Computational Aspects in Gene Expression Analysis

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

by

Noa Shefi

Supervised by Prof. Nir Friedman

November 2004

The School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

Acknowledgements

First and foremost, I would like to thank my supervisor prof. Nir Friedman for bringing me into the world of research and showing me the ways of science. I would also like to thank my dear friends in the lab: Gal Elidan, Dana Peer, Ariel Jaimovich, Tommy Kaplan, Yoseph Barash, Iftach Nachman, Ori Shachar, Matan Ninio, Omri Peleg, Hillel Fleischer, and Ilan Wapinski, for helping me in each and every step; without them, this work would have never been written. I am deeply thankful to Ronnen Segman, Naftali Kaminski and the rest of the PTSD team, for proving how fruitful an interdisciplinary collaboration can be. I am also grateful to Zohar Itzhaki for his part in the investigation of classification significance, and for working together. I am indebted to my family and friends for their patience and support along the way. And last but not least, I would like to thank Eli, for his wise advice, enormous support and infinite love.

Contents

Chapter 1

Introduction

1.1 From Genes to Microarrays

Molecular biology has made major steps towards the understanding of the human body in the last few decades. One of these steps was to understand the concept of DNA inheritance and protein synthesis. The DNA is a linear sequence of nucleotides (A, C, G and T), which carry all the information needed for the creation of life. Inheritance is possible due to the ability of the cell to replicate the DNA, and transform it to the inheriting cell. Unlike the dynamic creation and degradation of proteins and other molecules, the DNA sequence is static and does not change over time. It is built from two *complementary* strands that fit each other, where due to chemical association, if A occurs in one strand, T will occur in the other strand, and the same for G and C (the formation of this structure is called *hybridization*). The two strands create a *double helix*, which is folded into *chromosomes* (Figure 1.1).

According to the central dogma, the DNA encodes the inherited characters of the cell in short sequences called *genes*, which are transformed into the building blocks of life - the *proteins* [Lodish et al., 2000]. The proteins are shorter sequences, generated by repeating 20 *amino acids*, and are folded and form a 3D structure, which is crucial for their function.

How does the DNA sequence encode the protein sequence? Like in a regular text, it has ”words”: each triplet of nucleotides is translated to one amino acid; it also has ”sentences”: the gene has a beginning and an end, which define the *reading frame* of the sequence. The transformation from gene to protein is done in two steps. First is the *transcription* of gene to *mRNA*, which is an intermediate sequence, also built from four nucleotides (A, C, G and U), and is identical to the gene sequence (where U replaces T). The mRNA strand goes through further editing and is finally translated to the protein product (Figure 1.2). The opposite process which turns mRNA to DNA is called *reverse transcription*.

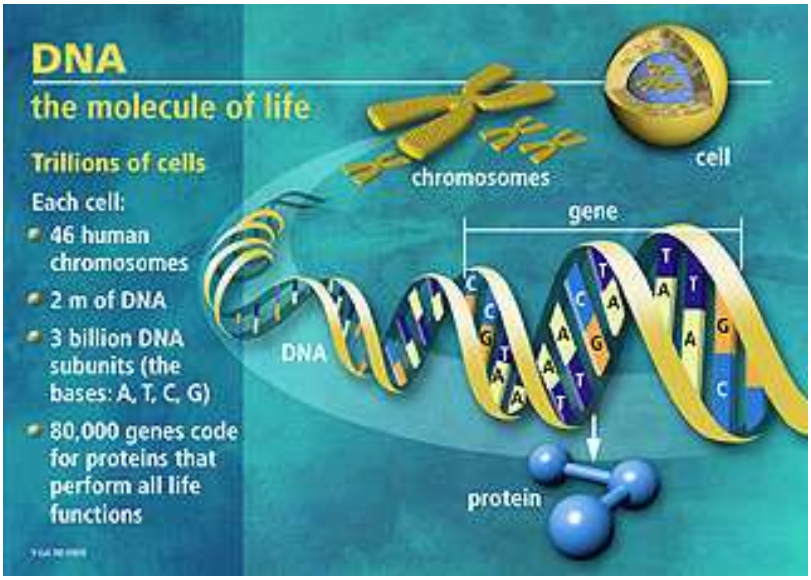


Figure 1.1: The DNA double helix, built from two hybridized complementary strands, and folded to create chromosomes. http://www.ucsc.edu/currents/00-01/07-03/dna.graphic.html

The transcription is done by the *RNA polymerase* protein complex. This "machine" binds to the DNA sequence and traverses it, while creating the mRNA sequence. However, the polymerase does not work alone; it has an ensemble of *transcription factors* (TFs) that regulate the transcription rate by binding to the sequence in specific *recognition sites*. Such binding can disturb the polymerase activity, resulting in transcription *inhibition*, or help it (for example by keeping open the double helix), resulting in transcription *activation*. Once it was created, the mRNA stays in the cytoplasm where it is translated to proteins, and after a while it is degraded. Therefore we say that the number of mRNA copies of a certain gene is proportional to the number of proteins of that gene. Nevertheless, the number of proteins in the cell is not necessarily identical to the number of active proteins: another regulation mechanism exists in the post-translational level, generated by chemical modifications on the proteins' residues, and by the binding of other proteins or ligands in the cell.

Following this paradigm, biological research has achieved tremendous discoveries. With the development of technology, the biological assays became more efficient, and today, computational biology research focuses on large scale assays. These assays can measure various parameters efficiently and accurately, for example: finding the DNA nucleotide sequence, measuring physical interactions between proteins and between proteins and DNA, and measuring the *expression* level of genes, i.e., how many mRNA copies of each gene, can be found in the cell. Analysis of large scale assays requires appropriate computational tools, and such tools will be presented in this work, mainly for the analysis of gene expression.

DNA microarrays is a technology that has reached technical maturity around 1998, and is

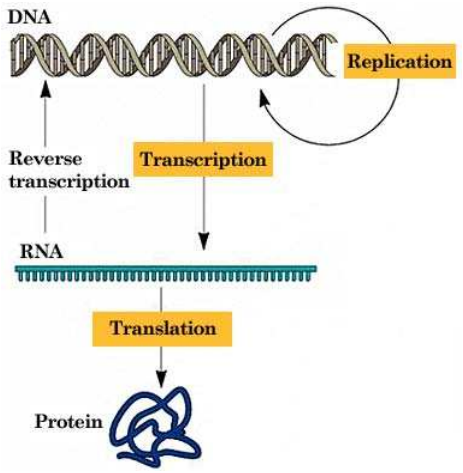


Figure 1.2: The central dogma in biology: the DNA is transcribed to mRNA, and the mRNA is translated to give proteins.http://cats.med.uvm.edu/cats_teachingmod/microbiology/courses/genomics/introduction/1_2_bgd_gen_pro_dog.html

now widely used in biological and medical research. It measures the number of mRNA copies of each known coding sequence, in a certain tissue and time point. In a microarrays experiment (Figure 1.3 [Brown and Botstein, 1999]) mRNA is extracted from two different samples. It is then reverse transcribed to give complementary DNA (*cDNA*), in the presence of fluorescently labeled nucleotides. To allow direct comparison of the abundance of every gene in the two samples, the cDNA pool from each sample is labeled with different fluorescent marker (green or red). The two pools are then mixed and put on a small silicon microarray, which contains thousands of spots with DNA probes; each spot contains probes which are specific to one gene. The individual transcript hybridizes specifically to the complementary probe in the array. Thus, the relative abundance of gene transcripts from one sample, compared to the those from the other sample, is reflected by the ratio of 'red' to 'green' fluorescence measured at the array element corresponding to that gene. This competitive hybridization is called a "dual channel" experiment, whereas in a "single channel" experiment, transcripts from a single sample are hybridized to the microarray, and then the total fluorescence is measured.

1.2 Computational Analysis of Microarrays

A single microarray sample can provide information about the mRNA expression levels in the sampled tissue. However, the question of expression changes between samples is more interesting than the absolute expression level in one sample. Thus, we usually consider several comparable samples, either time serialized or from distinct sources (e.g. tissue in several conditions). This kind of experiments enables to identify large scale transcriptional changes, that describe distinct active mechanisms in the different conditions. A typical dataset of a gene expression experiment consists of several dozens of conditions, and sev-

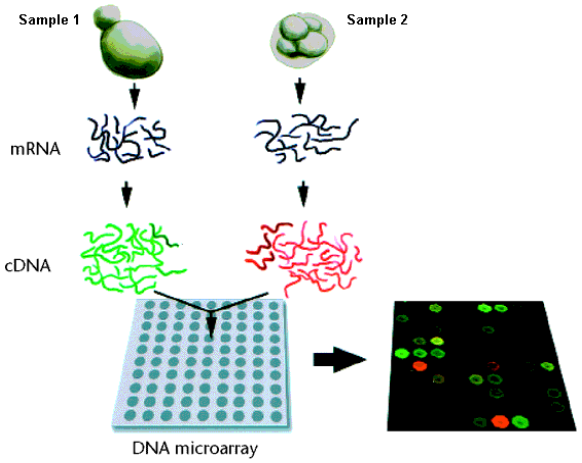


Figure 1.3: DNA microarray dual channel experiment

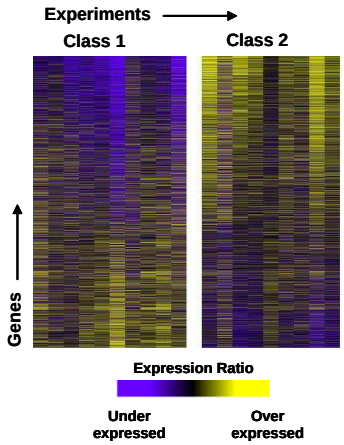


Figure 1.4: An expression matrix: entry (i, j) describes the expression value of gene i in experiment j. Values are in log-ratio, normalized in the average of each row.

eral thousands of genes (Figure 1.4).

The first question to be answered, is whether there is a group (or groups) of samples, that share a common expression pattern. Such a pattern may describe molecular properties associated with that group. Symmetrically, is there a group of genes that share a common expression profile over the samples? These genes are likely to be co-expressed in the cell, and may share a common function. To answer such questions, we use *clustering* methods, which divide either the samples or the genes to clusters. The genes (or samples) within the cluster have a similar expression pattern, which is different from the patterns in other clusters. Clustering methods are *unsupervised*, since the real classification of the features is hidden during the clustering. Some important works revealed in this way the existence of sub-types of diseases that were considered homogeneous before [Golub et al., 1999, Bittner

et al., 2000, Alizadeh et al., 2000].

Clustering of samples can point out the existence of expression patterns which are common to a group of samples, but the next computational challenge is to detect the genes that generate these patterns. Furthermore, we would like to find the pattern that is relevant to our biological question. For example, if we have two classes of samples from normal and cancerous tissues, we are interested in the genes that share a common expression pattern within each class, but are differentially expressed between the classes. These *differentially expressed genes* have biological and computational importance [Ben-Dor et al., 2000b]. They are biologically interesting because they reflect the difference between the sample classes, and therefore are probably relevant to the biological question. They are computationally important, when trying to classify samples of two classes, i.e., to predict to which class do the samples belong. According to the *Vapnik-Chervonenkis theory* [Kearns and Vazirani, 1994], the quality of the classification is dependent on the dimension of the problem. When the dimension of the problem increases, larger set of samples is required to keep on the classification quality[1]. In typical gene expression experiment, the sample set is small, relatively to the dimension of the problem, which is the number of genes. Therefore we would like to limit the dimension, by selecting only the features which are relevant to the problem. The differentially expressed genes are good candidates for such features, and are detected with various statistical tests.

Using differentially expressed genes we can learn a generalized pattern rule that predicts the classification of a new given sample [Ben-Dor et al., 2000a]. This can lead to the development of a molecular diagnosis tool, which is more sensitive than current diagnostic tools, and can be used in earlier stages of the disease. The hypothesis that is used to classify a new sample is called *classifier*. To build a classifier we learn a decision rule that associates class to expression pattern, using a training set of labeled samples. Then we test the classification ability of the classifier on a brand new set of unlabeled samples. An important step before learning a decision rule will be to reduce the dimension of the problem by selecting genes as described above. In contrast to clustering, classification is a *supervised* algorithm since the labels of the training set are known during the learning.

1.3 Information Integration for Learning Regulatory Networks

We can gain additional insight from expression profiles by integrating information from other sources, like sequence information, cellular function and location, physical interactions, evolutionary processes and more. These sources usually carry information on some of the genes or proteins, but not on all of them. Nevertheless, partial information can teach

[1] see Bishop [1995] for *the curse of dimensionality*

us a lot, and can be used to learn about the other genes or proteins. Consider for example the clustering method described above. Given a cluster of genes with similar profiles, we can look for a functional annotation that is common in the group, more than expected by chance. Such a test can validate the clustering result, and give us new information about unknown genes - by deducing that all genes with a similar profile, share that function (”guilt by association”).

Since expression level is mostly determined by a complex regulation system, genes that are co-expressed may be also co-regulated. The transcription factors that regulate expression bind to the genes in certain recognition sites in their promoters. The recognition sites are identified by sequence motifs. Therefore, an interesting test would be to check for the existence of a common motif in the gene promoters. Again, this test can validate the cluster we have found, and indicate potential transcription factors that regulate this cluster.
By connecting several parts of the huge biological puzzle we are able to reveal new information. Some works construct models that describe several aspects of biological processes at once. For example, describe a regulation network that consists of transcription factors and target genes, supported by various information sources such as expression data [Segal et al., 2003], physical interaction networks [Yeang et al., 2004], knowledge about regulation [Bar-Joseph et al., 2003], and sequence information [Segal et al., 2002]. These kind of models can be used both for evaluating observations, and for gaining new insights about the cell processes.

This work presents and discusses methods for gene expression analysis, and applies those methods in real-life medical practice: **Chapter 2** will review and discuss some of the existing methods for feature selection (2.1), and present a new selection method (2.1.8). It will review the concept of gene abundance analysis (2.4), and elaborate on the importance of its statistical validation. New methods that estimate the significance, using annotation from external sources and permutation test, will be presented for that purpose (2.5). **Chapter 3** will review clustering and classification methods. It will elaborate again on the importance of statistical validation and a correct learning procedure (3.6). **Chapter 4** will present and discuss a work that was done in collaboration with a medical team, in which we analyzed expression data that was measured in blood cells of post traumatic stress disorder patients (4.3). In this analysis we employ all the methods that are presented in the former chapters.

This work includes analysis examples of real expression data. Seven datasets are used for illustration purposes, but due to space limit, only part of the datasets are shown in each example. However, all the described phenomena are common to most or all of the datasets. The datasets describe human diseases, mainly cancer: **LUCA** - Lung cancer, **Breast** - Breast cancer, **Lymph** - Lymphoma, **Lukemia**, **Colon** - Colon cancer, **Adenocarcinoma** - Lung adenocarcinoma and **PTSD** - Posttraumatic Stress disorder. See appendix A for more details on the datasets.

All the methods described in chapters 1-4 were implemented. Their code is available in http://www.cs.huji.ac.il/∽ shefi/thesis/src/.

Chapter 2

Statistical Benchmark of Genes

2.1 Detecting Relevant Genes

Two samples with different biological characteristics (e.g. normal vs. tumor cells) are expected to have different gene expression profiles. However, most of the genes represented in the chip are not relevant to the question of interest, and are likely to have similar expression values in both samples. One of the important computational challenges is to identify the group of genes that are relevant to the question of interest and therefore are expected to have a distinct distribution of expression values in each class. These genes are potential targets for further investigation, as well as candidates for constructing diagnosis tools. For instance, a gene involved in proliferation process, is expected to be over-expressed in the tumor samples (which are characterized with aggressive proliferation of cells), compared to the normal samples.

To identify the relevant genes we make two assumptions: if the samples are divided into classes, we assume that the expression values of these genes will be driven from a different distribution in each class, and therefore they are *differentially expressed* between the classes. Secondly, we assume that the genes are independent from each other. As this assumption clearly simplifies the biology, it enables us to measure the relevance of each individual gene. Following these assumptions, our computational task is to identify potentially relevant genes based on their expression profiles.

Definition 2.1.1.: Assume we are given a dataset D, consisting of pairs $\langle x_i, l_i \rangle$ with $i = 1...M$, where the sample x_i is vector in R^N, $x_i[g]$ is the expression value of gene g in sample x_i, and l_i is the label of that sample. $l_i \in L = \{-, +\}$. Assume we have n negative samples and p positive, s.t. $M = p + n$.∎

For simplicity purposes, we focus here on two-labeled classification, which can be extended later to a larger number of classes.

Given a gene g, our *null hypothesis* is that the expression values of g in the two classes are driven from the same distribution. i.e., let D_i be the expression distribution of class $i = \{1, 2\}$, then

$$H_0 : D_1 = D_2$$

The *alternative hypothesis* suggests that each class has a different distribution:

$$H_1 : D_1 \neq D_2$$

The simplest test we use, is the *fold-change*, in which the null hypothesis is rejected if the ratio between the expression average in one class compared to the other, is big enough.

The following statistical scoring methods score the genes according to their ability to separate the samples. If the score is good enough, we will reject the null hypothesis, otherwise we will accept it.

The *parametric* scores, such as the students t-test, make assumptions about the form of the score distribution within each group. The *non-parametric* methods, such as TNoM, Info and Kolmogorov-Smirnov, do not make distributional assumptions about the expression values.

2.1.1 TNoM - Threshold Number of Misclassifications

A relevant gene g is expected to have different value distributions in the two classes, and therefore it can serve as a predictor: given a new sample x, we can check to which distribution $x[g]$ is closer, and classify x to this class. Specifically, we can find a threshold value that separates the class values, and then check if $x[g]$ is above or below the threshold. Formally we define the following *decision stump*:

$$l(x \mid t, d, g) = \begin{cases} d & x[g] > t \\ -d & x[g] < t \end{cases}$$

Where t is some threshold value and $d \in \{-1, +1\}$ is a direction parameter. Rewriting it, the prediction class is simply $sign(d \cdot (x[g] - t))$. The *error* of such a predictor gene is the sum of prediction errors over all samples:

$$Err(d, t|g, l) = \sum 1\{l_i \neq sign(d \cdot (x[g] - t))\}$$

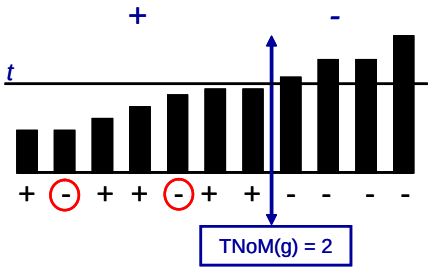


Figure 2.1: Example of threshold and direction that minimize the number of misclassifications.

Definition 2.1.2.: The TNoM score was defined by Ben-Dor et al. [2000a] as the number of errors made by the best decision stump, i.e.:

$$TNoM(j,l) = \min_{d,t} Err(d,t|g,l)$$

∎

To find the TNoM score, we arrange the expression values in increasing order s.t. $x_1[g] < ... < x_M[g]$. For each threshold t and direction d, we divide the vector into two groups - above and below t, and label the groups according to d. We then count the number of misclassified samples. The TNoM score is the minimal number of misclassifications, over all possible t and d (Figure 2.1).

2.1.2 Estimating Score Significance

An immediate question is whether a gene with a low TNoM score is "good enough" to reject the null hypothesis. According to the lemma of Neyman and Pearson [DeGroot, 1989], it means that the probability to find such a score under the null hypothesis is much smaller than the probability to find it under the alternative. The *p-value* of a score is the probability of a gene to attain such a score or a better one, under the null hypothesis. A low p-value means that genes with such a score are rare in a random dataset, therefore they probably carry some relevance to the studied classification.

Definition 2.1.3: Let S denote a scoring metric, and g denote a gene that received the score s, then:

$$pVal(g) = Prob(S(g) \leq s)$$

∎

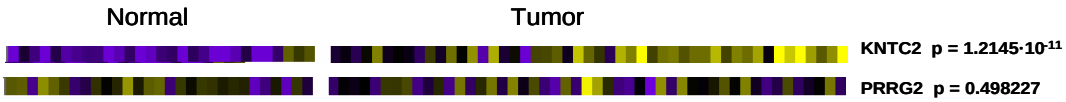


Figure 2.2: Expression pattern of two genes (LUCA data set), with their associated p-values. KNTC2 (kinetochore associated 2) is involved in spindle checkpoint signaling, during cell division. it is known to be highly expressed in cancer. PRRG2 is a proline-rich Gla (G-carboxyglutamic acid) polypeptide 2.

The p-value is calculated under the null hypothesis of random labels, while the size of each class is kept. This is a relaxed definition of the original null hypothesis $(D_1 = D_2)$, as we do not have the distribution parameters.

Practically, we choose to reject the null hypothesis, if the p-value is smaller than a certain threshold. The common convention is $pVal \leq 0.05$, or a stricter threshold, which is determined with various methods (Section 2.2). An example for a separating gene and a non-separating one, is shown in Figure 2.2. It is clear that KNTC2 is differentially expressed between the classes, as it is over expressed in the tumor samples, and under expressed in the normals. PRRG2 is not differentially expressed.

2.1.3 Calculating TNoM's p-Value

The combinatorial character of TNoM makes it amenable to rigorous calculations. Ben-Dor et al. [2000a] developed a recursive procedure that computes the exact distribution of TNoM scores, and it is outlined here:

Definition 2.1.4: Let

- $v = \{-, +\}^{n,p}$ be the labels of the ranked expression vector of gene g
- $p_v(i)$ be the number of samples labeled $+$ within first i samples
- $n_v(i)$ be the number of samples labeled $-$ within first i samples
- $\pi_v(i)$ be $p_v(i) - n_v(i)$

∎

Now let us look at the paths in R^2 through the points $(i, \pi_v(i))$ (Figure 2.3). There is a one to one and onto mapping from labels vector v to a path in the grid, which starts at $(0, 0)$ and ends at $(p + n, p - n)$, denoted by $\Pi[(0, 0) \rightarrow (p + n, p - n)]$. Given p and n, all paths are bounded by the paths of the perfect classifiers (Figure 2.3, diagonal lines).

The above definitions leads to the following conclusion (proven in [Ben-Dor et al., 2000b]):

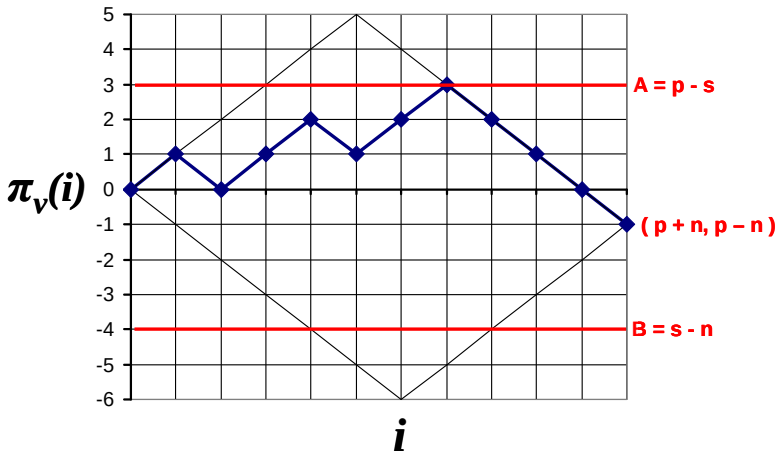


Figure 2.3: Path on the grid that describes the gene from Figure 2.1. $p = 5, n = 6, s = 2$

Theorem 2.1.5: TNoM$(v) \leq s$ iff there is an i such that $\pi_v(i) \geq p - s$, or $\pi_v(i) \leq s - n$.

So the probability to get a score s is the probability that a random path on the grid crosses those lines. Formally:

Marking $A = p - s, B = s - n$ then:

$$pVal(s) = Prob(TNoM(v) \leq s) = \frac{\nu(A, B)}{\binom{M}{p}}$$

Where $\nu(A, B) =$ number of paths $\Pi[(0, 0) \rightarrow (M, p - n)]$ that cross A or B, i.e., $\pi_v(i) \geq p - s$ or $\pi_v(i) \leq s - n$.

To calculate $\nu(A, B)$ we use the repeated reflection principle: The number of paths $\{\Pi[(0, 0) \rightarrow (p + n, p - n)]\}$ that cross A, is equal to the number of paths $\{\Pi[(0, 2A) \rightarrow (p + n, p - n)]$, where A is a mirror axis. Using this and the inclusion/exclusion principle, we count exactly the number of pathes that cross A or B at least once.

The TNoM score is intuitive and simple to implement. Due to its discrete character, its p-value distribution has a typical ”steps” form. Comparing the TNoM to the basic fold-change test, we find that TNoM is able to detect finer cases of separation (Figure 2.4).

However, TNoM still has drawbacks:

- Its p-values are distributed in discrete bins, and there is no way to determine which

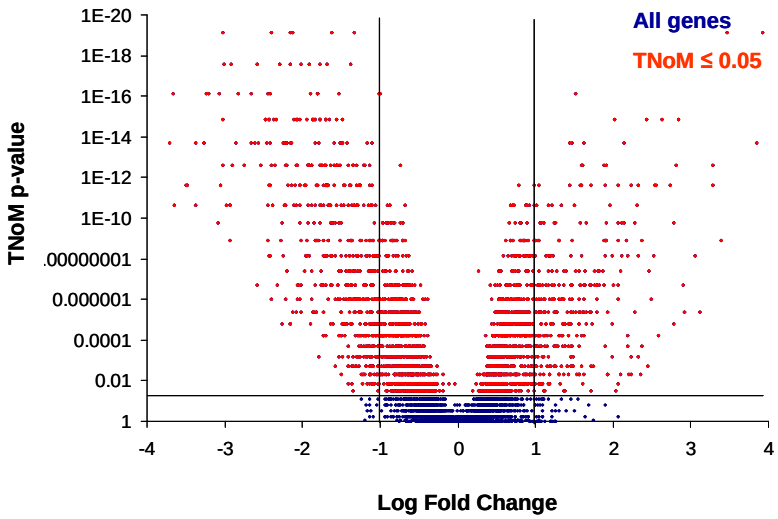


Figure 2.4: TNoM p-value distribution vs. fold-change ratio. Red dots mark genes with $pVal(TNoM) \leq 0.05$. (LUCA dataset)

gene is more informative within the bins. We would generally like a method which scores the genes on a more refined scale.

- The p-values have lower bound of $\frac{1}{\binom{p+n}{p}}$. In some cases even the best genes will not pass the correction thresholds (Section 2.2), and in the extreme case of low n and p, even the most informative gene will have p-value > 0.05.
- The type of error - false positive or false negative is not being considered in the current definitions: k false positive errors will have the same score as k false negative errors.
- The classification quality is not considered. i.e., TNoM does not distinguish a rule that makes k errors on a single class, and a rule that makes $k/2$ errors on each class. In this case the first rule will perform very badly on the class with the errors.

The Last problem of classification quality is addressed by the following mutual information and Kolmogorov-Smirnov scores, which are also non-parametric methods.

2.1.4 The Mutual Information Score

The *entropy* of a variable describes how much uncertainty do we have on it. When the variable has only one value, the entropy is minimal, and when it gets several possible values with uniform distribution, the entropy is maximal. The Info score calculates the relative entropy of a ranked label vector, i.e., how much uncertainty do we have on the

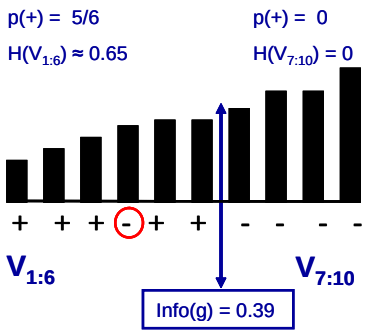


Figure 2.5: Example of Info score calculation

sample labels, given the expression values. To describe Info, first let us present the formal definition of entropy:

Definition 2.1.6: [Shannon, 1948] The entropy of a random variable X is:

$$H(X) = -\sum p(x) \cdot \log p(x)$$

where $x \in X$ ∎

Hence, the maximum entropy is accepted when the values of X are uniformly distributed.

Let v be the vector of expression values $x_1[g] \leq ... \leq x_M[g]$, which are divided into two sub vectors $v_{1:i}$ and $v_{i+1:M}$ using some threshold value like in TNoM.

Definition 2.1.7: [Ben-Dor et al., 2000b] The Info score of v is:

$$INFO(v) = \min_i \{\frac{i}{M} \cdot H(P_{1:i}) + \frac{M-i}{M} \cdot H(P_{i+1:M})\}$$

∎

Where $H(P_{1:i})$ is the entropy of labels distribution in the sub vector $v_{1:i}$. (Figure 2.5)

When the vector is divided into two sub vectors with homogenous labels, the Info score will be zero. The highest score will be received when $i = M/2$, and the labels are uniformly distributed between the sub vectors.

2.1.5 The Kolmogorov-Smirnov Score

The Kolmogorov-Smirnov (KS) test [DeGroot, 1989] estimates the distance between two distributions, according to an empirical sampling of these distributions.

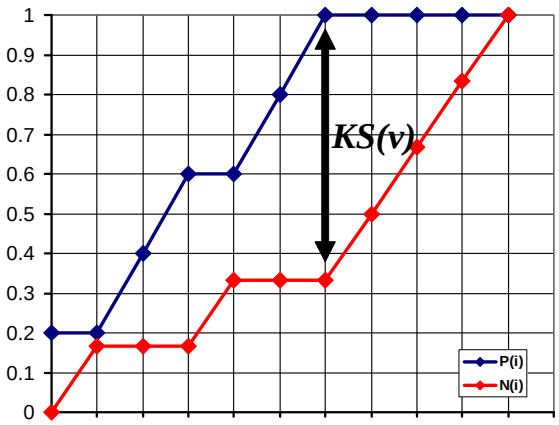


Figure 2.6: Calculation of KS score: $p_v(i)$ is the fraction of positives in the first i samples (blue) and $n_v(i)$ is the fraction of negatives in the first i samples (red).

Given a ranked vector v with p positive samples and n negative samples, let $p_v(i)$ and $n_v(i)$ be as in TNoM's definition. We will define two *empirical distribution functions* to be:

$$N(i) = \frac{n_v(i)}{n}$$
$$P(i) = \frac{p_v(i)}{p}$$

The KS score is:

$$KS(v) = \sup_i |N(i) - P(i)|$$

i.e., the maximal distance between the two function (Figure 2.6).

According to Gilvenko-Cantelli lemma, under the null hypothesis

$$\lim_{n,p\to\infty} KS(v) = 0$$

Therefore, the test rejects the null hypothesis if the distance is large enough. The exact threshold is determined according to the score p-value.

2.1.6 Mutual Information and KS p-Values

Both Info and KS calculate their p-values in a similar way to TNoM. However, in their case the definition of boundaries (A and B in TNoM) is more complicated. $INFO(v) > s$ iff

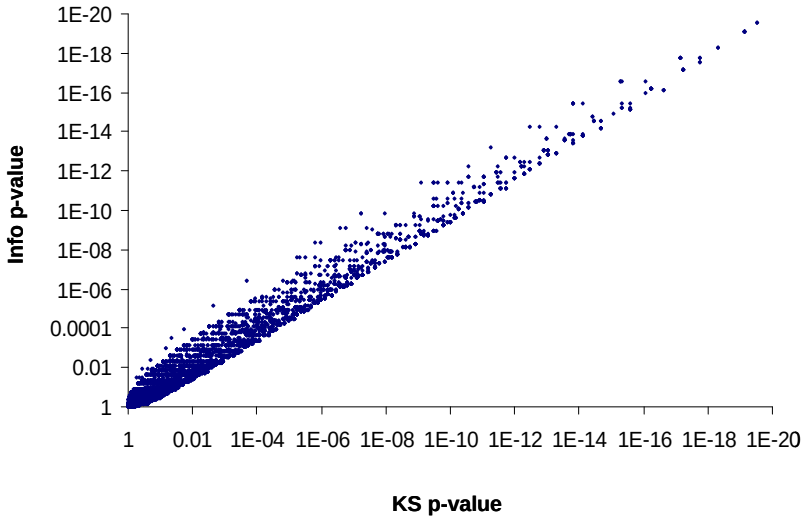


Figure 2.7: KS p-values vs. Info p-values (LUCA dataset)

the relative entropy of v is greater than s in each partition i. For a given s, Ben-Dor et al. [2001] calculate the area $\mathcal{R}(s)$ of the grid, in which the relative entropy is greater than s, and count the number of paths within $\mathcal{R}(s)$, using a dynamic programming. The p-value is:

$$Prob(INFO(v) \leq s) = 1 - \frac{\text{\# paths within } \mathcal{R}(s)}{\text{total \# paths}}$$

TNoM, Info and KS are close methods. They all look for the division of the vector that optimizes functions of $p_v(i)$ and $n_v(i)$. Considering this, it is not surprising to find out that KS and Info tend to "agree" about the score and give the genes similar p-values (Figure 2.7). When $n = p$ we can notice that the KS choice of i is identical to TNoM's: $TNoM(g) = \min_i\{p - \pi_v(i), p + \pi_v(i)\}$, and $KS = \sup_i |\frac{\pi_v(i)}{p}|$. Both reach the optimum in $\sup_i |\pi_v(i)|$.

The Info and the KS scores express the quality of the classification, and in this way they are better than TNoM. But they still carry some of the drawbacks that were listed for TNoM:

- Info and KS are both discrete, in the sense that they depend only on the ranked vector labels. The number of ways to order the labels vector is finite, resulting in bins of scores and a lower bound on the p-values. However, using the terminology of probabilities gives Info and KS the ability to have higher resolution, and to distinguish between genes that had identical TNoM scores. The typical distribution of Info and KS p-values looks much smoother than TNoM's.
- The type of error - false positive or false negative is not being considered.

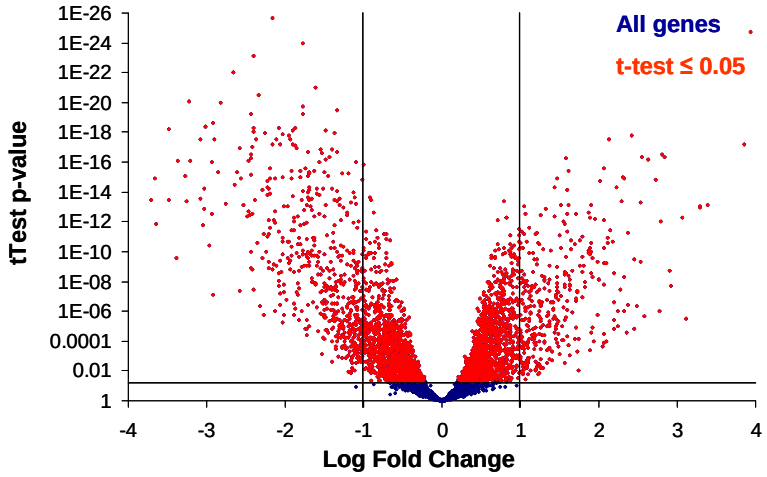


Figure 2.8: The t-test p-value distribution vs. fold-change ratio. Red dots mark genes with $pVal(t-test) \leq 0.05$. (LUCA dataset)

- All the three methods have one major drawback, as they do not take into consideration the *margin* of the separation, i.e., how much are the distributions far from each other. Genes that separate the distributions with different margins may have the same score, because only the labels are being considered, while the expression values are ignored. To address this problem we use the parametric *students t-test*.

2.1.7 The Student's t-test Score

This score considers two groups of expression values a and b, driven from normal distributions with means μ_a and μ_b and variance σ^2. The t-test score [DeGroot, 1989] decides whether the values were sampled from the same distribution (the null hypothesis), or from separate distributions (the alternative)[1].

To accept or reject the null hypothesis, the following statistic is calculated, and then its p-value is calculated according to the *t-distribution* [DeGroot, 1989].

$$t = \frac{\sqrt{(n_a + n_b - 2)}(\bar{X}_a - \bar{X}_b)}{\sqrt{(\frac{1}{n_a} + \frac{1}{n_b})(n_a + n_b)S^2}}$$

Where n_a, n_b are the number of samples in a and b respectively, $\bar{X}_a$, $\bar{X}_b$ are the estimators to the means, and S^2 is the estimator of the variance.

[1] Other version of t-test assumes different variance for each distribution.

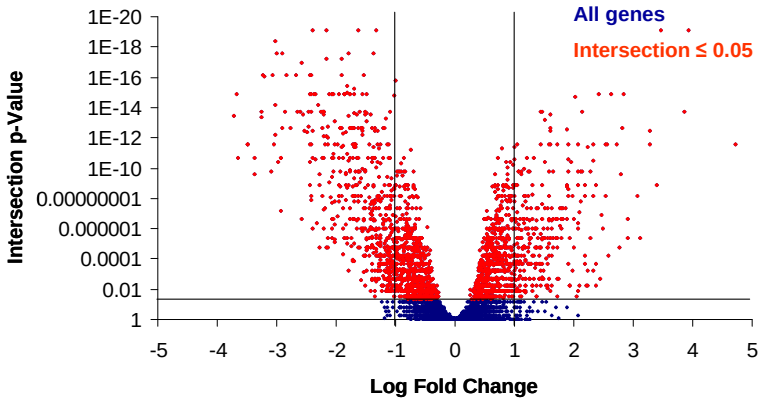


Figure 2.9: Intersection p-value distribution vs. fold-change ratio. Red dots mark genes with $pVal(Intersect) \leq 0.05$. (LUCA dataset)

Unlike TNoM, Info and KS, the score of t-test is continuous, which gives a good scaling of the genes (Figure 2.8). Genes that had identical score in the non-parametric methods, may have now different scores, that express the distance between the empirical distributions.

Still we are not satisfied with t-test, mainly because of the normal distribution assumption. This assumption is reasonable in cases where we have many samples, according to the central limit theorem [DeGroot, 1989]. However, when the number of samples is small, the distributions may not be normal.

2.1.8 Integrating Several Methods

Trying to reduce the number of errors among the selected genes, we constructed a method that intersects several scoring methods. The idea is that a gene which is significant according to several methods, is more likely to be truly relevant to the classification.

Formally, for a given gene g and scoring methods p-values $S_1 ... S_N$,

$$Intersect(g) = \max\{S_1, ..., S_N\}$$

The intersection of TNoM, Info and t-test results in a score which is neither discrete like TNoM, nor continuous like the t-test. It has a typical steps form (Figure 2.9) due to the integration of the scores.

The *intersection score* has a straightforward intuition, but does not have a well defined model as the other methods. The meaning of intersection p-value is not the probability to get such an intersection score, but the probability to get the score with the strictest method. Therefore it is better to treat it as a score between 0 to 1 and not as a p-value.

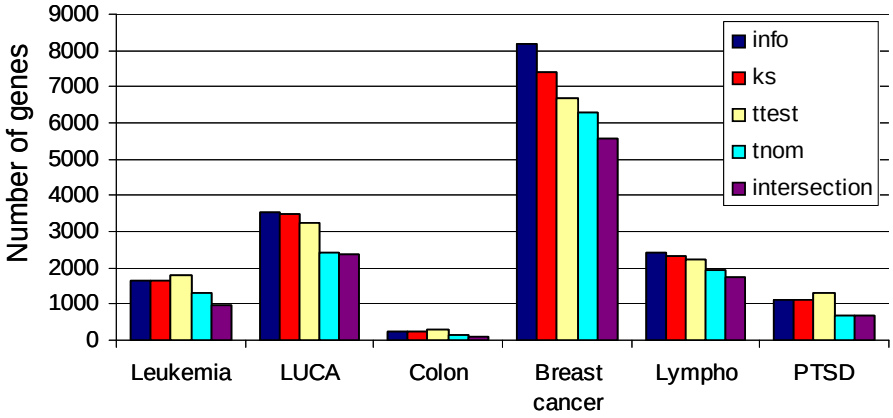


Figure 2.10: The number of informative genes with p-value ≤ 0.05, that were selected by the different methods in six datasets.

2.1.9 Scoring Methods Comparison

We have presented several methods, parametric and non-parametric. Generally we can say that the advantage of the non-parametric methods is that they do not assume anything on the sample distributions. Their major drawbacks are the discrete character, and the inability to score different margins. The t-test score solves the two last problems, but in the cost of a strong assumption on the distributions.

Comparing scoring methods is not an easy task. If we had information about the true relevant and irrelevant genes, we could calculate the sensitivity and specificity. However, we do not have such information: in real life many genes have indirect relevance to the classification, and there is no clear-cut between the relevant and the irrelevant genes. In addition, the information about the genes is usually lacking. As a result, we can only try and simulate the methods performance on synthetic data in which the expression values are being sampled either from a single distribution or from two distinct distributions. Good simulation, that describes well what happens in a real biological experiment, is a very ambitious goal.

A possible solution would be to estimate the ability to classify samples (see Section 3.3) with the different methods. Such experiment revealed that there is no method which is significantly better than the others, but it varies between the datasets (data not shown).

Nevertheless, we are able to compare two aspects: first is how much strict or relaxed is the method? Note that relaxed method will have higher sensitivity and lower specificity. Second, is how much do the scoring methods agree? If they always agree on the informative genes, then comparing them is meaningless.

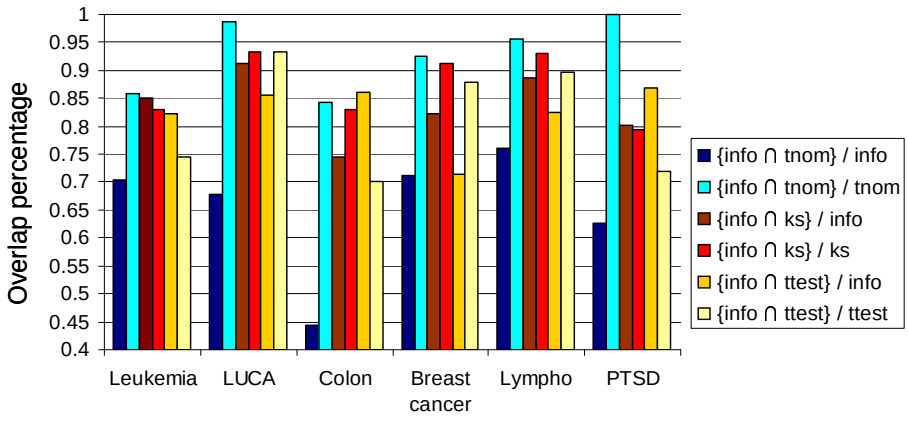


Figure 2.11: Overlapped fraction of selected informative genes (with p-value ≤ 0.05), according to the different methods in six datasets. (Appendix A)

To answer to the first question we measured the number of informative genes with p-values smaller than a certain threshold, according to the different methods, over six datasets, of Human diseases (Appendix A). The results (Figure 2.10 for $p \leq 0.05$) show that the TNoM score is stricter than the others, and that the Info and KS select a similar number of genes. The t-test score behaves differently over the datasets, sometime more strict and sometime more relaxed than others. Since the intersection score select the consensus genes of TNoM, Info and t-test, it obviously the strictest test.

The answer to the second question is strongly related to former results. TNoM is much stricter than Info, so most of the genes selected by TNoM (more than 84%, Figure 2.11), are selected also by Info. The relative part of Info genes which is common to TNoM, is low (44% to 76%). A second observation is that Info and KS agree on most of the genes - usually above 80%. This may result from the fact the both Info and KS are functions of $p_v(i)$ and $n_v(i)$. The overlap between t-test and Info varies from 70% to 93% of t-test genes, what may result from the inherited difference between the methods.

2.2 Statistical Corrections

Let us assume now we have 10,000 genes in the dataset, among them we find 600 genes with p-value ≤ 0.05. Under the null hypothesis of uniform p-values distribution, we would expect to get 5% of the genes with this p-value, i.e., 500 false positive genes[2]. Meaning that most of our selected genes are false. How do we set the threshold to filter out the false positive genes? One solution adapts the threshold to the number of features, using the

[2]Here the genes that are relevant to the biological classification are *positives*, and the irrelevant genes are *negatives*. Therefore the irrelevant genes among the selected ones are the *false positives*.

union bound principle:

Definition 2.2.1.: Let S be some scoring metric, with p-value $pVal_S$. Let P_0 be the p-values distribution under the null hypothesis, $p^* = \min_g pVal_S(g)$ is the p-value of the best gene. The Bonferroni Correction [Bonferroni, 1936] finds a threshold t, given a desired significance level α, and number of genes N:

$$P_0(p^* < t) \leq \sum_g P_0(pVal_S(g) < t) = N \cdot t \leq \alpha$$

$$\Rightarrow t \leq \frac{\alpha}{N}$$

∎

So instead of choosing genes with $pVal \leq \alpha$, we will only choose genes with $pVal \leq \frac{\alpha}{N}$.

The Bonfferoni correction, however, is a very strict correction. It bounds the chance that the best gene is false positive to α. But what about the second best gene? If it is also beyond the threshold we would be more surprised since under the null hypothesis, the ith gene should get p-value of $\frac{i}{N}$[3]. So now we would like to choose the ith gene, if it has p-value $\leq \frac{i}{N}\alpha$

This alternative approach is employed in the "False Discovery Rate" (FDR) correction [Benjamini and Hochberg, 1995].

Definition 2.2.2*:* FDR Correction:
Let $p_1 \leq ... \leq p_N$ be the ordered p-values of the genes. Find $\max_k$ s.t.

$$p_k \leq \frac{k}{N}\alpha$$

FDR threshold will be p_k ∎

This means that we choose all the genes with surprising low p-values. There is no rule to which correction should be used. This decision depends on the abundance of informative genes in the specific dataset, and the importance of false negative errors. Table 2.2 shows the number of genes selected in each data set, with $\alpha = 0.05$:

[3]Under the null hypothesis the probability to get the same score as the ith gene is $\frac{\{\#g:S(g)<S(g_i)\}}{N}$, and that denotes the p-value.

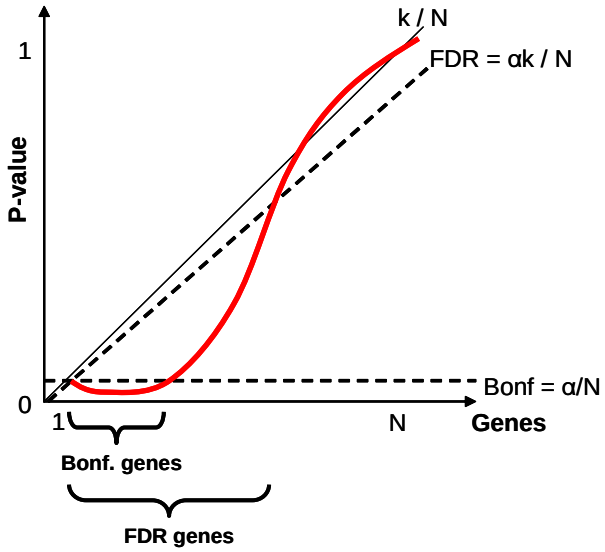


Figure 2.12: Bonfferoni and FDR corrections result in a different group of predicted positives genes, on a hypothetical p-values distribution.

Data Set	Classification	Genes passed Bonferroni	Genes passed FDR
Adenocar	Stg1vs3	5	129
Colon	NvsT	11	60
Lympho	DLCLvsNon	617	2003
PTSD	PvsC	9	538
Breast Cancer	BRCAvsAll	553	2604
Leukemia	NvsT	197	986
LUCA	NvsT	1099	3043

Figure 2.12 shows the principle difference between the corrections on hypothetical data: the red plot describes the p-values of the N genes. the break lines mark the Bonferroni and FDR thresholds, where the genes with p-value below these thresholds will be selected.

2.3 Dealing with Continuous Parameters

So far we have discussed how to find informative genes, which are differentially expressed between the classes. Sometimes, however, the question of interest is not the difference between dichotomic classes, but the change in expression along a continuous scale. Such scale may be time, for example Spellman et al. [1998] found a group of genes whose transcript levels vary periodically within the cell cycle of the yeast. Other parameter may be a clinical score that indicates the status of the tissue, or the severity of the disease. In these cases we look for genes with expression pattern that correlates with the external parameter. Such genes may be the reason for the parameter or its result, therefore they are biologically meaningful. To find those genes we calculate the correlation between the expression pattern of each gene and the external parameter; the correlation between variables is the degree to which the two are correlated. We will present here the non-parametric *Spearman Rank Correlation* and the parametric *Pearson correlation*.

2.3.1 The Spearman Rank Correlation

The simple Spearman Rank Correlation [Lehmann and D Abrera, 1998] estimates the distance between two vectors using their raking.

Definition 2.3.1: Let X and Y be random variables, and let $\{X_i\}, \{Y_i\}, i = 1..N$ be ranked vectors of values sampled from their distributions. Let r_{X_i} be the rank of sample X_i in the ranked vector, and so is r_{Y_i}. The rank correlation is:

$$r = 1 - 6 \sum_i \frac{d_i^2}{N(N^2 - 1)}$$

∎

Where $d_i = r_{X_i} - r_{Y_i}$, is the difference between the ranks of sample i in X and Y values.

2.3.2 The Pearson Correlation

Pearson correlation is a parametric method that estimates the linear relationship between two variables, while making assumptions on their distributions.

Definition 2.3.2.: Let μ_x, μ_y be the means of the variables X and Y, and σ_x, σ_y be their standard deviations. The *Pearson correlation* is simply:

$$\rho = \frac{E|(x - \mu_x)(y - \mu_y)|}{\sigma_x \sigma_y}$$

Given $\{X_i\}, \{Y_i\}$, we can estimate ρ. Let:

$$S_{xx} = \Sigma_i (X_i - \bar{X})^2$$
$$S_{yy} = \Sigma_i (Y_i - \bar{Y})^2$$
$$S_{xy} = \Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})$$

The estimator for ρ will be:

$$r = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

If the Pearson score is normally distributed around 0, we can easily find its p-value using the z table. However, with high correlation values, the distribution tends to have a negative skew. In these cases a transformation called *Fisher z' transformation* converts r to a normal distributed variable z' with standard error of $\frac{1}{\sqrt{N-3}}$. z' is used to calculate the p-value of r [http://davidmlane.com/hyperstat/A98696.html].

The use of Pearson correlation can be illustrated on the PTSD dataset, in which we have external parameter that describes the severity of the Post Traumatic Stress disorder. Measuring the correlation between expression and severity score, we have found many correlated and anti-correlated genes (Figure 2.13). Note that the correlation p-value is equivalent to a scoring method, therefore it can be employed for gene selection.

2.4 An Overabundance Analysis

In former sections we presented methods that score genes for their relevance to the biological attribute, and formulated the probability to get such a score by chance. To get a more comprehensive picture, we can measure the number of informative genes that were scored with a given p-value, and the probability to find this number by chance.

Looking over all p-values at once, we can get a good estimation for the separability of our data, i.e., to the information it carries. Dataset which has an abundance of informative genes, probably reflects a real biological phenomenon. We would like to detect such data, and to have a parameter which is comparable between datasets and between classifications.

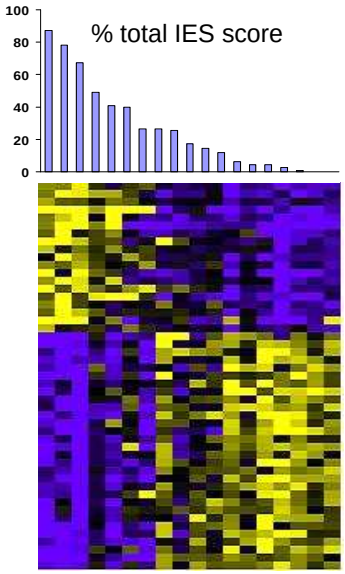


Figure 2.13: Expression patterns of genes with high correlation or anti-correlation ($r < -0.7$, $r > 0.7$) to the severity score (IES total, top histogram). The complete results are presented in Chapter 4.

2.4.1 Abundance Plots

Overabundance plot [Ben-Dor et al., 2001] describes for increasing p-value, the number of genes that were actually scored with this p-value or better, along with the number of genes expected by chance (Figure 2.14 (a)). The null hypothesis assumes uniform distribution of p-values, therefore the number of expected genes for p-value p, is $N \cdot p$. Looking over several classifications or dataset, we can easily detect the one with more informative genes (Figure 2.14 (b)).

Overabundance is simple and convenient methodology which gives a quantitative measurement to the separability of the dataset, or to its correlation to an external parameter, while other methods as the volcano plots, give only qualitative measurement (Figures 2.4, 2.8). Nevertheless, two problems in the current definition should be addressed:

- First, a quantitative parameter is needed to enable the comparison between overabundance plots, such as the maximal distance of the real plot from the expected plot. This problem is addressed, for example, in the max-surprise score [Ben-Dor et al., 2001].
- The second problem is the assumption that the scores are uniformly distributed under the null hypothesis. De facto, when estimating the scores distribution with random labels, we see that often they are not uniformly distributed. This may be due to the small number of samples and the dependencies between the genes. In the next section we address this problem by calculating empirical p-value.

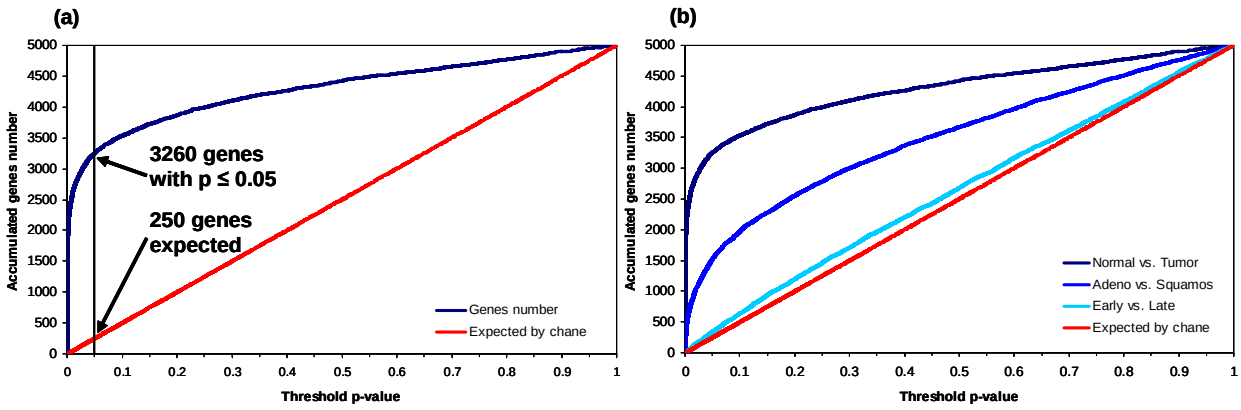


Figure 2.14: (a) Overabundance analysis of LUCA dataset. The blue plot describes for each p-value the number of genes that were scored with that p-value or less, according to the t-test method. Red plot describes the number of genes expected under null hypothesis. For instance, 3260 genes got $pVal \leq 0.05$, while 250 were expected by chance. (b) Overabundance plots with LUCA dataset, for three optional sample classifications.

2.5 Results Validation

Let us assume we have analyzed our data, and got a set of informative genes. A desirable goal would be to validate our results, both statistically and biologically.

2.5.1 An Empirical p-Value Estimation

The statistical validation estimates the significance of the results, i.e., the probability to get such results by chance. In the context of overabundance we will first ask, what is the probability to find k informative genes with a given p-value or better. Then we ask for the probability to get the whole overabundance plot.

To calculate the probability of finding k informative genes, we would like to have a background distribution. Since rigorous model of the background distribution is usually missing, we build a distribution model, under the null hypothesis of random labels.

Definition 2.5.1: Given dataset D consists of pairs $< x_i, l_i >$ with $i = 1...M$, the *random permutated dataset* D' is a new dataset consists of pairs $< x_i, l'_i >$, where $L' = l'_1...l'_M$ is a random permutation of the original labels vector $L = l_1...l_M$. ∎

In Random permutation test we repeat the examined test n times, each with different permutated dataset, to get n test results, sampled from the background distribution. Let $G_t(i)$ be the number of genes that got p-value better than t in the ith repeat, then the probability to find k informative genes is simply:

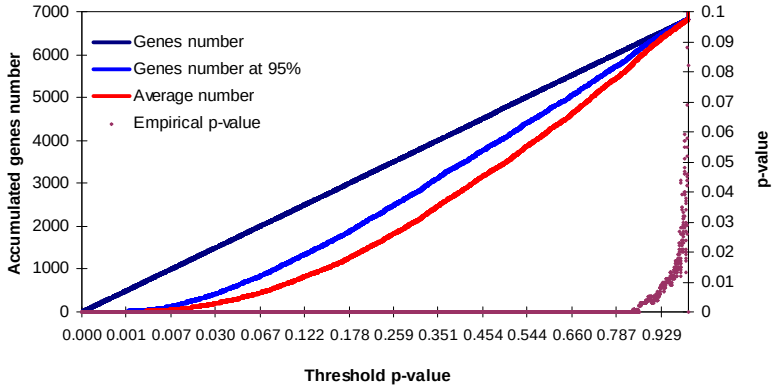


Figure 2.15: Empirical Overabundance plot: "Genes number" is the abundance of genes with the real labels. "Empirical p-value" marks for each point in the abundance graph the probability to find such amount of genes by chance. "Average" is the average number of genes over the random runs, which is the empirical expectation. "95%" marks the 95 percentile of the random plots. The graph was generated with Leukemia dataset, t-test scoring method and 1000 random permutations.

$$Prob(G_t = k) = \frac{\sum_i 1\{G_t(i) \geq k\}}{n}$$

To calculate the probability to find abundance of genes which correlate to an external parameter, we do the same permutation test, but instead of permuting the labels, we permute the external parameter values.

The significance of an overabundance plot is not so straightforward. We can calculate the probability of each and every point in the graph (Figure 2.15, "empirical p-value" plot), but that will not give us the probability of the plot as a whole. Since order relation is not well defined over the plots, we cannot estimate directly the p-value of a plot. What we can do is estimate the p-value of some parameter that describes the quality of the plot, such as the max-surprise or Sanov scores [Ben-Dor et al., 2001]. Another solution is to calculate the p-value in each point in the graph, and claim that the plot has p-value ≤ 0.05 if every point in the graph has p-value ≤ 0.05 (Figure 2.15, "95% plot").

The statistical validation of the results is essential, as sometimes there is an abundance of informative genes, but it has no statistical significance. In this case, random data would probably generate similar abundance (Figure 2.16).

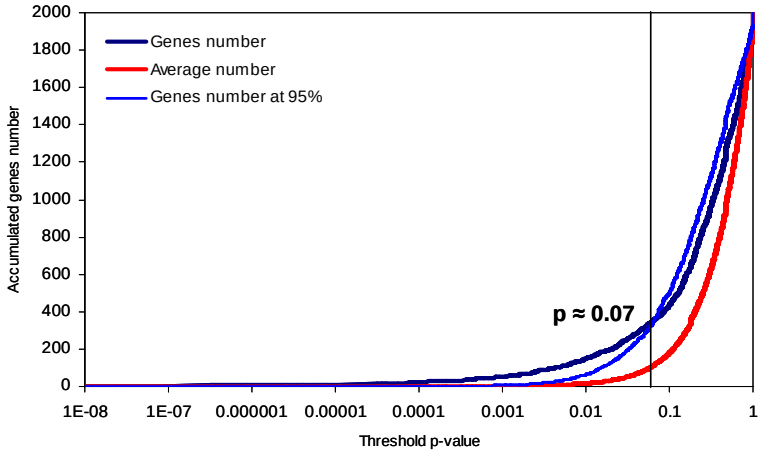


Figure 2.16: An example to overabundance which is not significant, starting from threshold p-value $\approx$ 0.07. The abundance of genes is below the 95% plot, meaning that more than 5% of the random runs got such abundance of genes, or more.

2.6 An Analysis of Annotation Enrichment

The biological validation addresses both the question of significance and the results interpretation. If we have an external parameter, which is completely independent in the parameters that were used for analysis, we can use it to estimate the significance of our set of genes, and to interpret them. Let us assume that the candidates have external annotations. If we detect a common annotation within the positive group (i.e. the selected genes), we can conclude that the group is meaningful, and learn about its members.

To emphasize let us take an hypothetic example of a dataset, which consists of healthy and tumor tissues. The genes which are differentially expressed between the two tissue types, are the positives. Let us assume we annotated all the genes according to their cellular role (e.g. restriction enzyme, kinase, G-protein, transcription factor etc.), and we found that among the positive genes, many are annotated as transcription factors. A reasonable conclusion would be that the molecular difference between healthy and tumor samples is related to transcriptional processes.

2.6.1 Statistical Significance of Annotations

In the example above we found many genes with a “transcription factor” annotation. The requested question here, is how much is ”many”? Fortunately the well established *hyper geometric* model describes the probability of our findings.

Definition 2.6.1: The Hyper Geometric Distribution [DeGroot, 1970]

Let us assume we have N features, of whom K are "special" and $N - K$ are not. We randomly choose (with no returns and repeats) a group of n features from this set. The probability to find k "special" features among the chosen n is:

$$Hyper(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

∎

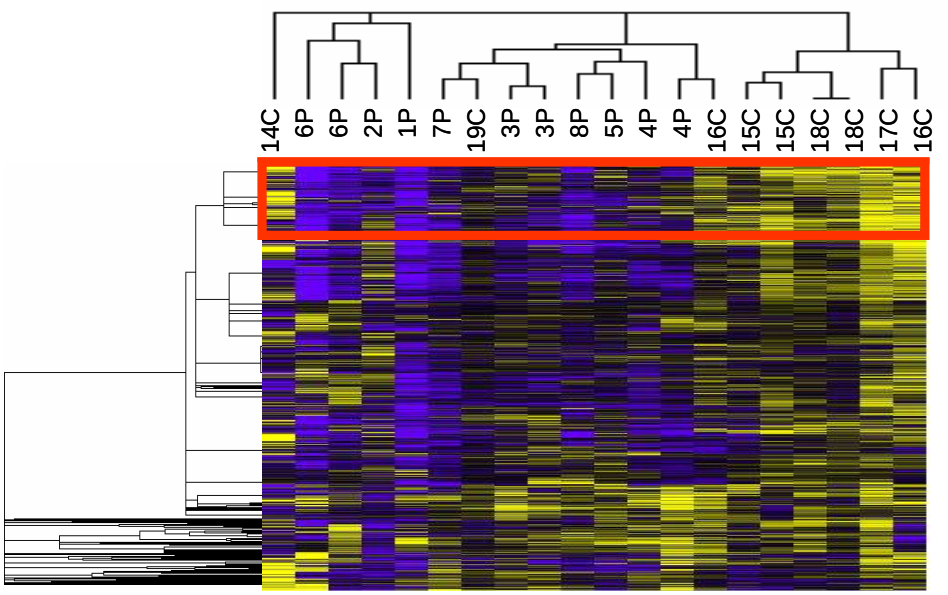
The intuition is very simple: the number of ways to choose k from K special features, and $n-k$ from $N-K$ non special, divided by the total number of ways to choose n features. In our case N is the number of genes in the dataset, K is the number of genes with a specific annotation, n in the number of informative genes, and k is the number of genes with that annotation, among the informative genes. In this way we compute the probability to find such annotation enrichment in our group of genes, under the null hypothesis of randomly chosen group. Figure 2.17 describes annotation analysis in a cluster of genes, obtained by implying double hierarchical clustering on the PTSD dataset (see Section 3.2). One cluster (red) was analyzed with GO annotations (Gene Ontology [Consortium, 2001]), and found to be enriched with annotations which are related to RNA and DNA processing.

2.6.2 Using Annotation Analysis to Integrate External Data

In the introduction we discussed the importance of integrating several data sources. In annotation analysis, the integration is immediate. The external parameter can come from various sources, each gives a different interpretation to the group of interest:

- The *Cellular role* annotation describes the molecular function of the genes, but it does not point on specific process that is being activated.
- The annotation of *molecular process* (e.g. cell-cycle, amino-acids synthesis, glycolysis etc.) is more relevant in this context.
- *Cellular localization* may point to special activity in one of the cell components.
- *Co-expression* annotation labels the genes according to the tissues in which they are known to be expressed. It would be interesting to find for example, that many genes that active in a metastasis are usually active in the tissue of the primary tumor and not in the hosting tissue.
- Taking the data integration ability another step, we can learn about new regulatory relationships. Say we annotate the genes according to the transcription factors that are known to regulate them. Annotation analysis reveals a TF that regulates many positive genes. We now can look for a sequence element which is common to all the regulated genes, and then look for appearance of that element among the other genes.

(a)



(b)

Annotation	% in cluster	% in background	p-Value
RNA splicing, via transesterification reactions	3.87	1.71	4.36E-04
heterogeneous nuclear ribonucleoprotein complex	1.22	0.25	4.66E-04
RNA metabolism	7.32	3.22	1.11E-06
RNA processing	6.1	2.71	1.10E-05
nucleic acid binding	13.22	9	6.30E-04
nuclear mRNA splicing, via spliceosome	3.26	1.27	2.56E-04
RNA splicing	4.27	1.91	2.57E-04
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	3.87	1.71	4.36E-04

Figure 2.17: (a) Result of applying double hierarchical clustering on the PTSD dataset. The algorithm cluster both the samples and the genes according to their expression patterns. (b) Annotation enrichment analysis of GO annotations, which was done to one of the genes cluster (red). The percentage of each annotation in the background and in the cluster is marked, together with the its p-value.

Chapter 3

Learning to Cluster and Classify Samples

3.1 Supervised and Unsupervised Learning

Until now we discussed methods that find and score differentially expressed genes. In these methods we assume the existence of a classification, and look for the informative genes accordingly. This approach, which ”forces” us to use the label information during the analysis, enables us to concentrate on the relevant classification while reducing other signals in the data. However, sometimes the interesting elements are unknown, and we would like to detect them. *Machine learning theory* formalizes this intuition by defining two approaches for learning [Vapnik, 1995]:

- In *supervised learning*, the learner has a supervisor which is ”omniscient”, i.e., it knows the true positives and negatives, and can calculate the error rate and give the learner feedback on its performance.
- In *unsupervised learning*, the learner has no supervisor, i.e., the true positives and negatives are unknown.

Both supervised and unsupervised learning are interesting and difficult tasks. Selection of informative genes and classification algorithms are supervised, since the labels of the samples are known and being used in the process (although gene labels are unknown). Clustering algorithms are unsupervised since the sample labels are not being used during the learning process.

3.2 Discovering Clusters

In a clustering procedure we detect groups of samples or groups of genes that share a common expression pattern. Such a pattern may describe molecular properties associated with that group, thus encompassing new biological knowledge. An example of sample clustering is the detection of disease sub classes [Alizadeh et al., 2000, Bittner et al., 2000, Golub et al., 1999]. Sample clustering may also reflect noise in the data, either artificial or biological (e.g. clusters of genders), and may indicate problems in the sample labels.

Gene clustering discovers groups of genes that share a common expression profile over the samples. These genes are likely to be co-regulated and co-expressed in the cell, and may share a common function. When clustering is applied both to samples and genes, we call it *double* clustering.

Definition 3.2.1: Given a dataset D, consisting of samples $\{x_i\}_{i=1...M} \in R^N$, a *clustering algorithm* divides the samples into k different clusters, so that close samples are in the same cluster. ∎

All clustering algorithms follow this concept, but they differ in the details, such as:

- The distance metric between the samples.
- The method for calculating distance to a cluster. For example, using the average of the cluster or the closest sample within the cluster etc.
- The method for sample traversal.
- The stop condition - which function is being optimized, if at all.

3.2.1 Hierarchical Clustering

The *Hierarchical clustering* [Johnson, 1967] is a simple algorithm: given a distance metric d and a method c for calculating distance between sample and cluster, it clusters the samples, starting from the closest pair. In the next stage it will cluster the second closest pair, which can be either two samples, a sample and a cluster, or two clusters. It will stop when it gets an hierarchical tree structure with a single root, where all M samples are the leaves of the tree (Figure 3.1 (a)). To get k clusters, we have to trim the tree so that k clusters are left disconnected (Figure 3.1 (b)). An example for hierarchical clustering, applied to the PTSD dataset, was shown in Figure 2.17 (a).

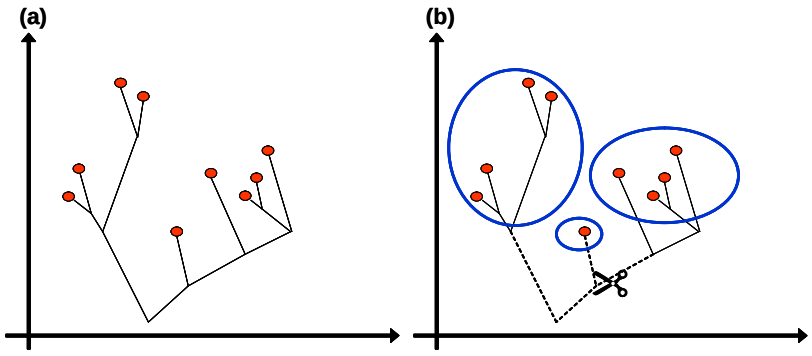


Figure 3.1: Hierarchical clustering of samples in a 2D space. (a) Building the tree (b) Trimming to generate disconnected clusters.

3.2.2 k-Means

The *K-means clustering* [Bishop, 1995] is a much more sophisticated method that gets as an input the distance metric d and the final number of clusters k. The target function for minimization is the sum of average distance squares within each cluster, over all clusters:

$$\min_{c_1..c)k} f(d, c_1...c_k) \text{ s.t.}$$

$$f(d, c_1...c_k) = \sum_{j=1...k} \frac{\sum_{i:x_i \in c_j} d < x_i, \mu j >^2}{|c_j|}$$

K-means starts by picking k random clusters, and calculating their means (Figure 3.2 (a)). Next, it iteratively improves the selection by re-assigning each sample to the cluster with the closest mean, which necessarily improves the target function (Figure 3.2 (b)). After each re-assignment the means are recalculated, and so on (Figure 3.2 (c)). The algorithm stops when no further improvement can be made.

In the *soft* version of k-means, each sample can be assigned to each cluster with a certain probability. In this case the target function summarizes the mean of distances within each cluster. Let p_{ij} be the probability of sample x_i to be assigned to cluster c_j, $\sum_j p_{ij} = 1$ then:

$$f(d, c_1...c_k, \bar{p}) = \sum_{j=1...k} \frac{\sum_{i:x_i \in c_j} p_{ij} d < x_i, \mu j >^2}{|c_j|}$$

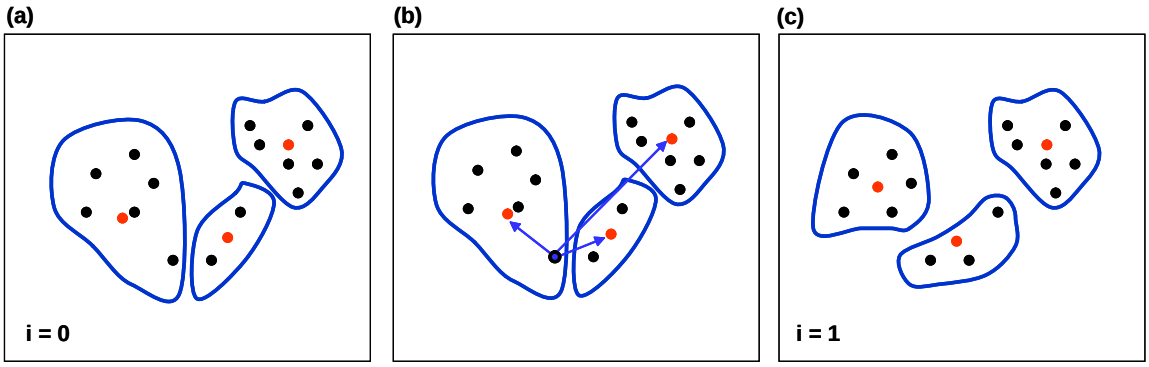


Figure 3.2: Single iteration of k-means: (a) Initial random clusters with means (red). (b) Calculation of distance from one sample (blue) to all clusters means. (c) Reassignment to the closest cluster, and recalculation of means.

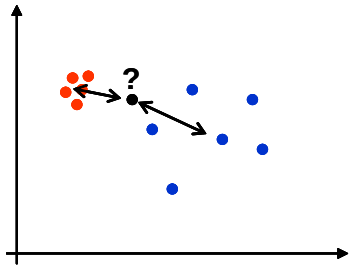


Figure 3.3: The unclassified sample is closer to the red class, but may belong to the blue class.

3.3 Classifying a New Sample

Let us assume that we have found two major clusters, which were interpreted as ”healthy” and ”diseased”, and now a new sample needs to be classified. We would naturally choose to classify it according to the closest cluster. If the samples were already labeled, we could divide them into the two classes and classify the new sample accordingly. The ability to classify a new sample serves as a statistical tool which evaluates the classification, and more importantly, can be used for real life clinical diagnosis of a new sample [M.J. van de Vijver, 2002]. Classification is not trivial, and the closest group is not always the correct answer. For instance, how would you classify the sample in Figure 3.3?

A simple decision rule was presented in the definition of TNoM: $l(x \mid t, g)$ will be ’+’ if $x[g] > t$, and ’-’ otherwise. This decision rule is a *linear separator* since in the dimension of gene g, it can be described as a straight line, where all the ’+’s are above it, and the ’-’s are below it.

Given a sample set D (*training set*), the classification algorithm learns a decision rule, which can be either a linear or a non-linear separator. Later when obtained a new sample x, it classifies x using the learned *classifier* decision rule. The classifier’s prediction ability

is evaluated by calculating the error rate on a new sample set (*test set*).

Definition 3.3.1: Given a dataset D, consisting of pairs $< x_i, l_i >_{i=1...M}$, $l_i \in \{1, 2...k\}$, a *classifier* is a function f_D that depends on the dataset D, and predicts the label of a given sample x: $f_D(x) = \hat{l}$. ∎

3.3.1 The Naive Bayesian Classifier

The *naive Bayesian classifier* [Duda and Hart, 1973, Ben-Dor et al., 2000a] is based on a probabilistic approach to the problem. It assumes that each class produces a different distribution of the expression values - $P(x|l_i)$. Given a new sample x, the classifier estimates the probability that the sample belongs to class j, as:

$$P(l_j|x) = \frac{P(x|l_j)P(l_j)}{\sum_{i=1..k} P(x|l_i)P(l_i)}$$

$P(l_j)$ is the prior probability of class j, which is estimated by the fraction of class j samples in the training set. $P(x|l_i)$ is determined using the training set samples. The naive Bayesian classifier prediction will be:

$$f(x) = arg \max_j P(l_j|x)$$

We will now see an explicit calculation, for the binary classification case:

$$f(x) = \log \frac{P(+|x)}{P(-|x)}$$

where the sign of the classifier determines the prediction, and the magnitude expresses the classifier's *confidence*.

$$\begin{aligned} f(x) &= \log \frac{P(+|x)}{P(-|x)} && (3.1) \\ &= \log \frac{P(+)}{P(-)} + \log \frac{P(x|+)}{P(x|-)} && (3.2) \\ &= \log \frac{P(+)}{P(-)} + \sum_g \log \frac{P(x_g|+)}{P(x_g|-)} && (3.3) \\ &= \log \frac{P(+)}{P(-)} + \sum_g (\log \frac{P(+|x_g)}{P(-|x_g)} - \log \frac{P(+)}{P(-)}) && (3.4) \end{aligned}$$

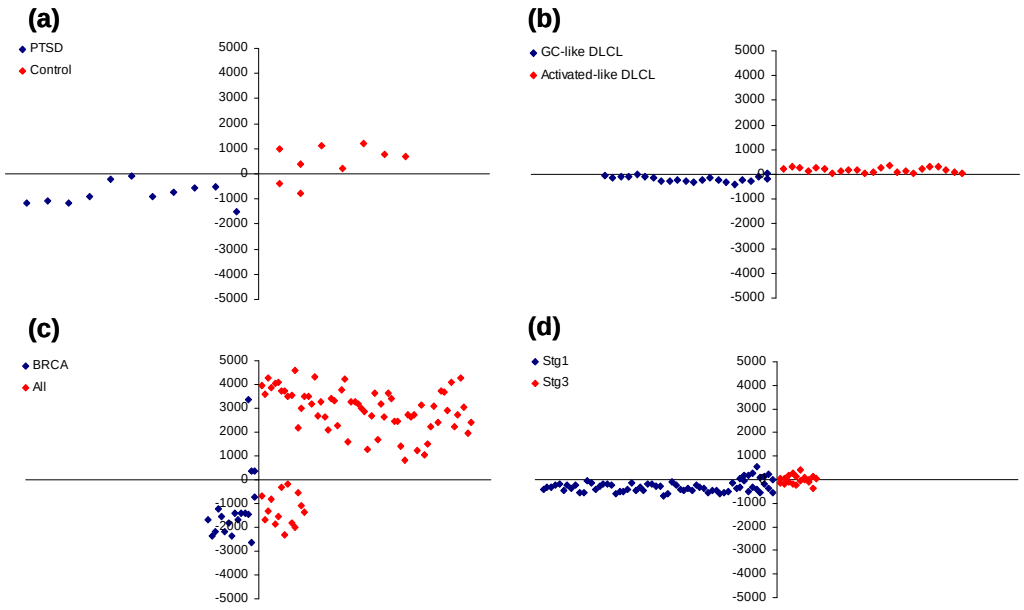


Figure 3.4: Naive Bayes classification confidence in several datasets: (a) PTSD, (b) Lymphoma, (c) Breast cancer and (d) Adenocarcinoma. The y axis denotes the confidence of the classifier in the sample predictions. Negative confidence means that the first class label (blue) was predicted, and positive confidence means that the second class (red) label was predicted. Therefore, blue dots with positive values and red dots with negative values, are classification errors.

where (2) is derived from the *Bayes rule*, (3) from the probability chain rule, assuming gene independence, and (4) again from the Bayes rule, used here for each gene separately.

Given a training set D, the algorithm will learn a classifier f, by calculating $P(+)/P(-)$ and $P(+|x_g)/P(-|x_g)$. To calculate $P(+|x_g)$ and $P(-|x_g)$, the algorithm finds a threshold t, that best separates the values of g into the two classes. The threshold t can be found with the TNoM or Info methods, using the training set samples. In the next step, the algorithm calculates the probabilities of '+' and '-' samples above the threshold and below the threshold. If the new sample $x[g] > t$ the algorithm will return the probabilities that were calculated from above the threshold, otherwise, it will return the probabilities from below the threshold. The final classifier is generated by summing the "votes" of all the genes.

An example of the performance of the naive Bayesian classifier on several datasets is presented in Figure 3.4. The y axis describes the confidence of the classifier, i.e., the proportion of the probabilities.

3.4 Validation of the Classifier

3.4.1 Train and Test Errors

To evaluate the classifier's performance, we would like to know its classification error rate, i.e., what is the probability to misclassify a new given sample.

Let $\mathcal{D}$ be the real distribution of the samples, let X be a sampling of $\mathcal{D}$, and let f_X be a classifier that was trained on X.

Definition 3.4.1.: The *generalization error* [Kearns and Vazirani, 1994], is the probability over $\mathcal{D}$ that the classifier will misclassify a sample.

$$Err(f) = Prob_{x \sim \mathcal{D}}(f_D(x) \neq l_x)$$

∎

Calculating this error is not feasible in most interesting problems[1], so we have to *estimate* it on the test set.

Definition 3.4.2.: The *empirical error* [Kearns and Vazirani, 1994] of a classifier is the fraction of the classifier's errors, estimated over a set of samples X:

$$Err(f_X(X)) = \frac{1}{|X|} \sum_i 1\{f_X(x_i) \neq l_i | x_i \in X\}$$

∎

The empirical error on the training set is the *train error*, and on the test set it is the *test error*. While the train error can be as small as we wish[2], the test error may vary. Since the test error is an estimator for the the generalization error, according to the weak law of large numbers [DeGroot, 1989] when the set size increases, the estimation is closer to the real error.

3.4.2 Cross Validation and Dimension Reduce

In order to get a good estimation of the classifier error, we would like to have a sufficient set of samples for training and testing, where "sufficient" depends on the number of genes, and would roughly be more than several hundreds of samples. However, typical experiment

[1] Since the sampling space is usually infinite.

[2] If the hypothesis class is PAC learnable, or in our case - the number of potential classifiers is finite [Kearns and Vazirani, 1994].

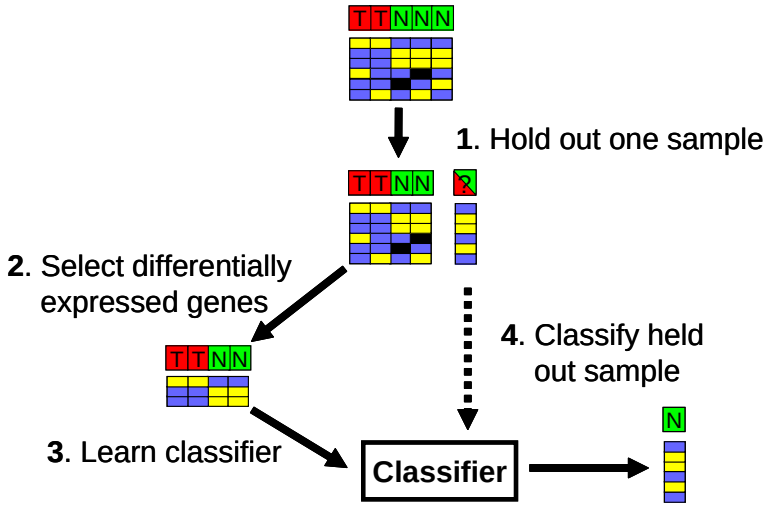


Figure 3.5: The leave one out cross validation (LOOCV) procedure: given labeled samples (e.g., Normal vs. Tumor), one sample is left out. Then the genes are selected and the classifier is learned with the rest of the samples. Finally the classifier gives its prediction to the left out sample.

consists of much fewer samples, and generating training and test set, reduces the effective number of samples even more. To answer this challenge we use two methods that reduce the problem dimension on the one hand, and increase the effective number of available samples on the other hand:

- *Feature selection* is used for reducing the problem dimension, by selecting genes to be the candidates for the classifier's learning. The most informative genes are selected according to the scoring methods, and in this way we reduce noise created by irrelevant genes. Note that the classifier is learned from the training set, and therefore the genes should also be selected only according to the training set.

- *k-Fold cross validation* is used to increase the effective number of samples. This procedure divides the sample set into k random groups and iterates on them. In each iteration it removes one group, learns a classifier f_{X_n} on the rest (training set) and tests it on the left out group. The error of the classifier is the sum of errors of the interior classifiers:

$$Err(f_X) = \frac{1}{|X|} \sum_{n=1..k} 1\{f_{X_n}(x_{n_i}) \neq l_{n_i} | x_{n_i} \in X\}$$

  When $k = 1$ this well known procedure is called *leave one out cross validation* (LOOCV) (Figure 3.5).

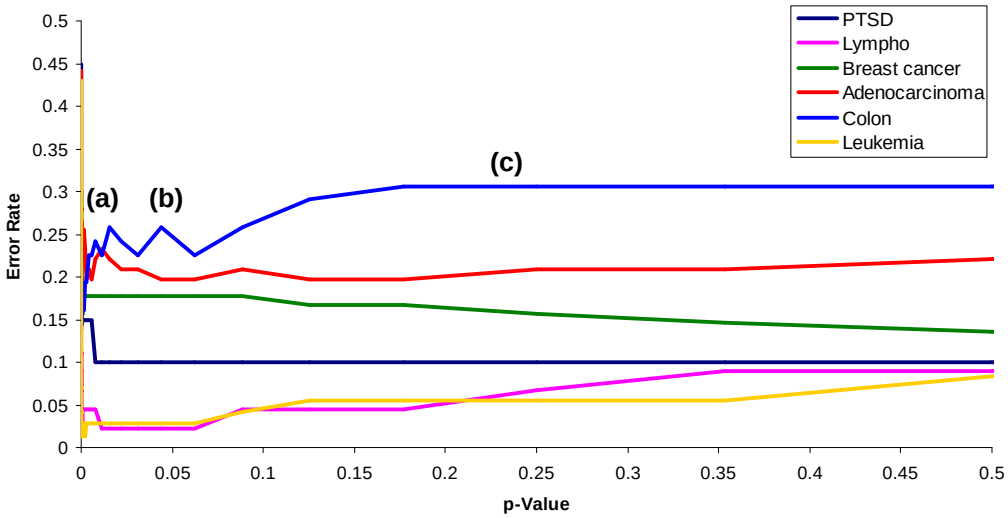


Figure 3.6: The error rate of Naive Bayes classifier with Info scoring method, over different datasets. The x axis denotes the threshold p-value, according to which the feature genes are selected. The y axis denotes the percentage of errors accepted with LOOCV procedure in each dataset.

3.5 Classification Results

Figure 3.6 presents the naive Bayesian classifier error rate for increasing threshold p-values, according to which the genes are selected. The error rate can vary between datasets, since some datasets are easier to classify than others. However, the error rate plot has several typical characteristics:

- In low p-values the plot is noisy (Figure 3.6 (a)). The number of selected genes in these p-values is relatively small, and each addition of genes has a great effect on the error rate.
- As the p-value increases, we can often detect a typical concaveness - a region of a low error rate, which does not change for increasing p-values (Figure 3.6 (b)). In this region an optimal set of separating genes is selected, and an addition of other genes does not affect the classification ability. Each dataset has a different typical optimal region.
- Increasing the p-values even more, the classifier is starting to perform badly, since many genes that are not relevant to the classification are being selected. Their noise overcomes the real signal that was clearer in the lower p-values, and the error rate increases (Figure 3.6 (c)).

3.6 Using the Classification Procedure Correctly

3.6.1 Over Fitting to the training set

We stressed before that the classifier performance should be evaluated on an independent set of test samples. It is very easy to understand that if the classifier is tested on the samples according to which it was learned, it will easily classify them, resulting in a presumably good ”generalization error”. But when checking the classifier on a real independent set, we will probably find out that the true generalization error is worse. This situation is an *over fitting* to the training set. It can be illustrated by a situation in which students are studying all year, and at the end of the year are given an exam. One student is given the same questions he saw during the year, and the other is given new questions. The first student will probably get a better grade, but his true understanding of the material was not examined.

This principle is true for both stages of learning - the feature selection and the learning of the decision rule. Both should be done only with the training set, and examined on an independent set. In the cross validation procedure, we first remove the test samples, and only then select features and learn a decision rule, with all the other samples.

Some gene expression studies were not aware of this principle, and selected the features before performing the LOOCV procedure. The result is a classifier which is tested on the very same samples according to which its features were chosen. The classification error which is claimed in those cases does not represent the true generalization error. Furthermore, it can be easily achieved by any classifier (as for the first student).

In conclusion, we hypothesized that the classifier that was learned with the misguided procedure is over fitted to the train data. In addition, the probability to get the same error rate by chance is higher than the probability to get the same error rate had we used the correct procedure.

3.6.2 p-Value of Classification

To prove our claim, we need to be able to measure the over fitting of a classifier. One way to measure it is simply to estimate the classifier error rate on a new test set. Returning to the students illustration, we will give them a truly new question, and compare their error rates to the error rates that were measured before.

An alternative way is to estimate the statistical significance of the classification. We estimate the p-value of the classification with the permutation test, that permutates the labels, repeats the learning procedure, and estimates the probability to get such an error rate or a better one, under the null hypothesis.

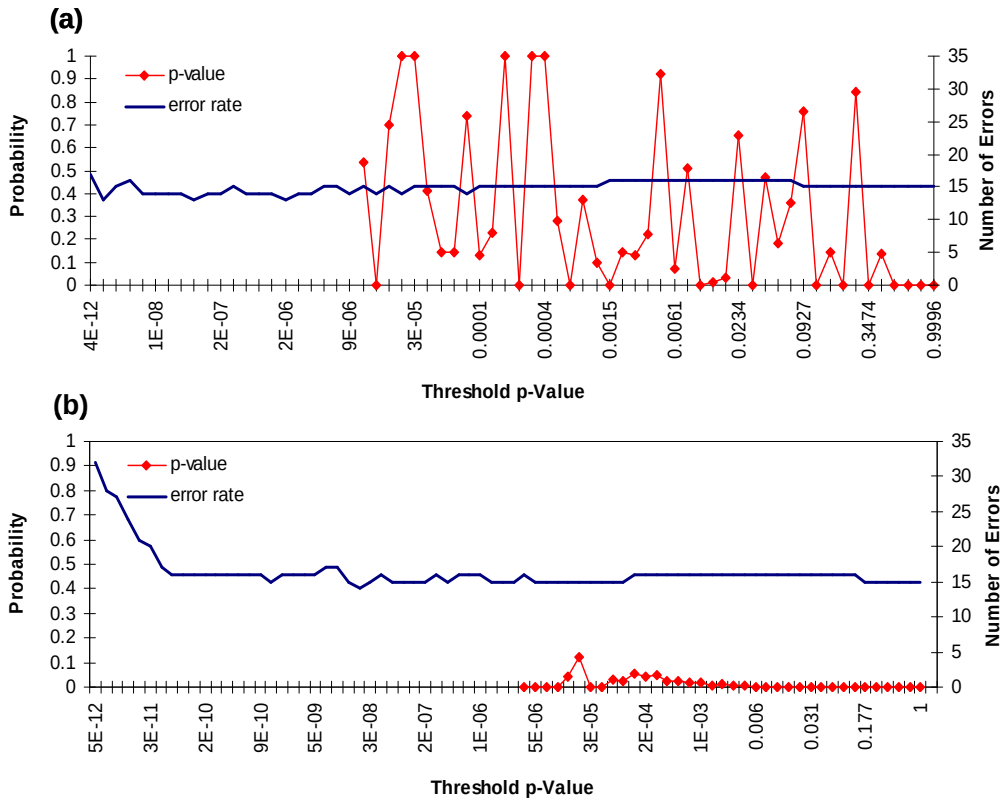


Figure 3.7: Classification error rate on the Breast dataset (BRCAvsAll). The x axis denotes the threshold p-value according to which the feature genes were selected. The **blue plot** denotes the number of errors accepted with that threshold, and the **red plot** denotes the empirical p-value of each error rate. Panel (a) describes the error rates and p-values when using the misguided learning procedure, and panel (b) describes the same, for the correct learning procedure. Classification was done with the naive Bayes classifier and t-test scoring method. The empirical p-value was calculated with permutation test with more than 300 random runs.

The following table presents an empirical p-value estimation for classifications of different datasets. As you can see, all the above classifications are statistically significant, as the probability to get such an error rate is usually ≤ 0.0001.

Dataset	**Classification**	**Errors Num**	**p-Value**
Adenocar	Stg1vs3	18	0.0005
Colon	NvsT	16	0.0001
Lympho	DLCLvsNon	3	0.0001
PTSD	PvsC	7	0.0001
Breast Cancer	BRCAvsAll	16	0.0001

The empirical p-value test was applied to the data from one study by L.J. van t Veer [2002] that used the misguided learning procedure. The results (Figure 3.7) clearly show that the classifier that was learned with the misguided procedure (panel a) produces non significant

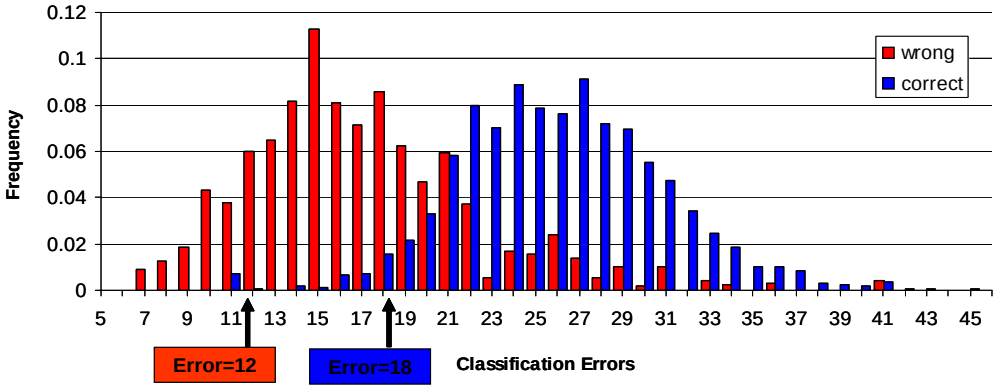


Figure 3.8: Classification error distribution that was calculated over 300 random permutation tests, with different p-values. Red bars mark the error distribution of the misguided procedure, Blue bars mark the distribution with the correct procedure. The real error rates are marked below the graph (Adenocarcinoma dataset, Stg1vs3).

classification results for most of the threshold p-values. The classifier that was learned with the correct procedure (panel b) preforms badly for very small p-values, but in contrast to the classifier in (a), for all p-values excluding one, the classification results are significant, i.e. $p \leq 0.05$.

Further investigation of the correct and misguided procedures was done by looking at the distribution of errors over random labels with the two methods. Figure 3.8 presents a histogram of the error rates that were received with 300 random permutations and different p-values. We can easily observe that the average error rate is lower with the misguided procedure, which supports our claim of ”easier” classification due to over fitting. A second observation is that although the error rate (with the real labels) in the misguided procedure is lower than in the correct procedure, we can observe again that the chance to get such error, which is equivalent to the position in the histogram, is higher in the misguided procedure.

To conclude, statistical validation is exhaustively used in feature selection and clustering, but rarely in classification. Using the simple estimation of empirical p-value we have shown that even a good classification may sometimes be insignificant.

Chapter 4

Detecting Psychiatric Disorder with Gene Expression Analysis

This chapter is excerpted from joint a research with Ronnen Segman, Arik Shalev and Tania Goltser-Dubner (Hadassa medical center, Jerusalem), Naftali Kaminski (Pittsburgh Universrity), and with Nir Friedman (Hebrew University).

The complete paper is available at www.cs.huji.ac.il/∽shefi/ptsd.pdf

4.1 PTSD - Post Traumatic Stress Disorder

Post-traumatic stress disorder is a maladaptive response to stressful events, consisting of re-experiencing of the traumatic event; avoidance and numbing; vigilance and hyper arousal [The American Psychiatric Association, 1994]. With a lifetime prevalence of 9-14%, PTSD is a common mental disorder [Breslau, 2001, Kessler et al., 1995, Yehuda, 2002]. Many survivors express PTSD symptoms at the early aftermath of traumatic events. With time, these symptoms subside in most survivors, but persist in a significant minority - one out of five, in the form of chronic PTSD [Breslau, 2001, Shalev, Zlotnick, 1999, Blanchard, 1997]. Early treatment might prevent PTSD [Bryant et al., 1999], but known risk factors and early PTSD symptoms do not effectively predict chronic PTSD, and therefore have limited use in guiding early treatment [Brewin et al., 2000, Freedman et al., 1999]. Additional difficulty is that the pathogenesis of PTSD is largely unknown.

Biological alterations may underlie the onset severity and persistence of PTSD symptoms [Kessler et al., 1995, Yehuda, 2002, Pitman et al., 2000]. Such alterations are likely to be associated with differential gene transcription, during or after exposure to the triggering

event. For example, acute stress exposure has been shown to induce long-term expression differences in the rat brain [Kaufer et al., 1999, Liberzon et al., 1999, Fujikawa, 2000]. While direct sampling of the brain is not possible in humans, peripheral blood cell gene expression may provide a surrogate indicator of differential response to stress and subsequent PTSD. Supporting this tenet, acute psychological stress is associated with immune activation [Aloe, 1994], and persistent immune alterations have been linked with chronic PTSD [Kawamura et al., 2001, Miller et al., 2001, Spivak, 1997, Maes, 1999]. Additionally, recent microarray studies in human CNS disorders (multiple sclerosis, stroke and seizure) as well as rodent models of such disease suggest specific gene expression signatures in peripheral blood mononuclear cells (PBMC) [Achiron et al., 2004, Tang et al., 2001].

Microarrays allow high throughput gene expression profiling of transcriptional reactivity. Applied to PBMCs, they may detect signatures of biological processes that underlie adaptive and pathological reactions to traumatic stress as they unfold over time. We hypothesized that the transcriptional response of peripheral blood mononuclear cells, will correlate with the development of PTSD among trauma survivors. Here we show that gene expression patterns evaluated four months after trauma identified survivors who either persistently manifested full criteria for acute and chronic PTSD at both one and four months respectively, or remained healthy at follow up. Signatures measured within hours of trauma correlated with later course, and expression patterns at both early and late time points correlated with core symptom trajectories among all survivors.

4.2 Experiment Design

Participants in the study were non physically injured trauma survivors, who were presented to the emergency room (ER), immediately following a traumatic event. 24 trauma survivors were included based on clinical assessment. fourteen of whom had *consistent phenotype* of either full diagnostic criteria of PTSD, or no formal clinical criterion for PTSD at any time. For part of the analysis we included additional ten subjects who showed *partial phenotype* at the ER, and had final clinical diagnosis only at 4 months. Peripheral blood samples were taken from each participant, in the ER, hours after the trauma, and 4 months later. Using oligonucleotide arrays (Affymetrix HU95A), we measured gene expression profiles from these samples. A total of 33 PBMC samples (18 M4 and 15 ER) were available for analysis, of whom 20 samples taken from the consistent phenotype group. After signal quantization and normalization, we identified a set of 4,512 active genes that were expressed and show some variance among the collected profiles. Analysis of the samples was done, and exhaustively validated with various statistical methods that were presented in this work.

4.3 Results

4.3.1 Gene expression signal distinguish PTSD and control

We first determined whether gene expression patterns could distinguish PTSD from control subjects. For this comparison we focused on the consistent phenotype group. Unsupervised hierarchical clustering distinguishes the clinical status at one and four months (Figure 4.1 a). When only M4 samples are analyzed, all subjects are classified into two clusters, one containing PTSD subjects and the other control subjects (Figure 4.1 b). Remarkably, a similar pattern (with one misclassified subject) is evident in clustering of samples taken at ER, hours after trauma (Figure 4.1 c), suggesting that gene expression patterns at the immediate aftermath of trauma can be informative of the later development of the PTSD phenotype.

Over abundance analysis identifies significant number of genes which are differentially expressed between PTSD and control samples - 656 compared to 103 expected by chance (Fig. 1d). Similarly, we find a significant amount of differentially expressed genes, much more than expected by chance, when we examine M4 samples or ER samples separately (Figure 4.1 e and f).

To further explore the predictive abilities of these gene expression signatures, we used the Nave Bayesian classifier, and LOOCV procedure. The classifier was able to correctly classify 8 out of 9 M4 samples (Figure 4.1 g) and 9 out of 11 ER samples (Figure 4.1 h). Evaluating the significance of these classifications using the random permutation test (with 1000 random runs) shows that the classification accuracy is significant with M4 samples $(p = 0.027)$, and nearly significant with ER samples $(p = 0.061)$.

4.3.2 Gene expression correlates with severity of PTSD symptoms

PTSD symptoms are grouped into 3 clusters: trauma re-experiencing (intrusive memory), avoidance behavior and hyper arousal. These symptoms were measured at month 4, using the ”Impact of Event Scale” (IES) clinical test that scores the symptoms severity. To investigate the persistence of symptom trajectories among all survivors, we correlated gene expression profiles of the entire sample of 24 subjects, with the composite IES score, and with the score of each of the three clusters. The correlation was measured with Pearson Correlation. Among the 18 available M4 samples, we found a significant overabundance of genes showing significant correlations $(p \leq 0.05)$ with the IES total score (Figure 4.2 a), as well as with each of the 3 symptom clusters (Figure 4.2 b-d). Among the 15 available ER samples, we also found a significant overabundance of genes that correlated with continuous IES scores measured four months later (Figure 4.3).

4.3.3 Affected trauma survivors show reduced expression of transcriptional enhancers and distinct immune activation

To gain better understanding of the informative and correlative genes that were found in the analysis, we examined their functional classifications. We identify several functional groups that are enriched in these signatures (Figure 4.4 a). Notably, we observe among the affected subjects, down regulation of genes encoding for transcriptional activation proteins, cell cycle and proliferation. We also observe distinct expression signatures for genes involved in immune activation, signal transduction and apoptosis. To attempt a quantitative analysis we calculated the enrichment of GO (Gene Ontology) annotations, and found a significant representations $(p < 0.0005)$ of genes involved in RNA metabolism and processing, as well as nucleotide metabolism (Figure 4.4 b).

4.3.4 Signatures are significantly enriched for genes that encode for neural and endocrine proteins

To further pursue how peripheral transcriptional response may be relevant to the neuropsychiatric process, we examined to what extent differentially expressed genes are also expressed in primary tissues involved in the mediation of neural and endocrine reactivity to stress. We assessed the enrichment of genes known to be expressed in primary tissues, in the signature of differentially expressed genes we identified above. Gene transcripts known to be expressed in brain amygdalar, and hippocampal regions, and the hypothalamic - pituitary adrenal (HPA) axis, were found to be significantly overabundant amongst the genes that distinguished trauma survivors with consistent PTSD (Figure 4.5 a). Significant increased representation of co-expressed genes was found also in the other signatures. Some of the genes showing differential expression patterns among affected trauma survivors play a major role in the neural and endocrine modulation of the stress response (Figure 4.5 b).

4.4 Discussion

We showed here that expression signatures in PBMCs, are informative of the development of PTSD, and its main symptom clusters. In contrast to the current notion [Nisenbaum, 2002, Barlow and Lockhart, 2002], the signatures remains detectable despite cellular heterogeneity of PBMCs, and despite the fact that the PBMCs are not the primary tissue in which the disease occurs.

Our results demonstrate that signatures in PBMCs contain information that is highly correlated with continuous symptom trajectories among all survivors regardless of threshold clinical designation. In addition, initial signatures are informative of later clinical course,

and could have a potential for guiding early detection and focused early intervention among survivors of trauma.

The results can be explained in several ways:

- Alterations of the immune system following psychological stress results in a perturbation of the PBMCs [Southwick, 1999, McEwen, 1998, Aloe, 1994]. We found differential transcriptional patterns of genes involved in immune activation, as well as regulators of proliferation differentiation and demise of leukocytes. It is unclear, though, whether the changes observed in PBMCs are merely informative of the development of PTSD or also bear relevance to its pathogenesis.

- Distinct changes in the *composition* of circulating white cells among affected survivors, may be an additional mechanism underlying the immediate expression changes observed here, resulting an advantage to measure a composite population of cells.

- Expression signatures among PBMCs may reflect in part genomic predisposition to develop PTSD, beyond the putative participation of immune system cells in this neuropsychiatric disorder. Genomic variation may drive related transcriptional reactivity among glial cells that share closer embryonal derivation to leukocytes or even among neuronal cells.

- Reduced hippocampal volumes have been described among PTSD patients [Gilbertson, 2002]. Altered neuroendocrine reactivity, signal transduction, and cellular proliferation and demise among neural and glial cells, have been implicated in hippocampal volume depletion [Kakiuchi, 2003, Gilbertson, 2002, Kim and Diamond, 2002], as well as in fear avoidance formation and memory consolidation processes [Schafe et al., 2001, McEwen, 2001]. It is thus tempting to suggest that our results may denote reduced potential for neural plasticity in response to stress among affected trauma survivors.

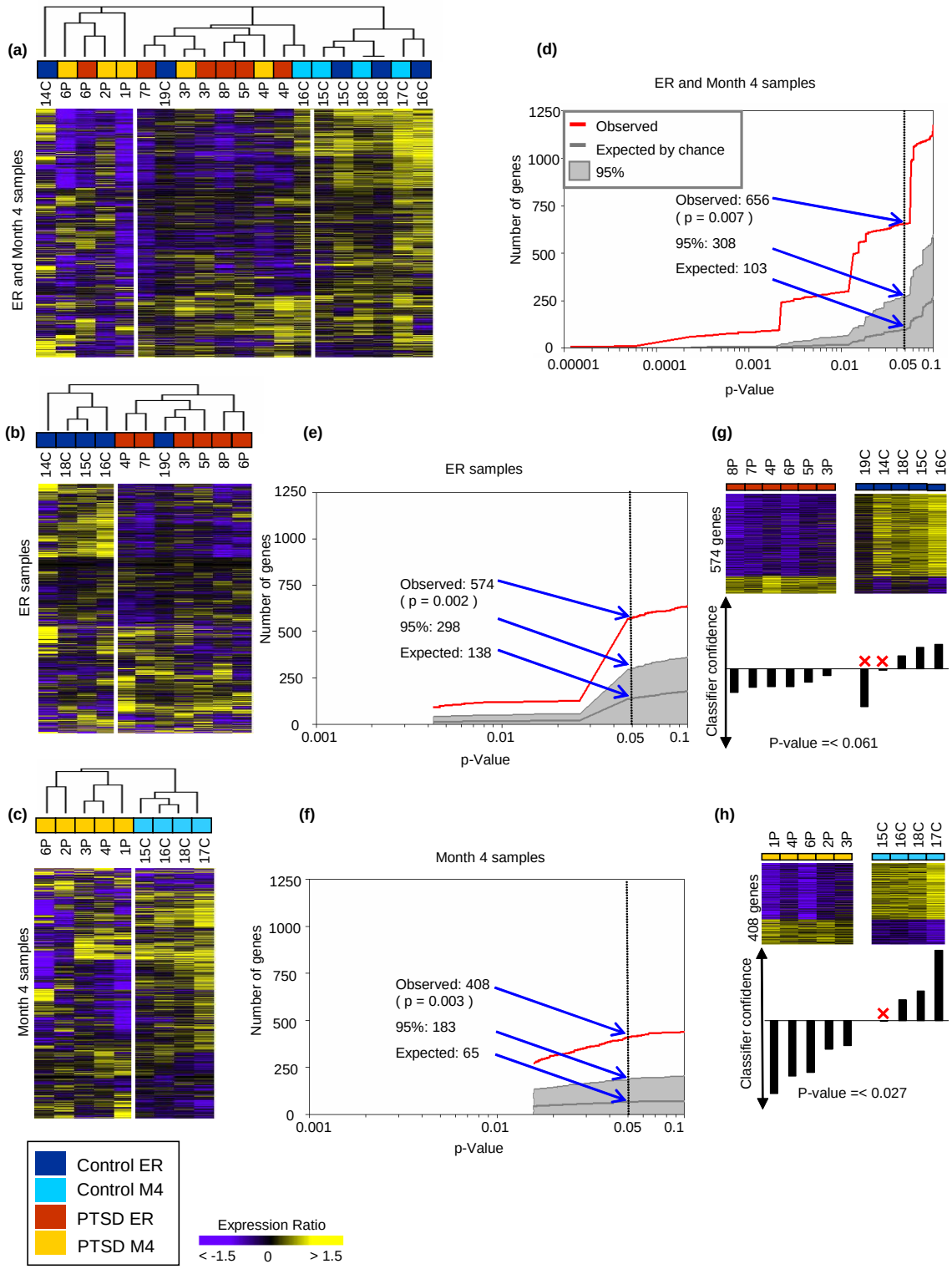


Figure 4.1: Unsupervised hierarchical clustering of 4,512 active genes from the entire sample set (a), at ER only (b) or at four months after trauma (c). The samples colors at the dendrogram bottom indicate the subject number, clinical status and time of sample harvest. (d,e,f) Overabundance plots: Red line - number of informative genes accepted with the real labels; Dark gray line the number expected by chance as calculated with random permutation test; Light gray area - the range on the 95th percentile. (g,h) Evaluation of the supervised classification of subjects phenotypes. Top: Expression profiles of differentially expressed genes. Bottom: Classification results from LOOCV. The sign of the outcome value indicates the predicted phenotype and the magnitude indicates relative confidence. Red cross marks denote misclassified examples.

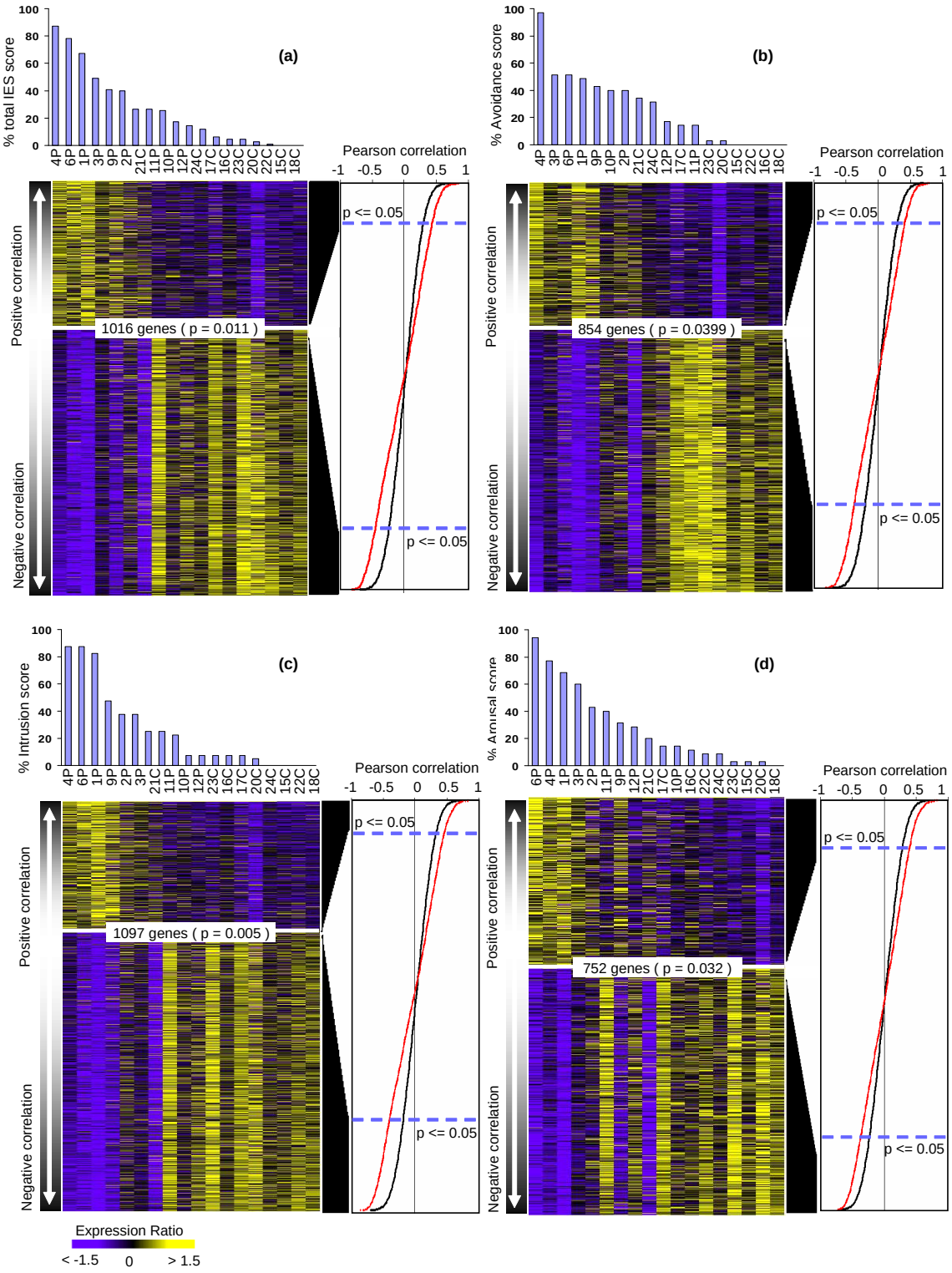


Figure 4.2: Shown is the expression of genes with significant positive or negative correlation to the Total IES score (a), Avoidance score (b), Intrusion score (c), and Arousal score (d). Correlation is measured in 18 month 4 samples. Each panel consists of three elements. Top: The scores of each of the 18 subjects who had month 4 samples. Bottom left: Expression levels of genes with significant $(p \leq 0.05)$ positive and negative correlations with the respective score. The number of correlated genes is shown together with its empirical p-value. Bottom right: Correlation coefficients of all 4512 active genes with the subject score. Red line - curve showing the Pearson correlation of each of the 4512 active genes with the subject score, when the genes are sorted in a decreasing order of correlation. Dark gray line - curve showing the expected sorted Pearson correlations according to the random permutation test.

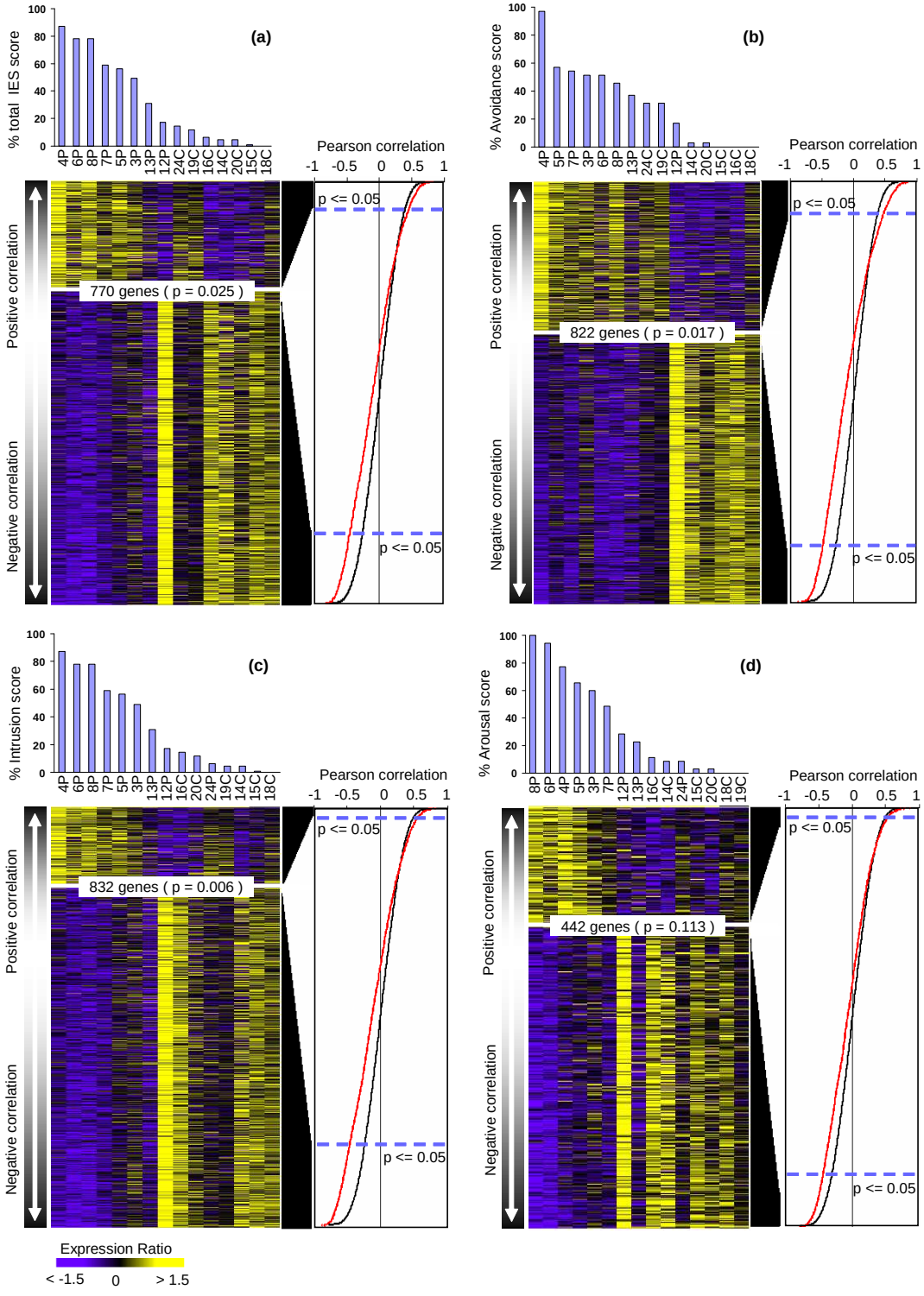


Figure 4.3: The same analysis as in Figure 4.2, but here with 15 samples from the ER.

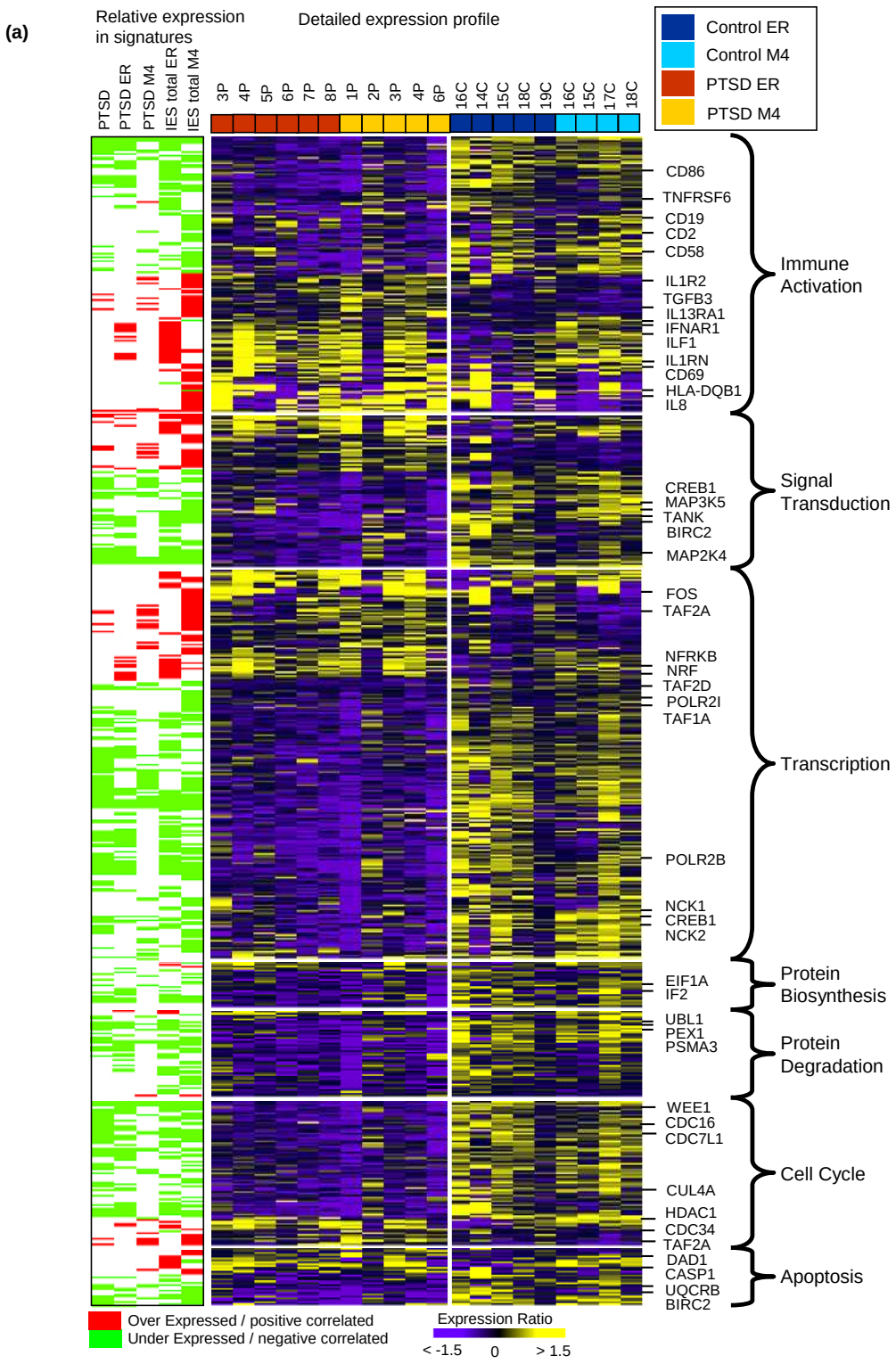


Figure 4.4: (a) Shown are expression profiles (middle) for genes from selected functional categories across samples from patients with consistent phenotypes. Genes are selected based on their participation in previously identified signatures (left): Green - the gene is down-regulated in PTSD or negatively correlated with IES Total score; Red - the gene is up-regulated in PTSD or positively correlated with IES Total score; White - the gene does not appear in the signature. (b) Enrichment of differentially expressed genes within Gene Ontology categories. Shown are functional categories that are significantly enriched for differentially expressed genes (following a False Discovery Rate correction). For each category, we show the percentage of differentially expressed genes within the category. The horizontal line marks the percentage expected by chance.

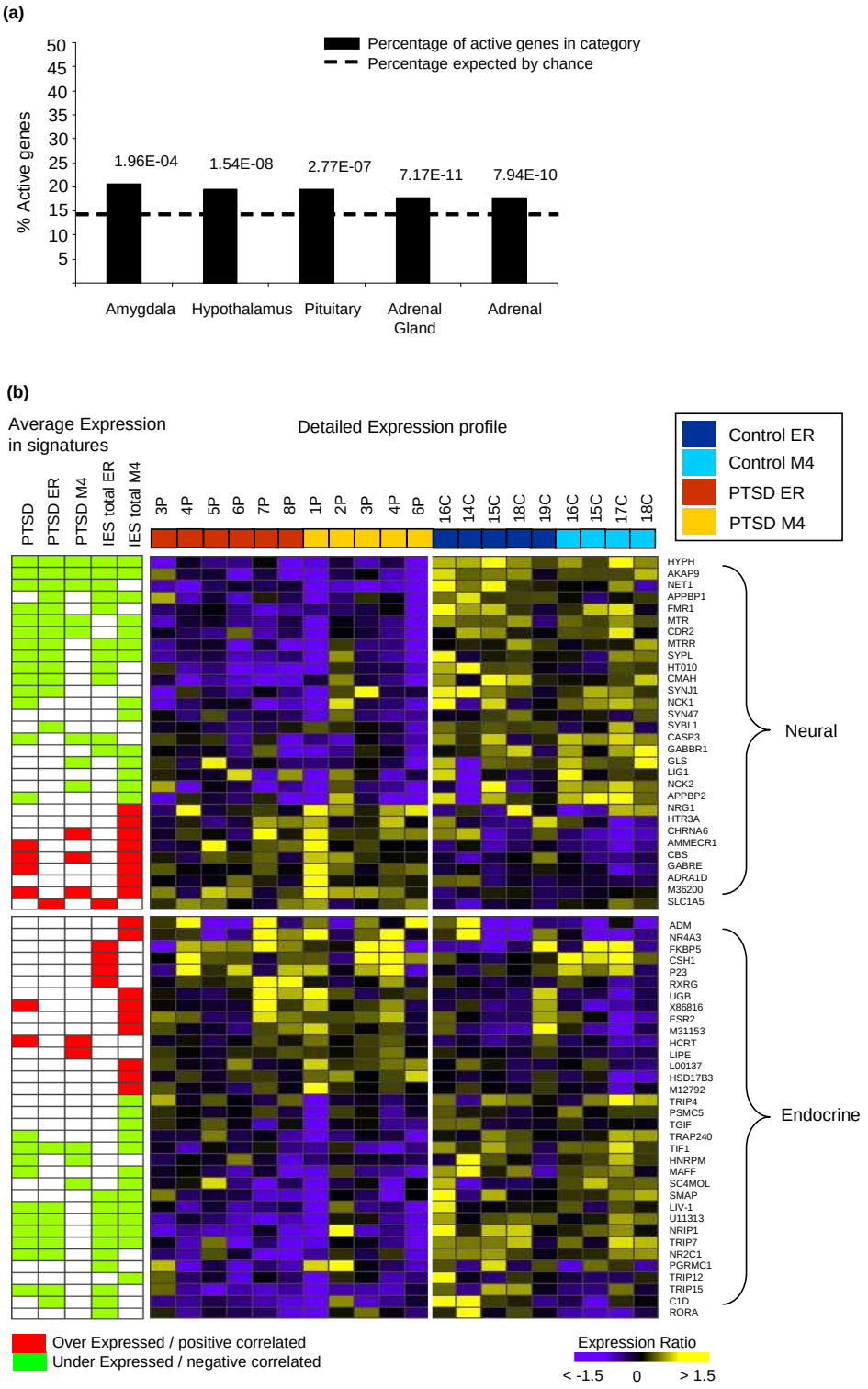


Figure 4.5: (a) Enrichment of differentially expressed genes within groups of genes known to be co-expressed in different brain areas. Annotations were determined using OMIM and UniGene databases. Shown are brain areas that are significantly enriched for differentially expressed genes (following a False Discovery Rate correction). For each annotation, we show the percentage of differentially expressed genes. The horizontal line marks the percentage expected by chance. (b) Expression profiles of neural and neuroenocrine genes that are known to be involved in modulation of the stress response.

Chapter 5

Conclusions

The new technique of DNA microarrays has had a great effect on biological research in the last few years, as it enables a large scale view of the transcriptional changes in the cell [Clarke et al., 2001, Slonim, 2002]. The large amount of data now available requires appropriate computational tools, which were indeed developed, drawing from the worlds of statistics, pattern recognition and machine learning.

Here we have presented several classical methods of gene expression analysis, including feature selection using gene scoring methods, overabundance analysis, clustering methods and classification. These methods enable us to detect new relevant genes, to identify previously unknown sample classifications and gene functionalities, and to develop medical diagnosis tools.

The power of applying theoretical methods on real data was emphasized in analysis in which we searched for molecular signals in the blood cells of people who suffer from post traumatic stress disorder. Using the aforementioned analysis methods, transcriptional changes were identified, even during the very early stages of the disorder. This work, as well as many others, shows that theoretical analytical methods may be fruitful in gaining new important medical insights.

Although the classical methods described are commonly used, some works are not fully aware of the need to validate results. However, statistical validation is crucial, as a non-significant result is eventually meaningless. A simple method that estimates the empirical p-value using a random permutation test was presented here, in the context of overabundance and classification validation. It was employed to show the importance of a correct procedure of classifier learning with k-fold cross validation. Incorrect cross validation leads to an over fitting of the classifier to the training set, which results in a low, yet insignificant, error rate.

In addition, annotation analysis was presented as an efficient tool for biological validation.

Such analysis achieves two goals: one is significance estimation, and the other is the interpretation of results. A great advantage of the annotation analysis is its ability to be applied to any kind of information available, both on the genes and on the samples. This allows us to gain insights in various different levels.

Besides DNA microarrays, interaction data and sequence data, other large scale assays are being developed, such as protein chips [Zhu et al., 2000, Zhou et al., 2004], and SNPs arrays - which identify Single Nucleotide Polymorphisms [Mei et al.]. The availability of such data creates opportunities to develop more complex models which integrate several data sources, and to create a richer and more accurate picture of reality.

In addition to that, other parameters in the cell may be measured, such as protein modifications, ligand concentrations, and the levels of microRNA (which were lately discovered to have a major regulatory role in the cell, [Bartel, 2004]). It would be a great computational challenge to create a model that gives a rich and accurate picture of reality while using these measurements. Such a model needs to describe both transcriptional and post-translational regulation, the protein functionality and the molecule synthesis.

Although there will be no replacement for the human researcher, computational analysis and automatic models may have a major contribution to biological and medical research. The latest technologies produce ample data and the analysis of which necessitates automatic methods. Computational biology had indeed made extraordinary progress in the last decade in supplying such methods, analyzing the various data starting from the DNA sequence, continuing through gene expression and protein structures, and - as of now - ending with system biology and evolution.

The challenge for computational biologists is nevertheless still formidable. As computer scientists we must continue to develop tools that will be standard instruments in the ”wet” lab, just like the pipet and the Petri plate. As biologists, we must continue to ask the seemingly unsolvable questions, which we will be able to solve in the future, on our lab’s PC.

Appendix A

Supplementary Information

1. **LUCA** - Human Lung cancer experiment (data not published, generated by Kaminski N. et al., Sheba medical center and Pittsburgh university). Data consists of 78 samples and 5004 active genes. **Classifications**:
   - **NvsT** - 32 normal samples vs. 46 tumor samples.
   - **AdenovsSq** - 28 tumor samples from type Adeno vs. 18 from type Squamous.
   - **EarlyvsLate** - 29 samples from early satge tumor vs. 17 from late stage tumor.
2. **Lympho** - Human Lymphoma experiment [Alizadeh et al., 2000]. Data consists of 96 samples, and 4026 active genes. **Classifications**:
   - **DLBCLvsALL** - 46 large b-cell lymphoma (DLBCL) samples vs. 50 normal samples that were taken from 8 different tissues.
   - **DLBCLSubClass** - 22 germinal center B-like DLBCL samples vs. 23 activated B-like DLBCL samples.
3. **Leukemia** - Human Leukemia experiment [Golub et al., 1999]. Data consists of 72 samples and 7129 active genes. **Classification**:
   - **AMLvsALL** - 25 Acute Myelogenous Leukemia (AML) samples vs. 47 Acute Lymphocytic Leukemia (ALL) samples.
4. **Colon** - Human Colon experiment [Alon et al., 1999]. Data consists of 62 samples and 2000 active genes. **Classification**:
   - **NvsT** - 20 normal samples vs. 38 cancerous.
5. **Breast** - Human Breast cancer [L.J. van t Veer, 2002]. Data consists of 96 samples, and 5664 active genes. **Classification**:

- **BRCAvsALL**: 18 samples from patients with BRCA1 mutation vs. 78 breast cancer patients without the mutation.

6. **Adenocarcinoma -** Human Lung Adeno carcinoma [D.G. et al., 2002]. Data consists of 86 samples, and 4968 active genes. **Classification**:

   - **Stg1vs3**: 67 tumor samples at stage 1 vs. 19 tumor samples at stage 3.

7. **PTSD** - Human Post traumatic stress disorder (data not published, see Chapter 4). Data consists of 20 samples and 4512 active genes. **Classification**:

   - **PvsC**: 11 PTSD samples vs. 9 control samples.

Bibliography

A. Achiron, M. Gurevich, N. Friedman, N. Kaminski, and M. Mandel. Blood transcriptional signatures of multiple sclerosis: unique gene expression of disease activity. *Ann Neurol*, 55:410–7, 2004.

A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, Jr. Hudson, J., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, L. M. Staudt, and et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11, 2000.

L. Aloe. Emotional stress induced by parachute jumping enhances blood nerve growth factor levels and the distribution of nerve growth factor receptors in lymphocytes. *Proc Natl Acad Sci U S A.*, 91:10440–4, 1994.

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96(12):6745–50, 1999.

Z. Bar-Joseph, G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo, F. Robert, D.B. Gordon, E. Fraenkel, Jaakkola T., D.K. Gifford, and R.A. Young. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, pages 1337–42, 2003.

C. Barlow and D.J. Lockhart. Dna arrays and neurobiology–what's new and what's next? *Curr Opin Neurobiol.*, 12:554–61, 2002.

D.P. Bartel. MicroRNA: genomics, biogenesis, mechanism and function. *Cell*, 116:281–97, 2004.

A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7:559–584, 2000a.

A. Ben-Dor, N. Friedman, and Z. Yakhini. Scoring genes for relevance. Technical Report 2000-38, School of Computer Science & Engineering, Hebrew

University, Jerusalem, 2000b. http://www.cs.huji.ac.il/∼nir/Abstracts/BFY1.html, and Technical Report AGL-2000-13, Agilent Labs, Agilent Technologies, 2000, http://www.labs.agilent.com/resources/techreports.html.

A. Ben-Dor, N. Friedman, and Z. Yakhini. Overabundance analysis and class discovery in gene expression data. Submitted for publication. A preliminary version appeared in *Fifth Annual International Conference on Computational Molecular Biology*, 2001 with the title "Class Discovery in Gene Expression Data", 2001.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J. Royal Statistical Society B*, 57:289–300, 1995.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, U.K., 1995.

M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–40, 2000.

E.B. Blanchard. Prediction of remission of acute posttraumatic stress disorder in motor vehicle accident victims. *J Trauma Stress*, 10:215–34, 1997.

C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilit? *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

N. Breslau. The epidemiology of posttraumatic stress disorder: what is the extent of the problem? *J Clin Psychiatry*, 62:Suppl 17:16–22, 2001.

C. Brewin, B. Andrews, and J. Valentine. Meta-analysis of risk factors for posttraumatic stress disorder in trauma exposed adults. *J Consult Clin Psychol*, 68:748–766, 2000.

P.O. Brown and D. Botstein. Exploring the new world of the genome with dna microarray. *Nat Genet*, 21:33–7, 1999.

R.A. Bryant, T. Sackville, S.T. Dang, M. Moulds, and R. Guthrie R. Treating acute stress disorder: an evaluation of cognitive behaviour therapy and supportive counseling techniques. *Am J Psychiatry*, 156:1780–1786, 1999.

P.A. Clarke, R. te Poele, R. Wooster, and P. Workman. Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. *Biochem Pharmacol*, 15:1311–36, 2001.

Gene Ontology Consortium. Creating the gene ontology resource:design and implementation. *Genome Res*, 11:1425–33, 2001.

M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

M. H. DeGroot. *Probability and Statistics*. Addison Wesley, Reading, MA, 1989.

Beer D.G., Kardia S.L., Huang C.C., Giordano T.J., Levin A.M., Misek D.E., Lin L., Chen G., Gharib T.G., Thomas D.G., Lizyness M.L., Kuick R., Hayasaka S., Taylor J.M., Iannettoni M.D., Orringer M.B., and Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*, 8:816–24, 2002.

R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

S.A. Freedman, T. Peri, D. Brandes, and A.Y. Shalev. Predictors of chronic ptsd -a prospective study. *Br J Psychiatry*, 174:353–359, 1999.

T. Fujikawa. A biphasic regulation of receptor mrna expressions for growth hormone, glucocorticoid and mineralocorticoid in the rat dentate gyrus during acute stress. *Brain Res.*, 874:186–93, 2000.

M.W. Gilbertson. Smaller hippocampal volume predicts pathologic vulnerability to psychological trauma. *Nat Neurosci.*, 11:1242–7, 2002.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, 1999.

http://davidmlane.com/hyperstat/A98696.html.

S.C. Johnson. Hierarchical clusterin schemes. *Psychometrika*, Sep;32(3):241–54, 1967.

C. Kakiuchi. Impaired feedback regulation of xbp1 as a genetic risk factor for bipolar disorder. *Nat Genet.*, 35:171–5, 2003.

D. Kaufer, A. Friedman, S. Seidman, and H. Soreq. Acute stress facilitates long-lasting changes in cholinergic gene expression. *Nature*, 393:373–7, 1999.

N. Kawamura, Y. Kim, and N. Asukai. Suppression of cellular immunity in men with a past history of posttraumatic stress disorder. *Am J Psychiatry*, pages 484–6, 2001.

M. J. Kearns and U.V. Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, Mass., 1994.

R.C. Kessler, A. Sonega, E. Bromet, M. Hughes, and C.B. Nelson. Posttraumatic stress disorder in the national comorbidity survey. *Arch Gen Psychiatry*, 52:1048–1060, 1995.

J.J. Kim and D.M. Diamond. The stressed hippocampus, synaptic plasticity and lost memories. *Nat Rev Neurosci.*, 6:453–62, 2002.

E. L. Lehmann and H. J. M. D Abrera. *Nonparametrics: Statistical Methods Based on Ranks, rev. ed.* Englewood Cliffs, NJ, 1998.

I. Liberzon, J.F. Lopez, S.B. Flagel, D.M. Vazquez, and E.A. Young. Differential regulation of hippocampal glucocorticoid receptors mrna and fast feedback: relevance to post-traumatic stress disorder. *J Neuroendocrinol*, 11:11–7, 1999.

M.J. van de Vijver Y.D. He A.A. Hart M. Mao H.L. Peterse K. van der Kooy M.J. Marton A.T. Witteveen G.J Schreiber R.M. Kerkhoven C. Roberts P.S. Linsley R. Bernards S.H. Friend L.J. van t Veer, H. Dai. Gene expression profiling predicts outcome of breast cancer. *Nature*, 415:530–6, 2002.

H. Lodish, A. Berk, P. Mastudaira, C.A. Kaiser, M. Krieger, M.P. Scott, L. Zipurskym, and J. Darnell. *Molecular cell biology, fourth edition*. WH Freeman and company, 2000.

M. Maes. Elevated serum interleukin-6 (il-6) and il-6 receptor concentrations in posttraumatic stress disorder following accidental man-made traumatic events. *Biol Psychiatry*, pages 833–9, 1999.

B.S. McEwen. Protective and damaging effects of stress mediators. *N Engl J Med.*, pages 171–9, 1998.

B.S. McEwen. Plasticity of the hippocampus: adaptation to chronic stress and allostatic load. *Ann N Y Acad Sci.*, 933:265–77, 2001.

R. Mei, P.C. Galipeau, C. Prass, A. Berno, G. Ghandour, R.K. Patil, N. Wolff, M.S. Chee, Reid B.J., and D.J. Lockhart. Genome-wide detection of allelic imbalance using human snps and high-density dna arrays. *Genome Res*.

R.J. Miller, A.G. Sutherland, J.D. Hutchison, and D.A. Alexander. Creactive protein and interleukin 6 receptor in post-traumatic stress disorder: a pilot study. *Cytokine*, 13:253–5, 2001.

L.J. van t Veer H. Dai A.A. Hart D.W. Voskuil G.J. Schreiber J.L. Peterse C. Roberts M.J. Marton M. Parrish D. Atsma A. Witteveen A. Glas L. Delahaye T. van der Velde H. Bartelink S. Rodenhuis E.T. Rutgers S.H. Friend R. Bernards M.J. van de Vijver, Y.D. He. A gene expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347:1999–2009, 2002.

L.K. Nisenbaum. The ultimate chip shot: can microarray technology deliver for neuroscience? *Genes Brain Behav.*, 1:27–34, 2002.

R.K. Pitman, A.Y. Shalev, and S.P. Orr. Post-traumatic stress disorder, emotion, conditioning and memory. in the new cognitive neurosciences (ed gazzaniga, m.s.). *MIT Press*, pages 1133–1148, 2000.

G.E. Schafe, K. Nader, H.T. Blair, and J.E. LeDoux. Memory consolidation of pavlovian fear conditioning: a cellular and molecular perspective. *Trends Neurosci*, pages 540–546, 2001.

E. Segal, Y. Barahs, I. Simon, N. Friedman, and D. Koller. From promoter sequence to expression:a probabilistic framework. In *RECOMB '02*. 2002.

E. Segal, M. Shapira, A. Regev, D. Peer, D. Botstein, Koller D., and Friedman N. Modules networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–76, 2003.

A.Y. Shalev. Prospective study of posttraumatic stress disorder and depression following trauma. *Am J Psychiatry*, 155:630–637.

C.E. Shannon. A mathematical theory of communication. *The Bell System Technical J*, 27: 379–423 and 623–656, 1948.

D. K. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet*, 32:502–8, 2002.

S.M. Southwick. Role of norepinephrine in the pathophysiology and treatment of posttraumatic stress disorder. *Biol Psychiatry.*, 46:1192–204, 1999.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9 (12):3273–97, 1998.

B. Spivak. Elevated levels of serum interleukin-1 beta in combat-related posttraumatic stress disorder. *Biol Psychiatry*, 42:345– 8, 1997.

Y. Tang, A. Lu, B.J. Aronow, and F.R. Sharp. Blood genomic responses differ after stroke, seizures, hypoglycemia, and hypoxia: blood genomic fingerprints of disease. *Ann Neurol*, 50:699–707, 2001.

Washington D.C. The American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition (DSM-IV)*. 1994.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

C.H. Yeang, T. Ideker, and T. Jaakkola. physical network model. *JCB*, 11:243–62, 2004.

R. Yehuda. Post-traumatic stress disorder. *N Engl J Med*, 346:108–14, 2002.

F.X. Zhou, J. Bonin, and P.F. Predki. Development of functional protein microarrays for drug discovery: progress and challenges. *Comb Chem High Throughput Screen.*, 6: 539–46, 2004.

G. Zhu, P. T. Spellman, T. Volpe, P. O. Brown, D. Botstein, T. N. Davis, and B. Futcher. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, 406: 90–4, 2000.

C. Zlotnick. Chronicity in posttraumatic stress disorder (ptsd) and predictors of course of comorbid ptsd in patients with anxiety disorders. *J Trauma Stress*, 12:89–100, 1999.