Summer 8-2007

EXPRESSION SEQUENCE TAGS ANALYSIS, ANNOTATION, TOXICOGENOMICS, AND LEARNING APPROACH

Mehdi Pirooznia
University of Southern Mississippi

The University of Southern Mississippi

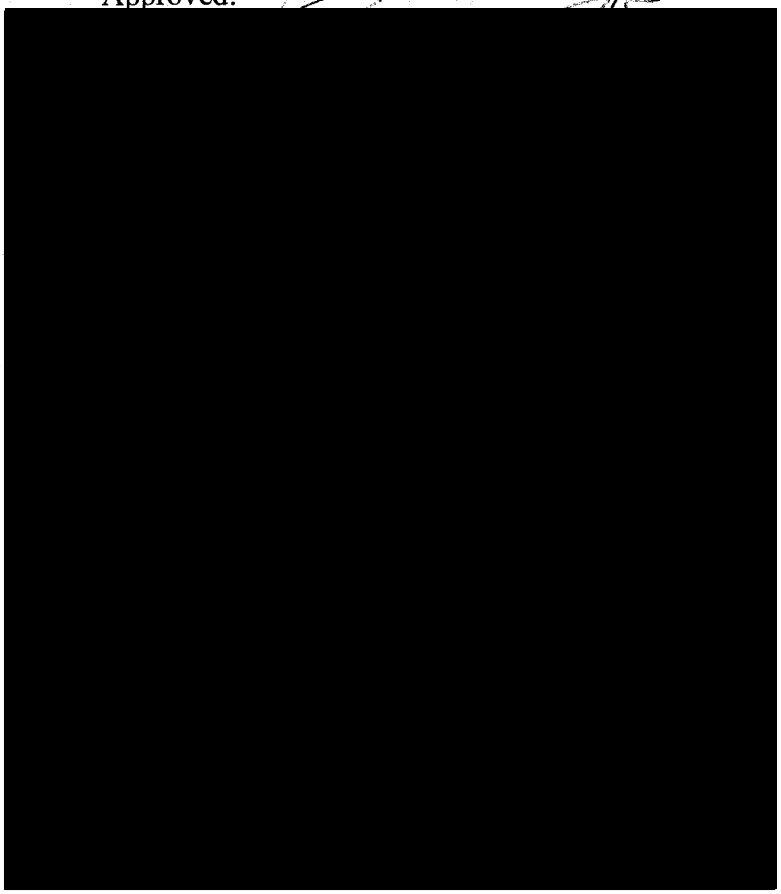
EXPRESSION SEQUENCE TAGS ANALYSIS, ANNOTATION, TOXICOGENOMICS, AND LEARNING APPROACH

by

Mehdi Pirooznia

A Dissertation
Submitted to the Graduate Studies Office
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved:



August 2007

The University of Southern Mississippi

EXPRESSION SEQUENCE TAGS ANALYSIS, ANNOTATION, TOXICOGENOMICS, AND LEARNING APPROACH

by

Mehdi Pirooznia

A Dissertation
Submitted to the Graduate Studies Office
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

August 2007

ABSTRACT

EXPRESSION SEQUENCE TAGS ANALYSIS, ANNOTATION,

TOXICOGENOMICS, AND LEARNING APPROACH

By Mehdi Pirooznia

August 2007

Genome sequence of many organisms is still unknown. Earthworm, *Eisenia fetida*, commonly known as compost worm, was described by Aristotle as "the intestine of the earth". Little is known about its genome sequence although it has been extensively used as a test organism in terrestrial ecotoxicology. In order to understand its gene expression in response to environmental contaminants, we cloned 4032 cDNAs or expressed sequence tags (ESTs) from two *E. fetida* libraries. Clustering analysis yielded 2231 unique sequences including 448 contigs (from 1361 ESTs) and 1783 singletons. We stored all the information along with Gene Ontology and Pathway information at a highly performed relational database called EST model database (ESTMD) an integrated Web-based database model.

To understand molecular mechanisms of the chronic, sublethal effects of 2,4,6-trinitrotoluene (TNT), a widely used ordnance compound of public concerns, we constructed a microarray consisting of 4,032 cDNA isolated from the earthworm *Eisenia fetida*. Based on the reproduction response to TNT, four treatments, i.e., control, 7, 35 and 139 ppm, were selected for gene expression studies. We performed an interwoven loop designed microarray experiment. Statistical data analysis identified that the expression of 109 significant transcripts. A down-regulation of chitinase genes and evidence blood disorders, weaken immunity and decrease digestion in *E. fetida* has been

reported. We also implemented a java application that allows easy evaluation of errors and the role of hybrid normalization methods to remove the systematic errors from the experiment's data.

Another important aspect of microarray analysis is classification of data and pattern recognition. Several classifications methods have been studied for the identification of differentially expressed genes in microarray data. However there is lack of comparison between these methods to find a better framework for classification, clustering and analysis of microarray gene expression results. We compared the efficiency of the classification methods. We reported that the choice of feature selection and classification methods substantially influence classification success. We also developed a java GUI application, called SVM Classifier, that allows SVM users to perform SVM training, classification and prediction.

DEDICATION

This dissertation is dedicated to Samaneh for years of love and support. Without her constant support and belief in my abilities, this project would have never been accomplished.

ACKNOWLEDGMENTS

I would like to thank the dissertation director, Dr. Youping Deng, and the other committee members, Drs. Shiao Wang, Mohamed Elasri, Chaoyang Zhang, and Jonathan Sun, for their advice and support throughout the duration of this project. I would especially like to thank Dr. Shiao Wang for his enormous patience and advice.

My special thanks must be expressed to Dr. Frank Moore, chair of the Department of Biological Sciences at The University of Southern Mississippi (USM) for his extreme support and advice.

Appreciation must also be expressed to my colleagues Tanwir Habib and Nicole Thompson for their efforts in reading the drafts and suggesting clarifications, and my other lab members, Puneet Bandi, Venkata Thodima, and Kuan Yang, for their help during this project.

The laboratory experiments including cDNA library construction, microarray experiment and RT-PCR have been implemented in collaboration with Engineer Research and Development Center (ERDC) at Vicksburg, therefore I would like to thank Drs. Edward J. Perkins (PI) and Ping Gong from ERDC who provided me the raw data.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure

LIST OF TABLES

Table

CHAPTER I

INTRUDUCTION

cDNA Library and Expression Sequence Tag Analysis

Earthworm (Eisenia fetida)

As key representatives of the soil fauna, earthworms are essential in maintaining soil fertility through their burrowing, ingestion and excretion activities (Liu *et al.* 2005). There are over 8000 described species worldwide, existing everywhere but in polar and arid climates (Bradham and others 2006). They are increasingly recognized as indicators of agroecosystem health and ecotoxicological sentinel species because they are constantly exposed to contaminants in soil (Rombke *et al.* 2005). The earthworm species (e.g., *Eisenia fetida*, *Eisenia andrei*, and *Lumbricus terrestris*) widely used in standardized acute and reproduction toxicity tests belong to the Lumbricidae family (phylum, Annelida; class, Clitellata; subclass Oligochaeta; order, Haplotaxida; superfamily, Lumbricoidea; family, Lumbricidae). *E. fetida and E. andrei* are two sibling species commonly found in North American composters and are sold commercially for fish bait. They have a life span of four or five years and are obligatorily amphimictic even though each worm has both male and female reproductive organs (Bundy *et al.* 2002).

Like many other ecologically important species, genomics research in earthworms legs far behind other model species such as *Mus musculus* and *Caenorhabditis elegans*. In the absence of full genome sequences, expressed sequence tags (ESTs) allow rapid identification of expressed genes by sequence analysis and are an important resource for comparative and functional genomics studies (Plant 2006). ESTs are often generated

from either end of randomly selected cDNA clones and provide valuable transcriptional data for the annotation of genomic sequences. Because of recent advances in biotechnology, ESTs are produced daily in large quantities, with nearly 42 million entries in the current GenBank db EST database (released 03/02/07). Nevertheless, it is still a challenging bioinformatics problem to analyze and annotate the often short, redundant, yet error prone EST sequences in an appropriate and efficient manner, especially when the genome sequence of the organism is unknown. Recent years have seen some EST projects undertaken with *L. rubellus* (Sturzenbaum *et al.* 2004) and *E. andrei* (Lee *et al.* 2005), which have generated 19,934 and 1,108 ESTs, respectively (db EST released 03/02/07). Before this study, there were only 96 nucleotide and 89 protein Entrez records found for *E. fetida.*

We cloned, sequenced and analyzed 4032 ESTs from *E. fetida.* We used suppression subtractive hybridization-PCR (SSH) to enrich cDNAs responsive to ten ordnance related compounds (ORCs). The objectives of this part of study were to make the *E. fetida* EST information publicly accessible by integrating it to our web-based EST model organism database (ESTMD) so that it can be shared with interested parties.

GOfetcher: A Complex Searching Facility for Gene Ontology

Biologists waste a lot of time and effort to search for information about genes in their research. This problem increases by the wide variations in terminology that may cause a huge redundancy in information. Gene Ontology (GO) (Gene Ontology Consortium 2001) has been widely used to characterize gene function annotation and classification. The GO project is a collaborative attempt to address the need for descriptions of gene products in different databases (Harris *et al.* 2004). Since 1998, the

GO consortium has grown to include many databases, including several of the world's major repositories for plant, animal and microbial genomes (Ashburner *et al.* 2000). The Gene Ontology has several benefits including long-term maintenance of annotation datasets and avoiding redundancy. Some model species research groups do not have an established database and/or time to commit to long-term maintenance of their datasets. Such groups can supply annotations to the central repository GO project (Harris *et al.* 2004).

The GO project has been developed based on three structured vocabularies, called ontologies, which describe gene products in terms of their associated "biological processes", "cellular components" and "molecular functions" (Wu *et al.* 2006).

One important aspect of the GO project is to develop tools that facilitate the creation, maintenance and use of ontologies. Several tools have been created for communicating and using the GO. They are divided into two categories: (1) Consortium tools, those that are developed within GO consortium and (2) Non-Consortium Tools, which are developed by other groups. Among the consortium developed tools is the AmiGO, which provides an interface to search and browse the ontology and annotation data. Several tools are included in second group, including GOProfiler, GORetriever, GOSlimViewer, and GOArray. These tools are either for searching and browsing or annotation of GO. They search perfectly either GO or their own specific database and the outcomes are often satisfactory. Some of them support batch searching, while the others only have single keyword searching.

However, one problem with most of these tools is that they suffer from lack of a comprehensive search facility. Data is increasing every day. Therefore, needs for fast and

accurate handling of the search queries especially for batch searching and the output increase, too. This problem particularly arises when one is handling several thousands of search queries such as annotation results of the Expressed Sequence Tags (EST) data for an organism.

Another problem is that the output formats are very limited. Almost all of these tools generate their output in either HTML or Text format. This might not seem very crucial if the search keywords are limited to tens or even hundreds, but when one has thousands of keywords searching simultaneously the output format would be an important issue.

In this project we developed a web application, GOfetcher, with a very comprehensive search facility for GO project and variety of output formats for the results to overcome these problems. It can be accessed at http://mcbc.usm.edu/GOfetcher.

ESTMD - An Integrated Web-Based EST Model Database

Bioinformatics has evolved into a multidisciplinary subject that integrates developments in information and computer technology that applied to Biological Sciences. It uses computer software for database creation, data management, data warehousing, data mining and networking. One of the main areas of bioinformatics is to design and develop Web-based applications containing biological database management, information retrieval, data mining and analysis tools to speed up and enhance biological research (Deng *et al.* 2006a; Latorre *et al.* 2006). In this part, we developed an integrated Web-based model to manage, analyze and retrieve Expressed Sequence Tags (ESTs) data that are partial sequences of randomly chosen cDNA obtained from the results of a single DNA sequencing reaction (Deng *et al.* 2006a).

Typically, processing ESTs includes raw sequence cleansing such as low quality, vector and adaptor sequence removal, assembly process to generate contig unique sequences, and unique sequence annotation and functional assignment (Nagaraj *et al.* 2007). Keeping track of and managing the information is a critical issue for many labs. Currently available EST database software, e.g. ESTAP (Mao *et al.* 2003) has many limitations. It mainly focuses on data analysis, and does not support GO retrieval. ESTIMA (Kumar *et al.* 2004), a tool for EST management in a multi-project environment, has limited services for detailed EST information search. RED (Everitt *et al.* 2002) provides only two simple search tools, keyword and GO term. Other tools such as ESTWeb (Paquola *et al.* 2003) and PipeOnline (Ayoubi *et al.* 2002) mainly focus on developing software packages designed for uniform data processing pipelines other than EST information management and presentation. Although ESTWeb package provides for reception of sequencing chromatograms, sequence processing such as base-calling, vector screening, comparison with public databases; and storage of data and analysis in a relational database, but none of them provide pathway searches so far.

We developed a high-performance Web-based application consisting of EST modeling and database (ESTMD) to facilitate and enhance the retrieval and analysis of EST information. We upgraded our previous developed EST model database (ESTMD version 1) (Deng *et al.* 2006a) and integrated the earthworm and Sheepshead EST information into the new version of ESTMD with many new features.

The ESTMD provides a number of comprehensive search tools for mining EST raw, cleaned and unique sequences, Gene Ontology, pathway information and a variety of genetic Web services such as BLAST search, data submission and sequence download

pages. The software is developed using advanced Java technology and it supports portability, extensibility and data recovery. It can be accessed at http://mcbc.usm.edu/estmd.

Toxicogenomics Study of Earthworm

The subclass Oligochaeta, commonly known as "earthworms", are the second largest group of the Annelida, with 3100 known terrestrial and aquatic species[1] in both freshwater and marine that make up about one third of the phylum (Jager 2004; Kuperman *et al.* 2004). Earthworms are enormously important in the construction and fertility maintenance of the soil. Some earthworm species are also used as ecotoxicological model organisms for which both acute and reproduction toxicity tests which are lengthy (14 to 56 d) (Alvarenga *et al.* 2007). In spite of the availability of a large database of earthworm toxicity tests, these tests only evaluate a few endpoints such as lethality, weight change, and juvenile counts.

2,4,6-trinitrotoluene (TNT) can be used as a test compound because it is one of the most widely used high explosives, and both the public and the military are highly concerned about its impacts on human health and wildlife. Continued production and use of such chemical compounds have resulted in contamination of related lands and facilities. For instance, many military installations have detectable TNT residues in soil ranging from 0.08 to 64,000 mg/kg (Hovatter *et al.* 1997), and thousands mg/kg of TNT in surface soils next to detonations are found at army fire training ranges (Jenkins *et al.* 2006). Health effects reported in people exposed to TNT include anemia, abnormal liver

[1] http://www.earthlife.net/inverts/oligochaeta.html

function, skin irritation, and cataracts (Sabbioni *et al.* 1996; Tchounwou *et al.* 2001). Hematological, biochemical, pathological and immunological effects of TNT were also demonstrated in animals (Johnson *et al.* 2000; Reddy *et al.* 2000). TNT may also target nucleic acids, proteins, and lipids, directly causing cell injuries or disrupting signal transduction. TNT has also been found inducing damage to spermatozoa in male rats through DNA damage mediated by its metabolite (Homma-Takeda *et al.* 2002). TNT's lethal and reproductive toxicities are also reported in various earthworm species (Kuperman *et al.* 2006). In the part of project, we investigated gene expression response to TNT toxicity in *Eisenia fetida* by using microarray technologies.

Microarray technologies are increasingly being used in biological and medical sciences for high throughput analysis of genetic information at genome levels (Wilkes *et al.* 2007). Microarray can be used for many areas including ecotoxicogenomics, a new discipline investigating impact of stressors on ecologically relevant organisms and its underlying mechanisms (Iguchi 2006). Like many other environmentally relevant species, the genomes of most earthworm species for instance, *Eisenia andrei,* and *Lumbricus terrestris* have not yet been sequenced. Only a few thousand single pass sequence reads of expressed sequence tags (ESTs) can be found in the public domain.

In the absence of the whole genome sequence, an efficient and economic method is to collect the expressed sequence tags (ESTs) with the purpose of constructing cDNA microarrays, which can be used to screen the transcriptomes process (Figure 1). In the absence of full genome sequences, expressed sequence tags (ESTs) allow rapid identification of expressed genes by sequence analysis and are an important resource for comparative and functional genomic studies. ESTs offer a quick and cost-effective

alternative of gene probe discovery (Adams *et al.* 1991). Selection strategies from high throughput to high accuracy are shown in the Figure 1:

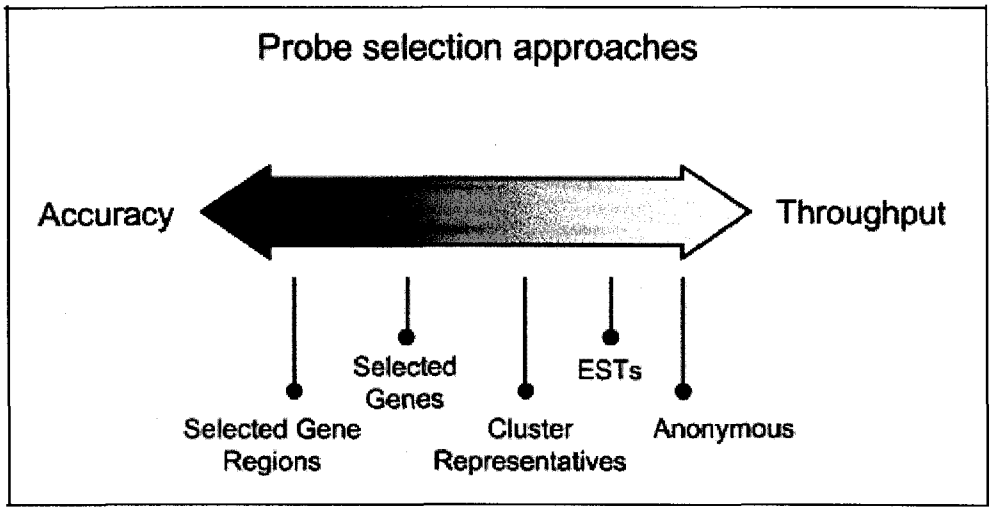


Figure 1. Selection strategies from high throughput to high accuracy - Picture taken from (Nijkamp and Parnham 2005)

EST selection for spotting on microarrays has been approached using three methods: (1) spotting ESTs without sequencing information (no annotation knowledge), (2) spotting only sequenced ESTs with annotations, and (3) selection on gene-oriented clusters. The choice of method typically reflects cost/benefit ratios and the stage of development of the EST collection (Kuster *et al.* 2007). In EST projects, the suppression subtractive hybridization (SSH) technique has often been used to enrich for differentially expressed transcripts independent of their abundance (Diatchenko *et al.* 1996; Liang *et al.* 2003). We also applied the SSH technique to isolate *E. fetida* genes responsive to ORCs exposure. SSH is based on a technique called suppression PCR. It combines normalization and subtraction in a single procedure. In normalization step, the abundance of cDNAs will be equalized within the target population and in the subtraction step; the common sequences between the target and driver populations will be excluded. It provides a 10-100 fold enrichment of differentially expressed mRNAs (Diatchenko *et al.*

1996). The specific objectives microarray study on toxicogenomics often are (1) to identify differentially expressed genes (in this study the earthworm *E. fetida*) as affected by exposure to chemical (TNT), (2) to identify or predict known or unknown toxicological modes of action for chemicals based on the gene expression profile; and (3) to generate new hypothesis of biological pathways involved in response to compound exposure. One ultimate goal is to develop mechanisms-based gene assays as rapid tools for assessing environmental risks associated with explosives exposure.

Experiment Design

Certain decisions of how many microarray slides will be used and which mRNA samples will be hybridized to each slide must be made in the preparation of the mRNA samples before carrying out a microarray experiment. Kerr and Churchill (2001) and Glonek and Solomon (2004) suggest efficient designs for the some of the common microarray experiments. The most commonly used design is the reference design. In this design, each condition of interest is compared with samples taken from a standard reference. This design allows an indirect comparison between the conditions, because the reference is common to all of the arrays. In contrast, a loop design compares two conditions via a chain of other conditions or multiple-pairwise (interwoven loop) fashion (Vinciotti *et al.* 2005). Table 1 shows the combination of varieties with dyes for the reference and loop design. Most studies on microarray design suggest that the loop design of microarray experiments is more efficient than the reference design (Churchill 2002; Glonek and Solomon 2004; Khanin and Wit 2005; Landgrebe *et al.* 2004). However one disadvantage of this method is that ratios observed across different pairwise comparisons are not immediately comparable, and visualization is more difficult (Townsend 2003).

Reference Design

V1	V2	V3	V4	V5
V6	V6	V6	V6	V6

Loop Design

V1	V2	V3	V4	V5
V2	V3	V4	V5	V1

Table 1. Combination of Varieties with Dyes for the Reference vs. Loop Design

Kerr and Churchill (2001) noticed that a loop design stops being optimal when there are more than eight conditions. Therefore it has been suggested that the optimal design could be a form of an interwoven loop design. Figure 2 shows an example of interwoven loop design for an experiment with nine conditions (or time points) and 18 slide arrays (Kerr and Churchill 2001; Vinciotti *et al.* 2005).

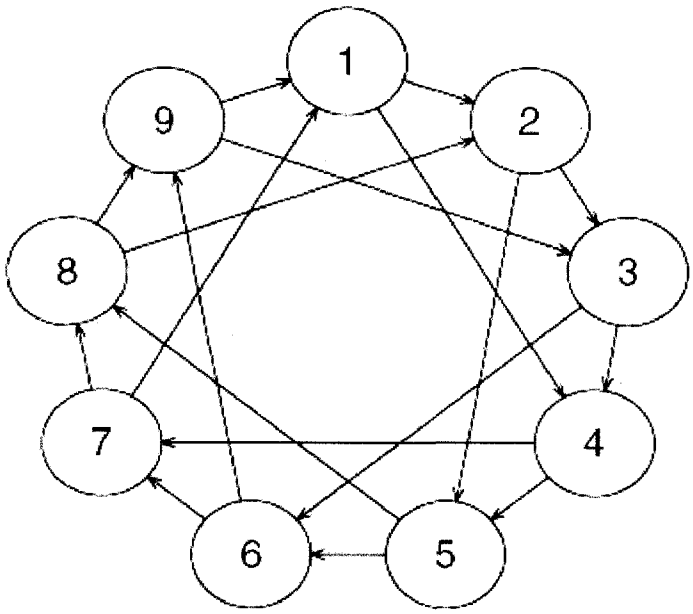


Figure 2. An example interwoven loop design with 18 arrays and 9 conditions

Efficiency of Hybrid Normalization of Microarray Gene Expression

The ability of simultaneous measurement of the expression of thousands of genes in cells has enabled microarrays to be widely used for the study of gene expression in biological research. However, microarrays assess gene expression indirectly by

monitoring fluorescence intensities of labeled target cDNA hybridized to probes on the arrays (Futschik and Crompton 2004). These florescence signals have to be adjusted from errors, which have been generated by microarray experiments, for instance errors are including artifacts and background intensities (Kim *et al.* 2002). At least two types of errors are included, the additive and the multiplicative noises (Sasik *et al.* 2002). Usually, background is considered as one of the additive noises to the signal and the variation between the signal-pixels is the representative multiplicative noise. In a typical spotted slide microarray experiment, the basic strategy is to isolate RNA from two sources, a control and an experimental sample (Kim *et al.* 2002; Zhang *et al.* 2006). mRNA samples are then converted into cDNA by reverse transcription. The transcripts are labeled with red or green fluorescent dyes such as Cy3 and Cy5 respectively and will be hybridized in optimal conditions. Hybridization will be followed by washing steps to minimize the unspecific bindings. Subsequently, after these processes, microarray will scan the hybridized probes at two different wavelengths, 532nm and 635nm for Cy3 and Cy5 respectively to detect the target labeled cDNA in pre made cDNA library (Taniguchi *et al.* 2001). Consequently, two unsigned 16 bit TIFF or BMP image files are generated from a scanner. The level of gene expression in two samples will be compared by measuring the ratio of fluorescent intensity of the two dyes. This technique is very effective in defeating the weak signal of microarray experiments. However, two samples with two different fluorescent dyes introduce dye bias in measurements.

Dye bias is a systematic error that could be removed from microarray data by using a normalization method (Yang *et al.* 2002). Yang *et al*, proposed various effective methods of normalization. One method introduced by the mentioned group is called self-

normalization (paired-slides) or dye-swap. The self-normalization method assumes that the hybridization bias in both slides is roughly the same. The advantage of paired-slide normalization is that it doesn't assume zero average of differential expression. Paired-slide normalization can be applied to dye-swap experiments, which is two hybridizations for two mRNA samples with dye assignment reversed in the second hybridization.

Since in real experiment it is not possible to calculate true intensity of spots, to study efficiency of this method of normalization a software simulation is required to generate both true and observed intensity of simulated dataset (Pirooznia and Deng 2007). True hybridization intensity is calculated then different additive and multiplicative errors are applied and normalized intensity logged ratio computed using background subtraction, self-normalization method and dye-swap technique. Results of this simulation study show the efficiency of hybrid normalization method in order to remove error rates from the intensity values.

Comparing Gene Lists Using Venn Diagrams. Microarray generates vast amounts of data, often in the form of large lists of genes differentially expressed between different sample sets. It leaves the researchers with the task of identifying the functional relevance of the observed expression changes. There are a number of methods to compare results from multiple microarray experiments (Kestler *et al.* 2005). Methods can be used to validate results from similar experiments performed under different conditions. One of the simplest but most effective of these procedures is to examine the overlap of resulting gene lists in a Venn diagram.

The Venn diagram is a graphic technique for visualizing set theory concepts. It uses overlapping circles and shading to indicate intersection, union and complement. It

was introduced in the late 1800s by English logician, John Venn (Venn 1880). Venn diagrams are used to show the mathematical or logical relationships between different groups of sets. A Venn diagram shows all the logical relations between the sets. Venn diagrams can provide much more information to the researcher. Full containment of one set in another, partial intersections and disjunctness can be seen at a glance with Venn diagrams.

Simple Venn diagrams are already being used in microarray data analysis software packages such as commercial GeneSpring® and SilicoCyte® or open source R-package limma to visualize intersections of up to three different lists of genes.

Machine Learning Approach of Microarray - A Comparative Study

Microarray technology allows scientists to monitor the expression of genes on a genomic scale. It increases the possibility of classification and diagnosis at the gene expression level. Many classification methods such as Neural Nets (Cowan and Sharp 1988), Bayesian Networks (Friedman *et al.* 2000; Schulman 1984), Decision Tree (Blower and Cross 2006) and Random Forrest (Diaz-Uriarte and Alvarez de Andres 2006) have been used in recent studies for the identification of differentially expressed genes in microarray data. However there is lack of comparison between these methods to find a better framework for classification, clustering and analysis of microarray gene expression.

Another issue that might affect the outcome of the analysis is that there is such a huge number of genes included in the original data that some of them are irrelevant to analysis. Thus, reducing the number of genes by selecting those that are important is

critical to improve the accuracy and speed of prediction systems. This process also known as feature selection or feature elimination (Jirapech-Umpai and Aitken 2005; Xing *et al.* 2001).

Supervised Classification

Supervised classification, also called prediction or discrimination, involves developing algorithms to priori-defined categories (Larranaga *et al.* 2006). Algorithms are typically developed on a training dataset and then tested on an independent test data set to evaluate the accuracy of algorithms.

Support Vector Machines (SVM). Support vector machines (Vapnik 1998) are a group of related supervised learning methods used for classification and regression (Noble 2006). The simplest type of support vector machines is linear classification (Brown *et al.* 2000). This type tries to draw a straight line that separates data with two dimensions. Many linear classifiers (also called hyperplanes) are able to separate the data (Byvatov and Schneider 2003). However, only one achieves maximum separation. A special feature of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin. Therefore they are also known as maximum margin classifiers (Pavlidis *et al.* 2004). Maximum-margin hyperplanes for a SVM trained with samples from two classes. Samples alongside the hyperplanes are called the support vectors.

Vapnik in 1963 proposed a linear classifier as an original optimal hyperplane algorithm (Vapnik 1998). The replacement of dot product by a non-linear kernel function allows the algorithm to fit the maximum-margin hyperplane in the transformed feature space. SVM finds a linear separating hyperplane with the maximal margin in this higher

dimensional space. $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ is called the kernel function. There are four basic kernels: linear, polynomial, radial basic function (RBF), and sigmoid (Vapnik 1998). Table 2 illustrates a formulation comparison of these four kernel types.

Kernel Type	$K(x_i, x_j)$
Linear	$x_i^T x_j$
Polynomial	$(x_i, x_j)^d$
RBF	$\exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$
Sigmoid	$\tanh(k(x_i x_j) + \vartheta)$

Table 2. Formulation of four basic kernels function

Decision Trees. In tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. There are advantages with decision tree algorithms: they are easily converted to a set of production rules, they can classify both categorical and numerical data, and there is no need to have a priori assumptions about the nature of the data. However multiple output attributes are not allowed in decision tree and algorithms are unstable. Slight variations in the training data can result it different attribute selections at each choice point within the tree. The effect can be significant since attribute choices affect all descendent subtrees (Wang *et al.* 2005). ID3 (Iterative Dichotomiser 3) is an algorithm used to generate a decision tree. Developed by J. Ross Quinlan (Quinlan 1993), ID3 is based on the Concept Learning System (CLS) algorithm. ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the n (where n is the number of possible values of an attribute) partitioned subsets to get their best attribute (Kinney and Murphy 1987). J48 is an improved version of ID3 algorithm. It

contains several improvements, including: choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs, and handling continuous attributes (Quinlan 1993).

Artificial Neural Networks (ANN). ANN is an interconnected group of nodes that uses a computational model for information processing. It changes its structure based on external or internal information that flows through the network (Chen *et al.* 2006). ANN can be used to model a complex relationship between inputs and outputs and find patterns in data.

The function $f(x)$ in ANN is defined as a composition of other function; $g(x)$, which itself can be as a composition of other functions (Anthony and Bartlett 1999). The dependencies between variables in a network can be viewed in two ways (Greer and Khan 2004):

1) Functional view: the input transformed to a less dimensional layer in each layer until it is finally transformed to a one dimensional output. This view is commonly used in optimization processes. These networks are commonly called feedforward.
2) Probabilistic view: all functions are dependent on randomly selected variables. This view frequently comes across in the context of graphical models. Networks with cycles are called recurrent (Narayanan *et al.* 2002).

Error backpropagation neural network is a feedforward multilayer perceptron (MLP) that is applied in many fields due to its powerful and stable learning algorithm (Ahmed 2005). The neural network learns the training examples by adjusting the synaptic weight according to the error occurred on the output layer. The back-propagation

algorithm has two main advantages: local for updating the synaptic weights and biases, and efficient for computing all the partial derivatives of the cost function with respect to these free parameters. A perceptron is a simple pattern classifier (Maclin *et al.* 1991).

RBF networks have two steps of processing. First, input is mapped in the hidden layer. The output layer is then a linear combination of hidden layer values representing mean predicted output. This output layer value is the same as a regression model in statistics (Casasent and Chen 2003). The output layer, in classification problems, is usually a sigmoid function of a linear combination of hidden layer values. Performance in both cases is often improved by shrinkage techniques, also known as ridge regression in classical statistics and therefore smooth output functions in a Bayesian network (Moody and Darken 1989).

Bayesian Networks. A bayesian network represents independencies over a set of variables in a given joint probability distribution (JPD). Nodes correspond to variables of interest, and arcs between two nodes represent statistical dependence between variables. Bayesian refers to Bayes' theorem on conditional probability (Schulman 1984). Bayes' theorem is a result in probability theory, which relates the conditional and marginal probability distributions of random variables. The probability of an event A conditional on another event B is in general different from the probability of B conditional on A. However, there is an explicit relationship between the two, and Bayes' theorem is the statement of that relationship (Dojer *et al.* 2006).

Naive Bayes is a rule generator based on Bayes's rule of conditional probability. It uses all attributes and allows them to make contributions to the decision as if they were

all equally important and independent of one another, with the probability denoted by the equation:

$$P(H \mid E) = \frac{P(E_1 \mid H).P(E_2 \mid H).....P(E_n \mid H)}{P(E)}$$

Where *P(A)* denotes the probability of event *A*, *P(A|B)* denotes the probability of event *A* conditional on event *B*, E_n is the n^{th} attribute of the instance, *H* is the outcome in question, and *E* is the combination of all the attribute values (Friedman *et al.* 2000; Langseth and Nielsen 2006).

Unsupervised Clustering

Cluster-analysis algorithms group objects on the basis of some sort of similarity metric that is computed for features. Genes can be grouped into classes on the basis of the similarity in their expression profiles across tissues, cases or conditions. Clustering methods divide the objects into a predetermined number of groups in a manner that maximizes a specific function. Cluster analysis always produces clustering, but whether a pattern is observed in the sample data remains an open question and should be answered by methods such as resampling-based methods.

K-Means Clustering. The k-means algorithm takes a dataset and partitions it into *k* clusters, a user-defined value. Computationally, one may think of this method as a reverse method of analysis of variance (ANOVA). The algorithm starts with *k* random clusters, and then move objects between those clusters with the goal to 1) minimize variability within clusters and 2) maximize variability between clusters (MacQueen 1967). In other words, the similarity rules will apply maximally to the members of one cluster and minimally to members belonging to the rest of the clusters. The significance

test in ANOVA evaluates the between group variability against the within-group variability when computing the significance test for the hypothesis that the means in the groups are different from each other. Usually, as the result of a *k*-means clustering analysis, the means for each cluster on each dimension would be examined to assess how distinct *k* clusters are. Obtaining very different means for most is perfect (Sun *et al.* 2006).

Expectation Maximization (EM) Clustering. An expectation-maximization (EM) algorithm finds maximum likelihood estimates of parameters in probabilistic models. EM performs repeatedly between an expectation (E) step, an expectation of the likelihood of the observed variables, and maximization (M) step, which computes the maximum expected likelihood found on the E step. EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters (Frank *et al.* 2004). By cross validation, EM can decide how many clusters to create.

The goal of EM clustering is to estimate the means and standard deviations for each cluster so as to maximize the likelihood of the observed data. The results of EM clustering are different from those computed by k-means clustering (Dempster *et al.* 1977). K-means assigns observations to clusters to maximize the distances between clusters. The EM algorithm computes classification probabilities, not actual assignments of observations to clusters.

Feature Selection

Feature selection methods can be divided into the *wrapper* model and the *filter* model (Kohavi and John 1997). The wrapper model uses the predictive accuracy of a mining algorithm to determine the goodness of a selected subset. Wrapper methods

generally result in better performance than filter methods because the latter suffers from the potential drawback that the feature selection principle and the classification step do not necessarily optimize the same objective function (Jirapech-Umpai and Aitken 2005). In gene selection, the filter model is often adopted due to its computational efficiency (Xing *et al.* 2001). Filter methods select a predictive subset of the features using heuristics based on characteristics of the data. Moreover, in the wrapper method, the repeated application of cross validation on the same data set might result in finding a feature subset that performs well on the validation data alone. Filter methods are much faster than wrapper methods and therefore are better suited to high dimensional data sets (John *et al.* 1994).

SVM-RFE. SVM-RFE (Guyon *et al.* 2002) is a feature selection method to filter out the optimum feature set by using SVM in a wrapper-style. It selects or omits dimensions of the data, depending on a performance measurement of the SVM classifier. One of the advantages of SVM-RFE is that it is much more robust to data overfitting than other methods.

This is an algorithm for selecting a subset of features for a particular learning task. The basic algorithm is the following: 1) initialize the data set to contain all features, 2) train an SVM on the data set, 3) rank features according to $c_i = (w_i)^2$, 4) eliminate the lower-ranked 50% of the features, 5) return to step 2. At each RFE step 4, a number of genes are discarded from the active variables of an SVM classification model. The features are eliminated according to a criterion related to their support for the discrimination function, and the SVM is re-trained at each step (Guyon *et al.* 2002).

Correlation based (CFS). In CFS features can be classified into three disjoint categories: namely, strongly relevant, weakly relevant and irrelevant features (Wang *et al.* 2005). Strong relevance of a feature indicates that the feature is always necessary for an optimal subset; it cannot be removed without affecting the original conditional class distribution. Weak relevance suggests that the feature is not always necessary but may become necessary for an optimal subset at certain conditions. Irrelevance indicates that the feature is not necessary at all. There are two types of measures for correlation between genes: linear and non-linear (Xing *et al.* 2001). Linear correlation may not be able to capture correlations that are not linear. Therefore non-linear correlation measures are often adopted for measurement. It is based on the information-theoretical concept of *entropy*, a measure of the uncertainty of a random variable (John *et al.* 1994).

Chi Squared. Another commonly used feature selection method is Chi-square statistic (χ^2) method (Liu and Setiono 1995). This method evaluates each gene individually by measuring the Chi-square statistics with respect to the classes. The gene expression numbers are first discretized into several intervals using an entropy-based discretization method. Then the Chi-square value of each gene is computed by

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(A_{ij} - \frac{R_i . C_j}{N}\right)^2}{\frac{R_i . C_j}{N}}$$

where m denotes the number of intervals, k the counts of classes, N the total number of patterns, R_i the number of patterns in the i^{th} interval, C_j the number of patterns in the j^{th} class, and A_{ij} the number of patterns in the i^{th} interval, j^{th} class. The genes with larger Chi-square statistic values are then selected as marker genes for classification.

Cross Validation

In order to perform to measure classification error, it is necessary to have test data samples independent of the learning dataset that was used to build a classifier. However, obtaining independent test data is difficult or expensive, and it is undesirable to hold back data from the learning dataset to use for a separate test because that weakens the learning dataset. V-fold cross validation technique performs independent tests without requiring separate test datasets and without reducing the data used to build the tree. The learning dataset is partitioned into some number of groups called "folds" (Chang and Lin 2001).

The number of groups that the rows are partitioned into is the 'V' in *V-fold cross classification.* Ten is the recommended and default number for "V". It is also possible to apply the *v-fold cross-validation* method to a range of numbers of clusters in *k*-means or *EM* clustering, and observe the resulting average distance of the observations from their cluster centers. Leave-one-out cross-validation involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data (Chang and Lin 2001).

CHAPTER II

MATERIALS AND METHODS

cDNA library and Expression Sequence Tag Sequence

Earthworm cDNA library construction[2]

Two earthworm cDNA libraries were constructed using SSH-PCR (Diatchenko *et al.* 1996). The first SSH library (Figure 3A) was made using pooled mRNA (10 μg) extracted from control unexposed worms against worms exposed to Cd (2.6 mmol/kg or 292 mg/kg), TNT (100 mg/kg), 2,6-DNT (54 mg/kg), RDX (50 mg/kg), or HMX (10 mg/kg). For the construction of the second library (Figure 3B), mRNA (10 μg) from worms exposed to Cu (293 mg/kg), Pb (8778 mg/kg), Zn (357 mg/kg), 2,4-DNT (100 mg/kg), and TNB (100 mg/kg) was run against mRNA from another set of control worms. Exposures (4-, 14-, or 28-d) were conducted in an Organization for Economic Cooperation and Development (OECD) artificial soil consisting of 70% sand, 20% kaolin clay, and 10% 2-mm sieved peat moss with an adjusted pH between 6.5 and 7.0. Chemical concentrations were selected at effective concentrations for 50% (EC_{50}) reduction in fecundity on the basis of our previous studies as well as published literature.

Exposed and unexposed earthworms were fixed in RNAlater (Ambion, Austin, TX) and stored at -80°C. Total RNA was extracted using RNeasy kits (Qiagen, Valencia, CA), and poly(A) mRNA was separated from total RNA using NucleoTrap mRNA purification kit (BD Biosciences, San Jose, CA).

---

[2] The cDNA libraries construction, and other laboratory works are performed by ERDC (Environmental Research and Development Center) at Vicsburg, MS

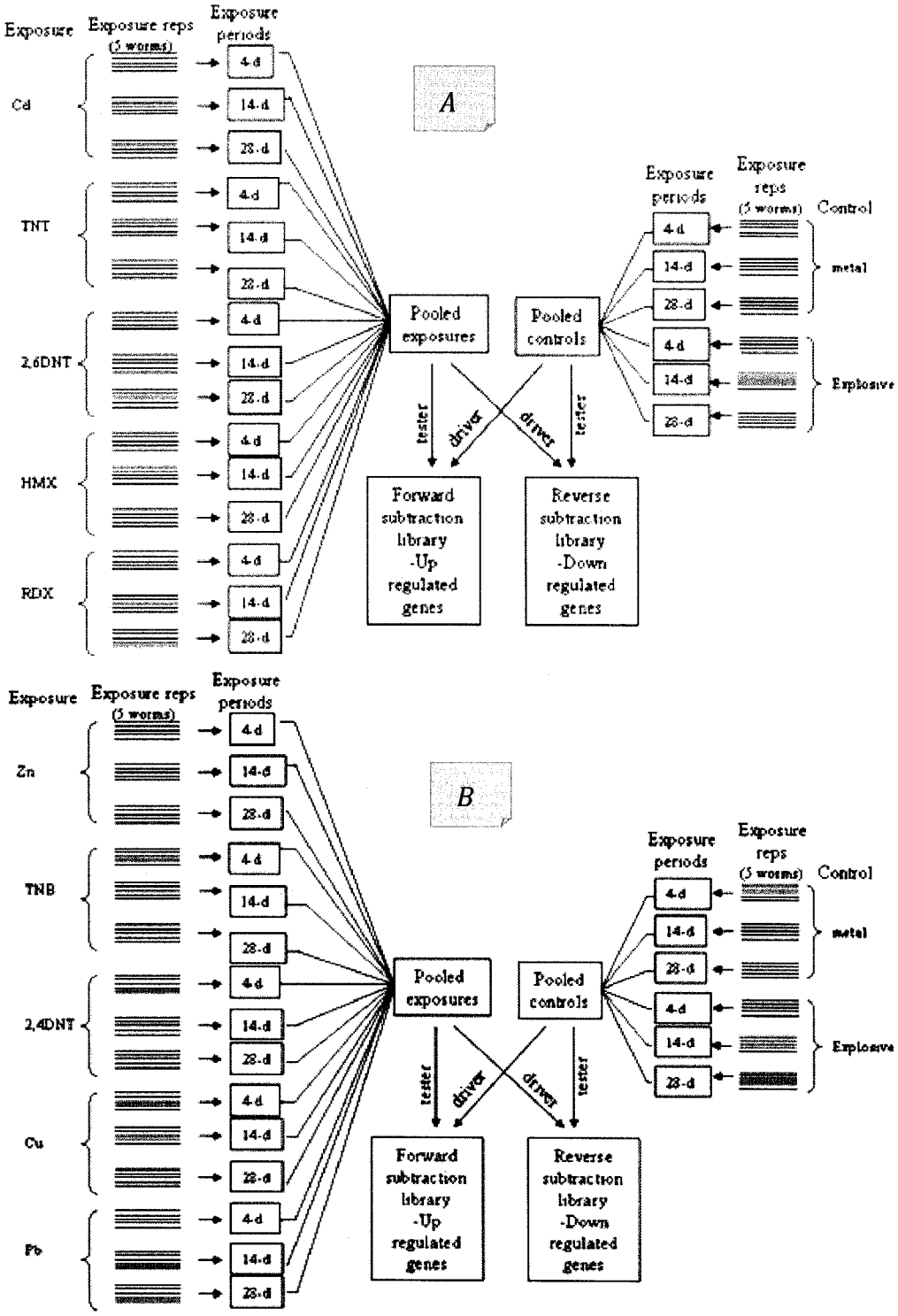
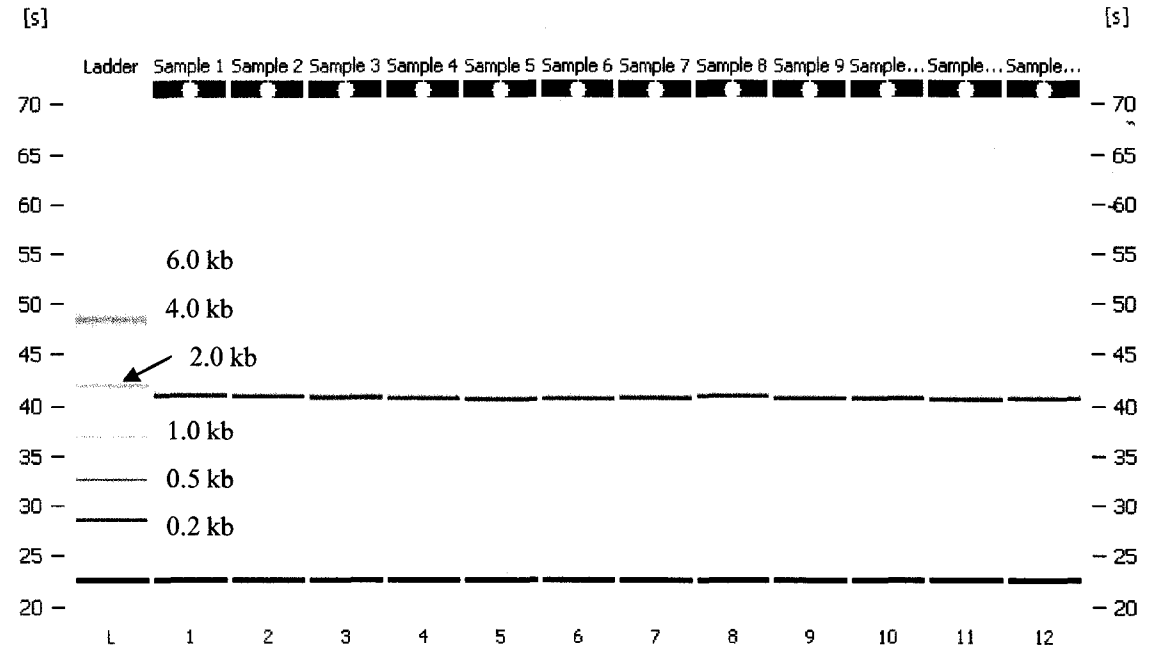


Figure 3. Scheme of RNA sample pooling for subtractive suppression hybridization cDNA library construction. 3A: the first library; 3B: the second library.

The integrity and concentration of mRNA were checked on an Agilent 2100 Bioanalyzer (Palo Alto, CA). The gel-like images generated by the Bioanalyzer show that both RNAs have only one bright band close to the 2 kb ladder band (Figure 4A&B), which is distinctive from the two bands seen with 18S and 26S RNA of mammalian RNA. A Clontech PCR-Select™ cDNA subtraction kit (BD Biosciences) was then used to enrich for differentially expressed genes (Figure 5).

4A



4B

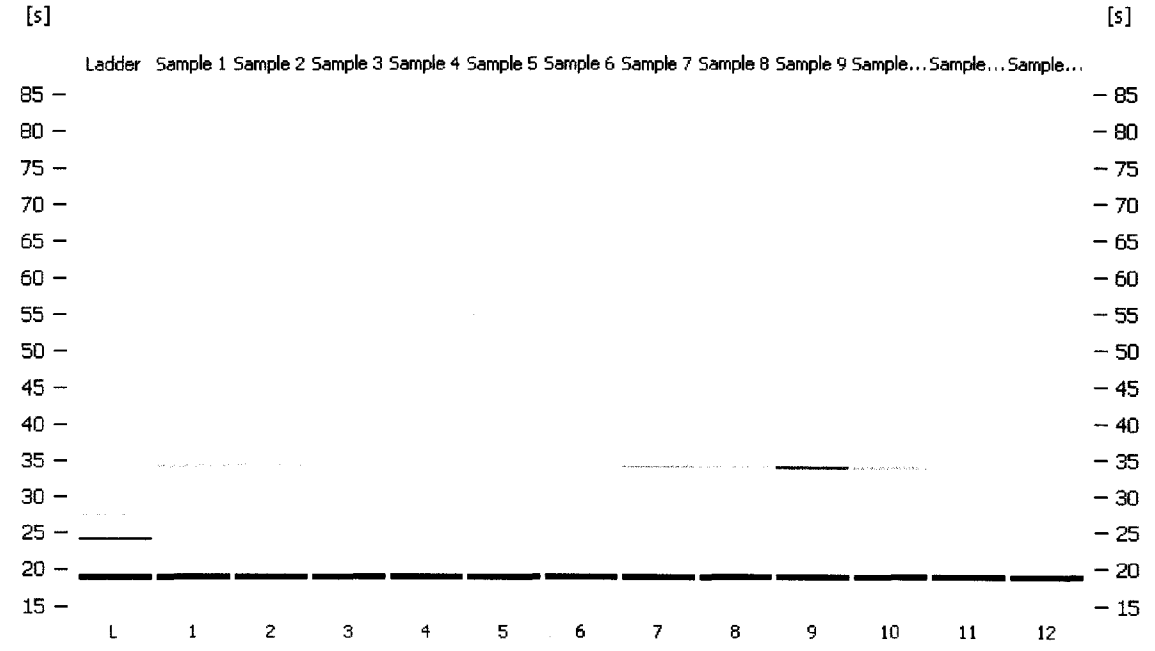


Figure 4. Earthworm total RNA (4A) and purified mRNA (4B) electrophoresis using Agilent 2100 Bioanalyzer

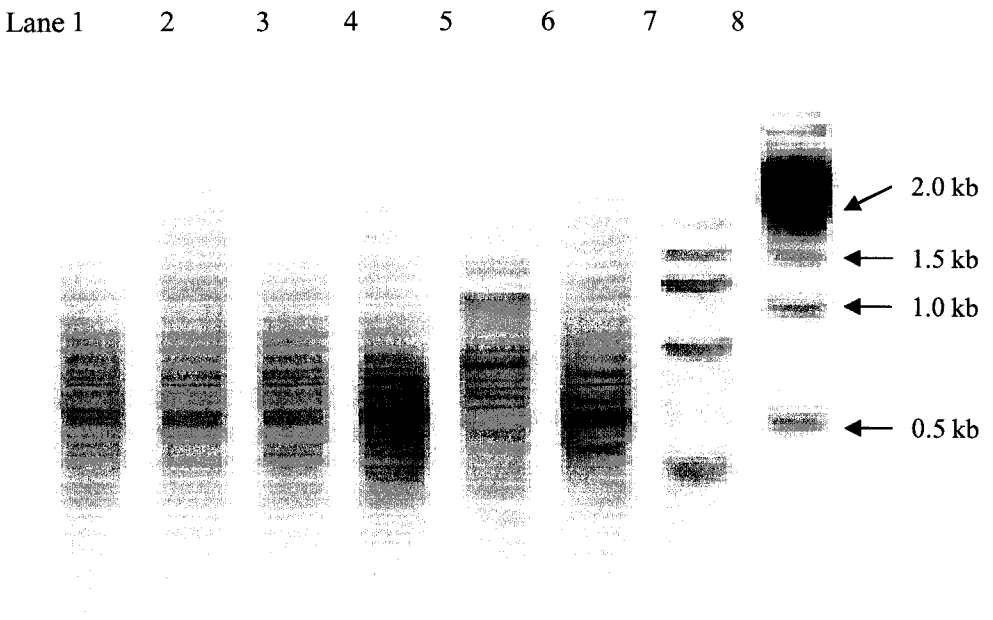


Figure 5. Subtracted and non-subtracted cDNAs electrophoresed on a 2% agarose/SybrGreen gel in 1X sodium borate buffer. Lane 1: forward subtracted earthworm (EW) cDNA; Lane 2: forward non-subtracted EW cDNA; Lane 3: reverse subtracted EW cDNA; Lane 4: reverse non-subtracted EW cDNA; Lane 5: subtracted human skeleton muscle (HSM) cDNA; Lane 6: non-subtracted HSM cDNA; Lane 7: control subtracted human skeleton muscle cDNA; Lane 8: 1kb DNA ladder.

EST Cloning and Sequencing

After the secondary PCR amplification, both forward and reverse subtracted PCR products of the two libraries were cloned using pCR2.1 or pCR4.0 vectors and Mach1-T1 chemically competent cells (Invitrogen, Carlsbad, CA). Positive colonies were picked and grown overnight at 37°C in LB media containing 50 μg/mL ampicillin in a 96-deep well block format. Half of the clone culture (300 μl) was archived with 300 μl of 60% glycerol and stored at -80°C. Two μl of the remaining clone culture was amplified in a 100-μl PCR reaction. After amplification, 8 μl of the PCR reaction was checked on a 96-well electrophoresis gel (2% agarose) for inserts of 100-2000 bps. Amplicons (cDNA inserts) were purified using Millipore Montage PCR 96 Cleanup Kit (Billerica, MA). We

checked the concentration of randomly selected purified cDNA using PicoGreen (Molecular Probes, Eugene, OR), which ranged from 100-500 ng/μl with an average of 240 ng/μl. Four μl of the purified cDNA (55 μl in total) was sequenced using BigDye® Terminator v3.1 and a 16-capillary ABI PRISM® 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA) according to manufacturer's instruction.

EST Data Processing

Many software programs are available that provide sequence cleansing and assembly. These include commercial software such as Sequencher (Gene Codes, Ann Arbor, Michigan, USA), and Aligner (CodonCode, Dedham, MA, USA), and open source software such as CAP and TIGR Assembler. With these software packages it is possible to quickly remove vector sequences from each EST clone and screen the ESTs for low-quality sequences. The high-quality and trimmed EST sequences then can be used to find overlap assembly of contiguous sequences. Sequence information was stored in ABI chromatograph trace files, and Phred was used to perform base-calling (Ewing and Green 1998). Phred read DNA trace data, called bases, assigned quality values to the bases, and wrote the base calls and quality values to output sequence files in either FASTA or SCF format. Quality values for the bases were later used by the sequence assembly program, Phrap (Nickerson *et al.* 1997), to increase the accuracy of assembled sequences. Phred uses simple Fourier methods to examine the four base traces in the data set to predict a series of evenly spaced locations. It determines where the true peak location would be if there were no compressions, dropouts, or other factors shifting the peaks from their locations. Then Phred examines each trace to find the centers of the observed peaks and the areas of these peaks relative to their neighbors. A dynamic programming algorithm is

used to match the observed peaks detected in the second step with the predicted peak locations found in the first step. It uses a quality value lookup table to assign the corresponding quality value. The quality value is related to the base call error probability by the formula $QV = -10 \times \log_{10}(P_e)$ where P_e is the probability that the base call is an error (Ewing and Green 1998).

Typically, sequence chromatograms have low-quality regions at the beginning and the end of each sequence read (Chou and Holmes 2001). One can automatically remove the low-quality ends when quality values are available. This process is called "end clipping" or "end trimming". There are two different end clipping methods (Chou and Holmes 2001): (1) maximizing regions with error rates below a given threshold, and (2) using separate criteria at the start and the end of the sequence. We chose the former method which was implemented in CodonCode Aligner[3] to remove low quality bases at both ends by setting quality score $QV \geq 20$ (or $P_e \leq 0.01$). Flanking vector/adaptor sequences should also be trimmed off because they can lead to incorrect assemblies or alignment. We input a custom-made vector/adaptor file into the Aligner to trim vector/adaptor sequences. Furthermore, we used VecScreen[4] to detect and then manually removed any residual and partial vector contamination in our ESTs.

Phrap[5] was used to assemble sequence fragments into a larger sequence by identifying overlaps between sample sequences. Samples that can be joined together are put into "contigs". The following greedy algorithm is used in Phrap. First, it finds

[3] http://www.codoncode.com/aligner/

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

[5] Phred/Phrap/Consed; Shotgun sequence assembly [http://www.phrap.org]

potential overlaps between samples by looking for shared 12-nucleotide "words" in the sequence. Then the pair of samples with the highest number of shared words is found. If the alignment is good enough, it would be kept as a new contig, and the consensus sequence would be calculated; otherwise, the alignment would be rejected, and the two samples would be left separated. Four criteria were used to determine whether to accept or reject an alignment: (1) minimum percent identity (the minimum percentage of identical bases in the aligned region) $\geq$ 70%; (2) minimum overlap length $\geq$ 25 bps, (3) minimum alignment score which is similar to (2) but takes any mismatches into account, $\geq$ 20 bps; and (4) maximum gap size $\leq$ 15 bps. Overall, these criteria were relatively relaxed if compared to more stringent settings such as 90% for minimum percentage identity or minimum overlap length $\geq$ 35 bps. If one sample has an insertion/deletion that is larger than 15 bps, the alignment will typically stop there, and the rest of the sample will be considered unaligned. The alignment process would then be repeated. If a sample is in a contig, the consensus sequence is then used for the contig. If the two samples are already in the same contig, the next pair is retrieved and analyzed. It repeats and continues the pairwise joins until all possible joins have been tried, or until the maximum number of merge failures in a row has occurred.

After assembly, all contigs with more than three ESTs were assessed for missassemblies using the assembly viewer Consed (Gordon *et al.* 1998). Contigs flagged for possible missassemblies were manually edited using Consed tools to remove potential chimeric ESTs or other suspect ESTs. Chimerism occurs because of multiple insert cloning or mistracking of sequence gel lanes. After assembly with Phrap, contigs with more than three ESTs were examined again in Consed to eliminate additional

missassemblies not resolved by Phrap. Any bps with a calculated quality value below 12 were changed to an N (unknown base) which were considered as suspect ESTs.

EST Comparative Analysis and Functional Assignment

Comparative analysis was performed using blastx through NCBI with the unique sequences (including the consensus sequences of assembled contigs and the singletons). Blastx searches were conducted on our local BLAST server against the NCBI's non-redundant peptide sequence database. The returned search results (100 best hits) were transferred automatically into a relational database. We discarded hits with an *E*-value > 10^{-5} and sorted out the remaining hits by organism name. To assign putative functions to the unique *E. fetida* sequences, we extracted the GO hierarchical terms of their homologous genes from the protein databases of the following four model organisms: *Mus musculus, Drosophila melanogaster, Caenorhabditis elegans,* and *Saccharomyces cerevisiae* (Gene Ontology Consortium 2001; Ashburner *et al.* 2000; Harris *et al.* 2004). Meanwhile, we also mapped the unique sequences to metabolic pathways in accordance with the KEGG (Ogata *et al.* 1999). Enzyme commission (EC) numbers were acquired for the unique sequences by blastx searching (*E*-value $\leq 10^{-5}$) the SWIR database, which is made up from three protein databases WormPep, SwissProt and Trembl. The EC numbers were then used to putatively map unique sequences to specific biochemical pathways (Deng *et al.* 2006b; McCarter *et al.* 2003). All the matched GO and pathway information was automatically stored in our local relational database. Figure 6 illustrates our pipeline for EST Cleansing and Assembly process.

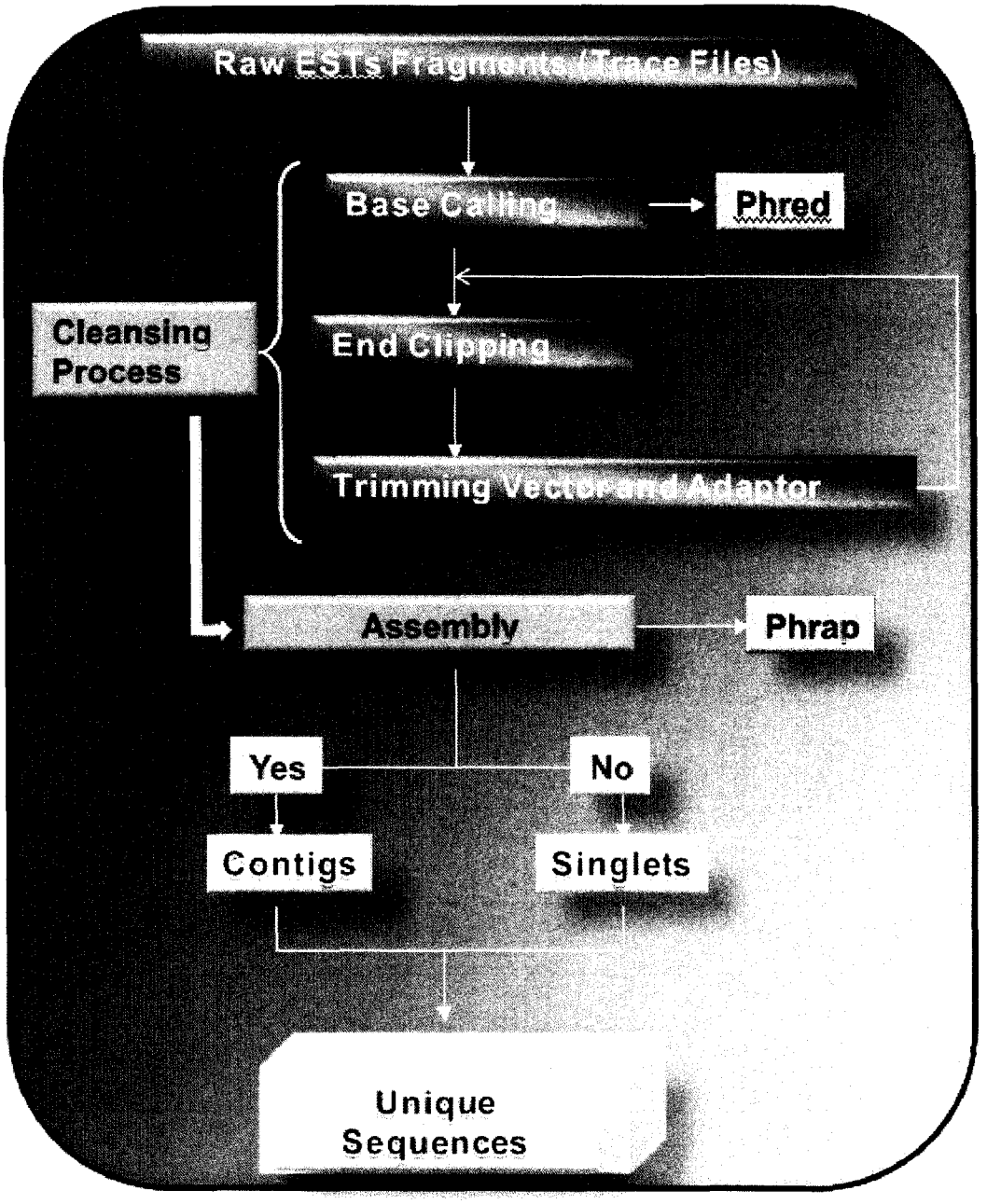


Figure 6. A Pipeline for Expressed Sequence Tag Cleansing and Assembly Process

ESTMD - EST Database Implementation and Web Application

To facilitate efficient management and retrieval of the EST information obtained from this project, we upgraded our previous developed EST model database (ESTMD version 1) (Deng *et al.* 2006a) and integrated the earthworm EST information into the new version of ESTMD. ESTMD is an integrated Web-based application consisting of client, server and backend database. The current implementation of ESTMD (version 2) has many new features. The main changes include further normalization of tables from 50 tables to 17, altering main tables to be capable of storing the information of multiple organisms, adding a new table (contigview) to store contigs' view information, using a 2D Java class for displaying contigs instead of a Perl script, and implementing the whole web application as a unified portable web module.

ESTMD is currently hosted on Suse Linux 10 and can be implemented in MySQL 4.0 or higher version. It has an integrated web-based application with a three-tier structure including client, server and backend database (Figure 6). The web-based interface of the database was created using HTML and JavaScript to evaluate the validation of the input on the client side and to reduce the burden on the server side. Apache 2.2 is used as the HTTP web server, while Tomcat 5.5 is the Servlets container. Both of these programs were developed and maintained on Linux and WinNT, ensuring that the database is transplantable and platform-independent. The server-side programs are implemented by Java 2 Enterprise Edition (J2EE) technologies. Servlet and JSP (JavaServer Pages) are used to communicate between users and databases and to implement a query.

Toxicogenomics Study of Earthworm

Array printing[6]

A total of 4032 purified cDNA clones were loaded on 384-well plates and dried down completely in a Vacufuge™ Concentrator 5301 (Eppendorf, Westbury, NY). The dried cDNA was re-suspended in 15 µl of 1× printing buffer (ArrayIt, Sunnyvale, CA). Each clone was spotted twice (i.e., in two super grids) on Ultra GAPS™ amino silane coated glass slides (Corning, Acton, MA) using 16 pins on a VersArray ChipWriter (Bio-Rad, Hercules, CA). Five alien cDNAs, i.e., PCR product 1 to 5 selected from SpotReport® Alien® cDNA Array Validation System (Stratagene, La Jolla, CA) prepared at 15, 30, 60, 125 and 250 ng/µl, were also spotted twice along with printing buffer and water as control spots. The total number of spots on each array was 8704 including 60 alien cDNA spots, 84 water spots, 256 blank spots and 240 printing buffer spots. After printing, arrays were incubated in a dessicator for two to three days and were then snap-dried on a hot plate after being rehydrated over a boiling water bath. The arrays were further immobilized using a UV Cross-linker (Stratagene) by applying 300 mJ per 10 arrays.

Earthworm toxicity test

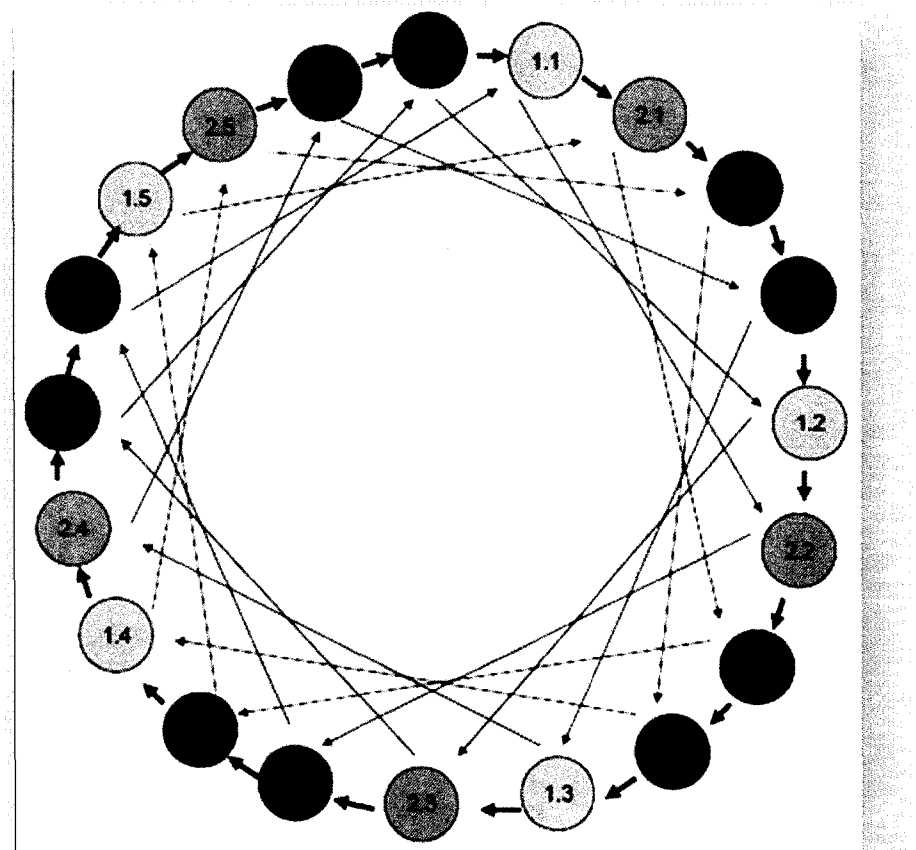
The earthworm reproduction toxicity test was conducted using a field soil in an environmental chamber with continuous lighting and temperature maintained at 21±1°C in accordance with the ASTM guideline (ASTM (American Society for Testing and Materials)). It had the following properties: pH 6.7, total organic C 0.7%, CEC 10.8 mEq/100 g. Appropriate amounts of TNT dissolved in acetone were spiked into air-dried

[6] Performed by ERDC (Environmental Research and Development Center) at Vicsburg, MS

soil to achieve the following nominal concentrations: 0 (solvent control), 2, 4, 7, 12, 22, 35, 55, 88, 139, and 220 mg/kg soil. Five mature worms were added in a jar containing 250 g (dry weight equivalent) of TNT-amended or non-amended soil. Each treatment had five replicate jars. Adult worms were counted, weighed and removed from the jar after 28-d exposure. Cocoon (both hatched and unhatched) and juvenile counts were conducted at day 56. At day 28, one of the worms removed from each jar was fixed in RNAlater (Ambion) to preserve RNA quality and integrity. Each worm was chopped into 8-10 pieces. The fixed samples were stored at -80°C. The rest of the worms were snap-frozen and stored at -20°C for enzymatic assays and other future uses.

Hybridization and array scanning

A total of 40 arrays were hybridized with the 20 cDNA probes in accordance with an interwoven loop scheme as shown in Figure 7 (Churchill 2002). Each biological replicate of cDNA samples was hybridized four times on four different arrays with twice labeled with the Cy3 and twice with the A647 fluorescence dye. After hybridization, arrays were scanned at 5-μm resolution using VersArray ChipReader (Bio-Rad).



Worm	Treatment (dose)	
Replicate	TNT 35ppm	7ppm
1	1.1	2.1
2	1.2	2.2
3	1.3	2.3
4	1.4	2.4
5	1.5	2.5

Figure 7. An interwoven loop hybridization schemes for 4 treatments with 5 independent biological replicates. Circles represent treatment samples. Sample code: 0.x = replicate x of solvent control worms; 1.x = replicate x of 10.6 mg TNT/kg soil treated worms; 2.x = replicate x of 2 mg TNT/kg soil treated worms; 3.x = replicate x of 38.7 mg TNT/kg soil treated worms; x = 1-5. Arrows represent array hybridizations between respective samples where the arrowhead indicates Alexa 647 dye labeling and the base of arrows indicate Cy3 dye labeling.

Overview of Data Analysis

Figure 8 illustrates an overview of the data analysis pipeline to find differentially expressed genes. There can be several filtering steps. When there are more than two conditions in the experiment, the data can be analyzed using two conditions routes.

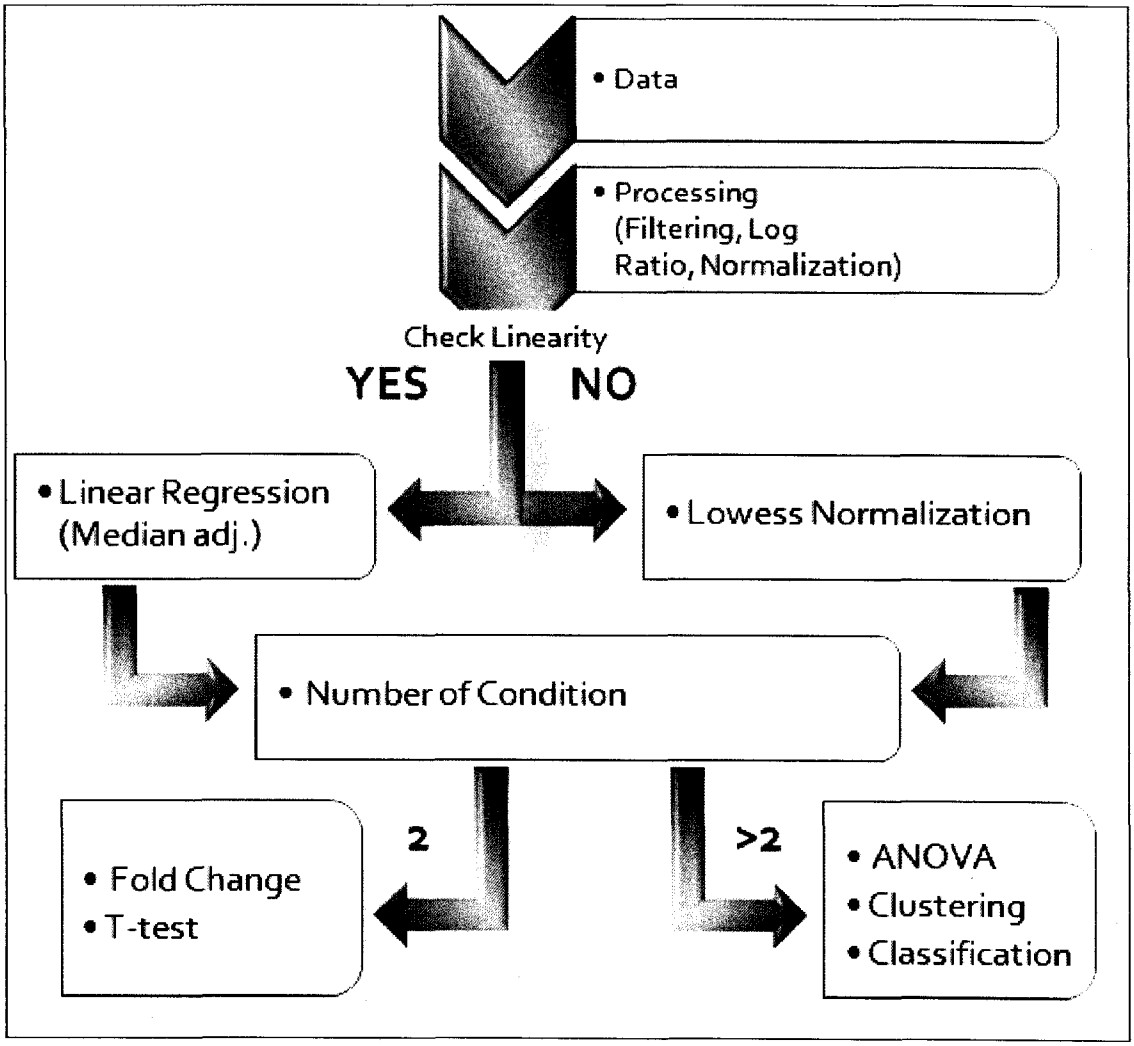


Figure 8. Overview of data analysis methods to find differentially expressed genes

Finding differentially expressed genes. One of the core goals of microarray data analysis is to identify which genes show good evidence of being differentially expressed. This goal can be accomplished in two parts. The first is rank them based on their distribution (log ratio

distribution). This is called fold change. The second is to choose a critical value, such as p-value in t-test or ANOVA, for the ranking significant statistics.

Fold change. Considering the foreground red and green intensities as *Rf* and *Gf* for each spot and the background intensities *Rb* and *Gb,* the background-corrected intensities will be *R* and *G* where *R* = *Rf-Rb* and *G* = *Gf-Gb*. M and A can be calculated as

$$M = \log R/G \quad and \quad A = \tfrac{1}{2} \log RG$$

It is convenient to use base 2 logarithms for *M* and *A* so that *M* is units of 2-fold change. On this scale, *M* = *0* represents equal expression, *M* = *1* represents a 2-fold change between the RNA samples, *M* = *2* represents a 4-fold change, and so on.

t-test. Briefly, the t-test looks at the mean and variance of the two distributions (e.g. control and treatment chip log ratios), and calculates the probability that they were sampled from the same distribution, $t = \frac{\overline{M}}{s / \sqrt{n}}$ where *s* is the standard deviation of the *M* values across the *n* replicates.

SAM. Tusher *et al,* (2001) have used penalized t-statistics of the form $t = \frac{\overline{M}}{(\alpha + s) / \sqrt{n}}$ where the penalty *a* is estimated from the mean and standard deviation of the sample variances.

Microarray data analysis

Raw gene expression data were acquired as spot and background signal intensity (mean and standard deviation) by processing scanned array images on VersArray Analyzer Software v. 4.5 (Bio-Rad). A spot was flagged out if (1) its raw signal intensity was below its background level, (2) it overlapped with other spots, or (3) it was stained or over-saturated. The filtered data was normalized by (1) subtraction of background intensity, (2) cross-channel LOWESS (local regression), and (3) centering to each

channel's median spot intensity. The effect of data normalization and transformation was reviewed graphically in M-A plots. Data points are distributed symmetrically about zero at all intensity values. Control spots including alien cDNAs, water, printing buffer and blank spots behaved as we expected. The spot intensity ratios (Cy3/A647) of alien cDNAs change with the concentration ratios of the spotting alien cDNA and the spiked alien mRNAs. This practice assures the quality of dynamic reverse transcription and hybridization.

Two statistics programs based on different algorithms were employed to identify significantly changed genes. The log ratios were analyzed using 2 Fold Change and t-test, and then compared with the normalized signal intensity data analyzed by Significance Analysis of Microarrays (SAM) (Tusher, Tibshirani, and Chu 5116-21).

Reverse-transcription quantitative PCR (RT-QPCR)[7]

Two-stage RT-QPCR was performed on selected transcripts to further confirm their relative expression in TNT-treated versus control worms. The same mRNA samples used for microarray hybridization were first reverse transcribed into cDNA in a 20-µl reaction containing 100 ng mRNA, random primers and SuperScript™ III reverse transcriptase (Invitrogen) following the manufacturers's instructions. The synthesized cDNA was diluted to 10 ng/µl. Each 20-µl reaction was run in triplicate and contained 6 µl of synthesized cDNA templates along with 900 nM primers and 500 nM Sybr Green PCR Master Mix (ABI). Cycling parameters were 95°C for 15 minutes to activate the DNA polymerase, then 40 cycles of 95°C for 15 seconds and 60°C for 1 minute. Melting dissociation curves were performed to verify that single products without primer-dimers

[7] Performed at ERDC (Environmental Research and Development Center) at Vicsburg, MS

were amplified. The raw fluorescence data were exported as clipped files. The starting concentrations of mRNAs and PCR efficiencies for each sample were calculated by an assumption-free linear regression on the Log(fluorescence) per cycle number data using LinRegPCR (Ramakers *et al.* 2003).

Efficiency of hybrid normalization of microarray gene expression: A simulation study

A simulation of microarray experiment was conducted to investigate the efficiency of hybrid normalization technique. The simulation was performed by generating a collection of fragments, evolving them into the mutated fragments, and hybridizing them together. Then considering start and mutated fragments as red and green spots of microarray experiment, the true intensity logged ratio of hybridization was calculated. Error values were added to the experiment and removed using dye-flip normalization technique in order to investigate the efficiency of this technique.

Markov process model design

The Markov process model can be thought of as a model that generates sequences of nucleotide, with a definite probability distribution. Since the total probability of all bases in the distribution must sum to one, the probability of one cannot increase without decreasing in another (Eddy 1998). The Broken Stick model (Macarthur 1957) is a well known model that leads us to random division of a fixed interval. This model can be used to generate random fragment lengths. According to the broken stick model, a stick is randomly and simultaneously broken into species. Therefore we used the log normal distribution to generate the fragments length, L:

$$L = \exp(r * \sqrt{\mu / c}) + \mu) \text{ Where } \mu = Log(256)/(1 + c/2)$$

where r was considered as a random Gaussian number. Parameters μ and σ^2 are in fact related. $\mu = C\sigma^2$ where μ is the mean, σ^2 is the standard deviation and C is the broken stick constant. The average (mean of distribution) of fragments length is considered 256 units because of the existence of four bases in DNA fragments and the possibility of 256 (4^4) different base-pair alignment positions.

Models of DNA evolution (nucleotide substitution models)

The point wise evolving mechanism was used to evolve a mutated sequence from the start sequence. Four nucleotide substitution models have been considered (Felsenstein 2003):

- JC69 model (Jukes and Cantor 1969)
- K80 model (Kimura 1980)
- HKY85 model (Hasegawa *et al.* 1985)
- TN93 model (Tamura and Nei 1993)

In Kimura model, rates differ between transitions (α, changes from one purine to another, or from one pyrimidine to another pyrimidine) and transversion (β, changes from one purine to one pyrimidine or vice versa). Jukes-Cantor model is simply the particular case of Kimura's two -parameter model which $\alpha = \beta$, so that kappa (the ratio of transition/transversion) = 1/2, with considering $\alpha + 2\beta = 1$. (Felsenstein 2003)

Therefore, the equation for Jukes-Cantor and Kimura two-parameter model turns out to be:

$$\Pr ob(transition \mid t) = \frac{1}{4} - \frac{1}{2}\exp(-\frac{2R-1}{R+1}) + \frac{1}{4}\exp(-\frac{2}{R+1}t)$$

$$\Pr ob(transversion \mid t) = \frac{1}{2} - \frac{1}{2}\exp(-\frac{2}{R+1}t)$$

One restriction in these two models is the fact that all four bases have equal expected frequencies (25% each). Two of the most widely used models that allow arbitrary base frequencies (πA, πC, πG, πT) are HKY and Tamura-Nei. Transition rate is divided into two rates in Tamura-Nei model, αR when the base is purine and αY when it is pyrimidine. General expression of Tamura-Nei transition probability between base j and i can be written as following:

$$Prob(j \mid i,t) = \exp((-\alpha_i + \beta)t)\delta_{ij}$$

$$+ \exp(-\beta t)(1 - \exp(-\alpha_i t))\left(\frac{\pi_j \varepsilon_{ij}}{\sum_k \varepsilon_{jk}\pi_k}\right)$$

$$+ (1 - \exp(-\beta t))\pi_j$$

Note that there are the standard "kronecker delta function" δ_{ij} which is 1 if i=j and 0 otherwise, and the "Watson-Kronecker" equivalent, ε_{ij}, which is 1 if i and j are both purine or both pyrimidine and 0 otherwise. We also have α_i ,which is αR or αY depending on whether i is purine or pyrimidine. Hasegawa-Kishino-Yano (HKY) is a special case of Tamura-Nei model when $\alpha R / \alpha Y = \pi R / \pi Y$ (Felsenstein 2003).

Binding Probability of DNA

First, a systematic way is needed to define a binding probability, which determines the accuracy of the calculations. Information collected from previous annotations or experiments (Lee *et al.* 2002), represented that signal intensity should be converted to a value between 0 and 1.

Binding probability of two DNA fragments mainly depends on two factors, binding energy and environment temperature (Le Pecq *et al.* 1975). Energy is in fact

hydrogen binding between nucleotide, so it will increase in GC rich fragments. Typical model of binding probability has been shown as

$$Binding\ \mathrm{Prob}[i][j] \propto e^{-\beta E_{ij}}$$

Where $\beta \propto \frac{1}{Temperature}$ and energy (E) is considered as

E = - (2 (No. of GC nucleotide matches) + No. of AT nucleotide matches)*

Intensity of spots

In two-color microarrays, the ratio of signal intensities of two hybridized samples is used as a relative measure of gene expression (Kane *et al.* 2000). In this experiment we considered channel intensity as:

*Intensity = Relative Abundance * Binding Probability * No. of Ts*

Because of equality of both started and mutated fragments in this experiment, we considered "relative abundance" value equal to 1. The *No. of Ts* is considered as a factor because the dye attaches to the T nucleotides and therefore the fluorescent intensity is depends on the *No. of Ts* in binding fragments.[8]

Normalization

As pointed out by Yang *et al.* (2002), the purpose of normalization is to remove systematic variation in a microarray experiment, which affects the measured gene expression levels. They mentioned a number of normalization methods such as: (I) within-slide normalization, (II) paired-slide normalization for dye-swap experiments, and (III) multiple slide normalization. Dye swap experiments are extended and well established in the microarray community (Black and Doerge 2002). Let's remind the

[8] cDNA labelling protocol (http://www.niaid.nih.gov/dir/services/rtb/docs/LABELING.PDF)

simulated start and mutated fragments labeled with Cy5 (red) and Cy3 (green). The following expression is considered for each spot.

$$M_i = \log_2(\frac{R_i}{G_i})$$

Using the same sequences, labeling is repeated but this time the dyes are swapped. We thus have,

$$M_i' = \log_2(\frac{R_i'}{G_i'})$$

We expect that the normalized log ratios of the two slides are equal in magnitude and opposite signs, that is,

$$\log_2(\frac{R}{G}) - c \approx -(\log_2(\frac{R'}{G'}) - c')$$

This equation is true if additive noises in R and G can be dropped. Here c and c' denote the normalization function for two slides. As suggested in Yang *et al.* under the self normalization if $c \approx c'$

$$c \approx \frac{1}{2}[\log(\frac{R}{G}) + \log(\frac{R'}{G'})] = \frac{1}{2}(M + M') = \frac{1}{2}\log(\frac{RR'}{GG'})$$

In this experiment we calculated observed or modified intensity (OI) by addition and multiplication of normal distributed amounts of errors:

$$OI = exp\ (LogE1) * TI + E2$$

where *E1* is the multiplicative error and *E2* is the additive error, which has been generated randomly with normal distribution.

Then normalized log ratio is calculated by

$$NormalizedLogRatio = \frac{1}{2}(Log(\frac{OI\_MutatedRed}{OI\_StartGreen}) + Log(\frac{OI\_MutatedGreen}{OI\_StartRed}))$$

Machine Learning Approach of Microarray

Support Vector Machine Classification of Microarray Data

High-density DNA microarray measures the activities of several thousand genes simultaneously and the gene expression profiles have been recently used for cancer and other disease classifications. The Support Vector Machine (SVM) (Vapnik 1998) is a supervised learning algorithm, useful for recognizing subtle patterns in complex datasets. It is one of the classification methods successfully applied to the diagnosis and prognosis problems. The algorithm performs discriminative classification, learning by example to predict the classifications of previously unclassified data. The SVM was one of the methods successfully applied to the cancer diagnosis problem in the previous studies (Brown *et al.* 2000; Guyon *et al.* 2002). In principle, the SVM can be applied to very high dimensional data without altering its formulation. Such capacity is well suited to the microarray data structure.

The popularity of the SVM algorithm comes from four factors (Pavlidis *et al.* 2004). 1) The SVM algorithm has a strong theoretical foundation, based on the ideas of Vapnik Chervonenkis (VC) dimension and structural risk minimization (Vapnik 1998). 2) The SVM algorithm scales well to relatively large datasets. 3) The SVM algorithm is flexible due in part to the robustness of the algorithm itself and in part to the parameterization of the SVM via a broad class of functions, called kernel functions. The behavior of the SVM can be modified to incorporate prior knowledge of a classification task simply by modifying the underlying kernel function. 4) Accuracy: The most important explanation for the popularity of the SVM algorithm is its accuracy. The underlying theory suggests explanations for the SVMs excellent learning performance, its

widespread application is due in large part to the empirical success the algorithm has achieved (Pavlidis *et al.* 2004).

Data is represented as a P-Dimensional vector. Our interest is to separate them with a P-1 dimensional hyperplane. This is a typical form of linear classifier. Obviously there are many linear classifiers that might satisfy our purpose. However, the best performance will be achieved if we can find maximum separation (margin) between the two classes. This means that we pick the hyperplane so that the distance from the hyperplane to the nearest data point is maximized. If such a hyperplane exists, it is clearly of interest and is known as the maximum-margin hyperplane and such a linear classifier is known as a *maximum margin classifier* (Vapnik 1998).

Kernel Types. Vapnik suggested a way to create non-linear classifiers by applying the kernel trick to maximum-margin hyperplanes. $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ is called the kernel function. Here training vectors x_i are mapped into a higher (probably infinite) dimensional space by the function Φ. There are four basic kernels: linear, polynomial, radial basic function (RBF), and sigmoid:

1. Linear: $K(x_i, x_j) = x_i^T x_j$

2. Polynomial: The polynomial kernel of degree d is of the form

$$K(x_i, x_j) = (x_i, x_j)^d$$

3. RBF: The Gaussian kernel, known also as the radial basis function, is of the form

$$K(x_i, x_j) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2})$$

Where σ stands for a window width

4. Sigmoid: The sigmoid kernel is of the form

$$K(x_i, x_j) = \tanh(k(x_i x_j) + \vartheta)$$

When the sigmoid kernel is used with the SVM one can regard it as a two-layer neural network.

SVM Types:

1) C-SVC: C-Support Vector Classification (Binary Case)

Given a training set of instance-label pairs (x_i, y_i), $i = 1 \ldots l$ where $x_i \in R^n$ and $y \in \{1, -1\}^l$, the support vector machines (SVM) require the solution of the following optimization problem:

$$\min_{\omega, b, \xi} \frac{1}{2}\omega^T \omega + C\sum_{i=1}^{l} \xi_i$$
$$y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term (Cortes and Vapnik 1995; Vapnik 1998). The decision function is:

$$\text{sgn}(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b)$$

2) nu-SVC: ν-Support Vector Classification (Binary Case)

The parameter ν ∈ (0, 1) is an upper bound of the fraction of training errors and a lower bound of the fraction of support vectors (Scholkopf *et al.* 2000). Given training vectors $x_i \in R^n$, $i = 1, \ldots, l$, in two classes, and a vector $y \in R^l$ such that $y_i \in \{1, -1\}$, the primal form considered is:

$$\min_{\omega,b,\xi} \frac{1}{2}\omega^T\omega - \vartheta\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i$$
$$y_i(\omega^T\phi(x_i)+b) \geq \rho - \xi_i$$
$$\xi_i \geq 0, \rho \geq 0$$

And the decision function is:

$$\text{sgn}(\sum_{i=1}^{l} y_i \alpha_i (K(x_i, x)+b))$$

3) epsilon-SVR: ε-Support Vector Regression (ε-SVR)

One extension of the SVM is that for the regression task. A regression problem is given for the training data set $Z = \{(x_i, y_i) \in X \times Y \mid i = 1,...,M\}$ and our interest is to find a function of the form $f: X \rightarrow RD$. The primal formulation for the SVR is then given by:

$$\min_{\omega,b,\xi} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*)$$
$$y_i(\omega^T x_i + b) \leq \varepsilon + \xi_i$$
$$\xi_i, \xi_i^* \geq 0$$

We have to introduce two types of slack-variables ξ_i and ξ_i^*, one to control the error induced by observations that are larger than the upper bound of the ε-tube, and the other for the observations smaller than the lower bound. The approximate function is:

$$\sum_{i=1}^{l}(-\alpha_i + \alpha_i^*)K(x_i, x)+b$$

4) nu-SVR: ν-Support Vector Regression (ν-SVR)

Similar to ν-SVC for regression, it uses a parameter ν to control the number of support vectors (Scholkopf *et al.* 1999; Scholkopf *et al.* 2000). However, unlike ν-SVC where C is replaced by ν, here ν replaces the parameter ε of ε-SVR. Then the decision function is the same as that of ε-SVR.

5) One-class SVM: distribution estimation

One-class classification's difference from the standard classification problem is that the training data is not identically distributed to the test data. The dataset contains two classes: one of them, the target class, is well sampled, while the other class is absent or sampled very sparsely. Schölkopf *et al.* (1999) have proposed an approach in which the target class is separated from the origin by a hyperplane. The primal form considered is:

$$\min_{\omega b,\xi} \frac{1}{2}\omega^T \omega - \rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i$$
$$\omega^T \phi(x_i) \geq \rho - \xi_i$$
$$\xi_i \geq 0$$

And the decision function is:

$$\text{sgn}(\sum_{i=1}^{l} \alpha_i (K(x_i, x) + \rho))$$

CHAPTER III

RESULTS AND DISCUSSIONS

Earthworm cDNA library and EST Sequence Analysis

We cloned a total of 4032 cDNAs from the two SSH libraries. We transformed and picked 2208 clones from forward subtracted cDNA pools and 1824 from the reverse subtracted cDNA pools. After running on 96-well gel electrophoresis, 216 clones were found to be false positives with no inserts or had more than one insert. The remaining 3816 clones were sequenced and produced 3144 good quality sequences with an average length of 310 bases. We batch-deposited them in the GenBank db EST under accession numbers EH669363-EH672369 and EL515444-EL515580. Clone sequences that were too short (<50 bases) or of poor quality (<50 good quality bases, see methods for quality criteria) were excluded from further analysis. The observed failure rate (18%) is typical for high-throughput sequencing (Deng *et al.* 2006b).

The deposited, cleaned sequences were further assembled into 2231 clusters (or unique sequences) on the basis of sequence similarity and quality. Nearly 80% or 1783 of the clusters produced were singletons, and 80% of the remaining 448 contigs (average length = 428 bases) were assembled from 2 or 3 clone sequences (Figure 9). The highest number of sequences assembled into one contig was 30.

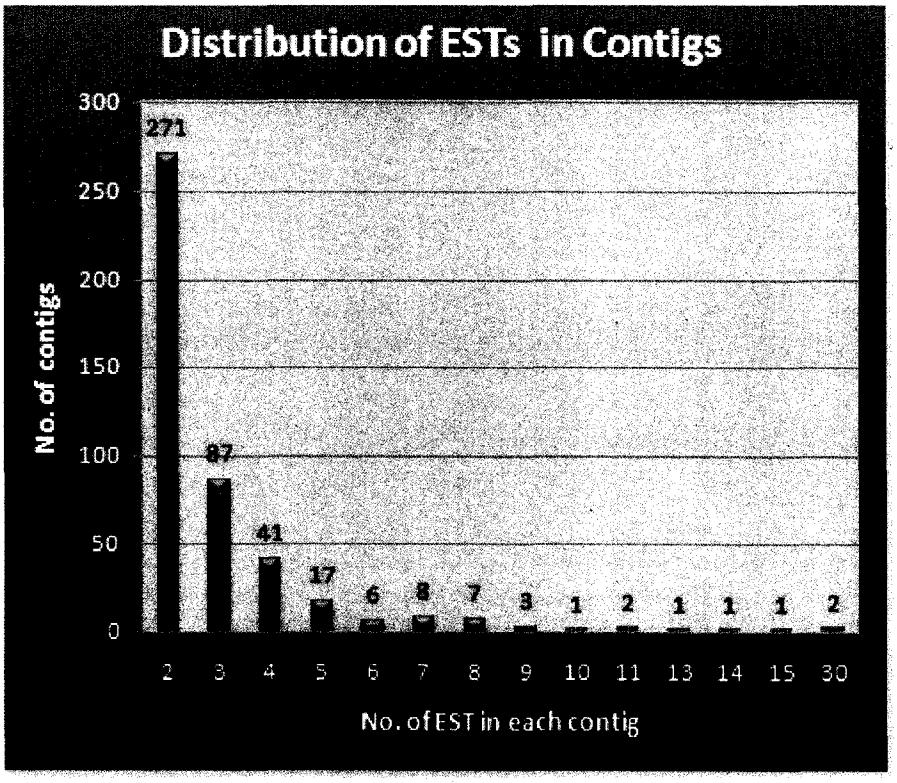


Figure 9. Distribution of 1361 good quality ESTs in 448 assembled contigs

The most represented putative genes in our libraries are Cd-metallothionein, cytochrome oxidase, chitotriosidase, actin, ATP synthase, Nahoda protein, lysozyme, SCBP (soluble calcium binding protein), ferritin, troponin T, lumbrokinase, and myohemerythrin (Table 3).

Contig	ESTs	Length	Accession Version #	bit	E-value	Identities	Organism	Description
Contig423	7	452	AAH69614.1	336	1.00E-30	64/137	Homo sapiens	CHIT1 protein
Contig424	7	480	CAE18118.1	205	2.00E-15	40/58	Lumbricus terrestris	SCBP3 protein
Contig426	7	659	AAW25147.1	171	5.00E-11	42/83	Schistosoma japonicum	SJCHGC00665 protein
Contig427	7	494	CAA48798.1	714	3.00E-74	132/135	Podocoryne carnea	actin
Contig428	7	230	BAC06447.1	195	3.00E-14	37/76	Haemaphysalis longicornis	chitinase
Contig428	7	230	NP_001020370.1	182	1.00E-12	36/73	Homo sapiens	chitinase 3-like 2 isoform c
Contig429	7	439	ABC60436.1	749	2.00E-78	145/146	Hirudo medicinalis	cytoplasmic actin
Contig431	8	397	AAX51817.1	383	5.00E-36	73/100	Diloma arida	actin
Contig434	8	579	CAA65364.1	971	1.00E-104	189/189	Lumbricus terrestris	Actin
Contig435	8	601	AAA96144.1	322	1.00E-28	62/134	Hirudo medicinalis	destabilase I
Contig436	8	810	XP_394202.2	217	3.00E-16	47/162	Apis mellifera	PREDICTED: similar to GA11808-PA
Contig436	8	810	EAL25702.1	216	4.00E-16	51/183	Drosophila pseudoobscura	GA11808-PA
Contig437	8	394	AAX77000.1	552	1.00E-55	110/122	Metaphire feijani	cytochrome c oxidase subunit 1
Contig438	9	1055	EAR81082.1	127	1.00E-05	27/60	Tetrahymena thermophila	hypothetical protein TTHERM_02141640
Contig440	9	472	NP_008244.1	439	2.00E-42	96/152	Lumbricus terrestris	ATP6_10599 ATP synthase F0 subunit 6
Contig442	11	449	CAA65364.1	760	1.00E-79	147/147	Lumbricus terrestris	Actin
Contig443	11	846	NP_008239.1	256	1.00E-20	57/105	Lumbricus terrestris	COX2_10599 cytochrome c oxidase subunit II
Contig444	13	894	AAH69614.1	614	4.00E-62	128/294	Homo sapiens	CHIT1 protein
Contig446	15	584	AAX62723.1	576	4.00E-58	122/166	Eisenia fetida	cytochrome oxidase subunit I
Contig448	30	488	CAA15423.1	246	5.00E-20	40/41	Eisenia fetida	metallothionein

Table 3. The most represented putative genes in the *Eisenia fetida* cDNA libraries

Using SSH-PCR, we enriched earthworm cDNAs responsive to exposure of ten ORCs that represent three classes of chemicals, i.e., nitroaromatics [2,4-dinitrotoluene, 2,6-dinitrotoluene, 2,4,6-trinitrotoluene (TNT), and trinitrobenzene], heterocyclic nitroamines (1,3,5-trinitroperhydro-1,3,5-triazine or RDX and 1,3,5,7-tetranitro-1,3,5,7-

tetrazocane or HMX) and heavy metals (Cd, Cu, Zn and Pb) (Figure 3). Exposure times varied from 4, 14, and 28 days to capture gene expression changes at different time points. This library construction strategy served our downstream purpose of making cDNA microarrays with the isolated cDNA clones even though we cannot identify which cDNA or groups of cDNAs responded to which compound and at which exposure time point using the raw EST data.

The combination of SSH-PCR and cDNA microarray analysis has been a widely used approach for identifying differentially expressed genes (Ghorbel *et al.* 2006; Yang *et al.* 1999) and characterizing mechanisms of action of known and suspected toxicants (Rim *et al.* 2004; Soetaert *et al.* 2006). The microarray studies have generated data enabling us to further identify differentially expressed transcripts and to elucidate sublethal toxicological mechanisms in *E. fetida* exposed to TNT alone or a mixture of TNT and RDX.

It is worth noting that the comparative sequence analysis (23%) and functional classification (7%) based on GO and KEGG analysis only found a small portion of the ESTs highly homologous ($E \leq 10^{-5}$) with well-annotated genes. Nevertheless, the functions of these ESTs are widely distributed representing 830 different GO terms and 99 different KEGG pathways. Notably, genes putatively involved in carbohydrate, energy and amino acid metabolism, cellular processes of endocrine, immune, nervous and sensory systems, signal transduction, DNA transcription, RNA translation and post-translation splicing are identified suggesting that the ten ORCs may have affected a wide range of important pathways.

From a candidate biomarker gene point of view, we found repeatedly the existence of some toxicant-specific *E. fetida* mRNAs in our libraries (Table 3). For instance, the expression of metallothionein (MT) mRNA, the most abundant transcript in our cDNA libraries, is reportedly a sensitive and early genetic biomarker of metal exposure (Brulle *et al.* 2006; Brulle *et al.* 2007; Demuynck *et al.* 2006; Galay-Burgos *et al.* 2003). Demuynck *et al.* demonstrated that a single exposure to 8 mg Cd/kg of dry soil for one day induced MT mRNA. Brulle *et al.* observed changes in MT mRNA expression as early as 14 hours after exposure. Copper is an essential element for the activity of a number of physiologically important enzymes including cytochrome c oxidase (COX), Cu/Zn-superoxide dismutase (SOD), and dopamine-beta-hydroxylase (DBH). However, exposure to a toxic level of copper can not only induce MT for Cu sequestration but also alter the expression of COX (Table 3). Further research is required to establish dose-dependent gene expression in both laboratory and field conditions.

Comparative Sequence Analysis

We used the 2,231 unique sequences to search non-redundant protein databases using blastx (Deng *et al.* 2006b; Wang *et al.* 2007). A total of 743 sequences (33% of all unique sequences) matched known proteins with cut-off expectation (*E*) values of 10^{-5} or lower, among which 71 (3%) had *E*-values between 10^{-100} and 10^{-50}, 309 (14%) between 10^{-50} and 10^{-20}, and 363 (16%) between 10^{-20} and 10^{-5} (Table 4).

	Contig		Singleton		Total	
Homology	N	%	N	%	N	%
$10^{-150} < E \leq 10^{-100}$	0	0	0	0	0	0
$10^{-100} < E \leq 10^{-50}$	38	8	33	2	71	3
$10^{-50} < E \leq 10^{-20}$	93	21	216	12	309	14
$10^{-20} < E \leq 10^{-5}$	78	17	285	16	363	16
Total meaningful match ($E \leq 10^{-5}$)	209	46	534	30	743	33
Less meaningful match ($E > 10^{-5}$)	165	37	715	40	880	40
No match (No hit)	74	17	534	30	608	27
Total	448	100	1783	100	2231	100

Table 4. Homology analysis of the 2231 unique *Eisenia fetida* EST sequences based on the results from BLASTX against NCBI's nr database.

A total of 880 unique sequences had less meaningful matches ($E > 10^{-5}$). The remaining 608 sequences (27%) had no matches. We also examined unique *E. fetida* sequences to determine similarity to the genes of four model organisms *Drosophila melanogaster, Mus musculus, Saccharomyces cerevisiae,* and *Caenorhabditis elegans.* A total of 830 blastx matches were found for 517 *E. fetida* unique sequences (23%) at the cut-off *E*-value of 10^{-5} (Table 5).

Organism Name	Number of matches	% of unique sequences
Drosophila melanogaster	265	12%
Mus musculus	447	20%
Saccharomyces cerevisiae	5	0.2%
Caenorhabditis elegans	113	5%
Total matches	830	
Total unique sequences	517	23%

Table 5. Comparison of significant homologous matches ($E \leq 10^{-5}$) to four model organisms of the 2231 unique *Eisenia fetida* EST sequences.

Some *E. fetida* ESTs matched genes conserved between the four organisms. More than 50% of the matches came from the mouse genome, whereas only five matches were found in the yeast genome. These results suggest that earthworms may be more evolutionarily distant from the yeast than from the other three organisms.

Functional Classification

We adopted the Gene Ontology (GO) annotation of the aforesaid four model organisms to interpret the function of the *E. fetida* ESTs (Deng *et al.* 2006b; Wang *et al.* 2007). Each unique sequence of *E. fetida* was assigned the same gene functions of the best blastx hit genes ($E \leq 10^{-5}$) in these model organisms' genome. The assigned GO terms for the unique sequences are categorized and outlined in Table 6 (biological process), Table 7 (molecular function), and Table 8 (cellular component). The most represented molecular function is "binding" accounting for 51% of the total 517 unique sequences assigned with at least one GO term (Table 5), whereas those for biological processes are "cellular process" (39%) and "physiological process" (40%) (Table 6). In terms of the final child GO categories, the most frequently assigned biological processes are "protein metabolism" (12.5%), "cellular macromolecule metabolism" (11.7%), and "cellular protein metabolism" (11%) under both cellular and physiological processes (Table 6), whereas those for molecular functions are "hydrolase activity" (11%) and "protein binding" (10%) (Table 7). The largest subcategory in cellular components is "intracellular organelle" (23.6%) under both the intracellular part and the organelle (Table 8).

Gene Ontology term	Unique sequences	Percentage of total matches
cellular process	328	39.52%
cell communication	52	6.27%
cellular physiological process	309	37.23%
cell organization and biogenesis	62	7.47%
cellular metabolism	255	30.72%
cellular biosynthesis	46	5.54%
cellular macromolecule metabolism	97	11.69%
cellular protein metabolism	92	11.08%
regulation of cellular physiological process	48	5.78%
transport	71	8.55%
regulation of cellular process	51	6.14%
development	51	6.14%
physiological process	331	39.88%
cellular physiological process	309	37.23%
cell organization and biogenesis	62	7.47%
cellular metabolism	255	30.72%
cellular macromolecule metabolism	97	11.69%
localization	53	6.39%
metabolism	272	32.77%
biosynthesis	70	8.43%
cellular metabolism	255	30.72%
cellular biosynthesis	46	5.54%
cellular macromolecule metabolism	97	11.69%
cellular protein metabolism	92	11.08%
macromolecule metabolism	181	21.81%
biopolymer metabolism	58	6.99%
cellular macromolecule metabolism	97	11.69%
macromolecule biosynthesis	34	4.10%
protein metabolism	96	11.57%
primary metabolism	164	19.76%
protein metabolism	104	12.53%
regulation of physiological process	51	6.14%
regulation of biological process	57	6.87%
response to stimulus	47	5.66%

Table 6. Distribution of Gene Ontology biological process terms assigned to *Eisenia fetida* unique sequences on the basis of their homology to the annotated genome of four model organisms.

Gene Ontology term		Unique sequences	Percentage of total matches
antioxidant activity		2	0.24%
binding		426	51.33%
	carbohydrate binding	18	2.17%
	cofactor binding	6	0.72%
	ion binding	84	10.12%
	lipid binding	5	0.60%
	metal cluster binding	3	0.36%
	neurotransmitter binding	3	0.36%
	nucleic acid binding	53	6.39%
	nucleotide binding	68	8.19%
	pattern binding	10	1.20%
	peptide binding	4	0.48%
	protein binding	90	10.84%
	tetrapyrrole binding	5	0.60%
	vitamin binding	2	0.24%
catalytic activity		194	23.37%
	helicase activity	4	0.48%
	hydrolase activity	94	11.33%
	isomerase activity	8	0.96%
	ligase activity	7	0.84%
	lyase activity	11	1.33%
	oxidoreductase activity	46	5.54%
	small protein activating enzyme activity	3	0.36%
	transferase activity	27	3.25%
enzyme regulator activity		16	1.93%
motor activity		4	0.48%
nutrient reservoir activity		2	0.24%
signal transducer activity		26	3.13%
structural molecule activity		47	5.66%
transcription regulator activity		16	1.93%
translation regulator activity		13	1.57%
transporter activity		33	3.98%

Table 7. Distribution of Gene Ontology molecular function terms assigned to *Eisenia fetida* unique sequences on the basis of their homology to the annotated genome of four model organisms.

Gene Ontology term	Unique sequences	Percentage of total matches
cell part	280	33.73%
intracellular part	224	26.99%
calcineurin complex	2	0.24%
cytoplasm	152	18.31%
cytoplasmic part	132	15.90%
intracellular organelle	196	23.61%
intracellular organelle part	97	11.69%
proteasome regulatory particle (sensu Eukaryota)	8	0.96%
proton-transporting ATP synthase complex	4	0.48%
respiratory chain complex I	3	0.36%
respiratory chain complex III	3	0.36%
ribonucleoprotein complex	35	4.22%
RNA polymerase complex	2	0.24%
ubiquinol-cytochrome-c reductase complex	3	0.36%
membrane	107	12.89%
membrane part	81	9.76%
protein serine/threonine phosphatase complex	2	0.24%
envelope	33	3.98%
extracellular matrix	10	1.20%
extracellular matrix part	6	0.72%
extracellular region	51	6.14%
extracellular region part	40	4.82%
membrane-enclosed lumen	2	0.24%
organelle	196	23.61%
intracellular organelle	196	23.61%
membrane-bound organelle	148	17.83%
non-membrane-bound organelle	68	8.19%
organelle part	97	11.69%
vesicle	8	0.96%
organelle part	97	11.69%
protein complex	102	12.29%
synapse	7	0.84%
synapse part	3	0.36%

Table 8. Distribution of Gene Ontology cellular component terms assigned to *Eisenia fetida* unique sequences on the basis of their homology to the annotated genome of four model organisms.

Pathway Assignment

We assigned the unique *E. fetida* sequences to a specific Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway based on their matching Enzyme Commission (EC) numbers. A total of 157 unique sequences (accounting for 7% of all unique sequences) including 28 contigs and 129 singletons matched enzymes with an EC number. Fifty-eight unique sequences are involved in two or more pathways. The remaining 99 pathway-assigned sequences are mapped to only one pathway. Eighty-two unique sequences (52% of total) containing 14 contigs and 68 singletons were assigned to metabolism pathways (Table 9). Amino acid metabolism has the highest number of assigned pathways, followed by carbohydrate metabolism, energy metabolism, translation, and signal transduction. Genes putatively coded by a singleton EW1_F1plate05_B07 (enoyl coenzyme A hydratase) and Contig 251 (thioredoxin peroxidase) are most versatile, which are mapped to 10 and 8 pathways, respectively.

KEGG pathway	No. of unique sequence	Percentage of total unique sequences*	No. of KEGG pathways mapped
Metabolism	**82**	**52%**	**57**
Carbohydrate Metabolism	35	22%	10
Energy Metabolism	28	18%	8
Nucleotide Metabolism	2	1%	2
Amino Acid Metabolism	18	11%	12
Metabolism of Other Amino Acids	10	6%	3
Glycan Biosynthesis and Metabolism	6	4%	8
Metabolism of Cofactors and Vitamins	9	6%	6
Biosynthesis of Secondary Metabolites	2	1%	1
Xenobiotics Biodegradation and Metabolism	6	4%	7
Genetic Information Processing	**28**	**18%**	**6**
Transcription	2	1%	2
Translation	17	11%	1
Folding, Sorting and Degradation	9	6%	3
Environmental Information Processing	**27**	**17%**	**10**
Membrane Transport	1	1%	1
Signal Transduction	14	9%	6
Signaling Molecules and Interaction	13	8%	3
Cellular Processes	**37**	**24%**	**18**
Cell Motility	9	6%	3
Cell Communication	13	8%	4
Endocrine System	4	3%	3
Immune System	5	3%	3
Nervous System	8	5%	2
Sensory System	3	2%	1
Development	3	2%	2
Human Diseases	**9**	**6%**	**8**
Neurodegenerative Disorders	6	4%	4
Metabolic Disorders	2	1%	2
Cancers	2	1%	2

Table 9. KEGG pathway mapping for *Eisenia fetida* unique sequences. The total number of mapped unique sequences is 157. A complete listing of the KEGG pathways mapped for 157 unique *Eisenia fetida* sequences in Appendix A.

ESTMD (EST Model Database) Web Application

We introduce a high-performance Web-based application consisting of EST modeling and database (ESTMD) to facilitate and enhance the retrieval and analysis of EST information. The ESTMD is a highly performed, web-accessible and user-friendly relational database (Deng *et al.* 2006a). It provides a number of comprehensive search tools for mining EST raw, cleaned and unique sequences, Gene Ontology, pathway information and a variety of genetic Web services such as BLAST search, data submission and sequence download pages. It facilitates and enhances the retrieval and analysis of EST information by providing a number of comprehensive tools for mining raw, cleaned and clustered EST sequences, GO terms and KEGG pathway information as well as a variety of web-based services such as BLAST search, data submission and sequence download. The application is developed using advanced Java technology (JSP and Servlets) and it supports portability, extensibility and data recovery. It can be accessed at http://mcbc.usm.edu/estmd/.

The main ESTMD tables are clone, contigview, est new, flybase, geneon, gomodels, pathway, term, uniseqhit, master_search and unisequence (Figure 10). Main sequence information including ECnumber, Labname, raw and clean sequence, and vector information are stored in the master_search table. Figure 10 shows the main schema of ESTMD database.

Figure 10. The main schema of ESTMD database.

Software Architecture

Apache2.2 acts as a HTTP server. Tomcat 5.5 is the servlet container used. Both of them are platform-independent and therefore can run on UNIX, Linux and Windows platform. The server-side programs are implemented using Java technologies. Java

Servlets and JSP (Java Server Pages) are used as interfaces between users and the database. Java 2D Graphic is used to generate and express contig view. The user interface of the database is created using HTML and JavaScript. JavaScript can check the validation of the users' input on the client side, which reduces some burden on the server side.

ESTMD is an integrated Web-based application consisting of client, server and backend database, as shown in Figure 11. ESTMD is a new integrated Web-based model that consists of (1) a front-end user interface for accessing the system and displaying results, (2) a middle layer for providing a variety of Web services such as data processing, task analysis, search tools and so on, and (3) a back-end relational database system for storing and managing biological data. It provides a wide range of search tools to retrieve original (raw), cleaned and unique EST sequences and their detailed annotated information. Users can search not only the sequence, gene function and pathway information using single sequence ID, gene name or term, but also the function and pathway information using a file including a batch of sequence IDs. Moreover, users can quickly assign the sequences into different functional groups using the Gene Ontology Classification search tool. ESTMD provides a useful tool for biological scientists to manage EST sequences and their annotated information.

The workflow process begins when users input keywords or IDs from the web interface and then submit them as a query to the server. The server processes the query and retrieves data from the backend database through the database connection interface. The results are processed and sent to the users in proper formats.

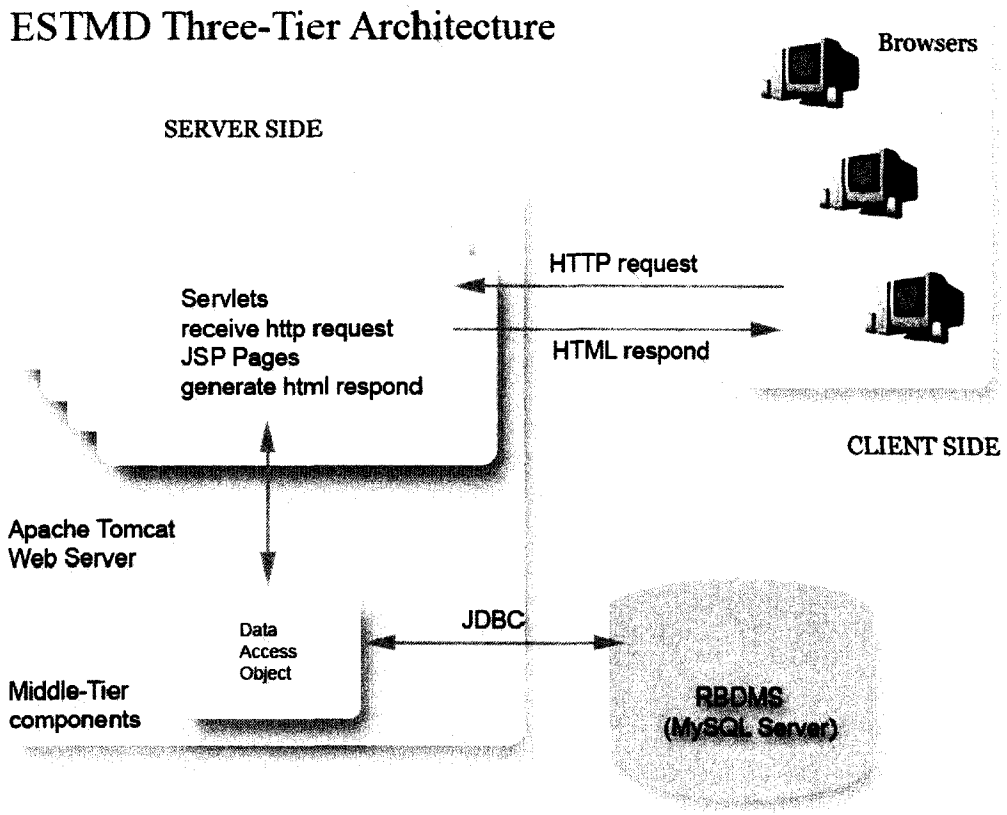


Figure 11. The software architecture of ESTMD

Web Services

The Web application provides a number of search tools and Web services, including search in detail, search by keyword, Gene Ontology search, Gene Ontology classification, and pathway search. Users may search the database by several methods. Users are also allowed to download data from or submit data to the database.

Search ESTMD

Users may search the database by gene symbol, gene name, or any type of ID (such as unique sequence ID, clone ID, FlyBase ID, Genbank ID or accession ID). The Web search interface is given in Figure 12. The keyword search returns results in a table rather than in plain text. The results include clone ID, raw sequence length, cleaned

sequence length, unique sequence ID, unique sequence length, gene name and gene symbol. It has a hyperlink to the contig view which uses color bars to show the alignment between contig and singlet sequences, as shown in Figure 13.

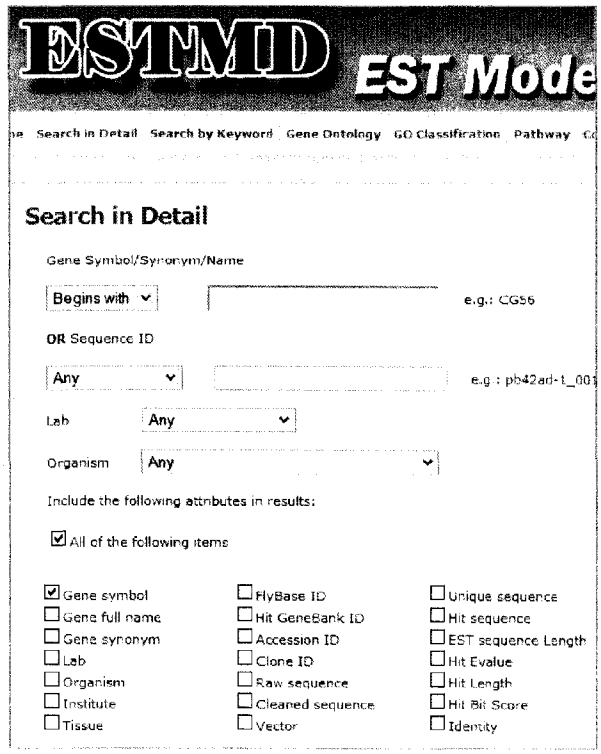


Figure 12. Web search interface showing fields for user input and attributes of results

Contig View

Users may input the contig sequence ID to see the alignments of the contig and all of the singlet sequences contained. This feature allows users to check if the contig is correct (Figure 13).

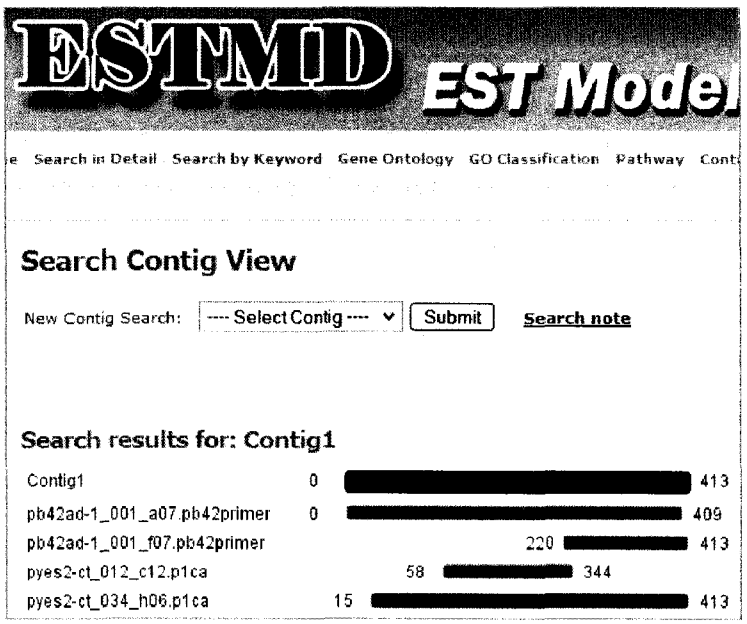


Figure 13. An example result of contig view.

Gene Ontology and Classification

ESTMD allows users to search Gene Ontology not only by a single gene name, symbol or ID, but also by a file containing a batch of sequence IDs. The file search capability in ESTMD allows users to get function information of many EST sequences or genes at one time instead of searching one by one. Users can search all the GO terms by selecting one molecular function, biological process or cellular component to submit their search. The result table includes GO ID, term, type, sequence ID, hit ID, and gene symbol. Classifying genes into different functional groups is a good way to know the gene function relationship. Another important feature of ESTMD is Gene Ontology Classification search. ESTMD defines a series of functional categories according to molecular function, biological process and cellular component. Users can classify Gene Ontology of a batch of sequences. The results show type, subtype, how many sequences

and percentage of sequences in this subtype (Figure 14). This feature is very useful for cDNA microarray gene function analysis. In this type of array, ESTs are printed on slides. Therefore, the Gene Ontology Classification tool in ESTMD can help automatically divide these ESTs into different functional groups.

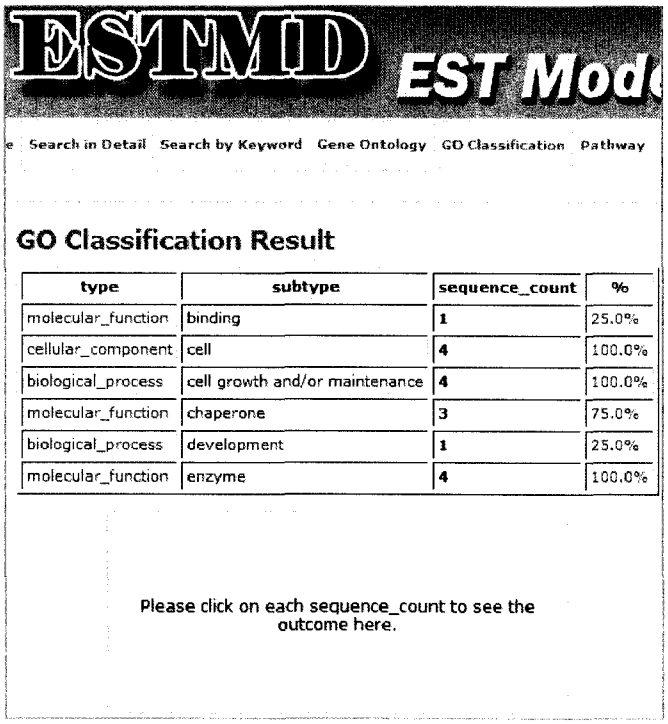
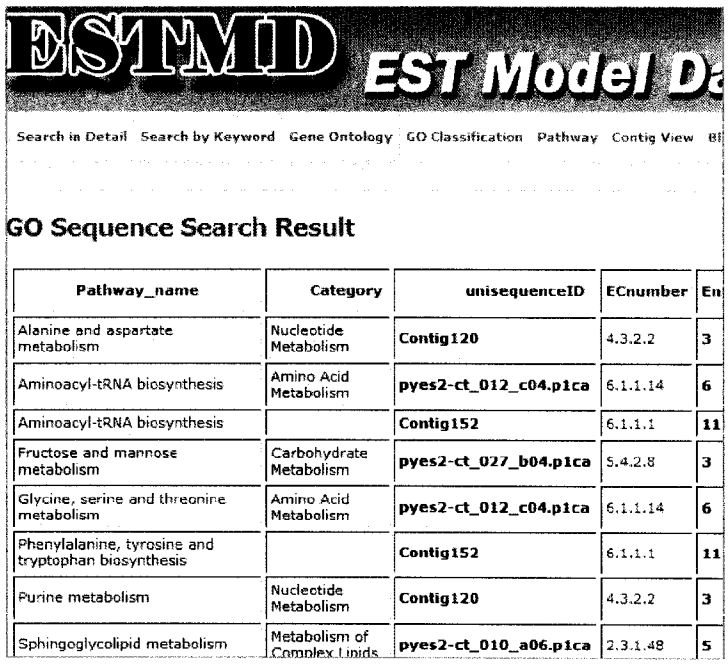


Figure 14. The results of classifying Gene Ontology from a text file which contains 4 sequence IDs.

Pathway

The Pathway page allows the search of a pathway by single or multiple gene names, IDs, EC numbers, enzyme names, or pathway names. File search is also provided on this page. The scope of the search may be the whole pathway or just our database. The results show pathway name, category, unique sequence ID, EC number, and enzyme count (Figure 15). The pathway information comes from KEGG metabolic pathway. We have downloaded, reorganized and integrated it into our database.



Pathway_name	Category	unisequenceID	ECnumber	En
Alanine and aspartate metabolism	Nucleotide Metabolism	Contig120	4.3.2.2	3
Aminoacyl-tRNA biosynthesis	Amino Acid Metabolism	pyes2-ct_012_c04.p1ca	6.1.1.14	6
Aminoacyl-tRNA biosynthesis		Contig152	6.1.1.1	11
Fructose and mannose metabolism	Carbohydrate Metabolism	pyes2-ct_027_b04.p1ca	5.4.2.8	3
Glycine, serine and threonine metabolism	Amino Acid Metabolism	pyes2-ct_012_c04.p1ca	6.1.1.14	6
Phenylalanine, tyrosine and tryptophan biosynthesis		Contig152	6.1.1.1	11
Purine metabolism	Nucleotide Metabolism	Contig120	4.3.2.2	3
Sphingoglycolipid metabolism	Metabolism of Complex Lipids	pyes2-ct_010_a06.p1ca	2.3.1.48	5

Figure 15. The results of a pathway search from a text file, ordered by Pathway, are shown. The blue texts mark hyperlinks on the items.

BLAST

The BLAST program (Altschul *et al.* 1990) is used to search and annotate EST sequences. The BLAST page allows users to do BLAST searches by choosing different databases. The databases contain raw EST sequences, cleaned EST sequences and assembled unique sequences, as well as NCBI GenBank nr (non-redundant), Swissprot amino acid, and gadfly nucleotide.

GOfetcher: A Complex Searching Facility for Gene Ontology

The GOfetcher Web application and search engine has been written in PHP programming language. Therefore, GOfetcher is platform independent and can run on any standard machine with a Web browser. It communicates with a local MySQL database in the backbone which stored the data.

Search capabilities

In this project we developed a web application, GOfetcher, with a very comprehensive search facility and variety of output formats for the results to overcome these problems.

The GOfetcher Database can be searched using any Web browser. The search options enable users to input simple as well as complex queries and search the GOfetcher. The advanced search panel allows users to define specific queries using Boolean operators connecting multiple fields for specific requirements. Each search returns a result list including species ID, species unique ID, symbol, GO term, name, and category as well as a summary of the distinct matching entries with the pie chart for categories. The user can also print or export query results in multiple formats including Excel, Word and XML. An online tutorial has been developed to describe the various features of the database with examples.

GOfetcher has three different levels for searching the GO:

1. Quick Search: It searches any keyword as a species ID, species unique ID, symbol, GO term, name, or category. Keywords should be separated by any comma delimited or whitespace such as space, tab, or line break. There is also option for searching "any words", "all words", or "exact phrase".
2. Advanced Search (Figure 16): In the "advanced search" tab user is able to search very complicated combination of keywords for the species ID, species unique ID, symbol, GO term, name, or category. Results could be the “exact match”, “contain”, “not contain”, or “starts with” keywords.

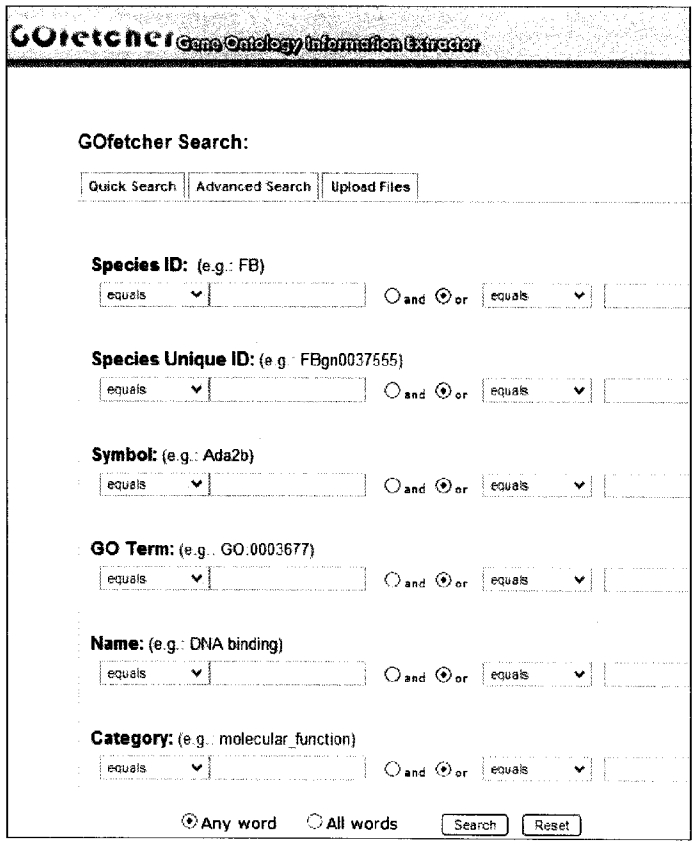


Figure 16. GOfetcher Advanced Search

3. Upload Files (Figure 17): In the "upload files" tab users can upload file(s) containing keywords which like quick search separated by comma or any white spaces. GOfetcher then searches for any words in the files and shows the results.

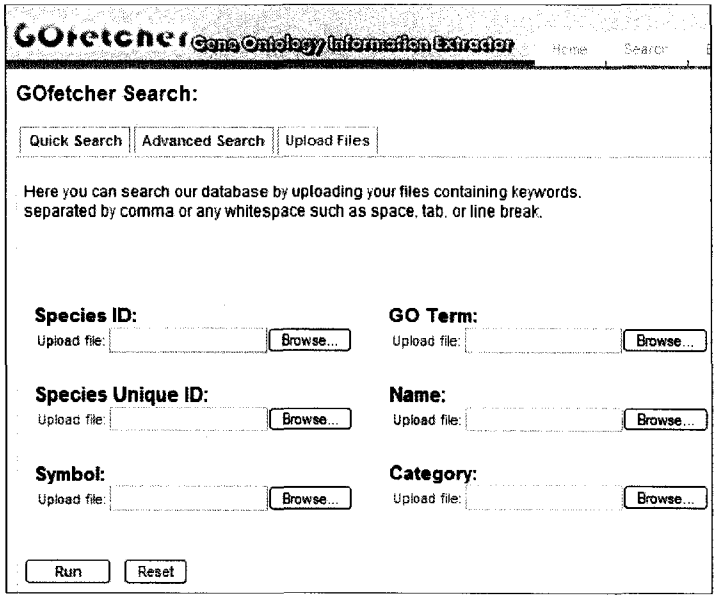


Figure 17. GOfetcher File Upload

Browse by Species

From the browse menu it is possible to browse the database by species. Currently our database includes 18 model organism's information including; *Arabidopsis thaliana, Bacillus anthracis, Caenorhabditis elegans, Campylobacter jejuni, Candida albicans, Drosophila melanogaster, Mus musculus, Oryza sativa, Rattus norvegicus, Saccharomyces cerevisiae,* and *Vibrio cholera.* Table 10 illustrates organisms and the annotation records in details. The total numbers of annotations are 847,510 records.

Search Results

If selected by the user a summary of the distinct matching entries with a pie chart for categories will appeare on the top of the search results page (Figure 18). The summary table contains unique numbers for species unique ID, symbol, GO term, and term name without any redundancy. By clicking on each number users will be able to view the related list.

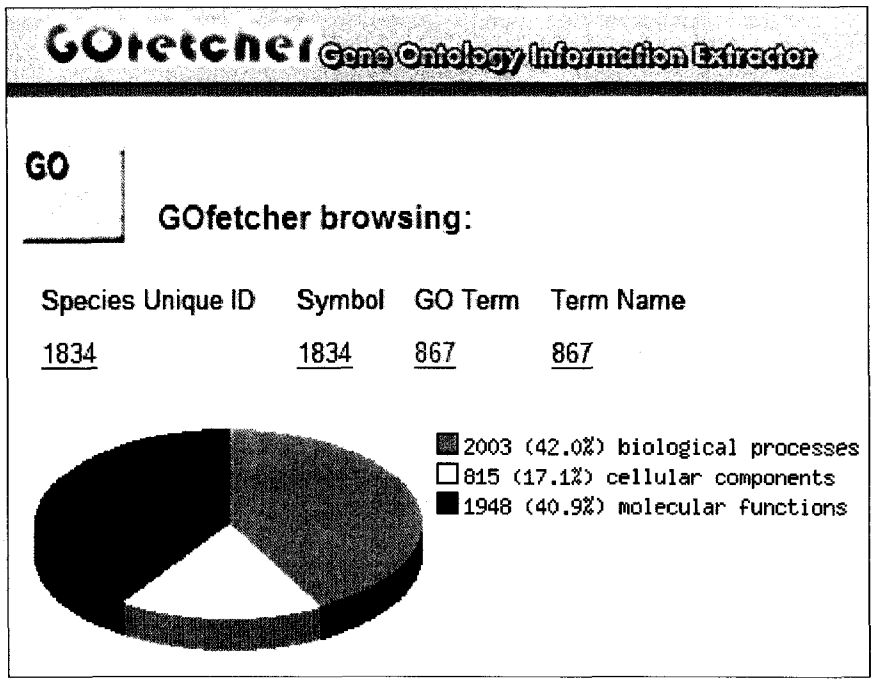


Figure 18. Distinct matching entries with a pie chart for categories

Results include the following:

1. Species ID: This is often a two or three letter abbreviation for a species, for instance FB for flybase and MGI for mouse.
2. Species Unique ID: specific Accession ID to a species, e.g. "MGI:1918918" for mouse and "FBgn0015567" for flybase Drosophila. Information about a specific organism from related external databases provided here.
3. Symbol: This is the gene name with access to NCBI gene database information
4. GO Term: Gene Ontology specific term ID with both tree and graphic view
5. Name: Gene Ontology specific term name with related information from gene ontology database (geneontology.org).
6. Category: One of the three organizing principles of GO which are "cellular component", "biological process" and "molecular function".

The output in GOfetcher can be saved into several different formats; Excel spreadsheet, Microsoft word document, comma-separated values (CSV), the Extensible Markup Language (XML) format, and printer friendly format. Although it is not our aim to be a visual-oriented tool, we provided a tree and graphical view for the results.

Fetching

Figure 19 illustrates the flow chart for the searching and fetching process. When search results appear, each record contains information fetched from related external databases. The GOfetcher extracts information from a variety of databases including MCBI, ArabidopsisDB, GeneDB, *Saccharomyces* Genome Database, FlybaseDB, Mouse Genome Informatics, Wormbase and TIGR Annotation. Table 10 shows the complete list of 18 organisms currently available through the fetching process.

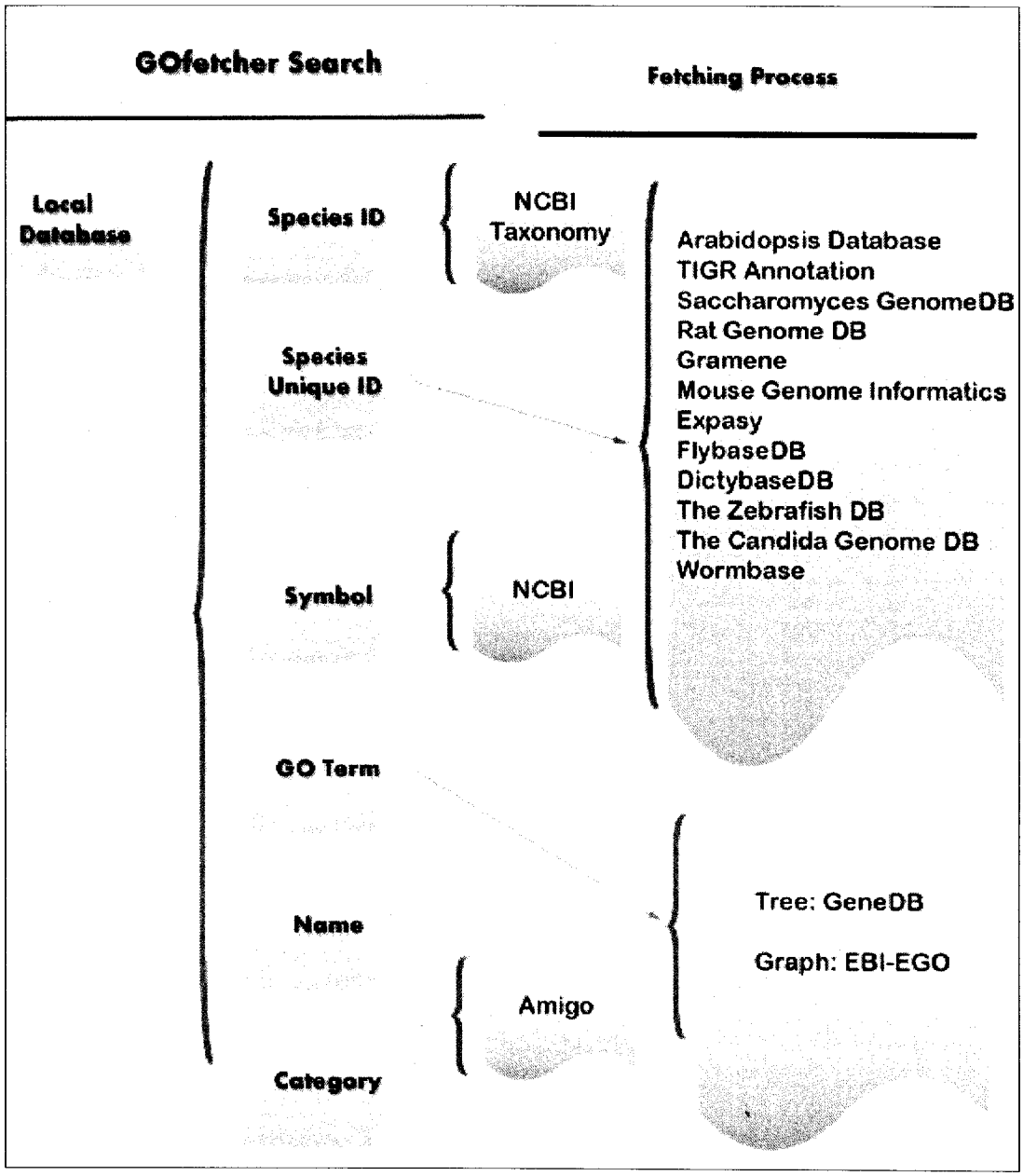


Figure 19. Flow chart for searching and fetching process

Species	ID	Database	Annotations
Arabidopsis thaliana	TAIR	TAIR/TIGR	101288
Bacillus anthracis Ames	BA	TIGR	13366
Caenorhabditis elegans	WB	WormBase	60303
Campylobacter jejuni	CMR	TIGR	4766
Candida albicans	CGD	CGD	2024
Coxiella burnetii	CBU	TIGR	5283
Danio rerio	ZFIN	ZFIN	55099
Dictyostelium discoideum	DDB	DictyBase	27056
Drosophila melanogaster	FB	FlyBase	57838
Homo sapiens	GH	EBI	37538
Listeria monocytogenes	LMO	TIGR	7214
Mus musculus	MGI	MGI	110141
Oryza sativa	GR	Gramene	80425
Plasmodium falciparum	PF	Sanger GeneDB	11706
Rattus norvegicus	RGD	RGD	122444
Saccharomyces cerevisiae	SGD	SGD	29458
Trypanosoma brucei	TB	Sanger GeneDB	14763
Vibrio cholerae	VC	TIGR	9610

Table 10. List of 18 organisms currently available through GOfetcher

Toxicogenomics Analysis of 2,4,6-Trinitrotoluene in *Eisenia fetida*

Microarray hybridization and data analysis

We used 40 arrays and chose an interwoven loop design (Figure 2) to accommodate the selected 20 worm samples. Each biological replicate of cDNA samples was hybridized four times on four different arrays with two swaps of Cy3 and A647 fluorescent dyes. The acquired array image and processed signal data of all 40 arrays have been deposited in the GEO database with a series number of GSE7024[9].

Scatter plot, illustrated sample in Figure 20 (See complete 40 slides in Appendix B), was used to identify the relationship between the two dyes and to check the hybridization quality. Log_2R is plotted against log_2G. Scatter plot is useful in the early stage of analysis as it can help to determine whether a linear regression model is appropriate and also if the normalization was effective. A correlation between the variables results in the clustering of data points along a line. MA plot (Figure 21) was also used to see the log-ratios and intensity-dependent effects at the same time (See all 40 slides in Appendix B).

[9] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7024

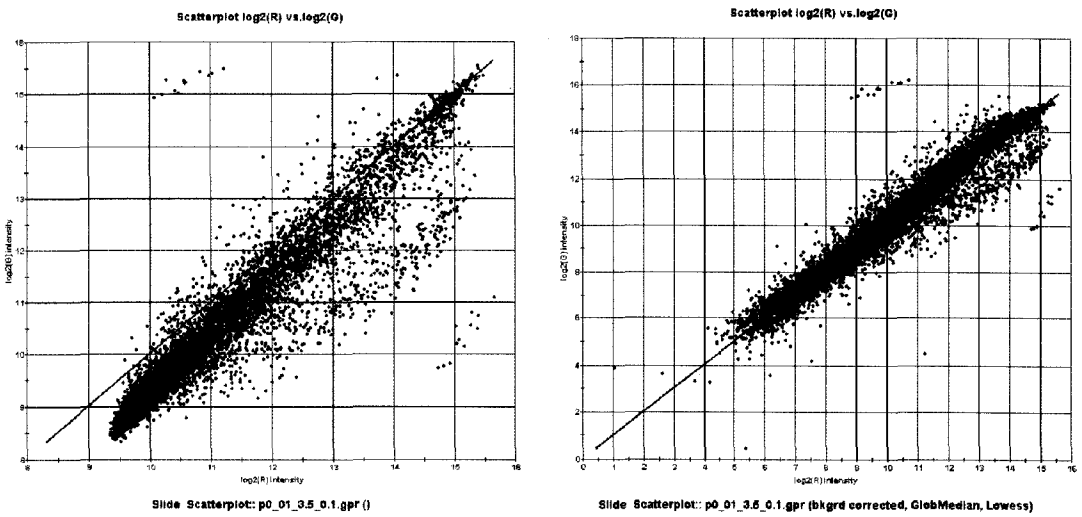


Figure 20. Scatter plot of array 1 – left, raw data and right, normalized data

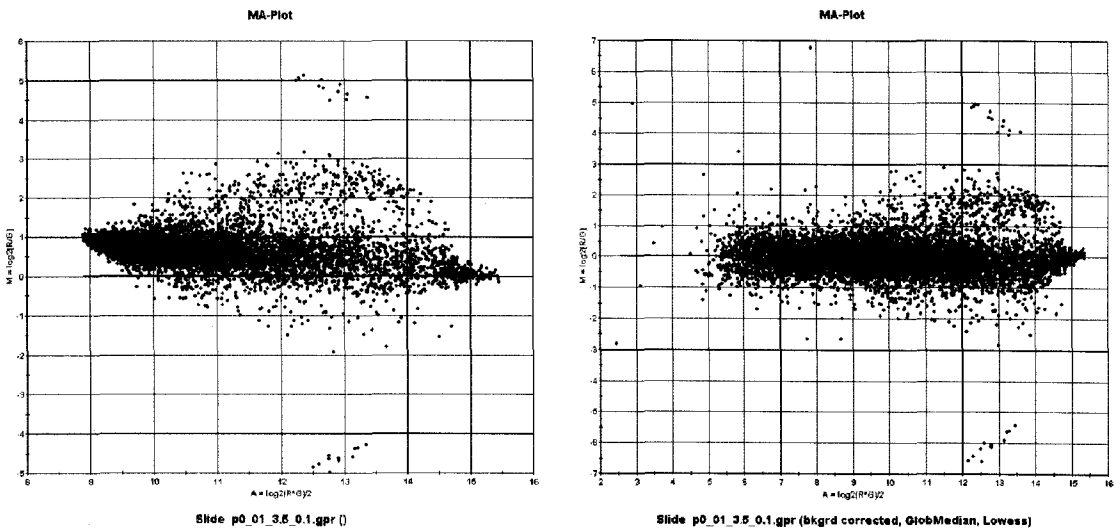


Figure 21. MA plot of array 1 – left, raw data and right, normalized data

We applied three stages of normalization to all 40 arrays. First background subtraction, then mean adjustments and finally lowess normalization have been applied. The box plot illustrated in Figure 23 shows the effectiveness of normalization methods comparing to Figure 22, the raw data. We also applied between array normalization in order to scale data for comparative analysis which is illustrated in Figure 4.

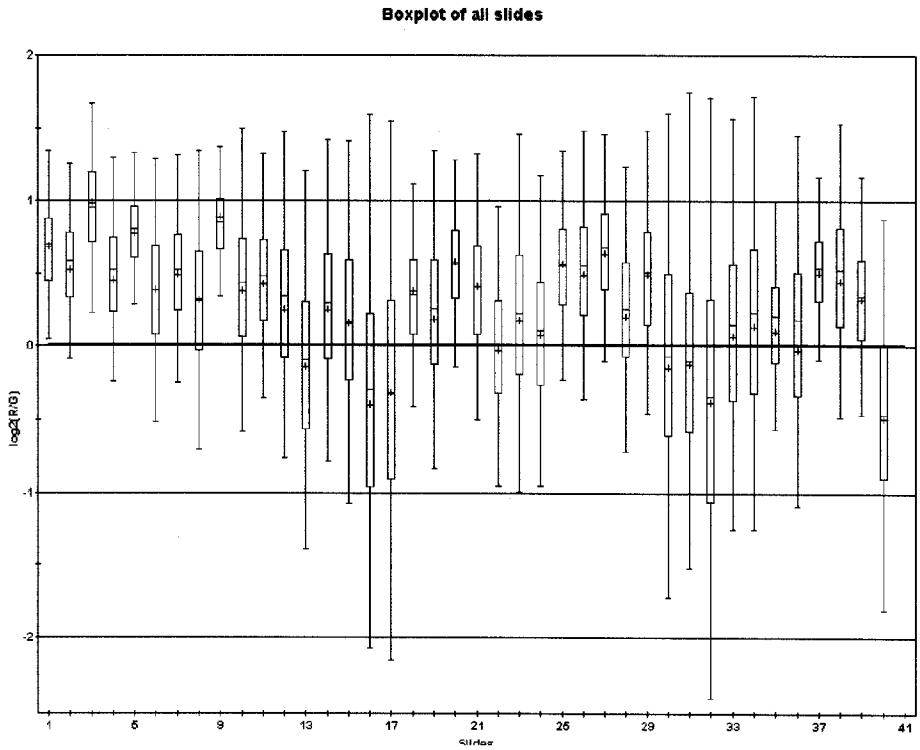


Figure 22. Box plot of 40 microarray slides (raw data)

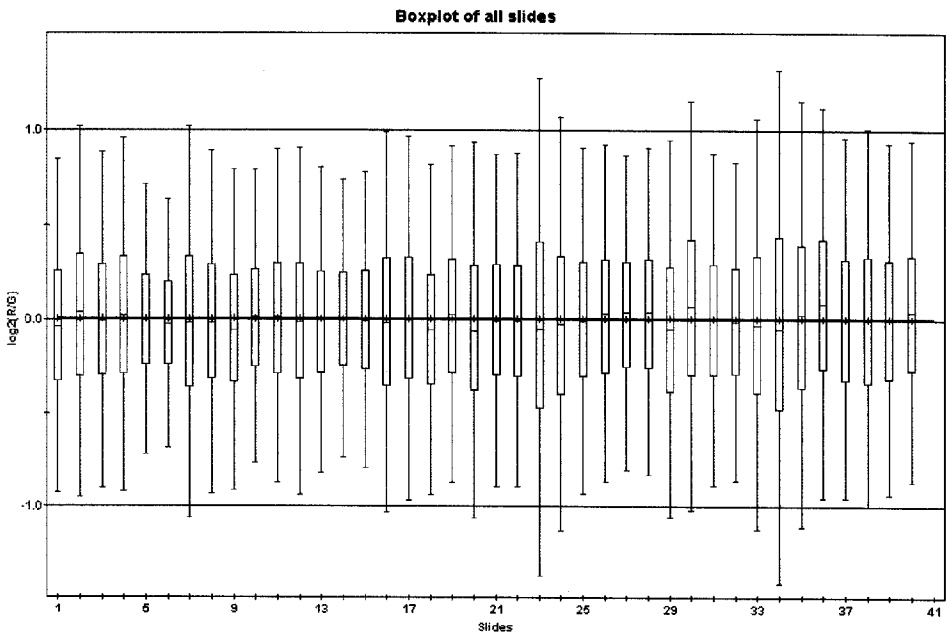


Figure 23. Box plot of 40 microarray slides (within array normalized data)

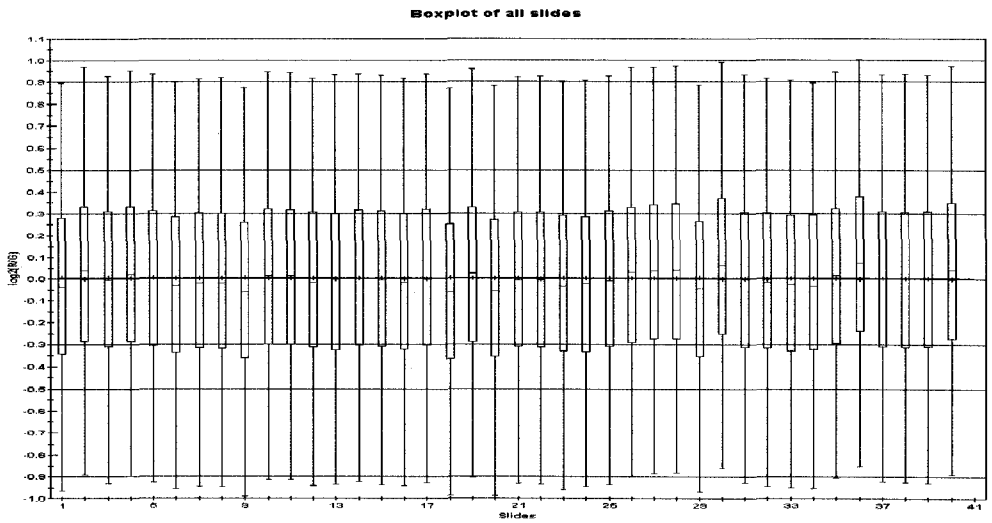


Figure 24. Box plot of 40 microarray slides (within and between array normalized data)

Significant transcripts were selected with measures of confidence based on t-test and p-value. A cut off of $p < 0.05$ and fold change > 1.5 was used. Assuming there are false positives among the differentially expressed genes, we also used Benjamini and Hochberg's false discovery rate (FDR) controlling approach (Benjamini and Hochberg 1995). We analyzed the same multiple class dataset using SAM. Venn diagram (Figure 25) was then employed to show the 109 overlapped sequences between SAM and t-test.

Figure 25. 109 overlapped sequences list between SAM and t-test

Out of the 109 common significant transcripts, 99 transcripts have blastx/tblastx matches in the GenBank non-redundant database. The 109 significant transcripts as well as their blastx results are shown in Appendix C. Among them we found several genes including 14 transcripts encoding chitinase and 7 transcripts encoding for ferritin (table 11).

Blood disorders: methemoglobinemia

Earthworms have one of the simplest blood circulatory systems. *E. fetid*a, like other annelids, possesses two completely different types of oxygen binding proteins: hemoglobins in the blood and hemerythrins in the vascular system and the coelomic fluid or in muscles (myohemerythrins). A major effect of TNT exposure is methemoglobinemia resulted from oxidation of hemoglobin (Reddy *et al.* 2000). Continued oxidation by TNT will create tissue hypoxia as the met- forms cannot bind and transport oxygen. In the gene expression experiments, we observed that expression of genes encoding ferritin, a globular protein complex and the main intracellular Fe(III)-storage protein, was down-regulated in TNT-exposed worms. Ferric iron can be reduced to Fe^{2+} in Fenton reaction to remove H_2O_2 by catalase or peroxidase (Boelsterli 2003).

Query ID	Accession Version #	Length	Score	bit	Evalue	Organism
Chitinase						
EW1_F1plate02_B05	BAD15061.1	477	77.8	190	1.00E-13	Paralichthys olivaceus
EW1_F1plate04_H08	AAH69614.1	454	117	293	1.00E-25	Homo sapiens
EW1_F1plate06_B12	AAH69614.1	454	94.7	234	1.00E-18	Homo sapiens
EW1_F1plate06_F04	AAH69614.1	454	120	301	2.00E-26	Homo sapiens
EW1_F1plate07_A11	AAH69614.1	454	120	301	2.00E-26	Homo sapiens
EW1_R1plate06_B02	AAH69614.1	454	131	329	1.00E-29	Homo sapiens
EW2_R1plate01_C02	BAC06447.1	929	82.8	203	4.00E-15	Haemaphysalis longicornis
EW2_R1plate02_H03	NP_446012.1	370	65.9	159	5.00E-10	Rattus norvegicus
EW2_R1plate03_H08	NP_001020370.1	311	73.2	178	3.00E-12	Homo sapiens
EW2_R1plate05_G04	AAH69614.1	454	120	300	4.00E-26	Homo sapiens
EW2_R1plate06_B03	AAB68960.1	497	72.8	177	8.00E-12	
EW2_R1plate06_F04	BAC06447.1	929	77	188	2.00E-13	Haemaphysalis longicornis
EW2_R1plate06_H08	BAC06447.1	929	77	188	2.00E-13	Haemaphysalis longicornis
EW2_R1plate07_D10	AAH69614.1	454	131	329	2.00E-29	Homo sapiens
ferritin						
EW1_F1plate05_B04	AAQ54709.1	172	57.4	137	2.00E-07	Amblyomma maculatum
EW2_R1plate02_G02	AAN63032.1	175	79.3	194	5.00E-14	Branchiostoma lanceolatum
EW2_R1plate03_B06	AAP83794.1	171	75.9	185	5.00E-13	Crassostrea gigas
EW2_R1plate03_C11	AAQ12076.1	206	79	193	6.00E-14	Pinctada fucata
EW2_R1plate05_C11	AAN63032.1	175	74.7	182	2.00E-12	Branchiostoma lanceolatum
EW2_R1plate05_G05	AAQ12076.1	206	70.5	171	2.00E-11	Pinctada fucata
EW2_R1plate10_C02	AAN63032.1	175	82	201	1.00E-14	Branchiostoma lanceolatum

Table 11. 14 transcripts encoding chitinase and 7 transcripts encoding for ferritin

Defense against fungal pathogens

Chitinases (EC3.2.1.14) are enzymes that catalyze the hydrolysis of the β-1,4-N-acetyl-D-glucosamine linkages in chitin polymers (Malaguarnera 2006). Chitinase may be involved in biological processes like cell wall chitin metabolism, chitin catabolism, digestion, immune response and response to fungus, and molting.

More than 10 transcripts putatively encoding human phagocyte-derived chitotriosidase (CHIT1) were prominently expressed in the earthworm and were consistently suppressed along with two other chitinase isoform genes in TNT-exposed worms. As a non-chitinous organism, worm phagocytes may produce and release the highly conserved chitinase which has been shown playing a role in defense against chitin-containing pathogens as a component of the innate immunity in human beings (van Eijk *et al.* 2005).

Confirmation of microarray results by Real time PCR[10]

To validate the microarray data, a real time PCR was performed to monitor gene expression. We selected 14 transcripts encoding chitinase and 7 transcripts encoding for ferritin. Particularly noticeable, as shown by both microarray and QPCR results in Figure 26 and 27, transcripts coding for chitinase and ferritin were consistently down-regulated in response to TNT exposure. There is a slight up-regulation at 2 mg/kg corresponding to the hormetic-like responses resulting from physical disturbances (van der Schalie and Gentile 2000).

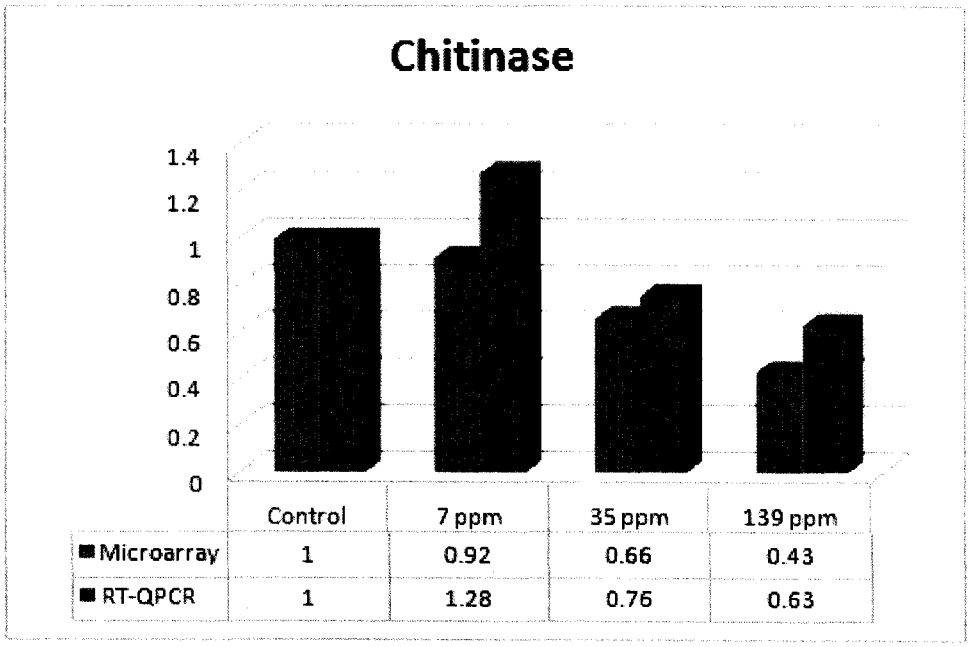


Figure 26. Microarray and QPCR expression results comparison for Chitinase

[10] Performed by ERDC (Environmental Research and Development Center) at Vicsburg, MS

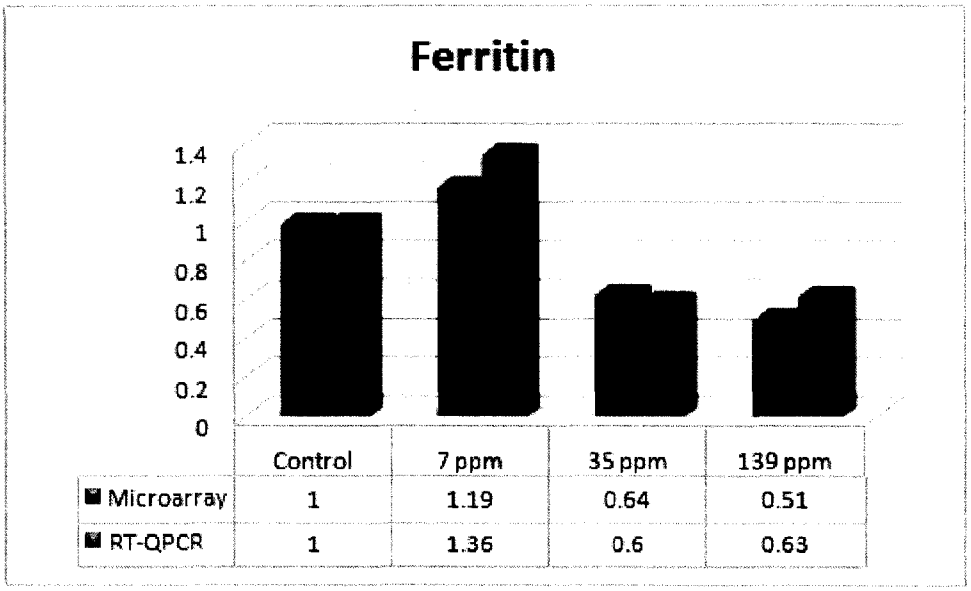


Figure 27. Microarray and QPCR expression results comparison for Ferritin

At the organism level, few significant effects such as mortality and growth were observed in the adult worms after 28-day exposure to up to 67 mg TNT/kg soil. The direct oxidation of lipids, proteins, nucleic acids can lead to cell injury or death when the oxidative stress of TNT overwhelms the antioxidant defense system (Boelsterli 2003).

In conclusion, a toxicogenomics approach was used to study molecular mechanisms involved in the sublethal toxicity of TNT in *E. fetida.* Some of the differentially expressed genes are potential candidates of new biomarkers, for which further screening and validation are required. Evidence obtained from this study strongly implies that many biological processes have been altered in response to TNT exposure, and that the affected pathways are related to blood disorders and defense mechanisms.

Efficiency of hybrid normalization of microarray gene expression: A simulation study

We created a java application, called MicroSim, with the user-friendly graphical user interface (GUI) shown in Figure 28 which allows easy evaluation of errors and the rule of self-normalization method to remove systematic errors from the experiment's data.

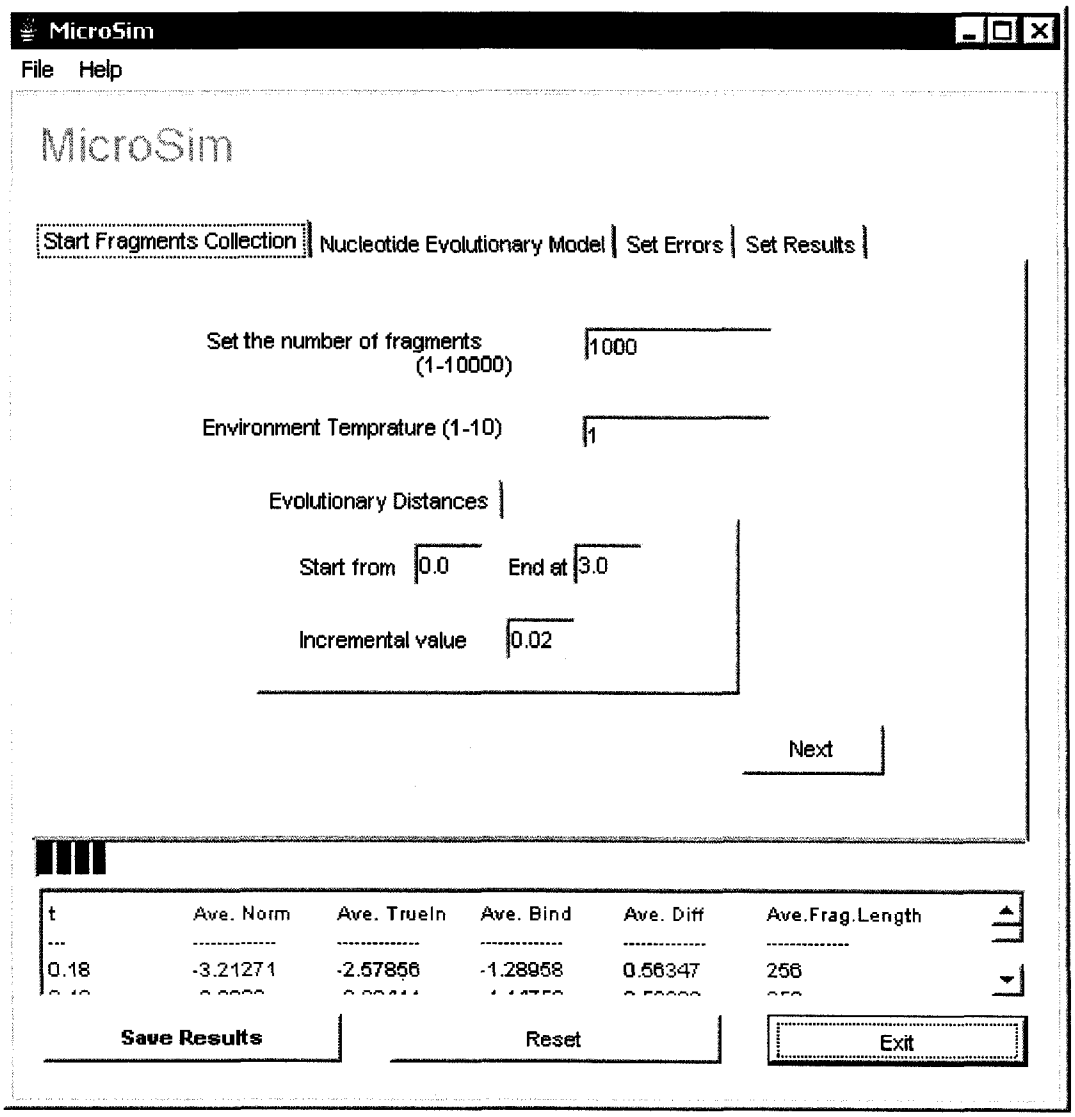


Figure 28. Main window of MicroSim

MicroSim generates a number of fragments with random lengths, and then using a range of evolutionary distances these fragments will be evolved to mutated fragments. Four nucleotide substitution models, including Jukes-Cantor, Kimura 2 parameters, Hasegawa-Kishino-Yano (HKY), and Tamura-Nei are implemented. Therefore, based on the chosen model different adjustable kappa (Transition/Transversion ratio) and base frequency will be applied. Then using start and mutated fragments as red and green spots, true intensity (logged ratio) of spots will be calculated and it would be stored. The user may adjust the additive and multiplicative errors and MicroSim computes the normalized intensity logged ratio using self-normalization method and dye-flip technique. The environment temperature, which can affect binding probability of fragments, is also adjustable.

Results can be saved into spreadsheet file that can be used directly by statistical software like Microsoft Excel ™ or SPSS to visualize and analyze data. MicroSim is able to run on any computer with java 1.2 (or higher version) runtime. It has been tested on WinXP with java 1.4.2, Linux Mandrake with java 1.2.1 and Linux Suse with java 1.4.2 virtual machine. Java language allows MicroSim to be portable on any platform. Figure 29 shows the Unified Modeling Language (UML) Class Diagram of the application. MicroSim is available for download from: http://mcbc.usm.edu/microsimapp.jar

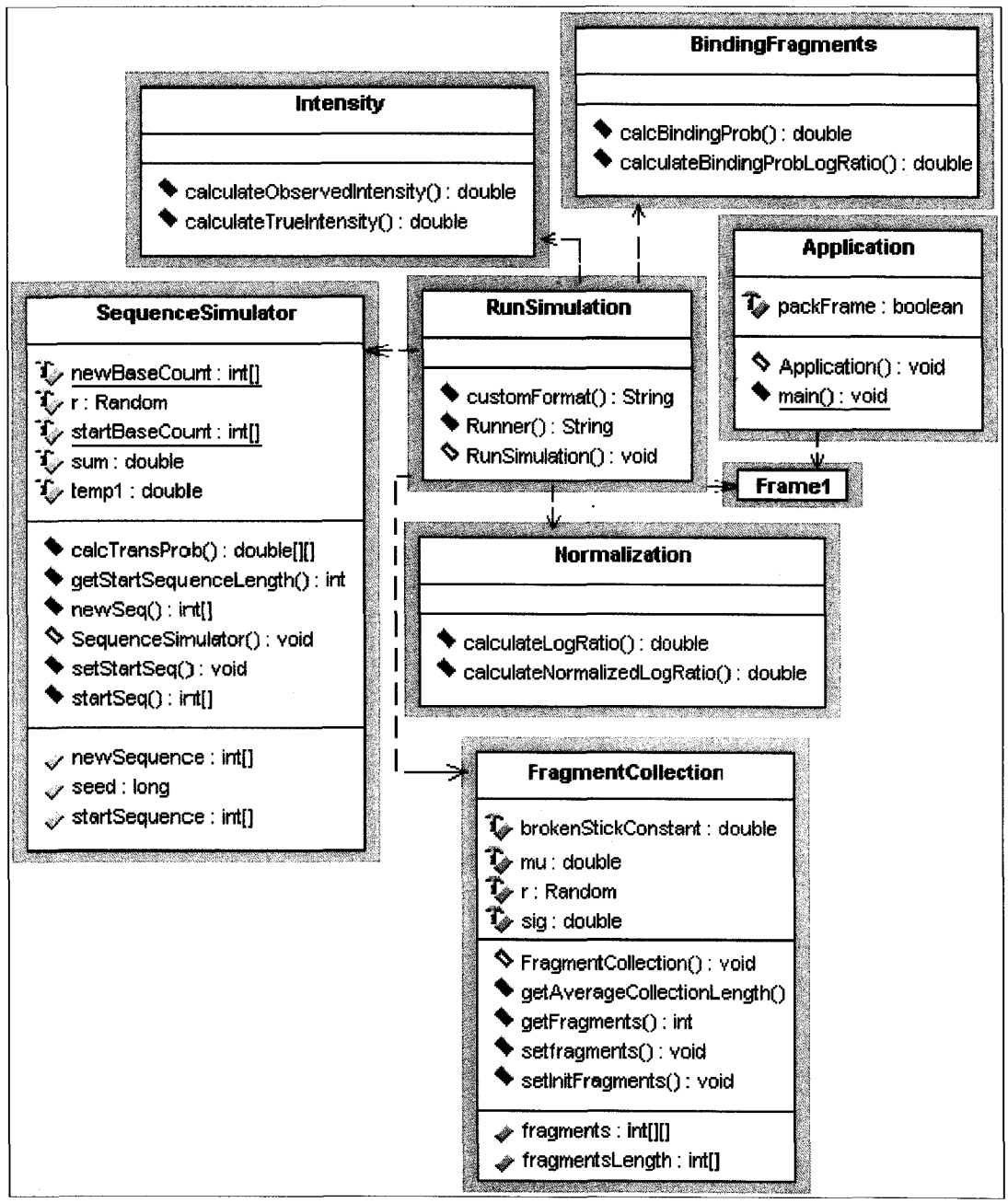


Figure 29. MicroSim UML Class Diagram

Using the nucleotide substitution method, simulated start sequences were evolved to the mutated sequences. Considering start and mutated sequence, as a red and green spot, binding probability and true intensity log ratio were calculated. The graph in Figure

30 clearly shows that there is a significant decrease of true intensity log ratio when evolutionary distance (t) increases.

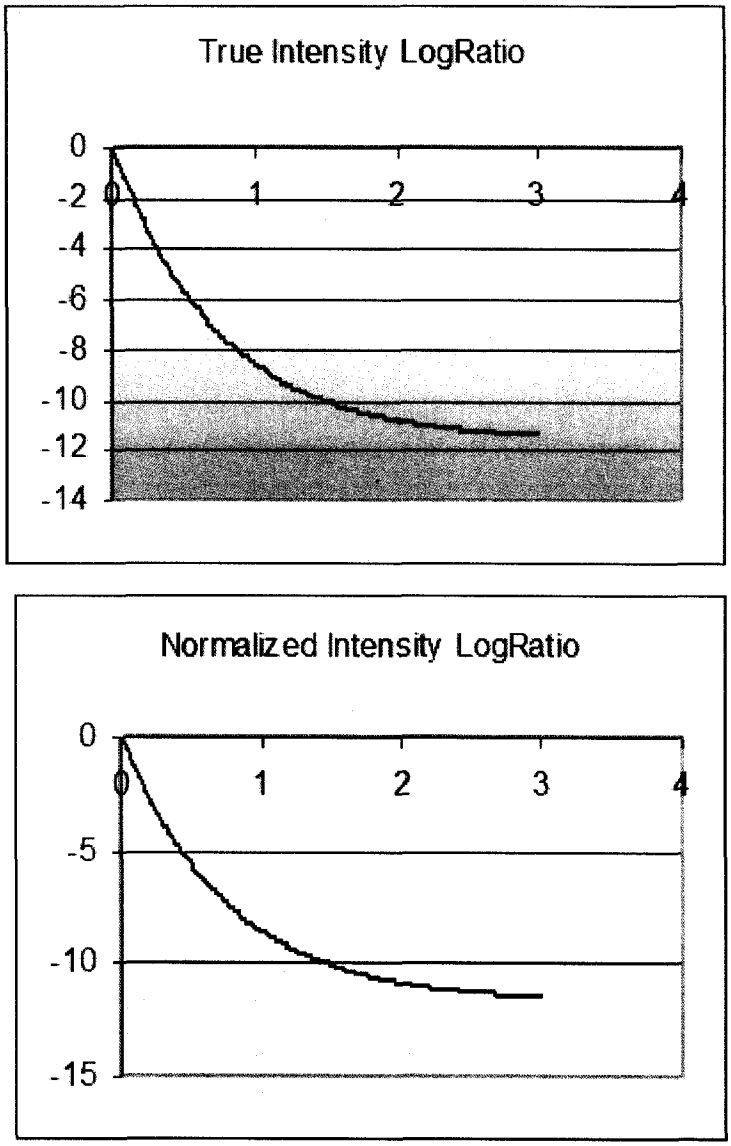


Figure 30. Dye-swap normalization: plot of comparison of true intensity log ratio (before normalization) (top graph) and normalized intensity log ratio (bottom graph). x axis is evolutionary distance (t) and y is the intensity log ratio. Kimura 2p model, kappa = 2, Temperature = 1, and Fragments numbers = 1000 with default errors value of application is applied.

Observed intensity was calculated by adding additive and multiplicative errors to the true intensity log ratio. The average of normalized log ratio was computed from true intensity by applying dye-flip normalization. The lower graph in Figure 30 shows the average of normalized intensity log ratio against evolutionary distances. It has also been plotted in Figure 31, which shows the average standard deviation of 0.05. Similarity in both pattern and value of true and normalized intensity log ratio can prove the efficiency of dye-swap normalization technique.

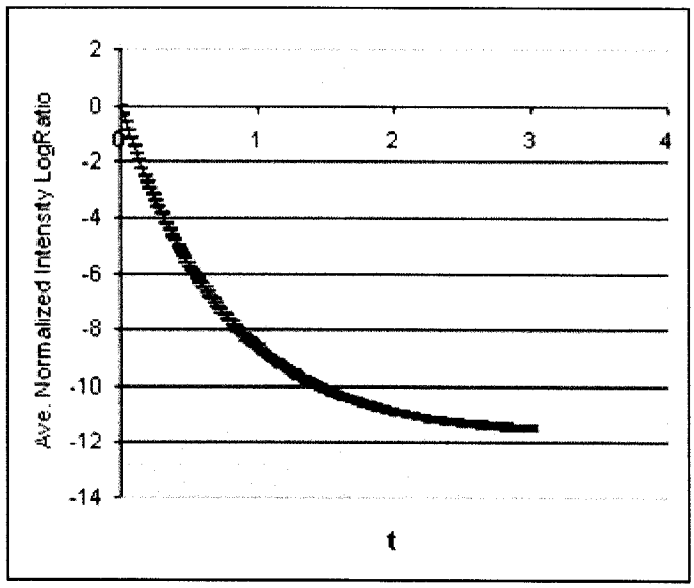


Figure 31. Plot of average of normalized intensity log ratio with 0.08<Standard Deviation<0.02 (t is the evolutionary distance between start and mutated fragments) - Jukes-Cantor model, Temperature = 1, and Fragments numbers = 1000 with application default errors value

In order to assess the efficiency, effect of temperature was also studied by applying different values of temperature to the experiment and calculating the average of binding probability log ratio. As it is shown in Figure 32, the average of the log ratio of the binding probability will be close to zero when the temperature increases.

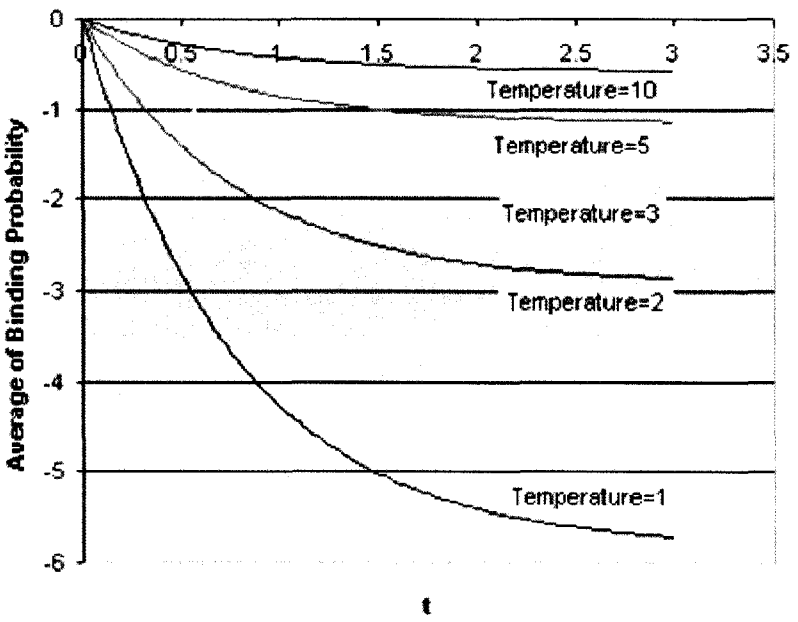


Figure 32. Plots of average of binding probability log ratio with different temperatures. (t is the evolutionary distance between start and mutated fragments) - Jukes-Cantor model and Fragments numbers = 1000 with default errors value are applied

We also studied the effect of kappa (transition/transversion ratio) by applying different values of kappa in Hasegawa-Kishino-Yano (HKY) model, which has been plotted in Figures 33 and 34. The results clearly show an increase in binding probability (Figure 34), and as a result of it, an increase in true and normalized intensity (Figure 33) log ratio when kappa increases. Tamura-Nei model were simulated with different base frequencies. The Graph in Figure 35 shows the plot of average normalized log ratios with different base frequency. The results show an increase in binding probability, true and normalized intensity log ratios when frequency of base T increases.

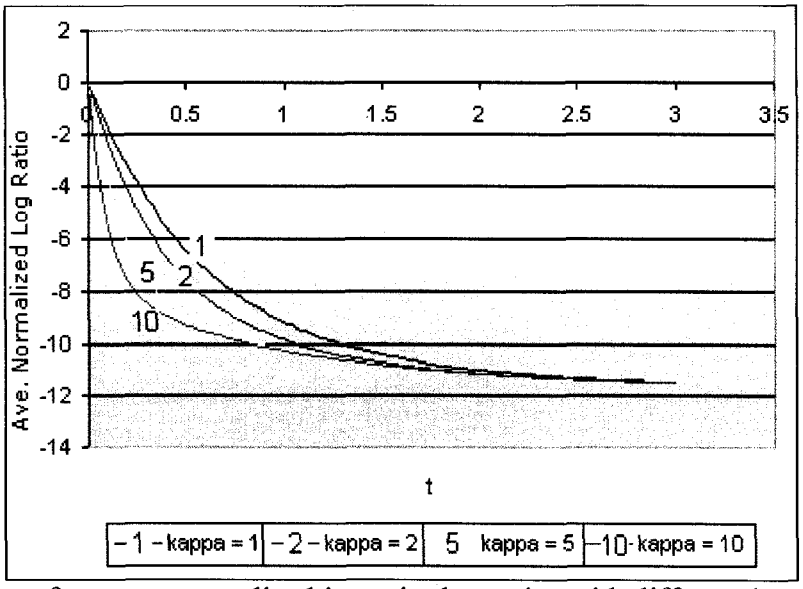


Figure 33. Plot of average normalized intensity log ratios with different kappa in HKY. (t is the evolutionary distance between start and mutated fragments) - HKY model, equal base frequency (25% for each) and Fragments numbers = 1000 with default errors value are applied

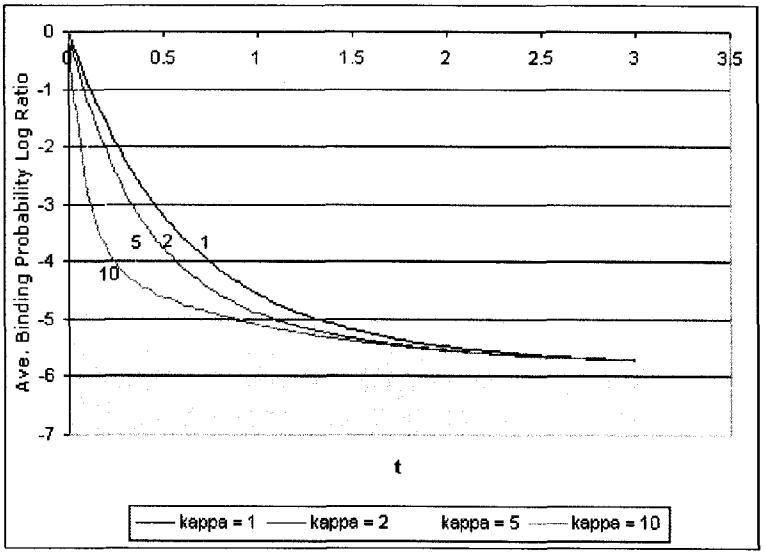


Figure 34. Plot of average binding probability log ratios with different kappa in HKY. (t is the evolutionary distance between start and mutated fragments) - HKY model, equal base frequency (25% for each) and Fragments numbers = 1000 with default errors value are applied

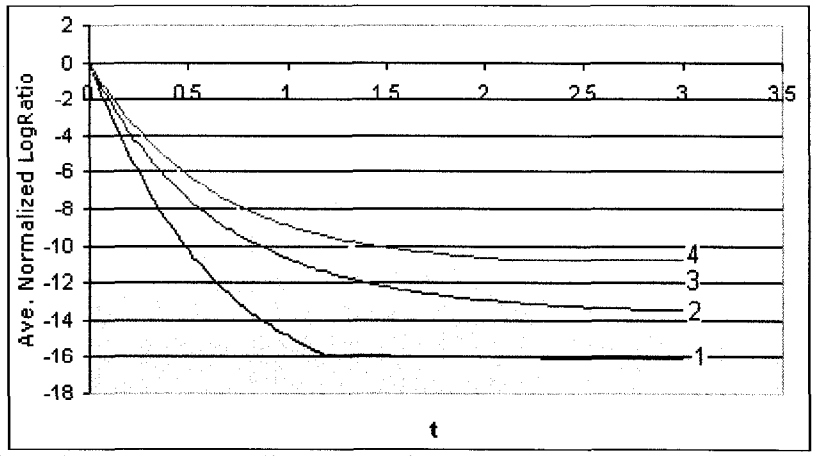


Figure 35. Plot of average normalized intensity log ratios with different base frequencies in Tamura-Nei Model. Fragments numbers = 1000 with default errors value are applied (t is the evolutionary distance between start and mutated fragments) - (1) Base frequency: A=10%, C=40%, G=40%, T=10% (2) Base frequency: A=19%, C=31%, G=31%, T=19% (3) Base frequency: A=25%, C=25%, G=25%, T=25% (4) Base frequency: A=35%, C=15%, G=15%, T=35%

GeneVenn – A Web Application for Comparing Gene Lists Using Venn Diagrams

Simple Venn diagrams are already being used in microarray data analysis software packages such as commercial GeneSpring® and SilicoCyte® or open source R-package limma to visualize intersections of up to three different lists of genes.

We proposed a web application creating Venn diagrams from two or three gene lists. It has been graphically designed and publicly available at http://mcbc.usm.edu/genevenn/.

The design of GeneVenn follows that of a two tier Web application. The UML class diagram including application's class variables and methods is illustrated in Figure 36.

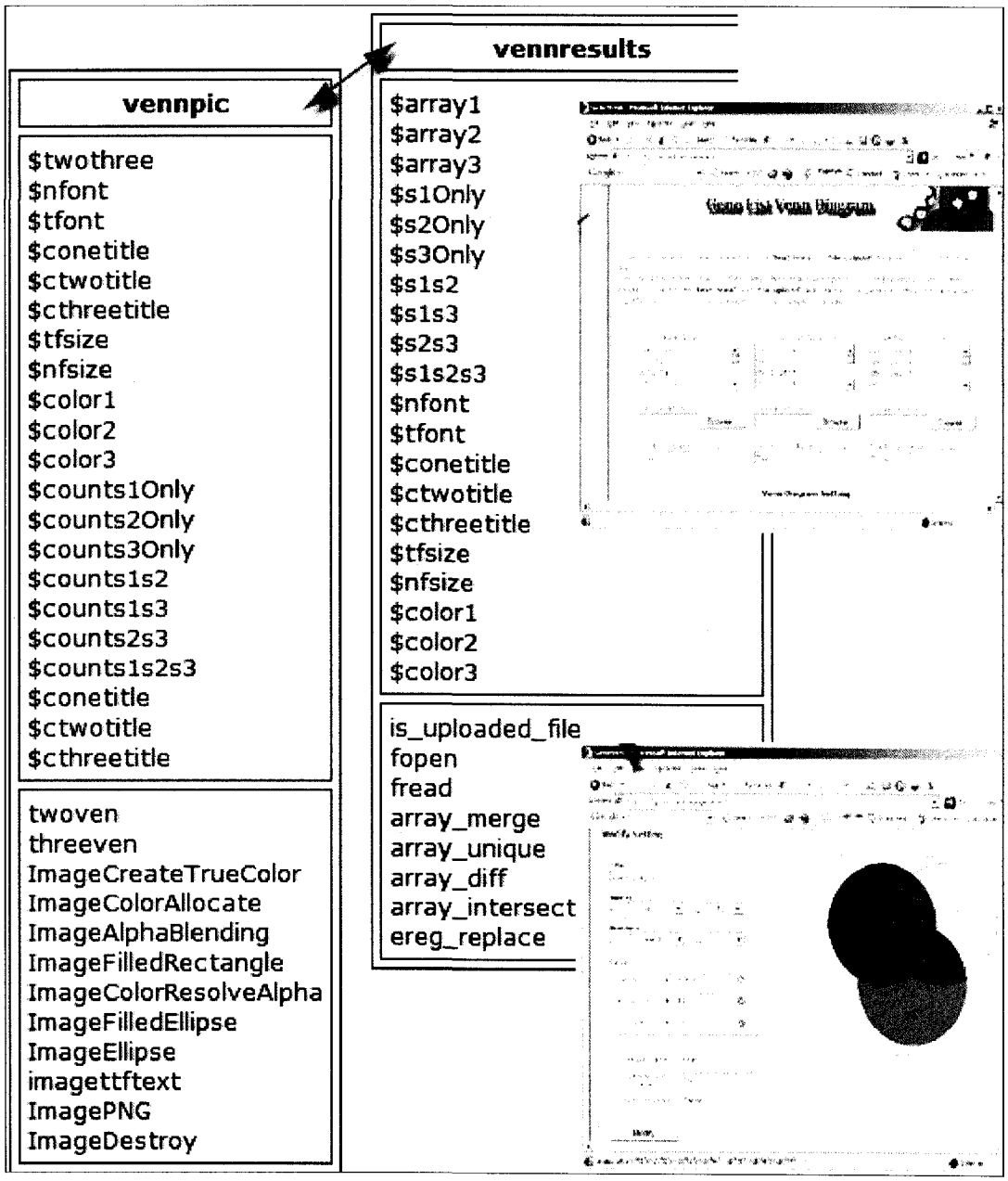


Figure 36. GeneVenn UML Class Diagram

The user-friendly web interface has been developed by using the PHP language, DHTML, and JavaScript. The application is currently running under an Apache web server version 2.2 based on Linux Suse 10.2 OS.

GeneVenn is relatively small but, nonetheless, effectively complete. In this, the initial welcome page has three text lists. A user is able to enter the gene names into these

text areas as well as upload gene list files to the server and process them automatically. If the user enters data in the text box and uploads a gene file for the same list, data will be merged and considered as a single list. It is also possible to select the number of diagrams (two or three), set up a name for each diagram, and a title for the results. Any white space such as a tab, space, line break or comma is accepted as a gene name delimiter. The result page processes the lists and creates a Venn diagram. Each area on the diagram has a hyperlink which shows the related gene list, and each gene name is linked to the related information in NCBI's Entrez Nucleotide database. Here, the user is able to modify every element of the generated diagram including font, color and name of the diagram.

A Comparative Study of Different Machine Learning Methods on Microarray Data

We compared the efficiency of the classification methods; SVM, RBF Neural Nets, MLP Neural Nets, Bayesian, Decision Tree and Random Forrest methods. We used v-fold cross validation methods to calculate the accuracy of the classifiers. We also applied some common clustering methods such as K-means, DBC, and EM clustering to our data and analyzed the efficiency of these methods.

Further, we compared the efficiency of the feature selection methods: support vector machine recursive feature elimination (SVM-RFE) (Duan *et al.* 2005; Guyon *et al.* 2002), Chi Squared (Liu and Setiono 1995), and CSF (Hall 1998; Wang *et al.* 2005). In each case these methods were applied to eight different binary (two class) microarray datasets. We evaluated the class prediction efficiency of each gene list in training and test cross-validation using our supervised classifiers. After features selection, their

efficiencies are investigated by comparing the error rate of classification algorithms applied to only these selected features versus all features.

The bioinformatics techniques studied in this project are representative of general-purpose data-mining techniques. We presented an empirical study in which we compare some of the most commonly used classification, clustering, and feature selection methods. We applied these methods to eight publicly available datasets, and compared how these methods perform in class prediction of test datasets. We reported that the choice of feature selection method, the number of genes in the gene list, the number of cases (samples) and the noise in the dataset substantially influence classification success. Based on features chosen by these methods, error rates and accuracy of several classification algorithms were obtained. Results reveal the importance of feature selection in accurately classifying new samples. The integrated feature selection and classification algorithm is capable of identifying significant genes.

Table 12 shows eight data sets used in this work.

Dataset	**Comparison**	**Variables (Genes)**	**Samples**
1. Lymphoma (Devos *et al.* 2002)	Tumor vs Normal	7129	25
2. Breast Cancer (Perou *et al.* 2000)	Tumor subtype vs Normal	1753	84
3. Colon Cancer (Alon *et al.* 1999)	Epithelial vs Tumor	7464	45
4. Lung Cancer (Garber *et al.* 2001)	Tumor vs Normal	917	72
5. Adenocarcinoma (Beer *et al.* 2002)	NP vs NN	5377	86
6. Lymphoma (Alizadeh *et al.* 2000)	DLBCL1 vs DLBCL2	4027	96
7. Melanoma (Bittner *et al.* 2000)	Tumor vs Normal	8067	38
8. Ovarian Cancer (Welsh *et al.* 2001)	Tumor vs Normal	7129	39

Table 12. Eight datasets used in the comparison experiment

Each dataset is publicly available and data were downloaded from microarray repositories from the caGEDA website from the University of Pittsburgh (Patel and Lyons-Weiler 2004):

- Lymphoma (De Vos *et al.* 2002), contains 25 samples of which came from normal vs. malignant plasma cells including 7129 genes
- Breast Cancer (Perou *et al.* 2000), 84 samples of normal vs. tumor subtypes including 1753 genes
- Colon Cancer (Alon *et al.* 1999), 45 samples of Epithelial normal cells vs. Tumor cells including 7464 genes
- Lung Cancer (Garber *et al.* 2001), contains 72 samples of which came from normal vs. malignant cells including 917 genes
- Adenocarcinoma (Beer *et al.* 2002), contains 86 samples of which came from survival in early-stage lung adenocarcinomas including 5377 genes
- Lymphoma (Alizadeh *et al.* 2000), 96 samples of DLBCL1 vs. DLBCL2 cells including 4027 genes
- Melanoma (Bittner *et al.* 2000), 38 samples of normal vs. malignant cells including 8067 genes
- Ovarian Cancer (Welsh *et al.* 2001), 39 samples of normal vs. malignant cells including 7129 genes

Preprocessing

We applied three steps of preprocessing to the datasets. First we applied baseline shift for the datasets by shifting all measurements upwards by a number of means (or averages). This process is then followed by performing global mean adjustment. The global mean of all intensities of all datasets is calculated. Then the difference between each individual mean and the global mean is calculated. This difference value is then

added to (or subtracted from) each individual expression intensity value on each dataset. The result is that all datasets now have the same overall mean.

Finally a log transformation is applied to the datasets. Log transformation has the advantage of producing a continuous spectrum of values.

Classification

We used Weka (Frank *et al.* 2004) and SVM Classifier (Pirooznia and Deng 2006) for applying classification, clustering and feature selection methods to our datasets. In house java program was used to convert dataset from delimited file format, which is the default import format for SVM Classifier, to ARFF (Attribute-Relation File Format) file, the import format for Weka. For the SVM we applied the following procedures. First we transformed data to the format of SVM software, ARFF for WEKA and Labeled for SVM Classifier. Then we conducted simple scaling on the data. We applied linearly scaling of each attribute to the range [−1, +1] or [0, 1].

We considered the RBF kernel and used cross-validation to find the best parameter C and γ. We used a "grid-search" (Chang and Lin 2001) on C and γ using cross-validation. Basically pairs of (C, γ) are tried and the one with the best cross-validation accuracy is picked. Trying exponentially growing sequences of C and γ is a practical method to identify good parameters, for example $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$.

The classification methods were first applied to all datasets without performing any feature selection. Results of 10-fold cross validation have been shown in Figure 37 and Table 13. In most datasets SVM and RBF neural nets performed better than other classification methods. In breast cancer data, SVM classification and RBF Neural Nets

had the best accuracy 97.6%, and overall they performed very well on all datasets. The minimum accuracy for RBF we calculated was 81.6% over the melanoma dataset. In the lung cancer dataset MLP Neural Nets also performed well and it was equal to SVM and RBF.

The lowest accuracies were detected from Decision Tree algorithms (both J48 and ID3). As it is shown in Figure 37, in most cases they performed poorly compared to other methods.

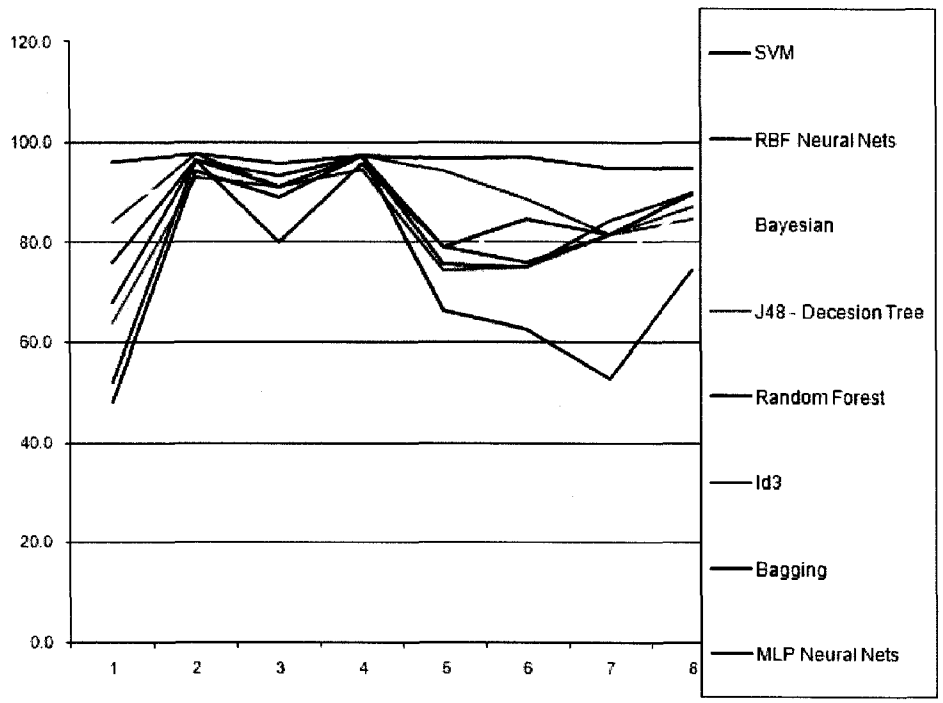


Figure 37. Percentage accuracy of 10-fold cross validation of classification methods for all genes

Bayesian methods also had high accuracy in most datasets. It didn't perform as well as SVM and RBF, with the lowest accuracy being 85.4% on Lymphoma datasets.

However, overall we have to mention that it seems in some cases performance of the classification methods depends on the dataset and a specific method cannot be concluded as a best method. For example, Bayesian and J48 Decision Tree performed

very well on colon and lung cancer, with 93% and 95% for Bayesian respectively and 91% and 94 % for J48, while RBF and MLP out performed on breast and lung cancer (97% and 96% respectively for MLP and 97% for both datasets for RBF).

We applied two class clustering methods to the datasets that are illustrated in Figure 38 and Table 16. As it is shown in Figures 38, we have a consistent performance of Farthest First in almost all datasets. EM performed poorly in Adenocarcinoma and Lymphoma datasets (54.7 and 54.2 respectively) while it was performing well in breast melanoma (81%).

The effect of feature selection

Pairwise combinations of the feature selection and classification methods were examined for different samples as it is shown in Tables 15 and 16 and Figure 38. The procedure is illustrated as a pipeline in Figure 39.

First we tested SVM-RFE, Correlation based, and Chi Squared methods on several gene numbers (500, 200, 100, and 50). Methods were mostly consistent when gene lists of the top 50, 100, or 200 genes were compared. We selected 50 genes because it performed well, consumed less processing time, and required less memory configurations comparing to others.

In almost all cases, the accuracy performance classifiers were improved after applying feature selection methods to the datasets. In all cases SVM-RFE performed very well when it applied with SVM classification methods.

In the lymphoma dataset SVM-RFE performed 100% in combination of SVM classification method. The Bayesian classification method performed well for SVM-RFE and Chi Squared feature selection methods with 92% accuracy in both cases.

CFS and Chi Squared also improved the accuracy of the classification. In the breast cancer dataset the least improvement is observed from applying Chi Squared feature selection methods with no improvement over SVM, RBF and J48 classification methods with 97%, 84%, and 95% respectively.

In the ovarian cancer dataset all feature selection methods performed closely. However the SVM-RFE had a slightly better performance comparing to other methods. We detected 100% accuracy with SVM-RFE feature selection with both SVM and RBF classification methods. We also observed high accuracies among MLP classification and all feature selection methods with 94%, 92%, and 92% for SVM-RFE, CFS, and Chi Squared respectively.

In the lung cancer datasets we can observe high accuracy in the Decision Tree classification methods (both J48 and ID3) with all feature selection methods.

Overall, we have to state again that although it is obvious that applying feature selection methods improve the accuracy and also particularly reduce the processing time and memory usage, but finding the best combination of feature selection and classification method might vary in each case.

1. Lymphoma (De vos et.al, 2002)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	0:0	0:1	2:3	1:1	1:2	1:1	4:5	3:6
CFS	50	1:0	2:1	3:3	2:1	4:4	2:2	3:6	3:6
ChiSquared	50	1:0	2:2	4:3	1:1	3:4	2:3	2:3	4:1
All features	7129	1:0	2:2	4:4	2:1	4:5	2:4	7:6	9:3
2. Breast (Perou et. al, 2000)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	0:0	1:0	1:1	3:1	4:1	1:1	2:1	1:0
CFS	50	1:0	1:0	2:1	3:2	3:1	1:1	1:1	1:0
ChiSquared	50	1:1	1:1	1:1	2:2	3:1	1:0	1:1	1:0
All features	1753	1:1	1:1	2:1	4:2	4:2	2:1	4:1	2:1
3. Colon (Alon et. al, 1999)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	0:0	0:1	1:0	1:0	2:0	3:1	1:1	1:1
CFS	50	1:1	2:1	1:1	2:0	1:1	2:2	1:1	1:0
ChiSquared	50	1:0	2:2	2:0	1:0	1:0	1:1	2:1	1:0
All features	7464	2:0	2:2	2:2	3:0	2:2	6:3	3:2	2:1
4. Lung (Garber et. al, 2001)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	0:0	0:1	1:0	1:1	1:1	1:0	1:0	1:0
CFS	50	1:1	1:1	1:0	1:0	1:1	1:1	1:1	1:1
ChiSquared	50	1:0	1:0	1:0	1:1	2:1	2:0	1:1	1:1
All features	917	2:0	2:0	1:1	2:1	2:2	2:1	1:1	1:1
5. Adenocarc. (Beer et.al, 2002)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	0:0	2:1	2:3	4:5	4:5	3:6	4:5	3:6
CFS	50	1:0	1:1	3:3	3:6	3:6	3:6	3:6	3:6
ChiSquared	50	1:0	2:2	4:3	5:5	3:5	5:5	2:3	5:5
All features	5377	2:1	3:2	15:6	15:6	15:7	14:4	17:13	12:6
6. Lymphoma (Alizadeh et al, 2000)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	0:0	0:1	2:3	4:5	4:5	3:6	4:5	3:6
CFS	50	1:1	2:3	3:3	3:6	3:6	3:6	3:6	3:6
ChiSquared	50	1:1	2:2	4:3	5:5	3:5	5:5	2:3	5:5
All features	4027	2:1	9:2	15:7	12:2	14:10	16:7	21:15	12:3
7. Melanoma (Bittner et. al, 2000)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	0:0	0:1	2:1	2:1	3:1	3:1	4:5	3:1
CFS	50	1:0	2:3	2:2	2:2	2:1	2:1	3:6	3:2
ChiSquared	50	1:0	2:2	3:2	2:3	2:2	2:2	2:3	3:2
All features	8067	2:0	4:3	4:2	6:3	4:3	4:3	15:3	5:2
8. Ovarian (Welsh et. al, 2001)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	0:0	0:1	1:1	1:1	1:1	1:1	2:1	3:1
CFS	50	1:0	3:2	1:2	1:1	1:1	1:1	2:2	2:1
ChiSquared	50	1:0	2:2	2:1	1:1	1:1	1:1	2:3	1:1
All features	7129	2:0	4:2	2:2	3:2	3:2	2:2	7:3	3:1

Table 13. 10-fold cross validation evaluation result of feature selection methods applied to the classification methods. X:Y pattern indicates X as the error rate in cancer samples and Y as the error rate in normal samples

1. Lymphoma (De vos et.al, 2002)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	100.00	96.00	80.00	92.00	88.00	92.00	64.00	64.00
CFS	50	96.00	88.00	76.00	88.00	68.00	84.00	64.00	64.00
ChiSquared	50	96.00	84.00	72.00	92.00	72.00	80.00	80.00	80.00
All features	7129	96.00	84.00	68.00	88.00	64.00	76.00	48.00	52.00
2. Breast (Perou et. al, 2000)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	100.00	98.81	97.62	95.24	94.05	97.62	96.43	98.81
CFS	50	98.81	98.81	96.43	94.05	95.24	97.62	97.62	98.81
ChiSquared	50	97.62	97.62	97.62	95.24	95.24	98.81	97.62	98.81
All features	1753	97.62	97.62	96.43	92.86	92.86	96.43	94.05	96.43
3. Colon (Alon et. al, 1999)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	100.00	97.78	97.78	97.78	95.56	91.11	95.56	95.56
CFS	50	95.56	93.33	95.56	95.56	95.56	95.56	95.56	97.78
ChiSquared	50	97.78	91.11	95.56	97.78	97.78	95.56	93.33	97.78
All features	7464	95.56	91.11	91.11	93.33	91.11	80.00	88.89	93.33
4. Lung (Garber et. al, 2001)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	100.00	98.61	98.61	97.22	97.22	98.61	98.61	98.61
CFS	50	97.22	97.22	98.61	98.61	97.22	97.22	97.22	97.22
ChiSquared	50	98.61	98.61	98.61	97.22	95.83	97.22	97.22	97.22
All features	917	97.22	97.22	97.22	95.83	94.44	95.83	97.22	97.22
5. Adenocarc. (Beer et.al, 2002)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	100.00	96.51	94.19	89.53	89.53	89.53	89.53	89.53
CFS	50	98.84	97.67	93.02	89.53	89.53	89.53	89.53	89.53
ChiSquared	50	98.84	95.35	91.86	88.37	90.70	88.37	94.19	88.37
All features	5377	96.51	94.19	75.58	75.58	74.42	79.07	66.28	79.07
6. Lymphoma (Alizadeh et al, 2000)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	100.00	100.00	94.79	90.63	90.63	90.63	90.63	90.63
CFS	50	97.92	94.79	93.75	90.63	90.63	90.63	90.63	90.63
ChiSquared	50	97.92	95.83	92.71	89.58	91.67	89.58	94.79	89.58
All features	4027	96.88	88.54	77.08	85.42	75.00	76.04	62.50	84.38
7. Melanoma (Bittner et. al, 2000)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	100.00	97.37	92.11	92.11	89.47	89.47	76.32	89.47
CFS	50	97.37	86.84	89.47	89.47	92.11	92.11	76.32	86.84
ChiSquared	50	97.37	89.47	86.84	86.84	89.47	89.47	86.84	86.84
All features	8067	94.74	81.58	84.21	76.32	81.58	81.58	52.63	81.58
8. Ovarian (Welsh et. al, 2001)	# Genes	SVM	RBF	MLP	Bayesian	J48	ID3	R. Forest	Bagging
SVM-RFE	50	100.00	100.00	94.87	94.87	94.87	94.87	92.31	89.74
CFS	50	97.44	87.18	92.31	94.87	94.87	94.87	89.74	92.31
ChiSquared	50	97.44	89.74	92.31	94.87	94.87	94.87	87.18	94.87
All features	7129	94.87	84.62	89.74	87.18	87.18	89.74	74.36	89.74

Table 14. Percentage accuracy of 10-fold cross validation of feature selection methods applied to the classification methods.

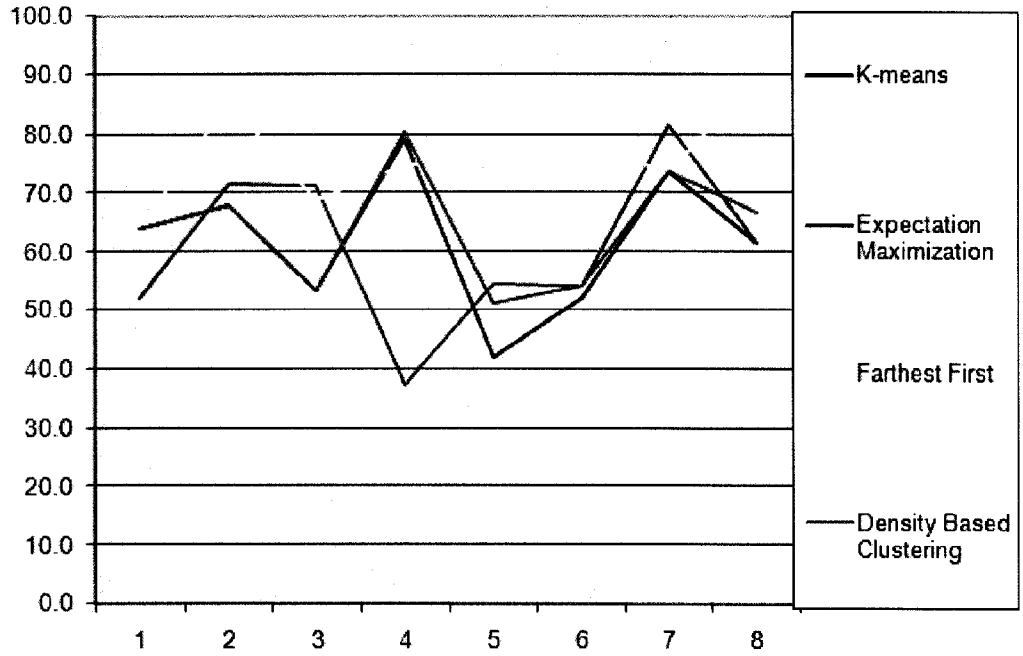


Figure 38. Percentage accuracy of 10-fold cross validation of clustering methods for all genes

Dataset	SVM	RBF	MLP	Bayesian	J48	Random Forest	Id3	Bagging
1. Lymphoma (Devos et.al, 2002)	96.0	84.0	68.0	88.0	64.0	76.0	48.0	52.0
2. Breast Cancer (Perou et. al, 2000)	97.6	97.6	96.4	92.9	92.9	96.4	94.0	96.4
3. Colon Cancer (Alon et. al, 1999)	95.6	91.1	91.1	93.3	91.1	80.0	88.9	93.3
4. Lung Cancer (Garber et. al, 2001)	97.2	97.2	97.2	95.8	94.4	95.8	97.2	97.2
5. Adenocarcinoma (Beer et.al, 2002)	96.5	94.2	75.6	75.6	74.4	79.1	66.3	79.1
6. Lymphoma (Alizadeh et al, 2000)	96.9	88.5	75.0	85.4	75.0	76.0	62.5	84.4
7. Melanoma (Bittner et. al, 2000)	94.7	81.6	84.2	76.3	81.6	81.6	52.6	81.6
8. Ovarian Cancer (Welsh et. al, 2001)	94.9	84.6	89.7	87.2	87.2	89.7	74.4	89.7

Table 15. Percentage accuracy of 10-fold cross validation of classification methods for all genes

Dataset	K-means	Expectation Maximization	Farthest First	Density Based Clustering
1. Lymphoma (Devos et.al, 2002)	64.0	52.0	64.0	64.0
2. Breast Cancer (Perou et. al, 2000)	67.9	71.4	85.7	67.9
3. Colon Cancer (Alon et. al, 1999)	53.3	71.1	68.9	53.3
4. Lung Cancer (Garber et. al, 2001)	79.2	37.5	75.0	80.6
5. Adenocarcinoma (Beer et.al, 2002)	42.0	54.7	74.4	51.2
6. Lymphoma (Alizadeh et al, 2000)	52.1	54.2	78.1	54.2
7. Melanoma (Bittner et. al, 2000)	73.7	81.6	73.7	73.7
8. Ovarian Cancer (Welsh et. al, 2001)	61.5	61.5	89.7	66.7

Table 16. Percentage accuracy of 10-fold cross validation of clustering methods for all genes

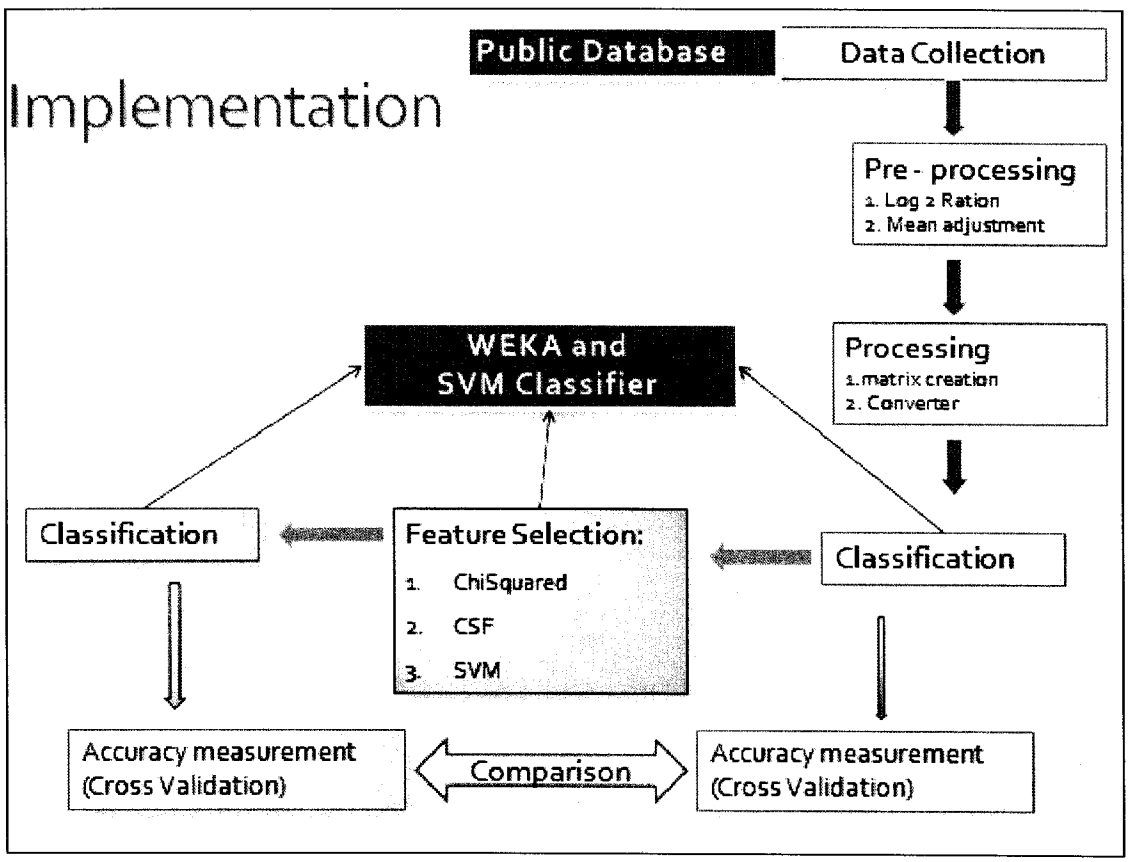


Figure 39. Overview of the machine learning comparison pipeline

SVM Classifier – A Java Interface for Support Vector Machine Classification of Microarray Data

We have developed a comprehensive graphical interface to implement the SVM algorithm, called SVM Classifier. This interface allows novice users to download the software for local installation and easily apply a sophisticated machine learning algorithm, support vector machine, to their data. We implemented a publicly accessible application that allows SVM users to perform SVM training, classification and prediction.

For details on using the software, sample dataset and explanations of the underlying algorithms, we refer readers to the project home page which is available at: http://mcbc.usm.edu/svm/.

SVM users might also be interested in a number of other licensed SVM implementations that have been described previously, including LIBSVM (Chang and Lin, 2001).

We used the SVM algorithms implemented by the Libsvm team, as a core. In order to maximize cross-platform compatibility SVM Classifier is implemented in java using standard swing libraries (Figure 40).

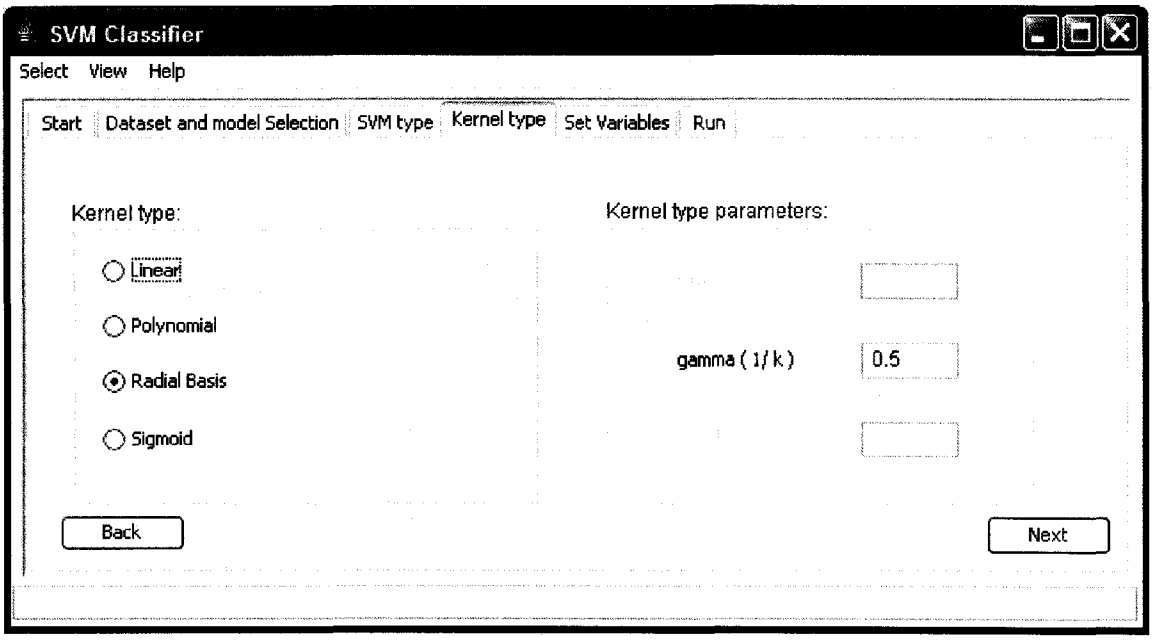


Figure 40. GUI of SVM Classifier

The open source, cross-platform Apache Ant and free edition of Borland JBuilder 2005 Foundation are used as building tools. Although developed on WinXP OS, SVM Classifier has been successfully tested on Linux and other Windows platforms, and will run on Mac OS9 with the Swing extension. Users are able to run SVM Classifier on any computer with java 1.4 runtime or higher version.

The application has two frames, the classification and the prediction frame. In both frames the data file format can be imported either as a labeled or delimited data file format.

In the classification frame the user will create a model from the training dataset for classification (C-SVC, nu-SVC), regression (epsilon-SVR, nu-SVR), and distribution estimation. In this frame the user is able to import the training dataset into the application, select the path to save the model file, select the appropriate SVM and kernel type and create a model for the dataset. The model file can be later used for prediction

purposes. There is also a choice for cross validation. The cross validation (CV) technique is used to estimate the accuracy of each parameter combination in the specified range and helps us to decide the best parameters for classification problems.

In the prediction frame the model will be applied to the test data to predict the classification of unknown data. We have also provided a tool for viewing the two dimensional data that can be accessed from the view menu bar.

We have demonstrated that support vector machines can accurately classify genes into functional categories based on expression data from DNA microarray hybridization experiments. Among the different kernel functions that we examined, the SVM that uses a radial basis kernel function provides the best performance.

We presented an evaluation of the different classification techniques presented previously. Data from Hedenfalk, *et al.* (Hedenfalk *et al.* 2001) is used in this study. The data consists of 22 cDNA microarrays, each representing 3226 genes based on biopsy specimens of primary breast tumors of seven patients with germ-line mutations of BRCA1, eight patients with germ-line mutations of BRCA2, and seven with sporadic cases. We took log2 of the data to perform the classification using the three kernels. We have achieved 100% accuracy in classification among the BRCA1-BRCA2 samples with RBF kernel of SVM. RBF kernel also shows better performance among all data as shown in Figure 41.

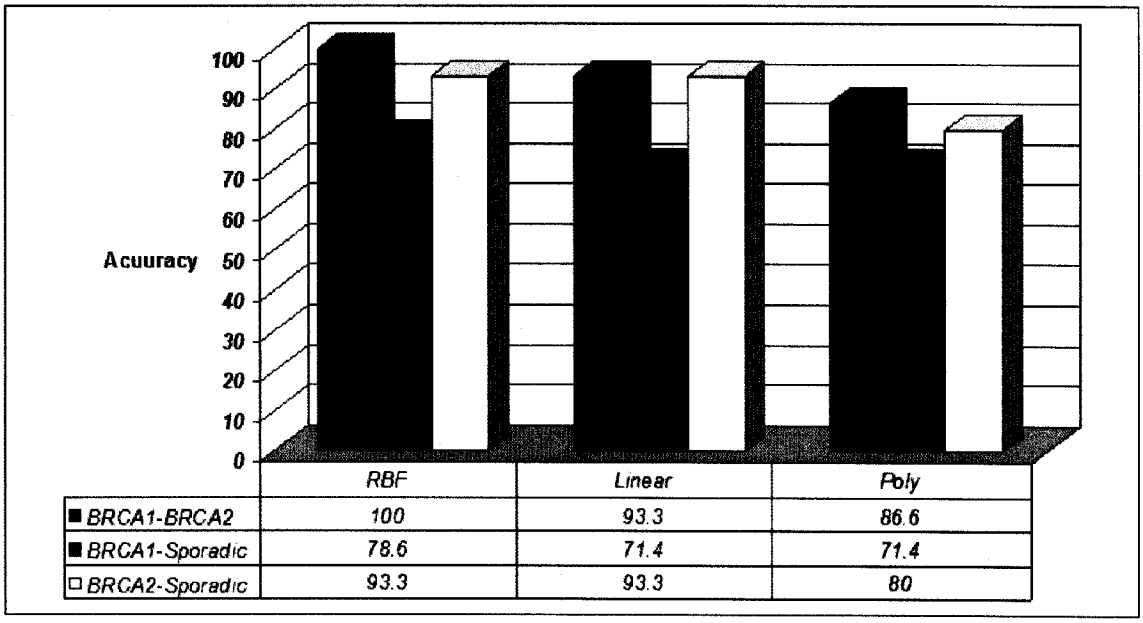


Figure 41. Classification accuracy shown with polynomial, linear and radial basis function kernel among the BRCA1–BRCA2, BRCA1-sporadic and BRCA2-sporadic breast cancer data

APPENDIX A

A complete listing of the KEGG pathways mapped for 157 unique *Eisenia fetida* sequences.

KEGG Pathway	# Mapping	Sequence ID	# Seq	% total
Carbohydrate Metabolism	10		35	22%
Glycolysis / Gluconeogenesis		EW1_F1plate01_F12, EW1_F2Plate20_G03, EW1_R1plate02_G06, EW2_F1plate03_F08, EW2_F1plate03_H05	5	3%
Citrate cycle (TCA cycle)		EW1_F1plate01_C07, EW2_R1Plate08_D07	2	1%
Pentose and glucuronate interconversions		Contig18	1	1%
Fructose and mannose metabolism		EW1_F1plate08_B05, EW1_R1plate03_G02, EW1_R1plate05_H11, EW2_F1plate02_G03	4	3%
Starch and sucrose metabolism		EW1_F2Plate20_H09	1	1%
Aminosugars metabolism		Contig125, Contig269, Contig275, EW1_F1plate02_G06, EW1_F1plate03_B11, EW1_F1plate04_H08, EW1_F1plate06_B12, EW1_F1plate06_H04, EW1_F1plate08_E04, EW1_R1plate06_B02, EW2_R1plate02_H03, EW2_R1plate03_B10, EW2_R1plate05_G04, EW2_R1plate07_H05, EW2_R1Plate08_B07, EW2_R1Plate10_G04, EW2_R1Plate11_C05	17	11%
Glyoxylate and dicarboxylate metabolism		EW1_F1plate01_C07	1	1%
Propanoate metabolism		EW1_F1plate05_B07, Contig321, EW1_R1plate03_D04	3	2%
Butanoate metabolism		EW2_F1plate03_C07, EW1_F1plate05_B07	2	1%
Inositol phosphate metabolism		EW1_F2plate14_A04	1	1%
Energy Metabolism	8		28	18%
Oxidative phosphorylation		Contig10, Contig58, Contig65, EW1_F1plate01_B01, EW1_F1plate02_F12, EW1_F1plate02_G07, EW1_F1plate04_C04, EW1_F1plate05_E04, EW1_F1plate06_E05, EW1_F1plate06_H02, EW1_F1plate07_F08, EW1_F1plate08_C02, EW1_F1plate08_E10, EW1_F2Plate19_B05, EW1_F2Plate20_D05, EW1_R1plate05_E07, EW2_F1plate02_D09, EW2_R1plate02_D06, EW2_R1plate07_D11	19	12%
Sulfur metabolism		Contig163	1	1%
Fatty acid metabolism		EW1_F1plate05_B07, EW1_F2Plate20_G03, Contig321, EW1_R1plate03_D04	4	3%
Bile acid biosynthesis		EW1_F2Plate20_G03	1	1%
Glycerolipid metabolism		EW2_F1plate02_G03, EW1_R1plate01_C09	2	1%
Glycerophospholipid metabolism		EW1_R1plate01_C09	1	1%
Ether lipid metabolism		EW2_R1plate07_D08	1	1%
Arachidonic acid metabolism		Contig18	1	1%
Nucleotide Metabolism	2		2	1%
Purine metabolism		EW1_F2plate11_E04, EW1_F1plate04_E06	2	1%
Pyrimidine metabolism		EW1_F1plate04_E06	1	1%
Amino Acid Metabolism	12		18	11%

Glutamate metabolism		EW2_F1plate03_C07	1	1%
Alanine and aspartate metabolism		EW2_F1plate03_C07	1	1%
Glycine, serine and threonine metabolism		Contig278	1	1%
Methionine metabolism		Contig278, Contig206, EW1_F2plate12_F12, EW1_R1Plate08_B05	4	3%
Valine, leucine and isoleucine degradation		EW1_F1plate05_B07, Contig356, Contig321, EW1_R1plate03_D04	4	3%
Lysine degradation		EW1_F1plate05_B07	1	1%
Arginine and proline metabolism		Contig236, EW2_R1Plate11_B03	2	1%
Histidine metabolism		EW1_R1plate02_G06, EW1_F1plate02_F11, EW2_R1plate01_E03	3	2%
Tyrosine metabolism		EW1_F2Plate20_G03, EW1_R1plate02_G06, Contig356, EW1_R1plate05_C11	4	3%
Phenylalanine metabolism		EW1_R1plate02_G06, EW1_R1plate05_C11	2	1%
Tryptophan metabolism		EW1_F1plate05_B07, Contig356, EW1_F2plate12_H08, EW2_R1plate07_C05	4	3%
Phenylalanine, tyrosine and tryptophan biosynthesis		EW1_F1plate01_F12, EW2_F1plate03_F08, EW2_F1plate03_H05	3	2%
Metabolism of Other Amino Acids	**3**		10	6%
beta-Alanine metabolism		EW2_F1plate03_C07, EW1_F1plate05_B07, Contig18, Contig321	4	3%
Selenoamino acid metabolism		Contig278, Contig206, EW1_F2plate12_F12, EW1_R1Plate08_B05, Contig163	5	3%
Glutathione metabolism		EW2_R1Plate11_D09	1	1%
Glycan Biosynthesis and Metabolism	**8**		6	4%
N-Glycan biosynthesis		EW1_F1plate09_D11, EW2_R1plate01_A06	2	1%
N-Glycan degradation		EW1_R1plate03_C01, EW1_R1plate03_F10	2	1%
Keratan sulfate biosynthesis		EW2_R1plate01_A06	1	1%
Glycosphingolipid biosynthesis - neo-lactoseries		Contig64, EW2_R1plate01_A06	2	1%
Glycosphingolipid biosynthesis - globoseries		Contig64	1	1%
Glycan structures - biosynthesis 1		EW1_F1plate09_D11, EW2_R1plate01_A06	2	1%
Glycan structures - biosynthesis 2		EW2_R1plate01_A06, EW2_F1plate02_G03, Contig64	3	2%
Glycan structures - degradation		EW1_R1plate03_C01, EW1_R1plate03_F10	2	1%
Metabolism of Cofactors and Vitamins	**6**		9	6%
Vitamin B6 metabolism		Contig356, EW1_F1plate02_E12, EW2_R1plate01_A08	3	2%
Nicotinate and nicotinamide metabolism		Contig356	1	1%
Pantothenate and CoA biosynthesis		EW1_R1plate07_F01, EW1_R1plate07_H09	2	1%
Folate biosynthesis		EW1_F2Plate20_H09	1	1%
One carbon pool by folate		EW1_F1plate02_F11, EW2_R1plate01_E03	2	1%
Retinol metabolism		EW1_R1plate02_G06	1	1%
Biosynthesis of Secondary Metabolites	**1**		2	1%
Limonene and pinene degradation		EW1_F1plate05_B07, EW1_F1plate08_H02	2	1%
Xenobiotics Biodegradation and Metabolism	**7**		6	4%
Caprolactam degradation		EW1_F1plate05_B07	1	1%
gamma-Hexachlorocyclohexane degradation		EW1_F1plate08_H02	1	1%
Ethylbenzene degradation		EW2_R1plate07_D08	1	1%
Benzoate degradation via CoA ligation		EW1_F1plate05_B07, EW1_F1plate08_H02	2	1%

Bisphenol A degradation		EW1_F1plate08_H02	1	1%
1- and 2-Methylnaphthalene degradation		EW1_F1plate08_H02, EW1_F2Plate20_G03	2	1%
Metabolism of xenobiotics by cytochrome P450		EW1_F2Plate20_G03, EW1_R1plate02_G06, EW2_R1Plate11_D09	3	2%
Transcription	**2**		2	1%
RNA polymerase		EW1_F1plate04_E06	1	1%
Basal transcription factors		EW1_F1plate05_B05	1	1%
Translation	**1**		17	11%
Ribosome		Contig164, Contig201, Contig312, Contig385, Contig78, EW1_F1plate09_E03, EW1_F2plate16_A07, EW1_F2plate16_A08, EW1_F2plate16_A09, EW1_F2plate16_A10, EW1_F2plate16_A11, EW1_F2plate16_A12, EW2_F1plate01_D07, EW2_F1plate03_H07, EW2_R1plate01_F09, EW2_R1plate03_D02, EW2_R1plate07_C03	17	11%
Folding, Sorting and Degradation	**3**		9	6%
Ubiquitin mediated proteolysis		EW2_R1plate02_D03	1	1%
Proteasome		Contig292, EW1_F1plate01_H09, EW1_F1plate04_D12, EW1_F1plate05_D04, EW1_F1plate07_B12, EW1_F1plate07_E08, EW2_R1Plate08_E10	7	4%
DNA polymerase		Contig52	1	1%
Membrane Transport	**1**		1	1%
ABC transporters - General		Contig66	1	1%
Signal Transduction	**6**		14	9%
MAPK signaling pathway		EW1_F1plate02_B07, EW1_F1plate07_C07, EW1_F1plate02_E08, EW2_R1Plate10_D02, EW1_R1plate04_D09	5	3%
Wnt signaling pathway		EW1_F1plate05_E07, EW2_F1plate03_B09, EW2_F1plate03_C09	3	2%
Notch signaling pathway		EW1_R1Plate08_E02, Contig116, EW1_R1plate03_B09	3	2%
TGF-beta signaling pathway		EW1_F1plate02_F09, EW2_F1plate03_B09, EW2_F1plate03_C09	3	2%
Calcium signaling pathway		Contig215	1	1%
Phosphatidylinositol signaling system		Contig215, EW1_R1plate01_C09	2	1%
Signaling Molecules and Interaction	**3**		13	8%
Neuroactive ligand-receptor interaction		EW2_F1plate03_A01, EW2_R1plate03_A02, EW2_R1plate04_B08, EW2_R1plate05_H01, EW2_R1Plate08_C09, EW2_F1plate01_D02	6	4%
Cytokine-cytokine receptor interaction		EW1_R1plate07_E02	1	1%
ECM-receptor interaction		EW1_F1plate01_F05, EW1_F1plate01_F11, EW1_F1plate04_B12, EW1_F1plate01_F03, EW1_F1plate02_F09, EW1_F1plate08_B06	6	4%
Cell Motility	**3**		9	6%
Regulation of actin cytoskeleton		EW1_F2plate13_B04, EW1_F2Plate20_C02, EW2_F1plate01_E07, EW2_R1Plate11_A12, EW2_R1Plate11_F10	5	3%
Cell cycle		EW1_F2plate13_E09, EW2_F1plate03_B09, EW2_F1plate03_C09	3	2%
Apoptosis		EW1_F2plate11_A09	1	1%
Cell Communication	**4**		13	8%
Focal adhesion		EW1_R1plate07_E02, EW1_F1plate01_F03, EW1_F1plate02_B07, EW1_F1plate02_F09,	6	4%

		EW1_F1plate07_C07, EW1_F1plate08_B06		
Adherens junction		EW1_R1plate07_E02	1	1%
Tight junction		Contig158, EW1_F2plate11_C07	2	1%
Gap junction		EW2_F1plate03_D07, EW2_R1plate01_F11, EW2_R1plate02_F04, EW2_R1Plate08_G05, EW2_R1Plate09_C05	5	3%
Endocrine System	3		4	3%
Insulin signaling pathway		Contig215	1	1%
PPAR signaling pathway		Contig321, EW1_R1plate03_D04	2	1%
GnRH signaling pathway		Contig215, EW2_R1Plate10_D02	2	1%
Immune System	3		5	3%
Complement and coagulation cascades		Contig214	1	1%
Toll-like receptor signaling pathway		EW1_F2plate11_A09	1	1%
Antigen processing and presentation		Contig178, Contig363, EW1_R1plate04_D09	3	2%
Nervous System	2		8	5%
Long-term potentiation		Contig215, EW2_F1plate03_A01, EW2_R1plate03_A02, EW2_R1plate04_B08, EW2_R1plate05_H01, EW2_R1Plate08_C09	6	4%
Long-term depression		EW2_R1plate02_F04, EW2_R1Plate09_C05	2	1%
Sensory System	1		3	2%
Olfactory transduction		EW2_R1plate02_F04, EW2_R1Plate09_C05, Contig215	3	2%
Development	2		3	2%
Dorso-ventral axis formation		EW1_R1Plate08_E02	1	1%
Axon guidance		EW1_F2plate13_D11, EW1_R1plate07_E02	2	1%
Neurodegenerative Disorders	4		6	4%
Alzheimer's disease		Contig116, EW1_R1plate03_B09	2	1%
Parkinson's disease		EW1_F2plate16_B03	1	1%
Huntington's disease		Contig278, Contig215	2	1%
Prion disease		EW1_R1plate04_D09	1	1%
Metabolic Disorders	2		2	1%
Type II diabetes mellitus		EW2_F1plate03_C07	1	1%
Maturity onset diabetes of the young		EW1_F2plate14_D06	1	1%
Cancers	2		2	1%
Colorectal cancer		EW1_R1plate07_E02	1	1%
Glioma		Contig215	1	1%

APPENDIX B

Plots of 40 Microarray slides

A. Scatter Plot of 40 microarray raw data

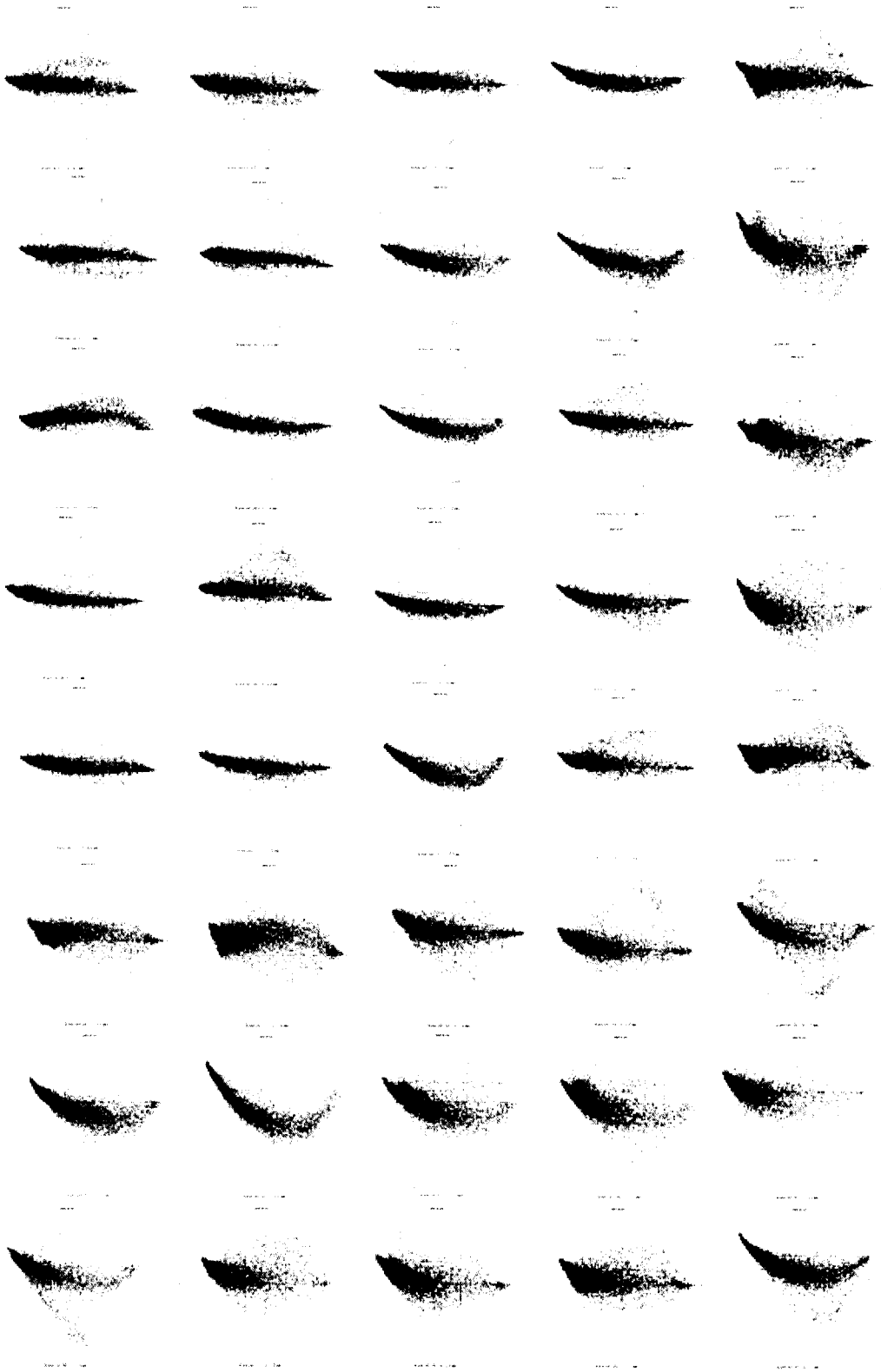
B. Scatter Plot of 40 microarray normalized data

C. MA Plot of 40 microarray raw data

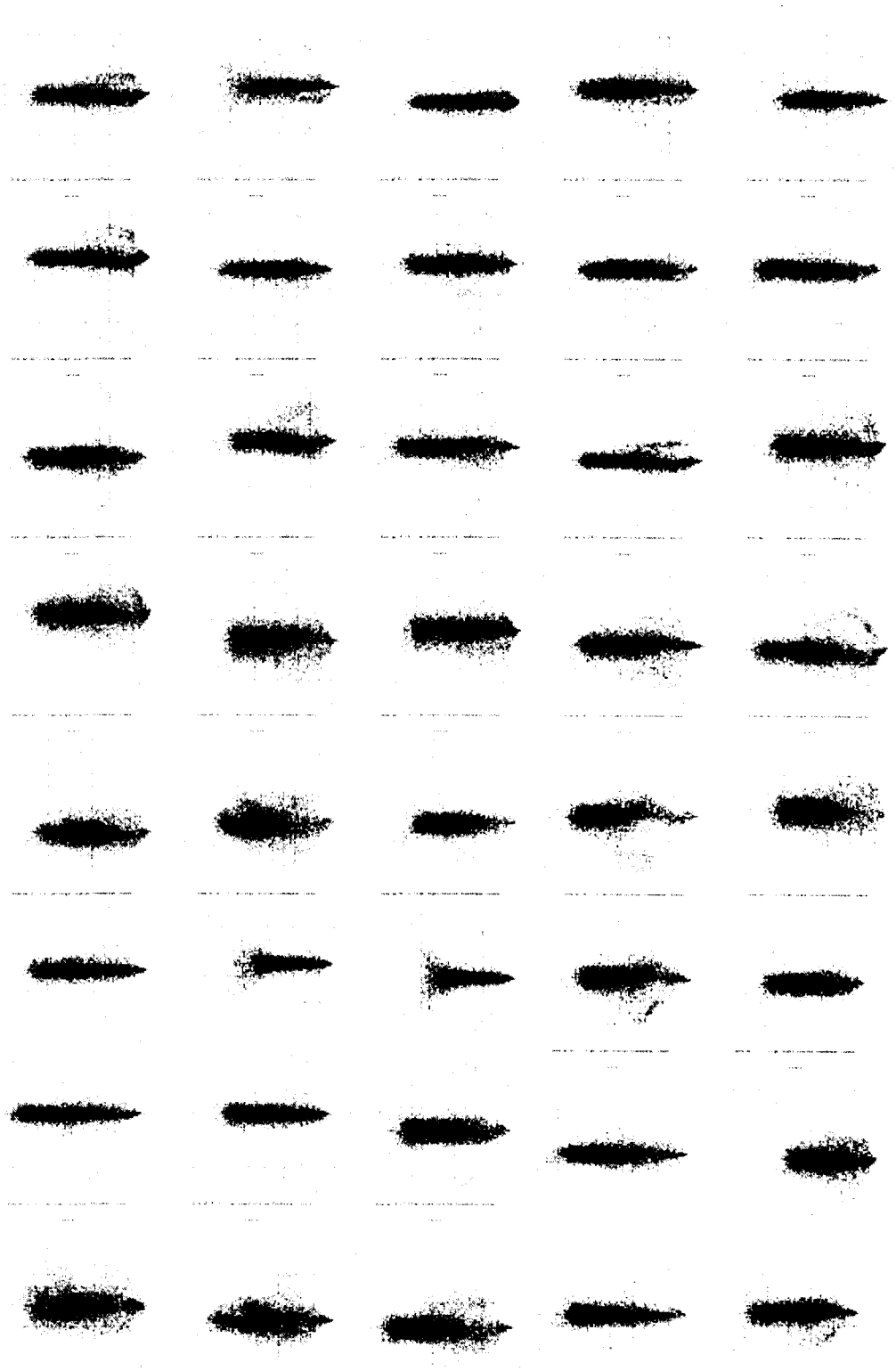
D. MA Plot of 40 microarray normalized data

A. Scatter Plot of 40 microarray raw data

B. Scatter Plot of 40 microarray normalized data



C. MA Plot of 40 microarray raw data



D. MA Plot of 40 microarray normalized data

APPENDIX C

109 significant overlapped sequences between SAM and t-test with their blastx results

Query ID	Length	Acc. Version #	Length	Evalue	Organism
EW1_F1plate01_A06	451	CAI08599.1	663	6.2	Azoarcus
EW1_F1plate01_F04	252	XP_708403.1	480	5.00E-12	Danio rerio
EW1_F1plate01_G01	401	XP_426056.1	566	6.00E-11	Gallus gallus
EW1_F1plate01_H12	250	******No hits			
EW1_F1plate02_A01	538	NP_524480.2	2146	4.1	Drosophila melanogaster
EW1_F1plate02_B01	265	ZP_00859789.1	316	6.2	Bradyrhizobium sp. BTAi1
EW1_F1plate02_B05	387	BAD15061.1	477	1.00E-13	Paralichthys olivaceus
EW1_F1plate02_B12	13	******No hits			
EW1_F1plate02_C04	252	XP_396925.2	909	1.00E-15	Apis mellifera
EW1_F1plate02_E05	306	AAS66770.1	408	0.002	Theromyzon rude
EW1_F1plate02_E11	516	BAC88577.1	938	0.11	Gloeobacter violaceus PCC 7421
EW1_F1plate02_E12	662	XP_785156.1	283	2.00E-10	
EW1_F1plate02_F08	300	BAD72193.1	373	0.05	Oryza sativa
EW1_F1plate03_C03	355	AAS07949.1	923	6.2	uncultured bacterium 463
EW1_F1plate03_E02	437	AAP99786.1	583	0.72	
EW1_F1plate03_G02	81	******No hits			
EW1_F1plate03_G07	281	AAH73276.1	489	8.00E-13	Xenopus laevis
EW1_F1plate04_A02	570	EAL25702.1	1216	1.00E-15	Drosophila pseudoobscura
EW1_F1plate04_A03	132	AAL76032.1	296	0.94	Aedes aegypti
EW1_F1plate04_B10	547	CAH10356.1	154	2.00E-04	
EW1_F1plate04_D04	608	XP_731877.1	100	1.00E-54	Plasmodium chabaudi chabaudi
EW1_F1plate04_H08	390	AAH69614.1	454	1.00E-25	Homo sapiens
EW1_F1plate05_B04	277	AAQ54709.1	172	2.00E-07	Amblyomma maculatum
EW1_F1plate05_C01	433	ZP_01137954.1	273	6.1	Acidothermus cellulolyticus
EW1_F1plate05_E04	530	EAA00151.2	110	5.00E-24	Anopheles gambiae str. PEST
EW1_F1plate05_E08	132	AAL76032.1	296	0.94	Aedes aegypti
EW1_F1plate05_E10	321	XP_789440.1	545	2.00E-06	
EW1_F1plate05_F09	460	P13579	151	3.00E-14	
EW1_F1plate05_H06	485	CAH10355.1	153	0.033	
EW1_F1plate05_H11	498	CAH10356.1	154	6.00E-05	
EW1_F1plate06_B12	358	AAH69614.1	454	1.00E-18	Homo sapiens
EW1_F1plate06_C04	377	P02218	145	3.00E-17	
EW1_F1plate06_F04	425	AAH69614.1	454	2.00E-26	Homo sapiens
EW1_F1plate06_G03	376	AAO81977.1	1004	0.33	Enterococcus faecalis V583
EW1_F1plate06_G08	423	CAH03250.1	179	4.7	Paramecium tetraurelia
EW1_F1plate06_H04	428	CAC87888.1	488	1.00E-26	Bufo japonicus
EW1_F1plate06_H05	692	CAC37630.1	2673	1.00E-48	Homo sapiens
EW1_F1plate07_A11	425	AAH69614.1	454	2.00E-26	Homo sapiens
EW1_F1plate07_B05	526	XP_514259.1	643	0.34	Pan troglodytes
EW1_F1plate07_B07	379	NP_568124.1	766	2.1	Arabidopsis thaliana
EW1_F1plate07_E02	501	XP_789440.1	545	3.00E-07	
EW1_F1plate07_E04	321	XP_387888.1	673	3.6	Gibberella zeae PH-1

EW1_F1plate07_G02	544	******No hits			
EW1_F1plate07_G12	396	ZP_01112228.1	292	0.24	Alteromonas macleodii 'Deep
EW1_F1plate07_H08	336	EAR99456.1	2046	8	Tetrahymena thermophila
EW1_F1plate07_H09	62	******No hits			
EW1_F1plate08_A02	683	******No hits			
EW1_F1plate08_D07	197	******No hits			
EW1_F1plate08_E05	396	AAD56953.1	128	0.004	Myxine glutinosa
EW1_F2plate13_F04	605	ZP_00471943.1	481	2.5	Chromohalobacter salexigens DSM
EW1_F2plate16_E08	602	AAR98305.1	328	0.5	Orf virus
EW1_R1plate03_H08	413	P02218	145	1.00E-55	
EW1_R1plate06_A02	343	XP_537030.2	493	3.00E-22	
EW1_R1plate06_B02	448	AAH69614.1	454	1.00E-29	Homo sapiens
EW1_R1plate06_B11	337	XP_954774.1	175	4.00E-05	Theileria annulata strain Ankara
EW1_R1plate06_C07	695	EAL23259.1	898	1.5	Cryptococcus neoformans var.
EW1_R1plate06_E02	393	******No hits			
EW1_R1plate07_A01	846	AAM15241.1	394	0.006	Arabidopsis thaliana
EW1_R1plate07_A05	368	CAC87888.1	488	3.00E-24	Bufo japonicus
EW1_R1plate07_C02	653	CAD29317.1	177	1.00E-28	Lumbricus terrestris
EW1_R1plate07_E10	494	ABC68595.1	618	4.00E-32	Paracentrotus lividus
EW1_R1plate07_E11	393	******No hits			
EW1_R1plate07_F12	685	XP_664503.1	569	1.9	Aspergillus nidulans FGSC A4
EW1_R1plate07_H02	772	CAG01990.1	703	1.00E-12	Tetraodon nigroviridis
EW2_F1plate01_A03	465	AAL59385.1	193	2.00E-10	Citrobacter freundii
EW2_F1plate01_E03	496	AAK57554.1	126	1.00E-06	Methanococcus voltae
EW2_F1plate02_F08	563	AAL59385.1	193	8.00E-05	Citrobacter freundii
EW2_F1plate03_G03	536	AAL59385.1	193	1.00E-08	Citrobacter freundii
EW2_F1plate03_H02	675	AAH73276.1	489	2.00E-12	Xenopus laevis
EW2_R1plate01_A08	644	XP_785156.1	283	7.00E-10	
EW2_R1plate01_C02	298	BAC06447.1	929	4.00E-15	Haemaphysalis longicornis
EW2_R1plate01_G08	597	CAC87888.1	488	3.00E-27	Bufo japonicus
EW2_R1plate01_G10	483	AAK57554.1	126	1.00E-08	Methanococcus voltae
EW2_R1plate02_A08	6	******No hits			
EW2_R1plate02_D08	256	CAD24436.1	269	0.005	Palmaria decipiens
EW2_R1plate02_F07	356	AAL59385.1	193	2.00E-06	Citrobacter freundii
EW2_R1plate02_G02	501	AAN63032.1	175	5.00E-14	Branchiostoma lanceolatum
EW2_R1plate02_G04	355	AAL59385.1	193	9.00E-09	Citrobacter freundii
EW2_R1plate02_G07	312	AAL59385.1	193	2.00E-06	Citrobacter freundii
EW2_R1plate02_H03	388	NP_446012.1	370	5.00E-10	Rattus norvegicus
EW2_R1plate03_A07	427	XP_708403.1	480	9.00E-15	Danio rerio
EW2_R1plate03_B06	389	AAP83794.1	171	5.00E-13	Crassostrea gigas
EW2_R1plate03_C09	336	AAR13226.1	242	5.00E-25	Eisenia fetida
EW2_R1plate03_C11	488	AAQ12076.1	206	6.00E-14	Pinctada fucata
EW2_R1plate03_F08	368	AAL59385.1	193	3.00E-06	Citrobacter freundii
EW2_R1plate03_F11	240	AAK57554.1	126	9.00E-07	Methanococcus voltae
EW2_R1plate03_H08	317	NP_001020370.1	311	3.00E-12	Homo sapiens

EW2_R1plate04_D02	297	ABD76397.1	242	4.00E-23	Eisenia fetida
EW2_R1plate04_D04	466	XP_227566.2	275	3.00E-19	Rattus
EW2_R1plate04_D05	690	AAK57554.1	126	8.00E-04	Methanococcus voltae
EW2_R1plate05_A10	402	AAL59385.1	193	9.00E-07	Citrobacter freundii
EW2_R1plate05_C11	611	AAN63032.1	175	2.00E-12	Branchiostoma lanceolatum
EW2_R1plate05_F03	287	ABD76397.1	242	2.00E-31	Eisenia fetida
EW2_R1plate05_G04	604	AAH69614.1	454	4.00E-26	Homo sapiens
EW2_R1plate05_G05	498	AAQ12076.1	206	2.00E-11	Pinctada fucata
EW2_R1plate05_G12	340	AAL59385.1	193	9.00E-07	Citrobacter freundii
EW2_R1plate05_H10	306	AAL59385.1	193	9.00E-07	Citrobacter freundii
EW2_R1plate06_A01	406	AAF61070.1	124	5.00E-36	Paralichthys olivaceus
EW2_R1plate06_B01	446	AAL59385.1	193	3.00E-04	Citrobacter freundii
EW2_R1plate06_B03	614	AAB68960.1	497	8.00E-12	
EW2_R1plate06_B05	518	AAL59385.1	193	0.024	Citrobacter freundii
EW2_R1plate06_F04	281	BAC06447.1	929	2.00E-13	Haemaphysalis longicornis
EW2_R1plate06_G12	427	AAL59385.1	193	2.00E-04	Citrobacter freundii
EW2_R1plate06_H08	291	BAC06447.1	929	2.00E-13	Haemaphysalis longicornis
EW2_R1plate07_C07	576	AAK57554.1	126	1.00E-05	Methanococcus voltae
EW2_R1plate07_D06	461	AAK57554.1	126	9.00E-07	Methanococcus voltae
EW2_R1plate07_D10	549	AAH69614.1	454	2.00E-29	Homo sapiens
EW2_R1plate07_F01	498	AAL59385.1	193	3.00E-04	Citrobacter freundii
EW2_R1plate10_C02	555	AAN63032.1	175	1.00E-14	Branchiostoma lanceolatum

APPENDIX D

GLOSSARY

Bias The word bias refers to all sources of systematic variations, for example: PCR/handling of clones, printing and/or tip problems, labeling and dye effects, uneven hybridization, scanner malfunction.

Biological replicates biological samples from independent sources, representing the same condition, e.g. liver tissue from individual mice of the same sex and strain.

Bonferroni correction Multiple-testing adjustment in which the significance-level is divided by the total number of tests

C-SVC C-Support Vector Classification

cDNA complementary DNA (cDNA) is single-stranded DNA synthesized from a mature mRNA template by reverse transcriptase often synthesized from a cellular extract.

Channel A channel is an intensity-based portion of an expression dataset. In some cases, such as Cy3/Cy5 array hybridizations, multiple channels (one for each label used) may be combined to create ratios.

Chromosomes Part of a cell that contains genetic information. A chromosome is a grouping of coiled strands of DNA, containing many genes. Most multicellular organisms have several chromosomes, which together comprise the genome. Sexually reproducing organisms have two copies of each chromosome, one from the each parent.

Class In experimental design, a *class* denotes a subset of the whole experiment. For example one single time-point out of a time-course experiment represents one class, containing all microarrays belonging to this time-point. An experiment can consist of any number of classes.

Control The reference for comparison when determining the effect of some procedure or treatment.

Covariate A covariate is a variable that is possibly predictive of the outcome under study.

COX Cytochrome c Oxidase

Cross-hybridization The hydrogen bonding of a single-stranded DNA sequence that is partially but not entirely complementary to a single-stranded substrate. Often, this

involves hybridizing a DNA probe for a specific DNA sequence to the homologous sequences of different species.

Cross-validation The cross-validation is the practice of partitioning a sample of data into subsets such that analysis is initially performed on a single subset, while further subsets are retained \blind" in order for subsequent use in confirming and validating the initial analysis.

Cy3, Cy5 Cyanine fluorescent dyes used in microarray experiments for labelling different samples of DNA. Cy3 can be visualized as green, Cy5 as red.

DBH Dopamine-Beta-Hydroxylase

DE Short form for *differentially expressed.*

Dendrogram A hierarchy representation by a dichotomous diagram, in which the end of a branch corresponds to an element and the level of a junction corresponds to the taxonomic distance from the two elements or the two groups that it connects.

Distribution A distribution is a graphic representation of the values of a variable. The line formed by connecting data points is called a frequency distribution. An important aspect of the description of a variable is the shape of its distribution. Typically, one is interested in how well the distribution can be approximated by the normal distribution.

DNA (DeoxyriboNucleic Acid) The molecule that encodes genetic information. DNA is a double-stranded polymer of nucleotides. The two strands are held together by hydrogen bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases: adenine (A), guanine (G), cytosine (C), and thymine (T).

DNT (2,4-DNT) 2,4-dinitrotoluene

DNT (2,6-DNT) 2,6-dinitrotoluene

Dye-swap pair Two slides comparing the same samples of RNA, one with normal and one with reversed dye-assignment.

ε-SVR ε-Support Vector Regression (epsilon-SVR)

Error In statistics, *error* refers to all kinds of unspecific variability (variability introduced in the measurement). That is different from the everyday-use to mean *mistake.*

Estimation The process of using sample statistics to estimate population parameters.

EST Expressed Sequence Tags

ESTMD Expressed Sequence Tags Model Database

Expression The conversion of the genetic instructions present in a DNA sequence into a unit of biological function in a living cell. Typically involves the process of transcription of a DNA sequence into an RNA sequence.

Fold change The ratio of RNA quantities between two samples in a microarray experiment.

Gene DNA which codes for a particular protein or a functional or structural RNA molecule.

GenBank The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of an international collaboration with the European Molecular Biology Laboratory (EMBL) Data Library from the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ).

Gene Expression Transcription of the information contained within the DNA into messenger RNA (mRNA) molecules that are then translated into proteins.

GO Gene Ontology

GUI Graphical User Interface

HMX octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine

Hybridization is the process of binding complementary pairs of DNA molecules. It is the act of treating a microarray with one or more labeled preparations from a specified set of conditions.

J2EE Java 2 Enterprise Edition

JSP JavaServer Pages

KEGG Kyoto Encyclopedia of Genes and Genomes

Meta-analysis Analysis involving several sources of microarray data (e.g. Affymetrix and Agilent data)

Microarray A microarray (or slide) refers to the physical substrates to which biosequence (cDNA or oligos) are attached. Microarrays are hybridized with labeled samples and then scanned and analyzed to generate data.

Microarray experiment An experiment studies a system under controlled conditions while some conditions are changed. In gene expression, one varies some parameter such

as time, drug, developmental stage, or dosage on a sample. The sample is processed and labeled with a detectable tag (Cy3, Cy5) so that it can be used in hybridization with microarrays.

Missing values may exist in microarray data. In this case the spot is empty(intensity= 0) or background intensity is higher than the spot intensity.

mRNA (messenger RNA) A specialized form of RNA that serves as a template to direct protein biosynthesis. The amount of any particular type of mRNA in a cell reflects the extent to which a gene has been expressed.

nu-SVC ν-Support Vector Classification

nu-SVR ν-Support Vector Regression (ν-SVR)

Normal distribution or Gaussian distribution, this is one of the most important statistical distributions, since experimental errors are often normally distributed. Further, the normal assumption simplifies many methods of data analysis.

Normalization The process of removing the effect of all sources of non-biological variation from microarray data, making them comparable.

Null hypothesis A hypothesis for which the effects of interest are assumed to be absent.

Commonly used as basis for setting up statistical tests.

Oligo (Oligonucleotide) Short sequence of nucleotides (less than 80 bp) single stranded to be used as probes or spots. Oligos are often chemically synthesized.

ORCs Ordnance Related Compounds

PCR (Polymerase Chain Reaction) allows the exponential copying of part of a DNA molecule using a DNA polymerase enzyme. PCR is the Exponential amplification of almost any region of a selected DNA molecule.

Protein A biological molecule which consists of many amino acids chained together by peptide bonds. Proteins perform most of the enzymatic and structural roles within living cells.

Probe is an easily detectable molecule which has the property to be located specifically either on another molecule, or in a given cellular compartment. A marker (enzyme, compound radioactive or fluorescent) can be associated with the probe which allows its detection. Generally the probe is a nucleic acid fragment (RNA or DNA).

Probeset Set of probes used in the microarray platform of Affymetrix. Even if, generally, a probeset corresponds to one gene, the expression of one gene may be measured by a set of probesets.

p-Value A measure of evidence against the null hypothesis in a statistical test.

Ratio Also referred to as a "fold change". A ratio refers to a normalized signal intensity generated from one feature in a given channel divided by a normalized signal intensity generated by the same feature in another channel.

RDX 1,3,5-trinitro-1,3,5-triazacyclohexane

Replication A replicate set refers to repeated experiments where the same type of array is used, and the same probe isolation method is used to get more statistically meaningful interpretation of results. Reproducing an experiment helps to verify its results.

RNA (ribonucleic acid) A class of nucleic acids that consist of nucleotides containing the bases: adenine (A), guanine (G), cytosine (C), and uracil (U). An RNA molecule is typically single-stranded and can pair with DNA or with another RNA molecule.

RT-PCR (Reverse Transcription Polymerase Chain Reaction) The most sensitive technique for mRNA detection and quantitation currently available. It uses upon the reverse transcriptase to amplify a sequence of RNA and to transform it into DNA. It is sensitive enough to enable quantitation of RNA from a single cell.

Sample A subset of a population. Usually, the size of the sample is much less than the size of the population. The primary goal of statistics is to use information collected from a sample to try to characterize a certain population.

Sensitivity The sensitivity of a binary classification test is a parameter that expresses something about the test's performance. The sensitivity of such a test is the proportion of those cases having a positive test result of all positive cases tested (TP / (TP+FN)).

Significance level The p-value that is regarded as providing sufficient evidence against a null hypothesis. If the p-value falls below the significance-level, the null hypothesis is rejected.

Skewness is a measure of the asymmetry of the probability distribution of a real valued random variable. A distribution has positive skew (right skewed) if the higher tail is longer and negative skew (left-skewed) if the lower tail is longer.

SOD Cu/Zn-superoxide Dismutase

Specificity The specificity of a binary classification test is a parameter that expresses something about the test's performance. The specificity of such a test is the proportion of true negatives of all the negative samples tested (TN/ (TN+FP)).

SSH Suppression Subtractive Hybridization

Statistical significance A result is statistically significant when it doesn't happen by chance.

Subgrid A sub area of a single microarray. Within one subgrid all spots are printed by the same print-tip.

SVM Support Vector Machine

Technical replicates Multiple hybridisations with RNA samples obtained from the same biological source.

TNB 1,3,5-trinitrobenzene

TNT 2,4,6-trinitrotoluene

Variable Numerical data are observations which are recorded in the form of numbers. Numbers are variable in nature. E.g., when measuring gene expression levels, the score will vary for reasons such as temperature, cell activity etc. For this reason, the gene expression level is called *variable.*

VC dimension Vapnik Chervonenkis dimension

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF *et al.*. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252(5013):1651-6.

Ahmed FE. 2005. Artificial neural networks for diagnosis and survival prediction in colon cancer. Mol Cancer 4:29.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X *et al.*. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403(6769):503-11.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A 96(12):6745-50.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403-10.

Alvarenga P, Palma P, Goncalves AP, Fernandes RM, Cunha-Queda AC, Duarte E, Vallini G. 2007. Evaluation of chemical and ecotoxicological characteristics of biodegradable organic residues for application to agricultural land. Environ Int 33(4):505-13.

Anthony M, Bartlett PL. 1999. Neural Network Learning: Theoretical Foundations: Cambridge University Press.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.*. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25-9.

Ayoubi P, Jin X, Leite S, Liu X, Martajaja J, Abduraham A, Wan Q, Yan W, Misawa E, Prade RA. 2002. PipeOnline 2.0: automated EST processing and functional data sorting. Nucleic Acids Res 30(21):4761-9.

Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG *et al.*. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 8(8):816-24.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A *et al.*. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406(6795):536-40.

Black MA, Doerge RW. 2002. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. Bioinformatics 18(12):1609-16.

Blower PE, Cross KP. 2006. Decision tree methods in pharmaceutical research. Curr Top Med Chem 6(1):31-9.

Boelsterli UA. 2003. Diclofenac-induced liver injury: a paradigm of idiosyncratic drug toxicity. Toxicol Appl Pharmacol 192(3):307-22.

Bradham KD, Dayton EA, Basta NT, Schroder J, Payton M, Lanno RP. 2006. Effect of soil properties on lead bioavailability and toxicity to earthworms. Environ Toxicol Chem 25(3):769-75.

Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., Haussler D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 97(1):262-7.

Brulle F, Mitta G, Cocquerelle C, Vieau D, Lemiere S, Lepretre A, Vandenbulcke F. 2006. Cloning and real-time PCR testing of 14 potential biomarkers in Eisenia fetida following cadmium exposure. Environ Sci Technol 40(8):2844-50.

Brulle F, Mitta G, Leroux R, Lemiere S, Lepretre A, Vandenbulcke F. 2007. The strong induction of metallothionein gene following cadmium exposure transiently affects the expression of many genes in Eisenia fetida: a trade-off mechanism? Comp Biochem Physiol C Toxicol Pharmacol 144(4):334-41.

Bundy JG, Spurgeon DJ, Svendsen C, Hankard PK, Osborn D, Lindon JC, Nicholson JK. 2002. Earthworm species of the genus Eisenia can be phenotypically differentiated by metabolic profiling. FEBS Lett 521(1-3):115-20.

Byvatov E, Schneider G. 2003. Support vector machine applications in bioinformatics. Appl Bioinformatics 2(2):67-77.

Casasent D, Chen XW. 2003. Radial basis function neural networks for nonlinear Fisher discrimination and Neyman-Pearson classification. Neural Netw 16(5-6):529-35.

Chang C-C, Lin C-J. 2001. LIBSVM: a library for support vector machines.

Chen CF, Feng X, Szeto J. 2006. Identification of critical genes in microarray experiments by a Neuro-Fuzzy approach. Comput Biol Chem 30(5):372-81.

Chou HH, Holmes MH. 2001. DNA sequence quality trimming and vector removal. Bioinformatics 17(12):1093-104.

Churchill GA. 2002. Fundamentals of experimental design for cDNA microarrays. Nat Genet 32 Suppl:490-5.

Cortes C, Vapnik V. 1995. Support-Vector Networks. Machine Learning 20(3):273-297.

Cowan JD, Sharp DH. 1988. Neural nets. Q Rev Biophys 21(3):365-427.

De Vos J, Thykjaer T, Tarte K, Ensslen M, Raynaud P, Requirand G, Pellet F, Pantesco V, Reme T, Jourdan M *et al.*. 2002. Comparison of gene expression profiling between malignant and normal plasma cells with oligonucleotide arrays. Oncogene 21(44):6848-57.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society 34:1-38.

Demuynck S, Grumiaux F, Mottier V, Schikorski D, Lemiere S, Lepretre A. 2006. Metallothionein response following cadmium exposure in the oligochaete Eisenia fetida. Comp Biochem Physiol C Toxicol Pharmacol 144(1):34-46.

Deng Y, Dong Y, Brown SJ, Zhang C. An Integrated Web-Based Model for Management, Analysis and Retrieval of EST Biological Information. Lecture Notes in Computer Science; 2006a; Harbin, China. Springer Berlin / Heidelberg. p 931-938.

Deng Y, Dong Y, Thodima V, Clem RJ, Passarelli AL. 2006b. Analysis and functional annotation of expressed sequence tags from the fall armyworm Spodoptera frugiperda. BMC Genomics 7:264.

Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED *et al.*. 1996. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci U S A 93(12):6025-30.

Diaz-Uriarte R, Alvarez de Andres S. 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7:3.

Dojer N, Gambin A, Mizera A, Wilczynski B, Tiuryn J. 2006. Applying dynamic Bayesian networks to perturbed gene expression data. BMC Bioinformatics 7:249.

Duan KB, Rajapakse JC, Wang H, Azuaje F. 2005. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience 4(3):228-234.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14(9):755-63.

Everitt R, Minnema SE, Wride MA, Koster CS, Hance JE, Mansergh FC, Rancourt DE. 2002. RED: the analysis, management and dissemination of expressed sequence tags. Bioinformatics 18(12):1692-3.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8(3):186-94.

Felsenstein J. 2003. Inferring Phylogenies. Sunderland, MA: Sinauer Associates.

Frank E, Hall M, Trigg L, Holmes G, Witten IH. 2004. Data mining in bioinformatics using Weka. Bioinformatics 20(15):2479-81.

Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian networks to analyze expression data. J Comput Biol 7(3-4):601-20.

Futschik M, Crompton T. 2004. Model selection and efficiency testing for normalization of cDNA microarray data. Genome Biol 5(8):R60.

Galay-Burgos M, Spurgeon DJ, Weeks JM, Sturzenbaum SR, Morgan AJ, Kille P. 2003. Developing a new method for soil pollution monitoring using molecular genetic biomarkers. Biomarkers 8(3-4):229-39.

Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI *et al.*. 2001. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 98(24):13784-9.

Gene Ontology Consortium 2001. Creating the gene ontology resource: design and implementation. Genome Res 11(8):1425-33.

Greer and Khan, 2004. Diagnostic classification of cancer using DNA microarrays and artificial intelligence. Ann. N.Y. Acad. Sci. v1020. 49-66.

Ghorbel MT, Sharman G, Hindmarch C, Becker KG, Barrett T, Murphy D. 2006. Microarray screening of suppression subtractive hybridization-PCR cDNA libraries identifies novel RNAs regulated by dehydration in the rat supraoptic nucleus. Physiol Genomics 24(2):163-72.

Glonek GF, Solomon PJ. 2004. Factorial and time course designs for cDNA microarray experiments. Biostatistics 5(1):89-111.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res 8(3):195-202.

Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46(1-3):389-422.

Hall M. 1998. Correlation-based Feature Selection for Machine Learning.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al.*. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32(Database issue):D258-61.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22(2):160-74.

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP *et al.*. 2001. Gene-expression profiles in hereditary breast cancer. N Engl J Med 344(8):539-48.

Homma-Takeda S, Hiraku Y, Ohkuma Y, Oikawa S, Murata M, Ogawa K, *et al.* 2002. 2,4,6-trinitrotoluene-induced reproductive toxicity via oxidative DNA damage by its metabolite. Free Radic Res 36: 555-566.

Hovatter PS, Talmage SS, Opresko DM, Ross RH. 1997. Ecotoxicity of nitroaromatics to aquatic and terrestrial species at army superfund sites. In: Environmental Toxicology and Risk Assessment: Modeling and Risk Assessment (Doane TR, Hinman ML, eds). West Conshohocken, PA:American Society for Testing and Materials, 117-129.

Iguchi T. 2006. Importance of development of ecotoxicogenomics in understanding molecular mechanisms of chemicals in developing animals. Nippon Eiseigaku Zasshi 61(1):11-8.

Jager T. 2004. Modeling ingestion as an exposure route for organic chemicals in earthworms (Oligochaeta). Ecotoxicol Environ Saf 57(1):30-8.

Jenkins TF, Hewitt AD, Grant CL, Thiboutot S, Ampleman G, Walsh ME, *et al.* 2006. Identity and distribution of residues of energetic compounds at army live-fire training ranges. Chemosphere 63: 1280-1290.

Jirapech-Umpai T, Aitken S. 2005. Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. BMC Bioinformatics 6:148.

John GH, Kohavi R, Pfleger K. 1994. Irrelevant Features and the Subset Selection Problem. International Conference on Machine Learning:121-129.

Johnson MS, Holladay SD, Lippenholz KS, Jenkins JL, McCain WC. 2000. Effects of 2,4,6-trinitrotoluene in a holistic environmental exposure regime on a terrestrial salamander, Ambystoma tigrinum. Toxicol Pathol 28(2):334-41.

Jukes T, Cantor C. 1969. Evolution of protein molecules . In H. Munro, editor, Mammalian Protein Metabolism, pages 21-132. : Academic Press.

Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ. 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res 28(22):4552-7.

Kerr MK, Churchill GA. 2001. Statistical design and the analysis of gene expression microarray data. Genet Res 77(2):123-8.

Kestler HA, Muller A, Gress TM, Buchholz M. 2005. Generalized Venn diagrams: a new method of visualizing complex genetic set relations. Bioinformatics 21(8):1592-5.

Khanin R, Wit E. 2005. Design of large time-course microarray experiments with two channels. Appl Bioinformatics 4(4):253-61.

Kim JH, Shin DM, Lee YS. 2002. Effect of local background intensities in the normalization of cDNA microarray data with a skewed expression profiles. Exp Mol Med 34(3):224-32.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16(2):111-20.

Kinney EL, Murphy DD. 1987. Comparison of the ID3 algorithm versus discriminant analysis for performing feature selection. Comput Biomed Res 20(5):467-76.

Kohavi R, John GH. 1997. Wrappers for feature subset selection. Artif. Intell. 97(1-2):273-324.

Kumar CG, LeDuc R, Gong G, Roinishivili L, Lewin HA, Liu L. 2004. ESTIMA, a tool for EST management in a multi-project environment. BMC Bioinformatics 5:176.

Kuperman RG, Checkai RT, Simini M, Phillips CT. 2004. Manganese toxicity in soil for Eisenia fetida, Enchytraeus crypticus (Oligochaeta), and Folsomia candida (Collembola). Ecotoxicol Environ Saf 57(1):48-53.

Kuperman RG, Checkai RT, Simini M, Phillips CT, Kolakowski JE, Kurnas CW. 2006. Toxicities of dinitrotoluenes and trinitrobenzene freshly amended or weathered and aged in a sandy loam soil to Enchytraeus crypticus. Environ Toxicol Chem 25(5):1368-75.

Kuster H, Becker A, Firnhaber C, Hohnjec N, Manthey K, Perlick AM, Bekel T, Dondrup M, Henckel K, Goesmann A *et al.*. 2007. Development of bioinformatic tools to support EST-sequencing, in silico- and microarray-based transcriptome profiling in mycorrhizal symbioses. Phytochemistry 68(1):19-32.

Landgrebe J, Bretz F, Brunner E. 2004. Efficient two-sample designs for microarray experiments with biological replications. In Silico Biol 4(4):461-70.

Langseth H, Nielsen T. 2006. Classification using Hierarchical Naive Bayes models. Machine Learning 63(2):135-159.

Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A *et al.*. 2006. Machine learning in bioinformatics. Brief Bioinform 7(1):86-112.

Latorre M, Silva H, Saba J, Guziolowski C, Vizoso P, Martinez V, Maldonado J, Morales A, Caroca R, Cambiazo V *et al.*. 2006. JUICE: a data management system that facilitates the analysis of large volumes of information in an EST project workflow. BMC Bioinformatics 7:513.

Le Pecq JB, Le Bret M, Barbet J, Roques B. 1975. DNA polyintercalating drugs: DNA binding of diacridine derivatives. Proc Natl Acad Sci U S A 72(8):2915-9.

Lee ML, Bulyk ML, Whitmore GA, Church GM. 2002. A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. Biometrics 58(4):981-8.

Lee MS, Cho SJ, Tak ES, Lee JA, Cho HJ, Park BJ, Shin C, Kim DK, Park SC. 2005. Transcriptome analysis in the midgut of the earthworm (Eisenia andrei) using expressed sequence tags. Biochem Biophys Res Commun 328(4):1196-204.

Liang L, Ding YQ, Shi YM. 2003. Suppression subtractive hybridization and its application in study of tumors. Ai Zheng 22(9):997-1000.

Liu H, Setiono R. 1995. Chi2: Feature selection and discretization of numeric attributes.

Liu X, Hu C, Zhang S. 2005. Effects of earthworm activity on fertility and heavy metal bioavailability in sewage sludge. Environ Int 31(6):874-9.

Macarthur RH. 1957. On the Relative Abundance of Bird Species. Proc Natl Acad Sci U S A 43(3):293-5.

Maclin PS, Dempsey J, Brooks J, Rand J. 1991. Using neural networks to diagnose cancer. J Med Syst 15(1):11-9.

MacQueen J. methods for classification and analysis of multivariate observations; 1967. p 281-296.

Malaguarnera L. 2006. Chitotriosidase: the yin and yang. Cell Mol Life Sci 63(24):3018-29.

Mao C, Cushman JC, May GD, Weller JW. 2003. ESTAP--an automated system for the analysis of EST data. Bioinformatics 19(13):1720-2.

McCarter JP, Mitreva MD, Martin J, Dante M, Wylie T, Rao U, Pape D, Bowers Y, Theising B, Murphy CV *et al.*. 2003. Analysis and functional classification of transcripts from the nematode Meloidogyne incognita. Genome Biol 4(4):R26.

Moody J. E. and Darken C. 1989. Fast learning in networks of locally-tuned processing units. Neural Computation 1, pp. 281-294.

Nagaraj SH, Gasser RB, Ranganathan S. 2007. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8(1):6-21.

Narayanan A, Keedwell EC, Olsson B. 2002. Artificial intelligence techniques for bioinformatics. Appl Bioinformatics 1(4):191-222.

Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25(14):2745-51.

Nijkamp FP and Parnham MJ, editors. 2005. Principles of Immunopharmacology. 2nd ed. Birkhäuser; p.200

Noble WS. 2006. What is a support vector machine? Nat Biotechnol 24(12):1565-7.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27(1):29-34.

Paquola AC, Nishyiama MY, Jr., Reis EM, da Silva AM, Verjovski-Almeida S. 2003. ESTWeb: bioinformatics services for EST sequencing projects. Bioinformatics 19(12):1587-8.

Patel S, Lyons-Weiler J. 2004. caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer. Appl Bioinformatics 3(1):49-62.

Pavlidis P, Wapinski I, Noble WS. 2004. Support vector machine classification on the web. Bioinformatics 20(4):586-7.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al.*. 2000. Molecular portraits of human breast tumours. Nature 406(6797):747-52.

Pirooznia M, Deng Y. 2006. SVM Classifier - a comprehensive java interface for support vector machine classification of microarray data. BMC Bioinformatics 7 Suppl 4:S25.

Pirooznia M, Deng Y. 2007. Efficiency of Hybrid Normalization of Microarray Gene Expression: A Simulation Study. ainaw 1:739-744.

Plant N. 2006. Expressed sequence tags (ESTs) and single nucleotide polymorphisms (SNPs): what large-scale sequencing projects can tell us about ADME. Xenobiotica 36(10-11):860-76.

Quinlan R. 1993. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers.

Ramakers C, Ruijter JM, Deprez RH, Moorman AF. 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 339(1):62-6.

Reddy G, Chandra SAM, Lish JW, Qualls CW, Jr. 2000. Toxicity of 2,4,6-trinitrotoluene (TNT) in hispid cotton rats (Sigmodon hispidus): hematological, biochemical, and pathological effects. International Journal of Toxicology 19: 169-177.

Rim KT, Park KK, Sung JH, Chung YH, Han JH, Cho KS, Kim KJ, Yu IJ. 2004. Gene-expression profiling using suppression-subtractive hybridization and cDNA microarray in rat mononuclear cells in response to welding-fume exposure. Toxicol Ind Health 20(1-5):77-88.

Rombke J, Jansch S, Didden W. 2005. The use of earthworms in ecological soil classification and assessment concepts. Ecotoxicol Environ Saf 62(2):249-65.

Sabbioni G, Wei J, Liu YY. 1996. Determination of hemoglobin adducts in workers exposed to 2,4, 6-trinitrotoluene. J Chromatogr B Biomed Appl 682(2):243-8.

Sasik R, Calvo E, Corbeil J. 2002. Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. Bioinformatics 18(12):1633-40.

Scholkopf B, Platt o, Shawe-Taylor J, Smola A, Williamson R. 1999. Estimating the support of a high-dimensional distribution.

Scholkopf B, Smola o, Williamson R, Bartlett P. 2000. New support vector algorithms. 12:1207–1245.

Schulman P. 1984. Bayes' theorem--a review. Cardiol Clin 2(3):319-28.

Soetaert A, Moens LN, Van der Ven K, Van Leemput K, Naudts B, Blust R, De Coen WM. 2006. Molecular impact of propiconazole on Daphnia magna using a reproduction-related cDNA array. Comp Biochem Physiol C Toxicol Pharmacol 142(1-2):66-76.

Sturzenbaum SR, Parkinson J, Blaxter M, Morgan AJ, Kille P, Georgiev O. 2004. The earthworm Expressed Sequence Tag project. Pedobiologia 47(5-6):447-451.

Sun BC, Ni CS, Feng YM, Li XQ, Shen SY, Dong LH, Yuan Y, Zhang L, Hao XS. 2006. Genetic regulatory pathway of gene related breast cancer metastasis: primary study by linear differential model and k-means clustering. Zhonghua Yi Xue Za Zhi 86(26):1808-12.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10(3):512-26.

Taniguchi M, Miura K, Iwao H, Yamanaka S. 2001. Quantitative assessment of DNA microarrays--comparison with Northern blot analyses. Genomics 71(1):34-9.

Tchounwou PB, Wilson BA, Ishaque AB, Schneider J. 2001. Transcriptional activation of stress genes and cytotoxicity in human liver carcinoma cells (HepG2) exposed to 2,4,6-trinitrotoluene, 2,4-dinitrotoluene, and 2,6-dinitrotoluene. Environ Toxicol 16(3):209-16.

Townsend JP. 2003. Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays. BMC Genomics 4(1):41.

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98(9):5116-21.

van der Schalie WH, Gentile JH. 2000. Ecological risk assessment: implications of hormesis. J Appl Toxicol 20(2):131-9.

van Eijk M, van Roomen CP, Renkema GH, Bussink AP, Andrews L, Blommaart EF, Sugar A, Verhoeven AJ, Boot RG, Aerts JM. 2005. Characterization of human phagocyte-derived chitotriosidase, a component of innate immunity. Int Immunol 17(11):1505-12.

Vapnik V. 1998. Statistical Learning Theory. New York: Wiley.

Venn J.1880. On the diagrammatic and mechanical representation of propositions and reasonings. London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 5th ser., vol. 10, pp. 168-171.

Vinciotti V, Khanin R, D'Alimonte D, Liu X, Cattini N, Hotchkiss G, Bucca G, de Jesus O, Rasaiyaah J, Smith CP *et al.*. 2005. An experimental evaluation of a loop versus a reference design for two-channel microarrays. Bioinformatics 21(4):492-501.

Wang J, Jemielity S, Uva P, Wurm Y, Graff J, Keller L. 2007. An annotated cDNA library and microarray for large-scale gene-expression studies in the ant Solenopsis invicta. Genome Biol 8(1):R9.

Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW. 2005. Gene selection from microarray data for cancer classification--a machine learning approach. Comput Biol Chem 29(1):37-46.

Welsh JB, Zarrinkar PP, Sapinoso LM, Kern SG, Behling CA, Monk BJ, Lockhart DJ, Burger RA, Hampton GM. 2001. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci U S A 98(3):1176-81.

Wilkes T, Laux H, Foy CA. 2007. Microarray data quality - review of current developments. Omics 11(1):1-13.

Wu X, Zhu L, Guo J, Zhang DY, Lin K. 2006. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. Nucleic Acids Res 34(7):2137-50.

Xing EP, Jordan MI, Karp RM. 2001. Feature selection for high-dimensional genomic microarray data. ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning:601-608.

Yang GP, Ross DT, Kuang WW, Brown PO, Weigel RJ. 1999. Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. Nucleic Acids Res 27(6):1517-23.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30(4):e15.

Zhang D, Zhang M, Wells MT. 2006. Multiplicative background correction for spotted microarrays to improve reproducibility. Genet Res 87(3):195-206.