4

GENOMIC SIGNAL PROCESSING OF DNA MICROARRAY DATA FOR THE ENHANCED PREDICTION OF AXILLARY LYMPH NODE STATUS OF BREAST CANCER TUMORS

Gordon S. Okimoto

Bioinformatics Shared Resource, Cancer Research Center of Hawaii, University of Hawaii, Manoa, 1236 Lauhala Street, Honolulu, Hawaii 96813

4.1 INTRODUCTION

Breast cancer is the second leading cause of cancer deaths in the women today (after lung cancer) and is the most common cancer among women, excluding nonmelanoma skin cancers. Early detection and more effective treatments have decreased the mortality rate from breast cancer in recent years [1]. Still, according to the World Health Organization, more than 1.2 million people will be diagnosed with breast cancer each year worldwide. The American Cancer Society estimates that each year

Systems Biology and Synthetic Biology Edited by Pengcheng Fu and Sven Panke

178,000 Americans will be diagnosed with breast cancer with 44,000 deaths expected. Moreover, breast cancer is the leading cause of death among women between 40 and 55 years of age and is the second overall cause of death among American women (exceeded only by lung cancer).

The term breast cancer refers to a collection of cells of the breast that undergo uncontrolled growth, differentiation, and proliferation. Such a collection of cells is known as a malignant breast cancer tumor. Malignant tumors penetrate and destroy healthy tissues of the breast. In addition, a group of cells within a malignant tumor may also break away and spread to other parts of the body. Breast cancer tumor cells that spread from one region of the body into another are called metastases. One goal of this chapter is to characterize the metastatic potential of breast cancer tumors in terms of their global gene expression profiles.

Clinically, the presence of metastatic breast cancer in axillary lymph nodes is the most significant factor in the overall survival of breast cancer patients [2,3]. Although the determination of lymph node status is routine, the surgical procedure is invasive, and the selection of lymph nodes for examination can introduce biases that result in false negative results. Hence, the ability to assess the lymph node status of a breast cancer tumor based on quantitative measurements derived from the tumor itself may obviate the need for axillary lymph node dissection and the morbidity associated with the procedure [4].

Previous attempts to correlate characteristics of primary breast cancer tumors such as S-phase fraction, tumor grade, ploidy, hormone receptor status, and ERBB2 overexpression with lymph node status have been less than successful in terms of the sensitivity and specificity required in clinical settings [5]. Multivariable gene expression profiling appears to have the analytical resolution necessary to complement the known clinical markers currently used for tumor characterization [6]. In addition, the genes, pathways, and predictive models that result from a global analysis of gene expression in breast cancer tumors provide biological hypotheses for highly focused studies to identify new molecular targets that may contribute to improved treatment and personalized care, and a deeper understanding of the systems biology underlying breast cancer metastasis and tumor growth [7].

In this chapter, modern signal processing and pattern recognition techniques that employ the wavelet transform (WT), singular value decomposition (SVD), and neural networks (NNs) are used to analyze microarray data to predict the spread of breast cancer to the axillary lymph nodes based solely on the gene expression profiles of the primary breast cancer tumor. In Section 4.2, background knowledge on breast cancer and genomic signal processing and a description of the main clinical problem of interest are provided; that is, assessing the distant spread of breast cancer to the axillary lymph nodes based on the molecular characteristics of primary tumor. In Section 4.3, a microarray data set based on normal tissue and breast cancer tumor samples is described. In Section 4.4, results of a prior analysis on the Huang data set by Huang et al. are summarized [4]. In Section 4.5, genomic signal processing techniques such as WT and SVD are defined and discussed. The expression data matrix is discussed in Section 4.6 and its connection to Bellman's curse of dimensionality. Experimental design issues for the current study are discussed in Section 4.7. Data

preprocessing and data quality issues are discussed in Section 4.8. In Section 4.9, the modeling of phenotypic variation using features extracted from the Huang breast cancer data set using genomic signal processing techniques is described. Validation of pattern recognition models derived from the Huang microarray data is described in Section 4.10. Section 4.11 summarizes the main results of the overall study. Finally, a discussion of the main results is presented in Section 4.12.

4.2 BACKGROUND ON METHODS AND APPROACH

The central dogma of molecular biology states that a gene is transcribed into messenger RNA (mRNA) that in turn is translated into protein [8]. Networks of interacting genes and proteins then give rise to emergent states and system dynamics on these states that characterize the complex biological processes in cells, tissues, organs, and organisms [7]. Although the central dogma has been modified somewhat over the years, the core idea is still valid—the flow of information from genes to mRNA to proteins—and the underlying information processing that it implies forms the basis for life, death, and disease.

A crucial step in the information processing described by the central dogma is the transcription of a gene into mRNA, a process also known as transcription or gene expression. The expression level of every known gene represents the global gene expression pattern of a biological sample. This pattern is in constant flux over space and time and, in particular, changes as a normal cell is transformed into a cancer cell [9]. In this light, it is reasonable to assume that global gene expression patterns of normal and cancerous cells are quite different.

An important goal in cancer systems biology is the proper quantification of differences in global gene expression between normal or cancer cells [4]. The genes that underlie such differences serve as explanatory variables of quantitative models of cancer that are predictive of clinical outcomes or discriminative between different cancer subtypes [10]. For the first time, biologists are now able to measure the expression of every known gene in a tissue sample using a technology called the DNA microarray or chip. In a marriage of integrated circuit manufacturing, nanotechnology, photonics, materials science, biochemistry, and molecular biology, DNA microarrays are able to measure the activity of thousands of genes simultaneously using thousands of distinct probes that are positioned randomly on the surface of a small glass slide, plastic wafer, or silicon chip [11–13]. Each probe is composed of millions of strands of DNA that are complementary to specific mRNA target strand that we wish to quantify. Fluorescent tags are attached to the mRNA strands contained in a special ‘‘cocktail’’ prepared from a biological sample. The chip is immersed in the cocktail for a period of time under stringent conditions to allow the different mRNA target strands to attach, or hybridize, to their complementary DNA probes. The amount of tagged mRNA that hybridizes to a specific probe is quantified based on the intensity of the light that is emitted by the fluorescent probe when illuminated by a beam of laser light. This measure of light intensity serves as a surrogate measure of expression for the gene associated with the probe.

Each probe is interrogated in this way and the individual expression measurements are assembled into a high-dimensional vector that in aggregate provides a global snapshot of gene activity in a given tissue sample [14].

When multiple chips are hybridized to different samples, we have a microarray experiment. In the context of this chapter, the microarray experiment of interest compares the global gene expression patterns of tissue samples composed of normal cells to tissue samples composed of cancerous cells. Since not all 25,000 or so genes of the human genome are associated with cancer, it is reasonable to assume that only a relatively small number of genes will be differentially expressed (DE) between the normal and tumor samples. Here, we view the collection of DE genes as a characterizing biological state of tumor cells in terms of gene expression [15].

DE genes that interact in the context of a known signaling or regulatory pathways can be used as features for pattern recognition applications that are capable of identifying cancer subtypes and predicting clinical outcomes prior to and during treatment [5,9]. Indeed, a list of DE genes most likely intersects with multiple signaling pathways that control the transformation of a normal cell into a cancer cell [16]. Deconvolution of this list into pathways of functionally related, interacting genes helps to elucidate causal mechanisms that may lead to a more personalized treatment of cancer through early diagnosis and drugs targeted to the specific genes in specific pathways [17].

As with any other sensor system, data collected by DNA microarrays are contaminated with significant amounts of systematic (low-frequency) and random (high-frequency) variation. The primary sources for such unwanted variation include experimental error introduced by the data acquisition process unique to DNA microarrays and biological variation that exists between different tissue samples. The resulting $p \times n$ data matrix of a typical microarray experiment, where p equals the number of genes and n equals the number of samples, is ill posed in that p is greater than n ($p \gg n$) by several orders of magnitude. This situation is analogous to having many more unknowns than equations in a system of linear equations whereby the system in question has no solution. Standard statistical analysis of such ill-posed data results in models that mistake noise for signal, and hence, fail to capture the underlying biological processes that give rise to the observed patterns of differential gene expression [18,19]. Finally, background noise in microarray data is multiplicative instead of additive, which can confound standard statistical analysis and modeling techniques [20]. Modern signal processing, pattern recognition, and machine learning techniques provide the means to properly analyze and model the noisy, high-dimensional data sets generated by microarray experiments [21].

4.3 THE HUANG BREAST CANCER DATA SET

The global transcriptional profiles of 37 primary breast cancer tumor samples were measured using Affymetrix U95-AV-5 GeneChip microarrays [4]. Each microarray profiled the steady-state mRNA levels of 12,625 genes simultaneously in a single tumor sample. Of the 37 samples that were profiled, 19 were labeled as ‘‘negative’’ or low-risk samples and 18 as ‘‘positive’’ or high-risk samples based on microscopic

examination of lymph node samples obtained by axillary lymph node dissection. Among ER positive patients, the high-risk (or positive) clinical profile was represented by metastases involving 10 or more lymph nodes. The low-risk (or negative) profile was defined by node negative patients of age greater than 40 years with tumor size less than 2 cm. The main hypothesis for this experimental design asserts the existence of global gene expression patterns capable of discriminating between the high- and low-risk tumor samples.

Microarray data were acquired using protocols established by Affymetrix Corporation for the U95-AV-5 GeneChip. The amount of starting total RNA for each GeneChip hybridization was 20 μg. First-strand cDNA synthesis was generated using a T7-linked oligo-dT primer, followed by second-strand synthesis. An *in vitro* transcription reaction was performed to generate the cRNA containing biotinylated UTP and CTP, which was subsequently chemically fragmented at 95°C for 35 min. The fragmented, biotinylated cRNA was hybridized in MES buffer (2-[*N*-morpholino] ethansulfonic acid) containing 0.5 mg/mL acetylated bovine serum albumin to Affymetrix GeneChip HumanU95Av2 arrays at 45°C for 16 h, according to the Affymetrix protocol. The arrays contained probes that measured the expression of over 12,000 genes and ESTs. Arrays were washed and stained with streptavidin phycoerythrin (SAPE, Molecular Probes). Signal amplification was performed using a biotinylated antistreptavidin antibody (Vector Laboratories, Burlingame, CA) at 3 μg/mL. This was followed by a second staining with SAPE. Normal goat IgG (2 mg/mL) was used as a blocking agent.

Each hybridized GeneChip was scanned using an Affymetrix GeneChip scanner, and the expression value for each gene was calculated using the Affymetrix Microarray Analysis Suite (v5.0), computing the expression intensities in ‘‘signal’’ units defined by the software. Scaling factors were determined for each hybridization based on an arbitrary target intensity of 500. Scans were rejected if the scaling factor exceeded a factor of 25, resulting in only one reject. Files containing the computed signal intensity value for each probe cell on the arrays, files containing experimental and sample information, and files providing the signal intensity values for each probe set, as derived from the Affymetrix Microarray Analysis Suite (v5.0) software, were generated and posted on the Huang study Web site.

4.4 RESULTS OF THE HUANG STUDY

Using *k*-means clustering, SVD, and statistical tree models, Huang et al. discovered a gene expression signature based on 200 genes that was able to discriminate between high-risk and low-risk samples with 90 percent accuracy [4,5]. Moreover, they showed that many of the genes that defined the prognostic signature mapped to biological processes related to breast cancer. In particular, an interferon-mediated immune response was identified in the list of DE genes significantly changed in expression between the positive and negative sample groups of the experiment.

In brief, the data analysis employed by Huang et al. first removed genes with fold change less than two and maximum intensity less than nine on a $\log_2$ scale. This

filtering step resulted in a reduction in the number of genes available for downstream processing from 12,625 to 7030. *K*-means clustering was then applied to the filtered genes to obtain 496 gene clusters. Singular value decomposition was used to extract the first principal component of each gene cluster. This principal component was called the ''metagene" associated with the gene cluster. The 496 metagenes were presented as input to a classification tree, where the sample space is recursively partitioned into subsets that best fit the data based on a Bayesian measure of association between metagenes and a binary variable encoding the lymph node status of the samples [5]. Lists of genes were generated from the top four metagenes having the largest marginal Bayes' association. The list was extended by adding additional genes that are highly correlated with any one of the top four metagenes.

Metagenes were discovered that were highly associated with lymph node status. These discriminative metagenes were capable of predicting lymph node status in individual patients with about 90 percent accuracy using the classification tree model based on the microarray data. The metagenes also defined distinct groups of genes that participated in biological processes related to metastatic breast cancer. It was concluded that gene expression patterns can be used to accurately predict the lymph node status of a primary breast cancer tumor based solely on the gene expression patterns of the tumor itself [4].

4.5 GENOMIC SIGNAL PROCESSING

An important goal of bioinformatics and systems biology in cancer research is to improve the diagnosis, prognosis, and treatment of cancer through more accurate disease classification and patient stratification using quantitative techniques that take full advantage of the genome-wide data generated by new technologies such as DNA microarrays [10,22,23]. This comprehensive approach to understanding cancer allows for the design of therapeutic strategies that are targeted to the specific cancer subtypes that are unique to an individual patient. The hope is that a deeper understanding of the molecular heterogeneity of cancer could potentially improve the effectiveness of existing treatment regimens based on the ability to predict therapeutic response and adverse effects, as well as suggest new strategies based on the identification of new molecular targets susceptible to pharmacological intervention [9,23].

By genomic signal processing (GSP) we mean the identification, isolation, and extraction of information from high-dimensional data, such as that produced by DNA microarrays, that are useful for modeling and/or explaining observed changes in well-defined clinical or biological phenotypes. In this chapter, we describe a number of GSP techniques that in aggregate enable the minimally invasive prediction of distant changes in lymph node status based solely on the gene expression profile of the primary breast cancer tumor.

To facilitate the application of GSP techniques, we view the prediction of lymph node status as a problem in pattern recognition where the raw data are preprocessed, informative genes are identified, feature patterns are extracted from the expression profiles of these genes, and finally a pattern recognition (PR) model is formulated

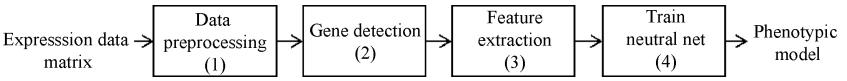

Figure 4-1 Modeling phenotypic variation using DNA microarrays. (1) Raw expression data matrix is preprocessed to remove systematic error and equalize noise. (2) Differentially expressed genes are selected from the preprocessed data matrix. (3) Feature patterns with reduced dimensionality and noise content are extracted from the data matrix of significant genes. (4) The extracted feature patterns are used to train a neural network to discriminate between phenotypic classes. The trained neural network constitutes a model of phenotypic variation defined in terms of gene expression.

based on the extracted feature patterns [24,25]. Figure 4-1 shows a high-level flowchart of the information processing chain used to formulate a NN model of breast cancer metastasis based on GSP features extracted from whole genome expression profiles. In brief, the following steps are involved: (1) the microarray data are normalized and equalized; (2) differentially expressed genes are detected; (3) feature patterns are extracted from the list of DE genes; (4) and a NN classifier is trained on the extracted feature patterns. An important step in the modeling process that is absent in Figure 4-1 is the objective assessment of predictive power of the resultant model using cross-validation techniques, which is discussed in Section 4.11.

4.6 THE EXPRESSION DATA MATRIX

Specifically, a $p \times n$ expression data matrix A_{raw} is formed where each of the n columns of A_{raw} represents the expression profile over p genes of a tumor sample. It follows that each of the p rows of A_{raw} represents the expression profile of a gene over the n samples of the microarray experiment [18]. We assume the n columns of A_{raw} are grouped so that the lymph node negative samples comprise first n_1 columns of A_{raw} for $j = 1,2, \ldots, n_1$ and the lymph node positive samples comprise the next n_2 columns of A_{raw} where $n = n_1 + n_2$.

Typically, $p \gg n$, (p much greater than n) where, for example, $p = 12{,}625$ and $n = 37$ for the Huang microarray data set. This situation is known as "Bellman's curse of dimensionality," which states that the number of samples needed to adequately model phenotypic variation grows exponentially with the number of input variables [24,26]. Hence, the Huang data matrix is mathematically ill posed for analysis using standard statistical approaches since the number of variables (genes) exceeds the number of equations (microarrays) by several orders of magnitude.

NN models based on a large number of input genes (and a relatively small number of samples) admit a large number of possible solutions that vary widely in terms of prediction performance, and hence, generalize badly from a finite set of training data to the general population that were unseen during training. Methods must be used to reduce the number of variables (i.e., dimensionality) without losing information that is relevant to solving the discrimination or prediction problem at hand [27,28]. Standard statistical techniques based on optimality arguments where the number of samples grow asymptotically without bound relative to the number of variables are inadequate

for the so-called "large p, small n" problems. Indeed, microarray experiments require statistical techniques based on asymptotics where the number of variables increases without bound relative to the number of samples [29]. Unfortunately, the statistical analysis of ill-posed problems is less well developed than for situations where the number of samples is plentiful and the number of variables small.

One approach is to use Bayesian statistics to constrain the space of possible solutions on a finite training set and automatically select parsimonious models that generalize to a larger population [21]. Another approach is to use signal processing techniques to select only highly informative features that reduce input space dimensionality, which in turn alleviates the negative impact of Bellman's curse on the ability of the derived model to generalize [30]. In this chapter, we describe methods more closely aligned to the latter approach where signal processing techniques are used to extract highly informative, low-dimensional features from expression data matrix. These feature patterns are then used to train NN classifiers that are capable of distinguishing benign breast cancer tumors from tumors that have spread to the axillary lymph nodes.

4.7 EXPERIMENTAL DESIGN

Global gene expression profiles are obtained using Affymetrix HU-95 GeneChip. Each hybridized GeneChip was "vectorized" into column vectors composed of 12,625 components, where each component represents the relative expression level of a single transcript on a given chip [4]. As described above, the vectorized chips were arranged to form the columns of a $12{,}625 \times 37$ expression data matrix A_{raw} of raw expression values where columns 1–19 represented the negative samples and columns 20–37 the positive samples. The data matrix A_{raw} was quantile normalized to obtain the preprocessed data matrix A_{nrm}.

A *sample response function* (SRF) for the Huang microarray experiment is a mapping $h{:}\{1,2,\ldots,n\} \rightarrow L$ defined on the columns of A_{nrm} where $L = \{-1,1\}$. Note that h reflects the phenotypic grouping of the samples such that $h(i) = -1$ for $1 \leq i \leq 19$ and $h(i) = -1$ for $20 \leq i \leq 37$. Note that h has the shape of a step function on the column indices of A_{nrm}. The ordered triple (A_{nrm}, L, h) represents the experimental design for the microarray experiment based on the Huang data set. Figure 4-2 visualizes the components of the microarray experiment (A_{nrm}, L, h). Here, Figure 4-2a is the step-like SRF defined by h for the microarray experiment, (A_{nrm}, L, h) and Figure 4-2b is a z-scored image of the data matrix $\log_2(A_{\text{nrm}})$ [31].

The fundamental hypothesis of microarray data analysis (FHMD) for (A_{nrm}, L, h) asserts the existence of a set of genes that are highly correlated with step function h shown in Figure 4-2a. For example, a numerical measure, t_g, can be computed for each gene g defined by

$$t_g = t(x_g, h) \equiv \frac{x_g^T h}{s_g} \tag{4-1}$$

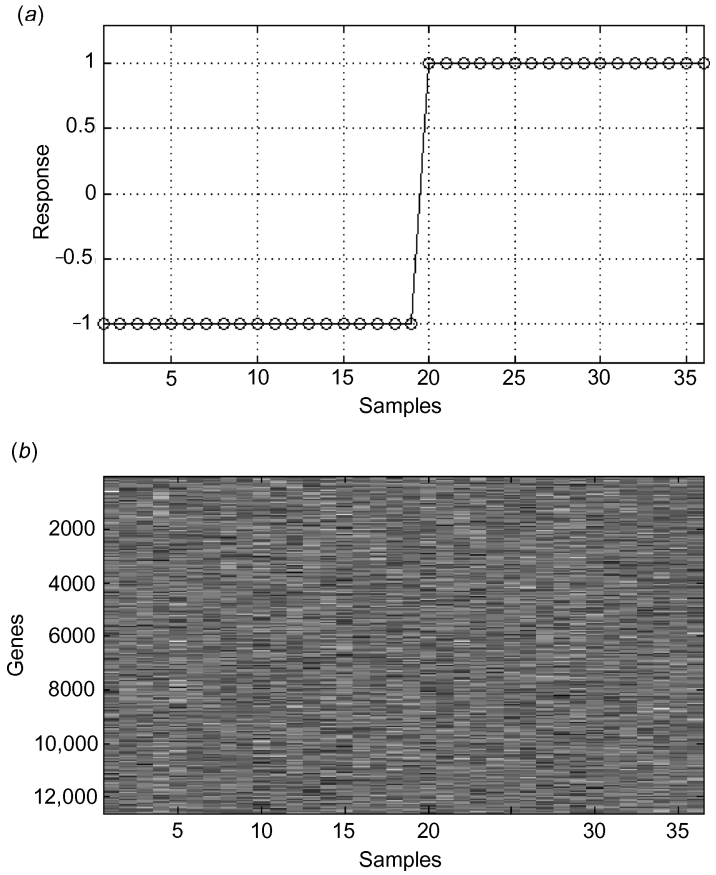

Figure 4-2 The experimental design for the Huang breast cancer microarray experiment. (a) Step-like sample response function h defined on the columns of A_{raw} that groups the columns of the data matrix into lymph node negative (columns 1–19) and positive (columns 20–36) sample groups. (b) The z-scored image of the $12{,}625 \times 36$ expression data matrix A_{raw} after quantile normalization and $\log_2$ transformation.

where (1) x_g is the expression profile of gene g over n samples; (2) s_g is the standard deviation of x_g; and (3) $x_g^T h$ is the correlation between x_g and h. Note that Equation 4-1 is equal to the correlation between x_g and h normalized by the standard deviation of x_g, which is also known as the t-score for g. The genes are ordered by absolute t-scores, and a subset of genes with the largest absolute t-scores is chosen based on some statistical threshold such as p-value or false discovery rate (FDR). The resulting list of genes is necessarily correlated with the unit step function h in accordance with Equation 4-1 and, in the this case, represents the genes that show the most consistent differential expression between the positive and negative samples of the Huang breast cancer data set. Such a set of genes is said to be differentially expressed between the two sample groups in accordance with the t-test.

4.8 DATA PREPROCESSING AND DATA QUALITY ASSESSMENT

The columns of A_{raw} are quantile normalized to facilitate comparison between the samples represented by the columns of A_{raw}, which results in the data matrix A_{nrm} [32]. Each entry of A_{nrm} is then $\log_2$ transformed to equalize variation over the entire range of expression values resulting in the $p \times n$ preprocessed data matrix $A_{\log 2}$ [20]. The preprocessed data matrices A_{nrm} and $A_{\log 2}$ form the basis for further downstream information processing depending on the algorithms used. For the purposes of this chapter, our focus will be on the normalized expression data matrix A_{nrm}. Note that quantile normalization essentially models and removes a low-frequency, correlated signal that corresponds to the systematic experimental error in the raw microarray data.

The primary goal of the preprocessing step is the removal of systematic nonrandom variation from the raw data to facilitate the comparison of gene expression across multiple chips. The quantile normalization procedure can be described in two steps:

- Create a mapping between ranks and expression values; that is, for rank k, find the n genes, one per array, that have rank k in terms of gene expression and compute their average expression over the n samples;
- For each gene on each array, replace the measured expression value with the rank-average expression for that gene.

Note that quantile normalization is an aggressive strategy that produces identical distributions for each array. On the contrary, quantile normalization is extremely fast, since it only requires a single sort of the data matrix, a computation of means across sorted rows, and a single pass through the data [33]. Note that other normalization schemes exist that employ nonparametric modeling techniques such as locally weighted polynomial regression (lowess) to characterize the systematic error in raw microarray data. One such method identifies genes that are invariant in terms of variation between normal and disease sample classes and models the low-frequency, correlated signal in these invariant genes using a lowess-type smoother. The resulting error model is then used to correct all the raw data for systematic error. Normalization based on lowess smoothing of invariant genes tends to be a less aggressive a procedure than quantile normalization.

Another important preprocessing step is the $\log_2$ transformation of the data matrix A_{nrm} to decouple variation in fold change from expression level. This decoupling makes the data appear more bell shaped and hence improves the performance of downstream statistical analysis algorithms designed to detect DE genes. Indeed, raw microarray data have essentially a log-normal distribution, which implies that the $\log_2$ transform of the data should be more or less normally distributed or at least unimodal [20]. Other more powerful variance stabilization methods have been proposed that view microarray data as having a normal distribution at low expression levels, a log-normal model at high intensities, and a mixture of both at intermediate intensities. The impact of such mixture models on classification and prediction performance is currently being evaluated.

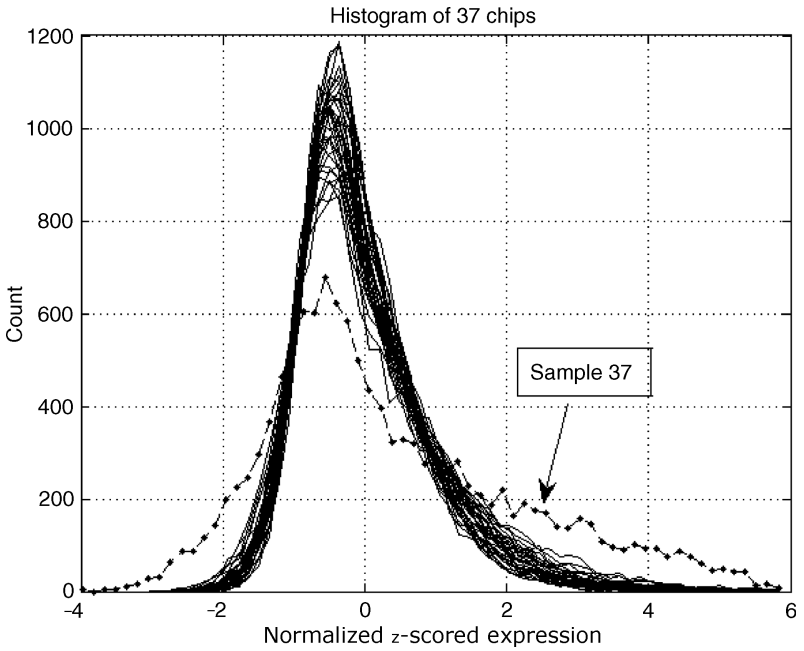

Figure 4-3 Histogram plots for quantile normalized, $\log_2$ transformed, *z*-scored microarray data from the Huang breast cancer data set. Note that most of the samples have similar histograms, while sample 37 is an outlier in distribution. Therefore, we removed sample 37 from the MANINI analysis since an outlier of that magnitude would distort the final performance results.

The chips that comprise the Huang microarray experiment are assessed for data quality using a number of standard statistical techniques. For example, histograms of all 37 columns of the quantile normalized data matrix A_{nrm} after z-scoring are shown in Figure 4-3. Note that sample 37 is clearly an "outlier" in distribution when compared to the other 36 chips of the experiment. Moreover, pairwise correlation analysis of the raw and normalized data indicates that sample 37 has relatively low correlation with the other 36 chips. These results suggest that the expression values for sample 37 were corrupted at some point in the data acquisition process. Hence, sample 37 was removed from this study, although we note that sample 37 was retained by Huang et al. in their study.

4.9 THE MODELING OF PHENOTYPIC VARIATION

The modeling of phenotypic variation in terms of gene expression is a pattern recognition problem that can be solved by mapping gene expression patterns directly to phenotypic states using NN classifiers [21]. Such models are known as discriminant classifiers. Note that it is not necessary to delineate the biological mechanism

underlying the observed variation in disease phenotypes since the predictive model is implemented based solely on the *association* between gene expression and phenotype [15]. The proposed formulation of a discriminant pattern recognition model for the prediction of lymph node status involves a four-step information processing chain shown in Figure 4-1. Each step of the processing chain requires the use of GSP techniques. Data preprocessing was described in the previous section. Details on the remaining steps of the processing chain shown in Figure 4-1 are described below.

The proposed GSP processing chain includes the following signal processing components: (1) Microarray Analysis of Intensities and Ratios (MANINI) detection algorithm for identifying DE genes, (2) pathway compression for data reduction, (3) wavelet transformation for the separation of signal from noise, (4) singular value decomposition for further dimensionality reduction and filtering, and (5) neural networks for encoding the information contained in features derived from microarray data using GSP techniques.

The signal processing steps outlined above can be combined in different ways leading to different information processing algorithms. For example, the SVD of the wavelet transformed data is known as wavelet/SVD (WSVD) signal processing. Alternatively, the SVD of the wavelet transform of the data matrix of genes confined to a specific pathway specified by Ingenuity Pathway Analysis (IPA) is denoted by WSVD/IPA signal processing. The following sections describe the different signal processing components and how they are combined to form analysis pipelines that lead to robust predictors of lymph node status based solely on the gene expression profiles of the primary breast cancer tumor.

4.9.1 The MANINI Detection Algorithm

An alternative to the *t*-test for the supervised detection of genes highly correlated to a given SRF is the so-called MANINI detection algorithm. The MANINI detector was specially designed to handle small numbers of samples often encountered in real-world case/control microarray experiments, since in the absence of a large number of chips, one is hard-pressed to do better than use fold change to detect DE genes [34]. MANINI was also designed to detect DE genes with expression profiles that are inconsistent or highly variable over the samples of the experiment [35]. There is a growing trend in microarray data analysis toward detecting genes in heterogeneous samples (e.g., tumor samples) with expression patterns that may be too inconsistent for more conventional detector designs such as the *t*-test [35,36].

For example, assume that a single signaling pathway is modulated between two biological conditions. Due to sample heterogeneity and biological variation, it may be that different components of the pathway are DE for different samples. In this case, individual genes that participate in the pathway would be difficult to detect using the *t*-test since the associated expression profiles would be highly variable over the samples of the experiment. The MANINI detector on the other hand would be able to select such genes for downstream ontological and pathway analysis, where different

ensembles of functionally related genes are assigned to the common pathway to which they belong [7,37].

Let x and y be $p \times 1$ column vectors representing the geometric averages of the control and disease chips, respectively. Let

$$M = \log 2(y) - \log 2(x) = \log 2\left(\frac{y}{x}\right)$$

and

$$A = \tfrac{1}{2}[\log 2(y) + \log 2(x)] = \log 2\left(\sqrt{xy}\right)$$

Figure 4-4 shows the Minus–Add (MA) scatter plot of M versus A where fold change is plotted versus average expression in $\log_2$–$\log_2$ space.

We note three things about the MA scatter plot: (1) the MA plot can be viewed as a visualization of differential expression; (2) genes located on the periphery of the MA data cloud are likely to be DE; and (3) the vertical variation of the MA data cloud is a function of expression level. This suggests a strategy for detecting DE genes by selecting only those genes that "live" on the edge of the MA data

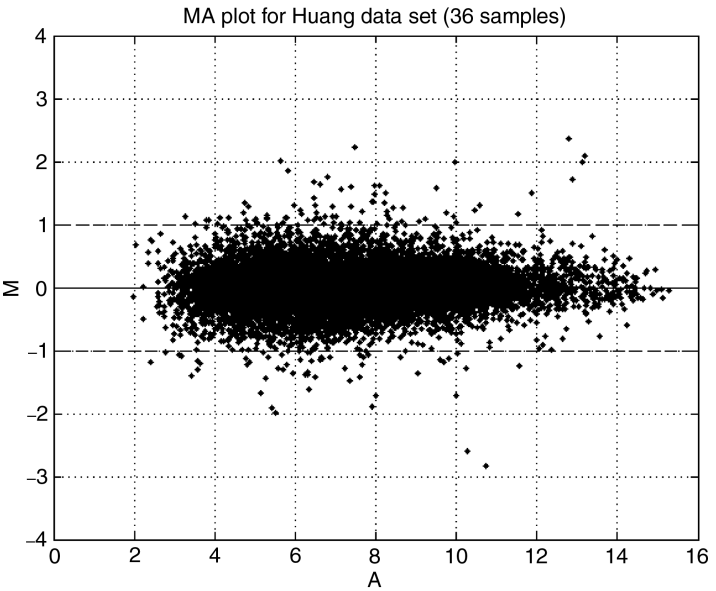

Figure 4-4 Minus–Add scatter plot for Huang breast cancer microarray data set. Each point represents the geometric average of expression (A-axis) and fold change (M-axis) for a gene in $\log_2$–$\log_2$ space. Horizontal dotted lines at $M = \pm 1$ represent constant twofold change in expression between positive and negative samples. Note the dependence of the variance of M on the intensity of A.

cloud based on a threshold that adapts to the spread of the cloud as a function of expression level [32,38]. Note that using standard twofold change as a constant threshold for differential expression over the entire range of expression values (as represented by the horizontal dashed lines in Fig. 4-4) is clearly inappropriate since too many genes are called DE at lower expression levels and too few genes are called DE at higher expression levels. A better strategy would be to adaptively threshold the genes of the MA scatter plot based on expression level.

The MANINI detector implements this idea by "binning" the horizontal axis of the MA plot into k quantiles. Each bin contains about the same number of genes that have similar expression intensities. This quantization scheme also implies that all genes in a bin have about the same degree of variation since variation is a function of intensity. For a given bin containing m genes, we assume the signal model $y_i = s_i + \eta$ for $i = 1,2, \ldots, m$. Here, s_i is the true expression level of the ith gene, and $\eta \sim N(0,\sigma)$ is normally distributed random variable with mean zero and known variance σ^2. Empirical studies show that this is a reasonable assumption for wide range of real-world data sets. It follows that differential expression within the bin can be modeled as a m-dimensional random vector, $y = [y_i,y_2, \ldots, y_m]^T$, where $E(y) = [s_1,s_2, \ldots, s_m]^T$ is sparse [29]. By sparse it is meant that most of the components of the true signal vector $E(y) = [s_1,s_2, \ldots, s_m]^T$ are zero.

In the field of wavelet denoising, Donoho and Johnstone showed that a signal contaminated by zero-mean Gaussian noise can be optimally filtered by thresholding the wavelet coefficients of the noisy signal [39]. The wavelet transform of a noisy signal localizes the information content of a signal simultaneously over time and scale. In this case, high-frequency noise is usually confined to the high-resolution scales and low-frequency coherent signal is concentrated in the low-resolution scales. Donoho and Johnstone found that by simply thresholding the higher resolution wavelet coefficients of the noisy signal and then applying the inverse wavelet transform, one can optimally estimate the true underlying signal assuming that it is sparse [29]. Thresholding in this manner to estimate a signal embedded in noise is called testimation. Note that the wavelet coefficients at a given scale of resolution form a Gaussian random vector where only a few of the coefficients are different from zero; that is, the wavelet coefficients at each scale form a sparse random vector.

Based on the properties of the MA scatter plot, the $\log_2$ ratios within an expression bin of a MA plot can be viewed as a sparse Gaussian random vector where only a small number of genes within the bin are truly DE. This observation suggests the application of the Donoho–Johnstone (DJ) universal threshold directly to the $\log_2$ ratios of a given intensity bin using

$$\hat{s}_i = \begin{cases} y_i & \text{if } |y_i| > \sigma\sqrt{2(1-\beta)\log(m)} \\ 0 & \text{otherwise} \end{cases} \tag{4-2}$$

to select those genes with "true" nonzero differential expression for $i = 1,2, \ldots, m$, where $\hat{s}_i$ is an estimate of the ith component of the true signal s and $0 < \beta < 1$ [29,39].

Note that $1-\beta$ represents a measure of the "sparseness" of y. It can be shown that this disarmingly simple procedure is asymptotically optimal in a statistical sense (i.e., it minimizes the maximum expected risk) as m grows without bound and that its application to a noisy high-dimensional data vector amounts to a Bonferroni-type correction for multiple comparisons [29].

Note the optimality of the estimate $\hat{s}$ depends on the number of variables m (or genes) growing without bound. This is in sharp contrast to the situation in classical statistics where the number of samples is assumed to grow without bound [18]. Hence, Equation 4-2 actually becomes more accurate when the number of genes is large, provided the random vector remains sparse. This is exactly the situation for most whole genome expression profiling studies and precisely the opposite of what is required for standard statistical algorithms to work properly. Hence, the MANINI detection algorithm takes advantage of the large number of genes interrogated in a typical microarray experiment by binning the genes into subgroups containing equal numbers of genes with similar expression levels. Since the total number of genes is large, each subgroup will have enough genes for the DJ universal threshold to work (for that particular subgroup). For example, the Affymetrix U133 Plus 2.0 GeneChip uses over 54,000 probe sets to interrogate over 47,000 transcripts that represent approximately 38,000 genes and gene variants. Binning the horizontal axis of the MA plot into 51 quantiles results in 50 subgroups of genes where each subgroup contains 1094 measurements with similar expression levels. Note also that each bin of the MA plot contains only a few genes that are truly DE; that is, the $\log_2$ ratios in each bin contain a sparse signal for DE. This allows the application of the DJ threshold to most microarray experiments where the signal for DE is sparse both locally and globally.

Note that Equation 4-2 was implemented for each bin using the mean absolute deviation (MAD) statistic in place of σ to provide a robust estimate of the variation within the bin. Genes that exceeded the Donoho–Johnstone threshold for the bin were called DE. The union of all genes called DE over all bins of the MA scatter plot resulted in a list of genes that are globally DE for the microarray experiment [31].

The MANINI detection algorithm was used to analyze the $12{,}625 \times 26$ $\log_2$ transformed, quantile normalized expression data matrix denoted by A. We summarize the MANINI results in Figure 4-5. Differentially upregulated genes are marked by up-triangles, differentially downregulated genes are marked by down-triangles, and genes unchanged in expression are represented by points. The black dashed lines located at $M = \pm 1$ represent constant thresholds for a twofold change in expression in either the up ($M = 1$) or down ($M = -1$) direction. The quantiles of A-axis are represented by dark and gray vertical bands shown in the body of the MA plot. The MANINI algorithm calls a gene within a given quantile, or bin, differentially expressed if its absolute M-value exceeds the DJ noise-adjusted threshold for that bin. The union of genes called DE over all bins represents the global signal for DE detected by MANINI.

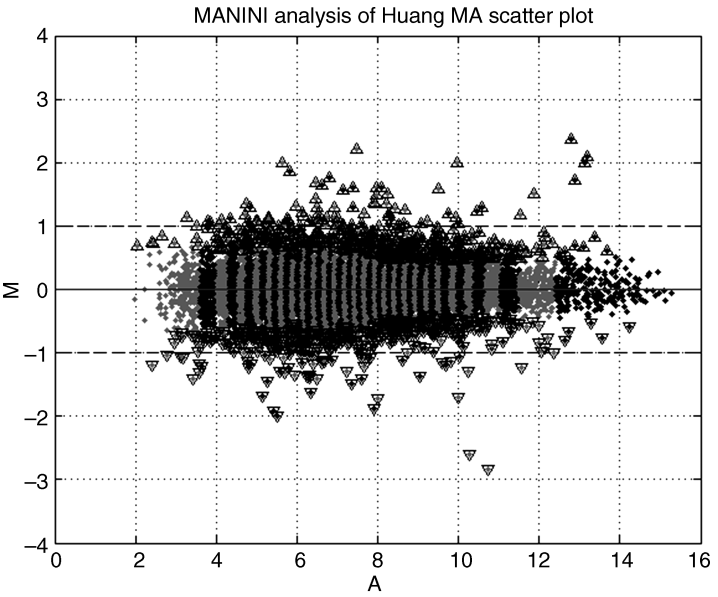

Figure 4-5 MA plot summarizing the results of a MANINI analysis of the Huang breast cancer microarray data set. Each point of the RI plot is a gene represented by average expression and fold change. Genes represented by up-triangles were called significantly upregulated by MANINI, while genes represented by the down-triangles were called significantly downregulated. The vertical bands in the body of the MA plot represent the 50 quantiles used to segment the *A* axis into disjoint bins containing approximately the same number of genes. Each bin represents a separate and distinct DE detection problem for genes that have comparable expression levels.

In Figure 4-6, we show scaled images of the data matrices U and D composed of genes that MANINI called significantly up- and downregulated, respectively. The rows of U shown in Figure 4-6a represents the samples of the experiment clustered by similarity of their expression profiles over genes called significantly upregulated by MANINI. A similar interpretation applies to the rows of D shown in Figure 4-6a. The results of the cluster analysis are summarized by a dendrogram shown on the left side of U and D. Moreover, the cluster structure over all samples is shown on the right side of Figure 4-6a and b, where the negative samples are labeled 1–19 and positive samples 20–36. Both data matrices U and D are quantile normalized, $\log_2$ transformed, and z-scored by rows.

Note that in Figure 4-6a and b, the samples to a large extent segregate by lymph node status with some erroneous classifications that are probably due to the inclusion of genes that are falsely called DE by the MANINI detector. This suggests that we may be able to identify a subset of genes that are able to do a better job of discriminating between positive and negative breast cancer tumor samples based on gene expression. We also note that although the MANINI detector is designed to detect genes with expression profiles that conform to a step-like response, it is also less sensitive to

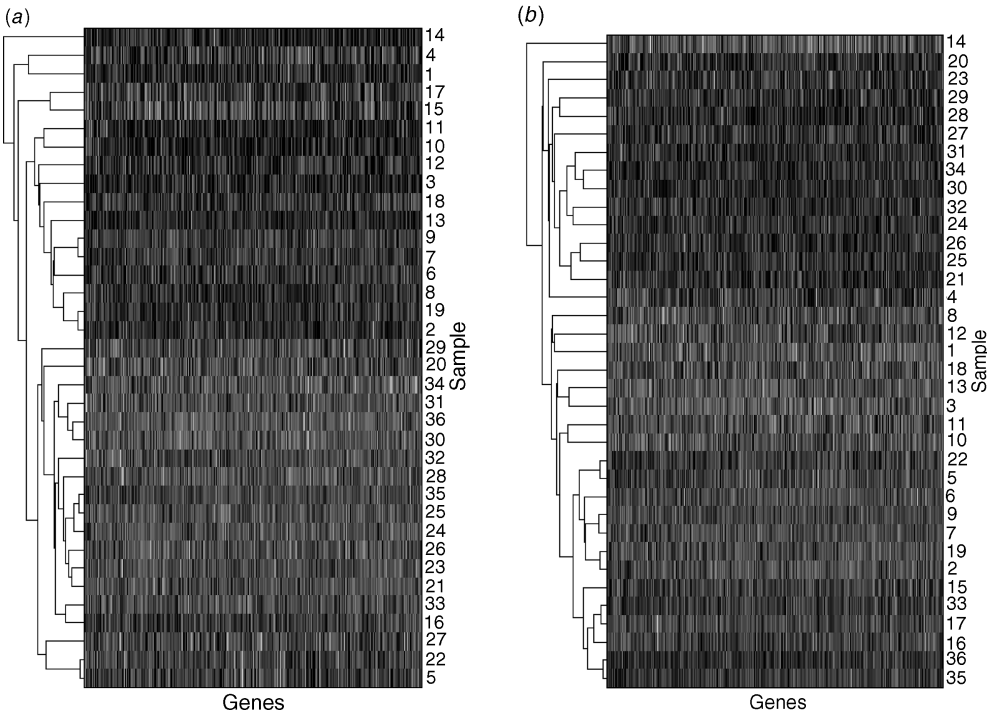

Figure 4-6 Scaled images of the expression data matrices for up- and downregulated DE genes called by the MANINI detector. The rows of each data matrix are hierarchically clustered by gene expression profile over the samples.(a) Data matrix composed of DE genes called upregulated by the MANINI algorithm. (b) Data matrix composed of DE genes called downregulated by the MANINI algorithm. Note that each set of genes approximately segregates the samples into two distinct clusters containing positive and negative breast cancer tumor samples.

deviations from the ideal response than standard statistics such as the two-sample t-score and hence will call a broader range of expression patterns as statistically significant.

4.9.2 MANINI and Signal Detection Theory

Let A_{nrm} be a $p \times n$ normalized expression data matrix and let r represent a one-to-one mapping of the column indices $\{1,2, \ldots, n\}$ of A_{nrm} to $\{-1,1\}$ defined by

$$r(i) = \begin{cases} -1 & \text{if the } i\text{th sample is a control} \\ 1 & \text{if the } i\text{th sample is a case sample} \end{cases}$$

Here, the function r is called a response function for the experiment. Figure 4-7 shows a response function equal to the unit step function h defined for a 64-chip microarray

experiment by

$$h(i) = \begin{cases} -1 & \text{if } i = 1, 2, \ldots, n_1 \\ 1 & \text{if } i = n_1 + 1, n_1 + 2, \ldots, n_1 + n_2 \end{cases}$$

where $n_1 + n_2 = n$. Note samples 1–32 are controls and samples 33–64 are cases. Let g_i denote the row expression profile of the ith gene of A_{nrm} for $i = 1,2, \ldots, p$. Then the ith gene is said to be differentially upregulated in the treated group if g_i is positively correlated with the step function h. Conversely, the ith gene is differentially downregulated in the treated group if g_i is negatively correlated

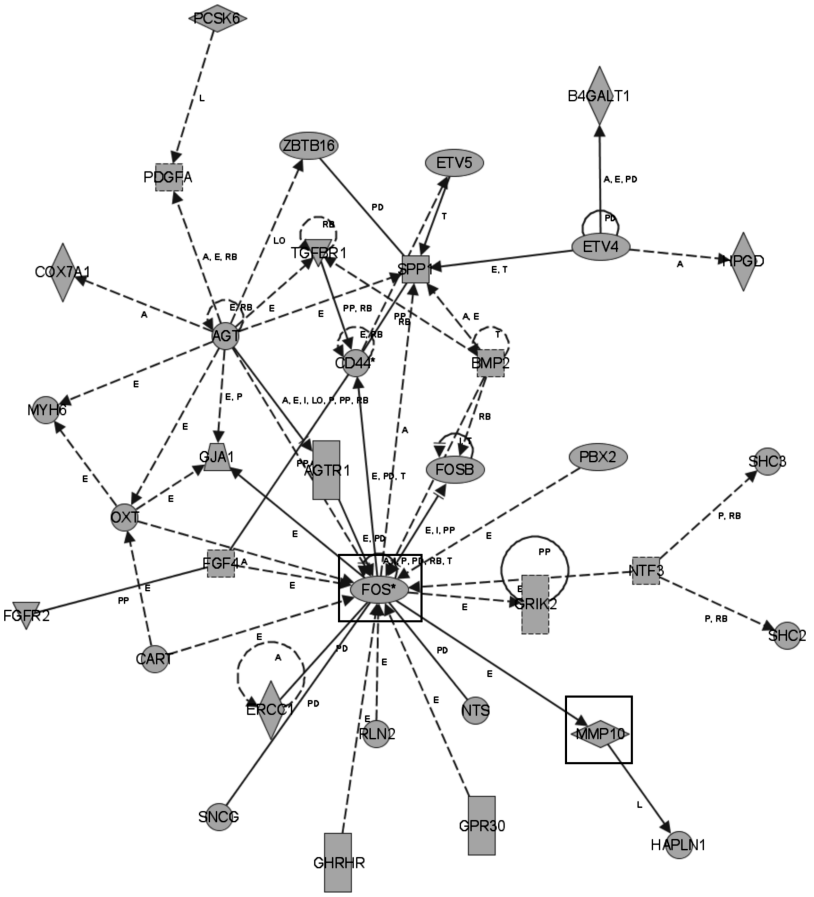

Figure 4-7 Top IPA network for downregulated genes. The network contains 35 genes with a score of 55 (p-value ~1.0E-55). An inferred function of the network is *Cancer*. The highly connected hub gene FOS has been implicated in the regulation of cell proliferation, differentiation and apoptosis, and MMP10 has also been implicated in tumor metastases.

with h. We note that transcript degradation in the control and/or case classes can also generate step-like gene expression profiles. For example, degradation of message for the ith gene in only the control samples can falsely produce a gene expression profile that suggests upregulation of the gene in question in the treated samples. We will assume that strict quality control of sample preparation and hybridization protocols will reduce message degradation and experimental variation to a minimum.

Let $y = [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^n$ be a noisy gene expression profile where each y_i represents the measured expression level of a gene in the ith sample for $i = 1,2,\ldots,n$. In the context of statistical hypothesis testing, let

$$H_0 : y = \eta$$

and

$$H_1 : y = h + \eta$$

where $\eta \in \mathbb{R}^n$ is a Gaussian, zero-mean, independent, identically distributed random vector representing noise in the data. Then the Neyman–Pearson Lemma states that uniformly the most powerful test for H_0 versus H_1 is defined by [40]

$$\delta_h(y) = \begin{cases} 1 & \text{if} \quad f(y) > \tau \\ \gamma & \text{if} \quad f(y) > \tau \\ 0 & \text{if} \quad f(y) > \tau \end{cases} \tag{4-3}$$

where

$$f(y) = \log[L(y)] = \log\left(\frac{p_1}{p_0}\right) = \log\left[\frac{p_\eta(y-h)}{p_n(y)}\right] = \sum_{k=1}^{n} h_k y_k$$

By Equation 4-3, the test δ_h is called the correlation detector for the known signal h in the noisy signal y. The step response h is called a template signal for δ_h. In other words, the best strategy for detecting the presence of the known template signal h in a noisy gene expression profile y is to correlate the two together [40,41]. Here, a large absolute correlation implies that h is present in y, otherwise h is not present. The key point here is to determine the threshold τ on f that optimizes the balance between sensitivity and the false positive rate defined by (1 – specificity). The response function need not be confined to be the unit step function, and in fact Equation 4-3 is quite general and holds for arbitrary y and r. But a step-like response is the standard template function for detecting what is commonly known as differential expression between the case and control classes of a microarray experiment. This step-like response represents a large and consistent difference in expression between the case and control samples of the experiment.

Note that fold change can be viewed as a correlation detector for the unit step function h in the noisy gene expression profile y. Indeed, we have

$$\begin{aligned} h^T y &= \sum_i h_i y_i = \sum_{\text{controls}} h_i y_i + \sum_{\text{cases}} h_i y_i = \sum_{\text{controls}} -y_i + \sum_{\text{cases}} y_i \\ &= \left(\frac{n}{2}\right)\left[\left(\frac{2}{n}\right)\sum_{\text{cases}} \log_2(x_i) - \left(\frac{2}{n}\right)\sum_{\text{controls}} \log_2(x_i)\right] \\ &= \left(\frac{n}{2}\right)\left[\log_2\left(\sqrt[\frac{n}{2}]{\sum_{\text{cases}} x_i}\right) - \log_2\left(\sqrt[\frac{n}{2}]{\sum_{\text{controls}} x_i}\right)\right] \\ &= \left(\frac{n}{2}\right)\log_2\left[\frac{\left(\sum_{\text{cases}} x_i\right)^{n/2}}{\left(\sum_{\text{controls}} x_i\right)^{n/2}}\right] = \left(\frac{n}{2}\right)\log_2\left(\frac{\text{geometric average of cases}}{\text{geometric average of controls}}\right) \\ &= \left(\frac{n}{2}\right)\log_2(\text{fold change}) \end{aligned}$$

and hence, fold change is an correlation detector of the unit step response in noisy data. Since the t-test for the difference in mean expression between two sample groups is

$$t(y) = \frac{\left(\frac{2}{n}\right)\sum_{\text{cases}} \log_2(x_i) - \left(\frac{2}{n}\right)\sum_{\text{controls}} \log_2(x_i)}{s} = \left(\frac{2}{ns}\right)\log_2(\text{fold change})$$

where s is the pooled sample standard deviation of y, it follows that the t-test is simply log fold change penalized for "within-group" variations in fold change through the estimated value for s.

Note that if the true response r deviates significantly from the step function h, then both fold change and the t-test become suboptimal tests for differential expression (as defined by h.) Note however, that the t-test is further penalized for large variations in fold change through s, that is, the t-test is biased toward step-like signals that are strong and consistent within each two-sample groups. Hence, in situations where the true response of a gene is highly variable over the samples, as in tumor samples with heterogeneous composition, the t-test will fail to detect these genes [42]. Also, genes that are up- or downregulated on only a fraction of the case samples may not be detected [36]. On the other hand, MANINI will detect genes that are modulated intermittently across the samples of the experiment since fold change by itself is not penalized for excessive variation. Finally, note that statistical validation of the resulting gene list is deferred until ontological (PANTHER) or pathway analysis (INGENUITY) can be conducted to determine the statistically significant functional categories and pathways contained in the gene lists. A gene is then called significant if it is contained in a significant functional category or pathway [31].

In summary, many genes have response profiles that deviate significantly from the step function due to intermittent up- or downregulation across the samples of the microarray experiment [43]. Hence, such genes will remain undetected by standard *t*-like tests that are designed to detect a consistent and strong step-like change in expression across the samples of the experiment. The MANINI detection algorithm attempts to circumvent this problem by selecting genes that exhibit high fold change relative to a noise-adjusted threshold that varies with expression level. This detection strategy exploits the observed relationship between variation in fold change and expression level in microarray data. Hence, a gene is not penalized for high variation so long as its average fold change exceeds a "universal" detection threshold that adapts to the noise background for the gene. Significant genes are subsequently defined as those genes that are contained in significant functional categories that signaling pathways that are contained in the gene lists as identified using IPA and Onto-Express [37].

4.9.3 Pathway Compression

We assume that the overrepresentation of known cancer-related pathways (as explained below) in the list of DE genes derived from the Huang breast cancer data set represents coherent structure that characterizes the underlying biology of lymph node positive breast cancer tumors in terms of gene expression. Conversely, we assume the absence of such coherent structure suggests that the gene list has little in common with what is known about gene function and is essentially composed of randomly selected genes representing mostly noise. We target genes contained in overrepresented pathways as means of "drilling down" to those genes that are at once the most biologically relevant and discriminative between positive and negative breast cancer tumors. This idea is called pathway compression since it serves to reduce the dimensionality of the resulting gene expression signature that will be used to train a NN model for classifying breast cancer tumors as lymph node negative or positive.

IPA was used to identify the biological networks that were perturbed in the Huang lymph node positive breast cancer samples in the context of what is currently known about mammalian biology derived from basic and clinical research [15,44]. Research findings presented in peer-reviewed scientific publications are manually encoded into a comprehensive knowledge base of gene function and gene–gene interactions. The IPA knowledge base contains over 200,000 full text scientific articles, a gene ontology of more than 9800 human, 7900 mouse, and 5000 rat genes that were manually curated and parsed from MEDLINE abstracts. A global interaction network of direct physical, transcriptional, and enzymatic interactions observed between mammalian orthologues as described in the literature—the so-called global "interactome"—was overlayed on the gene ontology. The resulting global interactome contained molecular interactions involving over 8000 orthologues with a high degree of connectivity. On average, individual genes have 11.5 interaction partners, of which 7.2 represent direct physical interactions.

Every gene interaction in an IPA network is supported by published articles. Furthermore, the original literature detailing the genetic interactions can be accessed

to further examine and verify the findings. The global interactome provides a framework for structuring existing knowledge regarding mammalian biology and enables the objective validation of experimental data in the context of known genome-wide interactions to identify significant functional pathways. This method is applicable to data of high-throughput platforms such as microarray expression profiling, polymorphism analysis, and proteomics.

Significant pathways contained in MANINI-derived gene lists were identified by first overlaying the genes identified as DE onto the global interactome. Focus genes were then identified as those genes having direct interactions with other MANINI-significant genes in the database. The specificity of connections for each focus gene was calculated by the percentage of its connections to other significant genes. The initiation and growth of pathways proceeded from genes with the highest specificity of connections, where each pathway had a maximum of 35 genes. Pathways of highly interconnected genes were identified by statistical likelihood based on the following formula:

$$\text{IPA\_Score} = -\log_{10}\left[1-\sum_{i=0}^{f-1}\frac{C(F,i)C(N-F,s-i))}{C(N,s)}\right]$$

where $C(n,k)$ is the binomial coefficient, N is the number of genes in the global interactome, F are the number of significant genes detected by MANINI, and s is the number of genes in the inferred pathway of which f are focus genes. Depending on the data set, pathways with a score greater than 5 (p-value $<$1.0E-05) are considered significant.

IPA was used to identify biologically significant pathways contained in the gene list derived by MANINI for the Huang breast cancer data set. A summary of an IPA analysis of 413 downregulated DE genes selected by MANINI is shown in Table 4-1. The list is composed of pathways that were found to be statistically overrepresented in the gene list based on gene function and gene–gene interactions contained in the IPA knowledge base. The networks were rank ordered by IPA significance score and gene descriptions, and the number of focus genes for each network was also provided. Only networks derived from the downregulated genes were targeted since they resulted in the most robust NN models. Note that the top two IPA networks have p-values on the order of 10^{-55}. Figure 4-7 shows a diagram of the top interaction network from the list (*dnet1*).

The network diagram for *dnet1* in Figure 4-7 details the internal interactions between the 35 genes that are contained in the network. Here, each node of the diagram represents a gene and each edge connecting two genes represents a documented interaction between them. We selected only the genes contained in *dnet1* for further downstream information processing. IPA pathway analysis can be viewed as a feature selection procedure where the genes in a significant IPA pathway are used as features for classifying the samples of the experiment. This gene selection process is known as pathway compression [31]. In fact, we show in Section 4.11 that genes contained in *dnet1* are able to accurately distinguish between the positive and negative samples of the Huang breast cancer data set when analyzed using WSVD signal processing and modeled using neural networks. Note that diseases and biological processes

Table 4-1 IPA analysis of genes called downregulated by MANINI

Rank	IPA Source	Focus Genes	Top Functions
1	55	35	Gene expression, cell-to-cell signaling and interaction, cancer
2	55	35	Gene expression, dermatological diseases and conditions, genetic disorder
3	16	16	Cellular development, cellular growth and proliferation, hematological system development and function
4	14	15	Cellular compromise, Dermatological diseases and conditions, gastrointestinal disease
5	14	15	Inflammatory disease, viral function, immunological disease
6	14	15	Protein synthesis, lipid metabolism, small molecule biochemistry
7	13	14	Cell signaling, cancer, cell death
8	13	14	Nervous system development and function, organ development, cancer
9	13	14	Organismal development, lipid metabolism, molecular transport
10	13	14	Carbohydrate metabolism, molecular transport, small molecule biochemistry
11	13	14	Nervous system development and function, cell-to-cell signaling and interaction, neurological disease
12	11	13	Cellular growth and proliferation, hair and skin development and function, cell signaling
13	11	13	Viral function, gene expression, cell cycle
14	10	12	Cellular movement, connective tissue development and function, cell cycle
15	10	12	Energy production, nucleic acid metabolism, small molecule biochemistry
16	9	11	Cell cycle, cellular assembly and organization, DNA replication, recombination, and repair

Each row represents a significant IPA gene network. Note the networks are ordered by *p*-value. The top network, *dnet*, has a *p*-value ~1.0E-55 and contains 35 genes from the MANINI gene list. An inferred function for *dnet1* based on the function of genes contained in the network includes *Cancer.*

associated with the gene network *dnet1* include *Cancer*, *Cell-to-Cell Signaling* and *Gene Expression*. Table 4-2 shows a list of the 35 genes contained in *dnet*1. Although, many cancer-related genes such as FOS and MMP10 are included in the *dnet1* gene list, and the network topology of *dnet1* suggests specific biological mechanisms that may have relevance to metastatic breast cancer, we are primarily concerned with how well the *dnet1* genes are able to discriminate between positive and negative breast cancer tumors, ignoring for now the underlying biology.

4.9.4 Continuous Wavelet Transform

Wavelet signal processing analyzes a noisy signal, for example, the expression profile of a gene over a range of scales using wavelets of different locations and time durations.

Table 4-2 Gene list for IPA network *dnet1*

Affy Tag	Name	Description
684_at	AGT	Angiotensinogen (serpin petidase inhibitor, clade A, member 8)
37983_at	AGTR1	Angiotensin II receptor, type 1
40960_at	B4GALT1	UDP-Gal betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 1
40367_at	BMP2	Bone morphogenetic protein 2
35457_at	CART	CART prepropeptide
2036_s_at	CD44	CD44 molecule (Indian blood group)
39031_at	COX7A1	Cytochrome *c* oxidase subunit Vlla polypeptide 1 (muscle)
1878_g_at	ERCC1	excision repair cross-complementing rodent repair deficiency, complementation group 1
2084_s_at	ETV4	ets variant gene 4 (E1A enhancer binding protein, E1AF)
34818_at	ETV5	ets variant gene 5 (ets-related molecule)
1408_at	FGF4	Fibroblast growth factor 4 (heparin secretory transforming protein 1, Kaposi sarcoma oncogen
1363_at	FGFR2	Fibroblast growth factor receptor 2 (bacteria-expressed kinase, keratinocyte growth factor receptor
2094_s_at	FOS	v-fos FBJ murine osteosarcoma viral oncogene homologue
36669_at	FOSB	FBJ murine osteosarcoma viral oncogene homologue B
32383_at	GHRHR	Growth hormone releasing hormone receptor
32531_at	GJA1	Gap junction protein, alpha 1, 43 kDa (connexin 43)
37447_at	GPR30	G-protein-coupled receptor 30
39573_at	GRIK2	Glutamate receptor, ionotropic, kainate 2
39618_at	HAPLN1	Hyaluronan and proteoglycan link protein 1
32570_at	HPGD	Hydroxyprostaglandin dehydrogenase 15-(NAD)
1006_at	MMP10	Matrix metallopeptidase 10 (stromelysin 2)
38602_at	MYH6	Myosin, heavy chain 6, cardiac muscle, alpha (cardimyopathy, hypertrophic 1)
35041_at	NTF3	Neurotrophin 3
33998_at	NTS	Neurotensin
32472_at	OXT	Oxytocin, prepro-(neurophysin I)
38295_at	PBX2	Pre-B-cell leukemia transcription factor 2
32001_s_at	PCSK6	Proprotein convertase subtilisin/kexin type 6
35703_at	PDGFA	Platelet-derived growth factor alpha polypeptide
31732_at	RLN2	Relaxin 2
35622_at	SHC2	SHC (Src homology 2 domain containing) transforming protein 2
1511_at	SHC3	SHC (Src homology 2 domain containing) transforming protein 3
36555_at	SNCG	Synuclein, gamma (breast cancer-specific protein 1)
34342_s_at	SPP1	Secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1)
32903_at	TGFBR1	Transforming growth factor, beta receptor I (activin A receptor type II-like kinase, 53 kDa)
39681_at	ZBTB16	Zinc finger and BTB domain containing 16

Shown are Affymetrix gene tags, gene name, and gene description. Cancer-related genes include FOS, MMP10, and FGF4. The interaction structure of these genes is shown in Figure 4-7.

In other words, the wavelet transform provides a *timescale* decomposition of a signal of interest [45]. Here, the term "time" is used loosely referring to some agreed-upon sequential ordering of multiple measurements (e.g., gene expression values) that may or may not reflect an actual temporal ordering. The main point is that all samples have their components ordered in the same way.

An underlying assumption of wavelet signal processing is that coherent signal and random noise "live" at different scales of resolution and hence are often well separated after wavelet transformation. Moreover, noise in real-world data sets are often better "equalized" after wavelet transformation, thus making the distinction between signal and noise even more pronounced in the wavelet transform domain. Indeed, it can be shown that the wavelet transform "diagonalizes" a scale-invariant signal in much the same way that the Fourier transform diagonalizes time-invariant signals. Since scale invariance generalizes time invariance, the wavelet transform can be viewed as a generalization of the Fourier transform. Researchers have confirmed that the improved signal/noise separation provided by the wavelet transform results in real-world pattern recognition applications with enhanced classification and predictive capabilities [27,28,46].

Wavelets at different scales and times $\psi_{s,t}$ are derived from a single "mother" wavelet ψ via scaling and translation operations. The wavelet transform of a given signal f is defined by correlating f with each wavelet $\psi_{s,t}$ and summing the correlations

$$\tilde{f}(s,t) \equiv \int_{-\infty}^{\infty} \psi\left(\frac{u-t}{s}\right) f(u)du = \int_{-\infty}^{\infty} \psi_{s,t}(u) f(u)du$$

where $\tilde{f}(s,t)$ is the CWT of f at scale s and time t with respect to the mother wavelet ψ. As a function of t for a fixed scale s, $\tilde{f}(s,t)$ represents information in the signal having frequencies that are localized to a spectral region that is centered on some frequency that depends on the fixed scale s. As a function of s for a fixed time t, $\tilde{f}(s,t)$ represents information in the signal of all frequencies that is localized in some temporal region centered on the fixed time t. Larger scale values capture coarse signal detail (e.g., global trends), while the smaller scale values capture finer detail (e.g., transient fluctuations and random noise). Hence, the CWT provides a means of characterizing both local and global variation in a single signal representation.

There are an infinite number of mother wavelets to choose from depending on the characteristic of the signal being analyzed. The mother wavelet used for the microarray data analysis in this study is known as the Daubechies mother wavelet. Computational studies using real and simulated data sets have shown that this particular wavelet results in the best classification performance on the Huang breast cancer tumor samples. This is mainly due to the shape of the Daubechies mother wavelet that simultaneously smoothes the expression profile for a given sample while capturing localized variations in the data over multiple scales of resolution.

Figure 4-8 shows the wavelet transform of quantile normalized, $\log_2$ transformed, z-scored expression profiles of a lymph node negative tumor sample (Fig. 4-9a) and a lymph node positive tumor sample (Fig. 4-9b). The expression profiles were generated by the downregulated genes in IPA network *dnet1*. In each case, the actual sample

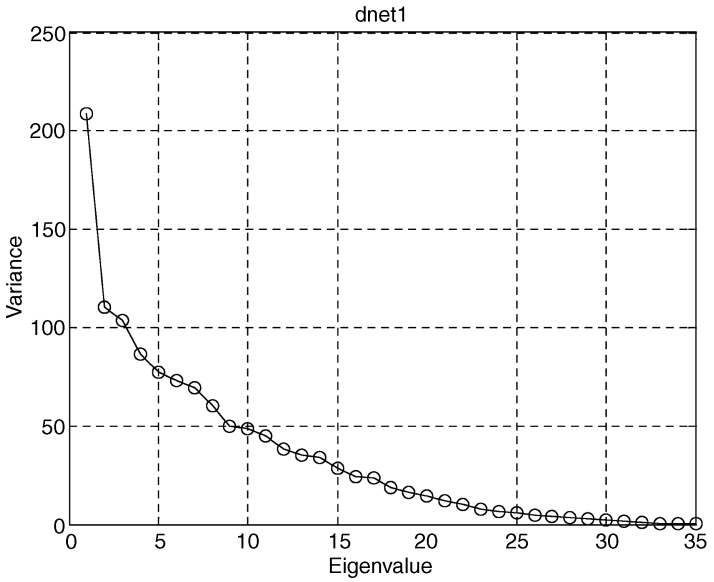

Figure 4-8 Eigenvalue plot for the data matrix of genes from IPA network *dnet1*. Each eigenvalue represents the variation in the direction of the corresponding eigencomponent. Eigenvalues 1–10 are most likely to correspond to coherent signal. Eigenvalues 11–36 probably correspond to noise.

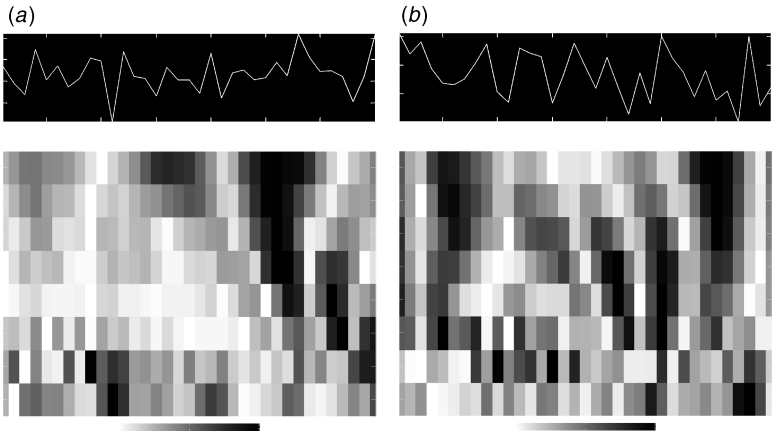

Figure 4-9 The continuous wavelet transform based on the Daubechies mother wavelet of order 4 for two sample expression profiles from the data matrix of pathway compressed genes.(a) The sample expression profile of a lymph node negative tumor over 35 genes is shown above an image of the CWT of the profile. The vertical axis of the CWT image represents scale, while the horizontal axis represents samples ordered by lymph node status with negative samples first. Note that the CWT of a one-dimensional signal is two-dimensional image. (b) The sample profile and CWT image of a lymph node positive tumor. Note how coefficients with high magnitude show different distribution over scale and samples. Coefficients contained in rows near the bottom of the image contain mostly noise, while coefficients in rows near the top of the image represent coherent signal characterizing differences between negative and positive tumor samples.

expression profile is shown above the CWT of the profile. Note that the CWT of a one-dimensional signal is a two-dimensional image. The vertical axis of each CWT image represents the scale of resolution of the CWT. Not that the scale of resolution becomes progressively coarser from the bottom up along the vertical axis. The horizontal axis represents the tumor samples ordered by lymph node status with the negative samples occupying columns 1–19 and the positive samples occupying columns 20–36. Hence, each row of the CWT image represents the information content of the profile at a particular scale of resolution, while each column represents the information content of a particular sample over eight scales of resolution.

The main idea of CWT signal processing is that random noise is concentrated at the higher scales of resolution near the bottom of the CWT image, while lower frequency coherent signal is concentrated in the coarser scales of resolution near the top. This separation of signal and noise by scale enhances subsequent compression and denoising of the data using SVD. Each CWT "image" is vectorized to generate a one-dimensional wavelet transformed expression profile that is used to form a column of a CWT data matrix A_{wave}.

In particular, let A_{dnet1} be the 35×36 data matrix of quantile normalized, $\log_2$ transformed, z-scored gene expression values defined by the genes of *dnet1*. Each 35-dimensional column of A_{dnet1} is wavelet transformed over 64 scales in increments of 8 using the continuous wavelet transform (CWT) based on the Daubechies mother wavelet of order 4 (Daub4) that results (after vectorization) in a 240×36 wavelet data matrix A_{wave}. SVD was used to compress A_{wave} down to a 12×36 data matrix $A_{\text{feats}} = [f_1, f_2, \ldots, f_{36}]$ as explained below. Each column f_j of A_{feats} represents a 12-dimensional vector of WSVD features that characterizes the jth sample of the experiment for $j = 1,2, \ldots, 36$. Previous research has shown that WSVD features significantly enhance the performance of pattern recognition algorithms in real-world applications [27,28]. In this study, we show that pathway compression coupled with WSVD signal processing enhances the classification of high- and low-risk breast cancer samples based solely on gene expression of the primary tumor using neural network classifiers.

4.9.5 Singular Value Decomposition

SVD is a classical statistical technique for characterizing the linear correlation that exists in a data matrix [47]. It is closely related to the Karhunen–Loeve transform (random processes), principal component analysis (matrix diagnonalization), and factor analysis (correlation structure of multivariate stochastic observations). SVD is used in many areas of science and engineering as a means of extracting features for pattern recognition, data compression, signal detection, and sample classification applications. Essentially, the primary goal of SVD is to find a linear transformation that maps a vector of noisy, correlated "time domain" measurements into a much smaller vector of denoised, uncorrelated feature components [27,28].

Let $A = [x_1, x_2, \ldots, x_K]^{\mathrm{T}}$ be a $p \times n$ expression data matrix with p-dimensional columns x_i composed of noisy, correlated expression measurements (superscript T is the matrix transpose operator). We desire a linear transformation $L : \mathbb{R}^p \rightarrow \mathbb{R}^k$ such

that $y_i = Lx_i$ is a compressed eigenarray of dimensionality k ($k \ll n$) composed of uncorrelated, denoised *eigenexpression* values. The *fundamental theorem of linear algebra* states that under very general conditions, there exists orthogonal matrices U and V and a diagonal matrix Σ such that $A = U\Sigma V^T$ [48]. Here, U is $p \times p$, U is $n \times n$, and S is $p \times n$. The columns of U are the *eigenarrays* of A, and they provide an orthonormal basis for $\mathbb{R}^p$. Similarly, the columns of V are the eigengenes of A, and they form an orthonormal basis for $\mathbb{R}^n$. The square of the diagonal entries of S are the eigenvalues, λ_i, of A and are ordered so that $\lambda_j > \lambda_{j+1}$ for $j = 1,2, \ldots, n-1$. We choose the first k eigenarrays of U that correspond to the k largest eigenvalues (where usually $k \ll n$) and form the matrix U_{trunc} with columns equal to the selected eigenarrays. It follows that $L : \mathbb{R}^p \rightarrow \mathbb{R}^k$ defined by $L = U_{\text{trunc}}^T$ is the linear transformation we seek since it maps a p-dimensional vector into a k-dimensional vector where $k \ll n < p$. The k components of $y_i = Lx_i$ are known as the principal components of x_i [47].

We note that the resulting feature vector y_i is denoised due to the truncation of those eigenarrays of U that are associated with the remaining $(n-k)$ eigenvalues. It is assumed that the truncated eigenarrays span a $(n-k)$-dimensional subspace containing the random noise component of the data. We note that the subspace spanned by the truncated eigenarrays may contain information that is useful for classification, and one needs to be careful that this information is not lost in the dimensionality reduction process. Usually, though, a visual analysis of a plot of the eigenvalues makes it clear where the threshold should be set using, say, Kaiser's rule [49].

Figure 4-10 shows a plot of the 36 eigenvalues obtained for the 240×36 wavelet data matrix A_{wave}. Note that the eigenvalue plot becomes linear starting at about the 12th eigenvalue, so that $\sum_{i=1}^{12} \lambda_i$ represents the variation associated with coherent signal, which accounts for 79 percent of the total variation in A_{wave} based on the top 12 eigenarrays. The sum of the remaining eigenvalues $\sum_{i=13}^{36} \lambda_i$ represents the variation associated with the noise, which accounts for the remaining 21 percent of the energy in the data. The above analysis of the eigenvalues suggests that we retain the first 12 eigenarrays (i.e., columns) of U to form a 240×12 transformation matrix U_{trunc}, which is used to "compress" the 240×36 data matrix A_{dnet1} down to a 12×36 data matrix y of WSVD/IPA feature vectors using

$$y = (U_{\text{trunc}})^T x$$

where x is a 240 component column vector of A_{wave}. Note that each of the 36 tumor samples is now characterized by 12 numbers instead of the original 12,625 expression measurements. This is a huge reduction in dimensionality with a theoretically minimal loss of information accompanied by a theoretically maximal reduction in noise [15].

4.9.6 Combining Wavelets, SVD, and IPA (WSVD/IPA)

WSVD/IPA signal processing combines wavelet signal processing SVD and IPA pathway compression to extract signal features from the Huang microarray data set. The WSVD/IPA feature patterns are then used to train a NN classifier to robustly

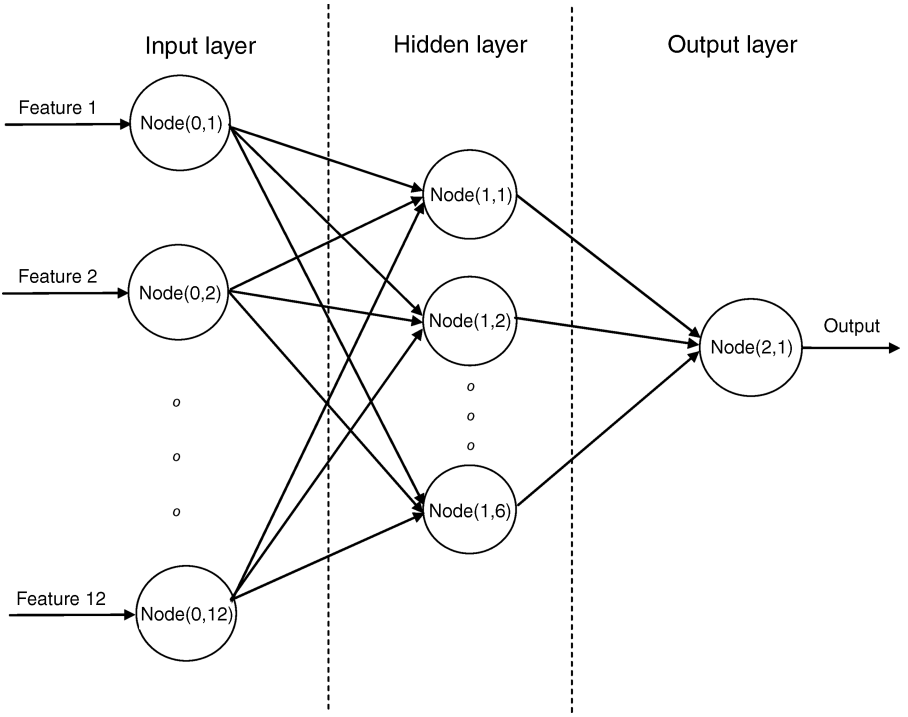

Figure 4-10 Network diagram for a feedforward multilayer perceptron. Each node represents a processing unit (artificial neuron) that computes an output according to Equation 4-4. Each arrow represents the passing of information from one node to another. Note that the output of a given node is passed as an input to the every node in the next layer. Nodes in the same layer do not pass information to each other. The diagram shown has one input layer of 12 nodes, one hidden layer of 6 nodes, and a single note in the output layer. A 12-dimensional feature vector (e.g., a WSVD/IPA expression feature pattern) is presented to the input layer of the FFMLP. The input cascades forward through the network layer by layer and eventually results in a single output value at the output layer of the FFMLP. This output vector is thresholded to determine the decision made by the FFMLP regarding the lymph node status of the tumor that is represented by the 12-dimensional input feature pattern.

predict lymph node status based on gene expression patterns from primary breast cancer tumors. The NN classifier is validated using leave-one-out cross-validation starting with the $12{,}625 \times 36$ raw data matrix A_{raw} with an outlier sample removed.

We basically follow the flowchart given in Figure 4-1. The columns of A_{raw} are first quantile normalization to obtain the normalized data matrix A_{nrm}. This normalization step facilitates comparisons between the microarrays of the experiment. The MANINI detection algorithm is then applied to A_{nrm} to obtain lists of up- and downregulated genes that are DE between the positive and negative samples of the Huang microarray data set. This completes the gene detection phase of the information processing chain.

Ingenuity Pathway Analysis is then used to extract statistically significant pathways from the MANINI gene lists. The genes in the most statistically significant IPA networks are then intersected with the MANINI gene lists to generate individual data

matrices of normalized, $\log_2$ transformed, z-scored gene expression profiles. This step is called IPA pathway compression, since a gene selected by MANINI is passed on for downstream processing only if it is also contained in a significant IPA network. Let A_{net} be the data matrix associated with a significant IPA network. The CWT is then used to transform the columns of A_{net} to obtain the 240×36 wavelet data matrix A_{wave}. The Daubechies mother wavelet over 64 scales in increments of 8 was used since computational experiments suggested that these parameters resulted in the best NN performance. Note the CWT better separates signal and noise in the wavelet domain and enhances data compression and denoising using SVD [39,50,51]. Application of the SVD to A_{wave} resulted in a 12×36 matrix of WSVD/IPA feature patterns $A_{\text{feats}} = [f_1, f_2, \ldots, f_{36}]$. Note the columns of A_{feats} represent 12-dimensional feature patterns (derived from the original 12,625-dimensional gene expression profiles) that characterize each of the 36 samples of the Huang data set. Note the greatly reduced dimensionality of the WSVD/IPA feature vectors f_j enhances NN modeling of lymph node status by alleviating the adverse impact of Bellman's curse of dimensionality. The WSVD/IPA feature vectors f_j for $j = 1,2, \ldots, 36$ were used to train a $12 \times 6 \times 1$ NN model with 12 input nodes, 6 hidden nodes, and 1 output node to discriminate between the positive and negative lymph node samples.

4.9.7 Neural Network Modeling of Lymph Node Status

NN models are useful for situations where there is much data, but a theory is lacking that explains the data [24]. The focus then shifts to patterns within the data that are associated with a quantifiable attribute measured for each sample. Neural networks are also known as data-driven models or machine learning models. In microarray data analysis, for example, we are usually given a finite number of ordered pairs (x,y) that associate a k-dimensional gene expression pattern $x \in \mathbb{R}^k$ to a unique phenotypic state $y \in \{0,1\}$. In this case, we wish to "discover" a mapping from expression patterns to phenotypic states $\{0,1\}$ that "generalizes" to new expression patterns not contained in the original data set [21]. This mapping allows the prediction of phenotypic states of new patterns that were "unseen" during training of the NN model. The discovered mapping can be implemented as a NN, which represents a massively parallel, highly distributed computational model of observed phenotypic variation based on gene expression patterns.

A NN is composed of elementary computational units that can be "connected" to each other to form a network that is capable of global information processing arising from the local interactions between the computational units. The strength of the interaction between two computational units is encoded in a "synaptic" weight that quantities the "strength" of the interaction between the two units [52]. The collection of all synaptic weights can be adjusted in parallel to realize almost any continuous mapping between two sets of variables. This emergent computational behavior can be highly nonlinear in nature, and as a result, a NN model is capable of solving very difficult classification and prediction problems that require complex boundaries between the different sample classes. The form of the resulting connection diagram is called the architecture of the network, and the computations performed by the

network are highly dependent on this architecture. That is, the modeler has control over how the connections are evolved so that the architecture is plastic and trainable. Indeed, the synaptic weights of the MLP can be adjusted using a machine learning algorithm such as backpropagation to realize an arbitrarily good approximation to any mapping between expression patterns and phenotype suggested by the data.

A neural network architecture known as the *feedforward multilayer perceptron* (FFMLP) is used to model phenotypic variation over the samples of the microarray experiment in terms of gene expression. The FFMLP architecture arranges the computational nodes of the network into a "hidden" layer and an output layer. There is an additional input layer of nodes, but these nodes merely pass the input values on to the hidden layer without computation. The hidden layer is described as such because it is shielded from direct contact with the outside world by the input and output layers of nodes. The FFMLP is fully connected in that every node of a given layer is connected to every node of the next layer by an arrow representing the direction of information flow. Finally, to every arrow is assigned an adaptive weight parameter that mediates the flow of information between the two nodes. Training data consisting of empirical input–output pairs are used to adjust the weights using a nonlinear optimization algorithm (e.g., error backpropagation) to approximate the input–output mapping that is "implied" by the data. In this way, a FFMLP can approximate any mapping between any two sets of quantities to an arbitrary degree of accuracy [49].

Figure 4-10 shows a network diagram for a two-layer FFMLP, where each node represents a computational unit and each arrow represents the flow of information from one node to another. In this case, the FFMLP has an input layer consisting of 12 nodes, a hidden layer consisting of 6 nodes, and an output layer consisting of a single node. Usually, the input layer is excluded from a count of layers that compose a given FFMLP. If the dimension of the input layer is k, then the training data set consists of ordered pairs $\{(x, y) \in \mathbb{R}^k \times \mathbb{R}^2\}$. For a given training data pair (x,y), the k-dimensional feature vector x is presented to the input layer of the NN and the resulting output vector $y_{\mathrm{NN}}(x)$ is compared with the target vector y where the dimensionality of y and y_{NN} are equal.

For this study, x is the WSVD/IPA feature vector for a chip hybridized to a sample from the Huang breast cancer data set and y is either 0.05 or 0.95, depending on whether the sample is negative or positive for lymph node involvement, respectively. The error between y and y_{NN} is propagated back through the NN to adjust the adaptive weights of the NN to reduce the error in the output layer using the *error back-propagation* algorithm. The learning process is iterated over all training pairs and repeated until the aggregate error for the training data is reduced to an acceptable level. Note that training is terminated in such a way as to balance accuracy on the training data with the ability of the FFMLP to generalize classification performance to data not seen during training. To enhance generalization to new data, the output of the FFMLP is smoothed using Bayesian regularization during training.

The input to the kth node of a given layer is a linear weighted sum of all inputs to the node from all nodes in the previous layer, given by

$$\sum_j w_{kj} x_j$$

where the sum runs over all nodes in the previous layer that sends a signal x_j to node k (we assume that a bias parameter is included in the summation) and w_{kj} is the adaptive weight for the connection between the two nodes [53]. The output of the kth node is obtained by transforming the weighted linear sum with a nonlinear activation function g using

$$z_k = g\left(\sum_j w_{kj} x_j\right) \tag{4-4}$$

For a vector of values presented to the input layer of the FFMLP, the output of each computational node in a given layer can be computed in a *feedforward* manner in terms of the outputs of all the nodes in the previous layer.

From a theoretical perspective, the FFMLP is known as a universal approximator; that is, it can uniformly approximate any continuous function on a compact domain to an arbitrary degree of accuracy provided the network has a sufficiently large number of hidden units and enough data. The key problem remains how to find suitable parameter values given a set of training data. There exist effective solutions to this machine learning problem based on both maximum likelihood and Bayesian approaches. The error backpropagation algorithm is probably the most widely used method to train a FFMLP.

In particular, the WSVD/IPA feature patterns x_j extracted for each sample were used to train a FFMLP model to discriminate between the positive and negative lymph node samples. The FFMLP had an architecture shown in Figure 4-10 compose of 12 input nodes, 6 hidden nodes, and a single output node. The 12-dimensional WSVD/IPA input feature vector is passed without any processing to the nodes of the hidden layer, where each hidden node computes an output value in accordance with Equation 4-4. The output of each hidden node is then fed into the single output node, which computes a weighted linear combination of inputs and transforms the result using the sigmoidal logistic function g. A fixed threshold (equal to 0.5) is applied to the output value to determine whether the input feature pattern was associated with a positive or negative breast cancer tumor.

The hidden nodes employed hyperbolic tangent activation functions with range confined to the interval $[-1,1]$. The single output node employed a sigmoidal logistic activation with range confined to the interval $[0,1]$. The Levenberg–Marquardt training algorithm with Bayesian regularization was used to train the FFMLP to output a value of 0.95 for lymph node positive samples and 0.05 for the lymph node negative samples. A sample was classified as lymph node positive if its associated FFMLP output exceeded a threshold of 0.5, otherwise it was classified as lymph node negative. The FFMLP was trained for 20 epochs with a targeted error goal of 0.005. Training was usually completed in less than 30 s, which facilitated validation of the GSP algorithms.

4.10 MODEL VALIDATION

The robustness and accuracy of FFMLP models trained on the Huang microarray data set was evaluated using leave-one-out cross-validation (LOOCV) analysis [4,19].

LOOCV analysis begins by removing, say, the kth column from the raw data matrix A_{raw}. This results in the column-reduced 12,625 × 35 data matrix A_{raw}^{k} from which WSVD/IPA features are extracted as described above to train a NN model $y_k(x)$, where x represents a 12-dimensional WSVD/IPA feature vector. Let x_k denote the WSVD/IPA feature vector associated with the kth column of A_{raw} that was left out. Recall that x_k was unseen during training of y_k and we want to see if y_k can correctly classify this sample. By design, we say that x_k is lymph node negative if $y_k(x_k) < 0.5$ and lymph node positive otherwise. The classification result is duly recorded and compared with the known lymph node status for the left-out sample. The entire process is repeated 36 times for each column of A_{raw}. The correct classification rate (CCR) is defined as the percentage of left-out samples that were correctly classified. Note that for each sample left out during the LOOCV analysis, a different set of downregulated genes is selected by the MANINI algorithm. This variation in the gene lists reflects the variation in the population of all breast tumor expression profiles. We utilize this variation to assess the robustness of NN models trained on such data.

Figure 4-11 shows a flowchart of the information processing used to validate feature patterns for classifying breast cancer samples into high- and low-risk groups using FFMLP classifiers. Major signal processing occurs in the orange boxes labeled ''preprocess data,'' ''select genes,'' ''extract features,'' and ''train classifier'' as shown in Figure 4-12. In particular, gene selection based on MANINI detection and IPA pathway compression occurs in the ''select genes'' box of the flowchart, and WSVD features are extracted in the ''extract features'' box. Note also the parallel chain (in green) that processes the ''left-out'' sample for eventual classification by the NN

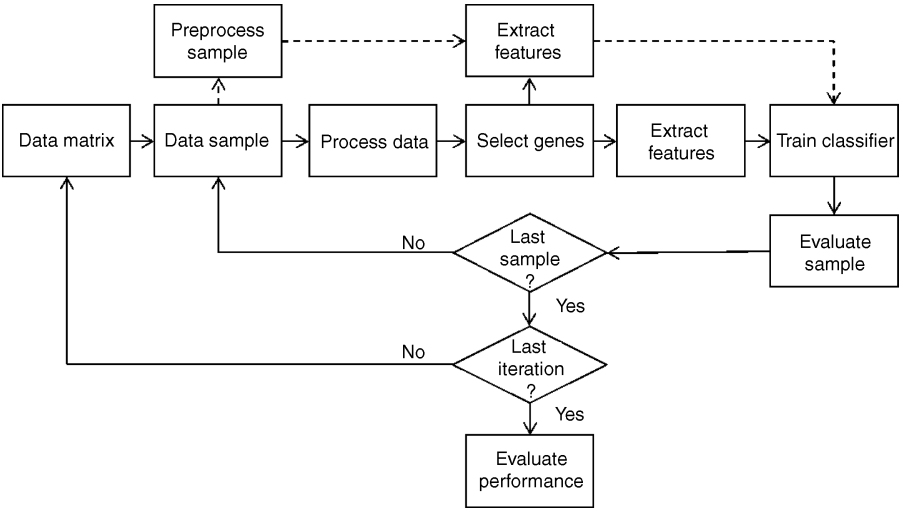

Figure 4-11 Flowchart for leave-one-out cross-validation of NN models of lymph node status based on pathway compression and WSVD signal processing. LOOCV analysis estimates the impact of sampling variation on the prediction performance of the neural network classifier. The robustness and accuracy of the proposed prediction model depends to a large extent on the quality of the feature patterns extracted from the raw data.

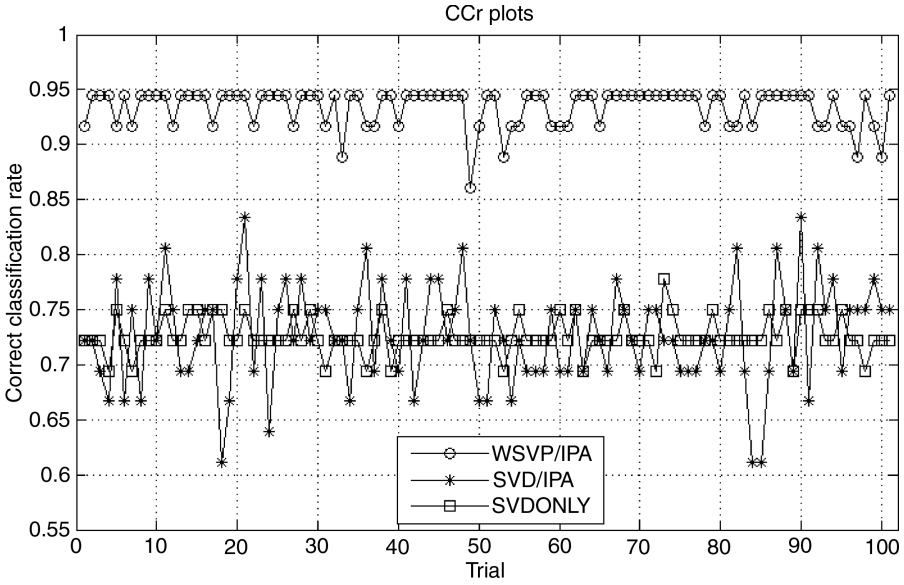

Figure 4-12 A comparison of WSVD/IPA, SVD/IPA, and SVDOnly features for predicting lymph node status using neural networks. LOOCV analysis of FFMLPs trained on each feature set was performed 100 times and plotted to assess the variation in CCR values. WSVD/IPA and SVD/IPA features were extracted from the 35 genes that were contained in the top IPA network for all genes called downregulated genes by the MANINI detector. SVDOnly features were extracted from all 413 downregulated genes detected by the MANINI detector. Note that the CCR plot for WSVD/IPA features (open circles) lies uniformly above the CCR plots for SVD/IPA (stars) and SVDOnly (open squares) features. Median CCR is 94 percent for WSVD/IPA and 72 percent for SVD/IPA and SVDOnly.

classifier trained on the remaining 35 samples. Because the algorithm used to train the FFMLP classifier is subject to entrapment in local minima, we repeated the LOOCV procedure 100 times to obtain some idea of the variability introduced into the training process due to sampling variation in the data. The maximum, median, and minimum CCR values over 100 LOOCV trails were used to evaluate the effectiveness of different signal processing algorithms in discriminating between positive and negative lymph node samples from the Huang data set.

4.11 RESULTS

In this section, we present results of a LOOCV evaluation of FFMLP models trained on WSVD/IPA features to discriminate between breast cancer tumor samples based on lymph node status. We also assessed and compared the performance of FFMLP models trained on two additional feature types derived using (1) SVD combined with pathway compression (SVD/IPA) and (2) SVD compression applied to all down-regulated genes (SVDOnly).

Figure 4-12 shows plots of 100 CCR values for FFMLP classifiers trained using WSVD/IPA (open circles), SVD/IPA (stars), and SVDOnly (open squares) feature patterns. Note that FFMLP models trained on WSVD/IPA feature patterns uniformly outperform models trained on SVD/IPA and SVDOnly features. Indeed, the median CCR value associated with WSVD/IPA features was 94 percent, while both SVD/IPA and SVDOnly features had median CCRs of 72 percent. Hence, wavelet signal processing improved FFMLP classification performance by 31 percent when used in conjunction with SVD analysis and IPA pathway compression. SVD/IPA and SVDOnly had the same median performance, but SVD/IPA had greater variation and attained a maximum CRR of 83 versus 78 percent for SVDOnly. This suggests that IPA pathway compression does not provide any improvement in classification performance over SVD alone. Table 4-3 summarizes the maximum, median, and minimum CCR values over 100 LOOCV trials for each feature type.

Overall, wavelet signal processing and pathway compression combined to significantly enhance the prediction of lymph node status when compared to more conventional signal processing based on SVD alone. Note the option of combining SVD and wavelet signaling processing (without the benefit of pathway compression) was not considered since this would have required wavelet processing of data vectors of length 413 instead of 35, which is computationally intensive. Hence, the success of wavelet signal processing in the context of this study was a direct result of the data reduction provided by pathway compression; that is, wavelets, pathway compression, and SVD worked hand in hand to reduce dimensionality without loss of information related to lymph node status, and thus significantly enhancing the classification of the Huang breast cancer samples.

Note wavelet signal processing together with pathway compression achieved a 94 percent CCR based on only 35 genes. This result compares favorably with the 90 percent CCR achieved by Huang et al. on the same data set using 200 genes and different statistical methodology. These results suggest that the lymph node status of breast cancer samples can be predicted in an accurate and robust manner using a relatively small number of genes when the appropriate signal processing and data

Table 4-3 Comparison of the classification performance of the different feature types over 100 LOOCV trials

Feature Type	Max CCR	Median CCR	Min CCR
WSVD/IPA	94%	94%	86%
SVD/IPA	83	72	61
SVDOnly	78	72	69

The maximum, median, and minimum CCR values are shown for each feature set. Note that FFMLP models trained on WSVD/IPA features attained a median CCR of 94 versus 72 percent for SVD/IPA and SVDOnly features. This result represents a 31percent improvement in prediction performance, suggesting that wavelet signal processing significantly enhances the classification of breast cancer tumors that have spread to the lymph nodes. Note the median CCR for both the SVD/IPA and SVDOnly feature sets were equivalent, although the CCR variation for the SVD/IPA features is greater than the CCR variation for the SVDOnly features. This suggests that pathway compression without wavelet processing does not result in any appreciable improvement in classification performance.

compression algorithms are utilized. In this particular case, ''appropriate'' means wavelet signal processing, SVD and pathway compression, and neural networks.

4.12 DISCUSSION

We have shown that combining different signal processing techniques, namely, MANINI detection, pathway compression, wavelets, and SVD results in feature patterns that are able to accurately and robustly discriminate between high- and low-risk breast cancer tumors using as few as 35 genes. Specifically, FFMLP classifiers trained on WSVD/IPA feature patterns showed a 31 percent increase in CCR in comparison to FFMLP classifiers trained on SVD/IPA or SVDOnly features.

A key step in the robust modeling of lymph node status was the selection of genes for downstream feature extraction and machine learning using the MANINI detection algorithm. Recall that the MANINI detector assigns genes having similar expression levels into ''bins'' where each bin defines a separate signal detection problem based on data that is approximately normally distributed. A ''test-and-estimate'' or testimation procedure based on the DJ universal threshold (from wavelet denoising theory) was applied to the fold change values of the genes in each bin. Genes that exceeded the threshold were called differentially expressed. The union of genes over all intensity bins was called differentially expressed over bins represents a global gene expression signature for metastatic breast cancer tumors.

Recent research has shown a deep connection between statistical testimation, sparse signal estimation, and multiple hypothesis testing with adjustments for multiple comparisons [29]. Hence, the MANINI detector can be viewed as an optimal estimator of a sparse signal that automatically accounts for the approximately 600 statistical tests performed simultaneously within each of the 50 bins of the MANINI detector. Note also that MANINI detection throws a wide net and detects genes based on similarity rather than absolute magnitude or consistency of expression. That is, MANINI may call a gene DE even though it is altered in expression on only a fraction of the positive samples of the Huang data set, so long as the average fold change over all samples exceeds the DJ threshold for the bin to which it belongs. The ability to detect genes that are ''intermittently'' DE is important when dealing with heterogeneous data sets often encountered in cancer research that contain weak signals for DE, or samples that are temporally and/or developmentally out of phase.

The observed increase in FFMLP classification performance using WSVD/IPA features is due in large part to the fact that pathway compression drastically reduces the number of genes that must be processed while at the same time preserving important information related to lymph node status [16]. Since significant IPA pathways extracted from the MANINI gene list are biologically relevant in terms of known gene function and gene–gene interactions as embodied in the IPA knowledge base, pathway compression provides a biologically driven method for selecting a small number of highly informative genes from which feature patterns can be extracted for diagnostic, prognostic, and predictive applications in the clinic. Because the resultant features are biologically based, the NN models trained on such features are likely to be more robust over a larger population of samples. The results shown in Table 4-3

suggests that NN models trained on gene expression signatures derived from pathway compression generalize well to a larger population of samples.

Table 4-3 also indicates that wavelet signal processing significantly enhances the prediction of lymph node status using NN models. Indeed, wavelet signal processing raised the median CCR from 72 to 94 percent for an overall improvement of 31 percent. The main reason for this result is that wavelet analysis of a sample's expression profile exhibits better separation between coherent signal and random noise in the CWT domain [45,52]. The subsequent analysis of the wavelet transformed data matrix using SVD results in a better estimate of the intrinsic dimensionality of the wavelet denoised data matrix. SVD compression in the wavelet domain has been used to solve a number of difficult pattern recognition problems, including the automated classification of underwater buried mines and the detection of cervical pre-cancer in three-dimensional hyperspectral images of the cervix [27,28].

Note that wavelet signal processing of microarray data was made feasible to a large extent by pathway compression. Indeed, without pathway compression, sample expression profiles of 413 genes would have to be wavelet transformed, which is computationally onerous. In contrast, after pathway compression, the resulting sample expression profiles are no more than 35 genes long, thereby enabling the efficient use of the CWT for data preconditioning and denoising. Moreover, the information processing described in this chapter applies wavelet signal processing to the columns of the data matrix instead of the rows. Hence, the resolution of the wavelet transform is limited not by the number of samples in the experiment, but by the number of genes included in the pathway-compressed data matrix. Since the number of genes is intrinsically large, this approach circumvents the problem of having too few samples for the wavelet transform to work properly.

NN classification performance can probably be improved by including genes contained in different IPA networks. Future work involves the use of genetic algorithms to globally search for the best combination of genes, significant IPA networks, wavelets, and model parameters that maximizes the CCR score for predicting lymph node status in the Huang breast cancer data set. Once the optimal genes and pathways for distinguishing positive and negative breast cancer tumors are identified, close examination of individual genes and their interactions with other genes contained in the selected networks could very well lead to new insights into the molecular mechanisms underlying metastatic breast cancer. Here, the overarching assumption is that predictive performance is equivalent to biological significance, thus enabling the use of machine learning models such as FFMLPs to identify the genes and pathways that are most biologically relevant to the spread of breast cancer to the axillary lymph nodes. Are there networks or combinations of networks that result in even more robust and accurate predictions of lymph node status than those produced by the IPA network *dnet1*? It is expected that modern signal processing, FFMLP/NN models, and genetic search algorithms will provide an answer to this question.

Note that IPA assigned the biological function of cancer to *dnet1*, and moreover, simulations suggest that the topology of *dnet1* with the FOS gene as a highly connected "hub" gene is invariant to minus-one perturbations of the data. The FOS gene family has been implicated in the regulation of cell proliferation, differentiation, and transformation. Other cellular roles include transformation, apoptosis,

growth, activation, motility, and cell cycle progression. FOS has also been associated with cardiovascular disease. Note that the ability to accurately classify breast cancer tumors according to lymph node status is quite different from attaining a deep understanding of the biological mechanisms underlying the spread of breast cancer. Be that as it may, a close examination of predictive networks and the genes they contain could well lead to a better understanding of the molecular mechanisms underlying metastatic breast cancer. Such mechanistic models could lead to new diagnostics and therapeutics that significantly improve the way breast cancer is treated and managed in the clinic.

ACKNOWLEDGMENTS

I would like to acknowledge my colleagues at the Cardiovascular Research Center and the Cancer Research Center of Hawaii and, in particular, Drs Charles Boyd, Richard Girton, Loic Le Marchand, and Patrick Fu for their support, encouragement, and many discussions during the research and writing of this chapter.

REFERENCES

1. Jemal A, et al. *Cancer Facts & Figures 2007*. Atlanta: American Cancer Society, 2007.
2. Veronesi U, Paganelli G, Galimberti V, Viale G, Zurrida S, Bedone M, Costa A, Decicco C, Geraghty JG, Luini A, Sacchini V, Veronesi P. Sentinel-node biopsy to avoid axillary dissection in breast cancer with clinically negative lymph-nodes. *Lancet* 1997; 349:1864–1867.
3. Jatoi I, Hilsenbeck SG, Clark GM, Osborne CK. Significance of axillary lymph node metastasis in primary breast cancer. *J Clin Oncol* 1997;17:2334–2340.
4. Huang E, et al. Gene expression predictors of breast cancer outcomes. *Lancet* 2003; 361:1590–1596.
5. West M, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 2001;98(20):11462–11467.
6. Ma XJ, Salunga R, Tuggle JT, Gaudet J, Enright E, Mcquary P, Payette T, Pistone M, Stecker K, Zhang BM, Zhou YX, Varnholt H, Smith B, Gadd M, Chatfield E, Kessler J, Baer TM, Erlander MG, Sgroi DC. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci USA* 2003;100:5974–5979.
7. Mircean C, et al. Pathway analysis of informative genes from microarray data reveals that metabolism and signal transduction genes distinguish different subtypes of lymphomas. *Int J Oncol* 2004;24:497–504.
8. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular Biology of the Cell*, 4th ed. New York: Garland Science, 2002.
9. van't Veer L, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–536.
10. Simon R. DNA microarrays for diagnostic and prognostic prediction. *Expert Rev Mol Diagn* 2003;3(5):587–595.

11. Schena M. Genome analysis with gene expression microarrays. *Bioessays* 1996;18(5): 427–431.
12. Kohane IS, Kho AT, Butte AJ. Microarrays for an Integrative Genomics. In: Istrail S, Pevzner P, Waterman M, editors. *Computational Molecular Biology.* Cambridge, MA: The MIT Press, 2003.
13. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251(4995):767–773.
14. Zweiger G. *Transducing the Genome*. New York: McGraw-Hill, 2001.
15. Mertens B. Microarrays, pattern recognition and exploratory data analysis. *Stat Med* 2006;22:1879–1899.
16. Slonim D. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* 2002;32:502–508.
17. Wong DJ, Chang HY. Learning more from microarrays: insights from modules and networks. *J Invest Dermatol* 2005;125:175–182.
18. Donoho D. High-dimensional data analysis: the curses and blessings of dimensionality. In: *Mathematical Challenges of the 21st Century.* University of California, Los Angeles: American Mathematical Society, 2000.
19. Simon R, et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003;95:14–18.
20. Rocke DM, Durbin B. A model for measurement error for gene expression arrays. *J Comput Biol* 2001;8(6):557–569.
21. Bishop C, editor. *Pattern Recognition and Machine Learning*, Information Science and Statistics, Jordan M, Kleinberg J, Scholkopf B,editors. New York: Springer Science + Media, 2006.
22. Vanchieri C. National Cancer Act: a look back and forward. *J Natl Cancer Inst* 2007;99(5): 342–344.
23. Yang SX, et al. Gene expression patterns and profile changes pre- and post-erlotinib treatment in patients with metastatic breast cancer. *Clin Cancer Res* 2005;11(17): 6226–6232.
24. Bishop C. *Neural Networks for Pattern Recognition*. Oxford, NY: Oxford University Press, 1998.
25. Valafar F. Pattern recognition techniques in microarray data analysis: a survey. *Ann N Y Acad Sci* 2002;980:41–64.
26. Bellman RE. *Adaptive Control Processes*. Princeton, NJ: Princeton University Press, 1961.
27. Okimoto GS, Lemonds D. Principal components analysis in the wavelet domain: new features for underwater object recognition. In: *Proceedings of SPIE on Aerosense: Detection and Remediation Technologies for Mines and Mine-like Targets IV, Orlando, FL*, 1998.
28. Okimoto GS, et al. New features for detecting cervical pre-cancer using hyperspectral diagnostic imaging. In: *Proceedings of SPIE on Clinical Diagnostic Systems, San Jose, CA*, 2001.
29. Sabatti C, Karsten SL, Geshwind DH. Thresholding rules for recovering a sparse signal from microarray experiments. *Math Biosci* 2002;176:17–34.
30. Schalkoff R. *Pattern Recognition: Statistical, Structural and Neural Approaches*. New York: John Wiley and Sons, Inc., 1992.

31. Okimoto GS. On the analysis of microarray data with applications to cancer and cardiovascular disease. Doctoral Dissertation for the Department Molecular Biosciences and Bioengineering (in Bioinformatics), 2006.
32. Quackenbush J. Microarray data normalization and transformation. *Nat Genet* 2002;32:496–501.
33. Ballman KV, et al. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics* 2004;20(16):2778–2786.
34. Quackenbush J, Computational analysis of microarray data. *Nat Rev Genet* 2001;2:418–427.
35. Lyons-Weiler J, et al. Test for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics* 2004;5110.
36. Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis. Stanford Technical Report, 2006.
37. Draghici S, et al. Global functional profiling of gene expression. *Genomics* 2003;81:98–104.
38. Mariani TJ, et al. A variable fold change threshold determines significance for expression microarrays. *FASEB J* 2003;17:321–323.
39. Donoho D, Johnstone I. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 1994;81:425.
40. Poor HV. *An Introduction to Signal Detection and Estimation*, Springer Texts in Electrical Engineering, Thomas JB, editor. New York: Springer Verlag, 1994.
41. Kay SM. *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory.* New York: Prentice Hall, 1998.
42. Dudoit S, et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report #578, University of California at Berkeley, 2000.
43. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;102(43): 15545–15550.
44. Zimmerman Z, Golden JB III. The return on investment for Ingenuity Pathways Analysis within the pharmaceutical value chain. Unpublished white paper, 2004.
45. Kaiser G. *A Friendly Guide to Wavelets*. Boston, MA: Birkhauser, 1995.
46. Parker M, et al. Initial neural net construction for the detection of cervical intraepithelial neoplasia by fluorescence imaging. *Am J Obstet Gynecol* 2002;187:398–402.
47. Kalman D. A singularly valuable decomposition: the SVD of a matrix. *College Math Journal* 1996;27(1):2–23.
48. Strang G. *Linear Algebra and Its Applications*. Orlando, FL: Harcourt Brace Javanovich, Inc., 1988.
49. Masters T. *Advanced Algorithms for Neural Networks: A C++ Sourcebook*. New York: John Wiley & Sons, Inc., 1995.
50. Mallat S. *A Wavelet Tour of Signal Processing*. New York: Academic Press, 1999.
51. Lio P. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics* 2003;19(1):2–9.
52. Masters T. *Signal and Image Processing with Neural Networks*. New York, NY: John Wiley and Sons, Inc., 1994.
53. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge, UK: Cambridge University Press, 1996.