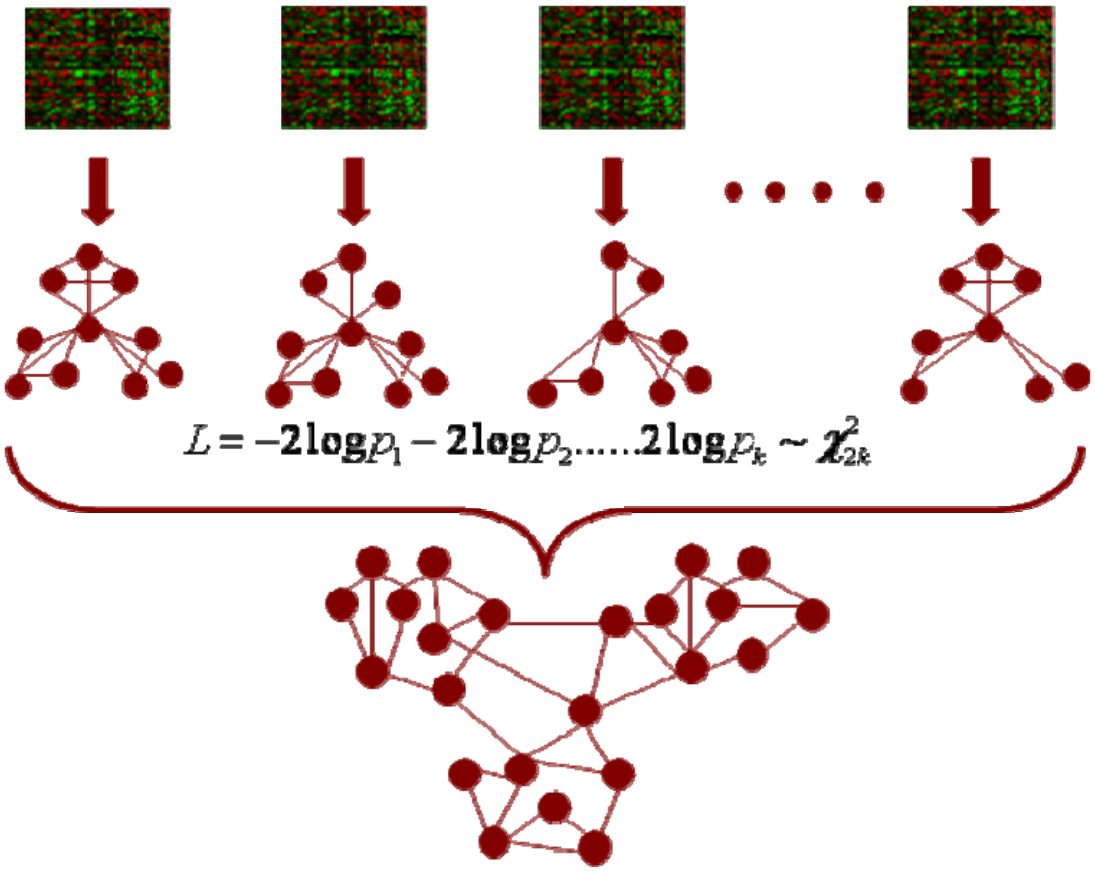

Figure 7: Network level meta analysis

One more difference between these two types of meta-analysis is the control of parameters at different levels. For example in case of dataset level meta-analysis, researchers have control at microarray data level once, the biological network is reconstructed, that will be the final network. Where as in case of network wide meta analysis, there is one more level of control at network level derived from individual microarray sources.

3. Meta-analysis Application: Functional Genomics

3.1. Gene function prediction

An immediate challenge in the post-genomic era is to assign an appropriate biological function to each protein encoded by the genome. There are different types of the functional annotation. A particular gene product can be characterized with respect to its molecular function at the biochemical level (e.g. cyclase or kinase, whose annotation is often more related to sequence similarity and protein structure) or the biological process which it contributes to (e.g. pyrimidine metabolism or signal transduction, which is often revealed in the high-throughput data of protein interaction and gene-expression profiles).

For many genes identified in a typical genome, their functions are not known. For example, only one-third of all 6200 predicted yeast genes were functionally characterized by genetic or biochemical techniques at the time when the completed sequences of yeast first become available in 1996. Even considering the additional 600 genes that could be identified based on homologs of known functions in other organisms, it left about 3500 genes with unknown functions. At this time, there are still about 2800 genes with unknown function. Many computational approaches for protein function assignments have been developed. The classical approach to infer function is based on sequence similarity, using sequence alignment tools such as FASTA and PSI-BLAST. In human

genome, about 60% of the predicted proteins were functionally annotated using this sequence-comparison method.

With an increasing number of completed genomes becoming available for comparative studies, new computational methods of protein function prediction (such as gene context (Huynen, et al., 2000), Rosetta-Stone method (Marcotte, et al., 1999), phylogenetic profiling method (Pellegrini, et al., 1999) have been developed. The underlying hypothesis of these method is that proteins evolving in a correlated fashion are functionally linked, i.e., belonging to the same physical complex, sharing the same pathway, or performing as an enzyme and its regulator. Gene context represents the positional association of genes, such as operon in prokaryotic genomes, which can be used to detect the functional association of proteins. The Rosetta-Stone method is applied to find two proteins A and B in one organism that are expressed as one fused protein in some other species. Because proteins A and B typically have no significant sequence similarity, this type of functional linkage may not be detected by a simple homology search. Phylogenetic profiling describes the pattern of presence or absence of a particular protein across a set of genomes, and uses such information to predict protein function.

When functioning, proteins rarely act in an isolated manner. Functionally related proteins often interact with each other. Hence, one possible approach to elucidate the function of an unknown protein is to investigate the functions of its interacting proteins. For this purpose, one can use the protein-protein interaction information to assign putative function for a hypothetical protein based on the '*guilt by association*' rule. For example,

if protein X (uncharacterized) is found to interact with proteins Y and Z, and both Y and Z are components of a DNA transcription processing machinery, then it is likely that protein X would also be involved in this process, perhaps being part of the complex containing Y and Z. Using the above approach, high-throughput protein-protein interaction data can provide a good coverage for many novel proteins whose functions cannot be assigned based on sequence comparison. Schwikowski et al. (Schwikowski, et al., 2000) collected 2709 published protein-protein interactions in yeast *S. cerevisiae* and clustered them based on their cellular roles and subcellular localizations annotated in the Yeast Proteome Database (YPD at http://www.proteome.com/YPDhome.html). They compiled a list of about 370 proteins with unknown functions that interact with at least one protein with known function. Among 29 of them, each has two or more interacting partners with the common function. To assign protein function by using protein-protein interaction data in a more systematic and rigorous way, a mathematical model based on the Markov random fields has been developed (Letovsky and Kasif, 2003). The MAGIC (Multisource Association of Genes by Integration of Clusters) approach to combine heterogeneous data for function assignment has been applied in yeast by Troyanskaya et al. (Troyanskaya, et al., 2003).

3.2. Break-down analysis

The problem of gene function prediction using microarray data can be divided into sub problems as shown in figure 8. Each of these sub problems have their own complexity and has been focus of research for many researchers in the part. The figure also shows specific research questions, which needs to be addressed in order to improve the gene

function prediction model. In our research work of applying meta analysis to the microarray data for gene function prediction, we analyzed these sub problems independently and justified our approach for suggesting hypothesis for each of these sub problems.

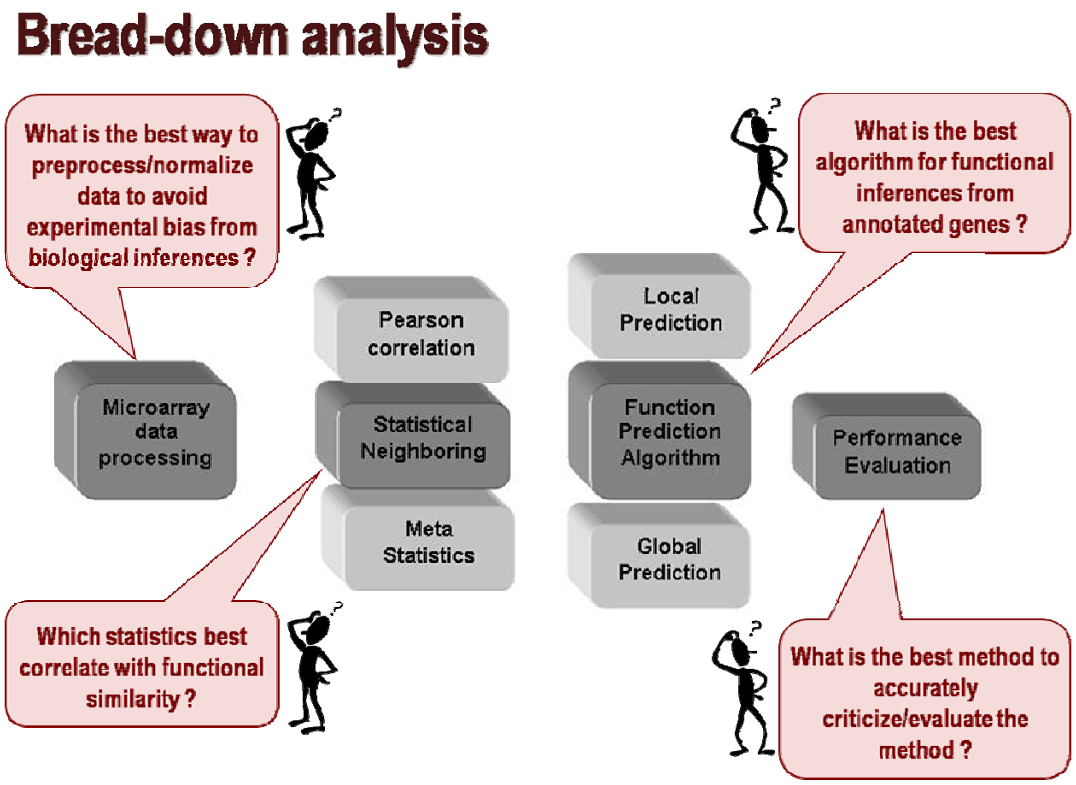

Figure 8: Top down analysis of problem of gene function prediction

This top down analysis is also important in the sense that assumptions and hypothesis used in each of these sub problems are not independent rather they are additive and affect the final hypothesis made for gene function prediction. The modular structure also provides us a way to compare different hypothesis associated with any of these modules while keeping rest of the module same. For example while comparing performance of function prediction with standard Pearson correlation coefficient method to that of meta

analysis, we kept all other hypothesis same. This analysis showed (as explained later) that meta analysis outperform the simple correlation method.

3.3. Hypothesis testing

In order to apply statistical meta-analysis in gene function prediction problem, we need to verify if the theoretical assumptions for significance test holds true. Here we took a parametric approach to obtain the combined p-value statistics (meta p-value), which is based on the assumption that the distribution of observed $\hat{t}$ statistics follows a student *t*-distribution with *n-2* degrees of freedom under the null hypothesis of no correlation between the gene pair. In order to examine this hypothesis, we randomly selected 7 yeast microarray datasets and randomly shuffled each array (column) independently and then calculate pair-wise correlation of same gene pair. After doing this shuffling for ~10000 times, we calculated ~10000 random gene pair correlations. The examination of the distributions of the observed $\hat{t}$ for all gene pairs for all datasets showed no obvious departure from this assumption as shown in Figure 9, which shows kernel density (distribution estimate) of the $\hat{t}$ statistics along with theoretical density. We tested student t-test as well as z-test for pair-wise correlation. We found that performance of t-test for randomly selected yeast microarray dataset is better than z-test. We assumed that the same will hold true for any dataset in general and selected t-statistics for calculating p-values from individual microarrays for gene pair correlation in further analysis. We also selected right-tailed p-value calculation as it has been previously shown that positive correlation is more related with function similarity where as negative correlation is not.

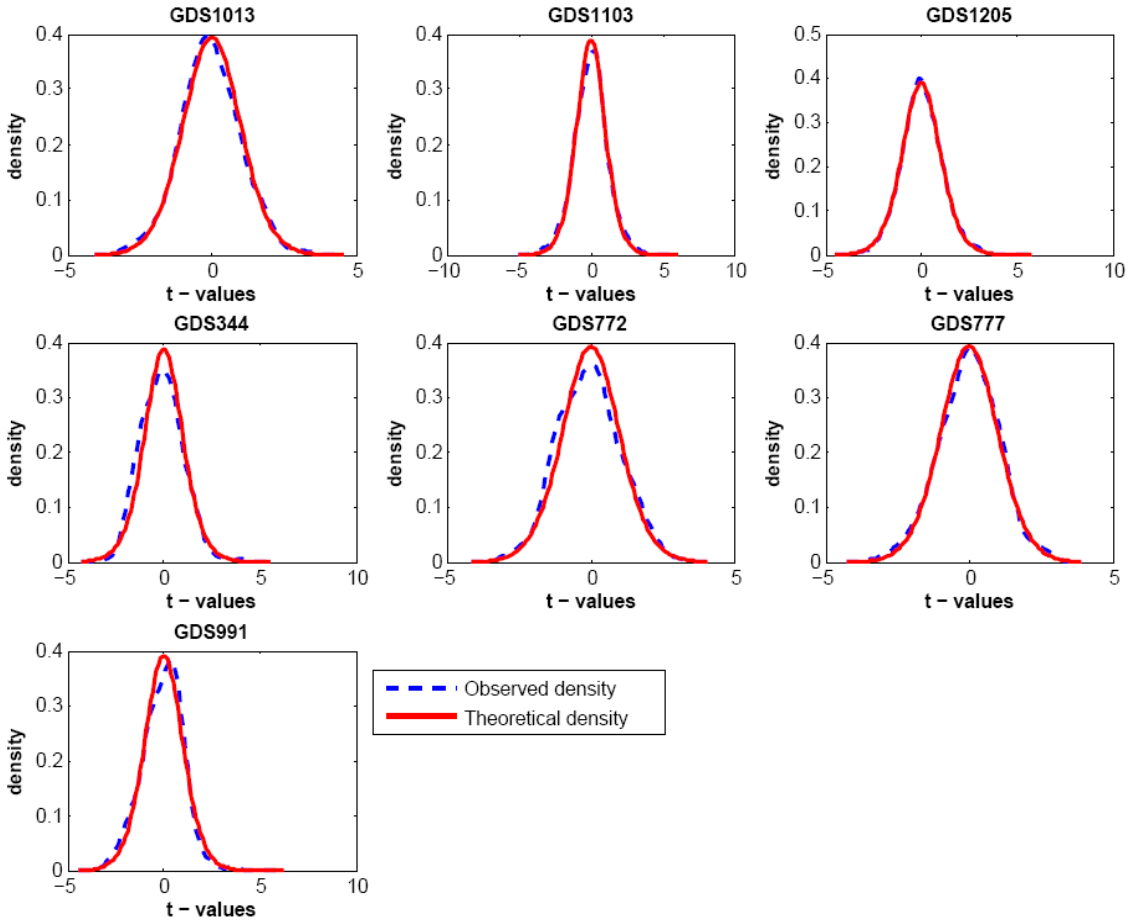

Figure 9: kernel density of observed $\hat{t}$ statistics (dashed/blue) with theoretical density (solid/red).

When this parametric assumption is a concern, individual p-values can be obtained by comparing the observed t-statistics to the ones generated by randomly permuting the rows within each column, and then the meta p-value can be obtained in the same permuted manner as done in Rhodes, et al (Rhodes, et al., 2002; Rhodes, et al., 2004). The meta p-value and the Pearson correlation coefficient will be used as co-expression statistics to calculate the conditional probability that two genes have the same function, which will in turn be used for gene function prediction.

3.4. Prototype development

In order to implement modular structure as explained in the earlier section, we proposed a software pipeline for gene function prediction. In this work, we combine the statistical meta-analysis with our previous gene function prediction methods (Chen and Xu, 2004; Chen and Xu, 2005; Joshi, et al., 2004; Joshi T, 2004; Joshi, et al., 2008) to predict gene functions explicitly on the genomic scale using multiple microarray datasets. Our method calculates the p-value of a Pearson correlation coefficient between two gene expression profiles based on the standard t-statistics in each single dataset. The p-values between two genes in single datasets are combined to obtain the joint meta-analysis p-value, which is used to quantify the posterior probability that two genes have the same function. Another approach to combine multiple microarray data is to calculate a meta correlation coefficient by combing the individual Pearson correlation coefficient. The function of a gene is predicted according to the posterior probabilities of its co-expressed gene neighbours in multiple datasets. We tested the method on both yeast and human data.

Figure 10 shows four major steps of operations performed for gene function prediction, i.e., microarray data pre-processing, statistical neighbouring and linkage graph construction, gene function prediction, and performance evaluation using a sensitivity-specificity curve. The details of each step are described below. The figure also shows important input data for each step and output produced from that step, which is used as input for the next step. We implemented various steps of our algorithmic prototype using the ANSI C language on both Linux and Windows platforms.

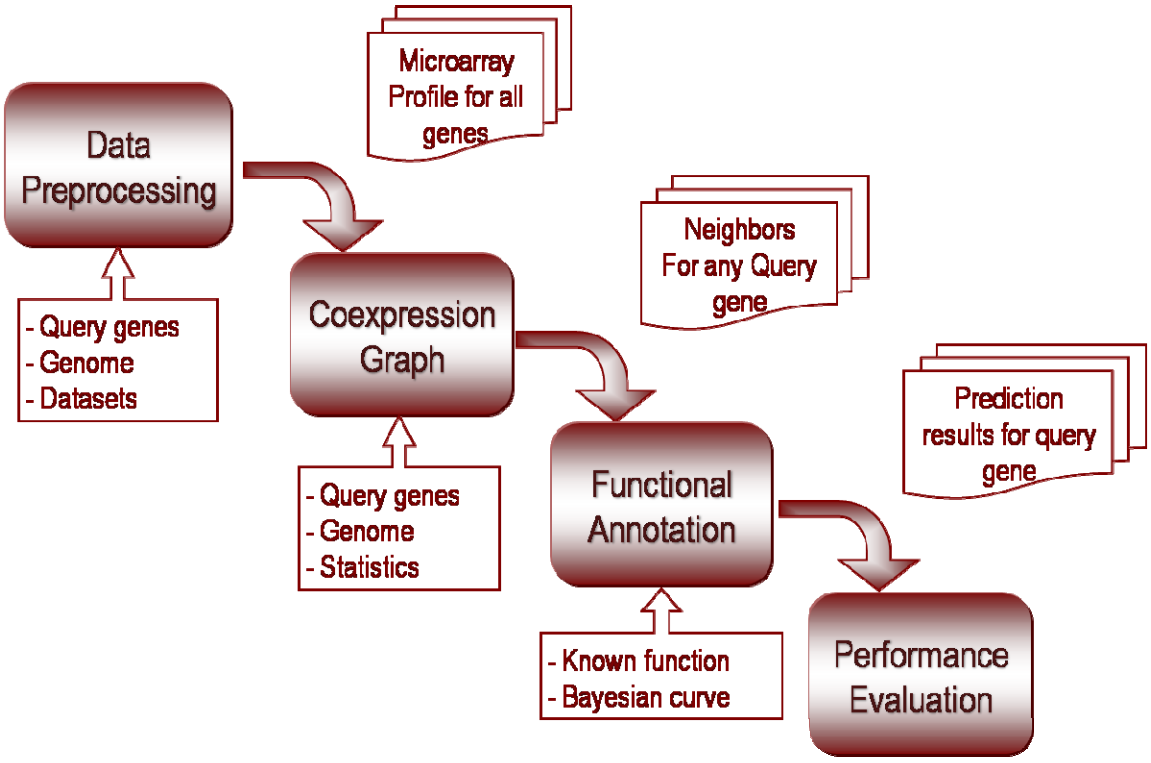

Figure 10: Modular framework for gene function prediction model

3.4.1. Microarray data processing

We selected the SOFT formatted microarray GDS datasets downloaded from the Gene Expression Omnibus (GEO; ftp://ftp.ncbi.nih.gov/pub/geo) for specific organism such as yeast or human, where a GDS dataset represents a reassembled collection of similarly processed, experimentally related hybridizations extracted from submitter-supplied records (Barrett and Edgar, 2006; Barrett and Edgar, 2006; Barrett, et al., 2005; Barrett, et al., 2007). When downloading the GDS microarray datasets from GEO, we performed the following steps:

1. **Download dataset-** We downloaded microarray datasets (released in November 2006) from the GEO FTP site. In particular, all the available GDS microarray datasets for *Saccharomyces cerevisiae* and *Homo sapiens* were retrieved. Since GDS datasets are already normalized (Barrett and Edgar, 2006; Barrett and Edgar, 2006; Barrett, et al., 2005; Barrett, et al., 2007), no further normalization is performed.

2. **Dataset selection-** Due to heterogeneous environment across different microarray platforms, not all the datasets of an organism were used for simplicity. In this study, we only used microarray data generated by the Affymetrix platforms, which have most number of datasets available. Given the variations within Affymetrix platforms, we further restricted the platform to GPL90 (Affymetrix Gene Chip Yeast Genome S98 Array YG-S98) for the yeast study and GPL96 (Affymetrix Gene Chip Human Genome U133 Array Set HG-U133A) for the human study since these two platforms have the highest numbers of arrays in respective organisms (GPL90 has 868 yeast arrays and GPL96 has 13,808 human arrays in total).

3. **Removing datasets with small number of arrays-** Datasets having small number of arrays (conditions or time points) is prone to noises in the statistical analysis. Therefore, only datasets that have enough number of array measurements (>11 conditions or time points for yeast and >50 for human) were selected for this study.

4. **Profile pre-processing-** In a selected dataset, if the expression profile of a gene contains any missing value or negative intensity value, that gene is removed. Furthermore, due to the noisy nature of the human genome datasets, the lower 30% genes that have the smallest maximum intensities are excluded for further consideration (Lee, et al., 2004). Expression levels of multiple probes corresponding to the same gene are merged by averaging them. Finally a logarithm transformation to all gene expression intensities is performed.

Once the microarray data pre-processing is done, these datasets are used to find statistically significant neighbours and generate the co-expression linkage graph for function prediction.

3.4.2. Statistical co-expression linkage graph

In order to find statistical neighbours and build co-expression linkage graph for a gene, we focussed on following topics.

- **Functional similarity measurement**

We followed the Gene Ontology annotation (Ashburner, et al., 2000) for functional description. In our study, we selected biological process functional annotation for function prediction, as biological processes are often better revealed in gene expression profiles than molecular functions and cellular components (Chen and Xu, 2004; Chen and

Xu, 2005). After acquiring the biological process functional annotation for the known gene product along with their GO Identification (ID), we assign a numerical GO INDEX to each GO identifier, which represents the hierarchical structure of the classification. The more detailed level of the GO INDEX, the more specific function assigned to a protein. The maximum level of GO INDEX is 13.

We quantify function similarity by comparing the level of GO INDICES that the two genes share. For example, if both gene-1 and gene-2 have annotated functions, assume gene-1 has a function represented by GO INDEX 2-1-8-1 and gene-2 has a function represented by GO INDEX 2-1-8-5-10. When compared with each other for the level of matching GO INDEX, they match with each other through 2-1-8, i.e., INDEX level 1 (2), INDEX level 2 (2-1) and INDEX level 3 (2-1-8). We found (based on the GO release as of November, 2006 and updated functional annotation information from the GEO site) that ~80% of the genes from yeast microarray datasets with the GPL90 platform and ~70% of the genes from human microarray dataset with GPL96 platform are annotated.

- **Co-expression statistics**

One of the important problems during the construction of functional linkage graph is find appropriate co-expression statistics. That is, how to measure the co-expression statistics between two genes given microarray expression profile, which highly correlated with functional similarity of these two genes. Since there is no answer for this problem as yet, we proposed meta statistics obtained from meta analysis and compared with Pearson correlation coefficient as obtained from single dataset.

Because genes involved in same pathway or part of the same protein complex are often co-regulated, a set of genes with similar functions often exhibit expression profiles that are correlated under a large number of diverse conditions or time points (Eisen, et al., 1998; Hughes, et al., 2000; Kim, et al., 2001; Segal, et al., 2003). Studies have shown a significant relationship between functional similarity and Pearson correlation coefficient for a given pair of genes (Chen and Xu, 2004; Chen and Xu, 2005; Joshi, et al., 2004; Joshi T, 2004; Joshi, et al., 2008). When we have multiple sets of microarray data, the Pearson correlation coefficients of all the datasets can be combined through meta-analysis (Lee, et al., 2004). One way to do so is to combine the individual p-values by calculating a meta p value. We evaluated the statistical significance of a Pearson correlation coefficient r for two gene expression profiles in a single dataset based on the standard t-statistics (Hogg RV, 2005):

$$p-value_i = P(T > \hat{t}_i), \text{ where } \hat{t}_i = \frac{r_i\sqrt{n_i-2}}{\sqrt{1-r_i^2}} \tag{3.1}$$

where T is a *t*–random variable with n_i-*2* degree of freedom and n_i is the number of conditions of the gene expression profiles. Note here we use the right-tailed p-value since our previous study (Chen and Xu, 2004; Chen and Xu, 2005; Joshi, et al., 2008) and Lee et al. (Lee, et al., 2004) showed that the negative correlation is less likely to be related to functional similarity. Therefore the p-value is monotone decreasing with the Pearson correlation and ranking gene pairs by either pvalue or the Pearson correlation gives the same result. Since we assume that the datasets are obtained independently, we apply the inverse chi-square method and obtain the meta chi-square statistics (Hedges and Olkin1985):

$$\hat{\chi}^2 = [-2\log(P_1) - 2\log(P_2) - ... - 2\log(P_n)] \tag{3.2}$$

where P_i is the p-value obtained from the i^{th} data set for a given gene pair defined in (3.1). When there is no linear correlation between a gene pair in any of the multiple datasets, the above chi-square statistics $\hat{\chi}^2$ follows a central chi-square distribution with degrees of freedom *2n* and hence the p-value for meta-analysis can be obtained by

$$meta - p - value = P(\chi^2_{2n} > \hat{\chi}^2), \tag{3.3}$$

where χ^2_{2n} is a chi-square random variable with *2n* degrees of freedom. For any gene pair of interest, we conclude that the gene pair is positively correlated in at least one of the multiple datasets at level *alpha* if the meta p-value is smaller than *alpha*.

- **Quantifying the gene function relationship based on co-expression statistics using Bayes' formula**

To quantify the functional relationship between a gene pair, we calculated using Bayes' formula the conditional probabilities of such gene pair sharing the same function at each GO INDEX level given a co-expression statistics, denoted by *M*, as proposed in our early study (Chen and Xu, 2004; Chen and Xu, 2005; Joshi, et al., 2004; Joshi T, 2004; Joshi, et al., 2008). In this paper, we use the Pearson correlation coefficient as the co-expression statistics for single dataset and the meta p-value or meta correlation coefficient for multiple datasets. Given a gene pair showing co-expression statistics M, the *posterior* probability that two genes sharing the same function at GO INDEX level S is

$$p(S|M) = \frac{p(M|S)\,p(S)}{p(M)} \tag{3.4}$$

where $p(M|S)$ is the conditional (*a priori*) probability that two genes are co-expressed in their expression profiles with statistics value M given that two genes have the same GO INDEX level S. The probability $p(S)$ is the relative frequency that a gene pair has similar functions at the given level of GO INDEX using the annotation data. The probabilities $p(M|S)$ and $p(S)$ are estimated based on the set of genes present in the given dataset platform of specific organism (yeast or human) whose functions have been annotated with the GO biological processes. The probability $p(M)$ is estimated by the relative frequency of co-expression statistics M over all gene pairs in the organism, which is calculated from the genome-wide gene expression profiles. This conditional probability will be used in calculating likelihood scores (defined later) for the set of predicted functions for each query gene. The predicted functions for each query gene are taken from the union of known functions of the neighbouring genes having direct link to the query gene.

- **Co-expression linkage graph**

The co-expression linkage graph connects gene pairs that have significant correlation based on the co-expression statistics M. For single datasets, we rank all the genes pairs using the p-value defined in (3.1) (ranking by the Pearson correlation coefficient will give the same result) and choose a fixed number of gene pairs from the top to produce the co-expression linkage graph. For multiple datasets, we rank all gene pairs based on the number of individual p-values that are significant at level 0.01 across multiple datasets (Lee, et al., 2004) and for gene pairs that have the same number of significant p-values, they are ranked by the meta correlation coefficient, which is called the meta correlation

coefficient method, or ranked by the corresponding meta chi-square statistics defined in (3.2) which is called the meta p value method (here we use meta chi-square instead of meta p-value since the meta p-value for many gene pairs are very close to zero and hard to distinguish computationally; both meta chi-square instead of meta p-value should result in the same order when the degrees of freedom for each gene pair is same). Then a fixed number of gene pairs are selected from the top to establish the co-expression linkage graph

Algorithm

Input query gene, microarray data, genome

FOR each query (*Q*)

FOR each gene (*G*) from genome (potential neighbour for *Q*)

FOR each microarray dataset (*M*)

IF *M* contains both *Q* and *G*

Select pair (*Q*, *G*)

Calculate p-value for pair (*Q*, *G*) from dataset (*M*)

ELSE

Skip dataset *M*;

END

END

Count #datasets with individual p-value<0.01 for pair (*Q*, *G*)

Calculate meta chi2 statistics & meta p-value for pair (*Q*, *G*)

Calculate meta correlation coefficient for pair (*Q*, *G*)

END

Group gene pairs using #datasets with individual p-value < 0.01

Sort groups in descending order of #datasets with p-value <= 0.01

FOR each group

Sort genes in a group in descending order of meta chi2.

Choose genes with meta chi2 > a cut-off value

Sort genes (within group) in descending order of meta corr.

Choose the genes with meta corr. > a cut-off value

END

Obtain all the statistical neighbours of the query gene (*Q*)

END

The number of gene pairs used to obtain the co-expression linkage graph can be decided in many ways. For instance, the user might simply use the top 200 gene pairs for function prediction Or Bonferroni correction can be used to obtain a threshold for individual p-value for single datasets. Or the magnitude of the Pearson correlation can be considered combined with the individual p-value as proposed by Lee et al. (Lee, et al., 2004). For the multiple datasets, the number of significant individual p-values follow a Binomial distribution ~Binomial (*n*, 0.01) if a gene pair is not correlated in any of the n datasets. Hence, this Binomial distribution can be used to obtain a cut-off value for the number of significant individual p-values. This threshold together with the cut-off value for the meta p-value using Bonferroni correction can be used to choose the gene pairs for the linkage graph. For the meta correlation coefficient method, one can choose meta

correlation coefficient above 0.7 just as suggested for single data analysis in Chen and Xu (2004). We chose the top 200, 400, 800, 1600, 3200 neighbours to obtain different co-expression linkage graphs and based on each linkage graph, we predicted gene functions and generated the corresponding sensitivity-specificity curves for meta-analysis and single data analysis. In this research work, we only show the gene function prediction results based on 200 neighbours.

3.4.3. Function Prediction algorithm

The statistical neighbours for each query gene can be obtained from the co-expression linkage graph and the union of all functions from the annotated neighbours are assigned to the query gene, each with a likelihood score (Joshi, et al., 2004; Joshi T, 2004; Joshi, et al., 2008). This method is a local prediction method compared to Boltzmann machine algorithm, which can improve the prediction power by obtaining predicted functions from genes that are not directly connected to the query gene on the co-expression linkage graph (Chen and Xu, 2004; Chen and Xu, 2005). The global prediction using Boltzmann machine algorithm is often useful for genomes that do not have many high-throughput data. However, for genomes like yeast and human where high-throughput data are rich, the local prediction method would be fast and reliable. For consistency, we only present local prediction results for both yeast and human data in this paper.

For each function that is assigned to the query gene, its likelihood score is calculated as

$$likelihood\ score(F) = 1 - \prod_{n=1}^{N} \{1 - P_n(S \mid M)\} \tag{3.5}$$

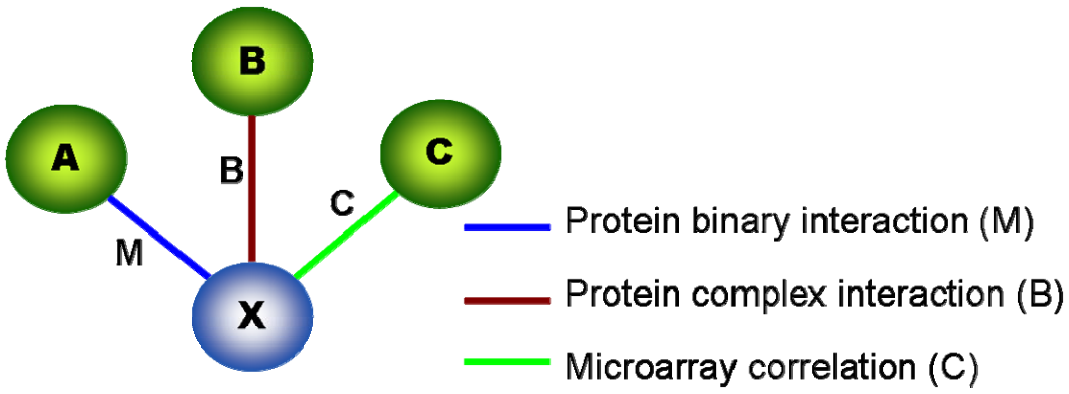

Figure 11: Local function prediction algorithm

$$P'(S_l|M) = 1 - \prod [1-P_j(S_l|M)], \qquad j = 1, 2, \ldots.nM \qquad (3.6)$$

$$P'(S_l|B) = 1 - \prod [1-P_j(S_l|B)], \qquad j = 1, 2, \ldots.nB \qquad (3.7)$$

$$P'(S_l|C) = 1 - \prod [1-P_j(S_l|C)], \qquad j = 1, 2, \ldots.nC \qquad (3.8)$$

where F is any given function, *N* is the total number of neighbours of the query gene that are annotated with the function F, and $P_n(S \mid M)$ is the conditional probability as defined in equation (4) for the n^{th} neighbour. This likelihood score is used to rank all the predicted functions for each query gene and a cut-off value can be chosen to decide the final predicted functions for each query gene. In this paper, we fix the number of predictions for each query gene to generate sensitivity-specificity curve. In particular, we use the top 50, 100, 200, 400, 800, 1600, or 3200 functions as the predicted functions for each query gene.

Algorithm for local prediction

Input query genes, annotation data and Bayesian probabilities

FOR each query gene *Q*

Read neighbours of *Q* based on the co-expression graph

FOR each neighbour *N*

IF *N* is annotated

Select *N* for function prediction of *Q*

ELSE

Skip this neighbour;

END

END

Build union of annotated neighbour's functions as predicted functions.

Assign likelihood score for each function in the union

Rank the predictions in decreasing order of the likelihood score

Choose the predicted functions at a given threshold

END

3.4.4. Performance Evaluation

Assessment of performance of different method on a standardized data set according to standardized performance criteria is the only way to draw meaningful conclusions about the strengths and weaknesses of the algorithms employed. For the testing purpose of our method, we randomly selected 500 genes from annotated yeast genes and 100 genes from annotated human genes as query genes. We predicted functions for each query gene once at a time and evaluated the sensitivities and specificities of the predictions of all query

genes using sensitivity-specificity curve. For each prediction scheme which corresponds to a particular co-expression graph and a specific cut-off value for the likelihood scores, the sensitivity and specificity are calculated according to the following definition. We consider assigning a function to a gene as a decision/prediction, which can be verified from the annotation data. There are two types of errors we can make: (1) we assign an incorrect function to a gene, which is the type I error or a false positive; and (2) we do not assign a known function to a gene, which is the type II error or false negative. On the other hand, if we assign a correct function to a gene, it is a true positive and if a gene does not have a function and we do not assign it, it is a true negative. We consider all query genes and all available GO IDs in the annotation data and summarize the results in the format of Table 2.

	Prediction: GO ID not assigned	**Prediction: GO ID assigned**
Known: GO ID not assigned	True Negative (TN)	False Positive (FP)
Known: GO ID assigned	False Negative (FN)	True Positive (TP)

Table 2: Decision table for function prediction

By changing the number of predictions selected for each query gene based on the likelihood scores at a fixed co-expression linkage graph, we can obtain a sensitivity-specificity plot, where

$$Sensitivity = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} TP_i + \sum_{i=1}^{K} FN_i}$$

$$Specificity = \frac{\sum_{i=1}^{K} TN_i}{\sum_{i=1}^{K} FP_i + \sum_{i=1}^{K} TN_i} \qquad (3.9)$$

where K is the number of query genes; and TP_i is the number of correctly predicted functions for gene i, FN_i is the number of known functions that are not predicted for gene i, and FP_i is the number of incorrectly assigned functions for gene i, and TN_i is the number of functions among all available GO IDs that are neither known nor predicted for gene i.

3.5. Implementation

We applied our method to yeast data as well as human data. With the small size of the genome and good annotations, yeast serves as the best model organism during the development of our method. Our result shows that function prediction based on multiple datasets using the proposed method improves significantly over using any single dataset.

3.5.1. Yeast datasets

We did a pilot study using 7 independent yeast microarray datasets from the GPL90 platform, including 116 experimental conditions in total for all the genes in yeast (Table 2, which shows the dataset ID, the number of conditions or time points, and the overall experimental condition.). We used microarray data of 5419 genes from the GPL90 platform, among which 4519 genes have Gene Ontology (GO) annotations. The GO biological process ontology data was downloaded from the GO web site http://www.geneontology.org, whereas the yeast genome GO annotation data was downloaded from the NCBI Gene Expression Omnibus (GEO) web site http://www.ncbi.nlm.nih.gov/geo/.

	Accession	#Cols	Experimental Condition
1	GDS 777	24	Nutrient limitation under aerobic and anaerobic condition effect on gene expression (growth protocol variation)
2	GDS 772	18	Histone deacetylase RPD3 deletion and histone mutation effect on gene regulation (genotype/variation)
3	GDS 344	11	Chitin synthesis (protocol variation)
4	GDS 1205	12	Ssl1 mutant for a subunit of TFIIH response to methyl methanesulfonate (genotype/variation)
5	GDS 1103	12	Leu3 mutant expression profiles (genotype/variation)
6	GDS 991	15	Phosphomannose isomerase PMI40 deletion strain response to excess mannose(dose variation)
7	GDS1013	24	IFH1 over-expression (time course)

Table 3: Selection of microarray datasets for the yeast study.

We plotted conditional probability of the GO functional similarity given an individual p-value (on the log scale) for a single dataset or given the meta p-value for the multiple datasets as shown in Figure 12. Although the curves did not differ substantially between a single dataset and the multiple datasets combined, the curve for meta p-value is much smoother than the curve from any single data, as the sample size for generating meta-analysis curve is higher. Hence, the estimate for the conditional probability in meta-analysis is more robust. We also found that there were many more statistically significant pairs using the same threshold for the meta p-values of multiple datasets than those of any single dataset. This suggests combing multiple datasets using the meta-analysis leads to more discerning power in establishing statistical neighbours for the query genes and hence, increases the sensitivity for function prediction.

To confirm this, we applied our function prediction method described in Materials and Methods to ~10% (500) randomly selected query genes from the yeast genome using either single datasets or multiple datasets. We compared the sensitivity-specificity plot for one single dataset and the one using all 7 datasets from Table 3. For this purpose, we selected top 200 neighbours for each query gene to generate the co-expression linkage graph using either a single dataset or 7 datasets.

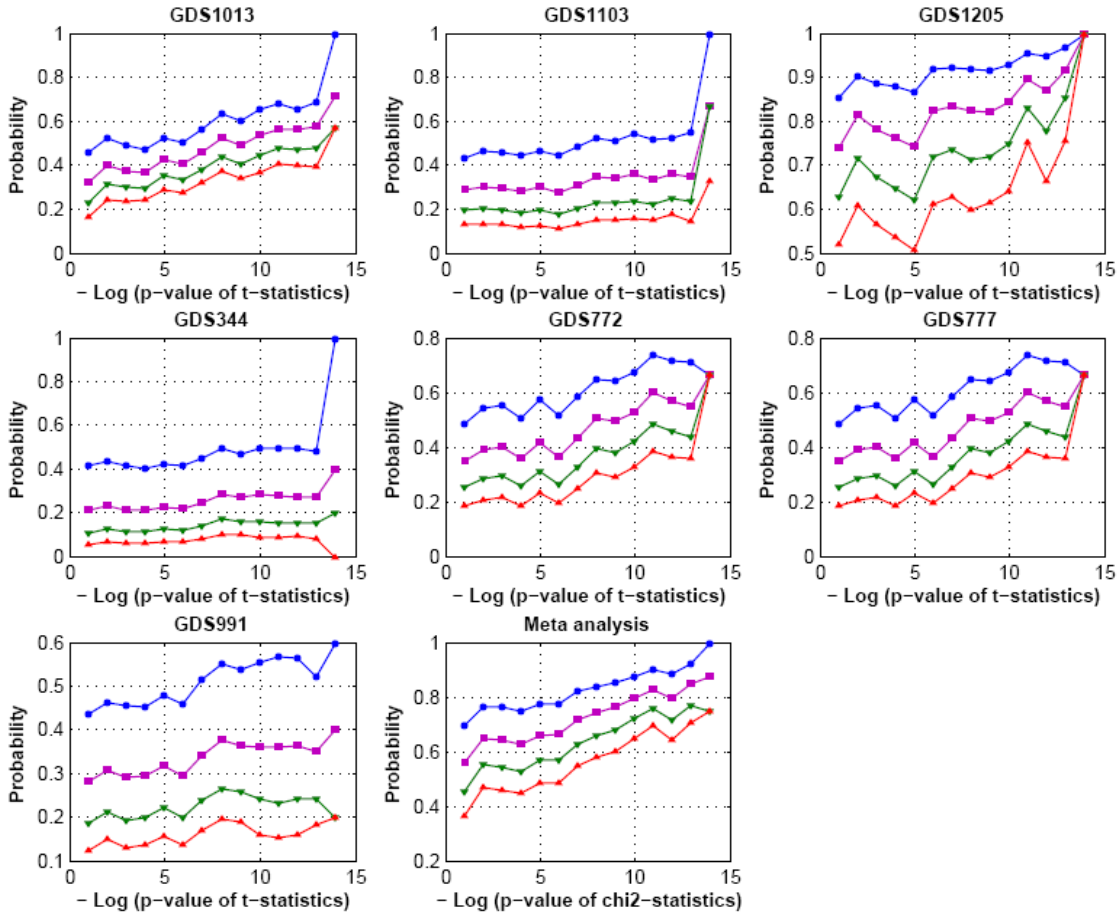

Figure 12: Conditional probability of functional similarity given an individual p-value (on log scale) for a single dataset or given the meta p-value for the multiple datasets for yeast.

Figure 13 shows performance comparison between single dataset (in blue) versus meta-analysis (in red) in yeast. In each plot, various cut-off values for the likelihood scores of the prediction functions for the query genes are used to generate different points in the sensitivity-specificity curve. In particular, the seven points correspond to using the top 50, 100, 200, 400, 800, 1600 and 3200 predictions for each query gene. According to figure 13 meta-analyses using all 7 datasets significantly improved the prediction accuracy over any individual dataset. The result suggests that the proposed method of combining multiple microarray dataset using meta-analysis works well.

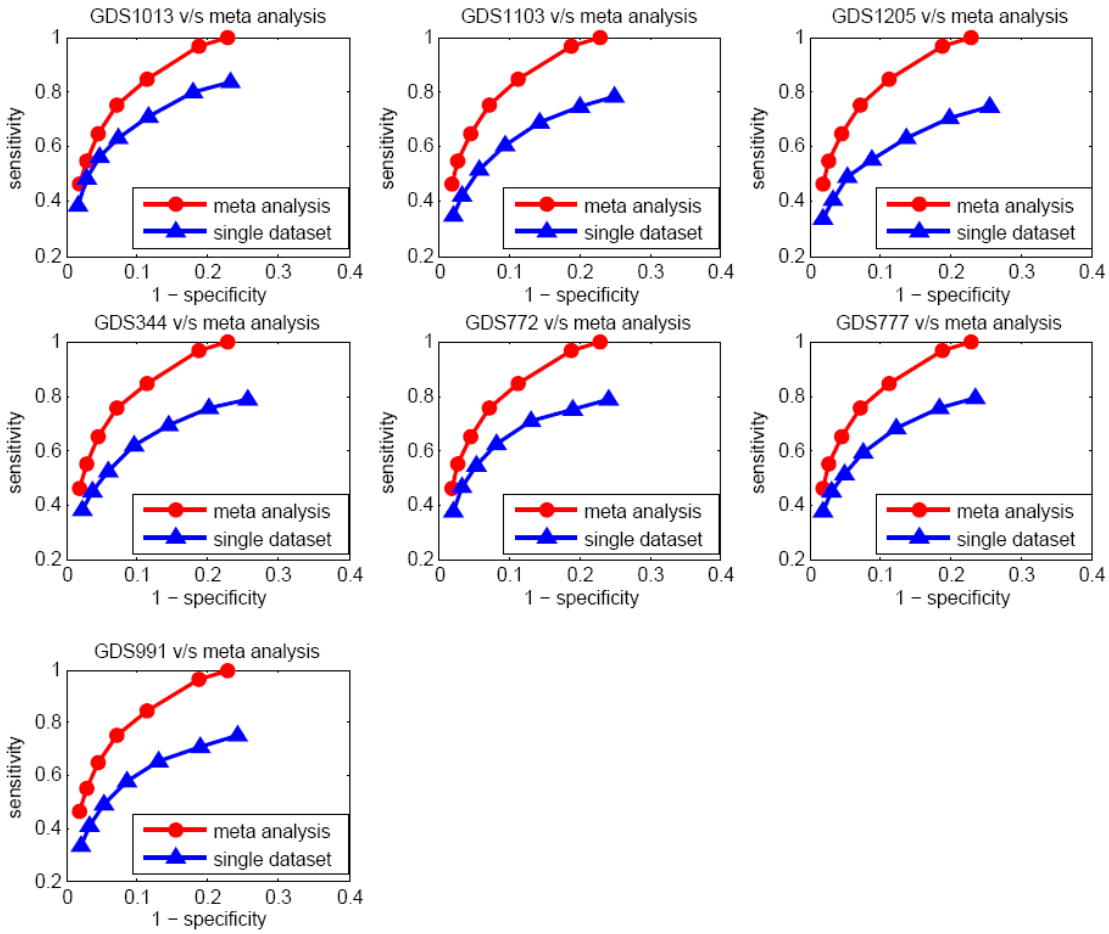

Figure 13: Performance comparison between single dataset versus meta-analysis in yeast.

3.5.2. Human dataset

We also applied our function prediction method using human microarray datasets. We randomly selected 100 genes from the annotated human genes. We predicted functions for these genes and compared the prediction performance between a single-dataset analysis and multiple-dataset analysis using sensitivity-specificity curves.

For the function prediction on human genes, we used the GPL96 microarray platform datasets (154 in total) consisting of 3198 arrays (see supplementary materials in

http://digbio.missouri.edu/meta_analyses). The GPL96 microarray platform provides 22,283 unique Affymetrix probes. Combining replicated genes with different probe identifiers, all the probes contain 13,955 unique genes in total. We selected 13 datasets (Table 4; it shows the dataset ID, the number of conditions or time points, and the overall experimental condition) from this platform, each of which has at least 50 arrays for our function predictions on human genes.

	Accession	**#Cols**	**Experimental Condition**
1	GDS1067	52	Plasma cell dyscrasias
2	GDS1209	54	Sarcoma and hypoxia
3	GDS1220	54	Malignant pleural mesothelioma
4	GDS1375	70	Cutaneous malignant melanoma
5	GDS1428	67	Neutrophil response to Anaplasma phagocytophilum infection time course
6	GDS1479	60	Carcinoma in situ lesions of the urinary bladder
7	GDS1615	127	Ulcerative colitis and Crohn's disease comparison peripheral blood mononuclear cells
8	GDS1975	85	Gliomas of grades III and IV (HG-U133A)
9	GDS2113	76	Pheochromocytomas of various genetic origins
10	GDS2190	61	Bipolar disorder dorsolateral prefrontal cortex
11	GDS534	75	Smoking-induced changes in airway transcriptome

12	GDS596	158	Large-scale analysis of the human transcriptome (HG-U133A)
13	GDS715	87	Acute myeloid leukemia cell differentiation induced by various drugs

Table 4: Selection of microarray datasets for the human study.

In order to find probabilistic relationship between functional similarity and statistical correlation, we studied all the 154 datasets by grouping them into 15 groups of 10 datasets each. The Bayesian conditional probability of functional similarity (y-axis) for given meta p-value (x-axis) for human organism using 15 groups of 10 datasets can be found in the supplementary documents at http://digbio.missouri.edu/meta_analyses. It shows that different datasets yield significantly different probability curves. For function prediction, we used probability curve generated separately for each single dataset and for meta-analysis probability curve, we generated using all 13 curated microarray data sets listed in Table 4.

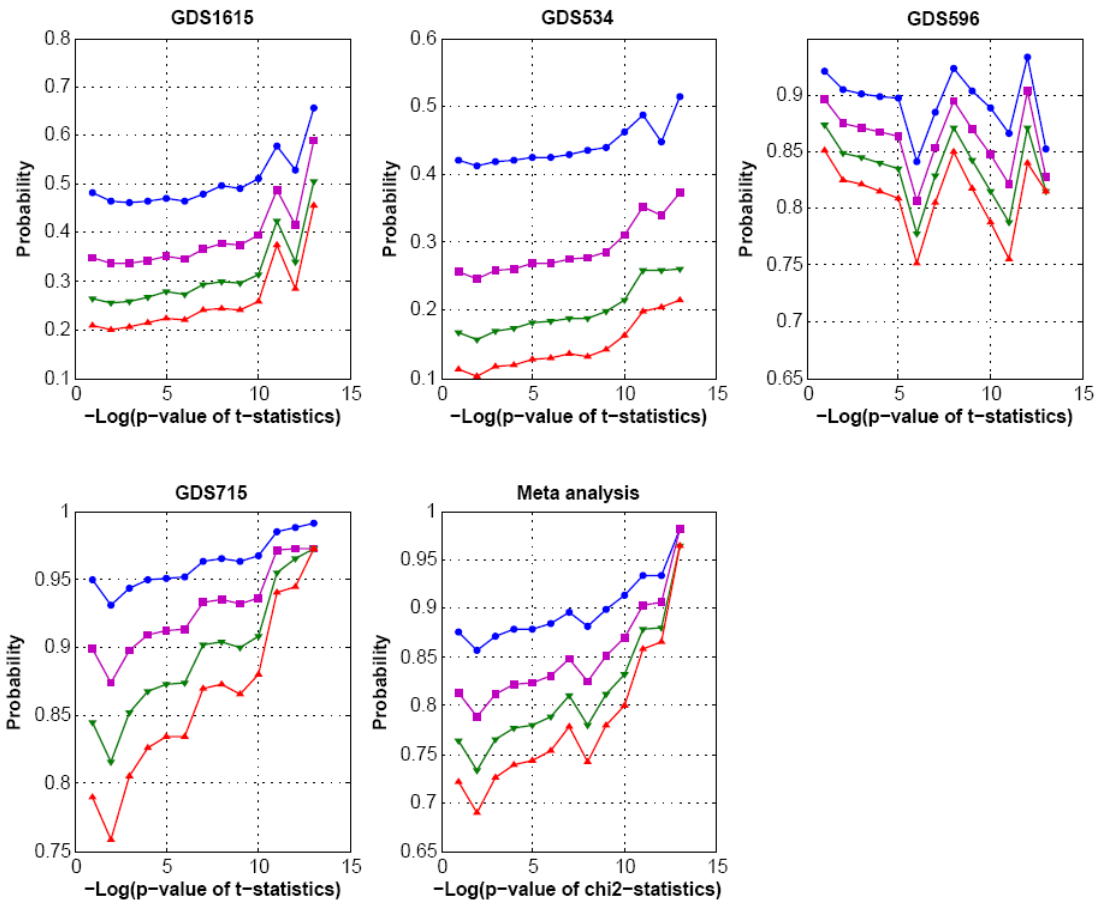

Figure 14: Conditional probability of functional similarity given an individual p-value (on log scale) for a single dataset or given the meta p-value for the 13 sets for human study.

In order to compare results from single dataset with that using meta-analysis, we randomly selected 4 single datasets, GDS596, GDS1615, GDS715 and GDS534, from the Table 3 to compare their individual prediction performance to that of combing all 13 datasets using meta-analysis. Note that GDS596 and GDS1615 have the largest number of arrays among the 13 datasets. Figure 14 shows conditional probability of functional similarity given an individual p-value (on log scale) for one of the 4 single datasets or given the meta p-value for the 13 datasets. The observation is consistent with that in the yeast data analysis, i.e., the curve for meta p-value is much smoother than the curve from

any single data, given that the sample size for generating meta-analysis curve is higher. However, this pattern is much strong in human data than yeast data. For example, the plot for GDS596 is highly fluctuated, which might indicate low information content for gene function prediction or high noise levels in the data set.

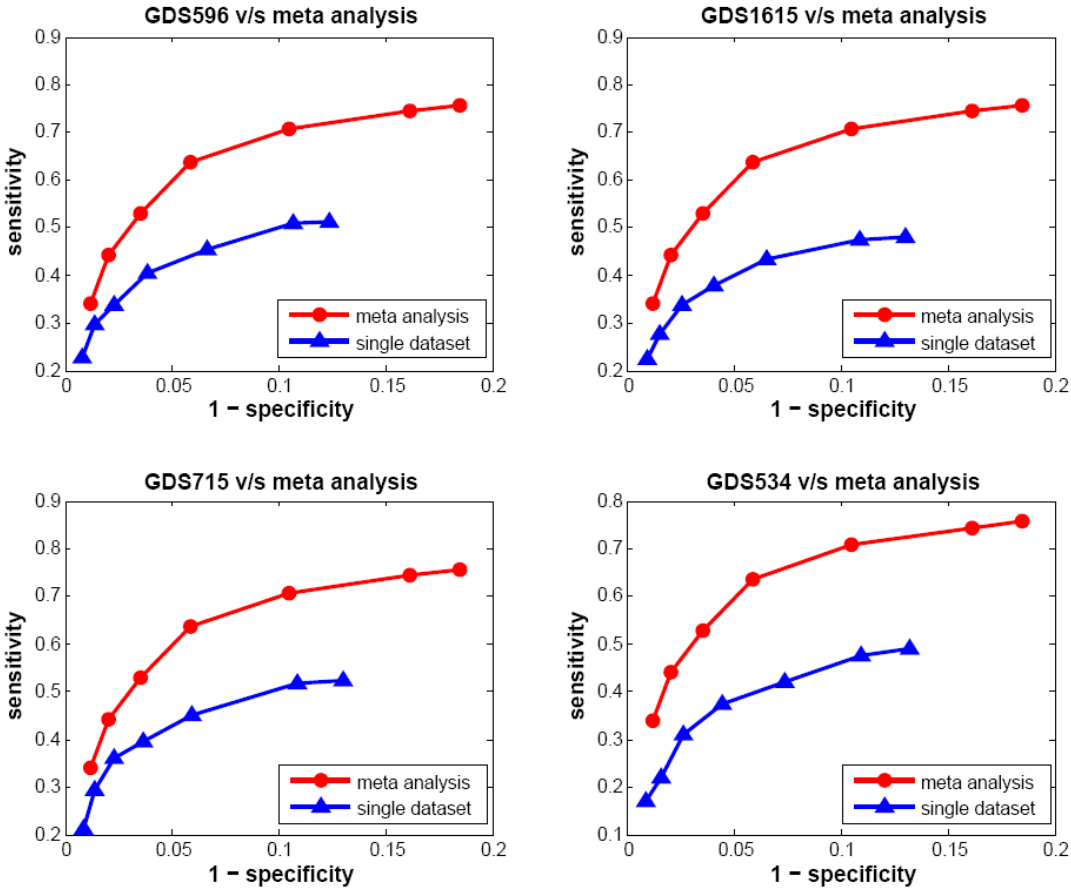

Figure 15: Prediction performance of single dataset (in blue) versus meta-analysis (in red) in human.

Figure 15 shows the sensitivity-specificity curves of function prediction applying either a single dataset or multiple datasets by using top 200 neighbours of each query gene for co-expression linkage graph in each case. The different point in sensitivity-specificity curve correspond to selecting top 50, 100, 200, 400, 800, 1600 and 3200 prediction in the form

of GO index for each query gene. As we expected, the meta-analysis performed significantly better than any single dataset analysis. We observed that the maximum sensitivity of function prediction using a single dataset is around 40% to 50%, whereas using meta-analysis goes to as high as ~80%. Both methods attain specificity ~70%. The prediction power (or sensitivity) never reaches to 1.0 in any of the cases. This is due to the fact that some functions for the query genes cannot be derived from their neighbours in the co-expression linkage graph.

3.5.3. Case study: Sin1 & PCBP2 interaction

When SIN1 (MAPKAP1) was used as the bait in a two-hybrid screen of a human bone marrow cDNA library, its most frequent partner was poly(rC) binding protein 2 (PCBP2/hnRNP-E2). PCBP2 associates with the N-terminal domain of SIN1 and the cytoplasmic domain of the IFN receptor IFNAR2. SIN1, but not PCBP2, also associates with the receptors that bind TNF. PCBP2 is known to bind to pyrimidinerich repeats within the 3' UTR of mRNAs and has been implicated in control of RNA stability and translation and selective capindependent transcription. RNAi silencing of either SIN1 or PCBP2 renders cells sensitive to basal and stress-induced apoptosis. Stress in the form of TNF and H_2O_2 treatments rapidly raises the cell content of SIN1 and PCBP2, an effect reversible by inhibiting MAPK14.

For this analysis, human microarray data from the NCBI Gene Expression Omnibus were analyzed to determine the datasets in which SIN1 and PCBP2 showed a significant (up or down) change in expression level. Then, the meta-analysis was performed on these

datasets to determine which genes were co-expressed with SIN1. The analysis created a statistical neighboring linkage network based on functional similarity score and its significance level. Close neighbors, i.e., genes that are co-expressed with SIN1 over time or in response to treatments, were assumed to have related functions of SIN1.

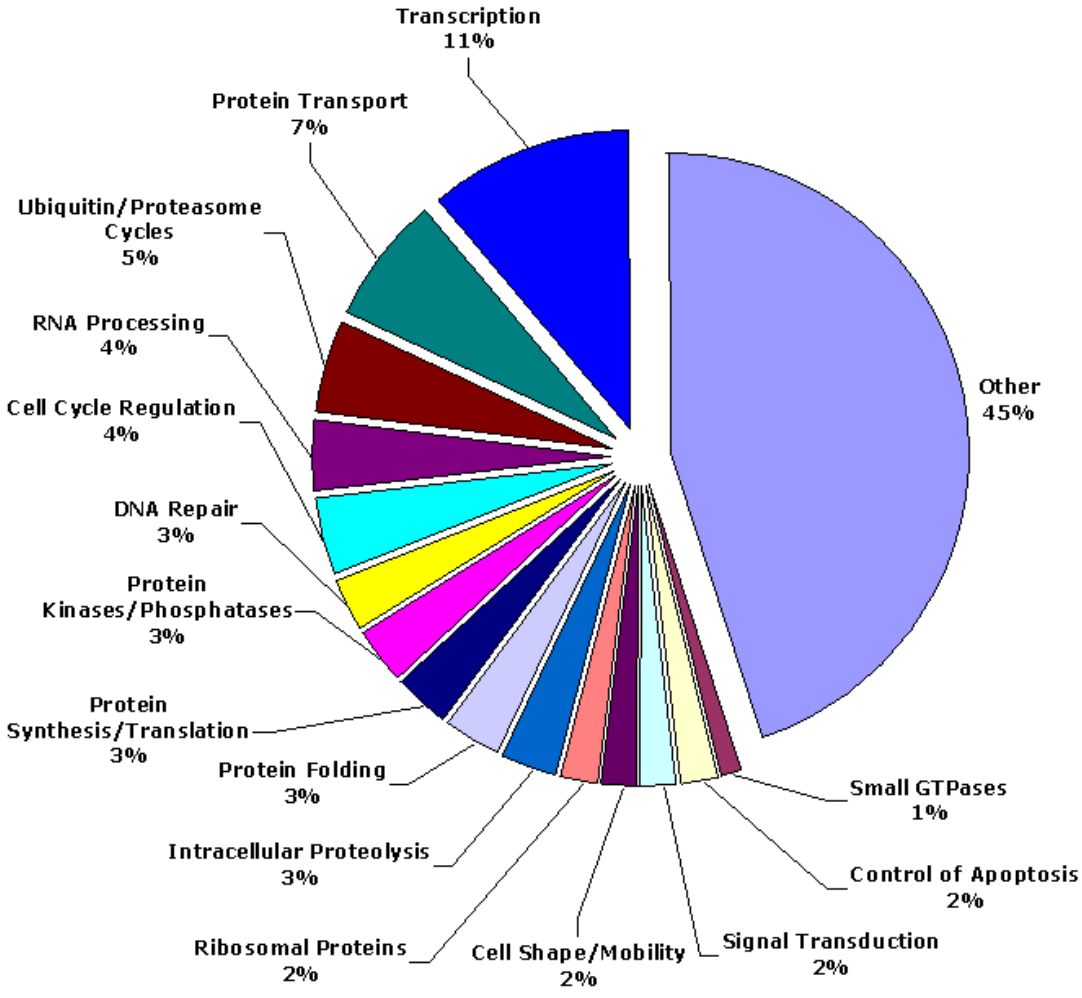

Figure 16: Classes of annotated genes that demonstrate expression similar to both SIN1 and PCBP2.

Here, the meta-analysis was confined to a single dataset microarray platform, GPL96, i.e., an Affymetrix Gene-Chip Human Genome U133 Array Set HG-U133A and used 13

curated microarray datasets, each of which had between 50 arrays and 154 arrays. The data were preprocessed and analyzed to provide two separate neighbor lists for SIN1 and PCBP2, respectively. The genes in common to each list with a significance level of $P < 0.01$ were then identified and ranked, based on associated confidence scores. The annotations of these identified genes were shown in Figure 16.

The meta-analysis of human microarray data supports the hypothesis that SIN1 plays a central, directive role in controlling apoptosis (Ghosh, et al., 2008). With few exceptions, genes and pathways regulated in concert with SIN1 are involved in reacting to various forms of stress. SIN1 appears to occupy an important node in a network of pathways that safeguard cells against environmental affronts and subsequently allow the cells either to die or to recover from damage. PCBP2, which is as vital as SIN1 in shielding against apoptosis, is also expressed co-ordinately with genes that encode large numbers of cell survival as well as cell death factors.

3.6. Discussion and Conclusions

In this research work, we developed various modules for pre-processing microarray data, identifying statistical neighboring, predicting gene function predictions and evaluating prediction performance using sensitivity-specificity curve. The strength of our function prediction model lies in its consistent performance across different organisms as shown in sensitivity-specificity curves. We have applied our method in various gene function predictions through collaborating with experimentalists. One example is the function predictions for stress-activated protein kinase interacting protein-1 (Sin1) and poly

(rC) binding protein-2 (PCBP2). Our predictions are partially verified experimentally and have been published elsewhere (Ghosh, et al., 2008).

The p-value for assessing the correlation between two gene expression profiles is calculated based on the distribution of t-statistics. In our paper, the sensitivity-specificity curve showed that the p-value based on one sided t-test works well for the data we tested. If t-distribution assumption is not correct, the p-value and the meta p-value can be obtained through random permutation in a way similar to what was done in (Rhodes, et al., 2002). The p-value approach only uses the binary information about the relationship of a gene pair, namely, it only tells whether a gene pair is significantly correlated or not and doesn't measure the strength of the relationship. Hence just combining the individual p-values will lose information in regard to the strength of the correlation. The meta correlation coefficient approach, which keeps the information about the strength of the relationship, does better than the binary method as verified by our results.

There are some limitations in the current studies. So far we have assumed different microarray datasets to be independent, which can be relaxed by defining some correlation terms and including them in meta-statistics. In function prediction algorithm, we could not afford jack-knife tests to rigorously separate the dataset used for calculating the Bayesian conditional probability from the test set for predicting gene function. We assume that the conditional probability estimation of functional similarity given the statistic value (individual p-value or meta p-value) does not change significantly even when some genes (up to ~10% in yeast and up to ~1% in human) are removed from the

training dataset. We tested a few cases and this assumption holds well (data not shown). Furthermore, our handling of the data was consistent between the single-dataset analysis and meta-analysis, and hence, our conclusion on better performance of meta-analysis does not depend on this assumption.

We believe that efficiency of this method can be improved by applying more sophisticated methods for microarray data pre-processing and normalization, better co-expressed pair identification, and integration with other types of data, such as protein interactions and phylogenetic profiles. As a next step, we are focusing on integrating microarray datasets from heterogeneous platforms and some more rigorous microarray data pre-processing techniques. For example, using a smaller subset of genes that are differentially expressed in a given microarray dataset may be more effective for gene function prediction than our current setting. We also plan to incorporate semantic similarity measurement technique and compare with our GO Index representation strategy.

4. Meta-analysis Application: Systems Biology

Transcription factors (TF) regulate downstream genes in response to environmental stresses in plants. Identification of TF target genes can provide insight on molecular mechanisms of stress response systems, which can lead to practical applications such as engineering crops that thrive in challenging environments. Despite various computational techniques that have been developed for identifying TF targets, it remains a challenge to make best use of available experimental data, especially from time-series transcriptome profiling data, to improve TF target identification.

In this study, we used a novel approach that combined kinetic modelling of gene expression with a statistical meta-analysis to predict targets of 757 TFs using expression data of 14,905 genes in Arabidopsis exposed to different durations and types of abiotic stresses. Using a kinetic model for the time delay between the expression of a TF gene and its potential targets, we shifted a TF's expression profile to make an interacting pair coherent. We found that partitioning the expression data by tissue and developmental stage improved correlation between TFs and their targets. We identified consensus pairs of correlated profiles between a TF and all other genes among partitioned datasets. We applied this approach to predict novel targets of known TFs. Some of these putative

targets were validated from the literature, for E2F's targets in particular, while others provide explicit genes as hypotheses for future studies.

Our method provides a general framework for TF target prediction with consideration of the time lag between initiation of a TF and activation of its targets. The framework helps make significant inferences by reducing the effects of independent noises in different experiments and by identifying recurring regulatory relationships under various biological conditions. Our TF target predictions may shed some light on common regulatory networks in abiotic stress responses.

4.1. Regulatory networks

Gene regulatory networks are critical for many organisms for organizing their genes to accomplish basic biological functions. Transcriptional regulation is an important step in controlling gene expression, and has been the target for many experimental explorations, such as transcript profiling by microarray analysis and quantitative PCR assays. Environmental stresses, such as drought, cold and chemicals, often have undesirable effects on plant developmental and physiological processes. Plants typically respond and adapt to these stresses through various transcriptional regulatory systems (Shinozaki, et al., 2003). Identification of these regulatory systems not only enhances our knowledge of biological processes in plants, but also helps a great deal in developing bio-engineered crops that can better sustain challenging environments (Kasuga, et al., 1999). It is traditionally a complex and time consuming exercise to elucidate gene regulatory relationship in plants, although several recent experimental approaches, such as cell-

based assays (Sheen, 2001), chemical inducible systems (Padidam, 2003; Padidam, et al., 2003) and chromatin immunoprecipitation (ChIP) assay (Wang, et al., 2002), have been developed that enable biologists to directly identify early targets of a small number of transcription factors. In parallel, advances in molecular biology and functional genomics, such as the completion of *Arabidopsis* and rice genome sequencing (Goff, et al., 2002; Lukowitz, et al., 2000) and high-throughput expression profiling experiments (Schena, et al., 1995), have generated large amount of data that makes it possible to develop computational approaches that predict regulatory relations.

4.2. Gene Regulation Model

It has been shown both in mRNA blotting and microarray experiments that activation of regulators under stress conditions usually occurs earlier than that of its targets (Haake, et al., 2002; Seki, et al., 2002). In a recent time series gene expression microarray experiment, Seki et al (Seki, et al., 2001; Seki, et al., 2002) determined the mRNA levels of a large number of genes in *Arabidopsis* under different stress conditions. The result suggested a clear time lag between the mRNA levels of a transcription factor, CBF and its known targets. This time lag complicated the identification of transcription factor-target relationship in time series transcription profiling data. In part, the time is used for an organism to translate the mRNAs of a transcription factor into proteins before the proteins can act on activating (or repressing) its targets. Due to the large amount of dynamic information on how the effect of upstream genes propagate in gene regulation networks, the time series transcription profiling data are extremely important resources

for detection of transcriptional factor-target relationship and learning the structures of the regulatory networks involved (Spellman, et al., 1998).

4.2.1. Kinetic Model for Time Lag in Regulation

A kinetic model is used to estimate the time delay in gene regulation. In eukaryotic cells, the effect of a TF is usually achieved in multiple steps, including transcription of the TF genes, transportation of the TF mRNAs out of the nucleus, translation of the transcripts, transportation of the TF proteins back to the nucleus, and binding of the TF proteins to the promoter regions of the target genes to achieve transcriptional regulation. Significant timing difference exists among changes in concentrations of the TF mRNA, the TF protein, and the mRNAs of its targets. A chemical kinetics model naturally fits this context by taking into account of the time lags among these events (Figure 17).

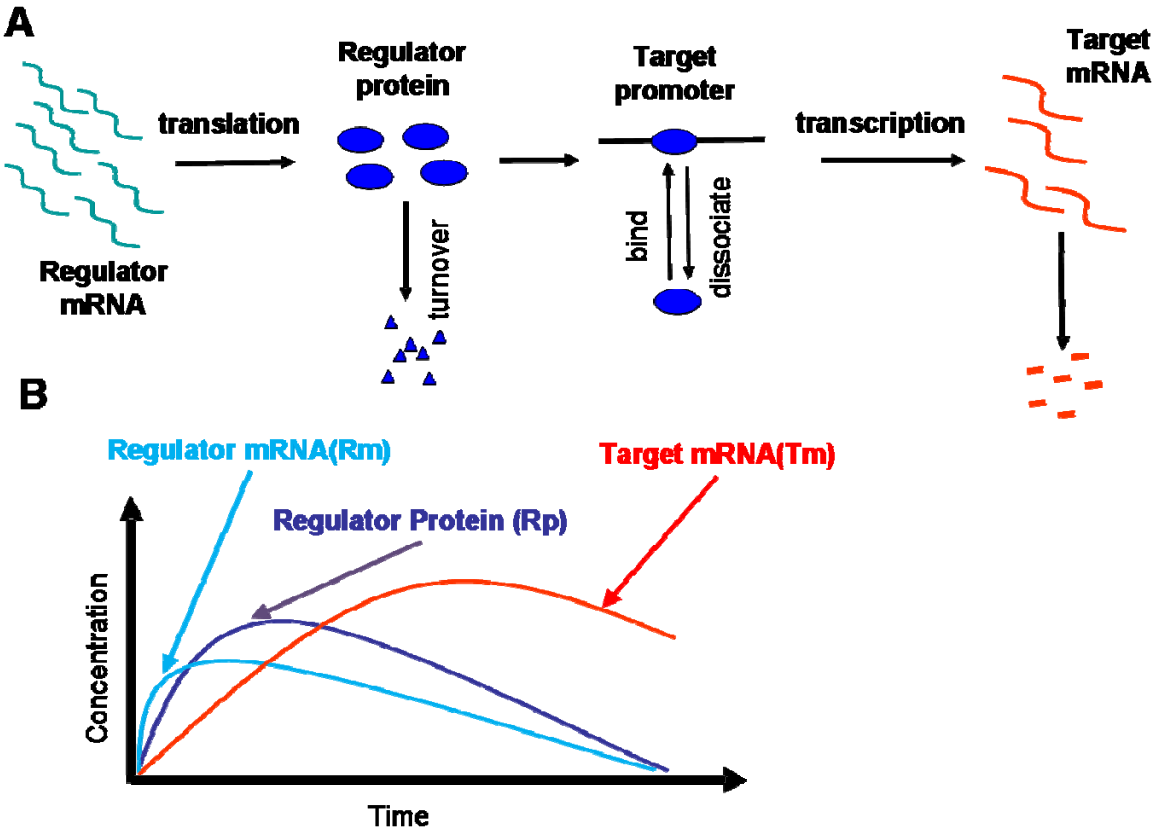

Figure 17:. Kinetic model for time lag in TF-target regulation.

Because the TF protein concentration is not measured directly in microarray experiments, it is treated as a hidden variable in our model to serve as the link between the measurable mRNA concentrations of a TF and its targets. More specifically, the TF protein concentration can be modeled by the following chemical kinetic equation without considering post-translational regulation:

$$\frac{dR_p}{dt} = K_{tran}R_m - K_pR_p \tag{4.1}$$

where R_p is the TF protein concentration; R_m is the TF mRNA concentration; K_{tran} is the apparent rate of mRNA translation, and K_p is the turnover rate of the TF protein.

Accordingly, the time course of the target mRNA concentration can be modeled with the following equation

$$\frac{dT_m}{dt} = B_t + f(R_p) - K_t T_m \tag{4.2}$$

where T_m is the concentration of the target mRNA; B_t is the basal transcription rate of the target gene; and K_t is the turnover rate of the target mRNA; $f(R_p)$ measures the regulated transcription rate, which is different for activators and repressors. For activators, it has the following Taylor first-order approximation when R_p is small (Chen, et al., 1999):

$$f(R_p) = f(R_p = 0) + \frac{d(f(R_p)}{dR_p}|_{Rp=0} R_p . \tag{4.3}$$

$f(R_p = 0)$ is equal to zero, assuming target gene transcription should not be activated when there is no TF protein. $\left.\frac{d(f(R_p)}{dR_p}\right|_{Rp=0}$ is the activation rate of TF protein on the target gene. If it is replaced by parameter K_{act} for simplicity, $f(R_p)$ takes the following form:

$$f(R_p) = K_{act} R_p . \tag{4.4}$$

The basal level target transcription rate should satisfy the following condition:

$$B_t + f(R_{pbasal}) - K_t T_{mbasal} = 0 \tag{4.5}$$

where R_{pbasal} and T_{mbasa} are the basal concentrations of the TF protein and target mRNA, respectively.

Usually, what is reported in transcription profiling experiment is not the absolute concentration of mRNA, but rather a fold change compared to basal transcription level of that gene. Thus, we define relative changes of R_m and T_m as $R_{m'}$ and $T_{m'}$

$$R_m' = R_m / R_{mbasal} - 1 \; ; \tag{4.6}$$

$$T_m' = T_m / T_{mbasal} - 1 \; . \tag{4.7}$$

Combining Equations (4.1), (4.2), (4.4), (4.5), (4.6) and (4.7), and considering the fact that $K_{tran} R_{mbasel} - K_p R_{pbaseal} = 0$ lead to the following second-order ordinary differential equation:

$$\frac{d^2(T_m')}{dt^2} + (K_t + K_p)\frac{d(T_m')}{dt} + K_t K_p T_m' = \gamma R_m' \tag{4.8}$$

where $\gamma = K_{act} K_{tran} R_{mbasal} / T_{mbasel}$.

Given all the model parameters, the relationship between the relative mRNA levels of TF and its target, R_m' and T_m', is defined by Equation (4.8). In other words, for the target gene of a TF, its relative mRNA level T_m' has to satisfy Equation (4.8), given the model parameters and the relative TF mRNA level R_m'. It is interesting to note that the TF protein concentration, a key variable in the original model equations, is not involved explicitly in the final equation relating to the relative mRNA levels of TF and target. To predict the target of a specific TF, we can solve Equation (4.8) to obtain the theoretical target behavior curve, and then find the genes with mRNA levels similar to the theoretical curve, which will be identified as the potential targets of that TF.

In the case of transcript expression profiling experiments under stress conditions, the initial conditions should be the following:

$$T_m'|_{t=0} = 0 \tag{4.9}$$

$$\frac{d(T_m')}{dt}|_{t=0} = 0 \tag{4.10}$$

Because the target gene mRNA and the TF protein should be at their basal levels at the onset of stress condition (*t=0*), it is apparent from Equations (4.2) and (4.5) that initial condition (4.10) should be satisfied.

To approximate R_m, a stepwise linear model can be fit as follows:

$$R_m'_i(t) = \alpha_i + \beta_i t \quad t_i \le t \le t_{i+1} \quad i = 0, \cdots, n-1 \tag{4.11}$$

where t_i is the i^{th} time point; and α_i and β_i are the parameters of stepwise linear function in each time interval, which are determined by the measured TF mRNA levels at the two adjacent time points. Equation (4.8) has analytic solution:

$$Tm_i(t)' = A_i e^{-K_t t} + B_i e^{-K_p t} + C_i + D_i t \quad t_i \le t \le t_{i+1}, \quad i = 0, \ldots, n-1 \tag{4.12}$$

where $D_i = \beta_i \gamma / K_p K_t$ *and* $C_i = \left[\alpha_i \gamma - (K_p + K_t) D_i\right] / K_p K_t$.

The contiguous restrictions on T_m' are stated in the following equations:

$$Tm_i'(t) = Tm_{i+1}'(t), \qquad when \;\; t = t_i \;\; i = 1, \ldots, n-1. \tag{4.13}$$

$$\frac{d(Tm_i'(t))}{dt} = \frac{d(Tm_{i+1}'(t))}{dt}, \qquad when \;\; t = t_i \;\; i = 1, \ldots, n-1. \tag{4.14}$$

After substituting Equation (4.12) into Equations (4.9), (4.10), (4.13) and (4.14), A_i and B_i can be obtained by solving sets of linear algebra equations, and are functions of α_i, β_i, γ, K_t and K_p.

4.2.2. Learning Model Parameters

For each TF-target pair, there are three parameters involved in Equation (4.8), the target mRNA turnover rate K_t, the active TF turnover rate K_p, and γ, which is equal to $K_{act}K_{tran}R_{mbasal}/T_{mbasel}$. K_{act} represents the strength of TF protein effect on the target gene and K_{tran} is the translation rate of TF mRNA. They lump together with the ratio of basal mRNA concentrations of TF and target to form parameter γ, which determines the magnitude of the relative target mRNA level but not its shape. The parameters K_t and K_p determine the shape of the relative target mRNA level, such as how fast the target gene responds to the TF.

For gene expression experiments under stress conditions in plants, the kinetics model can be trained with known TF-target pair reported in the literature (e.g., CBF and RD17 in Arabidopsis under cold stress) with a non-linear regression model. When the normalized expression profile of a target gene with its maximal response is considered, there is no need to keep γ as a free model parameter ($\gamma_1 = n\gamma_2$ leads to $T_{m1}' = nT_{m2}'$ when other parameters are kept the same in Equations (4.8), (4.9) and (4.10)). Therefore, only two parameters K_t and K_p are estimated from the non-linear regression model, and are used to predict other TFs and their targets in plant stress response. The theoretical target

mRNA expression profiles are calculated for all the genes annotated as TFs and are substituted in place of TF's profile during further computation. These theoretical target profiles of any TF are independent of actual targets of that TF as it is solely calculated based on kinetic model. According to the model, the theoretical target profile of any TF should match with profile of its actual targets. With this assumption, we can use correlation technique to find similarity between co-expression between these theoretical profile and rest of genes to find potential targets.

4.3. Implementation

We used a kinetic model combined with statistical meta-analysis to identify TF targets and reconstructed an Arabidopsis global regulatory network using large-scale expression profiles of 14,905 genes. We then evaluated our strategy by comparative and functional analysis of predicted E2F target genes and by comparing our method with other existing methods. Finally, we analyzed the reconstructed network to infer some novel features from the network.

- **Data Preparation**

We used publically available microarray data of *A. thaliana* from NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/) and TAIR (http://www.arabidopsis.org/). We preprocessed the data by removing genes with missing expression measurement across all the tissues and by averaging the replicated expression data. Consequently, we applied our method on 497 arrays in total measuring whole-genome response of Arabidopsis exposed to different durations and types of abiotic stresses. Some 14,905 genes from Arabidopsis

genome including 757 TFs were chosen for the analysis as each of these genes was differentially expressed in at least one of the stress conditions. The datasets consist of 27 different microarray experiments, out of which 10 experiments are time series.

- **Co-expression Statistics**

We used a statistical meta-analysis technique (Srivastava, et al., 2009) to identify highly correlated expression profiles from multiple microarray datasets. Using this technique, we evaluated the statistical significance (right-tailed p-value) of a Pearson correlation coefficient *r* for two expression profiles in a single dataset based on the standard t-statistics:

$$p-value = P(T > \hat{t}), \text{ where } \hat{t} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{4.15}$$

where *T* is a *t*–random variable with *n-2* degree of freedom and *n* is the number of conditions of the gene expression profiles. Since we assume that the datasets are obtained independently, we apply the inverse chi-square method and obtain the meta chi-square statistics:

$$\hat{\chi}^2 = [-2\log(P_1) - 2\log(P_2) - ... - 2\log(P_n)] \tag{4.16}$$

where P_i is the p-value obtained from the i^{th} data set for a given gene pair defined in Equation (4.15). When there is no linear correlation between a gene pair in any of the multiple datasets, the above chi-square statistics $\hat{\chi}^2$ follows a central chi-square distribution with degrees of freedom *2n* and hence the p-value for meta-analysis can be obtained by

$$meta\ \ p-value = P(\chi^2_{2n} > \hat{\chi}^2) \tag{4.17}$$

where χ^2_{2n} is a chi-square random variable with *2n* degrees of freedom. We calculate significance level of the gene pair from multiple datasets. The significance level of gene pair represents the count of datasets in which that gene pair has significant correlation (p-value < 0.01) based on Equation (4.15). We used meta p-value statistics (Equation 4.17) combined with significance level to rank potential targets for a TF (Srivastava, et al., 2009).

- **Regulatory Network Reconstruction**

The meta p-value combined with significance level and the Pearson correlation coefficient were used as co-expression statistics for finding putative targets for a TF. For a single dataset (without partitioning of microarray data), we ranked all the potential targets of a TF based on Pearson correlation coefficient and select targets such that TF-target correlation > 0.75 (medium size network) or 0.70 (large size network). For multiple datasets, we ranked all TF-target pairs based on the number of individual p-values that are smaller than 0.01 across multiple datasets; for pairs that have the same number of significant p-values, they were ranked by the corresponding meta chi-square statistics defined in Equation (4.16). Here we used meta chi-square instead of meta p-value since the meta p-value for many gene pairs are very close to zero and hard to distinguish computationally; both meta chi-square and meta p-value should result in the same order when the degrees of freedom for each gene pair is same. In the end, a fixed number of TF-target pairs were selected based on ranking.

In case of meta-analysis, number of target genes for a TF was determined in two methods, i.e., (1) selecting fixed number of targets from top (50 or 75) or (2) choosing targets form top-ranked genes that shows significance correlation as TF-target pair in at least certain number of microarray datasets used for meta-analysis. For example, we used significance cutoff 9 (out of 9 datasets) for small network and cutoff 8 (out of 9) for medium network and cutoff 7 (out of 9) for large network. The second method worked better in general.

4.4. Evaluation and Comparative Analysis

In order to conduct meta-analysis, we partitioned the datasets based on different attributes including tissue, experiment type and developmental stage. The tissue-specific partition of the microarray datasets produced totally 8 tissue types that have sample size of at least 9. We combined the rest of the samples into one group as combined tissues as shown in Table 5.

We defined the significance level of TF-target pair as number of tissues in which the TF-target pair is significantly co-expressed (p-value < 0.01) after time lag corrections using the kinetic model. We built three networks of ~2K, ~12K and ~59K edges, which correspond to significance levels of more than 9, 8 and 7 respectively. For further analysis, we used the network of ~12K edges to balance the size of network and tolerance of experimental errors in each tissue. This network consists of 12,300 regulatory interactions amongst 4,968 genes, in which 757 genes act as TFs as shown in figure 18. In this figure, blue (larger) nodes correspond to TFs and red (smaller) nodes correspond

to genes that are regulated by TFs. All the edges are marked by green. It is interesting to note that the distribution of the network is highly uneven. In some cases (e.g., lower right), a handful of TFs regulate many putative targets, while in other cases (e.g., left edge) many TFs form clusters among themselves.

	Tissue Group	Number of Samples	Number of Experiments
1	Seedling	180	9
2	Root	95	14
3	Shoot	68	10
4	Leaf	45	5
5	Flower	33	5
6	Seed	11	3
7	Shoot-apex	10	1
8	Protoplast	9	1
9	Combined rest	46	5

Table 5: List of all tissue groups used for meta-analysis.

- **Network Evaluation and Comparative Analysis**

In order to compare performance of meta-analysis with other methods for identifying TF targets, we prepared a benchmark dataset of TF-target pairs in Arabidopsis, which were obtained from the AGRIS database and AtRegNet (Palaniswamy, et al., 2006). The benchmark set has 348 pairs in total. Some of the well-known methods to identify TF

target and build regulatory network, including causal regression method, standard Pearson correlation method, and Graphical Gaussian model were used for comparative analysis.

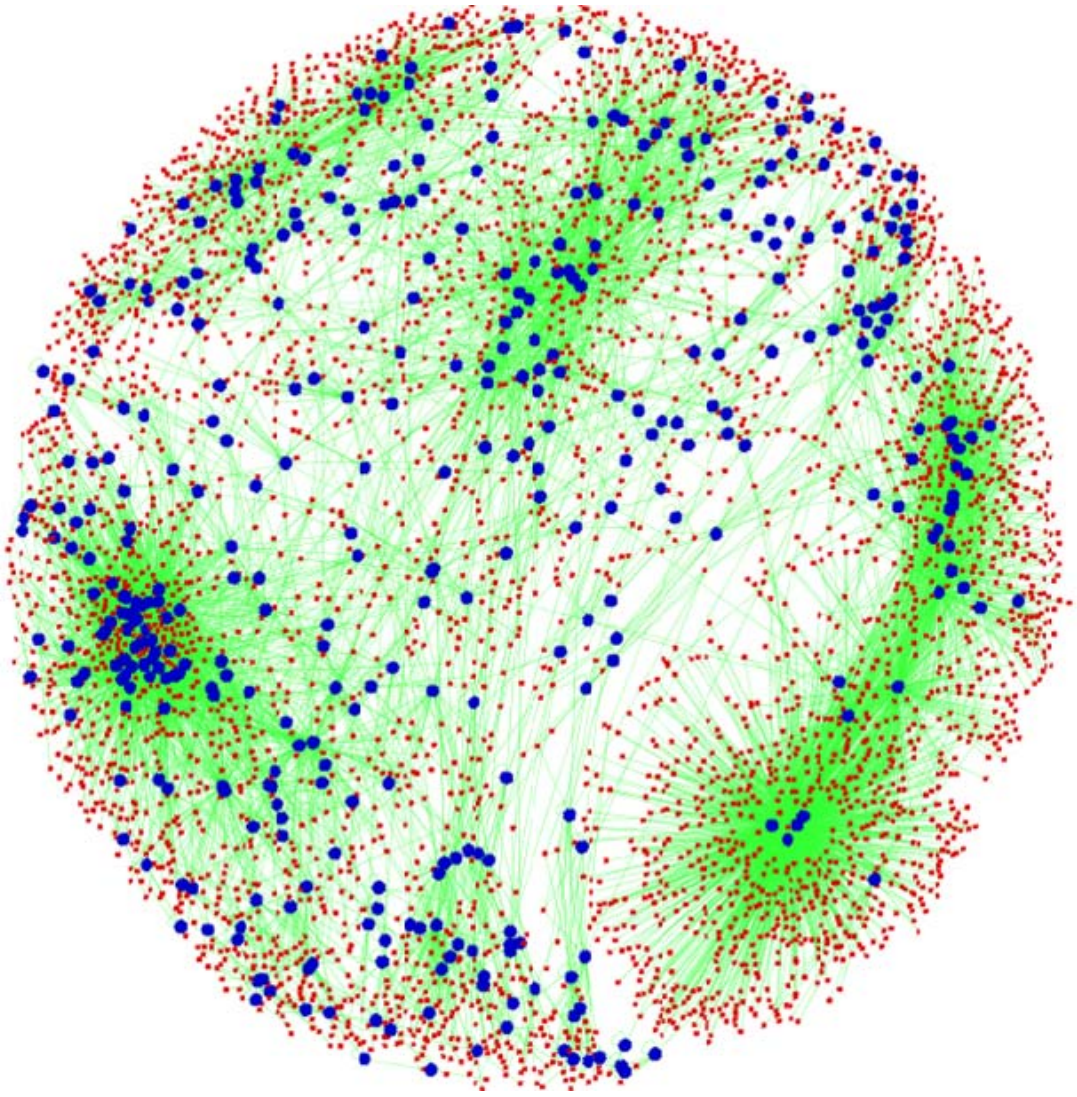

Figure 18: Global regulatory network with 4968 nodes (genes) and 12,300 edges for Arabidopsis.

In order to make direct comparison of various methods, we used the exactly same microarray datasets as input to these methods and also exactly the same benchmark.

While using Pearson correlation method, Graphical Gaussian model and regression method, we did not partition the data rather we followed the procedure as previously done in the literature. In case of microarray data partition and meta-analysis, we used three different ways to partition the microarray data, i.e., tissue based partition, experiment type based partition, and developmental stage based partition. For each type of partition, we identified genome-wide targets for the given set of TFs. While using other methods (Pearson correlation coefficient, causal regression and graphical Gaussian model), we input the microarray data as a single large dataset without partition and identified targets for the same list of TFs. Using these predicted TF-target pairs from each of the methods, we reconstructed two networks of different sizes that is, less than 40,000 edges and less than 70,000 edges. All the same category networks from different methods were then checked against the standard set to count the number of confirmed edges in these networks as shown in Table 6.

	Applied Method	**Network Size**	**Confirmed Edges**	**Ratio**
	Pearson Correlation (Cutoff=0.70)	35,253	25	7.09e-4
	Causal Linear Regression Model	30,000	5	1.66e-4
Graphical Gaussian Model	GeneNet: Static method	30,000	9	3.00e-4
	GeneNet: Dynamic Method	30,000	9	3.00e-4
Meta-analysis	Tissue-wide partition	12,300	35	28.5e-4

(Microarray data partition)	Experiment-wide partition	37,850	14	3.96e-4
	Development based partition	37,850	18	4.75e-4

(A)

	Applied Method	Network Size	Confirmed Edges	Ratio
	Pearson Correlation (Cutoff=0.70)	71,417	36	5.04e-4
	Causal Linear Regression Model	59,557	16	2.68e-4
Graphical Gaussian Model	GeneNet: Static method	68,624	10	1.46e-4
	GeneNet: Dynamic Method	68,658	10	1.45e-4
Meta-analysis (Microarray data partition)	Tissue-wide partition	59,676	57	9.55e-4
	Experiment-wide partition	56,775	18	3.17e-4
	Development based partition	57,339	22	3.84e-4

(B)

Table 6: Comparative analysis of Arabidopsis networks of ~40K (A) and ~70K (B) sizes.

The results show that our method with partitioning microarray data into tissue-specific datasets and then performing tissue-wide meta-analysis contains the most confirmed edges. Particularly in Table 6A, the network obtained using tissue-wide meta-analysis is 1/3 in size compared to other networks in the same category, but with more confirmed edges than any other network. The comparison clearly demonstrates that tissue-wide

partition performs much better than experiment-wide or development-based partition. It also shows that the tissue-wide meta-analysis could greatly improve network constructions over other methods. Interestingly, a simple method using Pearson correlation cutoff of 0.70, although not as good as meta-analysis, outperformed sophisticated methods of causal linear regression model and graphical Gaussian model. This may be because microarray data are often noisy and sophisticated methods could amplify noises to give incorrect predictions in gene regulatory relationships.

4.5. Case study of E2F transcription factor

In order to assess our TF target prediction with known regulatory mechanisms from the literature, we investigated Arabidopsis E2F family transcription factor "At2g36010", which represents a group of proteins that play a crucial role in the control of cell cycle progression and regulate expression of genes required for the G1/S transition. These include enzymes involved in nucleotide synthesis and DNA replication proteins (Bracken, et al., 2004; Ramirez-Parra, et al., 2003; Vandepoele, et al., 2005).

4.5.1. E2F-target identification and evaluation

Though it is clear that E2F is highly critical and conserved amongst high eukaryotes, only a few genes induced by E2F are experimentally verified in plants. Vandepoele et al. (Vandepoele, et al., 2005) combined microarray and promoter motif analyses to identify E2F-targets in plants. To do this, promoter regions of genes that were induced at the transcriptional level in Arabidopsis seedlings were searched for the presence of E2F-

binding sites. In another study, Ramirez-Parra et al. (Ramirez-Parra, et al., 2003) identified potential E2F-responsive genes by a genome-wide search of chromosomal sites containing E2F-binding sites. Using meta-analysis of tissue-specific microarray data, we identified 178 putative E2F-target genes (see Supplemental Dataset 1). Some of these were also predicted by either Vandepoele et al. (Vandepoele, et al., 2005) or Ramirez-Parra et al. (Ramirez-Parra, et al., 2003) as shown in Table 2. As the two other studies used different analytical approach to identify targets, these overlapping genes have more confidence to be true E2F target genes. In table 7 below following abbreviation is used to reference previous studies in the literature.

Ref [1]: (Ramirez-Parra, et al., 2003)

Ref [2]: (Vandepoele, et al., 2005)

	Locus ID	Symbol	Annotation	Ref [1]	Ref [2]
1	At1g08130	ATLIG1	DNA recombination / DNA repair / DNA replication	-	√
2	At1g07370	PCNA1	Regulation of DNA replication and cell cycle	-	√
3	At1g67630	POLA2	DNA synthesis and replication	√	√
4	At2g07690	-	DNA synthesis and replication	√	-
5	At5g66750	CHR1	Transcriptional control/chromatin modification	√	-
6	At1g78650	POLD3	DNA or RNA metabolism/ transferase activity	√	√
7	At4g14700	ORC1A	Cell cycle, Replication control, DNA synthesis	√	-
8	At1g09450	-	N-terminal protein myristoylation/ protein amino acid phosphorylation	√	-

9	At2g40550	ETG1	DNA replication	√	√
10	At1g67320	-	DNA replication, synthesis of RNA primer	-	√
11	At1g44900	-	DNA synthesis and replication, cell cycle control	√	√
12	At1g69770	CMT3	Chromatin silencing / DNA methylation	-	√
13	At2g21790	RNR1	DNA synthesis and replication	√	√
14	At2g16440	-	DNA replication initiation	-	√
15	At5g38110	ASF1B	Transcriptional control	√	√
16	At5g52950	ATIM	Putative protein	√	-
17	At5g18620	CHR17	Transcriptional control, chromatin modification	√	√
18	At5g52910	ATIM	Regulation of circadian rhythm	√	√
19	At2g24490	RPA2	Replication protein A-like	-	√
20	At2g29570	PCNA2	Error-prone postreplication DNA repair / replication	-	√
21	At2g31270	CDT1A	Chloroplast organization / DNA replication	-	√
22	At3g02820	-	Response to DNA damage stimulus / cell cycle	-	√
23	At3g18630	-	DNA repair	-	√
24	At3g25100	CDC45	Cell division control protein	-	√
25	At5g49010	SLD5	DNA replication initiation / GINS complex	-	√
26	At5g49160	MET1	DNA or RNA metabolism / other cellular processes	-	√
27	At5g62410	SMC2	Cell organization / DNA or RNA metabolism	-	√
28	At5g63960	-	DNA or RNA metabolism / nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-	√
29	At5g67100	ICU2	Negative regulation of flower development / leaf morphogenesis	-	√

30	At1g35530	-	helicase activity/ hydrolase activity / DNA binding	-	√
31	At3g02920	-	nucleic acid binding	-	√
32	At3g27640	-	nucleotide binding	-	√
33	At5g06590	-	Unknown	-	√
34	At5g63920	-	DNA metabolic process / DNA unwinding during replication	-	√

Table 7: Predicted E2F-targets evaluation from ~12K-size network with other studies.

4.5.2. GO-enrichment analysis

We also conducted functional enrichment analysis for the 178 E2F-target genes identified using meta-analysis. We used AmiGO's Term Enrichment tool, which is based on GO-TermFinder (Boyle, et al., 2004). We used all the annotated genes in TAIR (Reiser and Rhee, 2005) as the background set. We selected enriched gene groups with a p-value cutoff of 0.01 and the minimum number of gene products of 2. Our result (Table 3) supports that the E2F pathway is critical in regulating proteins, which are involved in cell cycle regulation, DNA replication, and chromatin dynamics. In addition, we identified other novel genes, which are involved in DNA methylation on cytosine, DNA repair, ribosome biogenesis, etc.

GO Term	Description	P-value	Number of Genes
GO:0006260	DNA replication	4.53E-29	23
GO:0006259	DNA metabolic process	1.97E-26	29

GO:0006261	DNA-dependent DNA replication	1.40E-13	12
GO:0006270	DNA replication initiation	6.44E-11	7
GO:0034645	Cellular macromolecule biosynthetic process	2.81E-10	47
GO:0009059	Macromolecule biosynthetic process	3.61E-10	47
GO:0034961	Cellular biopolymer biosynthetic process	7.72E-10	46
GO:0043284	Biopolymer biosynthetic process	9.65E-10	46
GO:0044260	Cellular macromolecule metabolic process	2.13E-09	60
GO:0043170	Macromolecule metabolic process	2.33E-09	61
GO:0034960	Cellular biopolymer metabolic process	3.61E-09	59
GO:0043283	Biopolymer metabolic process	4.82E-09	59
GO:0044249	Cellular biosynthetic process	8.07E-08	50
GO:0044238	Primary metabolic process	1.91E-07	65
GO:0009058	Biosynthetic process	4.48E-07	50
GO:0006139	Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6.32E-07	35
GO:0007049	Cell cycle	2.29E-06	12
GO:0044237	Cellular metabolic process	2.88E-06	65
GO:0009987	Cellular process	3.99E-06	77
GO:0008152	Metabolic process	1.64E-05	67
GO:0051052	Regulation of DNA metabolic process	2.18E-04	5
GO:0032776	DNA methylation on cytosine	1.04E-03	3
GO:0006412	Translation	1.80E-03	21
GO:0022402	Cell cycle process	1.98E-03	7

GO:0006281	DNA repair	4.07E-03	7
GO:0034984	Cellular response to DNA damage stimulus	4.29E-03	7
GO:0044267	Cellular protein metabolic process	4.83E-03	31
GO:0019538	Protein metabolic process	5.15E-03	31
GO:0042254	Ribosome biogenesis	5.18E-03	8
GO:0006974	Response to DNA damage stimulus	5.81E-03	7
GO:0022613	Ribonucleoprotein complex biogenesis	5.83E-03	8
GO:0044085	Cellular component biogenesis	8.30E-03	11

Table 8: GO term enrichment analysis of 178 predicted E2F-target genes.

4.6. Network feature analysis

- **Network Feature Analysis**

Using Cytoscape (Shannon, et al., 2003), we identified a few major hubs (nodes with many connections) from the medium sized network (~12K) using tissue-wide meta-analysis. In particular, we found regions of significant local density using the MCODE plugin (Bader and Hogue, 2003) from Cytoscape. Figure 19 shows an example of a major hub cluster, which represents 12 TFs including SCL13, ZAT6, AtERF-1 and anac062 each targeting many genes as found in Table 9 from further analysis. This sub-network contains 35 nodes and 362 edges.

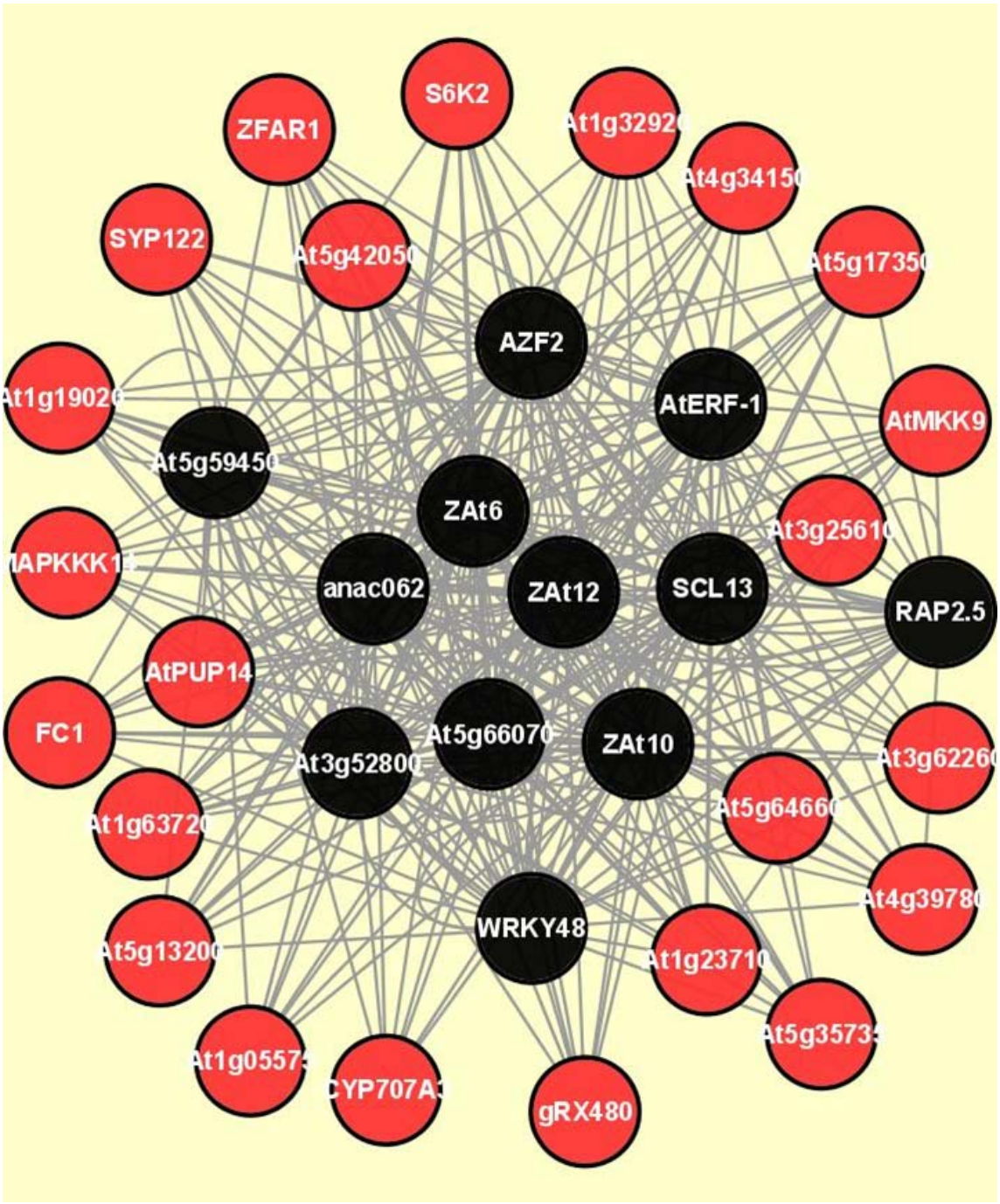

Figure 19: A cluster identified using MCODE. Black node is TF and red node is target gene.

Beside network feature analysis using Cytoscape, we analyzed TFs that target significantly more genes than other TFs across different tissues as shown in Table 19. Not surprisingly, most of these TFs are annotated with response to different stimulus in Arabidopsis, such as response to chitin and external stress, given that the microarray data we used were measured in response of Arabidopsis exposed to different abiotic stresses. Some of well-connected TFs are also present in hubs as recognized by the MCODE plugin from Cytoscape and are known to work together for gene regulation. For example, Zat6, Zat10, and Zat12 in the hub of Figure 19 are activated together in cold and osmotic stresses (Mittler, et al., 2006). WRKY33 and WRKY40 in Table 9 both function as activators of jasmonic acid-dependent defence pathways and repressors of salicylic acid signalling (Xu, et al., 2006).

4.7. Discussion and Conclusion

In this paper, we proposed a meta-analysis method for identifying TF targets. The novelty of the proposed method lies in combining two models that is (1) adjusting time lag between a TF and its target and (2) finding consensus regulatory interactions from different experimental sources/conditions including tissue types, developmental stages and experimental settings. Our study shows that tissue-wide partition performs much better than experiment-wide or development based partition for predicting TF targets. The method successfully identified more known TF-target pairs in Arabidopsis than other methods.

	Locus ID	Symbol	Annotation	Target
1	AT2G38470	WRKY33	Response to drought, heat, chitin, osmotic stress, salt, cold etc., defense response to fungus, bacterium	216
2	AT1G80840	WRKY40	Response to wounding, salicylic acid, chitin, defense response to bacterium, fungus etc	102
3	AT3G49530	anac062	Response to chitin	130
4	AT3G57150	NAP57	Pseudouridine synthesis	322
5	AT4G37490	CYCB1	Response to gamma radiation, regulation of cell growth	168
6	AT3G22780	TSO1	Regulation of meristem organization	134
7	AT4G17500	AtERF-1	Response to chitin, regulation of transcription, DNA-dependent	120
8	AT4G30930	NFD1	Embryo sac & pollen development, karyogamy, double fertilization forming a zygote and endosperm	518
9	AT5G59820	RHL41	Response to chitin, heat, UV-B, wounding, oxidative stress, cold, photosynthesis, hyperosmatic salinity response	122
10	AT4G17230	SCL13	Response to chitin	121
11	AT5G04340	ZAT6	Nucleic acid & zinc ion binding, transcription factor activity	139
12	AT1G27730	STZ	Response to abscisic acid, drought, light, cold, chitin, salt etc	128

Table 9: Global regulators from ~12K network having most target genes.

There are some limitations of this study. Like other approaches, our method will have both false positives and false negatives. It may not be able to distinguish TF targets from other co-expressed non-target genes, although meta-analysis across multiple tissues reduces such a possibility. From the meta-analysis point of view, tissue-wide meta-

analysis does not consider specific regulatory relations in particular tissue types. In plant some regulations are specific to different tissue types or developmental stages. Since such relations do not exhibit significant correlation across different microarray data, meta-analysis does not consider them. Nevertheless, meta-analysis is more robust to find correlations that are consistent across different tissues. Typically, global regulations are those that are fundamental for the existence of the tissues in general. In the context of our study, we only applied gene expression data of Arabidopsis exposed to different abiotic stresses.

It is known that there are common regulatory mechanisms for abiotic stresses. For example, certain heat-shock proteins are commonly elicited in response to various stress conditions in multiple plants (Sørensen, et al., 2003). Conserved regulatory mechanisms among responses to drought, salinity, and extreme temperature in Arabidopsis were identified, such as the DREB transcription factors (Mauch-Mani and Mauch, 2005). Characterizing common gene expression patterns under various abiotic stress conditions in plants can help elucidate these conserved regulatory mechanisms (Swindell, 2006). Hence, the meta-analysis that we provided on gene expression data under different abiotic stress treatments may shed some light on common regulatory networks in abiotic stress responses. In our future studies, we will explore more into meta-analysis of microarray data by applying different statistics like calculating meta correlation instead of chi-square statistics. Another dimension of improvement is to include inferences from other types of data such as promoter motif analysis.

4.8. Future work

We plan to extend our research work of applying meta analysis on microarray data for biological inferences to metabolomic network. According to Wikipedia, a metabolic network is the complete set of metabolic and physical processes that determine the physiological and biochemical properties of a cell. As such, these networks comprise the chemical reactions of metabolism as well as the regulatory interactions that guide these reactions. In this application, we will use meta analysis approach for studying metabolic responses to stress conditions in plants.

As shown in the example below, a metabolic profile looks similar to the gene expression profile. This is one of the main motivations behind application of meta analysis into construction metabolic network. Metabolomic network is becoming increasingly important these days to understand the biological processes and various phenotypes. For example many researchers are trying to screen the metabolic networks in Arabidopsis, which can help to understand the network behind various phenotypes and can help developing new methods for increase yield or improved resistant to abiotic stresses like drought etc. Below is an example of metabolic network and associated profile of metabolites.

5. High-throughput Oligo Design

5.1. Introduction

Various genome-scale sequencing project have generated vast amounts of sequence data. High-throughput data analysis and its study are one of the primary focuses for molecular biologist. Microarray is one of the most common tools for studying gene expressions on a large scale (DeRisi, et al., 1997; Duggan, et al., 1999). In cDNA microarray, typically each spot on the array contains sequence segment of a specific gene, which is amplified by PCR. The segment is expected to be gene specific to avoid cross-hybridization among genes sharing significant sequence identity. In another case, researchers may simply want to amplify gene-specific segments for a selected group of genes using reverse-transcriptase (RT)-PCR. In both cases, the problem can be formulated to choose a gene-specific segment for a gene in a genome and then design PCR primers according to some specifications. Such an objective is often achieved manually, e.g., using Primer3 (Rozen and Skaletsky, 2000) for primer design for a given sequence. Primer3 designs many possible primer pairs for a given sequence, but it does not guarantee their uniqueness in the whole genome. Therefore, a user has to manually run BLAST (Altschul, et al., 1997) for each PCR product against the genome to search to avoid cross-hybridization. Such manual approach cannot be applied to a large scale. PRIMEGENS (Srivastava, et al., 2008; Srivastava and Xu, 2007; Xu, et al., 2002) does not only fulfill this task but also automate the primer generation on the large scale. Furthermore, PRIMEGENS has a rigorous formulation, which has a much better chance to find gene-specific segment than a manual process.

5.2. PRIMEGENSv2 Framework

This chapter introduces the software package PRIMEGENS for designing gene specific probes and associated PCR primers on a large scale. Such design is especially useful for constructing cDNA or oligo microarray to minimize cross-hybridization. PRIMEGENS can also be used for designing primers to amplify a segment of a unique target gene using reverse-transcriptase (RT)-PCR. The input to PRIMEGENS is a set of sequences, whose primers need to be designed, and a sequence pool containing all the genes in a genome. It provides options to choose various parameters. PRIMEGENS uses a systematic algorithm for designing gene-specific probes and its primer pair. For a given sequence, PRIMEGENS first searches for the longest gene-specific fragment and then designs best PCR product for this fragment. The 2.0 version of PRIMEGENS provides a graphical user interface (GUI) with additional features. The software is freely available for any users and can be downloaded from http://digbio.missouri.edu.

Figure 20 gives some general idea about PRIMEGENS. The essence of PRIMEGENS is based on searching the sequence-specific fragment for any particular sequence. PRIMEGENS implements this task by finding the fragment of a given DNA sequence, which does not have high-sequence similarity with any other sequence in the given sequence pool (whole genome in general). If the given sequence is unique, then the whole sequence is considered as the sequence-specific fragment.

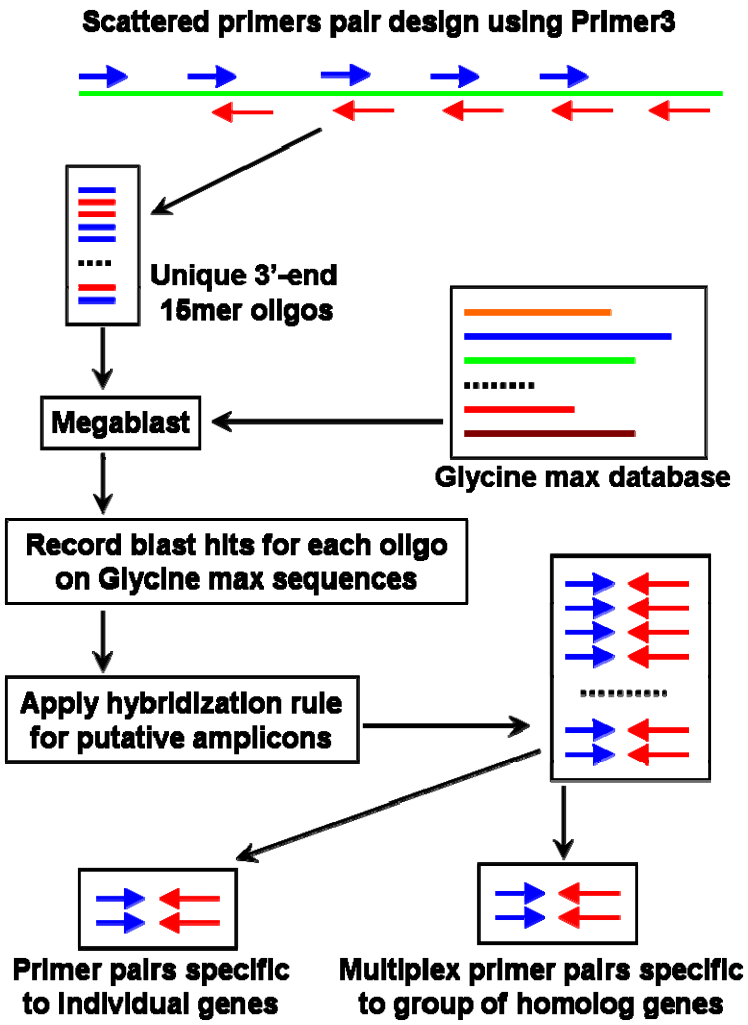

Figure 20: Basic PRIMEGENS model.

Otherwise, PRIMEGENS searches for the unique fragment based on the BLAST result for the query sequence. The optimal global alignment between the query sequence and each of its significant BLAST hits is performed. Based on the alignment, PRIMEGENS searches for the longest unique segment for the query sequence. Finally, it designs primers on the selected gene-specific fragment using Primer3. Figure 20 describes the detailed algorithm of the PRIMEGENS implementation.

PRIMEGENS 2.0 is available in the form of compressed format as *PRIMEGENSv2.zip* for Windows and *PRIMEGENSv2.tgz* for Linux. These packages are freely available for any users and can be downloaded from http://digbio.missouri.edu/primegens/. For installation, user should specify the location of PRIMEGENS folder by setting the environment variable *PRIMEGENS_PATH* with the absolute path of the software location. More detailed description about software installation can be found in *README.txt*. PRIMEGENS 2.0 (PRIMEGENS as the main folder) consists of following major directories and files:

1. bin/: console application executables
2. blast/: BLAST executables
3. doc/: documentation
4. include/: supporting resources
5. output/: output results
6. primer3/: primer3 executables
7. primerdesign/: graphical interface files
8. test/: testing resources
9. README.txt: instruction manual
10. primerdesign.jar: main Java executable for graphical interface

5.2.1. PRIMEGENSv2 Input

PRIMEGENS supports various input features according to the user requirements. Following are some descriptions about the inputs.

- **Sequence Pool**

To start primer design, a user needs to create a database file consisting of all the sequences in the FASTA format. The content format of the database file should look like as shown in Figure 21.

- **Sequence of Interest**

By default, PRIMEGENS searches for the unique sequence-specific fragment and primer pair relative to all the sequences present in database file. Alternatively, if the user is interested in a set of those sequences, a list of these sequences should be provided in to a separate file. This subset file can be either in the FASTA format or a list in which each line gives the name of the gene.

- **Saving Result Files**

PRIMEGENS generates various types of results in different files. This information may be useful subsequently; therefore, a user can specify any location on the local computer to save all the result files.

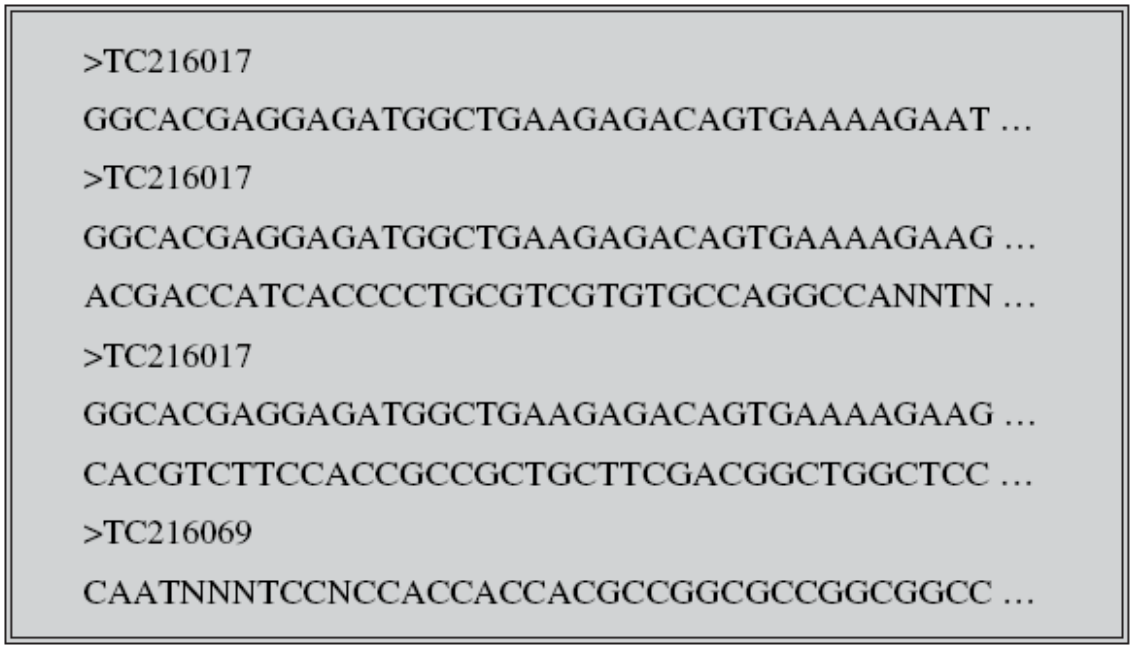

Figure 21: Input database format.

Execution Features (Command-Line Options)

Once the input files are selected, PRIMEGENS provides simple command line execution syntax for console version. The command syntax are same for any operating system as shown below.

> primegens.exe –q <query> –d <database> –p <database path>

> primegens.exe –q <query> –l <database-list> –p <database path>

5.2.2. PRIMEGENSv2 Output

PRIMEGENS supports permanent storage of primer design results. To generate organized results, PRIMEGENS creates various files and directories. Here is a brief description of all types of generated files and directories. For generality, it is assumed that a user has selected both *Database.txt* as the database file and *subset.txt* as the subset

file for PRIMEGENS. If *subset.txt* is not specified, it will select all the entries from *Database.txt* without sequences (only sequence IDs).

5.3. PRIMEGENSv2 Graphical User Interface

To use the GUI version of PRIMEGENS, the database file also has to be prepared first. The user can run the software by double clicking on the executable. Figure 22 provides a systematic workflow from the user perspective for primer design. The details for using the GUI version are explained in the following subheadings.

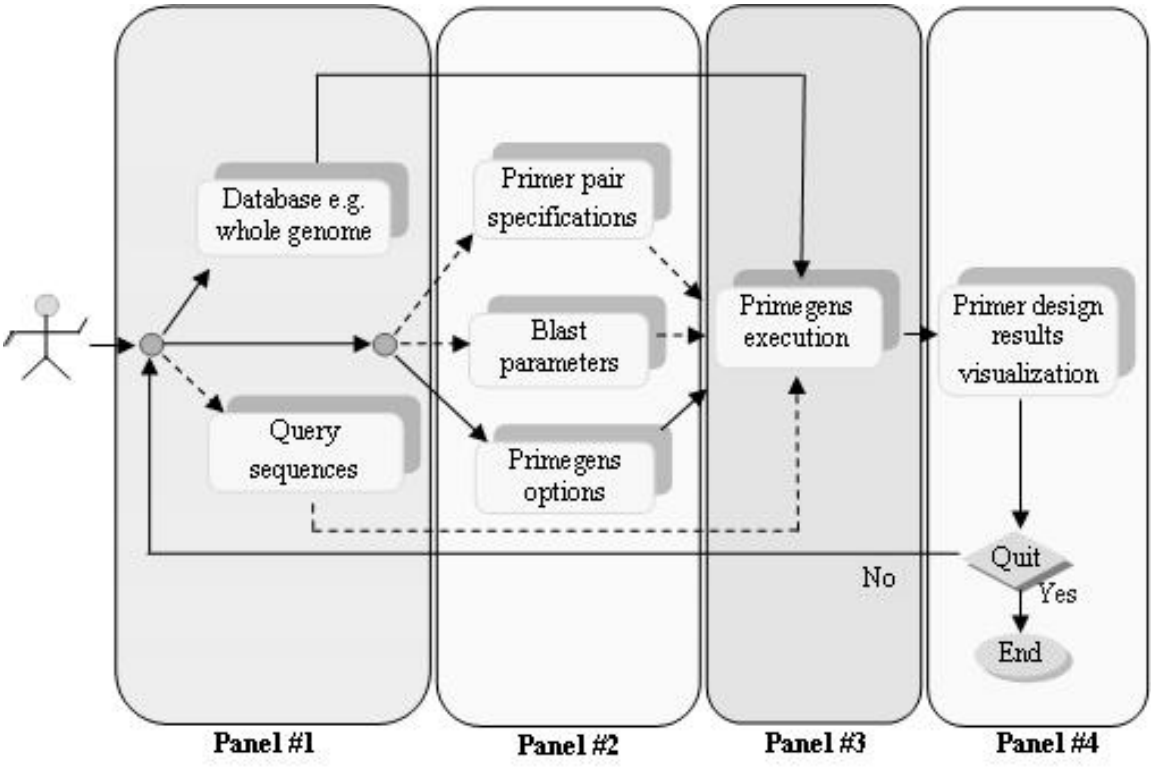

Figure 22: PRIMEGENS software user operation flow chart

5.3.1. Data input panel

Figure 23 shows the first panel, which will be visible when a user clicks to run PRIMEGENS. User should input the database, subset file (optional), and result storage location (optional). In other words, User can select primer design algorithm, query sequence format and input query sequence and database files. Once completed with input, the "next" button should be pressed for execution options.

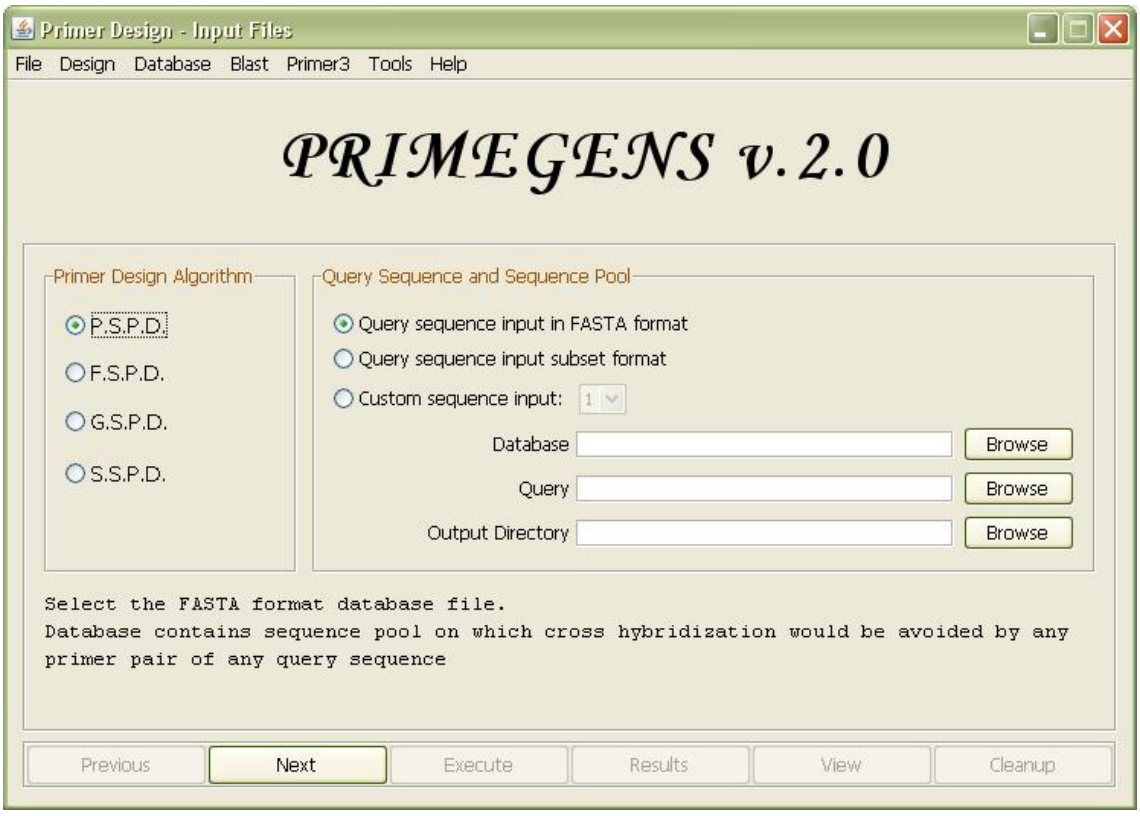

Figure 23: PRIMEGENS main input window

5.3.2. Execution option panel

This panel contains various execution options used by PRIMEGENSv2 while primers design. The execution features supported by the GUI map to all command-line features correspondingly. Figure 24 shows the option panel. Once user selects any attribute, the optional attribute value field shows the default attribute value, which can be modified then.

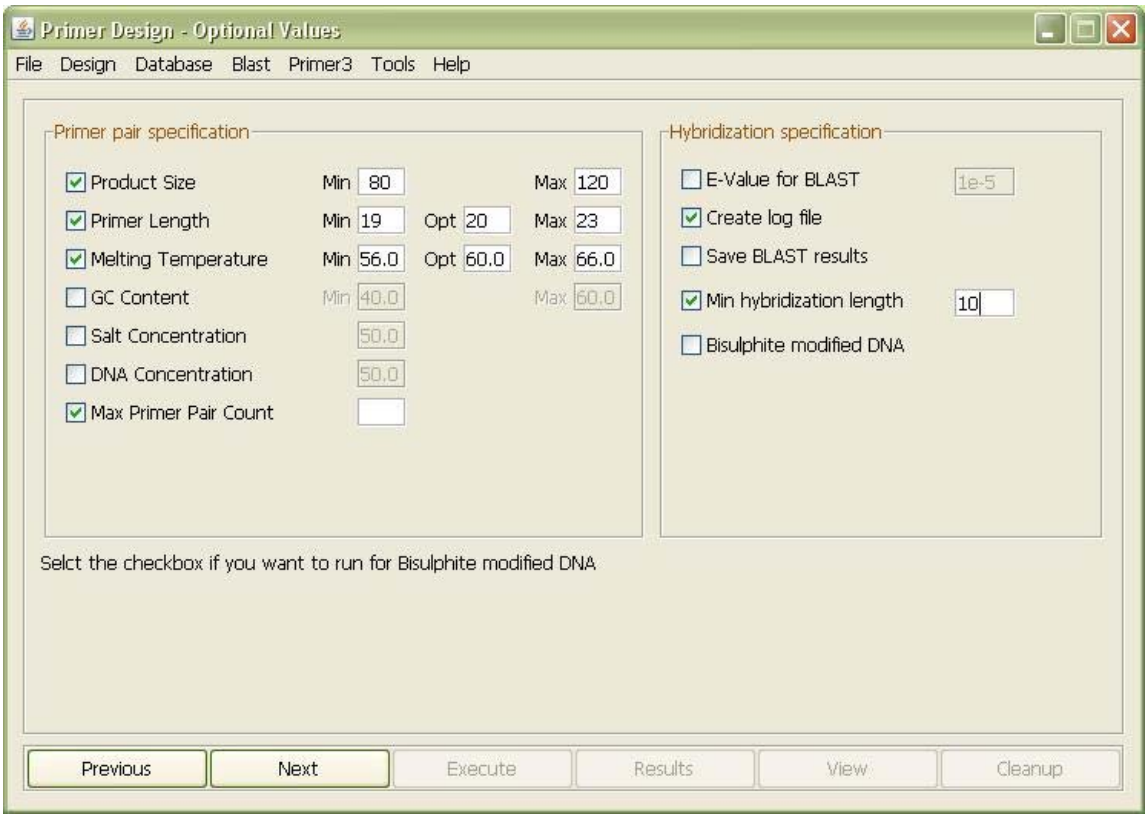

Figure 24: Primer design specification/option window

- **Primer Pair Specification**

In case a user has specific parameter requirements for primer pairs, he or she can specify those values in the primer-specification window. The user can click on the *Primer3* menu

to modify primer specifications. These modifications correspond to changes in the *append.txt* file in the command-line version.

5.3.3. Execution display panel

After specifying inputs and options, the software allows a user to open the execution window. Once the primer design is completed, the *result* and the *clean* buttons are activated. User can click on the *result* button to see the results and *clean* to remove all the temporary results from buffer and reset the software to the first window.

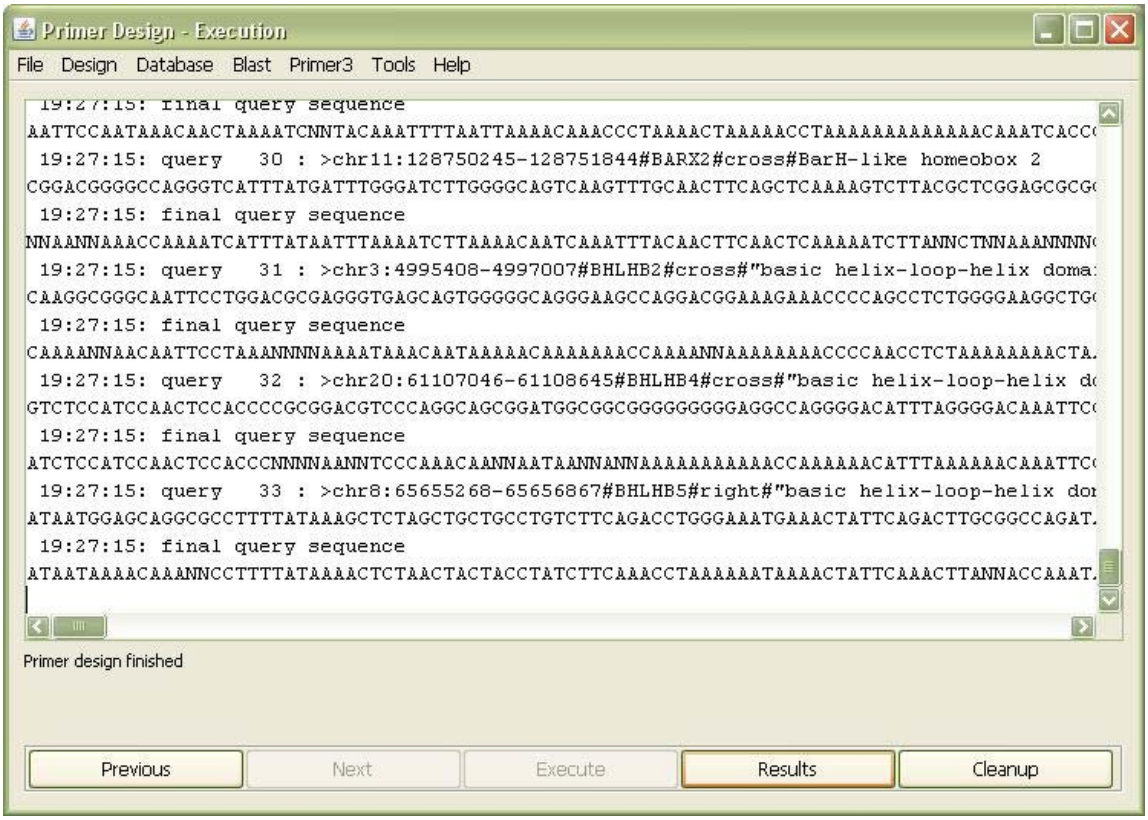

Figure 25: Primer design run time display

5.3.4. Result visualization panel

This window displays the primer design results generated by PRIMEGENS for all the sequences whose primers are found, including the gene-specific fragment and the global alignment between the sequence and its BLAST hits. The various buttons are self-explanatory. Figure 26 shows a sample result visualization window.

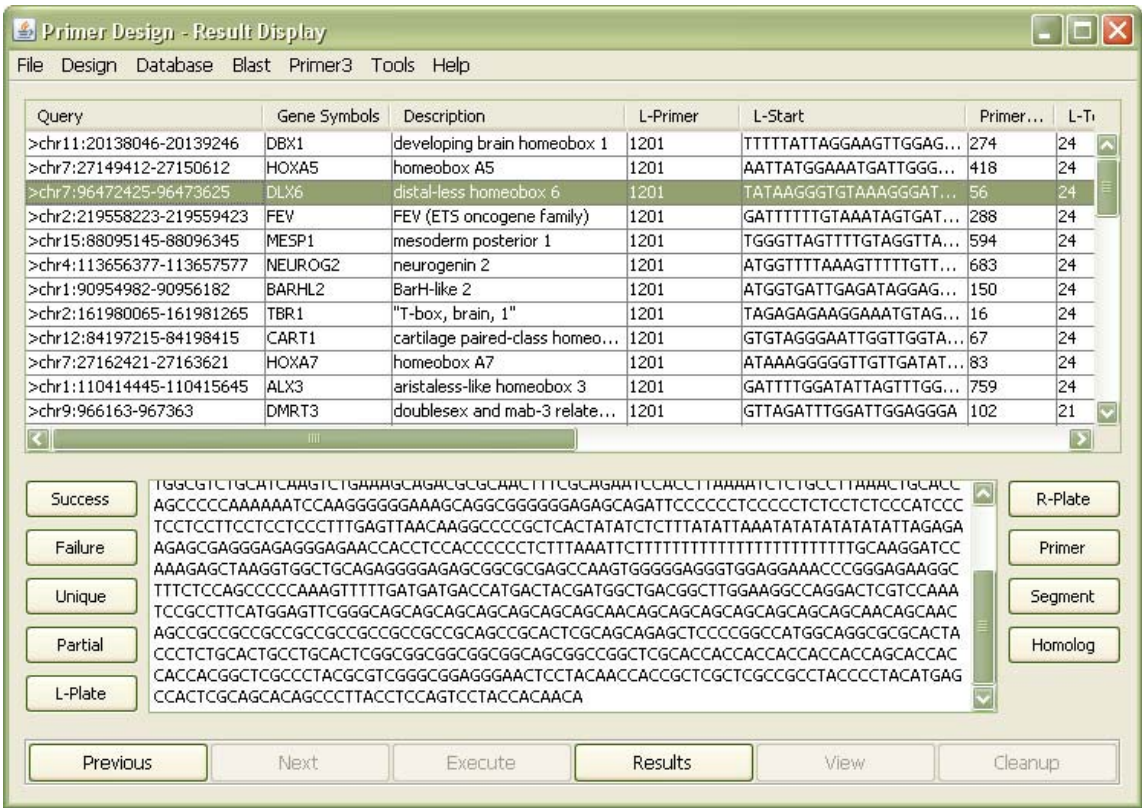

Figure 26: Alternate Primer design results display

5.4. Application of PRIMEGENS: Cancer Epigenetics

DNA methylation plays important roles in biological processes and human diseases, especially cancers. High-throughput bisulfite genomic sequencing based on new-generation of sequencers, such as the 454-sequencing system provides an efficient

method for analyzing DNA methylation patterns. The successful implementation of this approach depends on the use of primer design software capable of performing genome-wide scan for optimal primers from in silico bisulfite-treated genome sequences. We have developed a method, which fulfills this requirement and conduct primer design for sequences including regions of given promoter CpG islands. The developed method has been implemented using the C and JAVA programming languages. The primer design results were tested in the PCR experiments of 96 selected human DNA sequences containing CpG islands in the promoter regions. The results indicate that this method is efficient and reliable for designing sequence-specific primers. The sequence-specific primer design for DNA methylated sequences including CpG islands has been integrated into the second version of PRIMEGENS as one of the primer design features. The software is freely available for academic use at http://digbio.missouri.edu/primegens/.

5.4.1. Introduction

The importance of epigenetic effects in biological processes and diseases has been more and more recognized. Methylation of cytosine residues at CpG dinucleotides is the best studied epigenetic modification in mammalian genomes and is known to have profound effects on gene expression. Over the past three years, an international consensus has emerged in the epigenetics research community for the need of an organized Human Epigenome Project aimed at generating a high-resolution DNA methylation map of the human genome in all major tissues (Eckhardt, et al., 2006; Rakyan, et al., 2004). The recently initiated Human Epigenome Project (HEP) will provide a 'reference epigenome' by re-sequencing different normal tissues and adding 5-methylcytosine to the DNA

sequencing datasets (http://nihroadmap.nih.gov/epigenomics). The pilot human epigenome project (HEP) in Europe utilized direct sequencing of bisulfite PCR products to provide single methyl-cytosine resolution mapping of thousands of amplicons (Eckhardt, et al., 2006). In this method, the methylation present at any given CpG site is estimated by taking the average of all fragments (thousands) generated during PCR, which results in a more statistically robust representation of the methylation patterns as compared to sub-cloning. Recently, we applied an innovative massively parallel sequencing-by-synthesis method (454-sequencing) for ultra-deep bisulfite sequencing analysis of multiple tumor methylome (Taylor, et al., 2007). This highly parallel sequencing system has many potentially important applications, for example development of a high-throughput, large-scale bisulfite genomic sequencing approach that provides an efficient method for deeply exploring the human epigenome.

The successful implementation of above-mentioned approach depends on the use of automatic primer design program capable of performing genome-wide scans for optimal primers from in silico bisulfite-treated human genome sequences. Several methods have been proposed to address this issue partially. MethPrimer (Li and Dahiya, 2002) and PerlPrimer (Marshall, 2004) transform the target sequences according to the bisulfite treatment for primer design. These methods do not provide a mechanism to detect non-specific amplification in bisulfite PCR. Bisearch (Tusnady, et al., 2005) provides an important feature of similarity search for potential non-specific PCR product with the selected primer pairs on a bisulfite-treated genome. It uses a simple string matching search method to detect potential cross hybridization of a designed primer pair. The string matching search can find exact match of the primer but cannot detect highly similar

sequences (e.g., with 1 nt mis-match) in the genome to the primer, which could also be potential binding site for the primer. In addition, this methods is practically not suitable for analyzing the primer pairs for mispriming sites in case of high-throughput primer design, which is required for highly parallel sequencing system to develop high-throughput, large-scale bisulfite genomic sequencing. To address this issue, we developed an efficient method and integrated it into the second version of our software system PRIMEGENS (Srivastava, et al., 2008; Srivastava and Xu, 2007; Xu, et al., 2002), that is PRIMEGENS-v2. PRIMEGENS builds on third-party, open-source software tools like Primer3 (Rozen and Skaletsky, 2000) and BLAST (Altschul, et al., 1997) and has various new features for genome-scale primer design. PRIMEGENS has been widely used and cited by the research community (Haas *et al.,* 2003; Bertone *et al.,* 2005; He *et al*., 2005; Chen *et al*., 2006, Ehses *et al*., 2005). However, our early version did not have the feature of primer design for bisulfite sequencing. It did not have a Graphical User Interface (GUI) and it did not run under Windows O/S. We extended the PRIMEGENS algorithm to include sequence-specific primer design for bisulfite PCR and align these primers using Mega BLAST (Zhang, et al., 2000) to check cross-hybridization across *in silico* bisulfite-treated human genome. We also developed PRIMEGENS as a standalone tool with GUI to run under both Linux and Windows.

5.4.2. Implementation

The goal of our prime design experiment is to design primer for specific region of genes to cover both transcription start site (TSS) of the gene and part of a CpG island, which is located in the vicinity of the gene either in the promoter region or in the transcription

region. Recent studies show that methylation of CpG sites near the TSS is critical to the expression of hTERT gene in cancer cells (Zinn et al., 2007). Since the PRIMEGENS software will be likely most useful for the genome-wide bisulfite sequencing experiments such as the Human Epigenome Pilot Project (Eckhardt, et al., 2006; Rakyan, et al., 2004) we aimed to automate our primer design pipeline so that only a list of gene name is required to perform the primer design. Based on the gene name and the TSS site information associated with the gene, we can automatically archive the target sequences from the human genome database and design the primers based on a few parameters associated with the TSS such as the distance to the TSS. The main computational challenge in such a primer design is to avoid cross hybridization of the fragment-specific primers to the other place of thymine-rich methylated genome. To address this problem, we modified our previously developed PRIMEGENS algorithm as shown in Figure 1. The algorithm is composed of two basic components. The first component performs primer design using Primer3, which provides a set of primer pairs and the second component applies Mega BLAST to search the selected primer pairs against bisulfite-treated genomic sequence to find potential non-specific PCR products.

In order to design primer for any sequence, we first convert the target sequence and the complete human genome into bisulfite-treated sequences, where all the cytosine (C) sites in original sequence are converted into thymine (T) except places where cytosine is preceding guanine (G) known as methylation of the CG. In order to run Mega BLAST for the designed primers, we consider the bisulfite-treated human genome as a database. For each chromosome sequence, four variants are generated as a model of the bisulphite-

treated sequences (Pattyn, et al., 2006) : (1) bisulfite methylated forward sequence, (2) bisulfite methylated reverse sequence, (3) bisulfite unmethylated forward sequence, and (4) bisulfite unmethylated reverse sequence. As an example, suppose a fragment of human genome has nucleotide sequence "agctagccagtcga", then this fragment is modified to generate four variants as follows (Pattyn, et al., 2006) :

agctagccagtcga -- original

agttagttagttga -- unmethylated forward

agttagttagtcga -- methylated forward

ttgattggttagtt -- unmethylated reverse complementary

tcgattggttagtt -- methylated reverse complementary

- **Prepare Input Data**

Before starting the primer design, appropriate query sequences are generated. As a fact, the location of any CpG island with respect to TSS could be in three ways.

- CpG island completely on the left side of the TSS;
- CpG island completely on the right side of the TSS;
- CpG island contains the TSS.

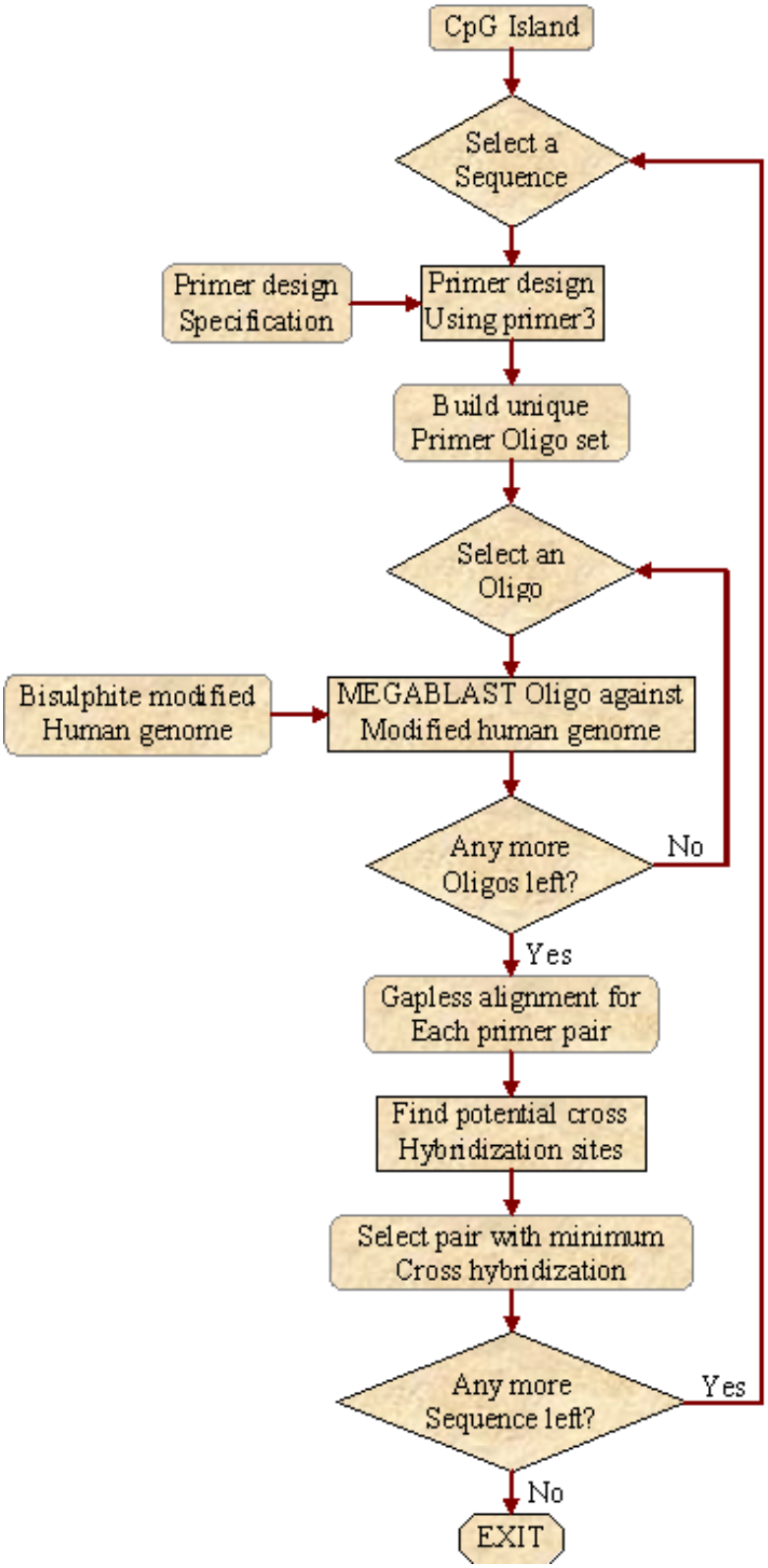

Figure 27: PRIMEGENSv2 algorithm flow chart

Based on the location of a CpG island, a fragment-specific query sequence can be designed so that most of the CpG island region along with the TSS is covered within the PCR product. As an example, Figure 28 explains a strategy to select a query sequence of 1600 nt with a product size ranging 1500 - 1600 so that any CpG island within the vicinity of 1400 nt from the TSS is covered. If a CpG island contains the TSS, the region of the CpG island that covers the TSS is selected. In case of a CpG island either in the left or right side vicinity of the TSS, the genome region of 1500 nt from the TSS toward the CpG island and additional 100 nt from the opposite site is selected as the fragment-specific fragment.

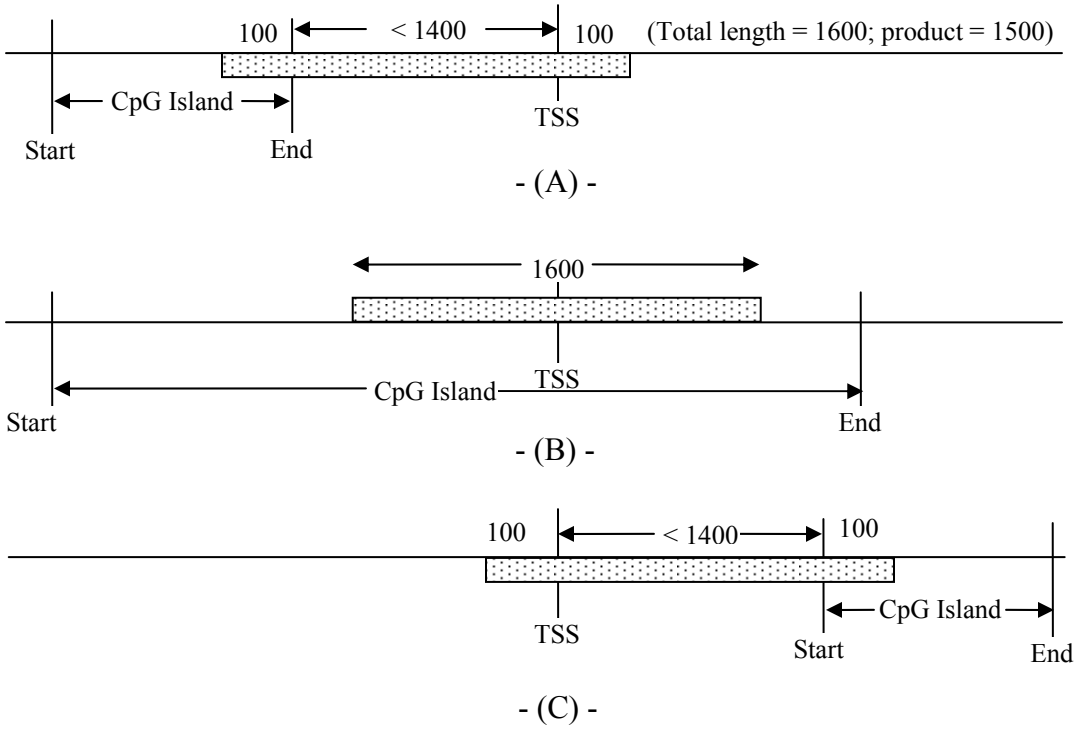

Figure 28: A CpG island can be located near the TSS in three different ways.

This strategy can only be applied when a CpG island is not far from the TSS, in particular, closer than 1400 nt to the either side of TSS. In case the TSS and the CpG island are far away from each other, we partition the selected region to generate reasonably long fragments; e.g. we cut the region into multiple and partially aligned fragments as shown in Figure 29.

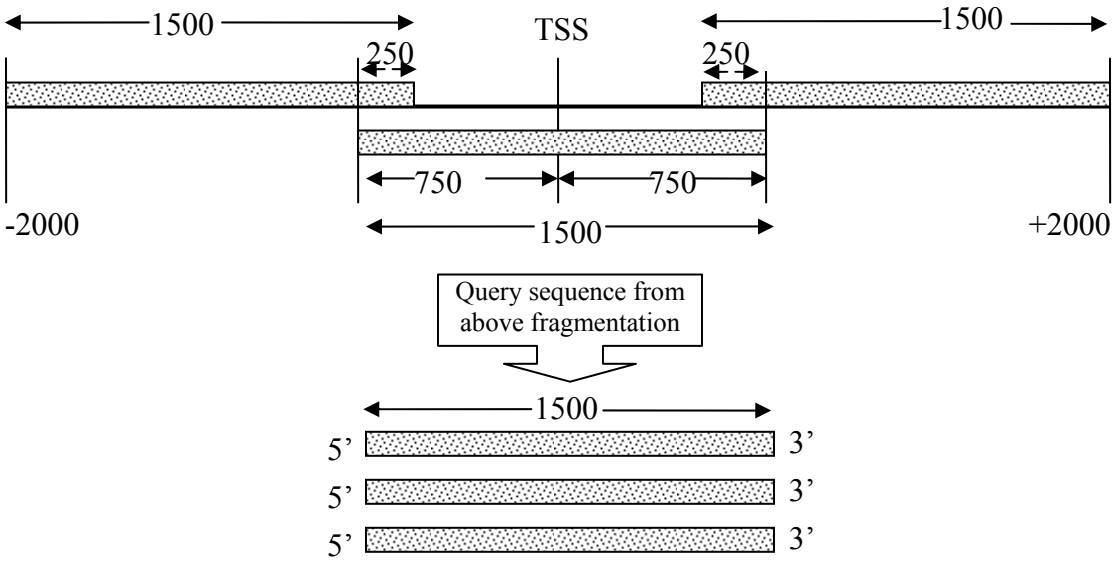

Figure 29: Partitioning method to cover the CpG island region located far from the TSS.

We designed primer pairs for 1012 cancer related genes and randomly selected 96 genes for experimental validation. We first downloaded 1012 CpG island sequences at promoter regions that associate with these genes from the UCSC genome web site http://genome.ucsc.edu and added extra 100 nt upstream (5') and 100 nt downstream (3') region for each sequence. We then performed in silico conversion for the sequences into bisulfite-treated sequences. All the in silico converted DNA sequences were stored in a single file in the FASTA format. Since the bisulfite modification will result in two single

strand sequences that are no longer complementary to each other, firstly the in silico bisulfite converted sense strand sequence is used. If no suitable primer pair is found, the antisense strand is used for designing primers.

5.4.3. Experimental Validation

We successfully designed primer pairs for 1012 query sequences. In order to validate primer design using PRIMEGENS, we randomly picked and synthesized 96 pairs of primers and performed bisulfite PCR using bisulfite-treated DNA in a 96-well PCR plate.

The figure 30 shows experimental validation of the PCR primers designed by PRIMEGEN. Genomic DNA isolated from RL cells was bisulfite-treated and subjected to PCR using primers designed by PRIMEGENS. 96 primer pairs were tested using the same touchdown PCR program in a 96-well PCR plate. The PCR products were examined on 2% agarose gels. 87 out of 96 primer pairs successfully yield expected unique products. * indicates the failed reactions or multiple PCR products. As shown in figure 30, 87 out 96 primers (91%) generated unique PCR products and all PCR products are similar in size as designed. We also used the designed 163 pairs of bisulfite PCR primers that amplify 111 polycomb target genes (Schlesinger et al., 2007; Ohm et al., 2007) and examined the methylation status of the 111 genes in lymphoma cell line RL using COBRA. Surprisingly, 81% (90 out of 111) of the polycomb target genes are methylated in a lymphoma cell line, RL.

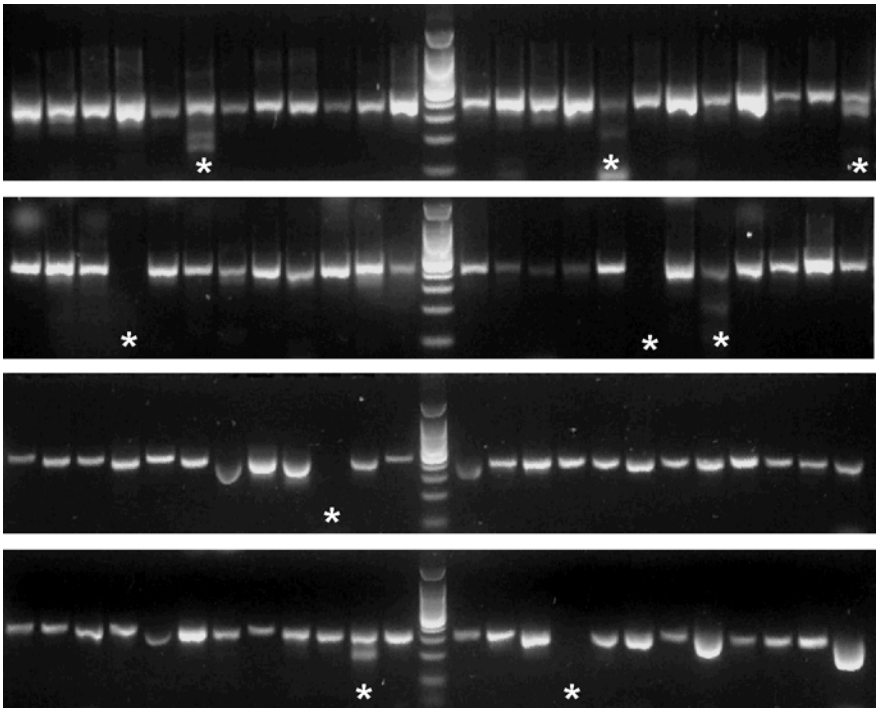

Figure 30: Experimental validation of the PCR primers designed by PRIMEGEN.

To compare with primer design without considering bisulfite-treated sequences, we also analyzed the percentage of generated primer pairs that appear unique via mega BLAST on regular human genome without bisulfite-treated human genome sequences. For all the 96 primer pairs used in the PCR experiment, we applied (1) the regular genome sequence to run mega BLAST and found total 91 (~94%) primer pairs showing unique hybridization (2) bisulfite-treated genome sequences to run mega BLAST and found total 81 (~84%) primer pairs showing unique hybridization. It is clear that ~10% of the primer pairs are excluded due to cross hybridization resulting from bisulfite modification on genome. Interestingly, the experimental success rate (~91%) is higher than what is predicted by PRIMEGENS (~84%). This is because PRIMEGENS uses very stringent criteria (e.g., to allow some nucleotide mismatches of primer as potential cross

hybridization). Some primer pairs are predicted to have possible cross-hybridizations, but the resulting PCR products are experimentally unfavourable or have too low yields to be observed.

5.4.4. Conclusion

In this paper, we present a method capable of designing sequence-specific primer pairs for bisulfite-treated sequences at the genomic scale. This method is incorporated into the second version of PRIMEGENS. It not only searches for appropriate primers but also checks for non-specific PCR amplification. In order to compare PRIMEGENS with other online available bioinformatics tool for primer design, we prepared a table for different features provided by these tools as below.

Tools	Free	Scale	Primer	Specificity	S/W	B/S	Citation
MethPrimer	Yes	One	Yes	No	Online	Yes	Li,L.C. *et al*. (2002)
Meth-BLAST	Yes	One	No	Manual	Online	No	Pattyn,F.et al. (2006)
PerlPrimer	Yes	One	Yes	No	GUI	Yes	Marshall, O.J. (2004)
BiSearch	Yes	One	Yes	Manual	Online	Yes	Tusnády, G.E. et al. (2005)
Methyl-Primer-Express	No	One	Yes	No	GUI	Yes	Applied Biosystems

Epi-Designer	Yes	More	Yes	No	Online	Yes	Sequenom Inc. (2007)
PRIME-GENS	Yes	More	Yes	Automatic	GUI	Yes	Srivastava, G.P. *et al*. 2008

Table 10: Summary of comparison of primer design software

This table summarizes various features provided by other primer design tools available online. The acronyms used for column headers are; a) Primer: feature of primer design available or not, b) Specificity: checking hybridization for designed primer, c) S/W: availability of the software, d) B/S: whether bisulphite primer-design available or not.

We see that while many other tools typically design primers for one gene at a time, PRIMEGENS can design primers for thousands of genes (fragments) by one run with automatic check of primer specificity. The successful rate for the PCR experiment is about 91%. This is a significant improvement compared to our earlier manual design process with the same parameters using Methyl-Primer Express software (Applied Biosystems), which only gave 57% successful rate under the same PCR condition. We designed 227 primers manually using the software, which is downloadable from Applied Biosystems websites. Among the 227 primer pairs, 130 primer pairs worked under a same PCR program without any optimization, this gives 57.3% success rate. 35 primer pairs worked under other PCR conditions, so total 165 out of 227 worked for PCR, which gave a success rate of 72.6%. 62 primer pairs failed for three PCR conditions are not tested further. There are other similar software available online, but to our knowledge Methyl Primer Express is best for such design.

PRIMEGENS uses very stringent criteria (e.g., to allow some nucleotide mismatches of primer as potential cross hybridization). Some primer pairs are predicted to have possible cross-hybridizations, but the resulting PCR products are experimentally unfavourable or have too low yields to be observed. The efficiency of similarity search lies in using Mega BLAST, which is well known as one of the fastest DNA sequence alignment algorithms. Since the typical memory requirement in Mega BLAST against the human genome is higher than what most desktop machines have, the human genome is split into chromosomes and all the four variants of each chromosome are saved, making a total of 96 chromosome files (24 chromosomes with 4 variants for each chromosome). Most running time is taken by Mega BLAST on 4 variants for each of the human chromosome sequence. Due to the scaling problem, users are not advised to use the software on a low-memory desktop machine. Later version of PRIMEGENS will better serve desktop users for bisulfite primer design, by having pre-computed data and indexing for the human genome bundled with the software. We also plan to develop a Web server for PRIMEGENS so that users can apply the software more easily. As more and more experimental studies are conducted for methylation, we expect our tool will benefit many users.

6. References

[1] Alter, O., Brown, P.O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling, *Proc Natl Acad Sci U S A*, **97**, 10101-10106.

[2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

[3] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

[4] Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics*, **4**, 2.

[5] Barrett, T. and Edgar, R. (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis, *Methods Enzymol*, **411**, 352-369.

[6] Barrett, T. and Edgar, R. (2006) Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*, *Methods Mol Biol*, **338**, 175-190.

[7] Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W. and Edgar, R. (2005) NCBI GEO: mining millions of expression profiles--database and tools, *Nucleic Acids Res*, **33**, D562-566.

[8] Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update, *Nucleic Acids Res*, **35**, D760-765.

[9] Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. and Sherlock, G. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics*, **20**, 3710-3715.

[10] Bracken, A.P., Ciro, M., Cocito, A. and Helin, K. (2004) E2F target genes: unraveling the biology, *Trends Biochem Sci*, **29**, 409-417.

[11] Brazhnik, P., de la Fuente, A. and Mendes, P. (2002) Gene networks: how to put the function in genomics, *Trends Biotechnol*, **20**, 467-472.

[12] Buck, M.J. and Lieb, J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, *Genomics*, **83**, 349-360.

[13] Chen, T., He, H.L. and Church, G.M. (1999) Modeling gene expression with differential equations, *Pac Symp Biocomput*, 29-40.

[14] Chen, Y. and Xu, D. (2004) Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae, *Nucleic Acids Res*, **32**, 6414-6424.

[15] Chen, Y. and Xu, D. (2005) Understanding protein dispensability through machine-learning analysis of high-throughput data, *Bioinformatics*, **21**, 575-581.

[16] Choi, J.K., Yu, U., Kim, S. and Yoo, O.J. (2003) Combining multiple microarray studies and modeling interstudy variation, *Bioinformatics*, **19 Suppl 1**, i84-90.

[17] Culhane, A.C., Thioulouse, J., Perriere, G. and Higgins, D.G. (2005) MADE4: an R package for multivariate analysis of gene expression data, *Bioinformatics*, **21**, 2789-2790.

[18] de la Fuente, A., Brazhnik, P. and Mendes, P. (2002) Linking the genes: inferring quantitative gene networks from microarray data, *Trends Genet*, **18**, 395-398.

[19] DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, **278**, 680-686.

[20] Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999) Expression profiling using cDNA microarrays, *Nat Genet*, **21**, 10-14.

[21] Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., Haefliger, C., Horton, R., Howe, K., Jackson, D.K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K. and Beck, S. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22, *Nat Genet*, **38**, 1378-1385.

[22] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A*, **95**, 14863-14868.

[23] Filkov V., S.S., Zhi, J. (2001) Identifying gene regulatory networks from experimental data. *In Proceedings of RECOMB*.

[24] Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data, *J Comput Biol*, **7**, 601-620.

[25] Ghosh, D., Srivastava, G.P., Xu, D., Schulz, L.C. and Roberts, R.M. (2008) A link between SIN1 (MAPKAP1) and poly(rC) binding protein 2 (PCBP2) in counteracting environmental stress, *Proc Natl Acad Sci U S A*, **105**, 11673-11678.

[26] Glazko, G., Gordon, A. and Mushegian, A. (2005) The choice of optimal distance measure in genome-wide datasets, *Bioinformatics*, **21 Suppl 3**, iii3-11.

[27] Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica), *Science*, **296**, 92-100.

[28] Grigoryev, D.N., Ma, S.F., Irizarry, R.A., Ye, S.Q., Quackenbush, J. and Garcia, J.G. (2004) Orthologous gene-expression profiling in multi-species models: search for candidate genes, *Genome Biol*, **5**, R34.

[29] Haake, V., Cook, D., Riechmann, J.L., Pineda, O., Thomashow, M.F. and Zhang, J.Z. (2002) Transcription factor CBF4 is a regulator of drought adaptation in Arabidopsis, *Plant Physiol*, **130**, 639-648.

[30] Hogg RV, M.J.a.C.A. (2005) *Introduction to Mathematical Statistics*.

[31] Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S.H. (2000) Functional discovery via a compendium of expression profiles, *Cell*, **102**, 109-126.

[32] Huttenhower, C., Hibbs, M., Myers, C. and Troyanskaya, O.G. (2006) A scalable method for integration and functional analysis of multiple microarray datasets, *Bioinformatics*, **22**, 2890-2897.

[33] Huynen, M., Snel, B., Lathe, W., 3rd and Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences, *Genome Res*, **10**, 1204-1210.

[34] Jiang, H., Deng, Y., Chen, H.S., Tao, L., Sha, Q., Chen, J., Tsai, C.J. and Zhang, S. (2004) Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes, *BMC Bioinformatics*, **5**, 81.

[35] Joshi, T., Chen, Y., Becker, J.M., Alexandrov, N. and Xu, D. (2004) Genome-scale gene function prediction using multiple sources of high-throughput data in yeast Saccharomyces cerevisiae, *OMICS*, **8**, 322-333.

[36] Joshi T, C.Y., Becker JM, Alexandrov N, Xu D (2004) Function Prediction for Hypothetical Proteins in Yeast *Saccharomyces cerevisiae* Using Multiple Sources of High-Throughput Data *Proceeding of the 8th World Multi-Conference on Systemics, Cybernatics and Informatics*. 17-20.

[37] Joshi, T., Zhang, C., Lin, G.N., Song, Z. and Xu, D. (2008) GeneFAS: A tool for prediction of gene function using multiple sources of data, *Methods Mol Biol*, **439**, 369-386.

[38] Kasuga, M., Liu, Q., Miura, S., Yamaguchi-Shinozaki, K. and Shinozaki, K. (1999) Improving plant drought, salt, and freezing tolerance by gene transfer of a single stress-inducible transcription factor, *Nat Biotechnol*, **17**, 287-291.

[39] Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J. and Harter, K. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses, *Plant J*, **50**, 347-363.

[40] Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N. and Davidson, G.S. (2001) A gene expression map for Caenorhabditis elegans, *Science*, **293**, 2087-2092.

[41] Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. and Pavlidis, P. (2004) Coexpression analysis of human genes across many microarray data sets, *Genome Res*, **14**, 1085-1094.

[42] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young, R.A. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science*, **298**, 799-804.

[43] Letovsky, S. and Kasif, S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach, *Bioinformatics*, **19 Suppl 1**, i197-204.

[44] Li, L.C. and Dahiya, R. (2002) MethPrimer: designing primers for methylation PCRs, *Bioinformatics*, **18**, 1427-1431.

[45] Lukowitz, W., Gillmor, C.S. and Scheible, W.R. (2000) Positional cloning in Arabidopsis. Why it feels good to have a genome initiative working for you, *Plant Physiol*, **123**, 795-805.

[46] Ma, S., Gong, Q. and Bohnert, H.J. (2007) An Arabidopsis gene network based on the graphical Gaussian model, *Genome Res*, **17**, 1614-1625.

[47] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function, *Nature*, **402**, 83-86.

[48] Marinelli, R.J., Montgomery, K., Liu, C.L., Shah, N.H., Prapong, W., Nitzberg, M., Zachariah, Z.K., Sherlock, G.J., Natkunam, Y., West, R.B., van de Rijn, M., Brown, P.O. and Ball, C.A. (2008) The Stanford Tissue Microarray Database, *Nucleic Acids Res*, **36**, D871-877.

[49] Markowetz, F. and Spang, R. (2007) Inferring cellular networks--a review, *BMC Bioinformatics*, **8 Suppl 6**, S5.

[50] Marshall, O.J. (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR, *Bioinformatics*, **20**, 2471-2472.

[51] Mauch-Mani, B. and Mauch, F. (2005) The role of abscisic acid in plant-pathogen interactions, *Curr Opin Plant Biol*, **8**, 409-414.

[52] Mittler, R., Kim, Y., Song, L., Coutu, J., Coutu, A., Ciftci-Yilmaz, S., Lee, H., Stevenson, B. and Zhu, J.K. (2006) Gain- and loss-of-function mutations in Zat10 enhance the tolerance of plants to abiotic stress, *FEBS Lett*, **580**, 6537-6542.

[53] Opgen-Rhein, R. and Strimmer, K. (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data, *BMC Syst Biol*, **1**, 37.

[54] Padidam, M. (2003) Chemically regulated gene expression in plants, *Curr Opin Plant Biol*, **6**, 169-177.

[55] Padidam, M., Gore, M., Lu, D.L. and Smirnova, O. (2003) Chemical-inducible, ecdysone receptor-based gene expression system for plants, *Transgenic Res*, **12**, 101-109.

[56] Palaniswamy, S.K., James, S., Sun, H., Lamb, R.S., Davuluri, R.V. and Grotewold, E. (2006) AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks, *Plant Physiol*, **140**, 818-829.

[57] Park, T., Yi, S.G., Shin, Y.K. and Lee, S. (2006) Combining multiple microarrays in the presence of controlling variables, *Bioinformatics*, **22**, 1682-1689.

[58] Pattyn, F., Hoebeeck, J., Robbrecht, P., Michels, E., De Paepe, A., Bottu, G., Coornaert, D., Herzog, R., Speleman, F. and Vandesompele, J. (2006) methBLAST and methPrimerDB: web-tools for PCR based methylation analysis, *BMC Bioinformatics*, **7**, 496.

[59] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc Natl Acad Sci U S A*, **96**, 4285-4288.

[60] Piatetsky-Shapiro G, T.P. (2003) Microarray Data Mining: Facing the Challenges, *SIGKDD Explorations*, **5**.

[61] Poole, R.L. (2007) The TAIR database, *Methods Mol Biol*, **406**, 179-212.

[62] Rakyan, V.K., Hildmann, T., Novik, K.L., Lewin, J., Tost, J., Cox, A.V., Andrews, T.D., Howe, K.L., Otto, T., Olek, A., Fischer, J., Gut, I.G., Berlin, K. and Beck, S. (2004) DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project, *PLoS Biol*, **2**, e405.

[63] Ramirez-Parra, E., Frundt, C. and Gutierrez, C. (2003) A genome-wide identification of E2F-regulated genes in Arabidopsis, *Plant J*, **33**, 801-811.

[64] Reiser, L. and Rhee, S.Y. (2005) Using the Arabidopsis Information Resource (TAIR) to find information about Arabidopsis genes, *Curr Protoc Bioinformatics*, **Chapter 1**, Unit 1 11.

[65] Reverter, A., Wang, Y.H., Byrne, K.A., Tan, S.H., Harper, G.S. and Lehnert, S.A. (2004) Joint analysis of multiple cDNA microarray studies via multivariate mixed models applied to genetic improvement of beef cattle, *J Anim Sci*, **82**, 3430-3439.

[66] Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. and Chinnaiyan, A.M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer, *Cancer Res*, **62**, 4427-4433.

[67] Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A.M. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression, *Proc Natl Acad Sci U S A*, **101**, 9309-9314.

[68] Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers, *Methods Mol Biol*, **132**, 365-386.

[69] Schafer, J. and Strimmer, K. (2005) An empirical Bayes approach to inferring large-scale gene association networks, *Bioinformatics*, **21**, 754-764.

[70] Schafer, J. and Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Stat Appl Genet Mol Biol*, **4**, Article32.

[71] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467-470.

[72] Schlicht, M., Matysiak, B., Brodzeller, T., Wen, X., Liu, H., Zhou, G., Dhir, R., Hessner, M.J., Tonellato, P., Suckow, M., Pollard, M. and Datta, M.W. (2004) Cross-species global and subset gene expression profiling identifies genes involved in prostate cancer response to selenium, *BMC Genomics*, **5**, 58.

[73] Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of Arabidopsis thaliana development, *Nat Genet*, **37**, 501-506.

[74] Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast, *Nat Biotechnol*, **18**, 1257-1261.

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

[76] Seki, M., Ishida, J., Narusaka, M., Fujita, M., Nanjo, T., Umezawa, T., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T., Satou, M., Akiyama, K., Yamaguchi-Shinozaki, K., Carninci, P., Kawai, J., Hayashizaki, Y. and Shinozaki, K. (2002) Monitoring the expression pattern of around 7,000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray, *Funct Integr Genomics*, **2**, 282-291.

[77] Seki, M., Narusaka, M., Abe, H., Kasuga, M., Yamaguchi-Shinozaki, K., Carninci, P., Hayashizaki, Y. and Shinozaki, K. (2001) Monitoring the expression pattern of 1300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray, *Plant Cell*, **13**, 61-72.

[78] Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A. and Shinozaki, K. (2002) Functional annotation of a full-length Arabidopsis cDNA collection, *Science*, **296**, 141-145.

[79] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res*, **13**, 2498-2504.

[80] Sheen, J. (2001) Signal transduction in maize and Arabidopsis mesophyll protoplasts, *Plant Physiol*, **127**, 1466-1475.

[81] Shinozaki, K., Yamaguchi-Shinozaki, K. and Seki, M. (2003) Regulatory network of gene expression in the drought and cold stress responses, *Curr Opin Plant Biol*, **6**, 410-417.

[82] Sørensen, J.G., Kristensen, T.N. and Loeschcke, V. (2003) The evolutionary and ecological role of heat shock proteins. In Letters, E. (ed). Blackwell Publishing, pp. 1025-1037(1013).

[83] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol Biol Cell*, **9**, 3273-3297.

[84] Srivastava, G.P., Guo, J., Shi, H. and Xu, D. (2008) PRIMEGENS-v2: genome-wide primer design for analyzing DNA methylation patterns of CpG islands, *Bioinformatics*, **24**, 1837-1842.

[85] Srivastava, G.P., Qiu, J. and Xu, D. (2009) Genome-wide functional annotation by integrating multiple microarray datasets using meta-analysis, *Int J Data Min Bioinform*, **(in press)**.

[86] Srivastava, G.P. and Xu, D. (2007) Genome-scale probe and primer design with PRIMEGENS, *Methods Mol Biol*, **402**, 159-176.

[87] Stevens, J.R. and Doerge, R.W. (2005) Combining Affymetrix microarray results, *BMC Bioinformatics*, **6**, 57.

[88] Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules, *Science*, **302**, 249-255.

[89] Swindell, W.R. (2006) The association among gene expression responses to nine abiotic stress treatments in Arabidopsis thaliana, *Genetics*, **174**, 1811-1824.

[90] Taylor, R.C., Patel, A., Panageas, K.S., Busam, K.J. and Brady, M.S. (2007) Tumor-infiltrating lymphocytes predict sentinel lymph node positivity in patients with cutaneous melanoma, *J Clin Oncol*, **25**, 869-875.

[91] Toh, H. and Horimoto, K. (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling, *Bioinformatics*, **18**, 287-297.

[92] Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. and Botstein, D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae), *Proc Natl Acad Sci U S A*, **100**, 8348-8353.

[93] Tusnady, G.E., Simon, I., Varadi, A. and Aranyi, T. (2005) BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes, *Nucleic Acids Res*, **33**, e9.

[94] Vandepoele, K., Vlieghe, K., Florquin, K., Hennig, L., Beemster, G.T., Gruissem, W., Van de Peer, Y., Inze, D. and De Veylder, L. (2005) Genome-wide identification of potential plant E2F target genes, *Plant Physiol*, **139**, 316-328.

[95] Wang, H., Tang, W., Zhu, C. and Perry, S.E. (2002) A chromatin immunoprecipitation (ChIP) approach to isolate genes regulated by AGL15, a MADS domain protein that preferentially accumulates in embryos, *Plant J*, **32**, 831-843.

[96] Warnat, P., Eils, R. and Brors, B. (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes, *BMC Bioinformatics*, **6**, 265.

[97] Wichert, S., Fokianos, K. and Strimmer, K. (2004) Identifying periodically expressed transcripts in microarray time series data, *Bioinformatics*, **20**, 5-20.

[98] Wille, A. and Buhlmann, P. (2006) Low-order conditional independence graphs for inferring genetic networks, *Stat Appl Genet Mol Biol*, **5**, Article1.

[99] Xu, D., Li, G., Wu, L., Zhou, J. and Xu, Y. (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis, *Bioinformatics*, **18**, 1432-1437.

[100] Xu, X., Chen, C., Fan, B. and Chen, Z. (2006) Physical and functional interactions between pathogen-induced Arabidopsis WRKY18, WRKY40, and WRKY60 transcription factors, *Plant Cell*, **18**, 1310-1326.

[101] Yugi, K., Nakayama, Y., Kojima, S., Kitayama, T. and Tomita, M. (2005) A microarray data-based semi-kinetic method for predicting quantitative dynamics of genetic networks, *BMC Bioinformatics*, **6**, 299.

[102] Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences, *J Comput Biol*, **7**, 203-214.

[103] Zhou, X.J., Kao, M.C., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O.M., Finch, C.E., Morgan, T.E. and Wong, W.H. (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data, *Nat Biotechnol*, **23**, 238-243.

7. Vita

Gyan Prakash Srivastava is a PhD student in the Department of Computer Science at University of Missouri, Columbia. His research interests include statistical analysis of microarray data for gene function prediction and regulatory network construction, data mining, database and bioinformatics studies of high-throughput sequencing analysis. He obtained his Bachelor of Technology (B. Tech.) from Indian Institute of Technology Kanpur (IIT-K) India in Electrical Engineering in 2002.

He was born in Lucknow, North of India in 1979 and completed his schooling in native language in local colleges. During his intermediate studies, he was selected in many prestigious engineering colleges through respective entrance examinations. In 1998 he joined IIT for further study and completed his bachelor degree in 2002.

Later he joined Center of Development of Telematics (C-DOT) as research engineer to carry out research in Telematics and also visited university of California Irvine, Developmental biology laboratory in 2004. He joined PhD program in 2005 under Dr. Dong Xu, who is department chair and McDowell endowed professor of computer science department in university of Missouri Columbia.