Summer 8-2009

Bridging Functional Genomics and Toxicogenomics Through DNA Microarrays in a Fish Model

Shuzhao Li
University of Southern Mississippi

The University of Southern Mississippi

BRIDGING FUNCTIONAL GENOMICS AND TOXICOGENOMICS THROUGH DNA MICROARRAYS IN A FISH MODEL

by

Shuzhao Li

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

Approved:

[redacted]

August 2009

The University of Southern Mississippi

BRIDGING FUNCTIONAL GENOMICS AND TOXICOGENOMICS THROUGH DNA MICROARRAYS IN A FISH MODEL

by

Shuzhao Li

Abstract of a Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

August 2009

ABSTRACT

BRIDGING FUNCTIONAL GENOMICS AND TOXICOGENOMICS THROUGH DNA MICROARRAYS IN A FISH MODEL

by Shuzhao Li

August 2009

In a case study of finding gene expression signatures for environmental stressors in *Cyprinodon variegatus*, this dissertation examines several important issues of applying DNA microarray technology to fish toxicogenomics. The most relevant disciplines, fish toxicogenomics and computational systems biology, are reviewed in Chapter 1. Chapter 2 reviews major aspects of DNA microarray technology.

On DNA microarrays, even for probes that target the same transcript, large variations are seen in the probe signals. These variations are partly dependent and partly independent on probe sequences. Chapter 3 estimates the sequence independent variation by combining experimental and computational approaches. Chapter 4 and 5 take on the central problem of sequence dependent variations, that is, modeling the physiochemistry of microarray hybridization. I propose a new competitive hybridization model, which demonstrates good success on publically available benchmark data. This new model leads the way to quantification of absolute target concentration, and brings critical insights into probe design and data interpretation of DNA microarrays. Our model relies on the accuracy of computing duplexing energy, which does yet not take into account secondary structures of probes and targets. I further explore the structural effects in Chapter 6.

After one obtains microarray data, the interpretation relies on existing knowledge of functional genomics, which come mostly from model organisms other than fish. As an effort to bridge this gap, a project to construct a genome-wide fish metabolic network, MetaFishNet, is launched. MetaFishNet is based on five fish genome sequences and the latest progress in metabolic modeling, especially the two high-quality human metabolic models. Chapters 7 to 9 describe the construction process of MetaFishNet. Chapter 10 demonstrates the two roles of MetaFishNet: a tool for interpreting high throughput expression data and a systems biology framework for hypotheses generation and study design.

Chapter 11 takes these methodological developments into the toxicogenomics of *C. variegatus.* We have constructed a *Cyprinodon* DNA microarray, and used it to profile the gene expression of larvae exposed to hypoxia, cadmium, chromium and pyrene. The result shows that specific markers can be identified for each stressor, and stressors can be classified by transcriptomic profiles. MetaFishNet enables us to perform Gene Ontology analysis and metabolic pathway analysis on these data. "Leukotriene metabolism" and "Xenobiotics metabolism" pathways appear to be upregulated by cadmium exposure.

FOR MY PARENTS

ACKNOWLEDGMENTS

I am grateful to Drs. Marius Brouwer and Alex Pozhitkov for their guidance and discussions. I thank Rachel Ryan, Steve Manning, Nancy Brown-Peterson and Jonathan Roling for their technical assistance.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

Introduction

Fish are abundant and diverse in aquatic ecosystems. Besides their economic and ecological importance, many fish species are of significant scientific value. Small fish species are widely used in ecological and pharmaceutical toxicology, developmental biology and genetics, evolutionary biology and as human disease models.

A prominent example in developmental biology is zebrafish, which gained popularity in a few decades for good reasons. Their embryos are transparent, and develop outside of the female body. Hundreds of eggs can be obtained from a handful of fish in the laboratory at controlled timing. Over the years, an array of molecular and genetic techniques has matured [1], including those for transgenesis, expression studies, forward and reverse genetics and in vivo imaging [2–5]. A large collection of mutants is maintained. In the past few years, the use of zebrafish as human disease models has spiked significant interests [6–8].

One of the many examples of using fish for evolutionary studies is the three-spine stickleback [9]. At the end of the last Ice Age about 10,000 to 15,000 years ago, thousands of fresh water habitats were created and colonized by ocean sticklebacks. These sticklebacks underwent rapid speciation, resulting in a large diversity of different sizes, shapes and physiological adaptations. This provides an excellent case for understanding the evolutionary process, complex traits in natural populations and the genetic basis of morphology.

Fish have been particularly important in ecotoxicology. The species used in toxicological studies include zebrafish (*Danio rerio*), medaka (*Oryzias latipes*), European flounder (*Platichthys flesus*), channel catfish (*Ictalurus punctatus*), sheepshead minnow (*Cyprinodon variegates*), mummichog (*Fundulus heteroclitus*), Atlantic salmon (*Salmo salar*), common carp (*Cyprinus carpio*), rainbow trout (*Oncorhynchus mykiss*), swordtail *Xiphophorus hellerii* and more.

Different species often have their unique, attractive features. For example, *Xiphophorus* is both a unique cancer model and a good evolutionary model of live-bearing vertebrates. Because small fish are currently the only vertebrate species that can be studied in high throughput, their future in modern biomedical sciences is brighter than ever.

1.1 Fish toxicogenomics

As molecular biology evolves towards genomic biology, the field of fish studies is also being transformed by newer methodologies. Thus far, genomes of five fish species have been completely sequenced. Among them, zebrafish and medaka are prominent laboratory models. Fugu and its close relative, green spot puffer, were sequenced primarily to help annotate human genome because of their very compact genome sizes - the comparison between these genomes helps to identify the genic units. The genomic sequence of green spot puffer also further substantiated the evolutionary genome duplication of teleost fish [10]. A brief summary on the genomic resources of model fish species is given in Table 1.1.

Among the aforementioned fish species, zebrafish is the only species for which commercial DNA microarrays are available. Therefore, microarray related literature leans heavily towards zebrafish. Various custom arrays have been constructed for a number of species and applied to ecotoxicological studies [11–13]. Nowadays, *in situ* synthesized DNA microarrays can be custom ordered from Nimblegen (now part of Roche) and Agilent Inc. at a reasonable cost. Given sufficient genomic resources, which will be no longer limiting as the cost of sequencing decreases drastically, mi-

Table 1.1: Current genomic resources on model fish species

species	common name	genomic resources	major research areas
Danio rerio	zebrafish	complete genome sequence, DNA microarrays, large collection of mutations	developmental biology, toxicology, disease models
Oryzias latipes	medaka	complete genome sequence, DNA microarrays, large genetic stocks	genetics, toxicology, developmental biology, carcinogenesis
Takifugu rubripes	fugu	complete genome sequence	comparative genomics, evolution
Tetraodon nigroviridis	spotted green puffer	low coverage genome sequence, not completely assembled	comparative genomics, evolution
Gasterosteus aculeatus	stickleback	complete genome sequence	species differentiation, comparative genomics
Xiphophorus maculatus	common platy	transcriptome sequences	genetics, toxicology, disease models, reproduction
Salmo salar	salmon	cDNA, ESTs, microarrays	reproduction, aging, toxicology
Oncorhynchus mykiss	trout	cDNA, ESTs, microarrays	carcinogenesis, toxicology

croarray tools are becoming more accessible.

In addition to toxicogenomics, proteomics and other omic technologies have become important fields of research in modern toxicology, and researchers have called for an integrated systems toxicology [14,15]. Similar methodologies are shared by ecotoxicology and pharmaceutical toxicology, part of drug discovery pipeline [16].

1.2 Computational systems biology

It is foreseeable that high throughput technologies will generate data at an exponentially increasing speed, and computational tools and models become a critical part of the emerging paradigm of systems biology. All the omic data need support of rigorous quantification methods and databases. Probably the methodologies around Sanger sequencing are a lot more mature than other still emerging technologies, as processing algorithms for these emerging technologies are often by themselves active research subjects. For example, the design and interpretation of DNA microarrays have been the subject of many controversies and debates [17–23]. The accuracy of major sequencing technologies is usually good, while the analyses of other omics data should always bear a cautionary note on false discovery rates.

The motivation of integrating different data types goes beyond the filtering of false positives. It helps us understand the multiple facets of living systems in the format of testable models. Models are tools that connect data with a predictive power. Molecular networks, including metabolic networks, signaling networks, protein interaction networks and gene regulatory networks have become pillar frameworks for systems biology [24–30].

Molecular mechanisms are the central piece that researchers are after in both disease studies and toxicology. Most of the emerging omics technologies have demonstrated great potential in finding biomarkers, understanding "mode of action" and dissecting causal pathways. They also present unique challenges to the fish models that have been discussed in this text.

Overall, the fish genomic resources are far behind other model organisms. With the sequencing costs dropping dramatically, a lot more sequences should become available in the near future. The upcoming genomic sequences will also lay the foundation for the applications of other omics technologies. Other types of data, for instance interactomes, may be more difficult to catch up. In most cases, we will have to bridge the functional genomic data from other model organisms via comparative genomics. The molecular networks in fish species can be initially constructed by ortholog mapping. These provisional models can then be refined from experimental data that are pertinent to the species. This practical solution points to a significant amount of computational work. As Waters and Fostel rightfully identified, "straightforward interpretation of global molecular-datasets derived from -omics technologies is currently constrained by the 'bioinformatics bottleneck' " [14]. An additional outcome from such computational modeling is the similarities and dissimilarities between model animals and human subjects, a salient point on how one extrapolates the experimental results.

This dissertation is a case study on the use of DNA microarray technology to interrogate the transcriptomic response of *Cyprinodon variegatus* (sheepshead minnow) to environmental stressors. The major obstacles involved in the process typify many of the challenges researchers face in this postgenomic era, in particular, the modeling of these highly parallel assays chemically and statistically, consistent data analysis, result interpretation and connecting to the existing knowledge body. I will give a brief review on the multiple aspects of DNA microarrays in Chapter 2. The rest of this dissertation consists of three major sections. Part I (chapters 3 to 6) focuses on the fundamental mechanisms of DNA microarrays. Part II (chapters

7 to 10) describes the construction of a fish metabolic network model, which is useful both as a tool to interpret high throughput expression data and a framework for further systems studies. Part III (chapter 11) takes these methodological developments into the toxicological study of sheepshead minnows.

Chapter 2

A compendious review of DNA microarrays

DNA microarrays are the best example of high throughput technologies in functional genomics. The technology has undergone a significant evolution, along with controversies [17,18,20,23,31–34] and improving standardizations [35–37].

The general mechanism of DNA microarrays is that polynucleotide probes of different sequences are attached to a surface at spatially defined positions, and labeled targets hybridize to their complementary probes to give a fluorescent readout. Due to miniaturization and parallelization, thousands of target genes can be simultaneously measured. The probes can be presynthesized then chemically attached to the chip surface, or synthesized in situ on the chip. A third array class comes from Illumia Inc., using micro-beads as reaction carriers. This dissertation will focus on the microarray platforms of surface hybridization.

Although the hybridization between probes and targets is based on Watson-Crick base pairing, there are important thermodynamic properties affecting the outcome. Thus, good modeling of the hybridization process is critical for designing probes and interpreting the signal intensities. This aspect of "upstream" informatics will be discussed in the second section of this chapter. I will first review the better known "downstream" informatics of microarrays.

2.1 Downstream informatics of DNA microarrays

The low level processing of microarray data involves image processing, normalization and sometimes filtering of data abnormality. Mature methods and implementations of these procedures can be easily found, and will not be discussed here. The important question of how to summarize gene expression abundance and infer differentially expressed genes led to a good number of statistical developments. The more popular programs include dChip [38], RMA [39], Limma [40] and PLIER [41]. Two aspects of these tools, summarizing probe signals and modeling statistical variants, are again subjects of later sections. Here, we shall give a brief review on the assessment of statistical significance, namely, how to determine differentially expressed genes.

Standard statistical techniques, such as t-test and ANOVA, can be applied to microarray data. However, since most microarray studies have a much smaller sample size relative to their feature size (number of probes or genes), this imbalance of dimensions severely compromises the statistical power [23]. In the case of ANOVA, variants may not be completely covered by experimental design [42]. Therefore, a "shrinkage" technique, that is, to borrow information from other genes towards gene specific statistics, is very useful and incorporated in popular programs such as SAM [43], Cyber-T [44] and MAANOVA [45].

Gene clustering is a common, unsupervised technique for dimension reduction, which is welcome because of the overwhelmingly large number of genes that are typically shown in microarray results. A good number of algorithms and implementations can be found for clustering microarray data. However, given the high noise level of microarrays, it should be noted that this technique is highly exploratory. Without knowing the "true" biological clusters, a definitive validation of clustering results is not possible [21, 22]. And "clusters" can always be found in any data at genomic scales.

Transcription profiles, as a measure of cellular and physiological states, have raised great interest in identifying biomarkers and predictive molecular signatures. Computationally, this type of studies generally fall into the category of supervised learning or supervised classification. A good example is to apply microarray data to classifying cancer types. Due to the complexity and multiple steps involved, early studies also raised concerns [46]. For instance, two high profiles studies published prognostic gene markers for breast cancer [47, 48]; however, little agreement was found between their results [49]. Differences in platforms and genetic backgrounds might be responsible to some extent. For laboratory studies, careful controls and validations should be implemented. A particular class of techniques, support vector machines (SVM), have shown good promises in this field [50–54]. I will apply SVM to analyzing our data from Cyprinodon microarrays.

The challenge of understanding gene expression data comes not only from their high noise level, but also from the fact that microarray data only provide a snapshot of cellular dynamics of mixed cell populations. Many molecular events are not captured by the measurements. It is therefore helpful to analyze the data in the context of certain types of functional units like signaling pathways or functional modules. With the help of Gene Ontology (GO) [55], contextual enrichment in gene categories can be examined for microarray data. Leading GO enrichment programs include GOminer [56, 57], DAVID [58, 59] and GOstat [60]. Most of these programs use Fisher's exact test to rank the significance of gene set enrichment.

Ultimately, biologists want to know what upstream and downstream molecular events are behind the experimental results, and how a pathway or a network of molecules function in concert. In this sense, good network models of molecular interactions are valuable references. They can be used for enrichment analysis in a fashion similar to GO tools. More importantly, they provide a framework to integrate multiple data types and connect to the existing knowledge body [61].

2.2 The Achilles' heel of DNA microarrays

Microarray probes produce different signals even when they interrogate the same target transcript in the same hybridization reaction (Figure 2.1). These variations are partly dependent and partly independent on probe sequences. As the technologies have improved significantly over the past decade, sequence independent variations, or technical errors, should also have decreased. Chapter 3 will address the sequence independent variations on the spotted array platform used in our lab. The more important aspect of sequence dependent variation comes from a deeper root, the lack of understanding of microarray hybridization.

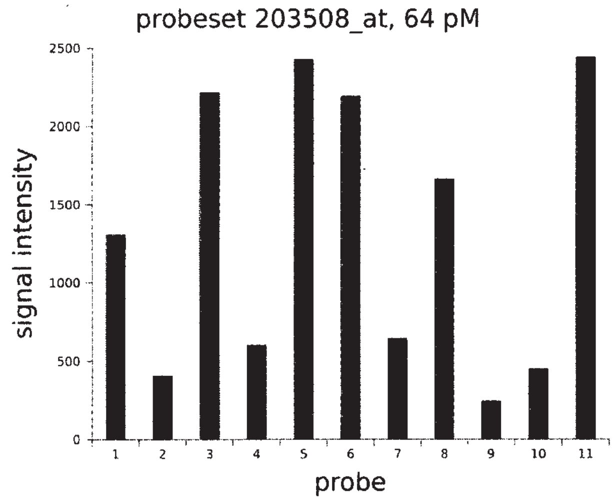


Figure 2.1: Microarray probes targeting the same transcript give very different signals. An example probe set from Affymetrix Latin square data [62].

Probes of different sequences have different thermodynamic features. However, how these thermodynamic properties translate into hybridization signals is not clear. Affymetrix arrays use mismatch probes as internal controls, but the usefulness of these mismatch probes is doubtful and they even give higher signals than perfectly matched probes [63, 64]. The current quantification of gene expression levels on DNA microarrays is usually a statistical estimation that is neither absolute value, nor comparable between different chip designs. This lack of understanding in microarray hybridization deprives a theoretical guide from probe design and data interpretation. As a result, current probe design either takes expensive errors and trials, or leaves hidden problems in subsequent processes. This problem in understanding hybridization is the "Achilles' heel of DNA microarrays", responsible for many of the issues in reproducibility and interoperability of microarray technology.

To model the physiochemistry of microarray hybridization, one has to account for both the thermodynamics and the kinetics. In connection to experimental data, a model has to explain the response of probe signals to different target concentrations.

Chemists started investigating this problem with simplified experiments, where a limited number of probes are used and complex backgrounds are avoided. In these simplified experiments, the hybridization process followed a Langmuir-like model. We will examine the applications and limitation of Langmuir models in the next few chapters. Chapter 3 focuses on the sequence independent variation on DNA microarrays; Chapter 4 through 6 focus on sequence dependent variations.

Part I

Computational modeling of DNA microarrays

Chapter 3

Sequence independent variation on DNA microarrays

Summary

It is a common assumption in the modeling of DNA microarray data that variation of spot signals is partly dependent and partly independent on probe sequences. The separate contributions of these two components, however, have not been previously characterized. A consensus from recent kinetic studies is that microarray hybridization follows a Langmuir or Sip model at the absence of a complex background. Based on this model, we were able to assess the sequence independent variation via a genetic algorithm. Our data on spotted DNA microarrays indicated a sequence independent variation at about 20%. Its major source is likely to be the variation of effective probe density. This method produces an experimental distribution of a key variable in microarray data, therefore provides a foundation for constructing parametric methods of statistical analysis. Our result also suggested that the sequence dependent variation was about 80%. If the sequence dependent variation can be successfully predicted, the quantifying power of current DNA microarray technology can still be largely improved.

3.1 Background

The understanding of experimental noise is obviously critical to quality assurance and data analysis, and has been the goal of a number of previous studies [33, 65–69]. However, these previous studies were based on data summarized at the gene level - none of them addressed the large variations at the probe level. Probe level variation is usually viewed as partly dependent and partly independent on probe sequences. Most statistical programs, including the popular dChip [38], RMA [39] and PLIER [41], contain parameters to model both types of factors. But their separate contributions to the overall variation, how much from sequence dependent factors and how much from sequence independent factors, have never been properly addressed, mainly because there was no plausible physical model to describe the sequence dependent factors; and fitting statistical models as a whole to biological data does not reveal this information. The importance of this question goes beyond the continued improvement of statistical analyses. The whole field of modeling microarray hybridization is based on the presumed importance of the sequence dependent factors. If the modeling efforts should improve DNA microarray technology [70, 71], how much they do so will probably have an impact on the future of the technology, which is already threatened by the next generation sequencing [72]. This chapter starts by assessing the sequence independent variations on DNA microarrays.

It has recently emerged that DNA microarray hybridization, when performed without a complex background, follows a Langmuir or Sip model [73–80]. At low target concentration, *i.e.*, far from saturation, both Langmuir and Sip models can be simplified to

$$S = \varepsilon \cdot k \cdot P \cdot T \quad , \tag{3.1}$$

where S is observed signal intensity, k probe specific binding constant, P effective probe density (also called hybridization density), T target concentration, and ε random error. Note that ε is multiplicative here, consistent with previous literature. As observed in experimental data, S can change several orders of magnitude for different T, therefore any additive random error would be trivial. Equation (3.1) may look similar to the multiplicative statistical models, but it is a physical model restricted to low target concentration. Microarray hybridization with a complex background may not follow Langmuir models [71]. But in either case, the linear assumption in the statistical models does not hold true.

In Equation (3.1), k is the only parameter that is dependent on probe sequence. T can be experimentally set to a constant. Then $\varepsilon \cdot P$ represents the sequence independent portion of the signal. We set forth designing experiments to assess the separate contributions of k and $\varepsilon \cdot P$ to the overall variation of S.

3.2 The experimental data

In order to assess the variation of k and $\varepsilon \cdot P$ effectively, we need data that are discriminative on k. In a matrix of spot signal intensities:

$$\mathbf{S} = \begin{pmatrix} S_{1,1} & S_{1,2} & \dots & S_{1,N} \\ S_{2,1} & S_{2,2} & \dots & S_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ S_{M,1} & S_{M,2} & \dots & S_{M,N} \end{pmatrix}$$

We let each column in S have the same probe sequence therefore the same k. Then the sequence independent variation should be random across the whole matrix, while k only differs between columns. Our experiments were carried out on spotted microarrays, where $N = 95$ and $M = 7$ (Figure 3.1). All these 95 probes were designed to hybridize to the same transcript at varying positions. Using a single target species in a constant flow chamber makes T in Equation (3.1) a constant. The M batches, each from a separate source plate, were prepared in different spotting concentrations. This is not necessary, but a reflection of the often needed calibration of spotting concentrations as well as the difficulty in a precise control of concentrations in real experiments. Although signal intensities tend to increase when we increase probe spotting concentration during the experiments, the precise relationship between probe density and probe spotting concentration is unknown.

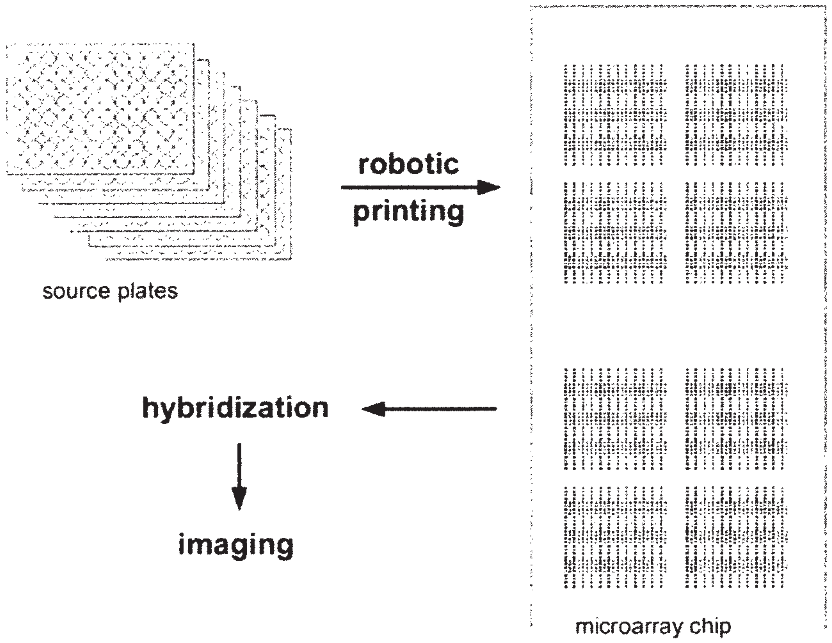


Figure 3.1: The schematic of experimental procedures for assessing variation. Seven source plates of different probe concentrations, each containing 95 probes and one dye control, were robotically spotted onto a microarray slide. With each spot repeated on a slide eight times, the resulting slide contained 5376 spots. The real spotting pattern is shown. Slides were then hybridized to labeled targets at a fixed concentration, and imaged for fluorescent signals. See text and Methods for details.

We performed three repeated experiments, with target concentrations 0.780 (slide C8), 0.780 (slide B9) and 1.365 (slide F9) $ng/\mu l$. The target concentrations were chosen based on two considerations: 1) hybridization signals should be distinctive from background; 2) the target concentration should be far from saturation and therefore satisfy the model in Equation (3.1). Extra hybridization experiments that used a wide range of target concentrations confirmed that these two requirements were met (Figure 3.2). The target concentration had no real effect in the following calculation because it is a single fixed value for each hybridization.

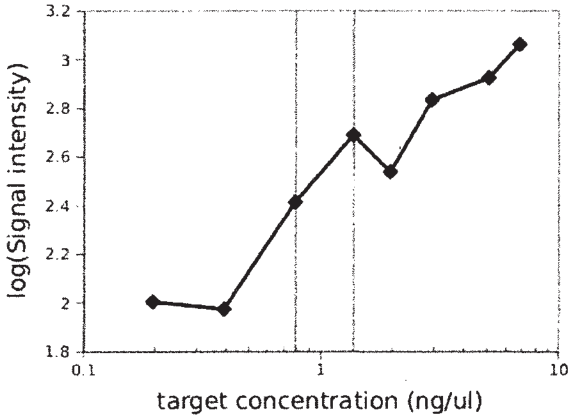


Figure 3.2: Choosing target concentrations. A typical response of hybridization signal to varying target concentrations. The target concentrations used in this study (0.780 and 1.365 $ng/\mu l$, purple vertical lines) were far from saturation.

The variation of spot size is common in this type of microarrays. During our image quantification, the spot size was automatically adjusted (Figure 3.3). Since the signal intensity of a spot was taken as the mean of pixel intensities, spot size did not participate in the model of Equation (3.1). The local background in our hybridizations was uniform (Figure 3.4). Thus, the hybridization quality was not affected by spatial disparity.

3.3 The mathematical formulation

We define "variation" as the coefficient of variation in this study. *E.g.*, the "variation" in a set of S_i is $\mathbf{var}_{i=1}^{N}(S_i) = \sigma/\mu$, where σ is the standard deviation and μ the mean. It is obvious that in Equation (3.1), the sequence dependent variation is $\mathbf{var}(k)$ and the sequence independent variation is $\mathbf{var}(\varepsilon \cdot P)$. In a real

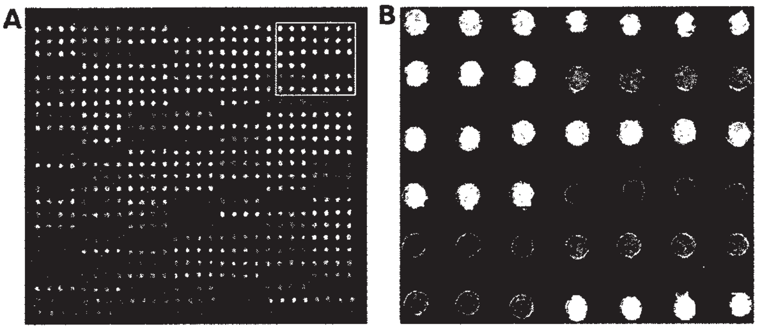


Figure 3.3: Spots sizes were adjusted for quantification. (A) Fluorescent image of a complete block of spots from slide B9, ready for quantification. (B) Close-up of the boxed area in (A). The blue circles in (B) defined the spot areas for quantification. Since spot signals were taken as the mean value of all its pixels and probe density was used in Equation (4.4), our analysis was not affected by spot size.

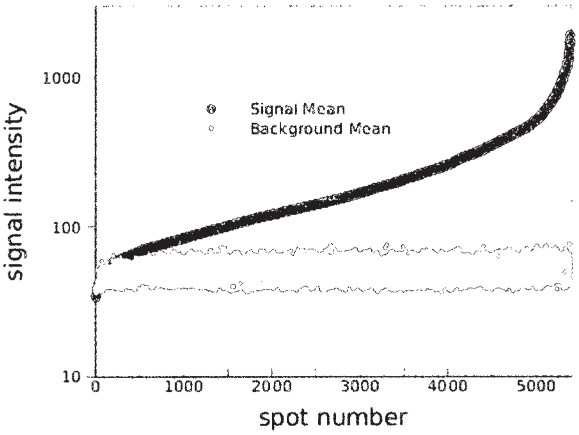


Figure 3.4: Local background is uniform, independent from spot signal intensity. All spots are sorted by signal intensities (blue dots), and purple dots are their corresponding local backgrounds.

microarray experiment, neither k or P can be readily measured. It should be emphasized that the effective probe density is different from raw probe density. Many probes are attached to the array surface yet fail to participate in hybridization, because of inaccessibility to targets or defective oligo synthesis. Thus, the measurement of effective probe density, P, has to involve target hybridization. Various techniques, ranging from optical methods to nanomechanics [73–75, 81–83], have been used to monitor hybridization in kinetic studies. However, all these methods are limited by specific experimental settings and instrument requirements. For example, the popular method of surface plasmon resonance requires a metal surface while most DNA microarrays in the market are manufactured on a glass surface. And none of these techniques can be easily adapted to the small feature size and high density of real DNA microarrays. In this study, we resort to a computational method.

We observe from Equation (3.1) that under constant T, $\mathbf{var}(\varepsilon \cdot P) = \mathbf{var}(S \cdot k^{-1})$. Let

$$\mathbf{V} = \sum_{i=1}^{M} \mathop{\mathbf{var}}_{j=1}^{N}(S_{i,j} \cdot k_j^{-1}) \quad , \tag{3.2}$$

$$\mathbf{U} = \sum_{i=1}^{M} \mathop{\mathbf{var}}_{j=1}^{N}(S_{i,j} \cdot f_j) \quad , \tag{3.3}$$

then the problem is defined as:

given a set of $\mathbf{S}$*, if a set of* f_j *can be found so that* $\mathbf{U}$ *is minimal, then* $\mathbf{U}$ *is also the minimum limit of* $\mathbf{V}$*.*

Because if $k_j^{-1} \neq f_j$, $\mathbf{V}$ will be only greater than $\mathbf{U}$. Thus, $\mathbf{var}(S \cdot f_j)$ constitutes the minimum estimation of sequence independent variation $\mathbf{var}(\varepsilon \cdot P)$. Since $\varepsilon \cdot P$ is independent from k or f, given an adequate sample size, the estimation above should approximate the true value. Likewise, $\mathbf{var}(f^{-1})$ can be used to estimate $\mathbf{var}(k)$. For a sizable number of probes ($N > M$), an analytical solution of f_j is difficult to obtain. We will solve it via a genetic algorithm.

3.4 Computing f_j with genetic algorithm

A rough approximation of f_j could be

$$\left(\frac{1}{M}\sum_{i=1}^{M} S_{i,j}\right)^{-1} \quad . \tag{3.4}$$

However, this approximation rarely gives the optimal f_j and is sensitive to probe specific noise. Computing f_j is a suitable job for Genetic Algorithms (GA). GA is a biologically inspired, stochastic search algorithm based on Schema Theorem. The algorithm starts with a random pool of candidate solutions named "chromosomes". The "chromosomes", like their biological analogs, mutate and cross with each other under an evolutionary pressure (monitored by a fitness function). Improved solutions emerge during this course of "evolution". Well-implemented GA is proven to produce optimal solutions reliably [84]. Standard implementation and terminology of GA are used in this paper. We use Equation (3.3) as the fitness function, each set of f_j as a chromosome. Thus, the goal of chromosome evolution is to minimize $\mathbf{U}$. After a number of generations of evolution, a set of optimal solution of f_j should be produced (see Method for details).

We first tested our GA program on synthetic data that were generated according to Equation (3.1). Our program successfully recovered the predefined variations of $\varepsilon \cdot P$ from the synthetic data (Figure 3.5A). Using the real data from our hybridization experiments, the GA program converged well and outperformed the rough estimation in Equation (3.4) in every case (Figure 3.5B). All repeated runs produced stable results.

3.5 Sequence independent variation

After a set of f_j is computed by our GA program, the sequence independent variation in each preparation group can be obtained as:

$$\underset{j=1}{\overset{N}{\mathbf{var}}}(S_{i,j} \cdot f_j) \quad .$$

For the three example slides, the results were summarized in Table 3.1. The average sequence independent variation was 20.5%. Similarly, the average sequence

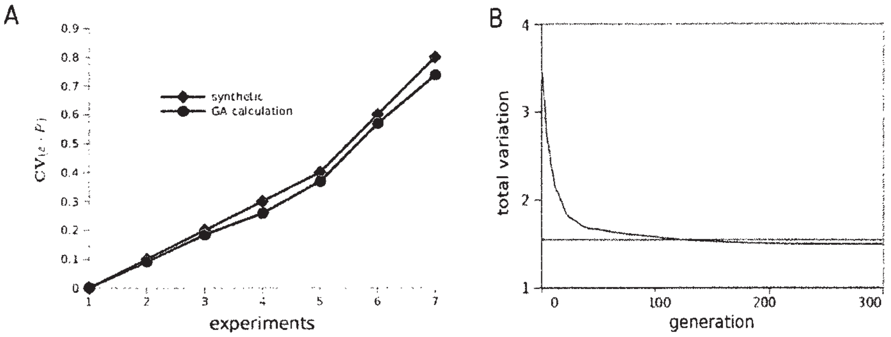


Figure 3.5: Performance evaluation of our Genetic Algorithm implementation. (A) Our program successfully recovered the variation in synthetic data. This GA approach sought the minimum limit of the variation, hence leading to a slight underestimation. (B) A representative performance graph of the GA (slide B9). Equation (3.3) was used as the fitness function on Y-axis. All GA runs converged in less than 300 generations (blue curve), and the results outperformed the rough estimation in Equation (3.4) (purple line). The solutions obtained by GA were also very different from those by Equation (3.4).

dependent variation was estimated by $\mathbf{var}(f^{-1})$ to be 82.1%.

A group with variation 20.9% is illustrated in Figure 3.6: signal intensity S is drawn in blue dots; the sequence independent variation is shown in purple line. Alternatively, the purple line in Figure 3.6 can be presented as a histogram in Figure 3.7. Some microarray programs try to avoid the sequence dependent variation by comparing spots of the same probe sequence. In this scenario, the variation of

Table 3.1: The estimated sequence independent variations, $\mathbf{var}(\varepsilon \cdot P)$, for three example slides. For each probe spotting concentration, the variation and standard deviation were obtained over three runs of GA algorithm.

spotting conc. (μM)	Slide B9	Slide F9	Slide C8
6.25	0.2375 ± 0.0005	0.2570 ± 0.0004	0.2596 ± 0.0005
9.38	0.1602 ± 0.0009	0.1883 ± 0.0007	0.1402 ± 0.0014
12.5	0.2183 ± 0.0006	0.1906 ± 0.0006	0.1904 ± 0.0003
18.75	0.1609 ± 0.0005	0.1809 ± 0.0007	0.1449 ± 0.0005
25	0.2145 ± 0.0004	0.1930 ± 0.0005	0.1451 ± 0.0004
37.5	0.2086 ± 0.0021	0.1882 ± 0.0008	0.2372 ± 0.0001
50	0.2934 ± 0.0008	0.2169 ± 0.0003	0.2949 ± 0.0011
average	0.21912 ± 0.0458	0.19391 ± 0.0267	0.20250 ± 0.0628

$\varepsilon \cdot P$ is the major concern, and, in the case of Figure 3.7, produces over 4% pair-wise comparisons of a fold change larger than two.

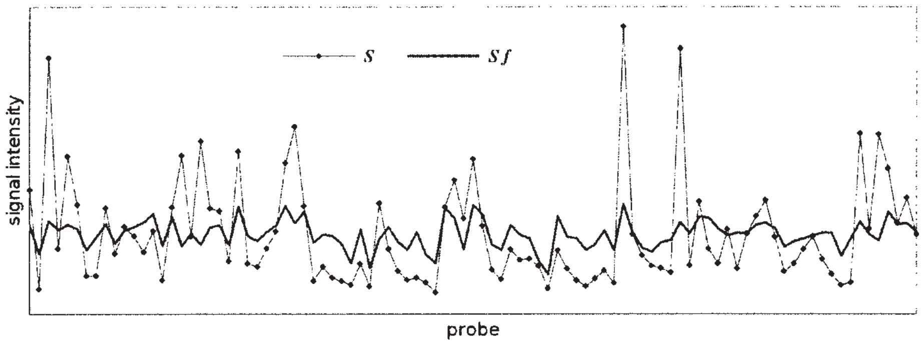


Figure 3.6: Sequence independent variation. Blue dots show the overall variation of signal intensities. The purple line $S \cdot f$ is an approximation of sequence independent variation, as explained in the text following Equation (3.3). This is an example from slide B9, group of spotting concentration 37.5 μM, $\mathbf{var}(S \cdot f) = 20.9\%$.

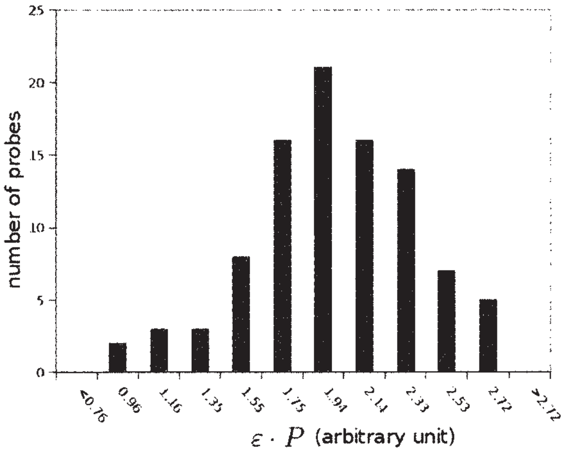


Figure 3.7: Distribution of $\varepsilon \cdot P$. Histogram representation of the purple line in Figure 3.6, with relative values of $\varepsilon \cdot P$ estimated by $S \cdot f$. If one compares $\varepsilon \cdot P$ among the 95 probes, 194 out of 4465 pair-wise comparisons, or over 4%, resulted in a fold change larger than two.

Repeated spotting (replicate spots) may increase accuracy by averaging the signal intensities of the repeats. We investigated this potential increase in Figure 3.8, where variations were estimated with the signal intensities averaged over 2 to

8 replicates. These replicates only slightly reduced the estimated variations, with 17.7% for 8 replicates.

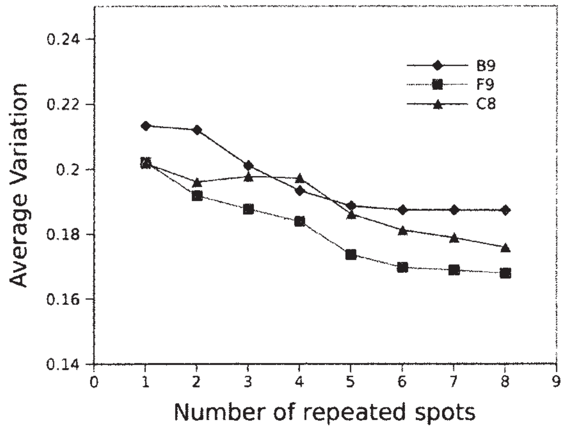


Figure 3.8: Effects from repeated spotting. When the averages of repeated spots are used, the estimated sequence independent variations slightly decreased. Blue diamonds: slide B9, purple squares: slide F9, yellow triangles: slide C8. The average variation with 8 repeated spots for all three slides was 17.7%.

3.6 Discussion

Although widely assumed in the modeling of microarray data, the separate impacts of sequence dependent and sequence independent variations have not been previously characterized. The consensus from recent kinetic studies can be formulated into Equation (3.1) at low target concentration, thus enabling us to assess the sequence independent variation, which was estimated at about 20% for our spotted DNA microarrays. This number is the minimum variation within a single chip for the particular platform, and should be a useful guide for experimental design and statistical analysis. Our data also suggested that probe sequences accounted for about 80% of the variation in probe signals - probably still higher for microarrays with better manufacturing methods. This highlights the importance of good physical models of microarray hybridization (to be discussed in the next two chapters).

If the sequence dependent variations can be predicted and the quantifying capacity of DNA microarrays is fully explored, the technology may still hold some cost advantage over the next generation sequencing.

The quality of microarray images in our study was good. We observed little spatial disparity, as shown by Figure 3 and Figure 4. Therefore, we think the dominating factor in the sequence independent variation is effective probe density. This agrees with the emphasis that previous studies placed on probe density, which may indeed confound the interpretation of microarray kinetic data [73, 75, 81, 85–89].

In the statistical analysis of microarray data, one can take a parametric approach or a resampling approach. Resampling approaches are now getting largely popular. Because the real statistical distributions of microarray variables, for example the sequence independent variation addressed here, are usually unknown, it is questionable to make inferences based on arbitrary assumptions [90]. On the other hand, resampling can also be biased on data of small sample sizes [23, 91, 92]. Because most resampling techniques (permutation, bootstrap, ect.) are computationally intensive, the running time could become inhibitory for large data sets. Our method in this chapter actually generates an experimentally measured variation distribution (e.g. Figure 3.7), which can provide a basis for constructing valid parametric methods. Then, such empirically supported parametric methods will avoid the caveats of resampling, and dramatically speed up the computation.

3.7 Methods

DNA Microarray Fabrication

DNA probes of 20mers were designed to complement *Cyprinodon variegatus* 18S rRNA, and synthesized with 5'-modification of Amine-C6 (Invitrogen Corp. Carlsbad, CA, USA). As illustrated in Figure 1, probes were prepared in 96-well format at concentrations of 6.25, 9.38, 12.5, 18.75, 25, 37.5 and 50 μM (one concentration per plate), and then deposited on SuperEpoxy slides (Thermo Fisher Scientific Inc. Waltham, MA, USA) with a VersArray ChipWriter Compact System (Bio-Rad Lab, Hercules, CA, USA). 95 probes were printed at 7 spotting concentrations with 8 on-slide repeats. An AlexaFluor 555 hydrozide dye (Invitrogen) was spotted along with each source plate as a quality control and reference for grid alignment in image processing. The resulted slide contained 5376 spots (including controls). Spotting was performed in 1x Micro Spotting Solution Plus(VWR, West Chester, PA, USA).

Hybridizations and data processing.

The *Cyprinodon* RNA target was directly labeled with AlexaFluor 546 using a ULYSIS Nucleic Acids Labeling kit (Invitrogen). Pre-hybridization, hybridization and all washing steps were performed in a-Hyb hybridization station (Miltenyi Biotec, Inc., Auburn, CA, USA). Printed microarray slides were pre-hybridized with BlockIt (VWR) blocking buffer for 1 h at 25 °C; then washed 5 times with 0.1% sodium sarkosylate at 25 °C for 2 min and finally 5 times with H_2O at 25 °C for 2 min. The hybridization mix contained 0.1 mM Na^+, 0.2 mg/ml of BSA and 0.1% Tween. The hybridization was conducted at 45 °C for 4h, after which the microarrays were washed 2 times with 2x SSC plus sodium sarkosylate at 25 °C for 2 min; 2 times with 2x SSC at 25 °C for 2 min and finally one time with 0.2x SSC at 10 °C for 1 min. The slides were scanned on a VersArray ChipReader 10 μm System scanner (Bio-Rad Lab). The images were quantified using ImaGene 7.0 software (BioDiscovery, Inc., El Segundo, CA, USA), with spot size automatically adjusted. Further data processing was done by custom Python scripts.

Genetic Algorithm for computing f_j.

The genetic algorithm (GA) in this study was implemented according to [84]. Each set of f_j was represented as a chromosome. The size of chromosome pool was 5000, crossover rate 0.7, mutation rate 0.01. Initial chromosomes were generated randomly. The fittest 1% chromosomes were retained in each generation. Our implementation was tested on synthetic data that were made according to Equation (3.1): both k and $\varepsilon \cdot P$ were drawn randomly from Gaussian distributions, k using standard deviations observed from real experiments while $\varepsilon \cdot P$ using predefined variations. For our experimental data, all GA runs converged in less than 300 generations, and repeated runs produced stable results. The data and the GA program used in this study are available upon request.

Chapter 4

A competitive hybridization model: part I

Summary

The previous chapter points out that the differences in signal intensities between microarray probes are mostly sequence dependent. However, the dominating theory of Langmuir models was only used under restrictions: low target concentration and no hybridization background. For real microarray experiments with complex backgrounds, Langmuir models have difficulties to correctly model the hybridization. Over the next two chapters, we will discuss the problems associated with both the Langmuir models and the thermodynamic (Nearest Neighbor) model in regard to microarray modeling. A new competitive hybridization model is then proposed to solve these problems. Our model uses a probe-specific dissociation constant that is computed with current Nearest Neighbor model and existing parameters, and only four global parameters that are fitted to Affymetrix Latin Square data. This model can successfully predict signal intensities of individual probes, therefore makes it possible to quantify the absolute concentration of targets. Our results offer critical insights into the design and data interpretation of DNA microarrays.

4.1 The Nearest Neighbor model

The free energy of polynucleotide hybridization in bulk solution has been successfully described by a Nearest Neighbor (NN) model. It has wide applications including the design of PCR primers. The NN model assumes the energy contribution of each nucleotide is determined by the combination with its immediate neighbors, and the overall hybridization energy can be obtained by stacking these nearest-neighbor counts:

$$\Delta G_d = \sum n_i \Delta G_i, \tag{4.1}$$

where ΔG_i is the energy contribution from one of the nearest neighbor combinations, $\{AA/TT,\ AT/TA,\ TA/AT,\ CA/GT,\ ...\}$, and n_i is the number of occurrences of each NN count. The NN model operates on an established set of empirically determined, position-independent parameters (10 parameters for DNA/DNA hybridization [93]; 16 for RNA/DNA hybridization [94,95]).

Although this NN model and its parameters have been well established for hybridization in solution, its application to microarray hybridization is a matter of debate. Experimental data from simple arrays seem to generally support the notion that hybridization energy on array surface correlates to that in solution [96–99]. However, the experimental conditions on these simple arrays differed considerably from "real" microarray experiments since they did not have complex background DNA/RNA species, and most had very limited number of probes. In "real" microarray settings, a few studies [100–102] suggested that binding energies based on the NN model had an important role in microarray hybridization, but no proof was adequate for a predictive model. The dominating opinion has been rather that NN model is not applicable to high density microarray hybridization, as it was either modified and re-parameterized [70, 103–106], or abandoned [63, 107, 108].

4.2 Langmuir-like models

Target concentration is a key parameter in microarray hybridization, which is incorporated via the kinetic models. The focuses have been on the Langmuir model and its derivatives [103–105, 107, 109–114]. Langmuir model is a generic mathematical form that also fits the description of first order chemical reactions, which is frequently used for probe-target binding on DNA microarrays:

$$\theta = \frac{T}{T+K}, \tag{4.2}$$

where θ is the fraction of occupied probes, T free target concentration, K dissociation constant.

According to the Langmuir model, all probes should saturate at the same level, which is clearly not the case in microarray hybridizations. Various modifications were proposed to accommodate this difference in saturation levels. A generic version may be written as

$$\theta = \frac{\chi T}{T+K}, \tag{4.3}$$

where χ is a probe specific factor. While a physical meaning of χ is difficult to obtain, some [104, 113] tried to explain χ through the washing step in microarray experiments. That is, all probes reach a same saturation level by the end of hybridization, but they lose the bound targets to different extents during the washing step. This "washing model" suggests a significant loss of signals upon each washing cycle. In experimental observations, the first washing cycle usually removes a considerable amount of partially bound targets, but it is clear that signal intensities do not decrease dramatically after extra washing cycles [115]. This contradicts the above "washing model". Furthermore, the Langmuir derivatives predict that, in response to increasing target concentrations, probes with higher binding affinities saturate first. In experimental observations, on the contrary, low affinity probes generally saturate first. Although Langmuir models seem to work well on simple surface hybridizations (as discussed in Chapter 3), no Langmuir derivative has adequately predicted probe signals in "real" experimental settings, such as those in the

Affymetrix Latin Square data with complex backgrounds [62].

The best prediction of probe signals to date was reported by Zhang et al. [70]. They accounted for both specific binding and nonspecific binding in the form of $\hat{T}/(1+K)$, where $\hat{T}$ is total target concentration, while fitting 83 parameters to the data. Mei et al. [108] also sought a linear composition of binding energy, where the single base energy contribution alone used 75 parameters. Over-parameterization has been a concern in all these previous studies and invited criticism on their general applicability [116].

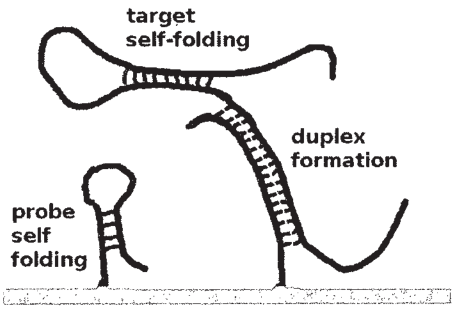
4.3 NN model is conditionally applicable

Both the Nearest Neighbor model (thermodynamics) and Langmuir-like models (kinetics) have disturbing issues in their applications to modeling DNA microarrays. It is easy to get trapped in the combined complexity. I will approach the problem step by step, beginning with the limitation of NN model.

The NN model does not take into account the secondary structures of polynucleotides. While on DNA microarrays, both probes and targets may form secondary structures by self-folding. As illustrated in Figure 4.1, while single-stranded probe (surface bound) and single-stranded target (in solution) form a double-stranded duplex, both probes and target molecules may form secondary structures by self folding. The secondary structure of probe will clearly interfere with the duplex formation. The secondary structure of target is a regional effect: it only affects duplex formation when the secondary structure is located in the very region complementary to the probe.

If we can select/enrich a subset of probes that are free from both probe and target folding, their hybridization can be simplified to:

$$^{SS}target +^{SS} probe \rightleftharpoons duplex \tag{4.4}$$



$$\begin{array}{ccc} {}^{DS}target & & {}^{DS}probe \\ \rightleftharpoons & & \rightleftharpoons \\ {}^{SS}target & + & {}^{SS}probe \rightleftharpoons duplex \end{array}$$

Figure 4.1: A simple model of microarray hybridization. A schematic presentation with three types of reactions: target self-folding, probe self-folding and the duplex formation between target and probe. In the grey box is the same model in chemical equations. Target structure is local - it only affects the hybridization when it is in the very region complementary to the probe. SS: single-stranded, DS: double-stranded.

with a binding constant K. At equilibrium, this type of reaction obeys the law of thermodynamics:

$$\Delta G_d = -RT \log K, \tag{4.5}$$

where ΔG_d is the Gibbs free energy of duplex formation, R the molar gas constant, T the absolute temperature; and K is the binding constant, reflected in measured signal intensities. Then a hypothesis can be tested on whether these probes free from structural effects follow the basic NN model in Equation (4.1).

An accurate prediction of the secondary structures is not easy. We will discuss it in detail in Chapter 6. As a first step, the secondary structure of probes can be investigated through their self-folding energy, ΔG_o (see method section in Chapter 5). The higher ΔG_o is, the less likely is the probe to form secondary structure. In the Affymetrix Latin Square data [62], about 45% of the probes can be selected by the criterion $\Delta G_o > -1$ kcal/mole. For these probes, a clear correlation appears between log signal intensities (SI) at the highest target concentration and the duplexing energy ΔG_d that are computed by the current NN model with existing parameters (Figure 4.2, $R^2 = 0.58$).

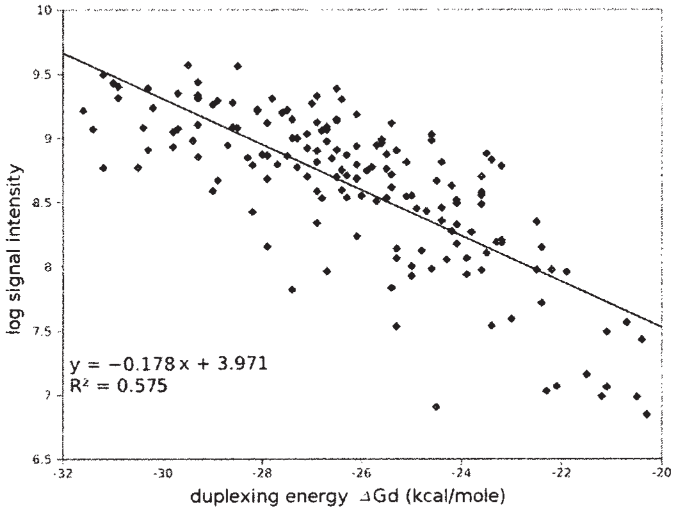


Figure 4.2: Correlation between $\log SI$-ΔG_d for probes with $\Delta G_o > -1$. Duplexing energy calculated by NN model correlates with signal intensity at 512 pM, the highest spike-in target concentration, for probes free of secondary structures ($\Delta G_o > -1$ kcal/mole, 160 probes from Affymetrix U133A data). Each dot represents a probe. Detailed data processing is described in the Method section of Chapter 5.

If the selection criterion is relaxed to $\Delta G_o > -2.5$ kcal/mole, 75% probes are included and the $\log SI$-ΔG_d correlation has $R^2 = 0.45$ (data not shown). However, the $\log SI$-ΔG_d correlation diminishes at lower target concentrations (not shown). These observations suggest that the current NN model offers a certain degree of predictability but they can not be accommodated by previous, Langmuir-like models. A new kinetic model is needed here.

4.4 A competitive hybridization model

We treat DNA microarray hybridization as two subprocesses, the binding of targets to probes and the dissociation of target/probe duplexes. Assuming that equilibrium is reached at the end of hybridization and the binding rate is the same for all target molecules (see below), the dissociation rate is governed by the duplexing energy between paired target/probe. A kinetic equilibrium between binding and dissociation should be observed.

Two types of targets are explicitly modeled: "specific targets" (perfect match) with probe-specific dissociation rate k_d, and "cross-hybridizing targets" with dissociation rate k_n. These cross-hybridizing targets are present in large quantities because partially matching sequences are abundant in a transcriptome. For the moment, we simplify them as a uniform mixture with a probe-nonspecific k_n.

The target/probe duplex formation is commonly believed to start with an initiation step, the base-pairing between a small number of nucleotide bases, and then extend to the rest of complementary regions [117, 118]. If the initiation step sets the rate limit, the binding rate should be hardly specific to probe sequences. We therefore assume a single binding rate, k_b, for all target molecules. How the specific factors [79, 80], including adsorption and electrostatics [119], steric and brush effects [120] and labeling [63, 121], come into play is not yet entirely clear. In this study, we postulate that the available area of probe spots is the limiting factor in adsorption, so that the binding is described as

$$\frac{\dot{n}_{in}}{N_A V} = (1 - \alpha - \beta) \cdot p \cdot k_b, \tag{4.6}$$

where $\dot{n}_{in}$ is the number of target molecules going into the exposed probes over a unit of time, N_A the Avogadro constant, V the volume of hybridization solution. On the right side, α is the fraction of probes bound to specific targets, β the fraction of probes bound to cross-hybridizing targets. p is the total number of probes in unit of molar concentration (for simplicity, as if they were dissolved in hybridization solution).

On the other hand, the dissociation is described by

$$\frac{\dot{n}_{out}}{N_A V} = \alpha \cdot p \cdot k_d + \beta \cdot p \cdot k_n, \tag{4.7}$$

where $\dot{n}_{out}$ is the number of target molecules leaving target/probe duplexes over a unit of time; k_d and k_n are dissociation rates for specific targets and cross-hybridizing targets, respectively.

At equilibrium between binding and dissociation,

$$(1 - \alpha - \beta) \cdot p \cdot k_b = \alpha \cdot p \cdot k_d + \beta \cdot p \cdot k_n \tag{4.8}$$

Equilibrium is established for both specific and cross-hybridizaing targets. The proportions of specific targets and cross-hybridizing targets are determined by their concentrations:

$$\alpha \cdot p \cdot k_d = \frac{\dot{n}_{in}}{N_A V} \cdot \frac{[T]}{[T] + [N]}, \tag{4.9}$$

$$\beta \cdot p \cdot k_n = \frac{\dot{n}_{in}}{N_A V} \cdot \frac{[N]}{[T] + [N]}, \tag{4.10}$$

where $[T]$ is the concentration of free specific targets, $[N]$ the concentration of free cross-hybridizing targets.

Equations (4.9) and (4.10) can be combined to express β as:

$$\beta = \frac{k_d[N]}{k_n[T]} \cdot \alpha \tag{4.11}$$

Then, Equations (4.8) and (4.11) give the fraction of specific binding

$$\alpha = \frac{1}{1 + k_d(\frac{1}{k_b} + (\frac{1}{k_n} + \frac{1}{k_b})\frac{[N]}{[T]})} \tag{4.12}$$

Here $[T]$, the concentration of free specific target molecules, is less than nominal spike-in concentration by the amount of probe binding:

$$[T] = \hat{T} - \alpha \cdot p \tag{4.13}$$

with $\hat{T}$ as the nominal spike-in concentration (total amount).

We assume the concentration of cross-hybridizing targets, $[N]$, is large and can be treated as constant in this model. Let

$$\gamma = (\frac{1}{k_n} + \frac{1}{k_b})[N], \tag{4.14}$$

then Equation (4.12) becomes

$$\alpha = \frac{1}{1 + k_d(1/k_b + \gamma/(\hat{T} - \alpha \cdot p))} \tag{4.15}$$

An analytical solution of Equation (4.15) is

$$\alpha = \frac{1}{p}\left(\Gamma - \sqrt{\Gamma^2 - \frac{p\hat{T}}{1 + k_d/k_b}}\right), \tag{4.16}$$

where

$$\Gamma = \frac{\hat{T}}{2} + \frac{p + \gamma k_d}{2(1 + k_d/k_b)} \tag{4.17}$$

It can be shown that the other analytical solution of Equation (4.15), which bears a plus sign before the square root, has no valid physical meaning and merits no further discussion.

So α, the fraction of probes bound to specific targets, is described by three global parameters: p, k_b and γ, one probe-specific parameter k_d and one variable $\hat{T}$. k_d can be expressed as:

$$k_d = e^{\frac{\xi \Delta G_d}{RT}}, \tag{4.18}$$

where R is the molar gas constant, T absolute temperature, ΔG_d the energy computed from NN model, ξ as a scaling factor to account for binding to immobilized probes.

The physical meaning of our model is clear. Both specific binding α and cross-hybridization β compete for the same probe sites. As a result, high affinity probes (small k_d) can achieve a higher fraction of specific binding, while low affinity probes (large k_d) saturate at a lower fraction. γ serves as a cross-hybridization factor. Our model shares the inspiration from two previous competitive kinetic models [77, 122], but differs in assumptions that are important to real experimental settings.

Experimentally, signal intensity is what is observed after washing, where most of cross-hybridized targets have been washed off:

$$SI = A \cdot \alpha \cdot p + \tau + \iota, \tag{4.19}$$

where SI is the observed signal intensity, τ the residual intensity from cross-hybridized targets, ι scanner bias, A the detection coefficient of fluorescence.

4.5 The $\log SI$-ΔG_d correlation is explained by the model

First of all, we shall demonstrate that our model is capable of explaining the $\log SI$-ΔG_d correlation at high target concentration in Figure 4.2.

Equation (4.15) can rearranged to a logarithmic form:

$$log\frac{\alpha}{1-\alpha} = -logk_d - log(1/k_b + \gamma/(\hat{T} - \alpha \cdot p)) \tag{4.20}$$

Note that the second item on the right side still contains the probe-dependent variable α. However, at high target concentration, the bound targets are minor comparing to free targets. This means, $\hat{T} \gg \alpha \cdot p$, and $\hat{T} \approx \hat{T} - \alpha \cdot p$. Hence, Equation (4.20) at high target concentration is approximated as:

$$log\frac{\alpha}{1-\alpha} = -logk_d - log(1/k_b + \gamma/\hat{T}) \tag{4.21}$$

For $0 < \alpha < 1$, a linear approximation can be drawn between $log\frac{\alpha}{1-\alpha}$ and $log\alpha$ over an extended range. As shown in Figure 4.3,

$$log\frac{\alpha}{1-\alpha} = 1.248 \cdot log\alpha + 0.702 \tag{4.22}$$

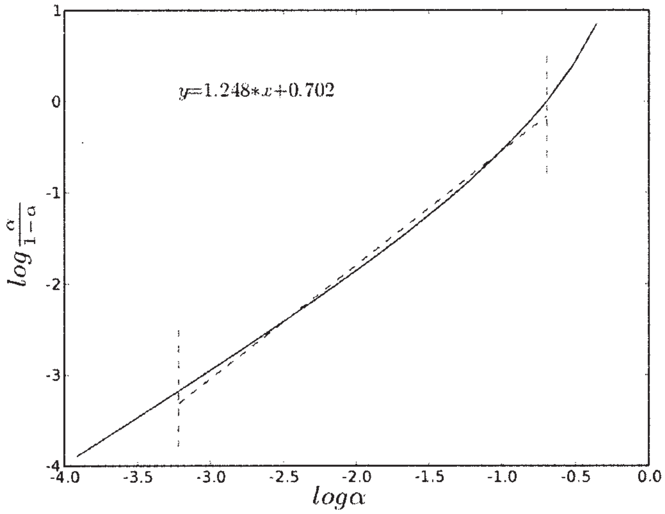


Figure 4.3: Linear relationship between $log\frac{\alpha}{1-\alpha}$ and $log\alpha$.

Combining Equations (4.21), (4.18) and (4.22), we get

$$log\alpha = -0.801 \cdot \frac{\xi \Delta G_d}{RT} - 0.801 \cdot log(1/k_b + \gamma/\hat{T}) - 0.563 \tag{4.23}$$

At high target concentration, both cross hybridization and scanner bias can be neglected. Therefore Equation (4.19) can be simplified to $SI = A \cdot \alpha \cdot p$. We substitute the α in Equation (4.23) with $SI/(A \cdot p)$:

$$logSI = -\frac{0.801\xi}{RT}\Delta G_d + C, \tag{4.24}$$

where $C = log(A \cdot p) - 0.801 \cdot log(1/k_b + \gamma/\hat{T}) - 0.563$, a constant for fixed $\hat{T}$. Thus, $\log SI$ is inversely correlated to ΔG_d. The observed $\log SI$-ΔG_d correlation is explained by our competitive hybridization model. At low $\hat{T}$, the premise $\hat{T} \approx \hat{T} - \alpha \cdot p$ is less valid; as a result, $\log SI$ is less correlated to ΔG_d. A similar effect may be created by a very low ΔG_d, where a large fraction of targets is bound to probes and taken out of solution.

A bonus here is the determination of ξ. Since the coefficient for ΔG_d in Equation (4.24) should equate the slope in Figure 4.2, we get $\xi = 0.140$.

Chapter 5

A competitive hybridization model: part II

5.1 Procedure of fitting model to Latin Square data

In DNA microarray experiments, signal intensities are measured in place of fluorescent densities of bound targets. However, common photomultiplier tube scanners usually carry a significant non-linearity for low signal intensities [123]. This means, the lower end of the Affymetrix data used in this study may deviate from the true fluorescent densities, a problem difficult to correct without knowledge of the specific instrument calibration data. Moreover, the signals from targets below 1 pM are hardly distinguishable from backgrounds, therefore, data from spike-in concentration 1 pM and above are used for our modeling.

From Equation (4.19), we define the observed value

$$\tilde{S} = SI - \tau - \iota \tag{5.1}$$

Here, the background levels τ are observed values in these Affymetrix data (signal intensities at zero spike-in concentration). Scanner bias ι is a relatively small number that has no significant effect on our model parameters. However, it is useful for stabilizing the small numbers in the fitting process. A coarse estimation of ι can be obtained by extrapolating the signal intensities at low target concentrations. We will use $\iota = -20$ in this study.

With the theoretical value

$$\widehat{S} = A \cdot \alpha \cdot p, \tag{5.2}$$

Equation (4.16) can be written as

$$\widehat{S} = A \cdot \left(\Gamma - \sqrt{\Gamma^2 - \frac{p\hat{T}}{1 + k_d/k_b}} \right), \tag{5.3}$$

where Γ is defined in Equation (4.17). In this equation, $\hat{T}$ is known, $\bar{S}$ the observed value for $\widehat{S}$, and k_d can be calculated from Equation (4.18). So we only need to fit four global parameters: A, p, k_b and γ.

We use a fitness function of weighted squares (similar to [107]). For a probe i, the fitting error is calculated as

$$E_i = \sum_t \frac{(\widehat{S}_{i,t} - \bar{S}_{i,t})^2}{\bar{S}_{i,t}}, \tag{5.4}$$

where $\bar{S}$ is observed signal intensity, $\widehat{S}$ the calculated value by Equation (5.3), t one of the nominal target concentrations $\hat{T}$ (1 - 512 pM). Our model in Equation (5.3) is fitted to the training data by minimizing the sum of E_i.

Both computationally and conceptionally, it is easier to use a saturation level in the unit of signal intensities, $P_0 = A \cdot p$. The overall fitting results are not very sensitive to P_0 beyond a certain value (Fig. 3), as other parameters can adapt to P_0. So we choose $P_0 = 30000$ here.

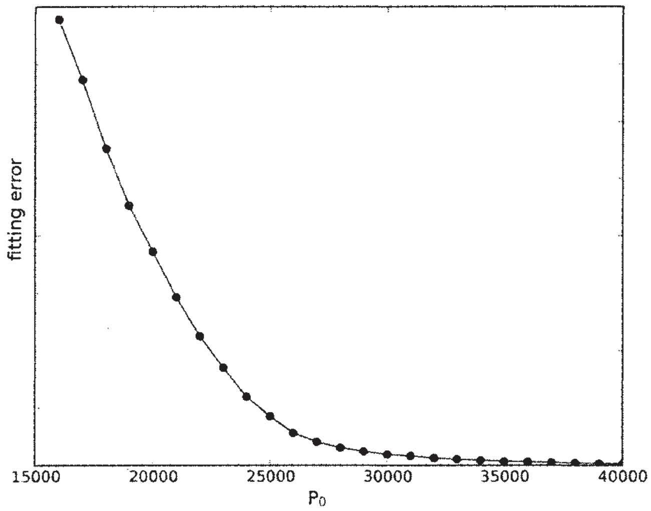


Figure 5.1: Model fitting with varying P_0.

5.2 Probe signal intensities can be successfully modeled

Figure 4.2 shows that ΔG_d (hence k_d) can be reasonably approximated by the current NN model for the probes free of secondary structures. We use half of these probes to fit our competitive hybridization model, and determine the four global parameters, A, p, k_b and γ. The evaluation is then performed on the rest of probes.

Figure 5.2 shows the fitting on two individual probes, the left one of low affinity (high ΔG_d) and the right one of high affinity. It reveals that our model simultaneously captures two important properties: low affinity probes have lower signal intensities and they saturate first as target concentration increases. Since all other parameters are the same for two probes, the differences are solely accounted for by ΔG_d.

The overall results indicate that our model captures the probe properties well. Figure 5.3A shows the modeling of individual probe signals on both the train-

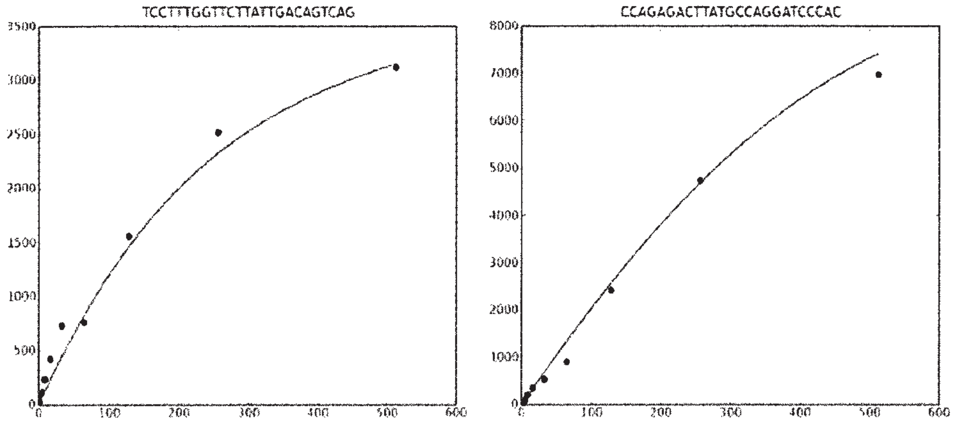


Figure 5.2: Model behavior on two example probes. Probe sequences are shown on top of charts. The blue dots are original data, red curves the prediction from our model; x-axis the target concentrations (pM), y-axis the signal intensities. Please note the different signal levels (3000 vs 7000) and curvatures.

ing data and evaluation data. Overall, the prediction on training data has $R^2 = 0.866$ (Figure 5.3B), and $R^2 = 0.880$ on evaluation data (Figure 5.3C). If we relax the probe selection criterion to $\Delta G_o > -2.5$, about 75% of total probes are included, with prediction $R^2 = 0.844$ (Figure 5.3D). The rest 25% of probes, which are presumably under stronger influence of secondary structures, can still be modeled with the same parameters but less accuracy at $R^2 = 0.735$.

In the previous, heavily parameterized models, the best prediction on $log\widehat{S}$ was correlation coefficient $r = 0.85$ in [108] and $r > 0.9$ in [70]. In comparison, our model of four parameters produces $r = 0.889$ for all probes, and $r = 0.919$ for 75% probes after a preliminary selection by secondary structures (i.e. Figure 5.3D). In conclusion, our competitive hybridization model can not only predict probe signals successfully, but also opens up paths to future improvements.

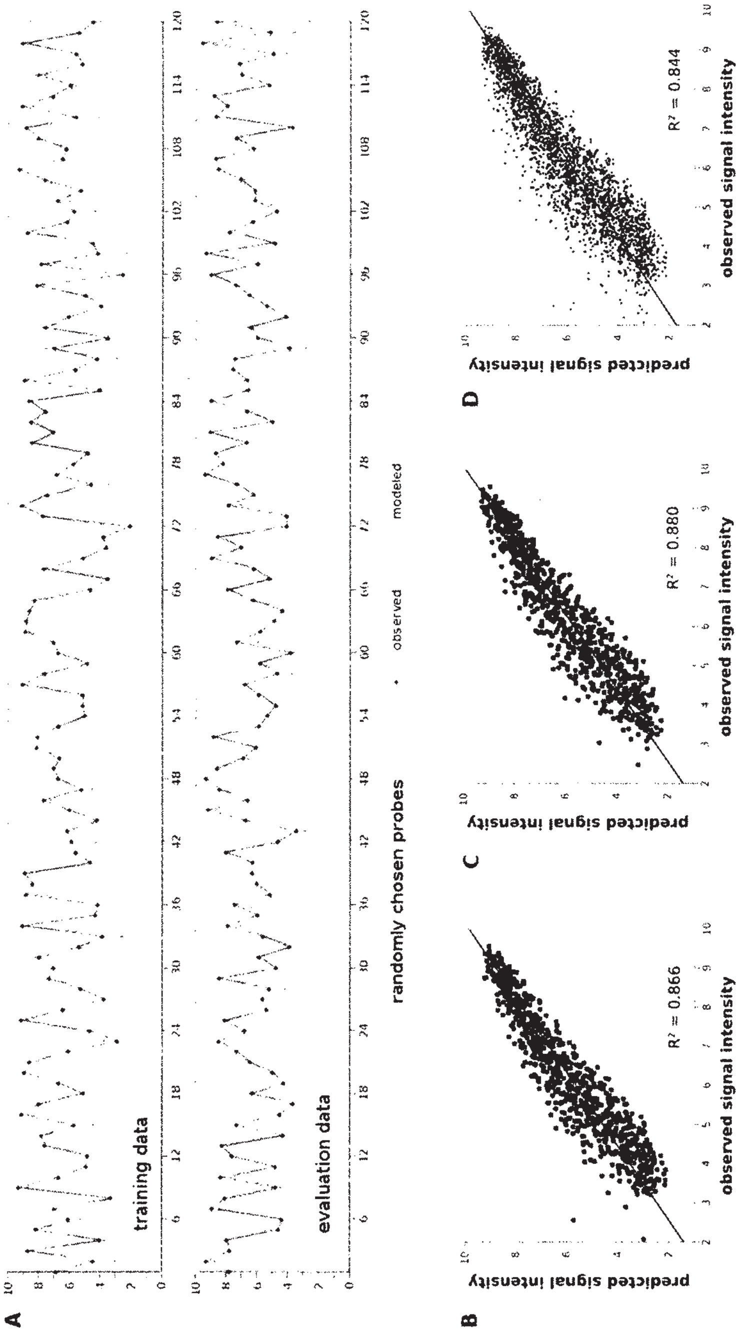


Figure 5.3: Our model can successfully predict probe signal intensities. (A) The prediction on randomly chosen probes at random target concentrations. Top: the training data; Bottom: the evaluation data. (B) Scatter plot of all training data. (C) Scatter plot of the evaluation data. The training data consist of half of the probes from Figure 4.2, and evaluation data from the other half. (D) Extended evaluation on 266 (75% of total) probes that satisfy $\Delta G_o > -2.5$. All signal intensities are in log scale. The parameters in this figure are $A = 33.408\ (pM)^{-1}$, $p = 898\ pM$, $k_b = 1.348\text{E-3}\ s^{-1}$ and $\gamma = 245500\ pM \cdot s$.

5.3 Prediction of target concentrations

With the four global parameters, target concentration can be calculated from Equation (4.15):

$$\hat{T} = \alpha p + \frac{k_d \gamma}{1/\alpha - k_d/k_b - 1} \tag{5.5}$$

If we substitute $\alpha = \bar{S}/Ap$,

$$\hat{T} = \frac{\bar{S}}{A} + \frac{k_d \gamma}{Ap/\bar{S} - k_d/k_b - 1} \tag{5.6}$$

Since k_d calculation is more accurate for probes free of secondary structures, we focus on nineteen out of the thirty probesets (transcripts) in this study that have five or more probes with $\Delta G_o > -1$. For these transcripts, Equation (5.6) is applied to calculate a target concentration from each probe. And the final concentration of a transcript is taken as the median of the data from its probes (Figure 5.4A). Figure 5.4B shows the prediction at gene level for all nineteen transcripts. In fact, comparable results can be obtained by using the few probes with $\Delta G_o > -1$ alone. At low concentrations, the predicted values in Figure 5.4B tend to be higher than the nominal concentrations. We think this is likely to be a reflection of scanner non-linearity in the low signal range, which can be corrected by an instrument calibration.

5.4 Discussion and Conclusion

In DNA microarray experiments, systematic variations stem from sample preparation and instrument operation. They are likely reflected in the global parameters of our model, A, k_b and γ. Therefore, batch variations can be expected in these parameters. Since our model has only 4 global parameters, they can be easily calibrated if control probes are built into array design. For instance, a set of

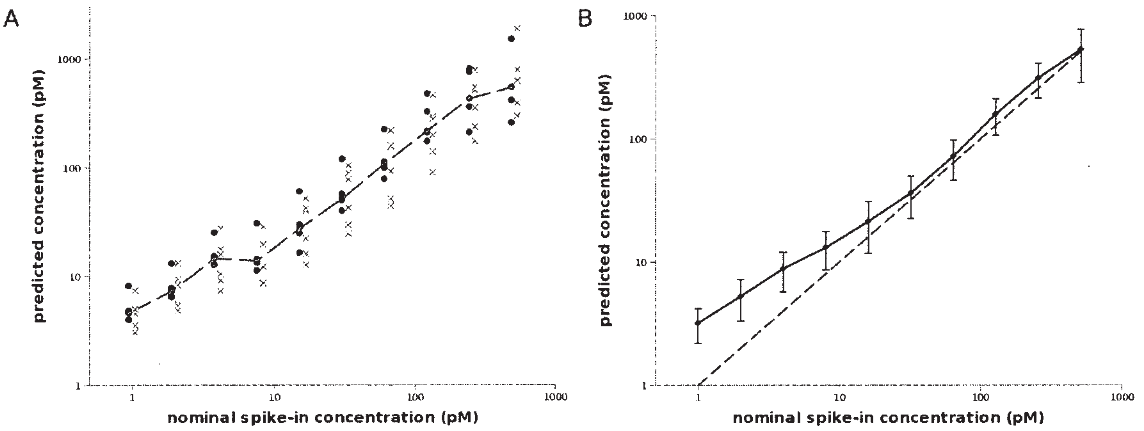


Figure 5.4: Prediction of transcript concentration. (A) Example of the 11 probes for transcript 205267_at. Dots are probes with $\Delta G_o > -1$, other probes in crosses (slightly shifted horizontally for clarity). The transcript concentration (dashed line) is taken as the median value of all probes. (B) Prediction of 19 transcripts that have five or more probes with $\Delta G_o > -1$. Error bars are standard deviations of the 19 transcripts. The predicted values bend away from the ideal line (dashed) at low concentrations probably because of scanner non-linearity.

targets complementary to the control probes can be spiked into the hybridization at various concentrations. Signal intensities of the control probes along with the known target concentrations can then be used to calibrate our model every time a hybridization experiment is performed.

We would like to emphasize that k_d is the only probe specific factor in our model, and therefore plays a pivotal role in model accuracy. The accuracy of k_d or ΔG_d in this paper is limited by the NN model, which is only a coarse approximation and affected by probe/target secondary structures. The structural effects will be discussed further in Chapter 6.

We assumed a constant cross-hybridization factor γ for all probes, which may not be the case. Further research on γ may improve the accuracy of our model. We did not deal with the background levels in this study, which are not important to signals at high target concentration but will affect signals at low concentrations. Background levels have a clear dependency on ΔG_d, and are well addressed in other studies [124, 125].

We have presented the first model of DNA microarray hybridization that explains probe signal intensities through sequence-based thermodynamic properties without excessive parameter fitting. This fills in the long standing knowledge gap in DNA microarray hybridization. Our model provides a mechanism of absolute quantification, and shall improve the quality control and reproducibility of the technology. With only four global parameters, this model can be easily calibrated through control features that are built into microarrays, and adopted in practice.

5.5 Methods

The Latin Square spike-in data U133A were retrieved from [62]. Probe information was obtained through [126], where only 30 of the 42 probesets were found. 365 probes matched to target sequences. Among them, 10 probes with very low signal intensities (under 900 at highest target concentration) were removed. In total, 355 probes are included in this study. Background was taken as the signal intensity at zero spike-in concentration, and subtracted from data at other concentrations. No normalization was performed on these data. The probe self-folding energy, ΔG_o, was computed by RNAStructure (version 4.5, function OligoWalk [127]). Duplexing energy, ΔG_d, was computed by the current Nearest Neighbor model with the parameters from [95].

Chapter 6

The structural effects in microarray hybridization

Summary

The last chapter showed that our kinetic model was successfully applied to microarray hybridization, and accurate computation of duplexing energy, ΔG_d, remained a key factor. We computed ΔG_d via the current Nearest Neighbor model and existing parameters. One clear limitation here is that the NN model does not take into account secondary structures. It is therefore important to screen for secondary structures on microarrays or incorporate them into computing ΔG_d. This chapter presents preliminary explorations of this research direction. We have found that the structural effects appear to suppress probe signals. A simple ranking method, combining folding energy and partition function, outperforms the ΔG_o method used in the previous two chapters in probe selection. An interesting note is that neither end of probes on Affymetrix platform participate well in microarray hybridization.

6.1 Probe selection and the $\log SI$-ΔG_d correlation

In order to consider the structural effects in modeling microarray hybridization, one has to ask the following questions: can we predict the secondary structures? Can we incorporate them into the computation of ΔG_d?

An accurate prediction of the secondary structures of long target molecules is still difficult today. In the Affymetrix Latin Square experiments, RNA targets were also fragmented before the hybridization. The exact effects of target secondary structures have been a matter of debate [104, 112, 128–130]. However, it is easier to identify the regions that are free of secondary structures, which are determined mainly by lack of base-pairing or weak bonding to the rest of the molecule. Such sequences are likely to remain structure-free after fragmentation. Similarly, probes that are predicted structure-free in solution should remain so on microarray surface. Therefore, a provisional method is used in this chapter to rank the likelihood of secondary structures in a probe/target pair. In short, the likelihood of both probe and target to form secondary structures were assessed via two methods: folding energy and McCaskill's partition function [131]. The ranking from these four parameters were then combined into a single ranking index (see methods for details).

In the Affymetrix experiments, the labeled targets are long RNA molecules and probes are 25mer DNA oligos. Thus, we computed probe duplexing energies, ΔG_d, by implementing the NN model in Equation (4.1) with RNA/DNA parameters from Wu *et al.* [95]. When signal intensities of all U133 probes are plotted against their ΔG_d (Figure 6.1A), a general correlation appears with p-value = 2.3E-23, but a poor $R^2 = 0.29$. After applying a criterion of selecting probes of minimal structural effects, clear correlation between $\log SI$-ΔG_d can be obtained. Figure 4.2 was the first example we have seen. Similarly, the selection by raw ΔG method and SSS method (see Methods) also produced clear correlations. A better result is shown in Figure 6.1B after my combined ranking method is applied ($R^2 = 0.67$, p-value = 3.4E-12). Comparable results were also observed on the U95 data (not shown).

This $\log SI$-ΔG_d correlation can be expressed as:

$$\log SI = \phi * \Delta G_d + \varepsilon, \tag{6.1}$$

where ϕ is the slope, ε a constant. This is an alternative of Equation (4.24), and in concordance with Equation (4.5), with SI being a proxy for K. Our ranking method allows different numbers of probes to be included with varying selection stringency, Their R^2 of fitting to Equation (6.1) is plotted in Figure 6.1C. The top 67 probes fit to Equation (6.1) with $R^2 = 0.61$, 135 probes with $R^2 = 0.55$.

Therefore, consistent with Chapter 4, a subset of probes free from structural effects show a good correlation between their duplexing energy ΔG_d and signal intensity. In selecting this subset of probes, our combined ranking method outperforms the simple ΔG_o method.

6.2 The open root and loose end

The constrained geometry on DNA microarrays leads to the question whether there is a positional effect in hybridization. This can be approached by removing a base at a certain position along the 25mer probe (equivalent to removing an NN pair out of the 24 counts), and examining the effect on the $\log SI$-ΔG_d correlation in Equation (6.1). We took the top 135 probes, for which the R^2 of fitting to Equation (6.1) was 0.55. When one of the positions was omitted in the NN model, a set of ΔG_d was computed on the remaining bases and R^2 was computed for their linear regression to $\log SI$. The result is shown in Figure 6.2A, clearly indicating that neither end of the probes is important for the hybridizations, and they do not participate in the duplex formation. We next recomputed the duplexing energy excluding the first and last bases of probes as ΔG_d^{23}. The R^2 was actually increased when ΔG_d^{23} was used in fitting to Equation (6.1). Similar to Figure 6.1C, Figure 6.2B

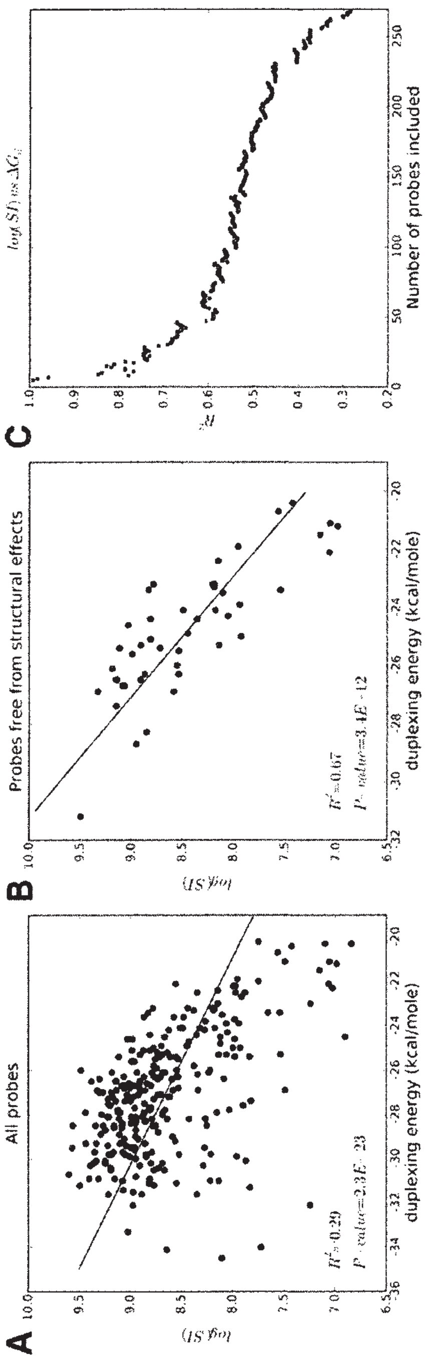


Figure 6.1: Improved selection of probes that are free from secondary structure by a combined ranking method. Similar to Figure 4.2, the $log(SI)$-ΔG_d correlation is used to measure the success of probe selection. Data are from U133 Latin Square experiment, using signal intensities at 512 pM spike-in concentration. (A) All 269 probes. (B) Probes free from structural effects (top 46 probes by our ranking method). (C) Depending on the stringency of probe selection, the R^2 of $log(SI)$-ΔG_d correlation versus different number of included probes.

shows the R^2 of varying number of probes, and the R^2 is consistently better with ΔG_d^{23} up to 168 probes, beyond which the background noise blurs the difference. In comparison to Figure 6.1B, R^2 for the top 46 probes goes up from 0.67 to 0.75 (Figure 6.2C).

Our results strongly suggest that the root of probes is not accessible in hybridizations, probably because Affymetrix synthesizes probes directly from the glass surface, leaving no room to accommodate the protruding ends of targets. Rather unexpectedly, the end of probes does not appear to hybridize well either. This raises an important issue as some microarray applications use the probe ends for primer extension or ligation. For this latter observation, we do not have a satisfactory explanation at this moment, but offer two speculations: (a) the movement of protruding tail of target in solution may introduce some disruptive energy to the target-probe duplex, causing the ΔG_i for the terminal base pairs to be greater than that used in the NN model; (b) probe synthesis may not be perfect and some probes did not really have the end bases. These positional effects agree with in previous notions [70, 102, 132]. In the rest of this paper, we use the computed ΔG_d^{23} with 23mers instead of duplexing energy from full 25mers.

6.3 The NN parameters

All the NN parameters to date have been determined in bulk solutions, which raised the question of their applicability to surface hybridization in microarrays. The set of parameters from Wu *et al.* [95] (temperature-independent) were used in this study and they worked reasonably well. Another set of RNA/DNA parameters were determined by the same group several years earlier [94]. These two sets of parameters work somewhat similarly, but the Wu parameters produce a better fit (compare Figure 6.2C with Figure 6.3A). The same advantage of ΔG_d calculated

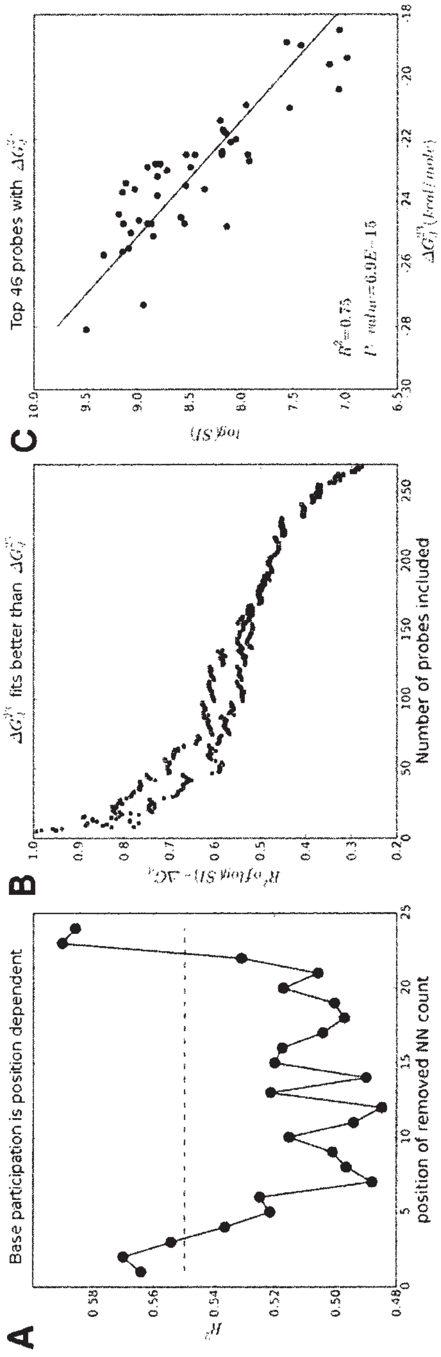


Figure 6.2: Neither end of the probes participates in duplex formation. (A) The effect on R^2 of $log(SI)$-ΔG_d correlation by removing one NN count, using top ranked 135 probes, which have original R^2 of 0.55 (the horizontal line). When bases on either end of probes are removed from ΔG_d calculation, the $log(SI)$-ΔG_d correlation gets better. (B) ΔG_d^{23} (red) computed without base 1 and 25 systematically fits better than ΔG_d^{25} (blue), referring to the same term of R^2. (C) Top ranked 46 probes, $R^2 = 0.75$, comparing to $R^2 = 0.67$ in Figure 6.1B.

with Wu parameters was observed on the U95 data (not shown).

Since DNA/DNA parameters are also used in practice, we include here the ΔG_d data calculated with common DNA/DNA parameters [93] for a comparison (Figure 6.3B). They deviate significantly from those in Figure 6.2C, suggesting that the distinction between DNA parameters and RNA parameters should be observed strictly.

If we fit the NN model in Equation (4.1) directly to these probes, a new set of 16 NN parameters can be derived on the measured signal intensities. The prediction by these parameters was evaluated in Figure 6.3C, with a $R^2 = 0.79$, offering only a slight improvement from Figure 6.2C ($R^2 = 0.75$). It is not clear whether there exists a set of "microarray NN parameters" that are really better than the Wu parameters. And these fitted parameters are similar to the two sets of experimentally determined parameters (not shown). There is room to improve the ΔG_d computation for microarray hybridization, which may require amendment of the NN model beyond simple parameter fitting.

6.4 Structural effects may suppress signal intensity

Now, let's take a close examination of the structural effects. Our probe ranking method is a combination of predicted structures and the confidence of predictions. Top ranked probes have high confidence of being free from structural effects. By going down the ranks, more probes with structural effects are included. In 6.2B, the R^2 of $\log SI$-ΔG_d correlation goes from 0.89 for the top 16 probes down to 0.46 for the top 231 probes.

We will refer the regression line for the top 16 probes as the "ideal line" (red in 6.4) for probes "free" of structural effects. It is immediately interesting to ob-

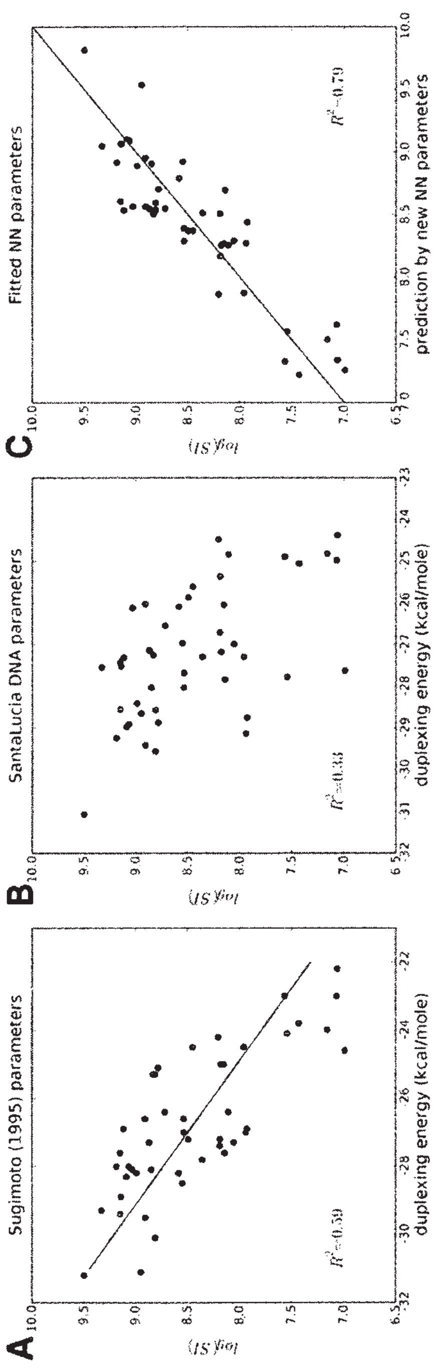


Figure 6.3: Comparison of NN parameters. All energies were based on 23mers, same probes as in Figure 6.2C. (A) Using the DNA/RNA parameters form [94]. (B) Using DNA/DNA parameters from [93]. (C) New NN parameters derived from data on these 46 probes (by fitting model 4.1 to $log(SI)$).

serve that this "ideal line" sets an upper boundary of probe signal intensities, as all intensities are around or below this line. This suggests that probe signal intensities are suppressed by structural effects. Please also note that our ranking method is tuned for selecting probes free of structural effects - not all probes of low ranks necessarily have strong structural effects.

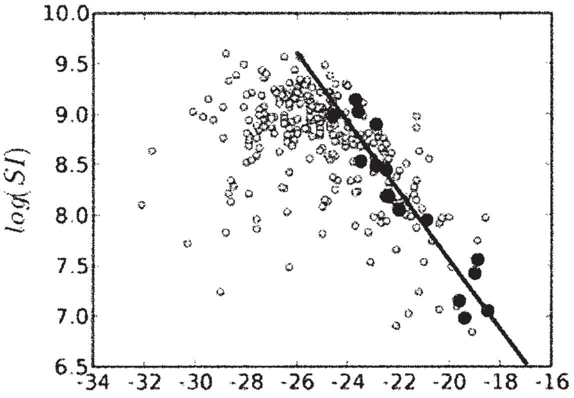


Figure 6.4: Structural effects suppress signal intensity. The top ranked 16 probes and their regression line in red, all other probes in yellow. The red line is referred as the "ideal line" for probes without structural effects, with a slope −0.34. All probes are either around or below this line. The clutter of more probes at lower ΔG_d reflects a bias in Affymetrix probe design.

6.5 Discussion

From our analysis of Affymetrix Latin Square data, the structural effects appear to suppress probe signal intensity. Such effects of secondary structures have been previously reported by surface chemistry experiments using limited number of probes [133–135]. Our result extends this observation into real microarray experiments. A recent study on the NimbleGen platform [102] suggested that the two most important factors on hybridization signal intensities, with both genomic DNA and cDNA hybridization experiments, were probe melting temperature and secondary structure. The melting temperature was also computed with the NN model and should correlate to ΔG_d. Our results consider the structural effects from both probes and targets. The NimbleGen experiments used DNA as targets, which have less secondary structures than RNA targets. Therefore, the structural effect in their case is dominated by probe structures. Thus, results from [102] align well with our conclusions.

This chapter used a prototype method for ranking the secondary structures of probes and their matching target regions. Better methods should be a goal of future research in this field. In particular, the current algorithms of computing polynucleotide secondary structures were developed for bulk solutions not microarray surface, which adds extra constraints in polymer structures. New algorithms for predicting secondary structures on microarray are desired components for developing new probe design methods. Because if one can just avoid most the secondary structures, it will improve the accuracy of ΔG_d calculation and therefore the quantitative results of our competitive kinetic model. On the other hand, the NN parameters may also need some adaptation to microarray surface. When secondary structures are present, they are expected to alter hybridization energy beyond the simple ΔG_d computed with the NN model. How to model such structural effects still presents a challenge to future research. Finally, we note that target structure and probe structure are not totally independent from each other. Enforcing a structure-free rule in microarray design may have multiple beneficiary consequences.

There have been many discussions about the background levels of microarrays. Backgrounds on the U133 data turned out to be less complex than the kinetic isotherms. There was a clear correlation between backgrounds and ΔG_d, as shown in Figure 6.5. Thus, simple background correction can be practically effective [124, 136], which is, however, beyond the scope of this study.

Chapters 3 through 6 have investigated both sequence independent and se-

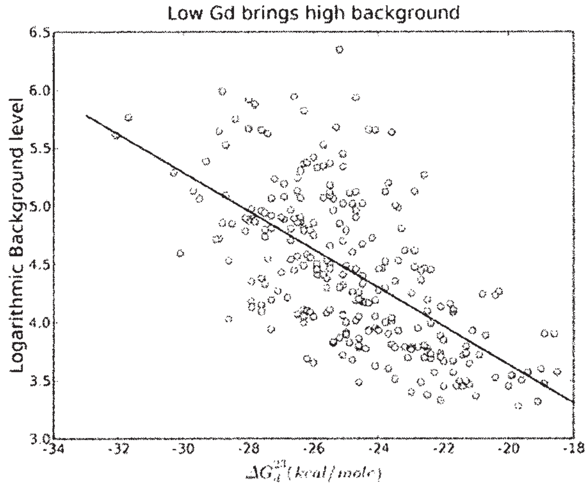


Figure 6.5: Lower ΔG_d comes with higher background. $R^2 = 0.39$, P-value = 1.2E-30. Background is taken as the signal intensities at 0 spike-in concentration.

quence dependent variations on DNA microarrays. The sequence dependent variation has been a dominating problem of DNA microarray technology, and has attracted significant modeling efforts. We have proposed a new competitive hybridization model that can successfully predict probe signals based on the empirical NN parameters. Our model, however, relies on the accuracy of computing duplexing energy between probes and targets, which still has difficulty to cope with secondary structures. Thus, an effective method of screening secondary structures on microarrays will be important to bringing our competitive hybridization model to real-world impacts. Future developments along this line should improve DNA microarray technology significantly, and reinforce its position in the face of stiff competitions. The potential of absolute target quantification is also very valuable to the reanalysis and integration of existing microarray data.

6.6 Methods

Data retrieval and processing

Affymetrix Latin square data were obtained as described in last chapter. The U95 data set was from an earlier experiment than the U133 data. Only probes with complete information were considered. For the U95 data, only 59 hybridizations with complex background were included in the data distribution, and only partial target sequences were available in the Affymetrix NetAffx Analysis Center. After removing the reportedly defective probes, 148 probes were included for the U95 data. Since the target structure prediction was compromised due to incomplete sequences in U95, and U95 data have a small number of probes, most of our discussions were led by the U133 data.

Energy and structure computation

The partition function developed by McCaskill [131] computes the base-pairing probabilities for each base in the studied sequence, and it often serves as a basis for RNA structure prediction programs. We summarized these base-pairing probabilities into an accessibility measure, the Single Strand Score (SSS), based on counting the projected single-stranded bases. A higher SSS indicates less secondary structure, hence better accessibility in the hybridization. An alternative, the popular energy minimization approach, was also used, where the local disruptive energy ΔG reflects the secondary structure [127]. A higher ΔG (less negative) also suggests better accessibility.

These two methods were applied to both probes and their targets. The computation for probes was based on their full length (25 bases). For the targets, the structure prediction was based on the maximal length of transcripts, while the accessibility parameters were calculated for the 25-base local region complementary to specific probes. We denote the disruptive energy for target structure as ΔG_t, and ΔG_o for probes. The obtained ΔG_t ranged from -50 kcal/mole to 0; ΔG_o was from -8 kcal/mole to 0, mostly in the upper range as probe self-folding was already considered a factor in probe design.

We are aware that computational prediction of secondary structures has limited accuracies. Moreover, the conditions in microarray hybridization differ significantly from those used in the prediction. Therefore, the SSS scores and energies should be interpreted solely as relative measures, not to be taken as absolute values. However, sequences free of secondary structures under the theoretical conditions should remain so as the hybridization conditions and fragmentation were unlikely to add new structures.

To have a more robust, single criterion of probe selection, the four parameters per probe, ΔG_t, SSS_{target}, ΔG_o and SSS_{probe}, were then combined into a simple ranking method. It is important to note that our goal is to enrich a subset of probes free from structural effects with minimal "false positives". Our method is very conservative and does not intend to include all probes that are factually free of structural effects. The "structural effects" herein, unless stated otherwise, will refer to the effects from both probe and target folding, though probes are the major unit of discussion.

For computing the base-pairing probabilities, we used the implementation of partition function in the Vienna RNA package (version 1.6.5, function RNAfold [137]). Then the Single Strand Score (SSS) is calculated as

$$SSS = \sum^{i} C_i, \tag{6.2}$$

where $C_i = 1$ for a low base-pairing probability and $C_i = -1$ for a high base-pairing probability; i =*1*, ..., N; as N is the length of studied region. *E.g.*, to compute

SSS_{target} for a probe with sequence "GAAGGCATGAAATTGTCTAGCAGAG", we first obtain the result of partition function for the full-length target transcript (in this case 203508_at). The matching region in target ("CUCUGCUAGACAAUU-UCAUGCCUUC") has an pseudo-structure summarized as

```
(((.(((((((...,,{...(((((
```

, where "." or "," and "{" denote no or weak pairing, "|", "(" and ")" strong pairing. The no or weak pairing base is counted as +1 for SSS, and the strong pairing base is counted as −1. Hence the SSS_{target} for this example is −6. The ΔG method is more sophisticated and should work better for short probes, while the SSS method is more robust and should work better for long targets.

For the free energy minimization approach, we used RNAStructure (version 4.5, function OligoWalk [127]), an extension of the Mfold algorithm, to compute the free energies ΔG_o and ΔG_t for each probe and its corresponding target region. So in total, we get four parameters for each probe: SSS_{probe}, ΔG_o, SSS_{target}, ΔG_t. Each parameter for each probe was then transformed into a ranking number between 1 to 10 by its rank in all probes. Combined, $Rank_{probe} = probe\_rank_{sss} + probe\_rank_{\Delta G_o}$, $Rank_{target} = target\_rank_{sss} + target\_rank_{\Delta G_t}$, and the overall ranking number is computed as (product is used for ranking confidence):

$$RN = Rank_{probe} \cdot Rank_{target} \tag{6.3}$$

The fitting of new NN parameters was carried out by singular value decomposition, using implementation in GNU Octave (`http://www.octave.org`). Other processing was done by Python (`http://www.python.org`) scripting and spreadsheet program Gnumeric (`http://www.gnome.org/projects/gnumeric/`). Some graphic plots were produced with Matplotlib library (`http://matplotlib.sourceforge.net/`).

Part II

Fish metabolic network

Chapter 7

Conceptualization of a fish metabolic network model

Chapters 2 through 6 discussed the multiple aspects of DNA microarray technology. Once one has obtained gene expression data and performed statistical analysis, the next challenge is how to interpret the data in the context of functional genomics. This last step is not only pertinent to microarrays, but common to most high throughput technologies. Since most data of functional genomics comes from other model species, significant bioinformatic development is needed to bridge the gap for fish toxicogenomics. Chapters 7 to 10 will discuss the construction of a fish metabolic network model and supporting utilities. The project is code-named MetaFishNet.

7.1 Why a fish metabolic network model

Models of molecular networks are valuable tools for understanding high throughput data, and the idea of visualizing expression data superimposed on network models has been implemented in a number of software packages [138–141]. Obviously, the quality of models relies on the source data, mostly experimentally determined molecular interactions. For example, chromatin immunoprecipitation in combination of DNA arrays (known as ChIP-chip) or high throughput sequencing (known as ChIP-seq) can identify protein-DNA interactions at genome scales [142, 143]. Large-scale protein-protein interaction data, based on techniques from yeast two-hybrid or protein arrays, are already available for yeast, fruitfly, C. elegans and humans [144–149]. However, these kinds of interaction data are still scarce for fish species. Network analysis tools for fish will have to rely on careful data mapping and integration, mostly from other species.

As much as we would like to have a comprehensive interactome model for fish, we are limited by both the supporting data and the practical resources. For this dissertation, I propose to construct a fish metabolic network model, MetaFishNet. Because metabolic data have been accumulated piecemeal over many decades, they are arguably more reliable than those from the yet maturing high throughput technologies. The value of metabolic models for interpreting expression data has been demonstrated [24, 28, 150].

There in fact exists a metabolic network model for zebrafish in the KEGG (Kyoto Encyclopedia of Genes and Genomes, [151]) database. However, the recent completion of genome sequencing of five fish species has paved the way for constructing a genome-wide fish metabolic network model. That is, all metabolic enzymes can be identified from complete genomes by sequence analysis; compounds can then be associated to enzymatic activities; a metabolic network can be constructed by linking these compounds. This type of *ab initio* construction of metabolic networks has been carried out for several bacteria and yeast [152–156], but not for any vertebrates. Given the higher complexity of vertebrates, an *ab initio* construction alone may produce too many errors. Good reference data are still essential. Fortunately, two high-quality human metabolic network models [24, 25] have been published recently. These high-quality models also demonstrated that the KEGG models could be improved significantly. Since both models contained many data from model species other than human, these two "human" models are better viewed as references for all vertebrates. With these latest developments, it is possible to construct a reasonable genome-wide fish metabolic network model.

Even with a network model in hand, substantial software development is still required to use the model, map sequences of different species, visualize data and connect with expression data. Thus, supporting utility programs are also an important part of this project.

Besides the interpretation of expression data, MetaFishNet will give a systematic review of our current understanding of fish metabolism, and address the question how the animal models resemble human subjects. Probably more importantly, it is a framework tool that will enable future studies of fish molecular networks by facilitating hypothesis generation and study designs.

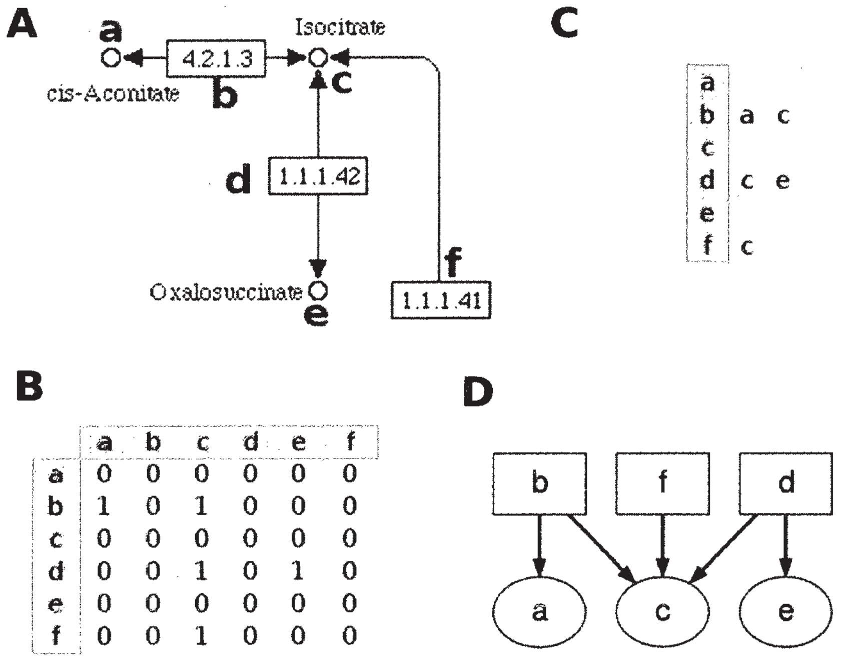
7.2 Networks and metabolic networks

This section gives an overview of the basic concepts of networks and metabolic networks. Networks are now frequently used to describe collective data of molecular interactions, including signaling, developmental and metabolic processes. The mathematical description of a network is a *graph* (not to be confused with the English word for visual presentation), defined as a set of nodes (also called vertices) and edges that connect the nodes. For example, in a protein interaction network, the proteins are modeled as nodes and their interactions are modeled as edges.

Both nodes and edges can have subtypes to accommodate various properties. In metabolic networks, there are two common types of nodes: enzymes and compounds; and edges represent their interactions. Thus, metabolic networks are usually modeled as a class of "bipartite network", where an enzyme (or enzyme complex) is connected to another enzyme only via a common compound. Depending on the context of individual reactions, this compound may be the product of one reaction and the reactant of another reaction. The connecting edges also bear directionality from such reactions. An edge can go both ways if the reaction is reversible

(Figure 7.1A, D).

In computer programming, a network can be represented either as an adjacency matrix or an adjacency list (Figure 7.1B, C). Both representations were used in the programs for MetaFishNet, depending on the situation. The number of edges attached to a node is called the "degree" of the node. Distribution of the degrees is often used to study the internal structure of networks. For example, the degrees in a random network follows Poisson distribution. When the degree distribution in a network follows a power law, this network is a "scale-free" network [157]. Most biological networks, including metabolic networks, have been shown to be scale-free [26, 158, 159].



	a	b	c	d	e	f
a	0	0	0	0	0	0
b	1	0	1	0	0	0
c	0	0	0	0	0	0
d	0	0	1	0	1	0
e	0	0	0	0	0	0
f	0	0	1	0	0	0

a		
b	a	c
c		
d	c	e
e		
f	c	

Figure 7.1: Representations of a network. (A) part of TCA cycle, as an example of metabolic network. (B) Adjacency matrix (C) Adjacency list (D) A different rendering of the same graph.

Modularity is a measure for community structures in a network. An influential mathematical definition of modularity was given by Newman and Girvan [160].

Ma et al. demonstrated that metabolic pathways can be detected by finding modularity in the larger networks [159]. Pathway organization was an important issue during the construction of our MetaFishNet model. A modularity finding algorithm by Newman [161] was adopted for this task, which will be discussed further in a later chapter.

A small number of molecules, for instance, water, oxygen and ATP, are present ubiquitously in metabolic reactions. They are called currency metabolites. They are usually removed from network analysis and visualization, which would be otherwise cluttered. What should be included in this list of currency metabolites is a matter of minor debates [153, 162–164], and somewhat dependent on the exact pathway context. They usually correspond well to the metabolites of highest connection degrees. We will see this is also the case in MetaFishNet (Table 7.1). The list by Holme and Huss [165], which is very close to that of Ma and Zeng [153], fits slightly better to our data and was adopted in this study. In previous metabolic models, currency metabolites may or may not be included in a reaction description. Thus, I excluded currency metabolites from reaction comparisons and network modularity analysis.

7.3 Construction strategy of MetaFishNet

The construction strategy of MetaFishNet is shown in Figure 7.2. It combines the integration of existing models and *ab initio* construction from whole genomes. We expect MetaFishNet to expand significantly from the KEGG zebrafish model, and this will be confirmed in a later chapter. In a nutshell, we will first merge the two high quality human models, then join them with the zebrafish model from KEGG. The combined result of these three models serves as a baseline reference. All metabolic genes/enzymes are then identified from the five complete fish genomes.

Table 7.1: The hub (most connected) compounds in human metabolic network correspond to the currency metabolites designated in previous studies. A new human metabolic network has been generated by merging the latest EHMN and UCSD models (See Chapter 9 for details), and the compounds of highest degrees are listed in the left column. I have adopted the list from Holme & Huss as currency metabolites for this study. Currency metabolites are often treated separately in studying network structures.

hub compounds in new data	Holme & Huss	Ma & Zeng
ATP	ATP	ATP
ADP	ADP	ADP
$NADPH$	$NADPH$	$NADPH$
$NADP^+$	$NADP^+$	$NADP^+$
NAD^+	NAD^+	NAD^+
$NADH$	$NADH$	$NADH$
P_i	P_i	P_i
PP_i	PP_i	PP_i
CO_2	CO_2	CO_2
H_2O	H_2O	H_2O
O_2	O_2	O_2
H^+	H^+	
NH_3		NH_3
		SO_4
H_2O_2		
CoA		
$Acetyl\text{-}CoA$		
UDP		
AMP		

For the enzymes that can be mapped to the baseline reference data, the reference reactions will be adopted. For the rest of enzymes, an *ab initio* construction will be performed. In the end, all data are reconciled into the first version of MetaFishNet, and supporting databases and computational utilities will be developed. An immediate application of MetaFishNet will be the interpretation of high throughput expression data. Certain levels of errors are expected in both genic annotations and sequence analysis. As a generic fish model combining five fish genomes, MetaFishNet should alleviate these problems.

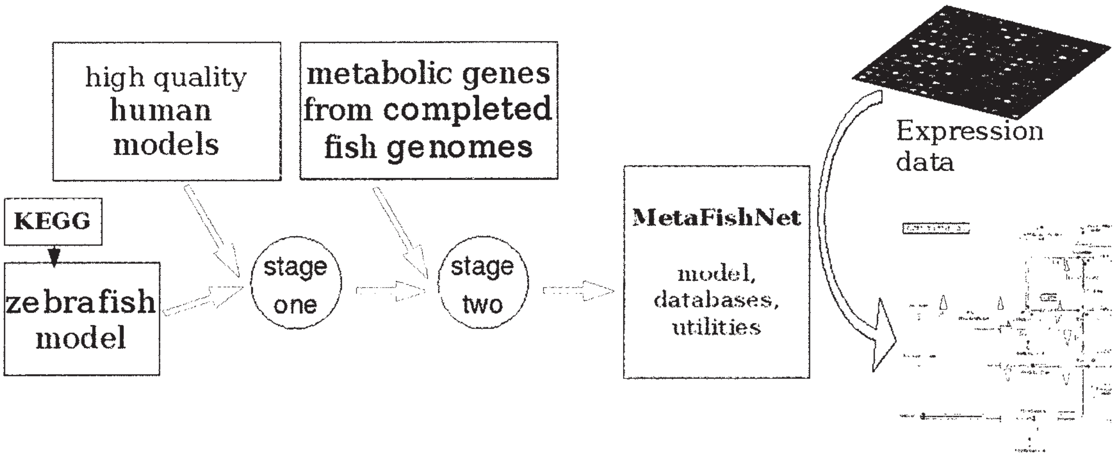

high quality human models
metabolic genes from completed fish genomes
KEGG
zebrafish model
stage one
stage two
MetaFishNet
model, databases, utilities
Expression data

Figure 7.2: Construction schematic of MetaFishNet.

In most cases, metabolic reactions from other species can only be mapped to a fish species through orthologous associations. This determines that the initial fish metabolic model has to center on genes/enzymes. In fact, even for the two high quality human models, the merging relies heavily on gene/enzyme identities, because metabolites have loose nomenclatures and are difficult to compare. Information on metabolites will be carried over from the baseline references, but needs to be verified and updated in the future.

A common method of ortholog identification is via best reciprocal similarity, i.e., two genes from different species have to appear as the top hits in a two-way homology search to be called as an ortholog pair. As most fish genomes underwent an extra duplication compared to the human genome, this one-to-one relationship from best reciprocal similarity has clear limitations. A better approach is to build ortholog/paralog relationships based on the phylogenetic tree of the gene family. This is computational intensive, but has been thoroughly implemented in Ensembl project [166]. Therefore, we will adopt the ortholog identifications from Ensembl.

Identification of metabolic genes is accomplished by Gene Ontology (GO) computation [55]. Zebrafish has a good GO annotation. But other fish species still need to be annotated in a GO conscious way. For this purpose, a new annotation tool, *SeaSpider*, will be created. Besides being used for the initial construction of MetaFishNet, *SeaSpider* is also used on the application side of MetaFishNet. When genes/clones from an arbitrary species are queried upon MetaFishNet, the only reliable identifier is often their sequences. In such cases, *SeaSpider* will be called on to map these sequences to MetaFishNet records.

MetaFishNet deals with a large amount of heterogeneous data. It is critical to write computational scripts quickly for data integration. There are also multiple components from sequence analysis, database utilities, visualization and web interface. I will use the programming language Python as a "glue" to streamline the project.

Chapter 8

SeaSpider, a GO centric annotation tool

The project MetaFishNet needs a sequence analysis tool for both the construction and applications. During the construction of MetaFishNet, the genes from five fish genomes were analyzed for Gene Ontologies (GO), then the metabolic genes were identified by their GO categories. For the applications of MetaFishNet, sequence comparison is often the only way to identify the genes submitted by users. Therefore *SeaSpider*, a GO centric annotation tool, was created for this study.

8.1 The program

Several GO annotation tools [167–172] already exist. Why do we need a new program here? The first reason is availability and usability. Web services from these tools are not always available and they can not be easily incorporated into our automated pipeline. The second reason is a different working strategy. All these GO annotation tools associate the sequences under investigation to existing annotation through searching for sequence homology. In our case, we prefer to utilize the zebrafish annotation as a primary source because gene functions do vary among species and zebrafish is as close as we can get. Only when annotation can not be found in zebrafish, we turn to generic annotations. The third reason is this customization can keep its annotation sources up-to-date by following the updates in Ensembl [166] and GO consortium.

The design of *SeaSpider* is illustrated in Figure 8.1. The sequence search program BLAST [173] from NCBI is used as a component. BLAST is not the fastest

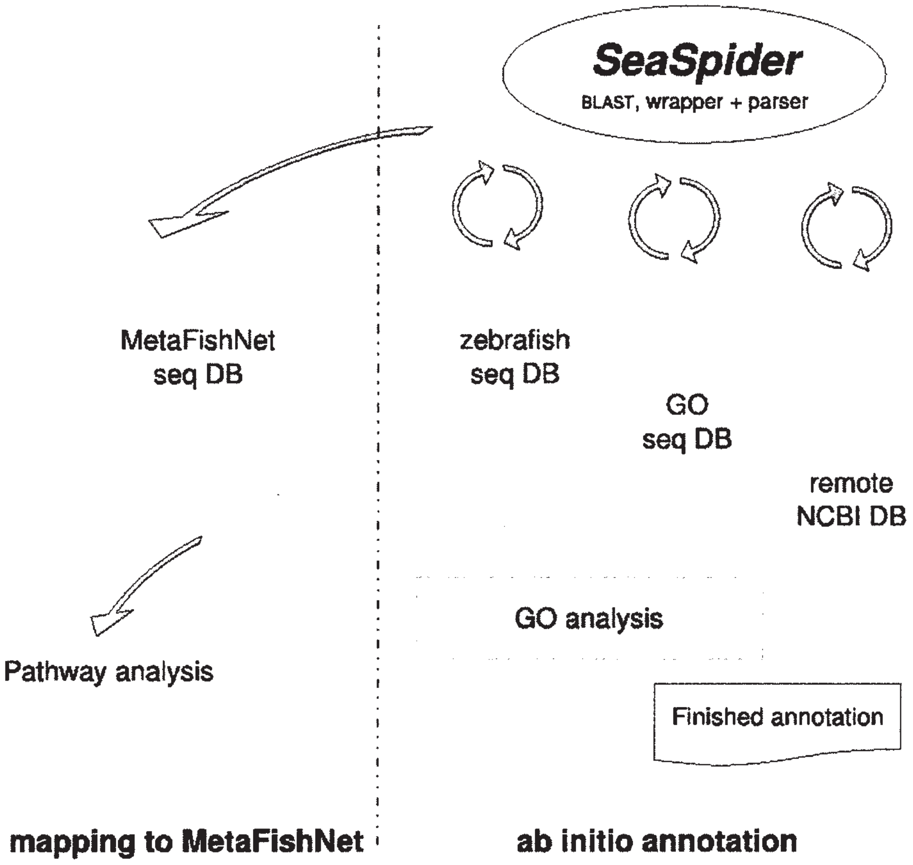


Figure 8.1: *SeaSpider* is used for both ab initio annotation and the mapping to MetaFishNet.

algorithm nowadays, but it is mature and a frequently adopted standard. *SeaSpider* wraps the input and output of BLAST, parsing the XML results. In reflection of its dual roles, *SeaSpider* can do *ab initio* annotation, which associates genes to their GO terms wherever possible, as well as a tool for mapping users' genes onto MetaFishNet model. Different databases are used for these two functions. For *ab initio* annotations, new sequences are searched against zebrafish sequence database first, then the generic GO sequence database. For sequences that do not have matches in these local databases, *SeaSpider* queries them further to NCBI remotely. The last step does not introduce GO information, but makes *SeaSpider* a competent standalone application for annotating new gene sequences. For the mapping to MetaFishNet, new sequences are searched against all the metabolic genes used in MetaFishNet, then taken to the next step of pathway analysis.

SeaSpider is organized as a Python package. It can be run directly from a command line Shell, or imported into other Python applications. The latter is how it is integrated with MetaFishNet applications, including access via a web interface.

8.2 Gene Ontology

Gene Ontology Consortium is a significant effort to unify biological vocabularies into a computable format. The whole set of Gene Ontology is modeled as a directed acyclic graph. When a GO term is assigned to a gene, the gene is automatically associated with all its upstream terms. They can come from all of the three major categories: biological process, molecular function and cellular component. It is common that a single gene is associated with dozens of GO terms. The relationships among these GO terms have to be tracked through the database provided by GO Consortium. Since intense database queries are involved and the size of the complete GO database is manageable (about 400 MB), we keep and use a local copy of GO database.

Zebrafish has good GO annotations, which came mostly from ZFIN (ZebraFish Information Network, [174, 175]) project. The gene sequences from genomes of medaka, Takifugu, Tetraodon and stickleback were annotated by *SeaSpider.* A gene is considered "metabolic" when it is associated with the GO term "metabolic process" and a next step will be taken to find its appropriate Enzyme Commission (EC) number.

For general annotation purpose, the thousands of GO terms can be overwhelmingly complex. GO consortium provides a trimmed version, GO slim, which is a short list of 131 most descriptive terms. It is a common practice to categorize the GO terms on a set of data, often with pie charts for visual presentation. However, this should be approached with caution. For example, the GO slim contains these terms: "nucleic acid binding", "DNA binding" and "RNA binding". It is not correct to place all three categories in parallel for statistics because the term "nucleic acid binding" actually includes "DNA binding" and "RNA binding". Therefore, *SeaSpider* is also shipped with a set of "GO sleek" with 87 terms, where the parent-child overlapping was removed from GO slim. These "GO sleek" terms are then broken down to the three major GO categories: biological process, cellular component and molecular function, because, for example, "extracellular space" (cellular component) should not be compared side by side with "protein kinase activity" (molecular function).

8.3 In action

With the annotation and GO analysis from *SeaSpider*, a complete list of metabolic genes were obtained from five fish genomes (Table 8.1). Most of these genes can be assigned to EC numbers. Multiple genes can have the same EC number, as they can be isozymes or differ in regulatory contexts, which are subject to further investigations. Some genes fail to get an EC number, because either the EC matching was of limited power or these genes are not *bona fide* metabolic enzymes - for example, transport proteins can also get a GO term under metabolism. It should be pointed out that the EC numbers in MetaFishNet are a result of electronic inference - the Nomenclature Committee of IUBMB actually requires experimental evidence for assigning an official EC number.

As previously mentioned, *SeaSpider* can also be used as a standalone tool. I used *SeaSpider* to annotate the sequences in our Cyprinodon subtractive library. The annotation strategy of *SeaSpider* proves to work well. From the local databases

Table 8.1: Metabolic Enzymes found in five fish genomes

species	number of metabolic genes	number of ECs
zebrafish	3853	654
medaka	3998	765
Takifugu	4103	771
Tetraodon	4424	782
stickleback	4324	791

alone (see Figure 8.1), SeaSpider found homologs for 1106 clones. If queried on GenBank alone, as an annotator usually does, *SeaSpider* returned homlogs for 870 clones. Combined, we obtained annotations for 1334 clones. The gene categories are shown in Figures 8.2, 8.3 and 8.4.

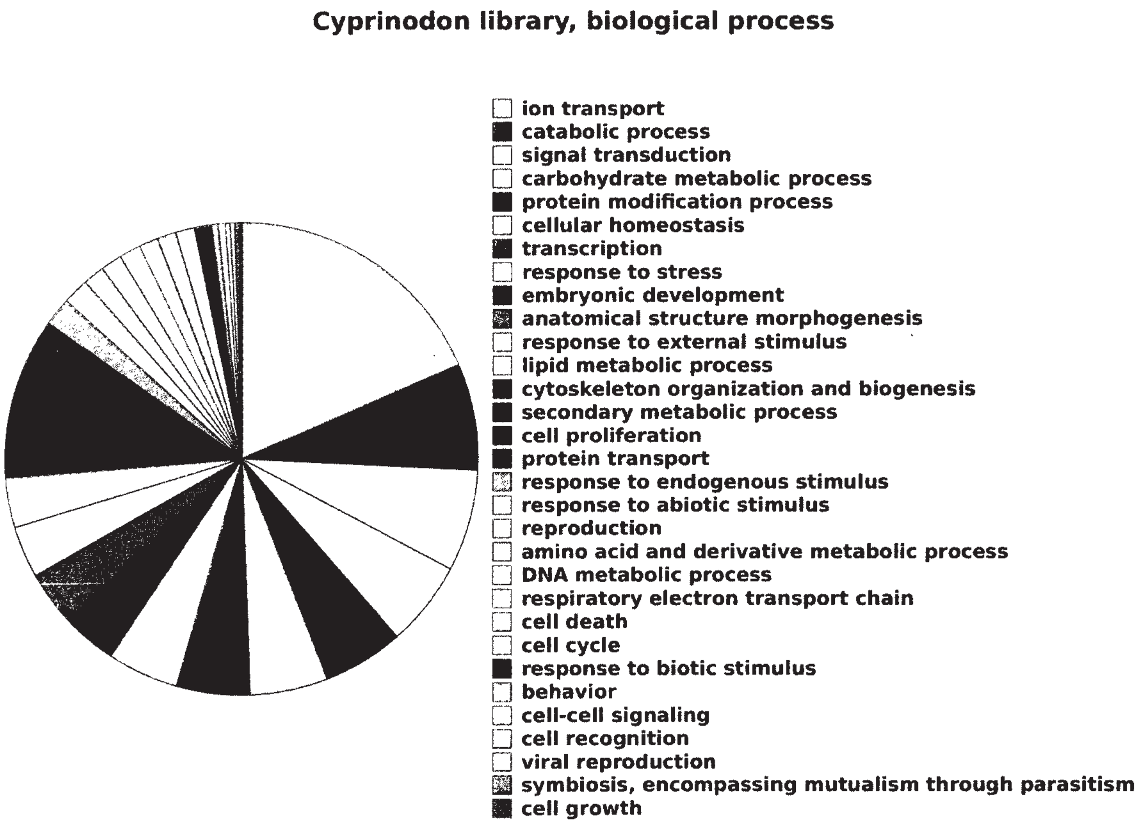


Figure 8.2: Gene categories in Cyprinodon library: biological process.

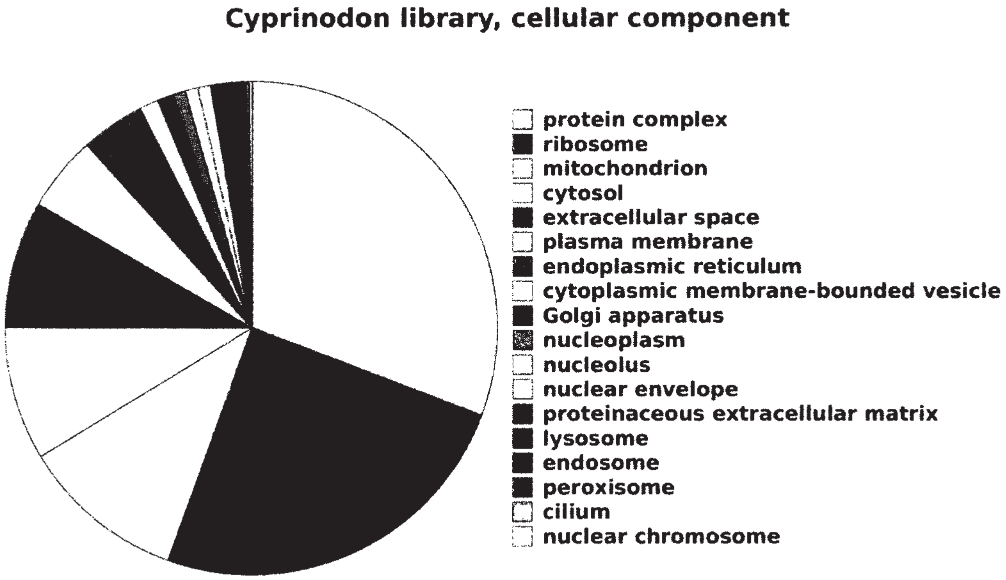


Figure 8.3: Gene categories in Cyprinodon library: cellular component.

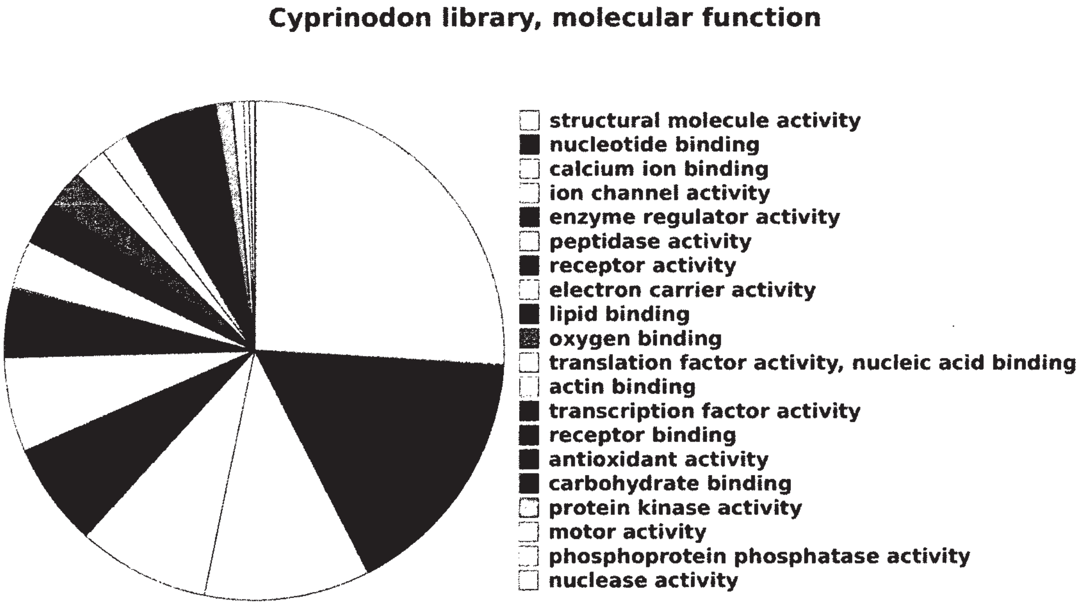


Figure 8.4: Gene categories in Cyprinodon library: molecular function.

8.4 Methods

SeaSpider

BLAST is installed locally. The component *blastall* is used for searching local databases, and the component *blastcl3* is used for querying NCBI remotely. I wrote a new Python wrapper and parser for BLAST. An existing parser is available from the BioPython project. However, the development of BioPython was lagging behind the BLAST version at the time I started my project. And the strong interdependency within BioPython poses a problem in software maintenance. For parsing XML in *SeaSpider*, I use *ElementTree*, which in this case is more memory efficient than standard DOM (Document Object Model) tools. *blastcl3* uses an older XML format that needs to be parsed differently. Python library *subprocess* is used to wrap BLAST. A newer version of this library should be used to avoid a problem in pipe buffering. Database queries are performed via Python binding to MySQL.

The homology criteria are enforced in *SeaSpider* as in previous literature: E-value under 10E-5 and over 33 nucleotides should be identical in an alignment. When one turns to NCBI for annotating new sequences, results often contain clauses like "Homo sapiens cDNA clone clone number ..." or "Tetraodon nigroviridis full-length cDNA". Such results contain little information regarding to the gene function. *SeaSpider* circumvents this problem by a preliminary semantic analysis. An annotation description is favored when it contains meaningful GO words, and disfavored when it contains words like "clone", "genome" or "library". This simple method makes the annotation result from *SeaSpider* more informative.

SeaSpider needs to record the status of its queries internally. This is achieved via Python *shelve*, which is a serialized object database. The most memory consuming part of SeaSpider is the parsing of BLAST results in large batches. E.g., a batch query of 500 sequences may use over 500 MB. This is not a real concern on modern computers, and the batch size can be decreased to accommodate less powerful hardware.

EC numbers

The Enzyme Nomenclature Database was retrieved from Swiss Institute of Bioinformatics (`http://www.expasy.org/enzyme/`, release on 11/04/2008). This Enzyme database contains EC numbers for verified enzymes, but few from fish species. I used two methods to infer EC numbers for fish metabolic genes. The first method was association via human orthologs. When an EC number was associated with the human ortholog of a fish gene, this EC number was carried over to the fish gene. The second method was association via Gene Ontology. GO database keeps external links to ECs, which were carried over during GO annotation by *SeaSpider*. The results from these two methods were then merged to create a list of EC numbers for fish metabolic genes. The Enzyme database only contains descriptions for complete EC numbers (e.g. 2.4.1.14). Sometimes, incomplete EC numbers (e.g. 2.4.-) have to be used. I further included the definitions of enzyme classes and subclasses to make a complete dictionary of all EC numbers in MetaFishNet. This dictionary provides enzyme names for visualized networks.

Chapter 9

Construction of MetaFishNet

We have introduced some basic concepts of metabolic network in Chapter 7. The construction of our MetaFishNet model involves entities of several organizational levels: enzymes, genes, compounds, reactions and pathways. A reaction is described by its compounds and enzymes. The exact cellular location of a reaction is valuable information, but still largely a subject of ongoing investigation. We choose not to include this information in the current version of MetaFishNet. Pathways are human-initiated organizations that are imposed on metabolic networks, mostly by conceptual conventions in biochemistry. The construction process is to reconcile data at all these levels. In this chapter, we will discuss the merging of the two high quality human models, then of the KEGG zebrafish model, and finally of the inferred reactions. An overview of the finished model is given in the end of this chapter.

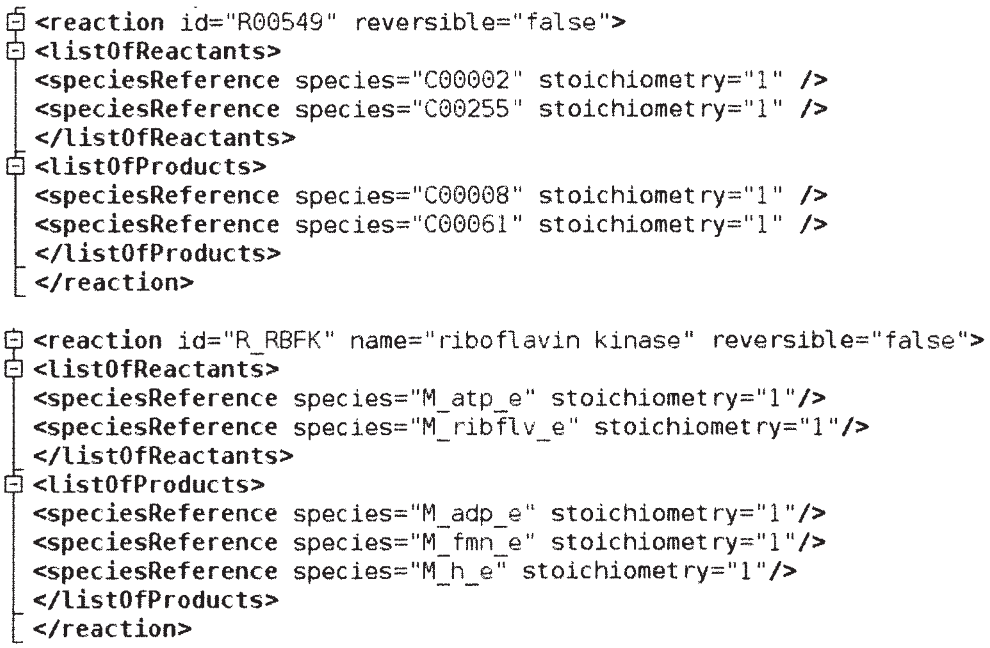
9.1 Merging the two high quality human models

Two high-quality human metabolic models were released in 2007. One was from Palsson's group at University of California at San Diego [24]. This UCSD model contained 1496 genes and 3311 reactions. These included over 1000 transport reactions, and reactions in different cellular compartments were treated as different reactions. The highlight of this work was the manual curation of literature supports, which was labor intensive but improved the data quality. The other human model came from University of Edinburgh [25]. This EHMN (Edinburgh Human

Metabolic Network) model has 2322 genes and 2824 reactions (excluding transport reactions). EHMN model included previous metabolic data from all major databases, and streamlined the identities of compounds. Pathways are specially reorganized in the EHMN model where original KEGG pathway organization became less satisfactory. Our MetaFishNet model will mostly inherit the pathway organization in EHMN model. It should be noted that both "human" models contained many data from other model organisms. Together, these two models offer probably the best annotated, genome-wide metabolic data for vertebrates, and therefore critical references for MetaFishNet. We will first merge these two human models.

The key here is a unified representation of reactions, because once all reactions are in place, the new network can be recovered by connecting the reactions. In practical terms, the unified representation means all enzymes are coded in EC numbers and all compounds are in KEGG compatible IDs (KEGG has one of the largest collections of compounds). The nomenclature of compounds is rarely consistent across literature. The EHMN project did a good job to reconcile them with KEGG IDs. For the compounds not found in KEGG, EHMN project assigned new IDs consistent with KEGG style.

Both UCSD and EHMN models are distributed in SBML (Systems Biology Markup Language, see Figure 9.1) and various formats. However, the completeness of the distributions has to be verified from case to case. SBML distributions usually have a complete definition of network relationships. But the definitions of enzymes and compounds have to be verified from their databases or flat files. I first translated all reactions from both models into unified ECs and compound IDs. Then, I compared the enzymes between pathways from the two human models. The pathways with significant overlaps should be merged. Table 9.1 shows the 49 pathways from UCSD model (91 total) to be merged into the corresponding pathways in EHMN model. Transport reactions from UCSD model were excluded. Two path-



```
<reaction id="R00549" reversible="false">
<listOfReactants>
<speciesReference species="C00002" stoichiometry="1" />
<speciesReference species="C00255" stoichiometry="1" />
</listOfReactants>
<listOfProducts>
<speciesReference species="C00008" stoichiometry="1" />
<speciesReference species="C00061" stoichiometry="1" />
</listOfProducts>
</reaction>

<reaction id="R_RBFK" name="riboflavin kinase" reversible="false">
<listOfReactants>
<speciesReference species="M_atp_e" stoichiometry="1"/>
<speciesReference species="M_ribflv_e" stoichiometry="1"/>
</listOfReactants>
<listOfProducts>
<speciesReference species="M_adp_e" stoichiometry="1"/>
<speciesReference species="M_fmn_e" stoichiometry="1"/>
<speciesReference species="M_h_e" stoichiometry="1"/>
</listOfProducts>
</reaction>
```

Figure 9.1: SBML descriptions of the same reaction from EHMN (top) and UCSD (bottom) models. SBML defines reactants and products clearly. But the identifiers still need a unified nomenclature. Note that the UCSD example contains an extra proton, which is considered a currency metabolite and not included in the EHMN example.

ways from UCSD model, "Nucleic acid degradation" and "Nucleotides" were dismantled (while the reactions were still included). In the merged result, pathway "CYP Metabolism" was merged into "Xenobiotics metabolism"; "Ascorbate and Aldarate Metabolism" and "Vitamin C metabolism" were merged to "Ascorbate (Vitamin C) and Aldarate Metabolism". In the case of EHMN pathway "Urea cycle and metabolism of arginine, proline, glutamate, aspartate and asparagine", it was oversized so that the several overlapping smaller pathways in UCSD model were kept instead. The rest of pathways were not affected.

Within each pathway, identical reactions were merged. Two reactions were considered identical when they have identical enzymes and identical compounds excluding currency metabolites, because currency metabolites might or might not be included in the original descriptions. In total, 2824 reactions from EHMN and

1859 reactions from UCSD were merged to 3953 reactions and 106 pathways.

Table 9.1: The pathways in UCSD model to be merged with corresponding EHMN pathways.

EHMN pathway	**UCSD pathway**
Aminosugars metabolism	Aminosugar Metabolism
Arachidonic acid metabolism	Eicosanoid Metabolism
Bile acid biosynthesis	Bile Acid Biosynthesis
Biopterin metabolism	Tetrahydrobiopterin
Butanoate metabolism	Butanoate Metabolism
C21-steroid hormone biosynthesis and metabolism	Steroid Metabolism
De novo fatty acid biosynthesis	Fatty acid elongation
Fructose and mannose metabolism	Fructose and Mannose Metabolism
Galactose metabolism	Galactose metabolism
Glycerophospholipid metabolism	Glycerophospholipid Metabolism
Glycerophospholipid metabolism	Triacylglycerol Synthesis
Glycine, serine, alanine and threonine metabolism	D-alanine metabolism
Glycine, serine, alanine and threonine metabolism	Glycine, Serine, and Threonine Metabolism
Glycolysis and Gluconeogenesis	Glycolysis/Gluconeogenesis
Glycosphingolipid metabolism	Sphingolipid Metabolism
Histidine metabolism	Histidine Metabolism
Lysine metabolism	Lysine Metabolism
Methionine and cysteine metabolism	Cysteine Metabolism
Methionine and cysteine metabolism	Methionine Metabolism
Methionine and cysteine metabolism	Taurine and hypotaurine metabolism
N-Glycan biosynthesis	N-Glycan Biosynthesis
O-Glycan biosynthesis	O-Glycan Biosynthesis
Pentose phosphate pathway	Pentose Phosphate Pathway
Phosphatidylinositol phosphate metabolism	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Phosphatidylinositol phosphate metabolism	Inositol Phosphate Metabolism
Porphyrin metabolism	Heme Biosynthesis
Porphyrin metabolism	Heme Degradation
Proteoglycan biosynthesis	Chondroitin / heparan sulfate biosynthesis

Continued on next page

EHMN pathway	UCSD pathway
Purine metabolism	IMP Biosynthesis
Purine metabolism	Purine Catabolism
Purine metabolism	Salvage Pathway
Pyrimidine metabolism	Pyrimidine Biosynthesis
Pyrimidine metabolism	Pyrimidine Catabolism
Selenoamino acid metabolism	Selenoamino acid metabolism
Squalene and cholesterol biosynthesis	Cholesterol Metabolism
TCA cycle	Citric Acid Cycle
Tryptophan metabolism	Tryptophan metabolism
Tyrosine metabolism	Phenylalanine metabolism
Tyrosine metabolism	Tyrosine metabolism
Tyrosine metabolism	Tyr, Phe, Trp Biosynthesis
Valine, leucine and isoleucine degradation	Valine, Leucine, and Isoleucine Metabolism
Vitamin A (retinol) metabolism	Vitamin A Metabolism
Vitamin B1 (thiamin) metabolism	Thiamine Metabolism
Vitamin B2 (riboflavin) metabolism	Riboflavin Metabolism
Vitamin B3 (nicotinate and nicotinamide) metabolism	NAD Metabolism
Vitamin B5 - CoA biosynthesis from pantothenate	CoA Biosynthesis
Vitamin B6 (pyridoxine) metabolism	Vitamin B6 Metabolism
Vitamin B9 (folate) metabolism	Folate Metabolism
Vitamin H (biotin) metabolism	Biotin Metabolism

9.2 Computing with KEGG

To extract the exact zebrafish network model out of KEGG turned out to be a technical challenge. KEGG bears a strong mark of its root as a visualization aid. So the graphical representation can hardly be separated from the logical, abstract network model. A paper in 2004 [176] stated that no automated way was available to extract network models from KEGG. KEGG offers an XML (Extensible Markup Language) distribution (called KGML) as well as a network API (application programming interface) to its databases. But both contain ambiguities. The KGML files are again designed around graphic display not molecular interactions (compare Figure 7.1 A and B: we would like to have B not A). In the API to KEGG databases, the definitions of element relationships are again based on the manually constructed graphs, like the KGML, mixed with visual elements. If one looks at the defined components separately, the compounds and enzymes can be obtained by association to reactions. The catch is, there seems no way to distinguish reactants from products. And both compounds and glycans are used as reaction components, but the overlapping nomenclature is not transparent. Therefore, a practical solution may be to combine KGML files and database API. KGML defines the scope of reactions and API confirms relationships. I wrote a program following this design. With this program, 101 zebrafish metabolic pathways from KEGG were retrieved and converted to definitive SBML models. Individual enzymes, compounds, reactions and pathways could then be easily extracted from these SBML files.

The stock zebrafish metabolic network from KEGG contained 101 pathways, 517 ECs and 1031 reactions. Among these reactions, 846 were already in the merged human network model. Thus, only 185 reactions from KEGG zebrafish network were added into our baseline reference. In a similar fashion to the merging of two human models, 54 pathways from KEGG zebrafish network were merged into pathways in the human model. In addition, the zebrafish pathway "Lipoic acid metabolism" was merged into "Lipoate metabolism" pathway.

Besides the zebrafish models, KEGG API was also used to fetch definitions of compounds. KEGG LIGAND database was used to query the compounds associated with enzymes. For the enzymes not included in existing models, this was the way to infer new reactions, which can be further organized into new pathways (*ab initio* construction).

9.3 Ab initio construction

All gene sequences from five fish genomes, with untranslated regions, were retrieved from Ensembl database. Gene ontology terms and descriptive annotations were available for a good number of zebrafish genes, while such information was scarce for other species. Human ortholog relationships were also obtained from Ensembl. All these data were recompiled into a local database for this project.

SeaSpider was used to assign GO terms to all fish genes wherever possible. For the genes associated with metabolism, Enzyme Commission (EC) numbers were then assigned wherever available. From the analysis of five fish genomes, we obtained 911 ECs. 561 of them could be found in the baseline reference network. For the remaining 350 ECs, their associated compounds were retrieved from KEGG wherever available. These enzyme-compounds interactions formed 260 newly inferred reactions. Since there was no way to distinguish reactants from products in these inferred metabolic data, the directions of these reactions were treated as unknown.

I also included glycans in the "compounds". Glycans were treated separately in KEGG, and the nomenclature was unclear. For example in KEGG, both "G11113" (a glycan ID) and "C00008" (a compound ID) denote ADP molecule. I then had to include "G11113" in the list of currency metabolites where ADP is a member.

These newly inferred reactions, plus the isolated reactions from the reference network, can be connected by shared compounds. This connected network shall be organized into new pathways. One way to derive new pathways from a network is by identifying network modules [159]. The identification of network modularity is a challenging and computational intensive problem. Newman proposed an attractive heuristic algorithm, which exploits an expression of modularity in terms of the eigenvectors of a characteristic matrix for the network [161]. Then, the network can be partitioned into modules by the signs of elements in the eigenvectors. I implemented this algorithm and used it to divide the inferred network reactions into new pathways. This modularity approach produced 45 "provisional pathways". Upon closer inspection, provisional pathway 36 was merged as part of pentose phosphate pathway, and provisional pathway 44 was dismantled to isolated reactions. In the end, 43 "provisional pathways" made their way into the current MetaFishNet version. I used a method of GO enrichment to assign possible functions to these provisional pathways. That is, when we check the most frequent metabolic GO terms associated with the enzymes in a pathway, they are likely to reflect the function of the whole pathway. This method was tested on the known pathways and it worked reasonably well. All these "provisional pathways" or "inferred pathways" follow a name format as "MetaFishNet prov_path ...".

9.4 Finished model

The data integration at reaction level is shown in Figure 9.2. The scope of MetaFishNet is defined by fish genomes, that is, what metabolisms fish may have are defined by the enzymes present in their genomes. In total, 3469 reactions and 169 pathways were included in the first version of MetaFishNet.

The core data are stored in a MySQL database (Figure 9.3), where the relationships among genes, enzymes, compounds, reactions and pathways are defined. I adopted the primary gene IDs from Ensembl. This MetaFishNet database also includes zebrafish gene IDs from GenBank and ZFIN, so that users can look up genes by these ID systems. However, as discussed earlier, fish genomics is still evolving and most gene identifications will have to be via sequence comparison by *SeaSpider.*

Besides the core database, MetaFishNet project has several components, as shown in Figure 9.4. A local copy of Gene Ontology database is maintained. The

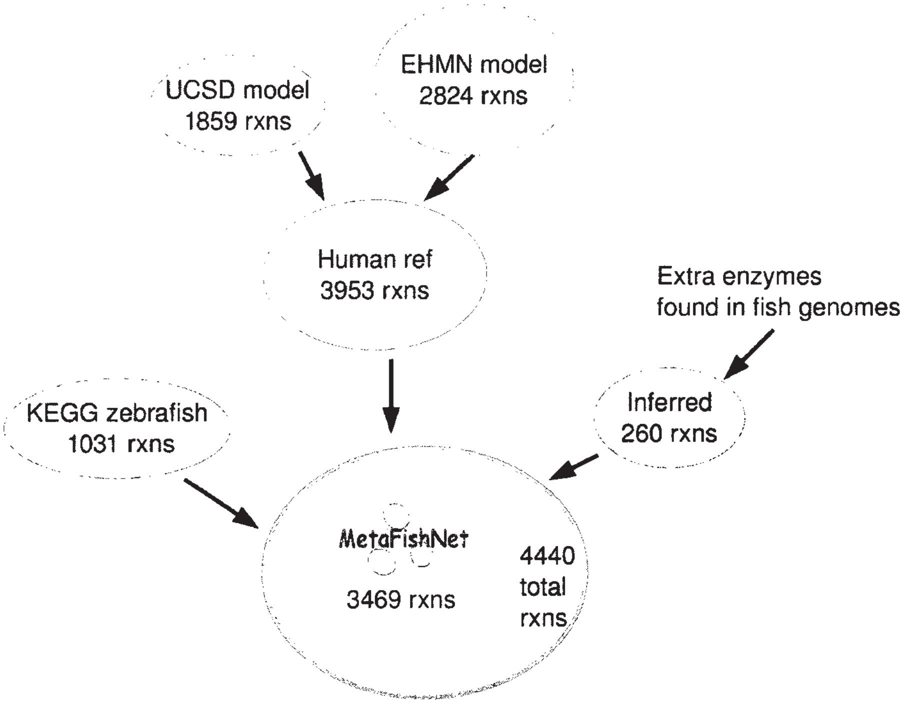


Figure 9.2: Data integration at reaction level for MetaFishNet. The UCSD and EHMN models were merged into a human reference network, which was then merged with the KEGG zebrafish model and newly inferred reactions based on genome sequences. The total reference model has 4440 reactions, while 3469 reactions are included in fish metabolic network. A number of spontaneous reactions (without enzymatic catalysis) are also included.

genic information from Ensembl is stored in a locally designed database, because it is used too frequently for remote API and the Ensembl database is too big and complex. The sequence databases serve *SeaSpider* for its sequence analysis. The utilities also include programs for visualization, expression data mapping and web interface.

9.5 Visualization

Programmatic visualization of networks/pathways is a challenging problem. The pathway graphs in KEGG were actually manually produced and many programs try to copy their layout. Since our MetaFishNet substantially outgrew

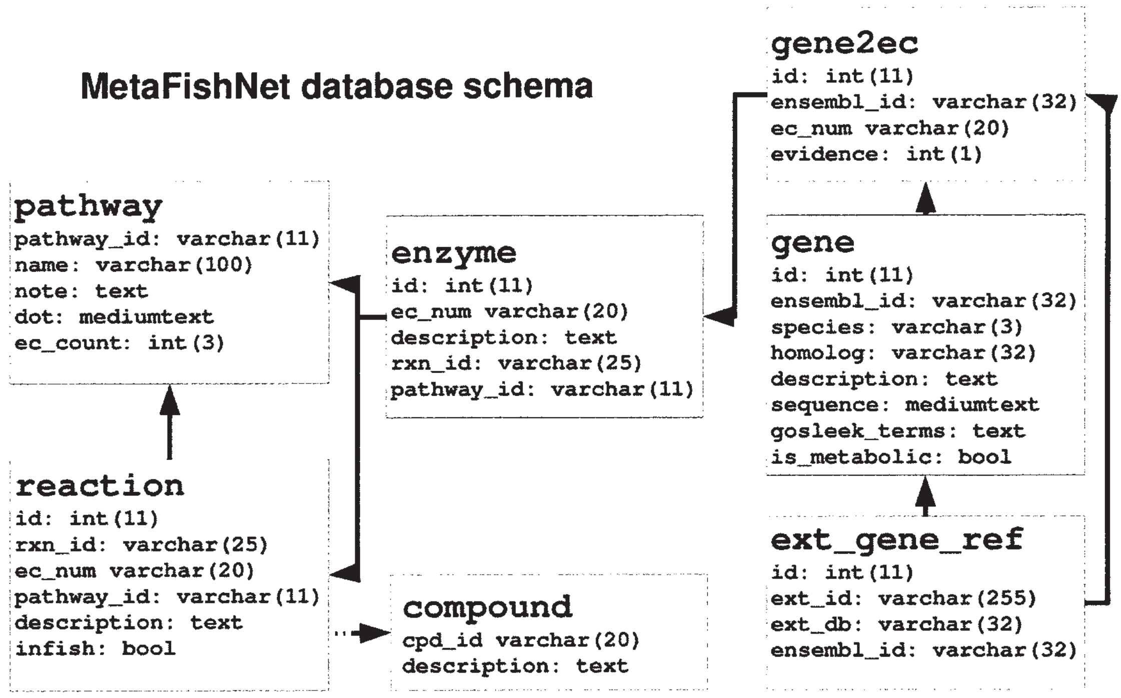


Figure 9.3: Database schema for MetaFishNet. The linkage between "compound" and "reaction" is not directly through attribute matching. A simple text parsing of reaction.description makes the connection. This trick saves storage space and improves database performance.

the KEGG pathways, it is not an option to use KEGG layouts. So the goal in this project is to visualize pathways anew and automatically. Tools like CellDesigner [177] still require laborious manual editing, less favorable for a genome-wide project. CytoScape [178] can do automatic layout. But it is not particularly suited for metabolic networks - difficult to balance global view and details, easily resulting in "hairball" pictures.

MetaFishNet project adopts the *dot* tool from Graphviz, which is a mature open source package from AT&T Research Labs (graphviz.org). Bindings of Graphviz are available in common programming languages. *dot* comes handy in Linux distributions - it is both a graph description language and a utility program. Graphviz and *dot* take care of the low level graphic generation, so that I could focus on the middle level of organizing nodes and network structures.

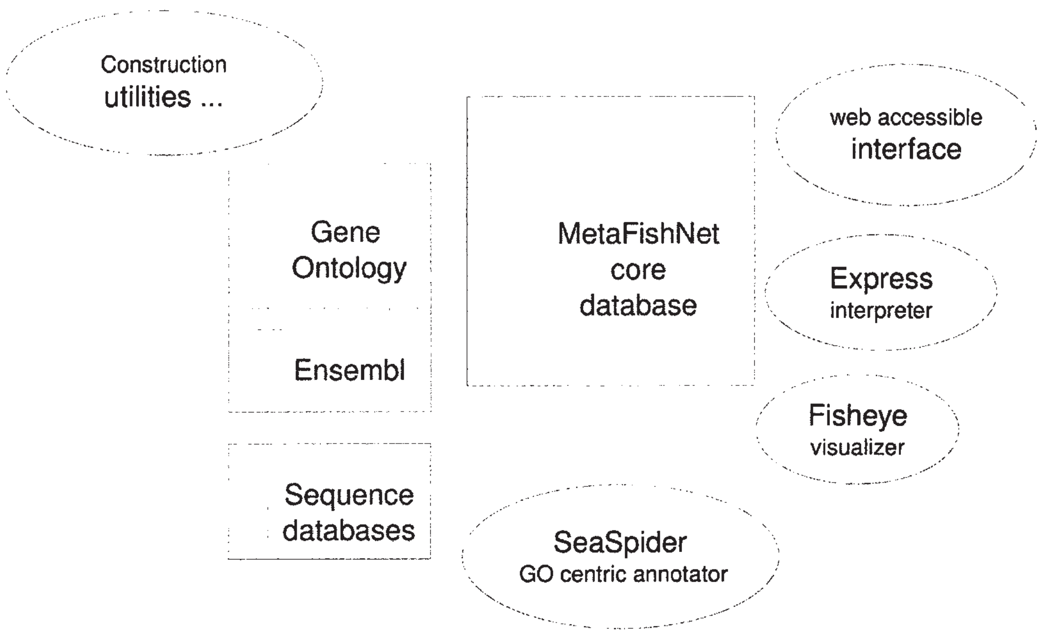


Figure 9.4: Major components of MetaFishNet.

My programs actually generate *dot* files as intermediates. They serve as archives, and allow modularized or manual manipulations. There are several tricks to improve the clarity. Long names are wrapped into multiple lines, and alternative names are used if the default names are insanely lengthy. Similar branches can be merged. One important feature of my programs is the "zoomglass" function. Some highly connected nodes will clutter the whole picture because they drag edges all over the place. "zoomglass" will break up such edges, color the nodes in yellow and attach them locally (e.g. the NADP+ in Figure 9.5). This technique reduces the complexity in the two-dimension space. It is especially useful for the "currency metabolites" in metabolic networks.

An example is given in Figure 9.5. Enzymes are drawn in filled ellipses, compounds placed as plain texts, and edges in directed arrows. When gene expression data are mapped to the pathway, upregulation is colored in red and downregulation in green, while violet is the default color and for no change. This version is a typical bipartite network, using both enzymes and compounds as nodes. My experimentation of using enzymes alone did not yield good clarity, because the enzyme/gene networks have too many edges. Overwhelming edge numbers are still a problem for larger pathways, even with the "zoomglass" function. An alternative might be to use reactions as layout units. These are subjects for further investigations.

9.6 Web interface and utilities

The data and programs of MetaFishNet will be made publicly available. Even so, it is a complex bioinformatic project and technically demanding for users to run a local version. It will be more useful if provided as a web service. And web browsers themselves are a convenient graphic user interface.

A few tasks in MetaFishNet project, including *SeaSpider* annotation and pathway graph generation, are computationally intensive. They do not make typical, responsive web applications. It is therefore reasonable to split the web interface into two layers: a frontend user interaction layer and a backend application service layer. For the latter, an application server is needed, while the former can be hosted on free web accounts like Google's App Engine (GAE). The two layers communicate through remote APIs.

The component of mapping expression data to MetaFishNet is shared between the standalone version and web version. User submitted genes are mapped to enzymes by either their IDs or *SeaSpider* sequence comparison. Then the relevant pathways are color coded, and graphs like Figure 9.5 are generated on the fly. If the regulation on a specific gene is true positive, all pathways containing that gene are valid candidates. In the case of preliminary analysis of high throughput data without independent validation, it is good to assign statistical significance to candidate pathways. The common method is Fisher's exact test, as in most Gene Ontology tools [56, 58, 60]. That is, a pathway contains a certain number of genes out of the whole genome. From this information, one can compute an expected number of genes in this pathway from a set of selected genes, such as a set of differentially expressed genes in microarray data. When the real number of regulated genes in the pathway exceeds this expectation, an "enrichment" occurs to this pathway. Similarly, each Gene Ontology term (category) is associated with a number of genes. The enrichment of GO category can be computed in the same way. Features of both MetaFishNet pathway enrichment and GO enrichment are implemented in MetaFishNet package.

9.7 Methods

Integration of heterogeneous data

Reactions from the two human models were extracted by a combination of parsing XML files (SBML) and flat files (corresponding to their original databases). Flat files of UCSD model were obtained from personal communication (Jan Schellenberger at Palsson lab). XML parsing was done with xml.dom.minidom implementation in Python libraries. And identifiers were extracted from flat files with the help of regular expression. Typically, four steps were involved in merging models:

1. Unifying all identifiers to compatible formats, e.g., all compounds to KEGG compatible IDs.
2. Comparing pathways, deciding what pathways to merge or change.
3. Comparing reactions, removing repetitive reactions.
4. Manual inspection of merged data. E.g., some pathways are functionally identical but differ significantly in source models. Such pathways require manual

merging.

The merging of the KEGG zebrafish model with the human model followed the same procedure. Reactions with enzymes found in fish genomes were included in MetaFishNet. The spontaneous reactions (without an enzyme) may be necessary for mass flow in metabolic pathways and were kept in MetaFishNet.

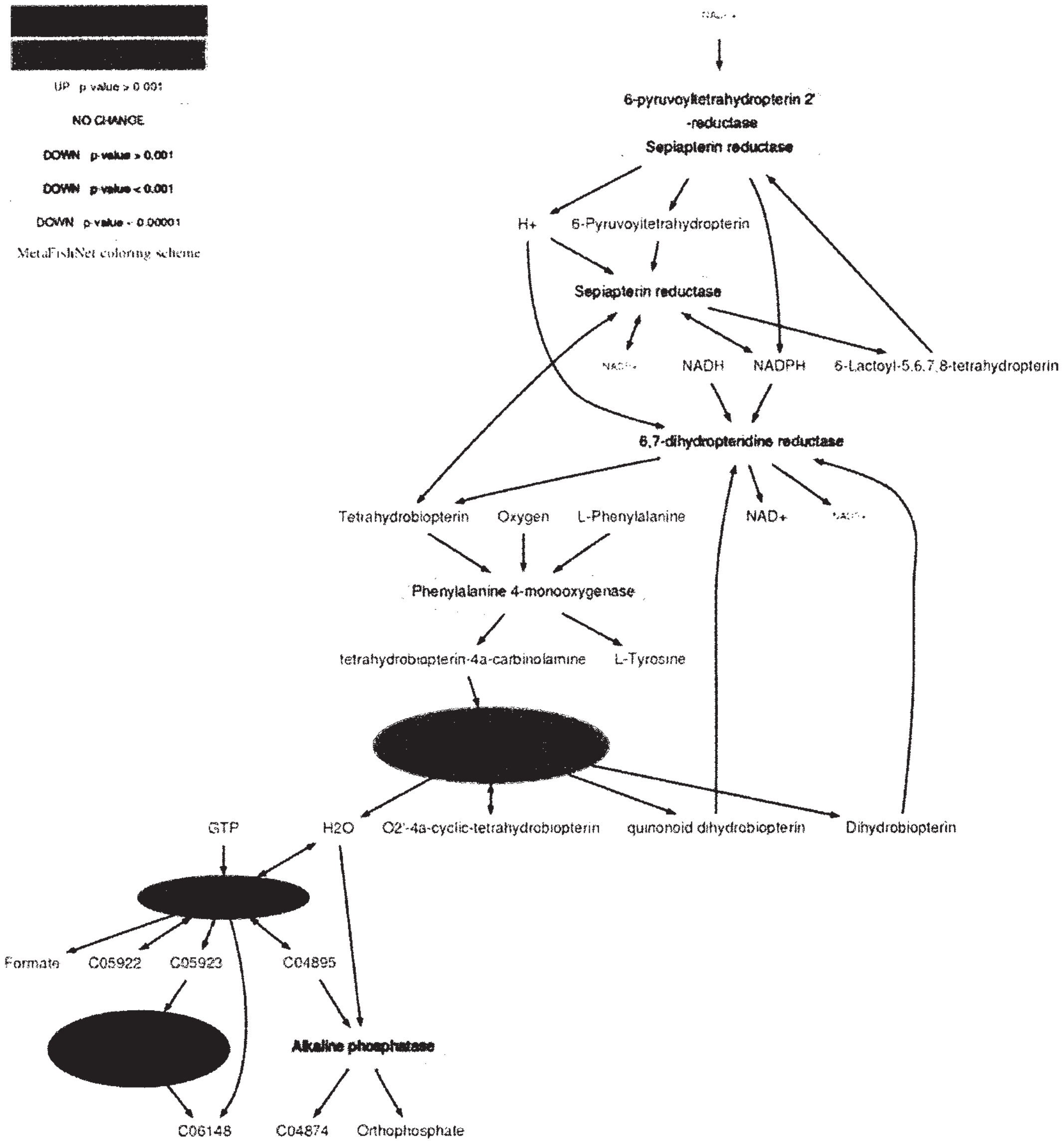
Several techniques were used to handle data from KEGG. XML files were retrieved by an automated FTP script. KEGG web API was called to map relationships between different entities, e.g., enzymes and compounds. My program of extracting network data out of KEGG combined the parsing of KGML and calling of web API.

Local MySQL server 5.0 runs on Linux (Ubuntu 8.04). All gene sequences (cDNA, UTRs included) of zebrafish, medaka, Takifugu, Tetraodon and stickleback were exported from Ensembl database via its BioMart feature. Their human homologs and zebrafish GO annotation were also retrieved from Ensembl, then reconstructed into local mysql databases. A local copy of complete Gene Ontology database (seqdblite) is maintained. And the core data for MetaFishNet were compiled into a new MySQL database (as in Figure 9.3).

Visualization and web utilities

Each of the 169 pathways was given a description file and an equivalent SBML file. These pathway files were then analyzed by my *Fisheye* program for producing *dot* files that contained all elements for final visualization. Even though the low level graph layout and generation are taken care by Graphviz and PyGraphviz library (`http://networkx.lanl.gov/pygraphviz/`), *Fisheye* is still fairly complex. It computes full sets of node types and edges, associates IDs to names, estimates zoom levels, and segments modules optionally. A *dot* file and a PNG graph were generated for each pathway. For expression data mapping, new graphs are generated on the fly. I use HSB (Hue, Saturation, Brightness) system for coloring. When expression data are mapped to MetaFishNet pathways, the corresponding nodes, according to their p-values, are directly assigned HSB values to control their colors.

The web service of MetaFishNet was prototyped on Django framework (`http://www.djangoproject.com`). It followed a standard "Model-Control-View" paradigm of software development. The "View" adopted the dynamic templating system from Django. The frontend web server and backend application server are connected via remote APIs. Python Imaging Library (PIL) was used to generate thumbnails for the web site.



MetaFishNet coloring scheme

MetaFishNet: Biopterin metabolism

Figure 9.5: Visualization of MetaFishNet, example of Biopterin metabolism pathway. Changes of gene expression can be overlaid to the network. In this example, violet is the default color for enzymes. When an enzyme is upregulated, it is coded in red, and downregulated ones in green. This way, expression data are put into the context of molecular networks (expressions in the example picture are fictional).

Chapter 10

Applications of MetaFishNet

Summary

This chapter starts by comparing the metabolic genes and pathways between fish and human. Close to half of the genes in fish have identifiable orthlogs in human. Because fish had an extra whole genome duplication during the evolution, the number of fish genes exceed their human orthologs by over 10%. However, the numbers of metabolic genes are a lot closer between fish and human, suggesting many metabolic pathways are evolutionarily conserved. An example is that the "proteoglycan biosynthesis" pathway, which is involved in a major class of enzyme deficiency diseases, may be identical between fish and human. The difference between fish and human pathways is also revealing. For example, MetaFishNet suggests fish have high concentration of omega-3 fatty acids because they lack some metabolizing enzymes in omega-3 fatty acids pathway. The use of MetaFishNet as a tool for interpreting gene expression data is demonstrated on a previously published data set. The reanalysis by MetaFishNet suggests that several metabolic pathways, including "Tyrosine metabolism", "Xenobiotics metabolism" and "3-Chloroacrylic acid degradation", are repressed in zebrafish liver tumor. This also opens up the possibility of using metabolites from these pathways as diagnostic markers.

10.1 Metabolic genes show less evolutionary diversity

It is now widely accepted that teleost fish underwent an extra round of genome duplication after their evolutionary separation from mammals [10,179]. Genome duplication is an important mechanism for generating gene diversity, as the extra copy can evolve more freely than the single copy before duplication. Only a small portion of these duplicated genes would gain new functionality and remain, while most duplicated genes get lost over time.

During the construction of MetaFishNet, we have collected a list of all fish metabolic genes, which can be compared to their human orthologs. We have noticed that the level of ortholog mapping differs between metabolic genes and other genes. As seen in Table 10.1, for the identifiable orthologs, most of the five fish have over 10% more genes than human, while the numbers among metabolic genes are significantly less. The final numbers may vary when the genomes are more accurately annotated. Still, these data suggest that metabolic genes are better conserved between human and fish than other genes. We think that metabolism was established early in evolution: by the time of the genome duplication for fish, the central metabolic machinery was already well tuned and left little room for changes. By implication, research on some fish metabolic pathways may be easily extrapolated to human. We will take a further step to examine the conservation of pathways.

Table 10.1: Comparisons between fish and human orthologs. An extra round of genome duplication produced more genes in fish than human. The number of total human orthologs found in a fish species is typically around 12,000.

species	extra duplicated genes (%)	extra duplicated metabolic genes (%)
zebrafish	15.35	0.58
medaka	8.87	1.50
Takifugu	12.16	3.75
Tetraodon	14.36	5.77
stickleback	11.90	4.48

10.2 Comparison between human and fish metabolic pathways

At the enzyme level, we have identified 911 enzymes from fish genomes. They overlap with the human data by 595 enzymes (65%, Figure 10.1). The true overlap may be greater because the EC numbers in fish were computationally inferred, and are not as well curated as human ECs. We can nonetheless start some comparison between human and fish at the pathway level.

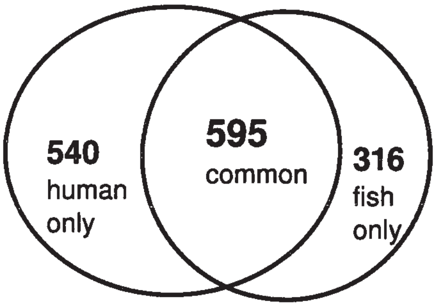


Figure 10.1: Among all metabolic enzymes, 595 are common between human and fish. The enzyme number in human is slightly higher, in spite of the fact that the number of metabolic genes in fish exceeds that in human. This is probably because human enzymes are better annotated, therefore we have better success of mapping genes to enzymes. The inferrence of fish ECs was described in Chapter 8.

The majority of pathways have over 50% enzymes in common between human and fish. Table 10.2 shows the most and least conserved pathways between human and fish, in term of the numbers of overlapping enzymes. Since most biomedical research in fish aims to extend the results to human, this pathway comparison reveals important information on how well fish may model human on a specific subject. For instance, fish may be a good model for studying vitamin B9, but probably a poor model for studying vitamins C and H.

In a sizable pathway, "proteoglycan biosynthesis", all 16 enzymes are common between human and fish. This suggests that the whole pathway may be identical between human and fish. Impairment of the proteoglycan biosynthesis pathway is actually responsible for a major class of enzyme deficiency diseases, mucopolysaccharidosis. Seven clinical types, including Hurler syndrome and Hunter syndrome, have been identified in this class, depending on defects of different enzymes in the pathway. Small fish are attractive disease models because of their high throughput capacity. Given the great similarity between human and fish in this pathway, fish may be a good model for studying mucopolysaccharidosis.

Omega-3 fatty acids are deemed essential nutrients, boosting a popular dietary preference for fish and fish oil consumption. But fish, just like humans, do not produce omega-3 fatty acids *per se* - they accumulate them from their diet, algae. How do fish accumulate omega-3 fatty acids in their body? The molecular mechanism of this question is still unidentified. An interesting hypothesis is, however, strongly suggested by our MetaFishNet model. As shown in Figure 10.2, compared to the 12 enzymes in human pathway for Omega-3 fatty acid metabolism, fish only have 7 of them. As a result, fish can easily process the metabolites in the top part and the bottom part of the pathway, but not the intermediate metabolites, which will then accumulate to a high level. In fact, these intermediate compounds include variants of most of the common omega-3 fatty acids, such as alpha-Linolenic acid,

Table 10.2: Comparisons between fish and human metabolic pathways: the most conserved and least conserved. The ratio is number of shared ECs over number of human ECs. Only pathways with three or more enzymes were considered. The hypothetical pathways inferred by modularity analysis (see Chapter 9) were excluded from comparisons.

Most conserved pathways				
pathway	human ECs	fish ECs	overlap	ratio
1- and 2-Methylnaphthalene degradation	2	3	2	1
Fatty acid oxidation	2	2	2	1
Glutathione metabolism	4	4	4	1
Hyaluronan Metabolism	3	3	3	1
Limonene and pinene degradation	3	4	3	1
Proteoglycan biosynthesis	16	16	16	1
Glycosphingolipid biosynthesis - ganglioseries	10	9	9	0.9
Glycosphingolipid metabolism	27	24	24	0.88
N-Glycan Degradation	8	7	7	0.87
Di-unsaturated fatty acid beta-oxidation	7	6	6	0.85
Glutamate metabolism	14	12	12	0.85
TCA cycle	18	15	15	0.83
Vitamin B9 (folate) metabolism	17	14	14	0.82
Linoleate metabolism	11	9	9	0.81

Least conserved pathways				
pathway	human ECs	fish ECs	overlap	ratio
Phytanic acid peroxisomal oxidation	13	5	5	0.38
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	3	1	1	0.33
Vitamin H (biotin) metabolism	6	2	2	0.33
Glyoxylate and Dicarboxylate Metabolism	7	2	2	0.28
Pentose and Glucuronate Interconversions	9	2	2	0.22
Ascorbate (Vitamin C) and Aldarate Metabolism	8	1	1	0.12

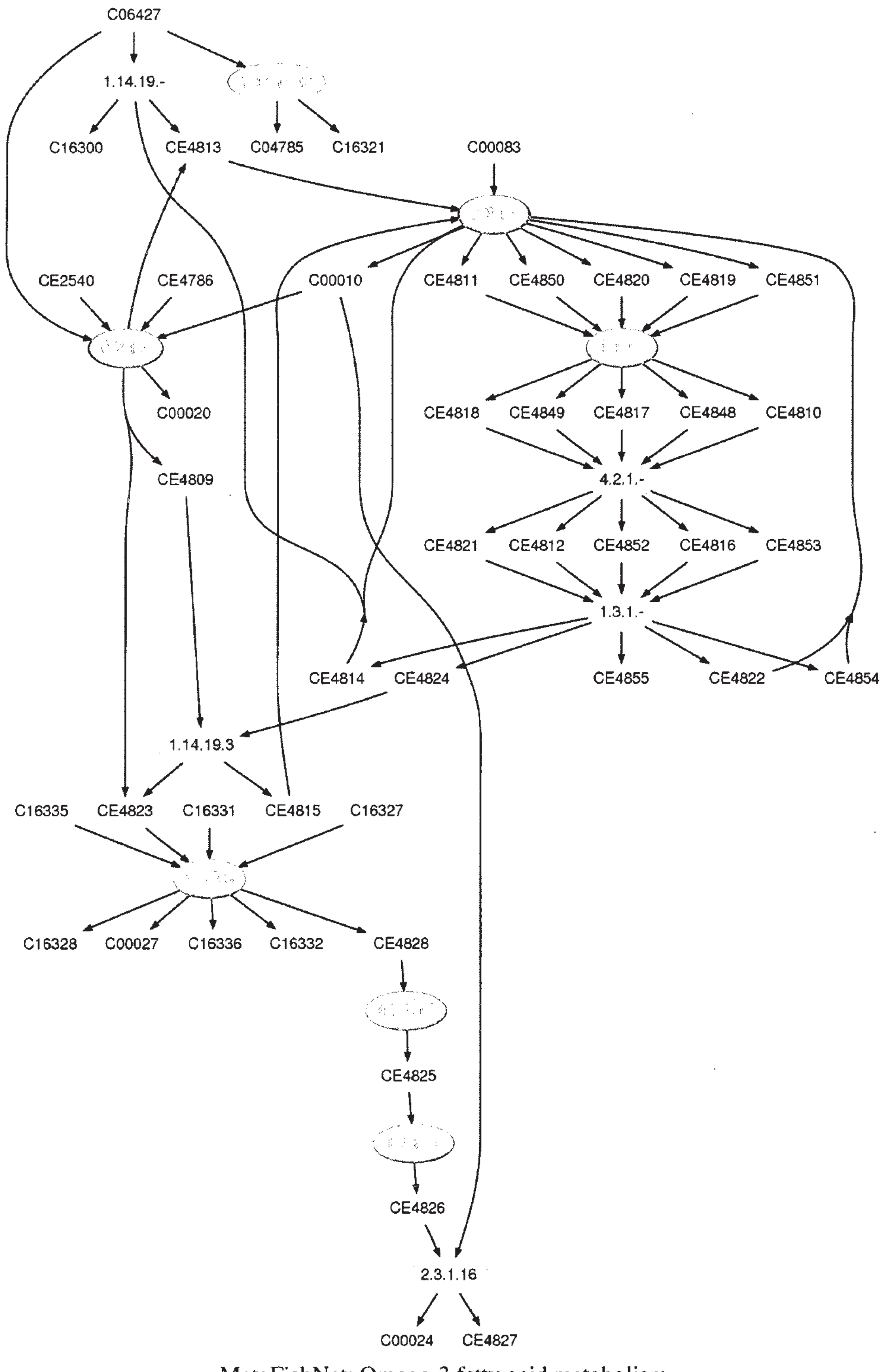
Stearidonic acid, Eicosatetraenoic acid, Eicosapentaenoic acid, Docosapentaenoic acid and Tetracosapentaenoic acid. It will be interesting to see if this computationally generated hypothesis will be supported by experimental data. We would also like to point out that, the real metabolic network is rather dynamic: different catalytic rates and tissue preferences may also have a role.

We have demonstrated that MetaFishNet is a powerful tool for generating new hypotheses. In fact, all of the 43 new pathways generated in this study can be viewed as hypotheses subjected to future tests. The current version of MetaFishNet was constructed by aligning enzymes. A desirable next step is to verify the metabolic reactions through systematic experiments. With this MetaFishNet framework in place, high throughput chemical genomics and metabolomics make such an approach very feasible. Then the model can be refined iteratively [180].

10.3 Zebrafish liver cancer misregulates several metabolic pathways

This section is a case study of applying MetaFishNet model to the analysis of gene expression data. Zebrafish has been long advocated as a model for cancer studies. Gong and coworkers conducted microarray experiments to examine the similarity between zebrafish and human liver tumors at the level of gene expression [181]. Although they found the overlapping of gene expression was statistically significant, in-depth data analysis was limited to Gene Set Enrichment Analysis (GSEA) and particular attentions to two signaling pathways (Wnt-beta-catenin and Ras-MAPK). We shall demonstrate here that MetaFishNet is a valuable new addition to the arsenal of data analysis.

The microarray data from Lam et al. were retrieved, and analyzed independently. Significance Analysis of Microarrays (SAM [43]) was used to select 219



MetaFishNet: Omega-3 fatty acid metabolism

Figure 10.2: Omega-3 fatty acid pathway. The enzymes that are found in fish are colored in forest green, and in the two gray shades. This shows that fish share enzymes with human the top part and the bottom part of this pathway, but lack enzymes that convert the intermediate metabolites, which are the source of omega-3 fatty acids important to human health.

Table 10.3: Metabolic genes in Figure 10.3.

Clone ID	GenBank ID	EC numbers	Fold change
5270127	AF170069	(1.11.1.7, 1.11.1.6)	0.174
5270152	AF309556	(1.14.19)	0.096
5269753	AF295407	(1.2.1.1, 1.1.1.1)	0.097
5269585	AF189238	(1.13.11, 1.13.11.5)	0.191
5270094	AF057713	(1.14.14.1, 1.14.14)	0.283
5269737	AF254954	(1.2.1.3, 1.2.1.5)	0.474
5270056	AF248042	(1.14.14, 1.14.14.1)	0.263

differentially expressed genes between tumor samples and controls (Figure 10.3). I chose SAM because of its more conservative nature [45], which also showed in a smaller number of selected genes in comparison to 2315 genes selected in the original paper.

Among these 219 genes, there are 7 metabolic genes, all downregulated in tumor samples (Table 10.3). These 7 genes are involved in a number of metabolic pathways, and the statistical significance on each pathway is assessed by Fisher's exact test (Table 10.4), as described in Chapter 9. The result suggests that several metabolic pathways are misregulated in zebrafish liver cancer. Among them, Xenobiotics metabolism and ROS Detoxification are cellular protective mechanisms; Fatty Acid Metabolism and Leukotriene metabolism may reflect the lipid processing role of liver. The pursuit of diagnostic metabolites is under very active investigation at this moment [183–188]. Our MetaFishNet pathway analysis suggests the possibility of using metabolites related to these pathways as diagnostic markers. Furthermore, the regulations can be visualized in the context of each pathway, as exemplified in Figure 10.4. The cross-board supression of these metabolic pathways in tumor samples is really intriguing.

Data analysis at the pathway level has two advantages: it is less susceptible to noise than analysis at the level of individual genes, and gives contextual insights to biological mechanisms [189–191]. MetaFishNet has demonstrated good promise

Table 10.4: Metabolic pathways that are affected in zebrafish liver cancer with $P-value < 0.05$.

MetaFishNet pathway	selected enzymes	enzymes in pathway	P-value
Tyrosine metabolism	7	39	0.000
Xenobiotics metabolism	3	8	0.000
3-Chloroacrylic acid degradation	2	2	0.000
Tryptophan metabolism	4	25	0.000
Fatty Acid Metabolism	3	12	0.000
Glycolysis and Gluconeogenesis	4	29	0.001
Leukotriene metabolism	3	17	0.001
Histidine metabolism	2	12	0.011
Ascorbate (Vitamin C) and Aldarate Metabolism	1	1	0.013
Androgen and estrogen biosynthesis and metabolism	2	15	0.016
Bile acid biosynthesis	2	16	0.019
ROS Detoxification	1	2	0.026
1- and 2-Methylnaphthalene degradation	1	3	0.039

to bring these advantages into fish studies.

10.4 Methods

The orthologous relationships of fish genes to human were retrieved from Ensembl database (see Chapter 9). Links of enzymes to diseases were based on Online Mendelian Inheritance in Man (OMIM, `http://www.ncbi.nlm.nih.gov/omim`) database. DNA microarray data from Lam et al [181] were retrieved from Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`) via accession number GSE3519. The arrays contained 16512 features, with 10 tumor samples and 10 control samples. Gene identifiers were converted from the internal IDs in Gong lab to GenBank IDs. Implementation of SAM in TM4 package [182] was used, with Delta value 2.38 and 200 permutations. The mapping of genes to MetaFishNet was described in Chapter 9.

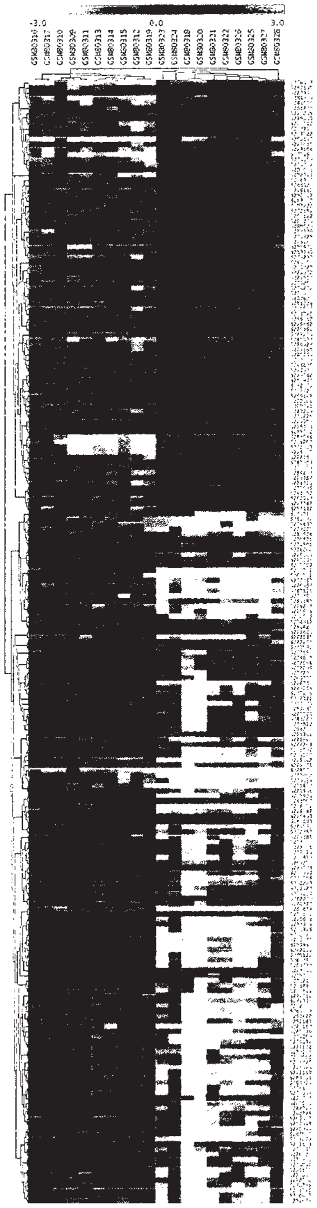


Figure 10.3: A reanalysis of Lam et al. data [181]. By comparing the zebrafish liver tumors to healthy controls, 219 statistically significant genes were selected. Figure was generated on TM4 [182] after a hierarchical clustering on these genes and conditions.

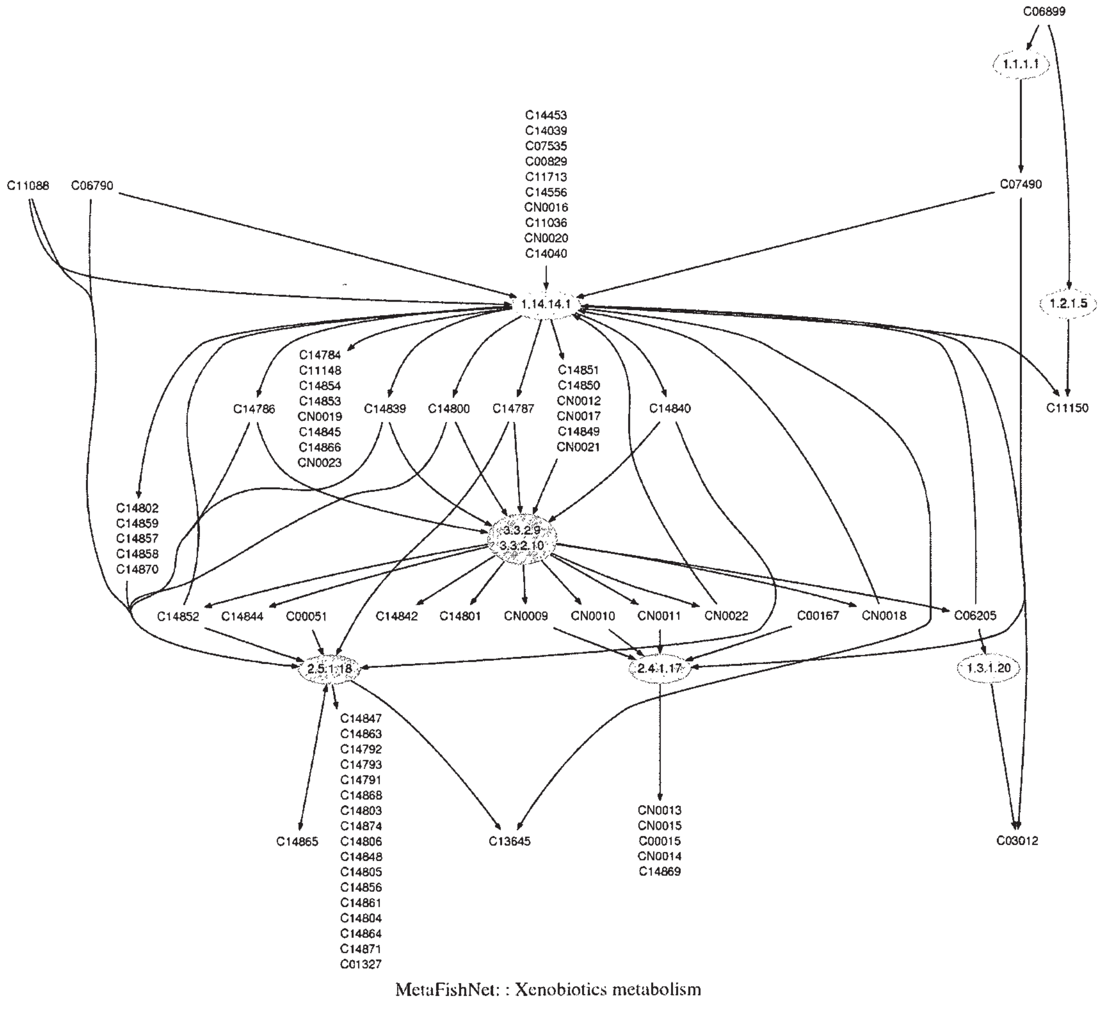


Figure 10.4: Three enzyems in Xenobiotic metabolism are downregulated in zebrafish liver cancer.

Part III

Cyprinodon toxicogenomics

Chapter 11

Toxicogenomic study with Cyprinodon microarrays

Summary

We have designed a *Cyprinodon* DNA microarray based on sequences from our SSH libraries. *Cyprinodon* larvae were exposed to hypoxia, cadmium, chromium and pyrene. RNA samples were extracted from these and control groups, labeled and hybridized to the DNA microarrays. A number of genes were differentially expressed in these groups, including metallothionein and Glutathione S-transferase, hallmark genes of toxic response. We performed hierachical clustering on 22 most statistically significant genes. They show good specificity to different exposures, and can correctly cluster most samples, except that those from chromium and pyrene exposures are too close to distinguish. A classification algorithm, support vector machine, showed more promising power, achieving 100% success in validations by a Jackknife approach. Our MetaFishNet package enabled us to carry out Gene Ontology analysis and metabolic pathway analysis on these data. Among the results, cadmium exposure appears to upregulate "Leukotriene metabolism" and "Xenobiotics metabolism" pathways. These data suggest that our *Cyprinodon* DNA microarrays reveal important response mechanisms to different stressors, and have the potential to become a diagnostic tool in environmental monitoring.

11.1 Study design

Sheepshead minnow (*Cyprinodon variegatus*) is a common, small estuarine fish that is found along the Atlantic and Gulf coasts of United States. They are easy to raise under laboratory conditions, and laboratory research can be extended to field studies. The Environmental Protection Agency has adopted *C. variegatus* as a model organism for studying pollution levels in estuarine waters. We would like to examine their transcriptomic response to common environmental stressors, and identify biomarkers for monitoring aquatic envrionmental quality. Towards this goal, we first generate a good pool of *Cyprinodon* gene sequences.

Since genes are expressed at very different levels, a sequencing project should normalize the expression libraries so that the resulting sequences are not dominated by the most abundant genes. Suppressive Subtractive Hybridization (SSH) is a technique to reduce the copy numbers of common genes shared by two sequence pools. We built SSH libraries from *Cyprinodon* embryos and larvae exposed to pyrene, and chronic and cyclic hypoxia; larvae exposed to cadmium, chromium and copper; and adults exposed to hypoxia. In addition, we prepared normalized and non-normalized cDNA libraries from adult *Cyprinodon* (Pozhitkov et al., in preparation). Combined, over 10,000 cDNAs were sequenced, and assembled into 1053 contigs and 3170 singlets. These sequences were annotated by the *SeaSpider* program from MetaFishNet project (Chapter 8).

A probe design program, ArrayOligoSelector [192], was used to generate an initial pool of microarray probes. Custom scripts were then applied to select probes with minimum secondary structure and reasonable distancing (see Methods). I tried to design four probes per clone. But because the quality of some sequences was not adequate, only 14494 probes could be designed for 4101 clones, averaging 3.5 probes per clone. All these 14494 probes were synthesized on microarray chips by Nimblegen Inc. with four replicates.

Our DNA microarrays were used to interrogate the gene expression of *Cyprinodon* in response to exposures of hypoxia, cadmium, chromium and pyrene. Hypoxia in marine water has become a global environmental concern [193, 194]. Cadmium and chromium are toxic heavy metals. Pyrene is one of the Polycyclic aromatic hydrocarbons (PAHs), a class of anthropogenic pollutants with a wide array of health impacts. RNA samples from these four exposure groups and control group were labeled and hybridized to microarrays. Biological triplicates were used from each experimental group, except four biological replicates were done from the hypoxia exposed group. In total, 16 single-channel hybridizations were performed.

11.2 Methods

Cyprinodon larvae exposures

Exposures and animal sampling were performed similarly as previously described [195, 196]. The study consisted of five exposure groups: seawater control, 1.5mg/L hypoxia, 10mg/L Cr(VI), 0.3mg/L Cd, and 40ppb pyrene. The exposure was conducted in an intermittent flow-through system. Concentrations of chromium, cadmium and pyrene in the exposure system were produced by precision syringe pumps. Hypoxia was maintained at the nominal oxygen concentration of 1.5 mg/L using an AquaController III from Neptune Systems (San Josa, CA). A total of 1200 larvae were used in this study. There were 5 treatments; 3 tanks/treatment; 4 cups/tank; and 20 larvae/cup. Each cup has 20 larvae that were randomly divided at the beginning of the exposure when fish were under 24 hours old. Fish were removed from the aquaria at day 7 from all the exposures. Whole larvae were instantly killed and stored in RNAlater (Ambion Inc., Austin, TX) for molecular analysis.

Cyprinodon DNA microarrays

ArrayOligoSelector was downloaded from `http://arrayoligosel.sourceforge.net`, and installed on a local Linux machine. I used ArrayOligoSelector to design 8 initial 25-mer probes per sequence, where a duplexing energy $\Delta G_d < -28$ kcal/mole was enforced. BLAT [197] was used for checking sequence similarity. The commands to run ArrayOligoSelector were

```
./Pick70_script1_contig UniqUcaseSep5.fas UniqUcaseSep5.fas 25 yes blat
./Pick70_script2 40 25 8 -28
```

where UniqUcaseSep5.fas was the input sequence file. With custom scripts, a further round of selection was made by probe folding energy $\Delta G_o < -3$ kcal/mole. Probes for the same target were distanced for at least 20 bases. All probes were verified against original sequences and identifiers. The resulting microarrays were ordered via NimbleGen Inc.

Sample labeling and microarray hybridization

Total RNAs were extracted from whole sheepshead minnow larvae by phenol/chloroform method. The samples were treated by DNase to remove DNA. The purified RNAs were then quantified by NanoDrop. Their integrity was verified by BioAnalyzer. The labeling of RNAs was carried out according to recommendation by Nimblegen Inc. In short, mRNAs were converted to double-strand cDNA. Cy3-labeled random nonamers were used as primers for DNA polymerase reaction, which produced labeled DNA targets off the double-strand cDNA. These labeled targets were purified and hybridized to microarray in batches of four samples. The fluorescent intensities from all 16 microarrays were normalized by a quantile method implemented in *limma* package [40]. Data at the probe level were averaged over on-slide replicates, with outliers removed. The expression values at the gene level were summarized as the geometric mean of its probe intensities.

Computational analyses

Custom scripts were used to summarize microarray data both at probe level and at gene level. TM4 package [182] was used to execute SAM, perform hierarchical clustering and generate heat maps. MAANOVA was run under R environment, using mixed ANNOVA models. GO enrichment and pathway analysis were performed by MetaFishNet package, as described in Chapter 9. SVM classification was performed with the help of *libsvm* library [198].

11.3 Differentially expressed genes

The first question we asked is how gene expression was affected by these exposures. By comparing the expression profiles of each exposed group to the control group, candidate genes were identified by Significance Analysis of Microarrays (SAM [43]), as listed in Table 11.1. The statistical significance in SAM tends to be very conservative, and this is reflected in the small numbers of genes in Table 11.1. In addition, the data were also analyzed by MAANOVA, a seasoned microarray data analysis program for statistical feature selection [45]. MAANOVA uses a mixed ANOVA model, and calculates the significance based on adjusted F-tests and data permutation. The lists of differentially expressed genes by MAANOVA under exposures of hypoxia, Cadmium, Chromium and pyrene are shown in Appendix Tables 3 to 6, respectively.

It is apparent that our data interpretation is impeded by the large number of unknown genes, as only about one quarter of clones in our *Cyprinodon* library have identifiable homologs. From the few genes with descriptions, we may get a peek into the cellular mechanisms in response to these stressors.

Table 11.1: Top candidate genes affected by exposure to each of the four stressors, identified by SAM with $FDR < 0.05$.

clone id	fold change	gene description
hypoxia exposure		
Contig374	0.87	None
C23_01_X125	1.18	None
cadmium exposure		
C07_04_A02	2.19	Oryzias javanicus metallothionein
Contig319	0.93	None
C15_02_C12	0.85	crystallin, gamma N1
Contig374	0.88	None
chromium exposure		
Contig627	1.77	None
Contig769	1.41	hypothetical protein LOC445282
C19_05_C05	1.38	Mus musculus placental protein 11 related (Pp11r)
Contig738	1.57	Betaine–homocysteine S-methyltransferase 1
C07_05_H08	1.68	Six-cysteine containing astacin protease 3.
pyrene exposure		
Contig627	1.39	None
Contig874	1.29	hypothetical protein LOC447843
Contig328	1.41	None
Contig663	1.33	amylase-3 protein

Hypoxia induced C03_04_B06, a cytochrome c oxidase subunit, and repressed C107_08_F04, a pyruvate dehydrogenase component (Appendix Table 3). The regulation of cytochrome oxidase subunits has been shown to be a key strategy cells use to cope with hypoxia [199]. The inhibition of pyruvate dehydrogenase by hypoxia has also been reported, as a way to redirect pyruvate from TCA cycle [200, 201]. Hemoglobin (Contig793) was also induced, consistent with the previous notion of increased oxygen utilizing efficiency under hypoxia [202–204]. During our microarray design, we specifically added probes for HIF1a. HIF1a is absent in Appendix Table 3. This is also consistent with the general notion that HIF1a is regulated at the protein level not at the mRNA level. The regulation of HIF1a should not be observed on DNA microarrays.

The list for cadmium exposure includes well known toxic response genes (Appendix Table 4). Cytochrome P450/CYP1A (CYP1A1_EF535032, C104_03_D08) are the classical enzymes in Phase I biotransformation, regulated by heavy metals at different levels of the aryl hydrocarbon receptor signaling pathway [205–207]. The Dimethylaniline monooxygenases (C200_01_X86, C24_01_X177) may have a similar role. Glutathione S-transferases (C200_01_X16, C13_03_A05, Contig125, C07_04_C03), or GSTs, are major Phase II enzymes in biotransformation [208]. This induction of GSTs by cadmium has been widely reported, and suggested to be a general response to increased oxidative stress [209]. S-adenosylhomocysteine hydrolase (C104_01_C09) may be involved in Phase III biotransformation.

Metallothioneins (C07_04_A02, Contig549), a major class of proteins to sequester toxic heavy metals, are strongly induced by cadmium. The induction of metallothioneins appear to be specific to cadmium not chromium, consistent with previous reports [210, 211]. Our data also showed metallothioneins were upregulated by hypoxia (Appendix Table 3). This agrees with a recent study in rat kidneys [212], where the authors showed a novel mechanism of metallothioneins in activating the HIF-HRE system through the ERK/mTOR pathway. The upregulation of crystallins by cadmium has been reported in astrocytes [213] and lens epithelial cells [210]. But crystallins appear to be downregulated by cadmium in our data, probably because of different tissue types (whole larvae in our study).

Chromium exposure (Appendix Table 5) also induced several glutathione S-transferases, yet at a weaker magnitude. The roles of Betaine-homocysteine S-methyltransferases (Contig738, Contig994, Contig729, Contig891) and six-cysteine containing astacin protease 3 (C07_05_H08) may be related to Phase III xenobiotic biotransformation, but require further investigation. Cytochrome P450 and Betaine-homocysteine S-methyltransferases also appear in the list for pyrene exposure (Appendix Table 6), suggesting a generic response to toxicity.

Overall, Table 11.1 and Appendix Tables 3 to 6 suggest that our DNA microarrays were capable of detecting specific biomarkers for different stressors, and provide a number of candidate genes, both previously known and unknown, for further investigation.

11.4 Enrichment of GO categories

Besides analysis of individual genes, particular Gene Ontological categories can be tested for their enrichment in differentially expressed genes. Although a few software programs perform GO enrichment analysis, they all rely on the GO annotation of the species under study. In the case of *Cyprinodon*, there is no prior GO annotation available. This makes such analysis rather difficult. Our MetaFishNet package takes on this problem by implementing features of both GO annotation (via *SeaSpider*) and statistical analysis of GO enrichment (see Chapter 9).

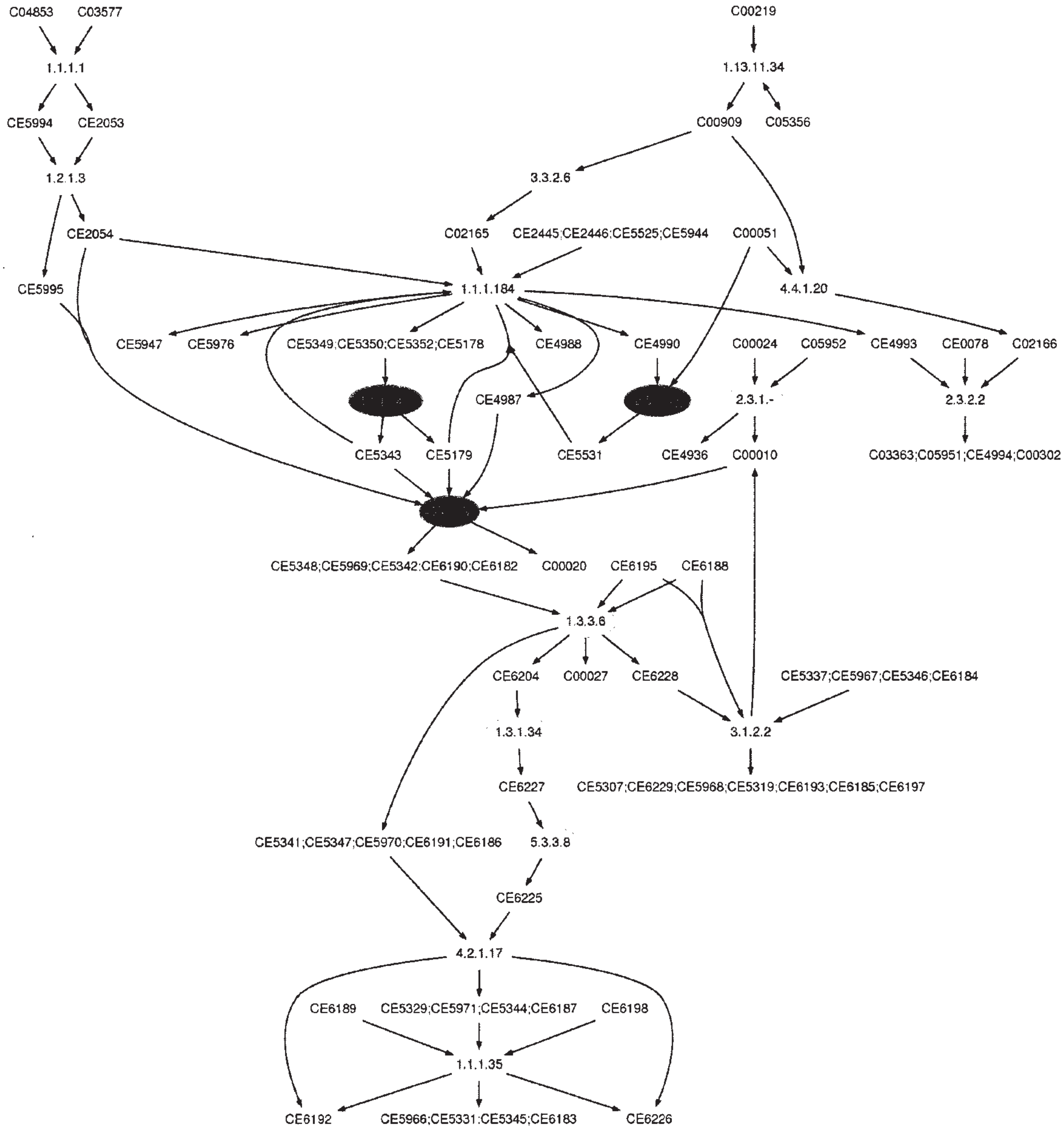
I use the lists of differentially expressed genes ("selected genes") in Appendix Tables 3 to 6 for GO enrichment analysis. From a list of "selected genes", if the number of genes in a particular GO category (term) exceeds the expected value by a significant P-value, this GO category is enriched in this gene list. The GO enrichments for each exposure are shown in Appendix Tables 7 to 10, respectively. In the enriched GO categories for cadmium exposure (Appendix Table 8), monooxygenase activity, reproductive development and xenobiotic response are on the top of list. Consistent with the lists of differential gene expression, GO enrichment analysis revealed that chromium and pyrene exposures significantly induced methyltransferase activity (Appendix Tables 9 and 10).

11.5 Effects on MetaFishNet pathways

Similar to GO enrichment analysis, MetaFishNet can also test the significance of effects on specific metabolic pathways, as already demonstrated in Chapter 10. Using the gene list from Appendix Table 3, hypoxia exposure only produced one hit by clone C107_04_F08 with P-value under 0.05. Since clone C107_04_F08 only showed very moderate fold change, it is likely that no MetaFishNet pathway is significantly affected in the statistics of hypoxia exposure. The results of MetaFishNet pathway analysis for cadmium, chromium and pyrene exposures are shown in Tables 11.2, 11.3 and 11.4, respectively.

From cadmium exposure, all three pathways in Table 11.2 were affected by Cytochrome P450s and Glutathione S-transferases. A third enzyme, C100_01_E04 (EC 6.2.1.3), was also upregulated in Leukotriene metabolism pathway. As this enzyme is not present in Xenobiotics metabolism pathway, the effect on Leukotriene metabolism pathway may suggest an inflammatory stimulus from cadmium exposure. The gene regulation in Leukotriene metabolism pathway is visualized in Figure 11.1.

Similar to what we have seen at gene expression level, chromium and pyrene



MetaFishNet: : Leukotriene metabolism

Figure 11.1: Cadmium exposure upregulates three enzymes (red) in leukotriene pathway.

exposures seem to stimulate common, non-specific responses at the pathway level (Tables 11.3 and 11.4). Oxidative Phosphorylation pathway suggests an effort to increase cellular energy production. The hits on pathway "Endohydrolysis of 1,4-alpha-D-glucosidic linkages in polysaccharides by alpha-amylase" came from alpha-amylase (Contig663). The hit on Purine metabolism pathway came from a hy-

Table 11.2: MetaFishNet pathways affected by cadmium exposure.

MetaFishNet pathway	selected enzymes	enzymes in pathway	P-value
Leukotriene metabolism	3	17	0.012
Xenobiotics metabolism	2	8	0.021
Fatty Acid Metabolism	2	12	0.045

Table 11.3: MetaFishNet pathways affected by chromium exposure.

MetaFishNet pathway	selected enzymes	enzymes in pathway	P-value
Oxidative Phosphorylation	2	2	0.000
Endohydrolysis of 1,4-alpha-D-glucosidic linkages in polysaccharides by alpha-amylase	1	2	0.041

pothetical protein (Contig881). The significance of effects on these pathways is marginal and demands verification.

Besides the data quality from microarrays, it should be pointed out that the power of our pathway analysis is still limited by library size and gene annotations. Among the 4101 clones on our microarrays, only about 350 clones have annotations related to any metabolic process. Nonetheless, MetaFishNet has provided a valuable tool for interpreting microarray data in the context of metabolic pathways.

Table 11.4: MetaFishNet pathways affected by pyrene exposure.

MetaFishNet pathway	selected enzymes	enzymes in pathway	P-value
Purine metabolism	3	53	0.011
Oxidative Phosphorylation	1	2	0.0175
Endohydrolysis of 1,4-alpha-D-glucosidic linkages in polysaccharides by alpha-amylase	1	2	0.017

11.6 Clustering by signature genes

A goal of this study is to identify molecular signatures that are capable of distinguishing different stressors. Previously, we examined the differentially expressed genes between an exposure group and the control group. In order to obtain a set of signature genes that are also distinctive among different treatments, we need to compare all five groups together. This analysis was also performed by MAANOVA. Twenty-two genes were selected with a P-value under 0.01. A hierarchical clustering was then performed on these genes and samples. As shown in Figure 11.2, most of the five experimental groups were clustered correctly. However, the clustering procedure failed to group chromium and pyrene treated samples correctly. This is partly because chromium and pyrene treatments do not generate very specific responses, as we have observed previously. Moreover, the similarity functions in clustering algorithms are ill-suited for classification. Another technique, SVM, is applied in the next section.

11.7 Classification by support vector machines

Support Vector Machines (SVMs) are a popular class of supervised machine learning techniques. They use certain kernel functions to map input data to a higher dimension, and search for an optimal hyperplane to separate classes. SVMs can use various kernel functions to capture a wide range of data types. Since microarray data are by themselves of high dimension, linear kernels usually work well in this field [198]. The feature unit for applying SVM can be either a gene or a probe. Because we have also made low-density microarrays that use a subset of the same probes from our high-density microarrays, I chose to use SVM on our data at probe level, so that the methodological development is relevant to both types of data. This choice may be beneficial to overall sensitivity, as the data summarized

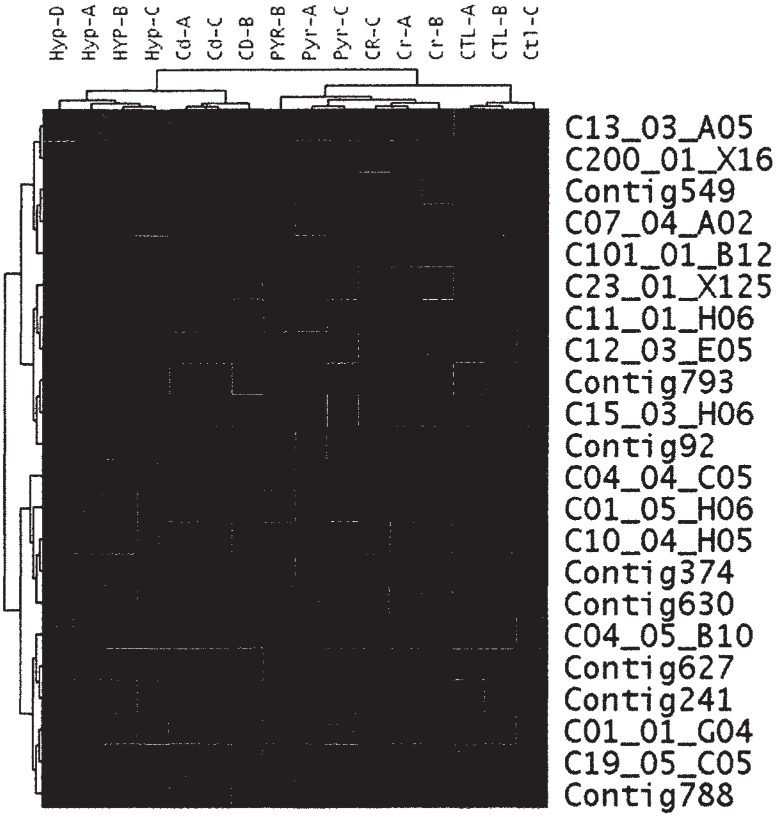


Figure 11.2: Hierachical clustering of statistically significant genes in the Nimblegen microarray experiment. The clustering of samples is shown on top of the heat map. Note that chromium and pyrene treated groups are not clustered correctly. Descriptions of the genes in this figure are given in Appendix Table 11.

at the gene level are a mixed result of both sensitive probes and under-performing probes.

In order to train a reliable classifier, a good sample size is desired. We have 16 microarray samples under 5 conditions. This is still a very small sample size, and not sufficient to do real cross validations. Therefore, I took a Jackknife approach here. That is, one sample is left out of each round of SVM training, and the prediction is then tested on this leftout sample; this procedure is repeated for every sample. The result is shown in Figure 11.3. When I used SVM on all 14494 probes, the success rate of classification was 31%. However, after a feature selection of 91 top probes by MAANOVA, the success rate reached 100%. That is, SVM correctly identified the treatment of all 16 microarray samples based on the signals of these 91 probes. For comparison, random sets of 91 probes produced only the background level of success rates. This result shows that feature selection is critical for increasing the signal/noise ratio.

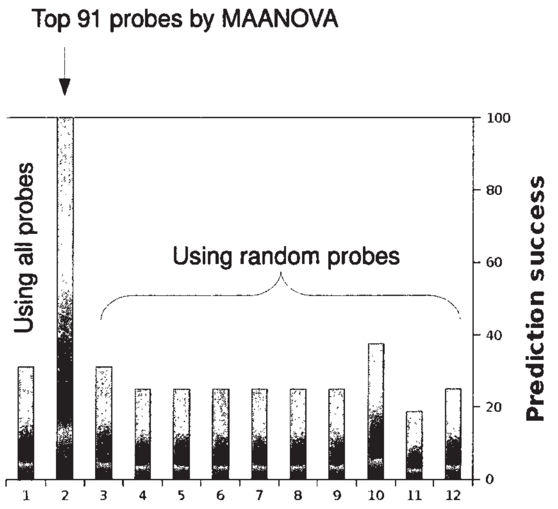


Figure 11.3: Jackknife validation of SVM classification. The success rate of prediction is the percentage of correctly identified samples.

11.8 Conclusion

We have constructed a *Cyprinodon* DNA microarray based on SSH libraries, and used the microarray to study the transcriptomic response of *Cyprinodon* to hypoxia, cadmium, chromium and pyrene. Specific gene expression signatures could be identified for different stressors at varying sensitivity. Our MetaFishNet package helps to perform GO enrichment analysis and metabolic pathway analysis. The result suggests that cadmium exposure affected pathways of Leukotriene metabolism and Xenobiotics metabolism. Finally, we have demonstrated that Support Vector Machines in combination with statistical feature selection are capalbe of classifying different stressors based on their transcriptomic profiles. A robust prediction tool may be developed in the future when a larger sample size and more data become available.

APPENDIX

Table A.1: All pathways in MetaFishNet, version 1.7.

pathway id	pathway name	number of enzymes
mfnpath101	1- and 2-Methylnaphthalene degradation	3
mfnpath102	3-Chloroacrylic acid degradation	2
mfnpath103	3-oxo-10R-octadecatrienoate beta-oxidation	4
mfnpath104	Alanine and Aspartate Metabolism	9
mfnpath105	Alkaloid biosynthesis II	3
mfnpath106	Aminophosphonate metabolism	1
mfnpath107	Aminosugars metabolism	18
mfnpath108	Androgen and estrogen biosynthesis and metabolism	15
mfnpath109	Arachidonic acid metabolism	23
mfnpath110	Arginine and Proline Metabolism	11
mfnpath111	Ascorbate (Vitamin C) and Aldarate Metabolism	1
mfnpath112	Atrazine degradation	1
mfnpath113	Benzoate degradation via CoA ligation	3
mfnpath114	Benzoate degradation via hydroxylation	0
mfnpath115	Bile acid biosynthesis	16
mfnpath116	Biopterin metabolism	6
mfnpath117	Blood Group Biosynthesis	3
mfnpath118	Butanoate metabolism	12
mfnpath119	C21-steroid hormone biosynthesis and metabolism	15
mfnpath120	C5-Branched dibasic acid metabolism	2
mfnpath121	Caprolactam degradation	1
mfnpath122	Carbon fixation	3
mfnpath123	Carnitine shuttle	1
mfnpath124	Chondroitin sulfate degradation	3
mfnpath125	CoA Catabolism	0
mfnpath126	Cyanoamino acid metabolism	1
mfnpath127	D-arg and D-orn metabolism	0
mfnpath128	D4&E4-neuroprostanes formation	0
mfnpath129	De novo fatty acid biosynthesis	10
mfnpath130	Di-unsaturated fatty acid beta-oxidation	6
mfnpath131	Dimethyl-branched-chain fatty acid mitochondrial beta-oxidation	4
mfnpath132	Diterpenoid biosynthesis	0

Continued on next page

pathway id	pathway name	number of enzymes
mfnpath133	Endohydrolysis of 1,4-alpha-D-glucosidic linkages in polysaccharides by alpha-amylase	2
mfnpath134	Fatty Acid Metabolism	12
mfnpath135	Fatty acid activation	2
mfnpath136	Fatty acid oxidation	2
mfnpath137	Fatty acid oxidation, peroxisome	1
mfnpath138	Fructose and mannose metabolism	13
mfnpath139	Galactose metabolism	11
mfnpath140	Geraniol degradation	1
mfnpath141	Glutamate metabolism	12
mfnpath142	Glutathione Metabolism	5
mfnpath143	Glutathione metabolism	4
mfnpath144	Glycerolipid metabolism	1
mfnpath145	Glycerophospholipid metabolism	37
mfnpath146	Glycine, serine and threonine metabolism	1
mfnpath147	Glycine, serine, alanine and threonine metabolism	30
mfnpath148	Glycolysis and Gluconeogenesis	29
mfnpath149	Glycosphingolipid biosynthesis - ganglioseries	9
mfnpath150	Glycosphingolipid biosynthesis - globoseries	6
mfnpath151	Glycosphingolipid biosynthesis - lactoseries	4
mfnpath152	Glycosphingolipid biosynthesis - neolactoseries	9
mfnpath153	Glycosphingolipid metabolism	24
mfnpath154	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	1
mfnpath155	Glyoxylate and Dicarboxylate Metabolism	2
mfnpath156	Heparan sulfate biosynthesis	1
mfnpath157	Heparan sulfate degradation	4
mfnpath158	Histidine metabolism	12
mfnpath159	Hyaluronan Metabolism	3
mfnpath160	Keratan sulfate biosynthesis	3
mfnpath161	Keratan sulfate degradation	7
mfnpath162	Leukotriene metabolism	17
mfnpath163	Limonene and pinene degradation	4
mfnpath164	Linoleate metabolism	9
mfnpath165	Lipoate metabolism	4
mfnpath166	Lysine metabolism	15
mfnpath167	Methionine and cysteine metabolism	24
mfnpath168	Mono-unsaturated fatty acid beta-oxidation	4

Continued on next page

pathway id	pathway name	number of enzymes
mfnpath169	N-Glycan Degradation	7
mfnpath170	N-Glycan biosynthesis	13
mfnpath171	Naphthalene and anthracene degradation	1
mfnpath172	Nitrogen metabolism	3
mfnpath173	Nucleotide Sugar Metabolism	3
mfnpath174	O-Glycan biosynthesis	4
mfnpath175	Omega-3 fatty acid metabolism	7
mfnpath176	Omega-6 fatty acid metabolism	9
mfnpath177	Oxidative Phosphorylation	2
mfnpath178	Pentose and Glucuronate Interconversions	2
mfnpath179	Pentose phosphate pathway	12
mfnpath180	Phosphatidylinositol phosphate metabolism	21
mfnpath181	Phytanic acid peroxisomal oxidation	5
mfnpath182	Polyunsaturated fatty acid biosynthesis	5
mfnpath183	Porphyrin metabolism	11
mfnpath184	Propanoate metabolism	8
mfnpath185	Prostaglandin formation from arachidonate	11
mfnpath186	Prostaglandin formation from dihomo gama-linoleic acid	3
mfnpath187	Proteoglycan biosynthesis	16
mfnpath188	Purine metabolism	53
mfnpath189	Putative anti-Inflammatory metabolites formation from EPA	8
mfnpath190	Pyrimidine metabolism	34
mfnpath191	Pyruvate Metabolism	6
mfnpath192	R Group Synthesis	0
mfnpath193	ROS Detoxification	2
mfnpath194	Riboflavin metabolism	1
mfnpath195	Saturated fatty acids beta-oxidation	12
mfnpath196	Selenoamino acid metabolism	10
mfnpath197	Sphingolipid metabolism	2
mfnpath198	Squalene and cholesterol biosynthesis	16
mfnpath199	Starch and Sucrose Metabolism	9
mfnpath200	Stilbene, coumarine and lignin biosynthesis	2
mfnpath201	Streptomycin biosynthesis	2
mfnpath202	Sulfur metabolism	0
mfnpath203	TCA cycle	15
mfnpath204	Trihydroxycoprostanoyl-CoA beta-oxidation	4
mfnpath205	Tryptophan metabolism	25
mfnpath206	Tyrosine metabolism	39
mfnpath207	Ubiquinone Biosynthesis	0
mfnpath208	Urea cycle/amino group metabolism	11

Continued on next page

pathway id	pathway name	number of enzymes
mfnpath209	Valine, leucine and isoleucine degradation	24
mfnpath210	Vitamin A (retinol) metabolism	6
mfnpath211	Vitamin B1 (thiamin) metabolism	3
mfnpath212	Vitamin B12 (cyanocobalamin) metabolism	1
mfnpath213	Vitamin B12 Metabolism	0
mfnpath214	Vitamin B2 (riboflavin) metabolism	3
mfnpath215	Vitamin B3 (nicotinate and nicotinamide) metabolism	13
mfnpath216	Vitamin B5 - CoA biosynthesis from pantothenate	4
mfnpath217	Vitamin B6 (pyridoxine) metabolism	2
mfnpath218	Vitamin B9 (folate) metabolism	14
mfnpath219	Vitamin D	0
mfnpath220	Vitamin D3 (cholecalciferol) metabolism	2
mfnpath221	Vitamin E metabolism	7
mfnpath222	Vitamin H (biotin) metabolism	2
mfnpath223	Vitamin K metabolism	1
mfnpath224	Xenobiotics metabolism	8
mfnpath225	aspartate and asparagine metabolism	39
mfnpath226	beta-Alanine metabolism	6
mfnpath227	MetaFishNet prov_path 1	3
mfnpath228	MetaFishNet prov_path 10	6
mfnpath229	MetaFishNet prov_path 11	4
mfnpath230	MetaFishNet prov_path 12	3
mfnpath231	MetaFishNet prov_path 13	6
mfnpath232	MetaFishNet prov_path 14	19
mfnpath233	MetaFishNet prov_path 15	4
mfnpath234	MetaFishNet prov_path 16	15
mfnpath235	MetaFishNet prov_path 17	1
mfnpath236	MetaFishNet prov_path 18	9
mfnpath237	MetaFishNet prov_path 19	5
mfnpath238	MetaFishNet prov_path 2	13
mfnpath239	MetaFishNet prov_path 20	18
mfnpath240	MetaFishNet prov_path 21	8
mfnpath241	MetaFishNet prov_path 22	4
mfnpath242	MetaFishNet prov_path 23	9
mfnpath243	MetaFishNet prov_path 24	1
mfnpath244	MetaFishNet prov_path 25	5
mfnpath245	MetaFishNet prov_path 26	12
mfnpath246	MetaFishNet prov_path 27	5
mfnpath247	MetaFishNet prov_path 28	7
mfnpath248	MetaFishNet prov_path 29	2

Continued on next page

pathway id	pathway name	number of enzymes
mfnpath249	MetaFishNet prov_path 3	7
mfnpath250	MetaFishNet prov_path 30	5
mfnpath251	MetaFishNet prov_path 31	8
mfnpath252	MetaFishNet prov_path 32	3
mfnpath253	MetaFishNet prov_path 33	2
mfnpath254	MetaFishNet prov_path 34	3
mfnpath255	MetaFishNet prov_path 37	3
mfnpath256	MetaFishNet prov_path 38	15
mfnpath257	MetaFishNet prov_path 39	7
mfnpath258	MetaFishNet prov_path 4	11
mfnpath259	MetaFishNet prov_path 40	3
mfnpath260	MetaFishNet prov_path 41	3
mfnpath261	MetaFishNet prov_path 42	3
mfnpath262	MetaFishNet prov_path 43	3
mfnpath263	MetaFishNet prov_path 45	4
mfnpath264	MetaFishNet prov_path 46	3
mfnpath265	MetaFishNet prov_path 5	8
mfnpath266	MetaFishNet prov_path 6	4
mfnpath267	MetaFishNet prov_path 7	21
mfnpath268	MetaFishNet prov_path 8	11
mfnpath269	MetaFishNet prov_path 9	18

Table A.2: Tentative functions of newly inferred pathways, assigned by enrichment of GO terms in the pathway.

pathway id	pathway name	Tentative functions
mfnpath144	MetaFishNet prov_path 1	glycolysis; lipopolysaccharide biosynthetic process; proteolysis
mfnpath145	MetaFishNet prov_path 10	glycolysis; glucose metabolic process; protein amino acid ADP-ribosylation
mfnpath146	MetaFishNet prov_path 11	glycolysis; glycogen metabolic process; glycogen catabolic process
mfnpath147	MetaFishNet prov_path 12	mitochondrial electron transport, NADH to ubiquinone; ubiquinone biosynthetic process; ATP synthesis coupled electron transport
mfnpath148	MetaFishNet prov_path 13	tRNA processing; RNA processing; rRNA processing
mfnpath149	MetaFishNet prov_path 14	purine ribonucleoside salvage; nucleoside metabolic process; tRNA modification
mfnpath150	MetaFishNet prov_path 15	glycolysis; nucleotide metabolic process; regulation of glycolysis
mfnpath151	MetaFishNet prov_path 16	regulation of transcription, DNA-dependent; positive regulation of transcription; fatty acid beta-oxidation
mfnpath152	MetaFishNet prov_path 17	proteolysis; ;
mfnpath153	MetaFishNet prov_path 18	fatty acid metabolic process; DNA recombination; DNA repair
mfnpath154	MetaFishNet prov_path 19	retinol metabolic process; phosphoinositide metabolic process; phosphatidylserine metabolic process
mfnpath155	MetaFishNet prov_path 2	protein amino acid phosphorylation; protein amino acid autophosphorylation; glycolysis
mfnpath156	MetaFishNet prov_path 20	mitochondrial electron transport, NADH to ubiquinone; ubiquinone biosynthetic process; fatty acid biosynthetic process
mfnpath157	MetaFishNet prov_path 21	chondroitin sulfate biosynthetic process; nitrogen compound metabolic process; glutamine biosynthetic process
mfnpath158	MetaFishNet prov_path 22	heme biosynthetic process; cellular aldehyde metabolic process; arachidonic acid metabolic process
mfnpath159	MetaFishNet prov_path 23	heme biosynthetic process; regulation of transcription, DNA-dependent; positive regulation of transcription

Continued on next page

pathway id	pathway name	Tentative functions
mfnpath160	MetaFishNet prov_path 24	glucocorticoid biosynthetic process; ;
mfnpath161	MetaFishNet prov_path 25	amino acid metabolic process; L-serine metabolic process; oxidation reduction
mfnpath162	MetaFishNet prov_path 26	dermatan sulfate biosynthetic process; hexose biosynthetic process; chondroitin sulfate biosynthetic process
mfnpath163	MetaFishNet prov_path 27	leukotriene biosynthetic process; iron incorporation into metallo-sulfur cluster; transcription antitermination
mfnpath164	MetaFishNet prov_path 28	ganglioside metabolic process; dermatan sulfate biosynthetic process; galactosylceramide catabolic process
mfnpath165	MetaFishNet prov_path 29	steroid biosynthetic process; ;
mfnpath166	MetaFishNet prov_path 3	purine nucleotide biosynthetic process; glycolysis; gluconeogenesis
mfnpath167	MetaFishNet prov_path 30	proteolysis; retinoic acid metabolic process; retinol metabolic process
mfnpath168	MetaFishNet prov_path 31	steroid biosynthetic process; glucocorticoid biosynthetic process; regulation of pentose-phosphate shunt
mfnpath169	MetaFishNet prov_path 32	androgen biosynthetic process; cholesterol metabolic process; estrogen biosynthetic process
mfnpath170	MetaFishNet prov_path 33	proteolysis; phosphorylation;
mfnpath171	MetaFishNet prov_path 34	aromatic amino acid family metabolic process; serotonin biosynthetic process; proteolysis
mfnpath173	MetaFishNet prov_path 37	heme biosynthetic process; tryptophan catabolic process to kynurenine;
mfnpath174	MetaFishNet prov_path 38	proteolysis; phosphate metabolic process; one-carbon compound metabolic process
mfnpath175	MetaFishNet prov_path 39	cellular aldehyde metabolic process; steroid biosynthetic process; exogenous drug catabolic process
mfnpath176	MetaFishNet prov_path 4	protein amino acid glycosylation; phospholipid biosynthetic process; nucleotide catabolic process
mfnpath177	MetaFishNet prov_path 40	bile acid biosynthetic process; oxidation reduction; androgen metabolic process
mfnpath178	MetaFishNet prov_path 41	proteolysis; nitrogen compound metabolic process; histidine catabolic process

Continued on next page

pathway id	pathway name	Tentative functions
mfnpath179	MetaFishNet prov_path 42	DNA methylation; norepinephrine biosynthetic process; histone H3-K9 methylation
mfnpath180	MetaFishNet prov_path 43	N-acetylglucosamine metabolic process; GPI anchor biosynthetic process; preassembly of GPI anchor in ER membrane
mfnpath182	MetaFishNet prov_path 45	pentose-phosphate shunt; phosphate metabolic process; phosphatidylinositol biosynthetic process
mfnpath183	MetaFishNet prov_path 46	protein amino acid phosphorylation; peptidyl-histidine phosphorylation; protein amino acid dephosphorylation
mfnpath184	MetaFishNet prov_path 5	regulation of transcription, DNA-dependent; transcription antitermination; transcription termination
mfnpath185	MetaFishNet prov_path 6	nucleotide-sugar metabolic process; glycogen metabolic process;
mfnpath186	MetaFishNet prov_path 7	chondroitin sulfate biosynthetic process; hyaluronan biosynthetic process; chondroitin sulfate proteoglycan biosynthetic process, polysaccharide chain biosynthetic process
mfnpath187	MetaFishNet prov_path 8	carbohydrate biosynthetic process; glycosphingolipid biosynthetic process; glycogen biosynthetic process
mfnpath188	MetaFishNet prov_path 9	protein amino acid O-linked glycosylation; chondroitin sulfate biosynthetic process; CTP biosynthetic process

Table A.3: Genes affected by hypoxia exposure, identified with MAANOVA, $P-value < 0.05$. Asterisk (*) marks the genes also identified by SAM with $FDR < 0.05$.

clone id	**fold change**	**P-value**	**gene description**
Contig374*	0.87	0.000	None
C23_01_X125*	1.18	0.000	None
Contig846	0.84	0.002	Zgc:85662 protein (Fragment).
C02_04_G01	0.92	0.003	None
C109_01_F04	1.22	0.003	None
C12_03_E05	1.66	0.004	Collagen alpha-1(X) chain precursor.
C04_04_C05	0.94	0.005	None
Contig1002	0.95	0.005	None
C09_04_D12	0.89	0.006	None
C107_05_D09	0.91	0.006	None
Contig302	1.27	0.007	None
C04_04_A03	0.92	0.008	None
C108_03_E02	0.93	0.008	None
Contig517	0.89	0.009	None
Contig759	0.84	0.010	40S ribosomal protein S18.
Contig763	0.95	0.010	None
C25_01_H03	0.92	0.010	Nematostella vectensis fibroblast growth factor a2 mRNA, complete cds
C03_03_D05	0.73	0.010	Lipocalin-type prostaglandin D synthase-like protein (Ptgds protein).
C105_01_A06	1.67	0.011	Transposable element Tcl transposase.
C07_03_E10	0.92	0.011	None
Contig962	1.45	0.012	None
C25_01_E09	0.94	0.012	hypothetical protein LOC436754
Contig788	1.13	0.012	60S ribosomal protein L29 (Cell surface heparin-binding protein HIP).
C14_04_E07	0.86	0.012	None
C03_01_D07	1.42	0.013	None
C23_01_X129	1.13	0.013	None
Contig328	1.61	0.013	None
C12_03_E09	1.22	0.015	None
C01_01_F08	0.91	0.016	None
Contig1026	1.36	0.016	Cyprinodon rubrofluviatilis mitochondrion, complete genome
C11_01_H06	1.26	0.017	None
Contig630	0.87	0.017	Zgc:114014.
C07_03_B09	0.85	0.018	None
C200_01_X76	0.91	0.021	None

Continued on next page

clone id	fold change	P-value	gene description
Contig213	1.16	0.021	Protein S100-A1 (S100 calcium-binding protein A1) (S-100 protein alpha subunit) (S-100 protein alpha chain).
Contig65	0.85	0.021	None
C02_04_C01	0.88	0.021	F.rubripes Huntington's disease gene homologue
C09_02_H10	1.14	0.022	None
C16_05_A03	1.16	0.022	None
Contig988	1.25	0.023	None
C108_03_A06	1.08	0.024	None
C19_03_H06	0.97	0.024	None
C12_04_B01	0.91	0.024	ISSod3, transposase.
C02_05_B08	1.21	0.025	None
C19_02_E07	1.07	0.025	None
Contig268	1.27	0.025	None
Contig38	1.17	0.025	None
C15_05_C06	1.06	0.025	None
C15_04_C02	0.81	0.025	ribosomal protein S5
Contig61	1.26	0.026	None
C06_01_B04	0.94	0.028	None
Contig697	0.92	0.028	None
C107_07_C02	1.07	0.028	None
C08_01_G12	0.97	0.028	None
C03_05_G05	0.91	0.029	None
C09_04_B10	1.21	0.029	None
C03_04_H04	0.89	0.029	None
C107_01_E10	0.96	0.029	None
C17_03_F09	0.95	0.029	None
Contig815	1.25	0.029	Protein S100-A1 (S100 calcium-binding protein A1) (S-100 protein alpha subunit) (S-100 protein alpha chain).
C11_05_B07	1.18	0.029	Danio rerio selenoprotein 15 (sep15), mRNA
Contig663	1.69	0.030	amylase-3 protein
C107_04_B09	0.93	0.030	None
C107_04_G11	0.89	0.030	Protein SDA1 homolog (SDA1 domain-containing protein 1).
C04_02_F05	0.94	0.030	None
C19_05_B11	1.21	0.030	None
C02_04_D04	0.86	0.030	Lateolabrax japonicus ribosomal protein L8 (L8) mRNA, complete cds

Continued on next page

clone id	fold change	P-value	gene description
C08_02_G06	0.96	0.030	None
C04_04_C08	0.95	0.031	None
C04_05_B06	0.93	0.031	None
C07_04_A02	1.23	0.032	Oryzias javanicus metallothionein (MT) mRNA, complete cds
Contig19	1.43	0.032	RNA-binding protein 5 (RNA-binding motif protein 5) (Tumor suppressor LUCA15).
C14_03_G07	0.75	0.032	Fundulus heteroclitus FLCE mRNA for hatching enzyme, complete cds
C03_01_E05	1.20	0.032	None
C03_02_E03	1.21	0.033	hypothetical protein LOC327336
C06_05_B09	0.86	0.033	None
C107_04_F08	0.97	0.033	Alpha-(1,3)-fucosyltransferase (EC 2.4.1.-) (Galactoside 3-L- fucosyltransferase) (Fucosyltransferase 7) (FUCT-VII).
C07_05_D12	1.15	0.034	Solute carrier family 25 member 39.
C23_01_X117	1.16	0.034	None
C19_04_F08	0.95	0.035	None
Contig969	1.25	0.035	None
C108_02_H06	0.93	0.035	B9 protein-like [Source:RefSeq_peptide;Acc:NP_001019544]
C107_02_A02	0.97	0.035	None
C09_03_D04	0.83	0.035	None
C01_05_H06	0.96	0.036	None
C11_04_F08	0.76	0.036	Cation transport regulator-like protein 1.
C10_04_G08	0.84	0.036	None
C108_01_E02	0.89	0.037	None
Contig674	0.95	0.037	None
C09_03_F03	1.12	0.038	None
C03_02_D02	0.81	0.038	None
C10_04_H05	0.91	0.039	None
C04_05_E12	0.76	0.039	None
C107_08_F04	0.87	0.039	Pyruvate dehydrogenase protein X component, mitochondrial precursor (Dihydrolipoamide dehydrogenase-binding protein of pyruvate dehydrogenase complex) (Lipoyl-containing pyruvate dehydrogenase complex component X).
C03_02_A04	0.95	0.039	None
C20_02_A06	1.05	0.039	None
C19_02_B01	1.35	0.040	None

Continued on next page

clone id	fold change	P-value	gene description
Contig319	0.96	0.040	None
C01_03_G04	0.91	0.041	None
C10_05_C10	0.93	0.041	hypothetical protein LOC550499
Contig705	0.68	0.041	None
C04_02_B12	1.15	0.041	hypothetical protein LOC393431
C11_02_F08	0.92	0.041	Zgc:136971.
C108_01_E03	0.79	0.041	None
C102_01_H07	0.95	0.042	None
C01_04_E02	1.07	0.042	None
C03_05_D09	1.39	0.042	None
C15_03_H06	1.25	0.042	None
C15_03_H12	0.86	0.042	Oryzias latipes hox gene cluster, complete cds, contains hoxCa
C04_04_B05	0.92	0.042	Paraneoplastic antigen Ma1 (Neuron- and testis-specific protein 1) (37 kDa neuronal protein).
Contig860	1.15	0.043	Proto galectin Gal1-L2.
C09_02_C08	0.98	0.043	Zgc:110116 protein
C01_01_E12	0.96	0.043	Takifugu rubripes DNA, highly conserved vertebrate non-coding sequence, element ID CNE870
C15_03_E10	1.11	0.043	MGC162578 protein.
C107_07_C04	0.92	0.043	None
C101_01_F11	0.91	0.043	None
C04_04_D12	0.93	0.043	None
C13_05_F08	0.71	0.043	None
C12_03_H12	1.11	0.043	None
Contig769	1.30	0.044	hypothetical protein LOC445282
Contig121	0.96	0.046	None
C04_01_F08	0.94	0.046	None
C11_01_B08	0.92	0.046	hypothetical protein LOC436863
C01_04_D10	0.85	0.046	None
C07_01_D10	1.20	0.046	Complement C3 precursor (HSE-MSF) [Contains: Complement C3 beta chain; Complement C3 alpha chain; C3a anaphylatoxin; Complement C3b alpha' chain; Complement C3c alpha' chain fragment 1; Complement C3dg fragment; Complement C3g fragment; Complement C3d fra
C03_01_D06	0.92	0.046	None
Contig41	1.40	0.046	hypothetical protein LOC406338

Continued on next page

clone id	**fold change**	**P-value**	**gene description**
C107_05_C10	0.96	0.047	None
Contig595	1.18	0.047	None
C04_05_C06	0.97	0.047	None
C11_01_G10	0.48	0.047	None
Contig793	1.23	0.048	alpha globin
C03_04_B06	1.12	0.048	cytochrome c oxidase, subunit VIIc
C12_04_A12	1.14	0.048	None
C12_02_H06	0.72	0.050	hypothetical protein LOC541504

Table A.4: Genes affected by cadmium exposure, identified with MAANOVA, P-value less than 0.05. Asterisk (*) marks the genes also identified with SAM, FDR less than 0.05.

clone id	fold change	P-value	gene description
C07_04_A02*	2.19	0.000	Oryzias javanicus metallothionein (MT) mRNA, complete cds
Contig319*	0.93	0.000	None
C15_02_C12*	0.85	0.000	crystallin, gamma N1
Contig374*	0.88	0.001	None
Contig549	1.58	0.001	Metallothionein-2 (MT-2).
C103_01_C12	1.24	0.001	Cyp3a65 protein (Fragment).
C200_01_X16	1.43	0.001	Glutathione S-transferase Mu 5
Contig39	1.17	0.003	Zgc:153093.
Contig328	1.38	0.003	None
Contig253	1.15	0.005	None
C06_04_F05	1.10	0.006	retinol binding protein 2a, cellular
C19_05_C07	0.94	0.006	None
C13_03_A05	1.33	0.006	ATGSTF3 (GLUTATHIONE S-TRANSFERASE 16); glutathione transferase
Contig967	0.88	0.006	crystallin, gamma M2d3
C100_01_C01	1.15	0.007	Murinoglobulin-1 precursor (MuG1).
Contig927	1.29	0.009	Zgc:113828.
C109_01_D01	0.97	0.010	small nuclear ribonucleoprotein D2
C200_01_X55	1.48	0.012	None
C100_01_G11	1.52	0.012	3-hydroxy-3-methylglutaryl-coenzyme A reductase (EC 1.1.1.34) (HMG-CoA reductase).
Contig697	0.89	0.013	None
C22_01_X67	0.92	0.013	None
C200_01_X76	0.91	0.013	None
Contig576	1.46	0.013	Corticosteroid-binding globulin precursor (CBG) (Transcortin) (Serpin A6).
Contig454	1.17	0.014	fibrinogen, B beta polypeptide
C11_02_E10	1.07	0.015	None
C100_01_B01	1.48	0.015	None
C11_01_E09	1.67	0.015	None
C09_04_G12	1.19	0.016	procollagen C-endopeptidase enhancer
C04_05_B06	0.92	0.016	None
C103_01_F12	1.16	0.016	Zgc:153219.
Contig214	0.87	0.017	None
C107_04_B09	0.92	0.017	None

Continued on next page

clone id	fold change	P-value	gene description
C19_02_B01	1.50	0.018	None
C04_04_C05	0.95	0.019	None
C100_01_F05	1.24	0.019	LOC791587 protein.
C07_03_E10	0.93	0.020	None
Contig314	1.34	0.021	None
Contig759	0.84	0.021	40S ribosomal protein S18.
Contig325	1.13	0.021	None
C09_04_B10	1.12	0.021	None
C200_01_X4	1.08	0.023	Zgc:77882 (Novel protein).
C04_05_C12	0.94	0.023	None
C14_04_E07	0.89	0.023	None
C107_01_A01	1.09	0.023	None
C04_02_F05	0.94	0.024	None
Contig338	0.95	0.024	None
C105_01_A06	1.60	0.025	Transposable element Tc1 transposase.
C09_01_E06	1.03	0.025	None
Contig125	1.14	0.025	Platichthys flesus partial mRNA for microsomal glutathione S-transferase (gst gene)
C20_01_G03	0.94	0.027	None
C12_04_B01	0.91	0.027	ISSod3, transposase.
C10_01_G08	0.92	0.028	None
C09_03_D04	0.80	0.028	None
Contig43	1.20	0.028	Beta-2-glycoprotein 1 precursor (Beta-2-glycoprotein I) (Apolipoprotein H) (Apo-H) (B2GPI) (Beta(2)GPI) (Activated protein C- binding protein) (APC inhibitor) (Anticardiolipin cofactor).
CYP1A1_EF535032	1.21	0.028	Cytochrome P450 1A1 (Cytochrome P450 1A).
C200_01_X86	1.25	0.028	Dimethylaniline monooxygenase [N-oxide-forming] 5(FMO 5) (Dimethylaniline oxidase 5).
C102_01_E06	1.19	0.029	3-hydroxy-3-methylglutaryl-coenzyme A reductase (EC 1.1.1.34) (HMG-CoA reductase).
C24_01_X177	1.34	0.029	Dimethylaniline monooxygenase [N-oxide-forming] 5 (FMO 5) (Dimethylaniline oxidase 5).
C107_07_C04	0.92	0.030	None
C16_04_B02	0.89	0.030	None
C27_01_H04	1.29	0.030	Selenium-binding protein 1.

Continued on next page

clone id	fold change	P-value	gene description
C07_04_C03	1.29	0.031	Kryptolebias marmoratus glutathione S-transferase theta-class (GST-T) mRNA, complete cds
Contig92	0.91	0.031	Protein S100-A1 (S100 calcium-binding protein A1) (S-100 protein alpha subunit) (S-100 protein alpha chain).
C12_04_D08	1.06	0.031	None
C104_01_C09	1.54	0.031	S-adenosylhomocysteine hydrolase
C20_02_C04	1.21	0.031	None
C02_05_H04	0.94	0.031	None
C08_02_H04	0.95	0.032	None
C108_03_E02	0.92	0.033	None
C02_02_B11	0.95	0.033	None
C100_01_E04	1.16	0.033	zgc:158482 (zgc:158482), mRNA
Contig657	1.07	0.033	guanine nucleotide binding protein (G protein), gamma transducing activity polypeptide 1
C200_01_X47	1.10	0.034	Sulfhydryl oxidase 1 precursor (EC 1.8.3.2) (Quiescin Q6).
Contig128	0.97	0.035	None
C09_01_G11	1.05	0.036	None
C109_01_F11	0.96	0.037	Trematomus bernacchii clone 47 immunoglobulin light chain isotype L2 mRNA, partial cds
C03_01_D07	1.39	0.037	None
C12_02_E10	0.96	0.037	None
C10_04_H05	0.89	0.038	None
C04_05_G04	1.03	0.038	None
Contig828	1.21	0.039	Fetuin-B precursor (Gugu) (IRL685) (16G2).
C20_02_C10	1.12	0.039	NMD3 homolog (S. cerevisiae), like
Contig663	1.24	0.039	amylase-3 protein
C100_01_D03	1.09	0.039	Ovostatin precursor (Ovomacroglobulin).
C102_02_C09	0.95	0.039	None
C107_07_C02	1.07	0.039	None
C11_01_B08	0.89	0.039	hypothetical protein LOC436863
C03_03_E06	0.92	0.040	None
C08_04_C04	0.95	0.040	None
C02_04_G01	0.91	0.040	None
C02_04_F02	0.94	0.041	None
Contig61	1.29	0.041	None

Continued on next page

clone id	fold change	P-value	gene description
C02_01_D11	0.96	0.041	None
C107_01_B08	0.91	0.041	None
Contig130	1.22	0.041	Zgc:103654.
Contig734	1.36	0.041	ribosomal protein, large P2
C108_01_C08	1.14	0.041	Collagen alpha-1(IV) chain precursor (Arresten).
C104_03_D08	1.18	0.041	ccytochrome P450, family 2, subfamily J, polypeptide 28
C04_04_D12	0.93	0.041	None
Contig279	0.93	0.041	hypothetical protein LOC550395
C15_01_A07	1.20	0.041	retinol binding protein 4, like
C03_05_G05	0.89	0.042	None
Contig625	1.33	0.042	None
Contig346	0.84	0.043	NADH dehydrogenase (ubiquinone) Fe-S protein 6
Contig906	1.20	0.043	Takifugu rubripes apolipoprotein E1 (apoe1), mRNA
C23_01_X135	1.28	0.043	Complement C3 precursor (C3 and PZP-like alpha-2-macroglobulin domain- containing protein 1)
Contig517	0.88	0.044	None
C108_02_B01	0.96	0.044	Gamma-secretase subunit Psenen (Presenilin enhancer protein 2 homolog).
C03_05_B08	1.14	0.045	proteasome (prosome, macropain) subunit, alpha type, 4
Contig997	1.06	0.046	40S ribosomal protein S7.
Contig630	0.90	0.046	Zgc:114014.
Contig738	1.31	0.046	Betaine–homocysteine S-methyltransferase 1
C09_05_G07	0.96	0.046	None
C19_01_B11	0.88	0.047	None
Contig643	1.28	0.048	None
C11_05_C08	1.15	0.049	Recoverin.
C19_04_F08	0.94	0.049	None
C15_03_E10	1.10	0.050	MGC162578 protein.

Table A.5: Genes affected by Chromium exposure, identified with MAANOVA, P-value less than 0.05. Asterisk (*) marks the genes also identified with SAM, FDR less than 0.05.

clone id	fold change	P-value	gene description
Contig627*	1.77	0.000	None
Contig769*	1.41	0.000	hypothetical protein LOC445282
C19_05_C05*	1.38	0.001	Mus musculus placental protein 11 related (Pp11r), mRNA
Contig738*	1.57	0.001	Betaine–homocysteine S-methyltransferase 1
C07_05_H08*	1.68	0.001	Six-cysteine containing astacin protease 3.
Contig663	1.69	0.002	amylase-3 protein
Contig994	1.20	0.005	Betaine–homocysteine S-methyltransferase 1
C03_05_C04	0.89	0.007	None
C20_03_B01	0.80	0.007	ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1
C108_01_H02	0.93	0.009	Zgc:73066.
C03_04_D01	1.12	0.011	None
Contig125	1.17	0.012	Platichthys flesus partial mRNA for microsomal glutathione S-transferase (gst gene)
Contig915	1.31	0.012	hypothetical protein LOC445282
Contig788	1.18	0.013	60S ribosomal protein L29 (Cell surface heparin-binding protein HIP).
C103_01_F12	1.14	0.013	Zgc:153219.
Contig874	1.55	0.013	hypothetical protein LOC447843
Contig464	1.26	0.013	None
Contig338	0.95	0.015	None
C108_01_G01	0.95	0.015	None
Contig778	1.35	0.016	Carboxyl ester lipase, like.
Contig729	1.55	0.017	Betaine–homocysteine S-methyltransferase 1
C10_01_H05	0.86	0.017	hypothetical protein LOC415253
Contig279	0.92	0.018	hypothetical protein LOC550395
Contig579	0.86	0.020	Alcohol dehydrogenase.
Contig391	1.31	0.020	hypothetical protein LOC641484
C11_01_B11	1.18	0.021	None
Contig891	1.39	0.023	Betaine–homocysteine S-methyltransferase 1
Contig517	0.88	0.023	None
C04_05_B06	0.90	0.023	None
C03_01_D07	1.44	0.024	None

Continued on next page

clone id	fold change	P-value	gene description
C25_01_H03	0.92	0.025	Nematostella vectensis fibroblast growth factor a2 mRNA, complete cds
Contig342	1.18	0.026	None
C19_05_B11	1.14	0.026	None
C19_05_C01	0.88	0.027	None
C107_07_D01	0.93	0.027	Cytochrome c oxidase subunit 1
Contig429	0.90	0.028	Perca flavescens calmodulin (CAL1) mRNA, complete cds
C200_01_X55	1.23	0.028	None
C11_05_D08	1.30	0.029	None
Contig595	1.17	0.029	None
C107_01_B08	0.90	0.029	None
Contig855	1.07	0.029	Kryptolebias marmoratus glutathione S-transferase theta-class (GST-T) mRNA
C14_03_A03	0.81	0.030	Heat shock cognate 71 kDa protein
Contig920	1.54	0.030	chymotrypsin C (caldecrin)
C03_02_B07	1.06	0.031	None
C13_03_A05	1.14	0.032	ATGSTF3 (GLUTATHIONE S-TRANSFERASE 16); glutathione transferase
C108_02_H05	1.28	0.032	None
C107_01_E09	1.16	0.034	None
C14_05_D12	0.79	0.035	None
Contig241	1.15	0.036	None
C01_02_F10	1.07	0.036	None
C12_05_E09	1.15	0.036	Zebrafish DNA sequence from clone DKEY-18F7 in linkage group 19 Contains the 5' end of the gene for a novel protein similar to vertebrate FK506 binding protein 10, 65 kDa (FKBP10), the gene for a novel protein similar to vertebrate cartilage associated protein (CRTAP), two novel genes and two CpG islands, complete sequence
Contig105	0.85	0.037	ribosomal protein L37
C19_05_H01	0.93	0.038	Zgc:136220 protein
C14_02_E05	1.07	0.038	None
C11_02_E10	1.12	0.038	None
C14_04_B08	1.13	0.039	None
C08_01_C10	1.23	0.039	None
C02_05_H04	0.93	0.039	None

Continued on next page

clone id	fold change	P-value	gene description
Contig347	1.17	0.040	Paralichthys olivaceus CPA1 mRNA for carboxypeptidase A1, partial cds
Contig764	0.86	0.040	None
Contig846	0.90	0.040	Zgc:85662 protein (Fragment).
C14_04_C12	0.84	0.041	None
C12_03_C04	0.87	0.042	nucleosome assembly protein 1, like 1
C02_03_E06	1.27	0.042	None
Contig793	0.77	0.042	alpha globin
Contig841	1.06	0.043	Diablo homolog, mitochondrial precursor (Second mitochondria-derived activator of caspase) (Smac protein) (Direct IAP-binding protein with low pI).
C04_02_E03	1.09	0.043	None
C11_01_E09	1.40	0.043	None
C107_08_D07	0.93	0.044	None
C200_01_X11	1.19	0.045	None
Contig643	1.30	0.045	None
C14_03_G07	0.76	0.045	Fundulus heteroclitus FLCE mRNA for hatching enzyme
Contig137	0.88	0.046	hypothetical protein LOC799058
C09_03_D04	0.82	0.046	None
C10_04_G08	0.83	0.047	None
Contig630	0.87	0.047	Zgc:114014.
C14_03_B07	0.94	0.047	None
C16_02_F03	1.08	0.047	None
C04_03_H06	1.19	0.048	ATPase, H+ transporting, lysosomal, V0 subunit c
C107_04_H02	1.11	0.048	None
C14_03_G06	1.08	0.048	None
Contig783	1.24	0.049	None
C15_02_C12	0.88	0.049	crystallin, gamma N1
Contig39	1.18	0.050	Zgc:153093.

Table A.6: Genes affected by pyrene exposure, identified with MAANOVA, P-value less than 0.05. Asterisk (*) marks the genes also identified with SAM, FDR less than 0.05.

clone id	fold change	P-value	gene description
Contig627*	1.39	0.003	None
Contig874*	1.29	0.008	hypothetical protein LOC447843
Contig773	1.09	0.008	Fatty acid-binding protein, intestinal (I-FABP) (FABPI).
Contig328*	1.41	0.009	None
C23_01_X125	1.11	0.010	None
C03_02_E08	1.06	0.011	None
C04_02_E03	1.05	0.012	None
Contig663*	1.33	0.012	amylase-3 protein
C12_03_E10	0.95	0.018	None
Contig374	0.89	0.024	None
C10_04_H05	0.86	0.024	None
Contig39	1.17	0.025	Zgc:153093.
Contig253	1.09	0.027	None
C107_08_E03	1.15	0.028	Claudin-15.
C100_01_E09	1.11	0.030	None
Contig881	0.86	0.030	hypothetical protein LOC554135
Contig840	1.10	0.031	Ela2 protein (Novel protein similar to elastase 2)
C02_04_G01	0.91	0.032	None
Contig738	1.38	0.032	Betaine–homocysteine S-methyltransferase 1 (EC 2.1.1.5).
Contig729	1.27	0.033	Betaine–homocysteine S-methyltransferase 1 (EC 2.1.1.5).
C103_01_F12	1.11	0.033	Zgc:153219.
C12_03_H12	1.21	0.034	None
Contig630	0.88	0.035	Zgc:114014.
C19_05_G10	1.34	0.035	None
Contig454	1.14	0.037	fibrinogen, B beta polypeptide
Contig805	1.22	0.042	None
C14_01_A01	0.96	0.042	None
C08_03_C06	1.97	0.043	None
C15_04_C02	0.82	0.045	ribosomal protein S5
C107_06_C09	1.22	0.045	None
CYP1A1_EF535032	1.25	0.047	Cytochrome P450 1A1 (Cytochrome P450 1A).
C107_04_G11	0.88	0.047	Protein SDA1 homolog (SDA1 domain-containing protein 1).

Continued on next page

clone id	**fold change**	**P-value**	**gene description**
C08_01_C10	1.09	0.048	None
C19_05_H02	0.95	0.049	None
Contig1019	0.85	0.049	troponin C, fast skeletal
C107_06_D11	1.09	0.050	None

Table A.7: GO categories affected by hypoxia exposure, P-value less than 0.05.

GO term	selected genes	genes in category	P-value
polysaccharide binding	2	2	0.001
pattern binding	2	2	0.001
carbohydrate binding	3	7	0.001
transposase activity	2	3	0.003
DNA recombination	2	3	0.003
sarcoplasm	2	3	0.003
sarcoplasmic reticulum	2	3	0.003
transposition	2	3	0.003
transposition, DNA-mediated	2	3	0.003
nucleolus	2	4	0.006
membrane-enclosed lumen	3	14	0.010
organelle lumen	3	14	0.010
nuclear lumen	2	8	0.026
hydrolase activity, acting on glycosyl bonds	2	8	0.026
hydrolase activity, hydrolyzing O-glycosyl compounds	2	8	0.026
amino sugar catabolic process	1	1	0.033
mitochondrial pyruvate dehydrogenase complex	1	1	0.033
pyruvate dehydrogenase complex	1	1	0.033
amylase activity	1	1	0.033
chitin metabolic process	1	1	0.033
laminin binding	1	1	0.033
glucosamine metabolic process	1	1	0.033
glucosamine catabolic process	1	1	0.033
N-acetylglucosamine metabolic process	1	1	0.033
N-acetylglucosamine catabolic process	1	1	0.033
amino sugar metabolic process	1	1	0.033
alpha-amylase activity	1	1	0.033
chitin binding	1	1	0.033
prostaglandin-D synthase activity	1	1	0.033
glycosaminoglycan binding	1	1	0.033
monosaccharide binding	1	1	0.033
chitin catabolic process	1	1	0.033
heparin binding	1	1	0.033
female pregnancy	1	1	0.033

Continued on next page

GO term	**selected genes**	**genes in category**	**P-value**
embryo implantation	1	1	0.033
galactoside binding	1	1	0.033
reproduction	2	9	0.033
DNA metabolic process	2	9	0.033
extracellular region	5	56	0.036
identical protein binding	2	10	0.041

Table A.8: GO categories affected by cadmium exposure, P-value less than 0.05.

GO term	selected genes	genes in category	P-value
biological process	41	860	0.000
monooxygenase activity	3	7	0.001
flavin-containing monooxygenase activity	2	2	0.001
development of primary sexual characteristics	2	2	0.001
reproductive developmental process	2	2	0.001
gonad development	2	2	0.001
sex differentiation	2	2	0.001
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NADH or NADPH as one donor, and incorporation of one atom of oxygen	2	2	0.001
reproductive structure development	2	2	0.001
germ cell migration	2	2	0.001
membrane fraction	4	17	0.001
molecular function	40	912	0.002
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	3	10	0.002
response to xenobiotic stimulus	2	3	0.002
transposition	2	3	0.002
DNA recombination	2	3	0.002
transposase activity	2	3	0.002
transposition, DNA-mediated	2	3	0.002
cell fraction	4	24	0.004
sexual reproduction	2	4	0.005
gamete generation	2	4	0.005
extracellular region	6	56	0.005
endoplasmic reticulum membrane	2	5	0.008
glutathione transferase activity	2	5	0.008
endoplasmic reticulum part	2	5	0.008
nuclear envelope-endoplasmic reticulum network	2	5	0.008
transferase activity, transferring alkyl or aryl (other than methyl) groups	2	5	0.008
response to chemical stimulus	4	29	0.009

Continued on next page

GO term	selected genes	genes in category	P-value
microsome	2	6	0.011
vesicular fraction	2	6	0.011
reproductive process	2	6	0.011
endoplasmic reticulum	3	20	0.019
cellular component	35	873	0.019
reproduction	2	9	0.026
receptor binding	2	9	0.026
DNA metabolic process	2	9	0.026
adenosylhomocysteinase activity	1	1	0.029
ATPase binding	1	1	0.029
amylase activity	1	1	0.029
selenium binding	1	1	0.029
alpha-amylase activity	1	1	0.029
protein thiol-disulfide exchange	1	1	0.029
thiol oxidase activity	1	1	0.029
hydrolase activity, acting on ether bonds	1	1	0.029
trialkylsulfonium hydrolase activity	1	1	0.029
S100 beta binding	1	1	0.029
S100 alpha binding	1	1	0.029
one-carbon compound metabolic process	1	1	0.029
toxin metabolic process	1	1	0.029
oxidoreductase activity, acting on sulfur group of donors, oxygen as acceptor	1	1	0.029
toxin catabolic process	1	1	0.029
flavin-linked sulfhydryl oxidase activity	1	1	0.029
wide-spectrum protease inhibitor activity	1	1	0.029
negative regulation of apoptosis	1	1	0.029
regulation of neuron apoptosis	1	1	0.029
neuron apoptosis	1	1	0.029
negative regulation of neuron apoptosis	1	1	0.029
signal transduction	4	43	0.034
cell migration	2	11	0.038
cell communication	4	45	0.039
biopolymer metabolic process	7	114	0.045

Continued on next page

GO term	selected genes	genes in category	P-value
endomembrane system	2	12	0.045

Table A.9: GO categories affected by chromium exposure, P-value less than 0.05.

GO term	selected genes	genes in category	P-value
S-methyltransferase activity	4	4	0.000
homocysteine S-methyltransferase activity	4	4	0.000
betaine-homocysteine S-methyltransferase activity	4	4	0.000
S-adenosylmethionine-dependent methyltransferase activity	4	6	0.000
transferase activity, transferring one-carbon groups	4	8	0.000
methyltransferase activity	4	8	0.000
serine-type endopeptidase activity	4	14	0.000
zinc ion transport	4	14	0.000
catalytic activity	16	284	0.000
serine hydrolase activity	4	15	0.000
serine-type peptidase activity	4	15	0.000
zinc ion binding	4	22	0.001
protease inhibitor activity	4	24	0.001
endopeptidase inhibitor activity	4	24	0.001
transition metal ion transport	5	44	0.002
enzyme inhibitor activity	4	27	0.002
endopeptidase activity	4	29	0.003
proteolysis	4	33	0.004
transferase activity	5	54	0.005
transition metal ion binding	5	54	0.005
peptidase activity	4	34	0.005
enzyme regulator activity	4	37	0.006
hydrogen ion transmembrane transporter activity	4	38	0.007
monovalent inorganic cation transmembrane transporter activity	4	40	0.009
inorganic cation transmembrane transporter activity	4	43	0.011
hydrolase activity	6	101	0.017
biological process	26	860	0.020
regulation of caspase activity	1	1	0.020
amylase activity	1	1	0.020
alpha-amylase activity	1	1	0.020
regulation of peptidase activity	1	1	0.020

Continued on next page

GO term	selected genes	genes in category	P-value
monohydric alcohol metabolic process	1	1	0.020
glycosaminoglycan binding	1	1	0.020
induction of apoptosis by intracellular signals	1	1	0.020
toxin metabolic process	1	1	0.020
positive regulation of caspase activity	1	1	0.020
induction of apoptosis by oxidative stress	1	1	0.020
heparin binding	1	1	0.020
female pregnancy	1	1	0.020
embryo implantation	1	1	0.020
toxin catabolic process	1	1	0.020
caspase activation	1	1	0.020
regulation of endopeptidase activity	1	1	0.020
alcohol dehydrogenase activity	1	1	0.020
ethanol metabolic process	1	1	0.020
electron transport	3	29	0.021
molecular function	27	912	0.022
cation transport	6	108	0.022
cellular component	26	873	0.023
heme-copper terminal oxidase activity	2	13	0.028
cytochrome-c oxidase activity	2	13	0.028
oxidoreductase activity, acting on heme group of donors	2	13	0.028
oxidoreductase activity, acting on heme group of donors, oxygen as acceptor	2	13	0.028
cation binding	6	114	0.028
ion transport	6	114	0.028
metal ion transport	5	89	0.035
metal ion binding	6	120	0.035
ion binding	6	121	0.036
chromatin assembly or disassembly	1	2	0.041
polysaccharide binding	1	2	0.041
induction of programmed cell death	1	2	0.041
apoptotic program	1	2	0.041
chromatin assembly	1	2	0.041
nucleosome assembly	1	2	0.041
DNA packaging	1	2	0.041
induction of apoptosis	1	2	0.041

Continued on next page

GO term	**selected genes**	**genes in category**	**P-value**
pattern binding	1	2	0.041

Table A.10: GO categories affected by pyrene exposure, P-value less than 0.05.

GO term	selected genes	genes in category	P-value
S-methyltransferase activity	2	4	0.000
homocysteine S-methyltransferase activity	2	4	0.000
betaine-homocysteine S-methyltransferase activity	2	4	0.000
S-adenosylmethionine-dependent methyltransferase activity	2	6	0.001
transferase activity, transferring one-carbon groups	2	8	0.002
methyltransferase activity	2	8	0.002
serine-type endopeptidase activity	2	14	0.006
zinc ion transport	2	14	0.006
serine hydrolase activity	2	15	0.007
serine-type peptidase activity	2	15	0.007
amylase activity	1	1	0.009
alpha-amylase activity	1	1	0.009
biological process	14	860	0.010
hydrolase activity	4	101	0.011
cation transport	4	108	0.014
zinc ion binding	2	22	0.016
cation binding	4	114	0.017
ion transport	4	114	0.017
metal ion binding	4	120	0.020
ion binding	4	121	0.020
cellular process	9	494	0.024
flagellin-based flagellum	1	3	0.026
ribosome biogenesis	1	3	0.026
response to xenobiotic stimulus	1	3	0.026
flagellum	1	3	0.026
ciliary or flagellar motility	1	3	0.026
endopeptidase activity	2	29	0.026
proteolysis	2	33	0.033
hydrogen-exporting ATPase activity, phosphorylative mechanism	1	4	0.035
nucleolus	1	4	0.035
cellular aromatic compound metabolic process	1	4	0.035
catalytic activity	6	284	0.035

Continued on next page

GO term	selected genes	genes in category	P-value
peptidase activity	2	34	0.035
metal ion transport	3	89	0.042
fatty acid binding	1	5	0.043
extracellular space	2	40	0.047

Table A.11: The genes shown in Figure 11.2.

clone id	P-value	gene description
C07_04_A02	0.000	Oryzias javanicus metallothionein (MT)
C200_01_X16	0.000	Glutathione S-transferase Mu 5 (EC 2.5.1.18) (GSTM5-5) (GST class-mu 5).
C12_03_E05	0.000	Collagen alpha-1(X) chain precursor.
Contig549	0.000	Metallothionein-2 (MT-2). [Source:Uniprot/SWISSPROT;Acc:Q7ZSY6]
C19_05_C05	0.000	Mus musculus placental protein 11 related (Pp11r)
C23_01_X125	0.001	
Contig793	0.001	alpha globin
Contig627	0.001	
Contig374	0.001	
C13_03_A05	0.001	ATGSTF3 (GLUTATHIONE S-TRANSFERASE 16); glutathione transferase
C11_01_H06	0.003	
C15_03_H06	0.003	
Contig92	0.003	Protein S100-A1 (S100 calcium-binding protein A1) (S-100 protein alpha subunit) (S-100 protein alpha chain).
C01_01_G04	0.004	
Contig241	0.005	
C04_04_C05	0.005	
C10_04_H05	0.006	
Contig630	0.006	Zgc:114014. zona pellucida glycoprotein 2-like
C101_01_B12	0.007	Low-density lipoprotein receptor-related protein 2 precursor (Megalin) (Glycoprotein 330) (gp330).
Contig788	0.007	60S ribosomal protein L29 (Cell surface heparin-binding protein HIP).
C01_05_H06	0.008	
C04_05_B10	0.008	

BIBLIOGRAPHY

[1] Zon, L. and Peterson, R. (2005) In vivo drug discovery in the zebrafish. *Nature Reviews Drug Discovery,* **4**(1), 35–44.

[2] Megason, S. and Fraser, S. (2007) Imaging in Systems Biology. *Cell,* **130**(5), 784–795.

[3] Sabaliauskas, N., Foutz, C., Mest, J., Budgeon, L., Sidor, A., Gershenson, J., Joshi, S., and Cheng, K. (2006) High-throughput zebrafish histology. *Methods,* **39**(3), 246–254.

[4] Goessling, W., North, T., and Zon, L. (2007) Ultrasound biomicroscopy permits in vivo characterization of zebrafish liver tumors. *Nature Methods,* **4**, 551–553.

[5] Keller, P., Schmidt, A., Wittbrodt, J., and Stelzer, E. (2008) Reconstruction of Zebrafish Early Embryonic Development by Scanned Light Sheet Microscopy. *Science,* **322**(5904), 1065.

[6] Area, S. and Index, A. (2007) Animal models of human disease: zebrafish swim into view. *Nature Reviews Genetics,* **8**(5), 353–367.

[7] Guyon, J., Steffen, L., Howell, M., Pusack, T., Lawrence, C., and Kunkel, L. (2007) Modeling human muscle disease in zebrafish. *BBA-Molecular Basis of Disease,* **1772**(2), 205–215.

[8] Feitsma, H. and Cuppen, E. (2008) Zebrafish as a Cancer Model. *Molecular Cancer Research,* **6**(5), 685.

[9] Colosimo, P., Hosemann, K., Balabhadra, S., Villarreal, G., Dickson, M., Grimwood, J., Schmutz, J., Myers, R., Schluter, D., and Kingsley, D. (2005) Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles. *Science,* **307**(5717), 1928–1933.

[10] Jaillon, O., Aury, J., Brunet, F., Petit, J., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature,* **431**, 946–957.

[11] Snape, J., Maund, S., Pickford, D., and Hutchinson, T. (2004) Ecotoxicogenomics: the challenge of integrating genomics into aquatic and terrestrial ecotoxicology. *Aquatic Toxicology,* **67**(2), 143–154.

[12] Ju, Z., Wells, M., and Walter, R. (2007) DNA microarray technology in toxicogenomics of aquatic models: Methods and applications. *Comparative Biochemistry and Physiology, Part C,* **145**(1), 5–14.

[13] Denslow, N., Garcia-Reyero, N., and Barber, D. (2007) Fish nchips: the use of microarrays for aquatic toxicology. *Molecular Biosystems,* **3**(3), 172.

[14] Waters, M. and Fostel, J. (2004) Toxicogenomics and systems toxicology: aims and prospects. *Nature Reviews Genetics,* **5**(12), 936–948.

[15] Heijne, W., Kienhuis, A., vanOmmen, B., Stierum, R., and Groten, J. (2005) Systems toxicology: applications of toxicogenomics, transcriptomics, proteomics and metabolomics in toxicology. *Expert Rev. Proteomics,* **2**(5), 767–780.

[16] Ryan, T., Stevens, J., and Thomas, C. (2008) Strategic applications of toxicogenomics in early drug discovery. *Current Opinion in Pharmacology,* **8**(5), 654–660.

[17] Knight, J. (2001) When the chips are down. *Nature,* **410**(6831), 860–861.

[18] Marshall, E. (2004) Getting the Noise Out of Gene Arrays. *Science,* **306**(5696), 630.

[19] Shi, L., Tong, W., Fang, H., Scherf, U., Han, J., Puri, R., Frueh, F., Goodsaid, F., Guo, L., Su, Z., et al. (2005) Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics,* **6**(Suppl 2), S12.

[20] Chen, J., Hsueh, H., Delongchamp, R., Lin, C., and Tsai, C. (2007) Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics,* **8**, 412.

[21] Yeung, K., Medvedovic, M., and Bumgarner, R. (2004) From co-expression to co-regulation: how many microarray experiments do we need?. *Genome Biology,* **5**(7).

[22] Garge, N., Page, G., Sprague, A., Gorman, B., and Allison, D. (2005) Reproducible clusters from microarray research: whither. *BMC Bioinformatics,* **6**(Suppl 2), S10.

[23] Allison, D., Cui, X., Page, G., and Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics,* **7**(1), 55.

[24] Duarte, N., Becker, S., Jamshidi, N., Thiele, I., Mo, M., Vo, T., Srivas, R., and Palsson, B. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences,* **104**(6), 1777.

[25] Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular Systems Biology,* **3**(135).

[26] Barabasi, A. and Oltvai, Z. (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics,* **5**(2), 101–113.

[27] Davidson, E. and Erwin, D. (2006) Gene Regulatory Networks and the Evolution of Animal Body Plans. *Science,* **311**(5762), 796–800.

[28] Vastrik, I., DEustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., deBono, B., Gillespie, M., Jassal, B., Lewis, S., et al. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biology,* **8**(3), R39.

[29] González, M. and Barabási, A. (2007) Complex networks: From data to models. *Nature Physics,* **3**, 224–225.

[30] Alon, U. (2007) Network motifs: theory and experimental approaches. *Nature Reviews Genetics,* **8**(6), 450.

[31] Kothapalli, R., Yoder, S., Mane, S., and Loughran Jr, T. (2002) Microarray results: how accurate are they?. *BMC Bioinformatics,* **3**, 22.

[32] Irizarry, R., Warren, D., Spencer, F., Kim, I., Biswal, S., Frank, B., Gabrielson, E., Garcia, J., Geoghegan, J., Germino, G., et al. (2005) Multiple-laboratory comparison of microarray platforms. *Nature methods,* **2**, 345–350.

[33] Larkin, J., Frank, B., Gavras, H., Sultana, R., and Quackenbush, J. (2005) Independence and reproducibility across microarray platforms. *Nature Methods,* **2**, 337–344.

[34] Draghici, S., Khatri, P., Eklund, A., and Szallasi, Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics,* **22**(2), 101–109.

[35] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C., Causton, H., et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics,* **29**(4), 365–372.

[36] Toxicogenomics Research Consortium (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods,* **2**, 351–356.

[37] MAQC Consortium (2006) The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology,* **24**, 1151–1161.

[38] Li, C. and Wong, W. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences,* **98**(1), 31.

[39] Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B., and Speed, T. (2003) Summaries of Affymetrix GeneChip probe level data.. *Nucleic Acids Research,* **31**(4), e15.

[40] Smyth, G., Ritchie, M., Thorne, N., and Wettenhall, J. (2005) Limma: linear models for microarray data. *Bioinformatics,* pp. 397–420.

[41] Hubbell, E. (2005) PLIER: An M-Estimator for Expression Array. *Affymetrix White Paper,*.

[42] Kerr, M. and Churchill, G. (2001) Statistical design and the analysis of gene expression microarray data.. *Genetical research,* **77**(2), 123.

[43] Tusher, V. et al. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences,* **98**(9), 5116–5121.

[44] Baldi, P. and Long, A. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. (2001).

[45] Cui, X., Hwang, J., Qiu, J., Blades, N., and Churchill, G. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates.. *Biostatistics,* **6**(1), 59.

[46] Michiels, S., Koscielny, S., and Hill, C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet,* **365**(9458), 488–492.

[47] van'tVeer, L., Dai, H., van deVijver, M., He, Y., Hart, A., Mao, M., Peterse, H., van derKooy, K., Marton, M., Witteveen, A., et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer.. *Nature,* **415**(6871), 530–6.

[48] Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Meijer-vanGelder, M., Yu, J., et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet,* **365**(9460), 671–679.

[49] Ein-Dor, L., Zuk, O., and Domany, E. (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences,* **103**(15), 5923–5928.

[50] Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares Jr, M., and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS,* **97**(1).

[51] Furey, T., Cristianini, N., Duffy, N., Bednarski, D., and Schummer, M. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics,* **16**(10), 906–914.

[52] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning, 46,* **1**(3), 389–422.

[53] Noble, W. (2006) What is a support vector machine?. *Nature biotechnology,* **24**, 1565–1567.

[54] Zhu, Y., Shen, X., and Pan, W. (2009) Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics,* **10**(Suppl 1), S21.

[55] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.. *Nat Genet,* **25**(1), 25–9.

[56] Zeeberg, B., Feng, W., Wang, G., Wang, M., Fojo, A., Sunshine, M., Narasimhan, S., Kane, D., Reinhold, W., Lababidi, S., et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol,* **4**(4), R28.

[57] Zeeberg, B., Qin, H., Narasimhan, S., Sunshine, M., Cao, H., Kane, D., Reimers, M., Stephens, R., Bryant, D., Burt, S., et al. (2005) High-Throughput GoMiner, an'industrial-strength'integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC bioinformatics,* **6**(1), 168.

[58] Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., and Lempicki, R. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome biology,* **4**(9), R60.

[59] Huang, D., Sherman, B., and R.A., L. (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc,* **4**(1), 44–57.

[60] Beissbarth, T. and Speed, T. P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics,* **20**(9), 1464–1465.

[61] Joyce, A. and Palsson, B. (2006) The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology,* **7**(3), 198–210.

[62] Affymetrix Inc. Affymetrix latin square data. *http://www.affymetrix.com/support/datasets.affx,* accessed 2008.

[63] Naef, F. and Magnasco, M. (2003) Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Physical Review E,* **68**(1), 11906.

[64] Pozhitkov, A., Stedtfeld, R., Hashsham, S., and Noble, P. (2007) Revision of the nonequilibrium thermal dissociation and stringent washing approaches for identification of mixed nucleic acid targets by microarrays. *Nucleic Acids Research,*.

[65] Tu, Y., Stolovitzky, G., and Klein, U. (2002) Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences,* **99**(22), 14031–14036.

[66] Bakay, M., Chen, Y., Borup, R., Zhao, P., Nagaraju, K., and Hoffman, E. (2002) Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformatics,* **3**(4), 1471–2105.

[67] Zakharkin, S., Kim, K., Mehta, T., Chen, L., Barnes, S., Scheirer, K., Parrish, R., Allison, D., and Page, G. (2005) Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics,* **6**, 214.

[68] Jones, L., Goldstein, D., Hughes, G., Strand, A., Collin, F., Dunnett, S., Kooperberg, C., Aragaki, A., Olson, J., Augood, S., et al. (2006) Assessment of the relationship between pre-chip and post-chip quality measures for Affymetrix GeneChip expression data. *BMC Bioinformatics,* **7**, 211.

[69] Klebanov, L. and Yakovlev, A. (2007) How high is the level of technical noise in microarray data. *Biology Direct,* **2**(1), 9.

[70] Zhang, L., Miles, M., and Aldape, K. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nature biotechnology,* **21**(7), 818–821.

[71] Li, S., Pozhitkov, A., and Brouwer, M. (2008) A competitive hybridization model predicts probe signal intensity on high density DNA microarrays. *Nucleic Acids Research,* **36**(20), 6585–91.

[72] Shendure, J. (2008) The beginning of the end for microarrays?. *Nature Methods,* **5**, 585–587.

[73] Peterson, A., Heaton, R., and Georgiadis, R. (2001) The effect of surface probe density on DNA hybridization.. *Nucleic Acids Research,* **29**(24), 5163.

[74] Nelson, B., Grimsrud, T., Liles, M., Goodman, R., and Corn, R. (2001) Surface plasmon resonance imaging measurements of DNA and RNA hybridization adsorption onto DNA microarrays.. *Anal Chem,* **73**(1), 1–7.

[75] Yu, F., Yao, D., and Knoll, W. (2004) Oligonucleotide hybridization studied by a surface plasmon diffraction sensor (SPDS).. *Nucleic Acids Research,* **32**(9), e75.

[76] Lee, H., Wark, A., Goodrich, T., Fang, S., and Corn, R. (2005) Surface enzyme kinetics for biopolymer microarrays: a combination of Langmuir and Michaelis-Menten concepts. *Langmuir,* **21**(9), 4050–4057.

[77] Bishop, J., Blair, S., and Chagovetz, A. (2006) A Competitive Kinetic Model of Nucleic Acid Surface Hybridization in the Presence of Point Mutants. *Biophysical Journal,* **90**(3), 831–840.

[78] Golovlev, V., Sun, Y., Fan, W., and McCann, M. (2007) Hybridization kinetics on microarray surfaces.. *Biotechnology J.,* **2**(8), 988–91.

[79] Levicky, R. and Horgan, A. (2005) Physicochemical perspectives on DNA microarray and biosensor technologies. *Trends in Biotechnology,* **23**(3), 143–149.

[80] Halperin, A., Buhot, A., and Zhulina, E. (2006) On the hybridization isotherms of DNA microarrays: the Langmuir model and its extensions. *J. Phys. Condens. Matter,* **18**, S463–S490.

[81] Michel, W., Mai, T., Naiser, T., and Ott, A. (2007) Optical Study of DNA Surface Hybridization Reveals DNA Surface Density as a Key Parameter for Microarray Hybridization Kinetics. *Biophysical Journal,* **92**(3), 999.

[82] Demers, L., Mirkin, C., Mucic, R., Reynolds 3rd, R., Letsinger, R., Elghanian, R., and Viswanadham, G. (2000) A fluorescence-based method for determining the surface coverage and hybridization efficiency of thiol-capped oligonucleotides bound to gold thin films and nanoparticles.. *Analytical chemistry,* **72**(22), 5535.

[83] McKendry, R., Zhang, J., Arntz, Y., Strunz, T., Hegner, M., Lang, H., Baller, M., Certa, U., Meyer, E., Guntherodt, H., et al. (2002) Multiple label-free biodetection and quantitative DNA-binding assays on a nanomechanical cantilever array. *Proceedings of the National Academy of Sciences,* **99**(15), 9783.

[84] Negnevitsky, M. (2005) Artificial Intelligence: A Guide to Intelligent Systems, Addison-Wesley, Harlow, England, .

[85] Chan, V., Graves, D., and McKenzie, S. (1995) The biophysics of DNA hybridization with immobilized oligonucleotide probes. *Biophysical Journal,* **69**(6), 2243–2255.

[86] Stevens, P., Henry, M., and Kelso, D. (1999) DNA hybridization on microparticles: determining capture-probe density and equilibrium dissociation constants.. *Nucleic Acids Research,* **27**(7), 1719.

[87] Watterson, J., Piunno, P., Wust, C., and Krull, U. (2000) Effects of Oligonucleotide Immobilization Density on Selectivity of Quantitative Transduction of Hybridization of Immobilized DNA. *Langmuir,* **16**(11), 4984–4992.

[88] Yao, D., Kim, J., Yu, F., Nielsen, P., Sinner, E., and Knoll, W. (2005) Surface Density Dependence of PCR Amplicon Hybridization on PNA/DNA Probe Layers. *Biophysical Journal,* **88**(4), 2745–2751.

[89] Jayaraman, A., Hall, C., and Genzer, J. (2007) Computer simulation study of probe-target hybridization in model DNA microarrays: Effect of probe surface density and target concentration. *The Journal of Chemical Physics,* **127**, 144912.

[90] Xu, R. and Li, X. (2003) A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data.. *Bioinformatics (Oxford, England),* **19**(10), 1284.

[91] Molinaro, A., Simon, R., and Pfeiffer, R. (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics,* **21**(15), 3301–3307.

[92] Fan, J. and Ren, Y. (2006) Statistical analysis of DNA microarray data in cancer research. *Clinical Cancer Research,* **12**(15), 4469.

[93] SantaLucia Jr, J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences,* **95**(4), 1460.

[94] Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M., and Sasaki, M. (1995) Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. *Biochemistry,* **34**(35), 11211–11216.

[95] Wu, P., Nakano, S., and Sugimoto, N. (2002) Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation. *FEBS Journal,* **269**(12), 2821–2830.

[96] Fotin, A., Drobyshev, A., Proudnikov, D., Perov, A., and Mirzabekov, A. Parallel thermodynamic analysis of duplexes on oligodeoxyribonucleotide microchips. *Nucleic Acids Research,* **26**(6), 1515–1521.

[97] Wick, L., Rouillard, J., Whittam, T., Gulari, E., Tiedje, J., and Hashsham, S. (2006) On-chip non-equilibrium dissociation curves and dissociation rate constants as methods to assess specificity of oligonucleotide probes. *Nucleic Acids Research,* **34**(3), e26.

[98] Fish, D., Horne, M., Brewood, G., Goodarzi, J., Alemayehu, S., Bhandiwad, A., Searles, R., and Benight, A. (2007) DNA multiplex hybridization on microarrays and thermodynamic stability in solution: a direct comparison. *Nucleic Acids Research,* **35**(21), 7197.

[99] Weckx, S., Carlon, E., De Vuyst, L., and Van Hummelen, P. (2007) Thermodynamic behavior of short oligonucleotides in microarray hybridizations can be described using Gibbs free energy in a nearest-neighbor model. *Journal of Physical Chemistry B,* **111**(48), 13583–13590.

[100] Matveeva, O., Shabalina, S., Nemtsov, V., Tsodikov, A., Gesteland, R., Atkins, J., and Journals, O. (2003) Thermodynamic calculations and statistical correlations for oligo-probes design. *Nucleic Acids Research,* **31**(14), 4211–4217.

[101] He, Z., Wu, L., Li, X., Fields, M., and Zhou, J. (2005) Empirical establishment of oligonucleotide probe design criteria.. *Appl Environ Microbiol,* **71**(7), 3753–60.

[102] Wei, H., Kuan, P., Tian, S., Yang, C., Nie, J., Sengupta, S., Ruotti, V., Jonsdottir, G., Keles, S., Thomson, J., et al. (2008) A study of the relationships between oligonucleotide properties and hybridization signal intensities from NimbleGen microarray datasets.. *Nucleic Acids Res,*.

[103] Binder, H. and Preibisch, S. (2006) GeneChip microarrayssignal intensities, RNA concentrations and probe sequences. *Journal of Physics: Condensed Matter,* **18**(18), S537–S566.

[104] Held, G., Grinstein, G., and Tu, Y. (2006) Relationship between gene expression and observed intensities in DNA microarrays–a modeling study. *Nucleic Acids Research,* **34**(9), e70.

[105] Bruun, G., Wernersson, R., Juncker, A., Willenbrock, H., and Nielsen, H. (2007) Improving comparability between microarray probe signals by thermodynamic intensity correction. *Nucleic Acids Research,* **35**(7), e48.

[106] Ono, N., Suzuki, S., Furusawa, C., Agata, T., Kashiwagi, A., Shimizu, H., and Yomo, T. (2008) An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays.. *Bioinformatics,* **24**(10), 1278–85.

[107] Hekstra, D., Taussig, A., Magnasco, M., and Naef, F. (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Research,* **31**(7), 1962–1968.

[108] Mei, R., Hubbell, E., Bekiranov, S., Mittmann, M., Christians, F., Shen, M., Lu, G., Fang, J., Liu, W., Ryder, T., et al. (2003) Probe selection for high-density oligonucleotide arrays. *Proceedings of the National Academy of Sciences,* **100**(20), 11237.

[109] Held, G., Grinstein, G., and Tu, Y. (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proceedings of the National Academy of Sciences,* **100**(13), 7575.

[110] Halperin, A., Buhot, A., and Zhulina, E. (2004) Sensitivity, Specificity, and the Hybridization Isotherms of DNA Chips. *Biophysical Journal,* **86**(2), 718–730.

[111] Abdueva, D., Skvortsov, D., and Tavare, S. (2006) Non-linear analysis of GeneChip arrays. *Nucleic Acids Research,* **34**(15), e105.

[112] Heim, T., Tranchevent, L., Carlon, E., and Barkema, G. (2006) Physical-Chemistry-Based Analysis of Affymetrix Microarray Data. *Journal of physical chemistry. B, Condensed matter, materials, surfaces, interfaces, & biophysical chemistry,* **110**(45), 22786–22795.

[113] Burden, C., Pittelkow, Y., and Wilson, S. (2006) Adsorption models of hybridization and post-hybridization behaviour on oligonucleotide microarrays. *Journal of Physics, Condensed Matter,* **18**(23), 5545–5565.

[114] Burden, C. (2008) Understanding the physics of oligonucleotide microarrays: the Affymetrix spike-in data reanalysed. *Phys Biol.,* **5**(1), 16004.

[115] Skvortsov, D., Abdueva, D., Curtis, C., Schaub, B., and Tavare, S. (2007) Explaining differences in saturation levels for Affymetrix GeneChip (R) arrays. *Nucleic Acids Research,* **35**(12), 4154.

[116] Wu, Z. and Irizarry, R. (2004) Preprocessing of oligonucleotide array data. *Nature Biotechnology,* **22**(6), 656–658.

[117] Bloomfield, V. e. a., (ed.) Nucleic Acids: Structures, Properties, and Functions, , .

[118] Christensen, U., Jacobsen, N., Rajwanshi, V., Wengel, J., and Koch, T. (2001) Stopped-flow kinetics of locked nucleic acid (LNA)-oligonucleotide duplex formation: studies of LNA-DNA and DNA-DNA interactions. *BIOCHEMICAL JOURNAL-LONDON-,* **354**(3), 481–484.

[119] Vainrub, A. and Pettitt, B. (2000) Thermodynamics of association to a molecule immobilized in an electric double layer. *Chemical Physics Letters,* **323**(1-2), 160–166.

[120] Halperin, A., Buhot, A., and Zhulina, E. (2005) Brush Effects on DNA Chips: Thermodynamics, Kinetics, and Design Guidelines. *Biophysical Journal,* **89**(2), 796–811.

[121] Zhang, L., Hurek, T., and Reinhold-Hurek, B. (2005) Position of the fluorescent label is a crucial factor determining signal intensity in microarray hybridizations. *Nucleic Acids Research,* **33**(19), e166.

[122] Zhang, Y., Hammer, D., and Graves, D. (2005) Competitive Hybridization Kinetics Reveals Unexpected Behavior Patterns. *Biophysical Journal,* **89**(5), 2950–2959.

[123] Shi, L., Tong, W., Su, Z., Han, T., Han, J., Puri, R., Fang, H., Frueh, F., Goodsaid, F., Guo, L., et al. (2005) Microarray scanner calibration curves: characteristics and implications.. *BMC Bioinformatics,* **6**(2), S11.

[124] Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association,* **99**(468), 909–917.

[125] Schuster, E., Blanc, E., Partridge, L., and Thornton, J. (2007) Estimation and correction of non-specific binding in a large-scale spike-in experiment. *Genome Biology,* **8**(6), R126.

[126] Affymetrix Inc. Netaffx analysis center.
http://www.affymetrix.com/analysis/index.affx, accessed 2008.

[127] Mathews, D., Disney, M., Childs, J., Schroeder, S., Zuker, M., and Turner, D. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences,* **101**(19), 7287–7292.

[128] Mir, K. and Southern, E. (1999) Determining the influence of structure on hybridization using oligonucleotide arrays. *Nature Biotechnology,* **17**, 788–792.

[129] Lane, S., Evermann, J., Loge, F., and Call, D. (2004) Amplicon secondary structure prevents target hybridization to oligonucleotide microarrays. *Biosensors and Bioelectronics,* **20**(4), 728–735.

[130] Ratushna, V., Weller, J., and Gibas, C. (2005) Secondary structure in the target as a confounding factor in synthetic oligomer microarray design.. *BMC Genomics,* **6**(1), 31.

[131] McCaskill, J. (1990) The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure. *Biopolymers,* **29**, 1105–1119.

[132] Zhang, L., Wu, C., Carta, R., and Zhao, H. (2007) Free energy of DNA duplex formation on short oligonucleotide microarrays. *Nucleic Acids Research,* **35**(3), e18.

[133] Sekar, M., Bloch, W., and StJohn, P. (2005) Comparative study of sequence-dependent hybridization kinetics in solution and on microspheres. *Nucleic Acids Research,* **33**(1), 366–375.

[134] Chien, F., Liu, J., Su, H., Kao, L., Chiou, C., Chen, W., and Chen, S. (2004) An investigation into the influence of secondary structures on DNA hybridization using surface plasmon resonance biosensing. *Chemical Physics Letters,* **397**(4-6), 429–434.

[135] Gao, Y., Wolf, L., and Georgiadis, R. (2006) Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nucleic Acids Research,* **34**(11), 3370.

[136] Gharaibeh, R., Fodor, A., and Gibas, C. (2008) Background correction using dinucleotide affinities improves the performance of GCRMA. *BMC bioinformatics,* **9**(1), 452.

[137] Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly,* **125**(2), 167–188.

[138] Dahlquist, K., Salomonis, N., Vranizan, K., Lawlor, S., and Conklin, B. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics,* **31**(1), 19–20.

[139] Salomonis, N., Hanspers, K., Zambon, A., Vranizan, K., Lawlor, S., Dahlquist, K., Doniger, S., Stuart, J., Conklin, B., and Pico, A. (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics,* **8**, 217.

[140] Nikitin, A., Egorov, S., Daraselia, N., and Mazo, I. (2003) Pathway studio-the analysis and navigation of molecular networks. *Bioinformatics,* **19**(16), 2155–2157.

[141] Ingenuity Systems. Ingenuity Pathway Analysis. *www.ingenuity.com,* accessed 2008.

[142] Kim, T., Barrera, L., Zheng, M., Qu, C., Singer, M., Richmond, T., Wu, Y., Green, R., and Ren, B. (2005) A high-resolution map of active promoters in the human genome. *Nature,* **436**(7052), 876.

[143] Johnson, D., Mortazavi, A., Myers, R., and Wold, B. (2007) Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science,* **316**(5830), 1497.

[144] Yu, H., Braun, P., Yildirim, M., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008) High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science,* **322**(5898), 104.

[145] Tarassov, K., Messier, V., Landry, C., Radinovic, S., Molina, M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S. (2008) An in Vivo Map of the Yeast Protein Interactome. *Science's STKE,* **320**(5882), 1465.

[146] Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., et al. (2005) Global analysis of protein phosphorylation in yeast. *Nature,* **438**(7068), 679–684.

[147] Rual, J., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G., Gibbons, F., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature,* **437**(7062), 1173–1178.

[148] Gandhi, T., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K., Mohan, S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., et al. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics,* **38**, 285–293.

[149] Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., et al. (2005) Protein interaction mapping: a Drosophila case study.. *Genome Res,* **15**(3), 376–84.

[150] David, H., Hofmann, G., Oliveira, A., Jarmer, H., and Nielsen, J. (2006) Metabolic network-driven analysis of genome-wide transcription data from Aspergillus nidulans.. *Genome Biol,* **7**(11), R108.

[151] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research,* **34**(Database Issue), D354.

[152] Schilling, C., Covert, M., Famili, I., Church, G., Edwards, J., and Palsson, B. (2002) Genome-scale metabolic model of Helicobacter pylori 26695. *Journal of Bacteriology,* **184**(16), 4582–4593.

[153] Ma, H. and Zeng, A. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.. *Bioinformatics (Oxford, England),* **19**(2), 270.

[154] Becker, S. and Palsson, B. (2005) Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. *BMC Microbiol,* **5**(8).

[155] Heinemann, M., Kummel, A., Ruinatscha, R., and Panke, S. (2005) In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network. *Biotechnology and bioengineering,* **92**(7), 850–864.

[156] Förster, J., Famili, I., Fu, P., Palsson, B., and Nielsen, J. (2003) Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. *Genome Research,* **13**(2), 244.

[157] Albert, R. and Barabási, A. (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics,* **74**(1), 47–97.

[158] Jeong, H., Tombor, B., Albert, R., Oltvai, Z., Barabasi, A., et al. (2000) The large-scale organization of metabolic networks. *NATURE-LONDON-,* pp. 651–653.

[159] Ma, H., Zhao, X., Yuan, Y., and Zeng, A. (2004) Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics,* **20**(12), 1870–1876.

[160] Newman, M. and Girvan, M. (2004) Finding and evaluating community structure in networks. *Physical Review E,* **69**(2), 26113.

[161] Newman, M. (2006) Modularity and community structure in networks. *Proceedings of the National Academy of Sciences,* **103**(23), 8577–8582.

[162] Wagner, A. (2001) The small world inside large metabolic networks. *Proceedings of the Royal Society B: Biological Sciences,* **268**(1478), 1803–1810.

[163] Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I., and Dandekar, T. (2002) Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae.. *Bioinformatics,* **18**(2), 351–61.

[164] Huss, M. and Holme, P. (2007) Currency and commodity metabolites: Their identification and relation to the modularity of metabolic networks. *Systems Biology, IET,* **1**(5), 280–285.

[165] Holme, P. and Huss, M. (2008) Currency metabolites and network representations of metabolism. *Arxiv preprint arXiv:0806.2763,*.

[166] Hubbard, T., Aken, B., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. (2007) Ensembl 2007. *Nucleic Acids Research,* **35**(1), D610–D617.

[167] Al-Shahrour, F., Diaz-Uriarte, R., and Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics,* **20**(4), 578.

[168] Zehetner, G. (2003) OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic acids research,* **31**(13), 3799.

[169] Khan, S., Situ, G., Decker, K., and Schmidt, C. (2003) GoFigure: Automated Gene Ontology annotation. *Bioinformatics,* **19**(18), 2484–2485.

[170] Groth, D., Lehrach, H., and Hennig, S. (2004) GOblet: a platform for Gene Ontology annotation of anonymous sequence data. *Nucleic acids research,* **32**(Web Server Issue), W313.

[171] Chalmel, F., Lardenois, A., Thompson, J., Muller, J., Sahel, J., Léveillard, T., and Poch, O. (2005) GOAnno: GO annotation based on multiple alignment.. *Bioinformatics (Oxford, England),* **21**(9), 2095.

[172] Conesa, A., Gotz, S., Garcia-Gomez, J., Terol, J., Talon, M., and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* **21**(18), 3674–3676.

[173] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* **25**(17), 3389–3402.

[174] Sprague, J., Doerry, E., Douglas, S., and Westerfield, M. (2001) The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research. *Nucleic Acids Research,* **29**(1), 87.

[175] Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D., Mani, P., Ramachandran, S., et al. (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic acids research,* **34**(Database Issue), D581.

[176] Gentry, J., Carey, V., E., G., and R., G. (September, 2004) Laying out pathways with Rgraphviz. *R News,* **4**(2), 14–18.

[177] Funahashi, A., Morohashi, M., Kitano, H., and Tanimura, N. (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico,* **1**(5), 159–162.

[178] Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research,* **13**(11), 2498.

[179] Vandepoele, K., De Vos, W., Taylor, J., Meyer, A., and Van dePeer, Y. (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences,* **101**(6), 1638–1643.

[180] Ideker, T., Thorsson, V., Ranish, J., Christmas, R., Buhler, J., Eng, J., Bumgarner, R., Goodlett, D., Aebersold, R., and Hood, L. (2001) Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science,* **292**(5518), 929.

[181] Lam, S., Wu, Y., Vega, V., Miller, L., Spitsbergen, J., Tong, Y., Zhan, H., Govindarajan, K., Lee, S., Mathavan, S., et al. (2005) Conservation of gene expression signatures between zebrafish and human liver tumors and tumor progression. *Nature biotechnology,* **24**, 73–75.

[182] Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., et al. (2003) TM4: a free, open-source system for microarray data management and analysis.. *Biotechniques,* **34**(2), 374.

[183] Kind, T., Tolstikov, V., Fiehn, O., and Weiss, R. (2007) A comprehensive urinary metabolomic approach for identifying kidney cancer. *Analytical Biochemistry,* **363**(2), 185–195.

[184] Podo, F., Sardanelli, F., Iorio, E., Canese, R., Carpinelli, G., Fausto, A., and Canevari, S. (2007) Abnormal choline phospholipid metabolism in breast and ovary cancer: molecular bases for noninvasive imaging approaches. *Current medical imaging reviews,* **3**(2), 123–138.

[185] Kaddurah-Daouk, R., Kristal, B., and Weinshilboum, R. (2008) Metabolomics: a global biochemical approach to drug response and disease.. *Annual review of pharmacology and toxicology,* **48**, 653.

[186] Abate-Shen, C. and Shen, M. (2009) Diagnostics The prostate-cancer metabolome. *Nature,* **457**(7231), 799–800.

[187] Spratlin, J., Serkova, N., and Eckhardt, S. (2009) Clinical Applications of Metabolomics in Oncology: A Review. *Clinical Cancer Research,* **15**(2), 431.

[188] Shlomi1, T., Cabili, M., and Ruppin, E. (2009) Predicting metabolic biomarkers of human inborn errors of metabolism. *Molecular Systems Biology,* **5**(263).

[189] Segal, E., Friedman, N., Kaminski, N., Regev, A., and Koller, D. (2005) From signatures to models: understanding cancer using microarrays. *Nature genetics,* **37**(6 suppl).

[190] Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., Stichweh, D., Blankenship, D., Li, L., Munagala, I., et al. (2008) A modular analysis framework for blood genomics studies: Application to systemic lupus erythematosus. *Immunity,* **29**(1), 150–164.

[191] Lynn, D., Winsor, G., Chan, C., Richard, N., Laird, M., Barsky, A., Gardy, J., Roche, F., Chan, T., Shah, N., et al. (2008) InnateDB: Facilitating systems-level analyses of the mammalian innate immune response. *Molecular Systems Biology,* **4**(1).

[192] Bozdech, Z., Zhu, J., Joachimiak, M., Cohen, F., Pulliam, B., DeRisi, J., et al. (2003) Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray.. *Genome Biology,* **4**(2), R9.

[193] Stramma, L. (2008) Expanding Oxygen-Minimum Zones in the Tropical. *science,* **1153847**(655), 320.

[194] Shaffer, G., Olsen, S., and Pedersen, J. (2009) Long-term ocean oxygen depletion in response to carbon dioxide emissions from fossil fuels. *Nature Geoscience,* **2**(2), 105–109.

[195] Hendon, L., Carlson, E., Manning, S., and Brouwer, M. (2008) Molecular and developmental effects of exposure to pyrene in the early life-stages of Cyprinodon variegatus.. *Comparative biochemistry and physiology. Toxicology & pharmacology: CBP,* **147**(2), 205.

[196] Brouwer, M., Brown-Peterson, N., Hoexum-Brouwer, T., Manning, S., and Denslow, N. (2008) Changes in mitochondrial gene and protein expression in grass shrimp, Palaemonetes pugio, exposed to chronic hypoxia.. *Marine environmental research,*.

[197] KENT, W. (2002) BLAT-the BLAST-like alignment tool. *Genome research,* **12**(4), 656–664.

[198] Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[199] Fukuda, R., Zhang, H., Kim, J., Shimoda, L., Dang, C., and Semenza, G. (2007) HIF-1 regulates cytochrome oxidase subunits to optimize efficiency of respiration in hypoxic cells. *Cell,* **129**(1), 111–122.

[200] Kim, J., Tchernyshyov, I., Semenza, G., and Dang, C. (2006) HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. *Cell Metabolism,* **3**(3), 177–185.

[201] Papandreou, I., Cairns, R., Fontana, L., Lim, A., and Denko, N. (2006) HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption. *Cell Metabolism,* **3**(3), 187–197.

[202] Jelkmann, W. and Metzen, E. (1996) Erythropoietin in the control of red cell production.. *Annals of anatomy= Anatomischer Anzeiger: official organ of the Anatomische Gesellschaft,* **178**(5), 391.

[203] Ebert, B. and Bunn, H. (1999) Regulation of the erythropoietin gene. *Blood,* **94**(6), 1864.

[204] Roesner, A., Hankeln, T., and Burmester, T. (2006) Hypoxia induces a complex response of globin expression in zebrafish (Danio rerio). *Journal of Experimental Biology,* **209**(11), 2129–2137.

[205] Rifkind, A. (2006) CYP1A in TCDD toxicity and in physiology-with particular reference to CYP dependent arachidonic acid metabolism and other endogenous substrates. *Drug Metabolism Reviews,* **38**(1-2), 291–335.

[206] Reynaud, S., Raveton, M., and Ravanel, P. (2008) Interactions between immune and biotransformation systems in fish: A review. *Aquatic Toxicology,* **87**(3), 139–145.

[207] Anwar-Mohamed, A., Elbekai, R., and El-Kadi, A. (2009) Regulation of CYP1A1 by heavy metals and consequences for drug metabolism. *Expert Opin Drug Metab Toxicol,* **5**(5), 501–21.

[208] van derOost, R., Beyer, J., and Vermeulen, N. (2003) Fish bioaccumulation and biomarkers in environmental risk assessment: a review. *Environmental Toxicology and Pharmacology,* **13**(2), 57–149.

[209] Casalino, E., Sblano, C., Calzaretti, G., and Landriscina, C. (2006) Acute cadmium intoxication induces alpha-class glutathione S-transferase protein synthesis and enzyme activity in rat liver. *Toxicology,* **217**(2-3), 240–245.

[210] Hawse, J., Cumming, J., Oppermann, B., Sheets, N., Reddy, V., and Kantorow, M. (2003) Activation of Metallothioneins and α-Crystallin/sHSPs in Human Lens Epithelial Cells by Specific Metals and the Metal Content of Aging Clear Human Lenses. *Investigative ophthalmology & visual science,* **44**(2), 672–679.

[211] Loumbourdis, N., Kostaropoulos, I., Theodoropoulou, B., and Kalmanti, D. (2007) Heavy metal accumulation and metallothionein concentration in the frog Rana ridibunda after exposure to chromium or a mixture of chromium and cadmium. *Environmental Pollution,* **145**(3), 787–792.

[212] Kojima, I., Tanaka, T., Inagi, R., Nishi, H., Aburatani, H., Kato, H., Miyata, T., Fujita, T., and Nangaku, M. (2009) Metallothionein is upregulated by hypoxia and stabilizes hypoxia-inducible factor in the kidney. *Kidney International,* **75**(3), 268–277.

[213] Head, M., Hurwitz, L., and Goldman, J. (1996) Transcription regulation of alpha B-crystallin in astrocytes: analysis of HSF and AP1 activation by different types of physiological stress. *Journal of cell science,* **109**, 1029.