Inference of Disease-Related Molecular Logic from Systems-Based Microarray Analysis

Vinay Varadan, Dimitris Anastassiou*

Department of Electrical Engineering and Center for Computational Biology and Bioinformatics (C2B2), Columbia University, New York, New York, United States of America

Computational analysis of gene expression data from microarrays has been useful for medical diagnosis and prognosis. The ability to analyze such data at the level of biological modules, rather than individual genes, has been recognized as important for improving our understanding of disease-related pathways. It has proved difficult, however, to infer pathways from microarray data by deriving modules of multiple synergistically interrelated genes, rather than individual genes. Here we propose a systems-based approach called Entropy Minimization and Boolean Parsimony (EMBP) that identifies, directly from gene expression data, modules of genes that are jointly associated with disease. Furthermore, the technique provides insight into the underlying biomolecular logic by inferring a logic function connecting the joint expression levels in a gene module with the outcome of disease. Coupled with biological knowledge, this information can be useful for identifying disease-related pathways, suggesting potential therapeutic approaches for interfering with the functions of such pathways. We present an example providing such gene modules associated with prostate cancer from publicly available gene expression data, and we successfully validate the results on additional independently derived data. Our results indicate a link between prostate cancer and cellular damage from oxidative stress combined with inhibition of apoptotic mechanisms normally triggered by such damage.

Citation: Varadan V, Anastassiou D (2006) Inference of disease-related molecular logic from systems-based microarray analysis. PLoS Comput Biol 2(6): e68. DOI: 10.1371/journal.pcbi.0020068

Introduction

The expression levels of thousands of genes, measured simultaneously using DNA microarrays, provide information useful for medical diagnosis and prognosis [1,2]. However, their ability to provide significant help towards devising therapeutic approaches has not yet been demonstrated. This failure can be partly attributed to the fact that traditional gene selection techniques typically end up producing a "list of genes" that are correlated with disease, without providing insights into the mutual interrelationships of these genes.

Gene selection techniques from microarray analysis are often based on individual gene ranking depending on a numerical score measuring the correlation of each gene with particular disease types. The expression levels of the highest-ranked genes tend to be either consistently higher in the presence of disease and lower in the absence of disease, or vice versa. Such genes usually have the property that their joint expression levels corresponding to diseased tissues and the joint expression levels corresponding to healthy tissues can be cleanly separated into two distinct clusters. These techniques are therefore convenient and powerful for classification purposes between disease and health, or between different disease types, but they are incompatible with a systems biology viewpoint, because they do not identify systems of synergistically interacting genes, whose joint expression state predicts disease. Rather, microarray clustering techniques tend to produce clusters of co-regulated genes.

Sophisticated machine learning classification approaches, in which a hypersurface on a high-dimensional space serves as a classification boundary separating the gene expression points into classes of tissues, have also been successfully used. Certain nonlinear transformations are typically used to define the shape of the hypersurface, but the performance of the algorithms is limited by the ability to identify and use the optimum such transformation. Furthermore, the set of selected genes is typically not combined with an easily interpretable interrelationship among its members, which could otherwise provide biological insight about the combined role of these genes.

To address such problems, several efforts have recently been made to analyze expression data at the level of biological modules, rather than individual genes [3–9]. However, it has proved difficult, so far, to infer modules of multiple synergistically interrelated genes directly from microarray data.

In this paper, we present a systems-based approach (Entropy Minimization and Boolean Parsimony [EMBP]) that identifies modules of genes jointly associated with disease from gene expression data. The technique also produces a simple logic function connecting the combined expression levels in each gene module with the presence of disease. Roughly speaking, the goals of EMBP analysis are, first, to identify the smallest module of genes whose joint expression levels can predict the presence of disease with high accuracy, and, second, to identify the simplest logic function connect-

Editor: Michael Eisen, University of California Berkeley, United States of America

Received December 27, 2005; **Accepted** May 4, 2006; **Published** June 16, 2006

A previous version of this article appeared as an Early Online Release on May 4, 2006 (DOI: 10.1371/journal.pcbi.0020068.eor).

DOI: 10.1371/journal.pcbi.0020068

Abbreviations: EMBP, Entropy Minimization and Boolean Parsimony

* To whom correspondence should be addressed. E-mail: anastas@ee.columbia.edu

Synopsis

Diseases such as cancer are often associated with malfunctioning pathways involving several genes. Identifying modules of such genes and how the genes in each module interact with each other is helpful toward understanding the nature of these diseases. Here the authors provide a novel computational method for discovering such modules of genes merely from two sets of gene expression data, one from healthy tissues and one from tissues suffering from a particular disease. The method is based on the concept of identifying sets of genes whose joint expression state predicts the presence or absence of a particular disease with minimum uncertainty. Once such gene sets have been identified, we can then further use the microarray data to determine the "logic" that connects the genes' individual expression states related to the outcome of the disease. In turn, this logic may give us valuable insight into the nature of the pathways and how we may target some elements of these pathways for therapeutic purposes. The authors apply this methodology in a particular example and conclude that prostate cancer is often associated with cellular damage from oxidative stress combined with the inhibition of the apoptotic mechanisms normally activated by such damage.

ing these genes to achieve this prediction. We applied EMBP analysis on a prostate cancer dataset [10], and validated the resulting gene modules and logic functions on a different dataset [11].

Results

Binarization of Gene Expression Data

We first binarize microarray expression data into two levels. Although the EMBP methodology can be generalized to account for multiple expression levels, binarization of expression data simplifies the presentation of the concepts in this paper and provides simple logical functions connecting the genes within the found modules.

Rather than independently binarizing each gene's expression level, which would be more appropriate for an individual gene ranking approach, we chose to use single thresholds for all genes. This approach is consistent with the fact that we seek to find global interrelationships among genes and that the microarray data have already been normalized across all tissues and all genes. Therefore, a choice of high threshold will identify the genes that are "strongly" expressed, while a choice of a low threshold will identify the genes that expressed even "weakly." We performed EMBP analysis across several thresholds and we focused on the threshold choices that provided best performance, as described in the following sections.

Entropy Minimization

Following binarization, each gene is assumed to be either expressed or not expressed in a particular tissue, and we also assume that there are two types of tissues, either healthy ones or tissues suffering from a particular disease. The latter assumption can also be generalized to include more than two types of tissues, or modified to be used for classification among several disease types.

Thus, given M genes and K tissues, an $M \times K$ binary "expression matrix" E is defined so that $E(i,j)$ is 1 if gene i is expressed in tissue j, and 0 otherwise. Furthermore, a K-vector c is defined so that $c(j)$ is 1 if tissue j is diseased and 0 if it is healthy. For each gene module of size n, there are 2^n possible gene expression states, and for each state S we can count the number $N_0(S)$ of times that the state appears in a healthy tissue, and the number $N_1(S)$ of times that it appears in a diseased tissue. We can then create a table with 2^n rows corresponding to the gene expression states, which we refer to as the "state-count table" in which each row contains the two counts N_0 and N_1 for the corresponding state. Table 1 shows two examples of such state-count tables for $n = 4$.

We first address the following problem:

Given a number n, identify the set of n genes whose combined expression levels predict the presence or absence of disease with minimum uncertainty.

We refer to this problem as the "entropy minimization" problem, because we quantify the uncertainty with the information-theoretic measure known as conditional entropy [12] after creating a probabilistic model in which probabilities are equal to relative frequencies derived from the counts $N_0(S)$ and $N_1(S)$, so that the presence of disease and the gene expression states are random variables.

In the following, we define the conditional entropy and explain in what sense it measures uncertainty. Given a discrete random variable satisfying a probability distribution $\{p_i\}$, the entropy $-\sum_i p_i \log_2 p_i$ is, in rough terms, the "average length of the shortest description" of the value of the variable [12]. More formally, if we have a sequence of independently drawn symbols, all of which obey an identical probability distribution, then it is a result proven by Shannon [13] that the entropy of that probability distribution measures the minimum average number of bits per symbol required to describe their values. Similarly, the conditional entropy of a random variable, given another variable, is defined as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable [12], and it measures the average length of the shortest description of the value of one random variable given the value of another. For our purposes, assume that the binary random variable C describes the presence or absence of a particular disease, and that S is the random variable (binary n-vector) describing the expression state of a particular gene set of size n. The quantity that we wish to minimize is the conditional entropy, $H(C|S)$, of C given S.

Specifically, using the counts N_0 and N_1, we define

$$P(S) = \frac{N_0(S) + N_1(S)}{K} \tag{1}$$

as the probability of encountering expression state S in a tissue chosen at random, and

$$Q(S) = \frac{N_0(S)}{N_0(S) + N_1(S)} \tag{2}$$

as the probability of disease in a tissue, given that its expression state is S, where Equation 2 is applied for the states that have been encountered at least once.

If we know the expression state S for a particular tissue, then the uncertainty of determining whether or not disease exists in that tissue is measured by the entropy $H(Q(S))$, where the function H is defined by

$$H(q) = -q\log_2(q) - (1-q)\log_2(1-q) \tag{3}$$

Note that the function $H(q)$ becomes close to 0 for values of q that are close to either 0 or 1, and takes a maximum value of 1

Table 1. Two Examples of State-Count Tables and the Corresponding Normalized Conditional Entropies

Example 1						**Example 2**					
a	**b**	**c**	**d**	***N*** N_0	N_1	**a**	**b**	**c**	**d**	N_0	N_1
0	0	0	0	0	0	0	0	0	0	0	14
0	0	0	1	0	0	0	0	0	1	5	0
0	0	1	0	0	0	0	0	1	0	8	0
0	0	1	1	0	0	0	0	1	1	2	0
0	1	0	0	0	0	0	1	0	0	1	19
0	1	0	1	1	0	0	1	0	1	0	13
0	1	1	0	12	21	0	1	1	0	0	2
0	1	1	1	10	10	0	1	1	1	0	1
1	0	0	0	0	0	1	0	0	0	4	0
1	0	0	1	2	1	1	0	0	1	12	0
1	0	1	0	0	0	1	0	1	0	4	0
1	0	1	1	0	0	1	0	1	1	6	0
1	1	0	0	0	0	1	1	0	0	1	3
1	1	0	1	8	3	1	1	0	1	2	0
1	1	1	0	2	1	1	1	1	0	0	0
1	1	1	1	15	16	1	1	1	1	5	0
H = 0.951						H = 0.088					

H indicates the normalized conditional entropy value.

DOI: 10.1371/journal.pcbi.0020068.t001

for $q = 0.5$, which is consistent with its interpretation as uncertainty. The average overall uncertainty of determining whether or not disease is present is then measured by the conditional entropy of the presence of disease given the expression state for the gene set:

$$\sum P(S)H(Q(S)) \qquad (4)$$

where the summation is over all 2^n states S with $P(S) > 0$. Again, this formula is consistent with the intuitive interpretation of uncertainty, because it becomes small whenever the probabilities $Q(S)$ become close to either 0 or 1 for the more frequently encountered states.

Finally, to ensure that the range of possible values for the conditional entropy extends from 0 to 1, we normalize by dividing by $H(Q_{\text{null}})$, the entropy corresponding to the probability of disease in a randomly chosen tissue. (In the case of the prostate data that we use, this probability is equal to 52/102). For simplicity, in the sequel we will often refer to the normalized conditional entropy as just "entropy."

The conditional entropy, as defined above, depends on the counts N_0 and N_1 for the 2^n states. Its interpretation as a measure of uncertainty is illustrated in the example of Table 1, which contains two state-count tables that were created using the binarized expression matrix for the 102 prostate tissues used in this paper, and a threshold of 15. The state-count table on the left corresponds to a choice of four genes a, b, c, and d, selected at random. The resulting value of the normalized conditional entropy of 0.951 is typical for random choices of gene sets. On the other hand, the state-count table on the right corresponds to the gene set for which we found the minimum normalized conditional entropy of 0.088, consisting of genes a: *COL4A6,* b: *CYP1B1,* c: *SERPINB5,* and d: *GSTP1,* to be discussed later in this paper. In this latter gene set choice, as shown in Table 1, the reduced entropy is manifested by the fact that the statistics are skewed for nearly all states. For example, all 13 tissues corresponding to state 0101 are cancerous, and all 12 issues corresponding to state 1001 are healthy.

The entropy minimization problem consists of identifying the gene set with the minimum conditional entropy, as defined above, among all subsets of size n of the full set of M genes. The number of these subsets is equal to $\binom{M}{n}$ and becomes large for $n \geq 3$, making the exhaustive search method impractical. As explained in the Materials and Methods section below, however, this problem can easily be addressed using heuristic search optimization methods.

If the conditional entropy for a particular gene set is found to be exactly 0, this implies that the joint expression levels of the members of that gene module determine the existence of disease with absolute certainty under the assumption of the probabilistic model derived from the relative frequencies. This happens whenever, for all 2^n states in the corresponding state-count table, at least one of the counts N_0 and N_1 is 0. In our experiments, we have found that when this occurs, a large number of states are only encountered once or twice. We cannot make any reliable association of disease based on these rarely encountered states, and including them in our model will result in "overfitting," so we treat them as noise and ignore them, in favor of the states which predominantly correspond to disease. Our definition for such states is that they have been encountered at least three times and that the number of corresponding diseased tissues is at least four times larger than the count of corresponding healthy tissues, i.e., $N_1 \geq 3$ and $N_1 \geq 4N_0$. Therefore, whenever we find the entropy for a gene set to be exactly 0, we decrease the size of the gene set by 1, and select the minimum-entropy gene set of that size. Thus, the output of the EMBP analysis contains a gene module for which the conditional entropy is close, but not equal, to 0.

Calculating the conditional entropy of a particular set of genes of length n involves evaluating the counts $N_0(S)$ and $N_1(S)$ for each of the 2^n states. However, the number of states for which these counts are non-zero cannot be larger than the number K of tissues. Thus the computational cost of calculating the entropy of a gene set remains bounded regardless of the size of the gene set.

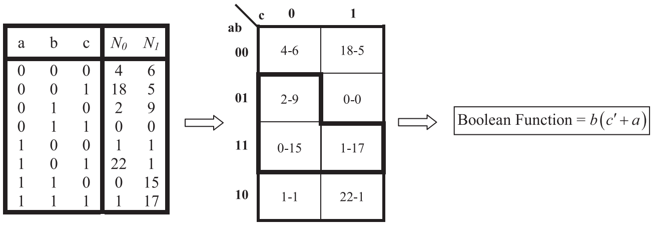
Boolean Parsimony

Once a gene module has been identified, and the expression states for that module that are predominantly associated with disease have been determined as described above, we then address the following problem: Given the gene expression states associated with disease, find the simplest logical rule that connects the expression levels in the gene module with the presence of disease.

We refer to this problem as the "Boolean parsimony" problem, because the logical rule will be identified by the "most parsimonious Boolean function." Our definition for this logic function is one containing the operators AND, OR and NOT, which minimizes a "cost," defined as the total number of logic variables appearing in the expression. In Boolean algebra [14], each logic variable can take the value of either 0 (false) or 1 (true), the operator AND corresponds to multiplication, and the operator OR corresponds to addition. We use the symbol of prime (′) following the logic variable to designate the operator NOT. For example, $ab+a'b'+ab'$ means (a AND b) OR [(NOT a) AND (NOT b)] OR [a AND (NOT b)] and the "cost" (as defined above) of this Boolean function is 6, because each of the variables a and b appears three times. This Boolean expression happens to be logically equivalent to $a+b'$, meaning: a OR (NOT b). The latter expression is more parsimonious than the former, because its "cost" is equal to 2, as each of the letters a and b appears once.

The reason for the need of Boolean parsimony is that the biological role of each gene becomes more immediately clear if the Boolean expression contains the corresponding logic variable either once or only a few times. We selected the above definition of Boolean parsimony because the logic functions AND, OR, and NOT often have straightforward potential biological interpretations.

The problem is easily solved manually when the size of the gene set is less than 5, as in the examples of this paper, using Karnaugh map logic design methodology [15] (Figure 1). Otherwise, Boolean minimization programs such as Espresso [16] can be used. Most of them retain the "sum of products" structure of the Boolean expression, but further minimization is desirable and possible using heuristic algorithms [17].



a	b	c	N_0	N_1
0	0	0	4	6
0	0	1	18	5
0	1	0	2	9
0	1	1	0	0
1	0	0	1	1
1	0	1	22	1
1	1	0	0	15
1	1	1	1	17

ab \ c	0	1
00	4-6	18-5
01	2-9	0-0
11	0-15	1-17
10	1-1	22-1

Boolean Function = $b(c'+a)$

Figure 1. Example of Estimating the Boolean Function for a Gene Module

The state-count table for a set of three genes is used to create the 4 × 2 table shown on the right, known as a "Karnaugh map." The binary coordinates in the rows and columns of Karnaugh maps are arranged according to the "Gray code" (consecutive coordinates differ by one bit only) for easier derivation of the logical function. Each entry of the Karnaugh map corresponds to one of the states and the numbers shown are equal to the counts N_0 and N_1 for that state. In this example, there are three entries (010, 110, 111) identified as predominantly associated with a particular disease. States 010 and 110 can be jointly described by the function bc′. Similarly, states 110 and 111 can be jointly described by the function ba. The overall Boolean function describing the area highlighted by bold lines in the Karnaugh map is $bc'+ba = b(c'+a)$.

DOI: 10.1371/journal.pcbi.0020068.g001

The computational cost of Boolean Parsimony is insignificant compared to that of Entropy Minimization. For example, Espresso [16] takes less than a minute on a standard Pentium III processor running at 3 GHz for Boolean functions involving many tens of variables. The BDS algorithm [17] used to find the most parsimonious Boolean function also takes less than a minute for functions having many tens of Boolean variables.

Prostate Cancer EMBP Analysis

We used two different prostate cancer datasets. The first prostate cancer microarray expression data [10] contain gene expression profiles for 102 prostate tissues, of which 52 were cancerous and 50 were healthy and is available in the public domain from http://www-genome.wi.mit.edu/MPR/prostate. The gene expression profiles in scaled average difference units were produced using HG-U95A Affymetrix microarrays with probes for 12,600 genes (Affymetrix, Santa Clara, California, United States). This dataset will henceforth be referred to as the "EMBP dataset," because we used it to apply EMBP analysis. A second independently derived dataset, also containing scaled average difference units referring to 34 tissue samples of which 25 were cancerous and nine were healthy was also obtained from the public domain [11] (http://www.gnf.org/cancer/prostate) and used for validating the gene modules and logic functions estimated over the EMBP dataset. This latter dataset will be referred to as the "validation dataset."

Gene module and Boolean function identification using the EMBP dataset. Each threshold choice for binarizing the continuous-valued data of the EMBP dataset will produce potentially different results. The "optimum" threshold can be defined and evaluated as the one that yields the minimum overall entropy. In our case, we found that this would be equal to 60. However, EMBP analysis is not meant to produce a unique gene set as its output. Rather, it can reveal a rich content of information from the microarray data by using different threshold choices. The combined information resulting from several such gene sets can further help provide insight into related pathways.

We used several thresholds ranging from −30 to 225 in steps of 15 and estimated the minimum entropy gene modules for each of them. For each threshold value we considered gene modules of size $n = 1$, 2, 3, and 4. For $n = 4$, we found that the entropy values occasionally went down to precisely 0 due to overfitting. On the other hand, we found several gene sets of size 3 with entropy values less than 0.20, which is low—note that $H(0.97) = 0.20$, meaning that if the conditional entropy is 0.20, then, on the average, each state is associated with either cancer or health with probability 97%. Therefore, we selected $n = 3$ to be the number of genes included in these gene modules. Our results of are shown in Figure 2A. The thresholds for which the minimum entropy values were below 0.20 for $n = 3$ are 30, 45, 60, 75, and 90. The minimum entropy gene modules for these thresholds along with their entropies are listed in Table 2. In the following, we use the official gene symbols, and Table 3 contains a legend

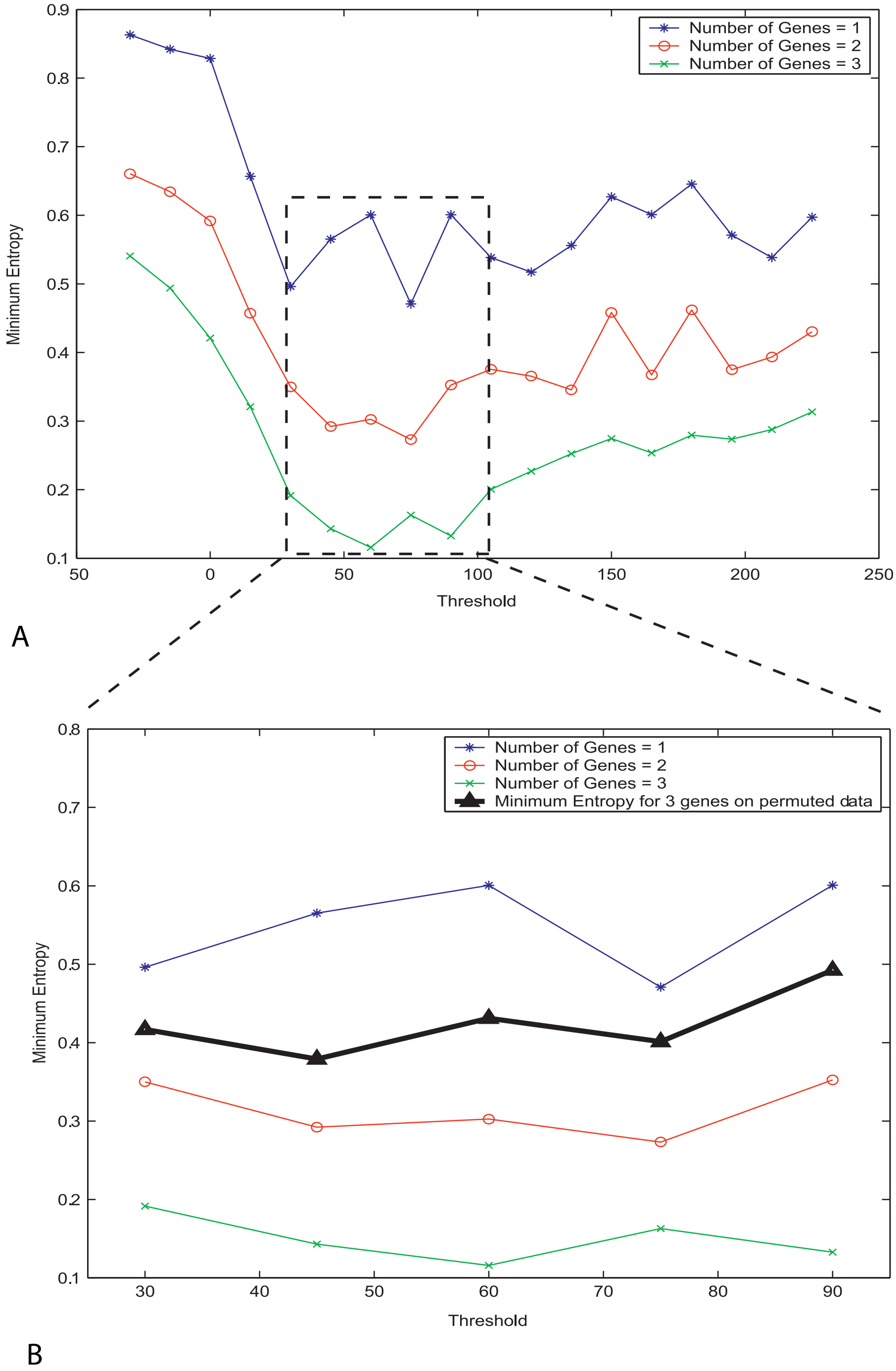


Figure 2. Minimum Entropy across Different Thresholds

(A) Estimates of the minimum entropy values for gene modules of size $n = 1$, 2, and 3 across various thresholds.

(B) Minimum entropy values for subset of thresholds (colored lines) along with the estimated means of the entropies over randomly permuted data for $n = 3$ (black line).

DOI: 10.1371/journal.pcbi.0020068.g002

with the corresponding accession numbers, aliases, and brief gene descriptions.

To evaluate the significance of these minimum entropy values, we performed entropy minimization over ten random permutations of the tissue class labels. In other words, while keeping the number of healthy and cancerous tissues constant to 50 and 52, respectively, we randomly assigned healthy (0) and cancerous (1) labels to the individual tissue profiles. The entropy minimization algorithm was performed on the randomly permuted data, and the average minimum entropies for $n = 3$ were estimated for the thresholds 30, 45, 60, 75, and 90 for the same expression matrix of the EMBP dataset. The estimated averages of the entropies are shown in the heavy black line in Figure 2B. Notably, the entropy values for the randomly permuted data for $n = 3$ are much higher than those estimated on the actual dataset, and even

Table 2. List of the Minimum Entropy Gene Modules for Different Thresholds

Threshold	Gene Module			Entropy
	a	b	c	
30	*SPINK2*	*TMSL8*	*RBP1*	0.19155
45	*HPN*	*ENTPD1*	*NELL2*	0.14302
60	*NCF4*	*HPN*	*PGM1*	0.11587
75	*HPN*	*MCM3AP*	*GSTP1*	0.16287
90	*HLA-DQB1*	*FNBP1*	*DF*	0.13267

DOI: 10.1371/journal.pcbi.0020068.t002

significantly higher than the entropy values of the actual data with $n = 2$, indicating that the gene modules identified by entropy minimization on the actual data have real biological meaning, rather than being due to chance. To further emphasize this fact, we performed entropy minimization over 40 random permutations of the tissue class labels using the E matrix for threshold 60, chosen because it yields the lowest overall entropy. We observed that the 40 entropy values derived from the permuted data consistently fit a normal distribution using any of the chi-square, Lillie and Geary tests [18]. The mean and standard deviation of the entropy values were 0.4557 and 0.0433, respectively. Using these statistics, we estimated the *p*-value of the minimum entropy on the actual data, defined as the probability of the minimum entropy being not more than 0.11587 purely by chance to be extremely small, equal to 2×10^{-15}.

We then estimated the most parsimonious Boolean functions for these five gene modules, which were selected because the corresponding thresholds are close to each other and spread around the "optimum" one, equal to 60, as mentioned above. Figure 3 contains the Karnaugh maps from which the functions were derived, together with the corresponding Boolean functions and their accuracy if these simple functions are used for classification on the EMBP dataset. For convenience, these Boolean functions are also formulated in words in Figure 3, where "presence" and "absence" of a gene refer, for simplicity, to the presence or absence of mRNA from the gene. Furthermore, we found that gene *ENTPD1* in Figure 3B can be replaced by gene *HIST1H1E,* and that gene *NCF4* in Figure 3C can be replaced by gene *KRT6E.* In both cases, these substitutions yield identical results.

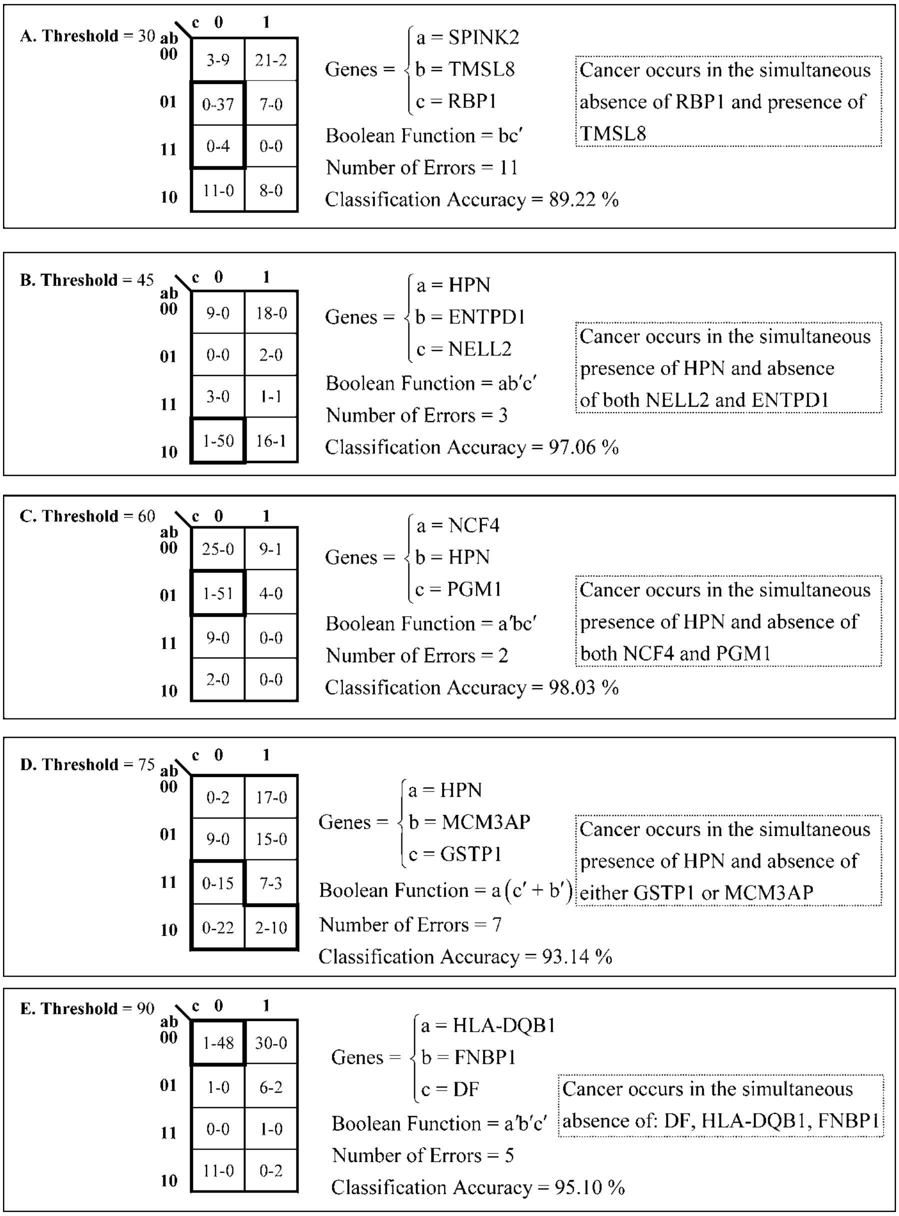
The genes mentioned in Figure 3 should not be seen as individual "prostate cancer-related genes," which, in traditional approaches, are found to be either consistently overexpressed or consistently underexpressed in prostate cancer. Instead, each of the identified genes should be seen as a member of a synergistic gene module, as evidenced by the formulation of the corresponding Boolean function. To further clarify the fundamental difference between the two approaches, we mention the following "notable facts" for each of the five identified gene modules, derived from simple observation of the counts in each Karnaugh map, each of which could provide hints for its biological explanation:

(1) Absence of *RBP1,* if accompanied by either presence of *TMSL8* or absence of *SPINK2,* is associated with cancer in 50 out of 53 such tissues. However, in the simultaneous absence of *TMSL8* and presence of *SPINK2,* absence of *RBP1* is not

Table 3. List of Genes That Were Included in the Estimated Gene Modules along with Their Accession Numbers and Gene Description

Symbol	Accession Number	Alias/Description
COL4A6	D21337	*Collagen, type IV, alpha 6*
CYP1B1	U03688	*Cytochrome P450, family 1, subfamily B, polypeptide 1*
DF	M84526	*Adipsin, D component of complement*
ENTPD1	AJ133133	*Ectonucleoside triphosphate diphosphohydrolase 1*
FNBP1	AB011126	*KIAA0554, Formin binding protein 1*
GSTP1	U12472	*Glutathione S-transferase pi*
HLA-DQB1	M81141	*Major histocompatibility complex, class II, DQ beta 1*
HPN	X07732	*Hepsin, transmembrane protease, serine 1*
HIST1H1E	M60748	*H1F4, Histone 1, H1e*
KRT6E	L42611	*Keratin 6E*
MCM3AP	AB011144	*KIAA0572, MCM3 minichromosome maintenance deficient 3 (S. cerevisiae) associated protein*
NCF4	AL008637	*P40PHOX, neutrophil cytosolic factor 4 (derived from precise chip probe)*
NELL2	D83018	*NEL-like 2 (chicken) protein*
PGM1	M83088	*Phosphoglucomutase 1*
RBP1	M11433	*Cellular retinol binding protein 1*
SERPINB5	U04313	*Maspin, Serpin peptidase inhibitor, clade B (ovalbumin), member 5*
SPINK2	X57655	*Serine peptidase inhibitor, Kazal type 2 (acrosin-trypsin inhibitor)*
TMSL8	D82345	*TMSNB, Thymosin-like 8*

DOI: 10.1371/journal.pcbi.0020068.t003



A. Threshold = 30

ab \ c	0	1
00	3-9	21-2
01	0-37	7-0
11	0-4	0-0
10	11-0	8-0

Genes = {a = SPINK2, b = TMSL8, c = RBP1

Boolean Function = bc′

Number of Errors = 11

Classification Accuracy = 89.22 %

Cancer occurs in the simultaneous absence of RBP1 and presence of TMSL8

B. Threshold = 45

ab \ c	0	1
00	9-0	18-0
01	0-0	2-0
11	3-0	1-1
10	1-50	16-1

Genes = {a = HPN, b = ENTPD1, c = NELL2

Boolean Function = ab′c′

Number of Errors = 3

Classification Accuracy = 97.06 %

Cancer occurs in the simultaneous presence of HPN and absence of both NELL2 and ENTPD1

C. Threshold = 60

ab \ c	0	1
00	25-0	9-1
01	1-51	4-0
11	9-0	0-0
10	2-0	0-0

Genes = {a = NCF4, b = HPN, c = PGM1

Boolean Function = a′bc′

Number of Errors = 2

Classification Accuracy = 98.03 %

Cancer occurs in the simultaneous presence of HPN and absence of both NCF4 and PGM1

D. Threshold = 75

ab \ c	0	1
00	0-2	17-0
01	9-0	15-0
11	0-15	7-3
10	0-22	2-10

Genes = {a = HPN, b = MCM3AP, c = GSTP1

Boolean Function = $a(c' + b')$

Number of Errors = 7

Classification Accuracy = 93.14 %

Cancer occurs in the simultaneous presence of HPN and absence of either GSTP1 or MCM3AP

E. Threshold = 90

ab \ c	0	1
00	1-48	30-0
01	1-0	6-2
11	0-0	1-0
10	11-0	0-2

Genes = {a = HLA-DQB1, b = FNBP1, c = DF

Boolean Function = a′b′c′

Number of Errors = 5

Classification Accuracy = 95.10 %

Cancer occurs in the simultaneous absence of: DF, HLA-DQB1, FNBP1

Figure 3. Karnaugh Maps Leading to Boolean Functions across Different Thresholds

Shown are the Karnaugh maps for the minimum entropy three-gene modules for five choices of binarization threshold, the parsimonious Boolean functions derived from the Karnaugh maps, and the classification accuracy of these Boolean functions over the EMBP dataset. Furthermore, we found that gene *ENTPD1* in Figure 3B can be replaced by gene *HIST1H1E* and that gene *NCF4* in Figure 3C can be replaced by gene *KRT6E*. In both cases, these substitutions yield identical results, but are not shown in the figure, for simplicity.

DOI: 10.1371/journal.pcbi.0020068.g003

associated with cancer. On the contrary all 11 such tissues are healthy.

(2) Presence of *NELL2* is associated with health in 37 out of 39 such tissues, even if *HPN* (normally associated with cancer) is present. Simultaneous presence of *HPN* and *NELL2* is associated with health in 17 out 19 such tissues.

(3) Presence of *NCF4* is associated with health in all 11 such tissues, even if *HPN* (normally associated with cancer) is present: Simultaneous presence of *HPN* and *NCF4* is associated with health in all nine such tissues. The same formulation is true if *NCF4* is replaced by *KRT6E.*

(4) If either *HPN* is present or *MCM3AP* is absent, then the absence of *GSTP1* is associated with cancer, as all such 39 tissues are cancerous. However, if *HPN* is absent and *MCM3AP* is present, then the absence of *GSTP1* is not associated with cancer, as all such nine tissues are healthy.

(5) In the absence of *HLA-DQB1,* absence of *DF* is associated with cancer in 48 tissues out of 50, and presence of *DF* is associated with health in 36 tissues out of 38. However, in the presence of *HLA-DQB1,* absence of *DF* is instead associated with health, as all 11 such tissues are healthy.

Validation of the gene modules and Boolean functions using the validation dataset. Although the classification performance of the Boolean functions from EMBP analysis is high over the dataset upon which the results were derived (Figure 3), it is important to validate these results over

Table 4. Classification Accuracy of EMBP Analysis Results over the Validation Dataset

EMBP Dataset Threshold	Gene Module			Boolean Function	Validation Dataset Threshold	Classification Accuracy (%)	Specificity (%)	Sensitivity (%)
	a	b	c					
30	*SPINK2*	*TMSL8*	*RBP1*	bc′	342.26	85.29	77.78	88
45	*HPN*	*ENTPD1*	*NELL2*	ab′c′	467.33	94.12	100	92
60	*NCF4*	*HPN*	*PGM1*	a′bc′	592.40	94.12	100	92
75	*HPN*	*MCM3AP*	*GSTP1*	a(c′ + b′)	717.47	97.06	100	96
90	*HLA-DQB1*	*FNBP1*	*DF*	a′b′c′	842.54	85.29	77.78	88

DOI: 10.1371/journal.pcbi.0020068.t004

previously unseen gene expression profiles. We tested the five gene modules of Figure 3 using their corresponding Boolean functions on the "validation dataset" [11]. For that task, we needed to binarize the expression levels of the validation dataset.

Because of the simplicity of the Boolean functions shown in Figure 3, which are applied to only three genes, *any* choice of threshold for this task yielding good classification results would be remarkable, as a moment's thought would be convincing that a "random" Boolean function cannot have good classification results regardless of the threshold choice. We chose, nevertheless, to find a single transformation formula mapping the EMBP dataset thresholds to the validation dataset thresholds. This is not a straightforward task, because of the lack of standardization in the conditions involved when measurements are made in different laboratories. For example, in our case, the median intensity of the validation dataset was close to ten times that of the EMBP dataset.

We used a simple transformation of the form $y = ax + b$ where x represents thresholds over the EMBP dataset and y represents thresholds over the validation dataset. To estimate the coefficients a and b, we averaged the gene expression values over all tissues for the 12,600 genes common to both datasets. We thus obtained two vectors x and y of length 12,600, whose elements were the mean gene expression levels across all tissues belonging to the EMBP and validation datasets, respectively. Using these two vectors, we calculated the least-squares estimate for the coefficients a and b. The mean value and the 95% confidence bound for the two coefficients were found to be: $a = 8.25 \pm 0.088$ and $b = 92.12 \pm 9.06$. We transformed the thresholds using several values of a and b within the 95% confidence bounds, and selected the values yielding the highest found classification performance, which were $a = 8.338$ and $b = 92.12$. Table 4 summarizes the results for each of the five Boolean functions outlined in Figure 3 over the validation dataset.

Remarkably, the classification accuracy of the simple three-gene (two-gene in one case) Boolean functions in the validation dataset were consistently high, exceeding 90% in most cases, indicating that EMBP analysis accurately extracted universally valid prostate cancer–related features. By comparison, the validation results of the *k*-nearest neighbor-based classification in [10], from which we extracted our input data, using the same training and testing data [11], had classification accuracy of 77% for four-gene models and 86% on 16-gene models, while even our worst accuracy across different thresholds for three-gene classification was 85.29% (see Table 4).

As mentioned earlier, because of overfitting, the classification accuracy on the validation data would not increase had we used four-gene modules for the threshold choices in Table 4. Indeed, the resulting Boolean functions were the same as the Boolean functions for $n = 3$, since the Boolean Parsimony procedure found that the fourth gene in the modules was irrelevant. To illustrate the concept of overfitting, we also evaluated the classification performance of the four-gene modules on the validation datasets by skipping the Boolean Parsimony procedure, instead using the individual states of the four-gene modules as predictors of cancer. In other words, we predicted each state to be cancerous or benign based on the "majority count" calculated from the values of N_0 and N_1 in the training dataset. The classification performance of the four-gene modules across thresholds 30, 45, 60, 75, and 90 was found to be 85.29%, 94.11%, 91.17%, 82.35%, and 79.41% respectively. In other words, the performance of the four-gene modules either worsens or remains the same compared to the three-gene EMBP modules in Table 4. Thus the Boolean Parsimony procedure addresses the problem of overfitting by its very nature.

Biological Interpretation of EMBP Analysis Results

The genes in the modules resulting from EMBP analysis are not co-regulated, because, if they were, then each of them alone would provide much of the information that all of them provide; therefore a different gene would be a more appropriate partner, as it would provide complementary information. Nevertheless, these genes are typically related by a shared common "theme" in which they are playing synergistic roles. For example, two genes may appear because they are both required for the activation of a particular cancer-causing pathway. Of course, the cause-and-effect relationship connecting disease and the presence of particular genes in a gene module is not clear from the results of quantitative analysis alone, and the Boolean functions can only be seen as approximations when they are based on a relatively small set of input data, as in our case.

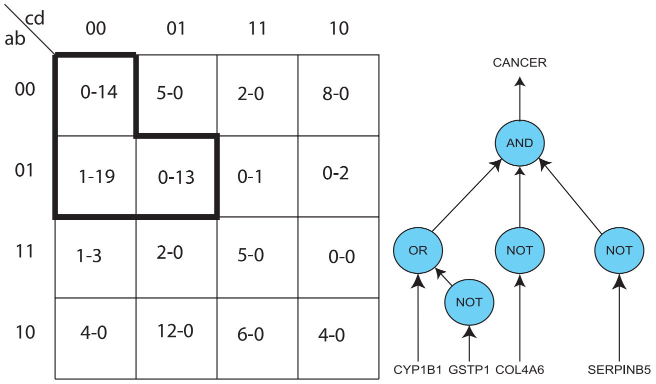
Coupled with additional biological knowledge, however, the results of EMBP analysis can help infer disease-related pathways, which, in turn can help develop therapeutic interventions. This methodology uses the clues provided by the results to create speculative assumptions involving additional genes. Assuming that each gene module has a "story" to say, we can then attempt to combine all these "stories" into an integrated scenario combining many genes. In this section, we provide two examples of this methodology.

We first focus on the three-gene module with the lowest overall conditional entropy (0.1159) (Figure 3C), consisting of genes *HPN, NCF4,* and *PGM1*. Hepsin (HPN) is a serine protease that is overexpressed in most prostate cancers [19]. Recent evidence indicates [20,21] that hepsin converts single-chain pro-hepatocyte growth factor into biologically active two-chain hepatocyte growth factor. The hepatocyte growth factor (HGF) is a ligand for Met, a known proto-oncogene receptor tyrosine kinase, suggesting [20,21] that this functional link between hepsin and the HGF/Met pathway may be related to tumor progression. Furthermore, HGF protects cell against oxidative stress–induced apoptosis [22,23]. These results suggest that hepsin may promote tumor progression by inhibiting the apoptotic mechanisms that are normally activated in cells after they become cancerous as a result of damage caused by oxidative stress.

Interestingly, both of the other members of the module *(NCF4* and *PGM1)* have also been related to oxidative stress, strengthening the above speculative interpretation. Phosphoglucomutase is inhibited under oxidative stress [24]. The absence of *PGM1* (as in the Boolean function of Figure 3C) could therefore result from oxidative stress. On the other hand, *NCF4,* also known as *P40PHOX,* is known to down-regulate [25,26], under some conditions, the NADPH oxidase, a phagocyte enzyme system that creates a superoxide-producing "oxidative burst" in response to invasive microorganisms. In this case, local oxidative stress would result from the reduced levels of *P40PHOX* activity.

Taken together, the above observations suggest a speculative scenario consistent with the Boolean function of Figure 3C: The absence of *NCF4* (perhaps as a result of mutation), if accompanied by other unknown factors, permits activation of the NADPH oxidase, which could be aberrant, i.e., not necessarily responding to the presence of invasive microorganisms. If this happens, then the resulting oxidative burst, evidenced by *PGM1* down-regulation, is damaging to the cell and is normally accompanied by triggering apoptotic mechanisms, which, however, are inhibited by the activated *HGF* resulting from the presence of hepsin. The damaged surviving cell may then become cancerous as a result of additional mutations. This interpretation may also shed light on the problem of the "hepsin paradox" [27]: Although hepsin is overexpressed in the vast majority of prostate cancers and is thought to promote tumor progression, it is unexpectedly underexpressed in metastatic lesions. It could be that the additional mutations in the damaged cells have already inactivated the apoptotic mechanisms, at which stage the expression of hepsin is not needed anymore for the "protection" of the cell, and cancer becomes more invasive. Furthermore, as noted above, the same conditional entropy (0.1159) with the same Boolean function results if we replace gene *NCF4* with gene *KRT6E (keratin 6E).* It is known that mutations in keratin genes can prime cells to oxidative injury [28]. In that case, *KRG6E* is absent due to its mutation, and the resulting oxidative injury is not stemming from NADPH oxidation, but is still manifested by the absence of *PGM1,* and the apoptotic mechanisms are still inhibited by the presence of hepsin.

There are many more gene modules that are revealed by EMBP analysis in addition to those indicated in Figure 3, and their complete presentation and interpretation is beyond the scope of this paper. We note, however, that the "theme" of oxidative stress combined by inhibition of apoptosis is encountered in several of them. Therefore, the credibility of this speculative interpretation is strengthened, because it can be consistently made multiple times with respect to other gene modules. Confirming the plausibility of this hypothesis, independent studies have also linked oxidative stress to prostate cancer [29]. Here we present one more such gene module that we found, which is particularly relevant in that respect.



ab \ cd	00	01	11	10
00	0-14	5-0	2-0	8-0
01	1-19	0-13	0-1	0-2
11	1-3	2-0	5-0	0-0
10	4-0	12-0	6-0	4-0

a: COL4A6 (Collagen type IV, alpha 6)
b: CYP1B1 (Cytochrome P450, family 1, subfamily B, polypeptide 1)
c: SERPINB5 (Maspin)
d: GSTP1 (Glutathione S-transferase pi)

Boolean function: (b+d')a'c'

Figure 4. Karnaugh Map and Tree Representation of the Most Parsimonious Boolean Function for the Minimum Entropy Four-Gene Module at Threshold = 15

Cancerous tissues are associated with the absence of both *COL4A6* and *SERPINB5* accompanied by either the presence of *CYP1B1* or absence of *GSTP1*.

DOI: 10.1371/journal.pcbi.0020068.g004

This module resulted from a threshold choice of 15. As seen in Figure 2A, the conditional entropy of the optimum three-gene module (shown outside the dotted box) is relatively high, exceeding 0.3. Therefore, we found the lowest-entropy four-gene module. The result is shown in Figure 4, and is the one that we had also used to generate Table 1 with corresponding minimum entropy of 0.088. Gene a is *COL4A6 (collagen type IV alpha 6),* gene b is *CYP1B1 (cytochrome P450, family 1, subfamily B, polypeptide 1),* gene c is *SERPINB5 (maspin, or serpin peptidase inhibitor, clade B, member 5),* and gene d is *GSTP1 (glutathione S-transferase pi)*. The most parsimonious Boolean function containing the entries of the Karnaugh map predominantly associated with cancer is (b+d′)a′c′.

The meaning of this function is the following: Cancerous tissues are associated with the absence of both *COL4A6* and *SERPINB5* accompanied by either the presence of *CYP1B1* or absence of *GSTP1*. Figure 4 also shows the Boolean function represented by a tree structure, which can be useful for providing insight into the flow of information in pathways when the number of genes in a module is high.

The common theme in this cluster is easily detected by the simultaneous presence of two genes, which have the following products: On the one hand, CYP1B1 is an oxidative enzyme induced by, and metabolizing, various substances, several of which are toxins, and on the other hand GSTP1 is another enzyme whose role is to detoxify by catalyzing conjugation of glutathione. Products of oxidative metabolism are "natural" substrates for the glutathione transferases [30]. Therefore, consistent with the observation that CYP1B1 and GSTP1 are coordinated in sequential reactions [31], it is reasonable to assume that, in normal prostate tissues, part of the role of GSTP1 is to respond to oxidative stress by detoxifying procarcinogens that have been activated by oxidative enzymes such as CYP1B1.

CYP1B1 has been found overexpressed and regulated by promoter hypomethylation in prostate cancer [32] and several cancer therapeutic options associated with *CYP1B1* are under consideration [33]. On the other hand, *GSTP1* has been found underexpressed in prostate cancer, often as a result of hypermethylation of its regulatory sequences [34,35]. Interestingly, there are three known gene variants of *GSTP1,* and the gene variant *(GSTP1*C)* that we include in the Boolean function is the one, among the three, that was shown to contain functional retinoic acid response elements (RAREs) in its introns, confirmed by the observation that retinoic acid treatment significantly increased *GSTP1*C* gene expression in glioblastoma cells, a fact that may contribute to a better understanding of the molecular regulation of the *GSTP1* gene in human cells [36] and provide possibilities for therapeutic intervention. We note that we have identified gene *RBP1 (retinol binding protein 1)* as a member of several of the gene modules that we found (one of which also appears in Figure 3), and it was recently hypothesized that the protective role of *RBP1* is due to inhibition of the PI3K/AKT survival pathway through a retinoic acid receptor–dependent mechanism [37].

A third molecule in the module of Figure 4 is *SERPINB5,* or *maspin,* a serine protease inhibitor. Its product has been identified as a tumor suppressor, originally in breast cancer [38]. The absence of its expression has also been linked to promoter methylation [39]. Given the identified theme of the gene module, we would like to see in what ways maspin is related to oxidative stress. Strikingly, precisely this role has recently been proposed for maspin in prostate cancer tissues [40]: that maspin may inhibit oxidative stress–induced generation of reactive oxygen species (such as free radicals) by interacting with glutathione S-transferase, thus preventing adverse effects on tumor genetics. This hypothesis is consistent with the entries of Figure 4, indicating that a combination of *GSTP1* and *SERPINB5,* and not *GSTP1* alone, as noted earlier, is needed to counteract the expression of *CYP1B1*. Indeed, although all 13 tissues associated with state 0101 are cancerous ($N_0 = 0$ and $N_1 = 13$), all five tissues associated with state 1111, in which maspin is expressed in addition to *GSTP1,* are healthy ($N_0 = 5$ and $N_1 = 0$).

More specifically, maspin was found to interact with three molecules: The two molecules in addition to glutathione S-transferase were two "stress proteins" (heat shock proteins HSP70 and HSP90), suggesting that intracellular maspin may be primarily involved in cellular response to stress stimuli [40] by inducing apoptosis [41]. Furthermore, heat shock proteins are known to be induced by oxidative stress and they may play oncogenic roles by "protecting" cells from apoptosis. For example, it has been found that reactive oxygen species play important roles in the activation of HSF1 (heat shock factor 1, the primary transcription factor responsible for the transcriptional heat stress response in mammalian cells) and in the accumulation of mRNA of the previously mentioned genes *HSP70* and *HSP90* in the ischemic-reperfused heart [42].

From the observations so far, it follows that GSTP1 and maspin may work synergistically to counteract oxidative stress, perhaps by maspin inducing apoptosis whenever GSTP1-induced detoxification is not successful. Interestingly, maspin, a serine protease inhibitor, and hepsin, a serine protease that we previously speculated to play an apoptosis-inhibitive role, were found to be inversely expressed in prostate cancer [43]. Even without the presence of *CYP1B1,* the simultaneous absence of both GSTP1 and maspin appears to be sufficient to cause cancer, as indicated by the entries at states 0000 and 0100 in Figure 4.

The only remaining gene in that module is *COL4A6,* which encodes one of the six subunits of type IV collagen, the major structural component of basement membranes. The association that we found, using EMBP analysis, between *COL4A6* and prostate cancer is remarkable, because *COL4A6* expression is missing in nearly all cancerous tissues as evidenced by the Boolean function. This absence of *COL4A6* expression in prostate cancer has also been observed before [44].

There are two identified splice variants of *COL4A6:* splice variant A and splice variant B, each of which uses a distinct promoter [45]. The specific gene in our microarray data is splice variant B. It was recently found [46] that splice variant B is regulated by HSF1 (heat shock factor 1, mentioned earlier) [42]. Specifically, HSF1 binding was detected at the promoter of splice variant B [46], suggesting that *COL4A6* plays a novel role in the heat shock response. Therefore, the possibility exists that one promoter confers tissue specificity for *COL4A6*'s role in the basement membrane and the other promoter regulates the gene for its potentially novel role [46]. This novel role is unclear, but collagen IV has already been found to regulate pathways related to oxidative stress through ERK activation [47,48].

Together with the fact, noted earlier, that maspin interacts with *HSP70* and *HSP90,* these observations further strengthen the case that the gene module is related to oxidative stress including heat shock–related stress. Interestingly, *HSP70* has been found [49] overexpressed in some prostate cancers and in the plasma levels in prostate cancer patients [50].

Therefore, a plausible speculative scenario for this module is the following: *CYP1B1* expression is induced by some toxins and, at least in some *CYP1B1* polymorphisms [51], they are activated and contribute to a type of cancer-causing oxidative stress. This oncogenic effect can be countered by the expression of *GSTP1* and *maspin,* resulting in either detoxification or apoptosis. In particular, absence of *maspin* may inhibit the apoptotic mechanisms normally activated in these cases. The role of the prominent observed down-regulation of *COL4A6* in the cancerous tissues is unclear, but it may be related to the activation of the heat shock response as a result of the oxidative stress, and it may indicate that collagen plays an important and yet unrecognized role in prostate cancer.

Discussion

The most notable feature of the EMBP method is that it is systems-based, in the sense that it considers the synergistic contributions of sets of genes, rather than individual genes, or sets of co-regulated genes as in microarray clustering techniques. As a result, the optimal gene module of size $n-1$ may not be a subset of the optimal gene module of size n, because the n members of the latter module may interact synergistically toward predicting disease in a manner that cannot be achieved if any one of the n members is removed. EMBP analysis actually discovers such gene modules mechanistically, and can therefore complement other gene-set–based techniques, such as "Gene Set Enrichment Analysis" [4], in which the gene sets are defined based on prior knowledge, such as published information about biochemical pathways.

Discrimination between healthy and diseased tissues, or between different types of diseases, is not the main purpose of EMBP analysis, although it can be used for this task. If we wanted to focus on classification, then we would use many genes, connect their expression levels in sophisticated ways, rather than through one extremely simple Boolean function, and use their continuous expression levels instead of binarizing them, so that we can extract additional information. Nevertheless, we still validated our results for classification accuracy using the binarized data from sets of three genes connected by simple Boolean functions, applied on totally new microarray data, independently generated in a different lab. The resulting high classification accuracy of typically more than 90% (Table 4), confirms that the extracted logic function connecting these genes is valid in the testing set as well, and therefore it is likely to be a universal property in prostate cancer tissues. In particular, the "notable facts," mentioned earlier, resulting from our Boolean functions, indicate a precise functional relationship among the genes leading to disease, which is valid on both the training and the testing datasets.

We tested the effect of sample size on the results by performing EMBP analysis on several randomly selected subsets of the training data using equal numbers (10, 20, 30, and 40) of samples from normal and cancerous tissues. The minimum entropy results are shown in Figure 5. We observe that the average minimum entropy values for $n = 1$, 2, 3, and 4 increase with the number of samples. Furthermore, it is also seen that for lower number of samples, the entropy reaches zero for lower values of n when compared with higher sample datasets. However, the entropy values appear to increase at a low rate as n becomes close to 50. Thus, we expect that for prostate cancer analysis, having more than a 100 samples from each category would not add additional information.

As mentioned earlier, the genes in the modules from EMBP analysis are not co-regulated, but they are co-operative. This is in sharp contrast to the gene modules resulting from clustering or bi-clustering approaches [52–58], the very nature of which is to cluster genes with similar expression patterns. Several of these clustering approaches use entropy extremization and other information theoretic measures, but in totally different contexts compared with our purposes.

For example, in reference [56], entropy minimization is used to reduce the amount of "disorder" within each cluster so that data points within a cluster are similar to each other. In reference [57], the mutual information between different clusters is minimized so that expression profiles falling into different clusters are maximally different from each other. In reference [58], pairwise mutual information is maximized to cluster genes together so that the genes within each cluster have maximum relevance with each other. In all such clustering approaches, the principal strategy is to cluster together genes that have similar expression levels given a particular phenotype. In contrast, the genes within an EMBP module are chosen such that we minimize the uncertainty of assigning correct phenotype labels from the joint expression values of the genes in that module. Thus the genes within the EMBP gene set can have very different expression patterns given the phenotype.

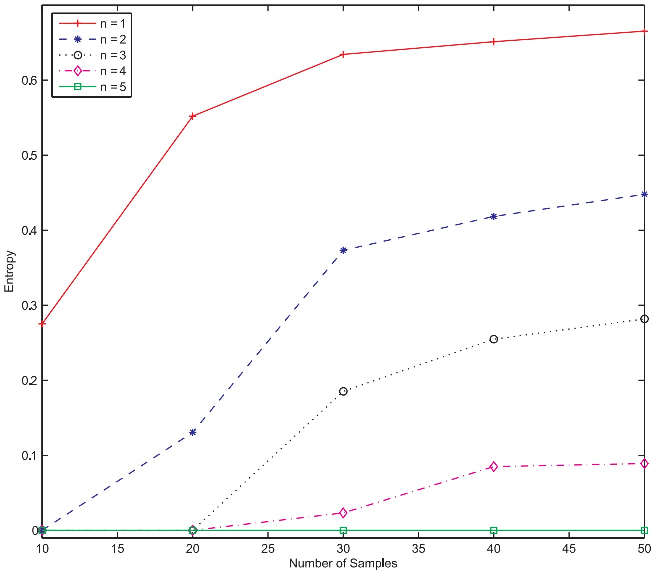


Figure 5. Minimum Entropy Values Averaged over Four Trials across Different Sample Sizes for Threshold = 15

The minimum entropy values increase with sample size for any given n, but they tend to saturate for large numbers of samples.

DOI: 10.1371/journal.pcbi.0020068.g005

We believe that EMBP analysis provides an opportunity for fruitful cross-disciplinary collaboration in which biologists use the "clues" resulting from the computational results to infer potential pathways, which they can validate with genetic experiments, as well as suggest further computational experiments. For example, if we wish to identify which genes play synergistic roles with another particular gene in terms of causing disease, we could "freeze" the presence of that gene and identify the other genes in a module minimizing the entropy. Furthermore, our systems-based approach can immediately suggest novel potential therapeutic methods that would not be possible with traditional individual-gene approaches, for example, targeting simultaneously two genes that appear in the same Boolean function by combining two already existing drugs targeting each of these genes.

We hope that EMBP analysis will prove to be a significant new tool for medical research working synergistically with the future efforts of diseased tissue genome sequencing. To illustrate with an example, if a Boolean function $a'b'c'$ is found when analyzing expression data of a particular cancer, this would suggest the possibility that the three genes "gene a," "gene b," and "gene c" may cause cancer when all are inactivated, perhaps due to their mutation or to hypermethylation of their promoter, as we previously discussed regarding *KRT6E* and *GSTP1*. In turn, this observation may provide motivation to sequence these genes in cancerous tissues. For EMBP analysis to be significantly effective, in addition to diseased tissues, hundreds of healthy tissues must also be profiled for each tissue type. Inclusion of data from healthy tissues has so far not been emphasized, mainly because microarray data were thought to be more useful for classifying among disease subtypes, rather than detecting disease from non-disease, which is often achievable by normal biopsies. Furthermore, it is important to use large and high-quality input datasets obtained under standardized conditions.

Materials and Methods

Entropy minimization. We used a combination of two heuristic optimization techniques to search for minimum entropy gene sets, allowing sufficient time for each of them to converge. The first technique is the following: Starting from a randomly chosen gene set of size *n,* at each step of the iteration, we modify the "current" gene set by replacing one of its genes, chosen at random, with a new gene, also chosen at random from the entire gene set *M.* If the conditional entropy of the new gene set is lower than that of the current gene set, then the new gene set replaces the current gene set. The process terminates when the conditional entropy is 0, or when the current gene set remains unmodified for a particular large number of steps. To avoid selecting a local minimum, we repeat the same iterative algorithm several times, starting from different initial conditions of the same size, and select the gene set that yields the overall lowest conditional entropy. We then increase the size of the gene set to $n+1$ and repeat the whole process, making sure that one of the chosen initial conditions contains the previously found gene set. This technique typically converges to some choice of near-optimum results.

To reduce the chance that the found solution corresponds to a local, rather than global minimum, we also used simulated annealing [59] to search in the space of all subsets of size *n.* In an "annealing" process, a melt, initially disordered and at high "temperature" *T,* is slowly cooled. As cooling proceeds, the system becomes more ordered and approaches a "frozen" state at $T=0$. In our case, we started from a randomly chosen gene set of size *n* and replaced a randomly chosen gene in the set by another randomly chosen gene from the entire set of *M* genes. If the conditional entropy of the new gene set is found to be lower than that of the current gene set, we replace the current set with the new set. If, however, the conditional entropy of the new gene set is higher than that of the current gene set, we allow replacement of the current gene set with a probability p_accept that is proportional to the temperature T_k at the time and inversely proportional to the amount by which the conditional entropy of the new gene set is higher than that of the current gene set. Thus, the value of p_accept at any given temperature T_k is given by:

$$p\_accept(T_k, \delta H) = \exp\left(-\frac{\delta H}{T_k}\right) \tag{5}$$

where δH is the increase in conditional entropy due to the random change in the gene set.

The parameters of the simulated annealing algorithm that we used were:

$$p\_init = 0.3 \tag{6}$$

which is the initialization value of p_accept),

$$T\_init = \frac{-\langle \delta H \rangle}{\log_e(p\_init)} \tag{7}$$

where $\langle \delta H \rangle$ is the average increase in conditional entropy due to a random change in a gene set,

$$L_k = 1,500 \tag{8}$$

which is the total number of transitions at the k^{th} temperature,

$$\eta_{\min} = 500 \tag{9}$$

which is the minimum number of acceptances at the k^{th} temperature, and

$$k_{\max} = 1,000 \tag{10}$$

which is the total number of temperature values to be tried over the course of the algorithm.

The "cooling scheme" defines the rate at which the temperature falls over the length of the algorithm. We adopted the exponential cooling scheme defined as $T_{k+1} = \alpha T_k$, where $\alpha = 0.98$. The simulated annealing algorithm allows for significant increases in the conditional entropy at higher temperatures, searching coarsely in the space of subsets of genes. As the temperature falls, the probability of accepting even small increases in conditional entropy is reduced, thus searching only in the local neighborhood of the conditional entropy value.

An average run of the simulated annealing algorithm for estimating the minimum conditional entropy gene-set for $n=3$, on a Pentium III processor running at 3 GHz is around 30 min.

Supporting Information

Accession Numbers

The National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) accession numbers for the genes discussed in the paper are as follows: *COL4A6* (D21337), *COL4A6* splice variant A (NM__001847), *COL4A6* splice variant B (NM__033641), *CYP1B1* (U03688), *DF* (M84526), *ENTPD1* (AJ133133), *FNBP1* (AB011126), *GSTP1* (U12472), *HIST1H1E* (M60748), *HLA-DQB1* (M81141), *HPN* (X07732), *KRT6E* (L42611), *MCM3AP* (AB011144), *NCF4* (AL008637), *NELL2* (D83018), *PGM1* (M83088), *RBP1* (M11433), *SERPINB5* (U04313), *SPINK2* (X57655), and *TMSL8* (D82345).

Acknowledgments

Author contributions. VV and DA conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, and wrote the paper.

Funding. This research was funded by DA's research share from Columbia University's allocation policy of patent royalties from an invention unrelated to this paper.

Competing interests. The authors have declared that no competing interests exist.

References

1. Ramaswamy S, Golub TR (2002) DNA microarrays in clinical oncology. J Clin Oncol 20: 1932–1941.
2. Garber K (2004) Genomic medicine: Gene expression tests foretell breast cancer's future. Science 303: 1754–1755.
3. Segal E, Friedman N, Kaminski N, Regev A, Koller D (2005) From signatures to models: Understanding cancer using microarrays. Nat Genet 37: S38–S45.
4. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
5. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, et al. (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267–273.
6. Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36: 1090–1098.
7. Lamb J, Ramaswamy S, Ford HL, Contreras B, Martinez RV, et al. (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. Cell 114: 323–334.
8. Huang E, Ishida S, Pittman J, Dressman H, Bild A, et al. (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. Nat Genet 34: 226–230.
9. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, et al. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 101: 9309–9314.
10. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, et al. (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1: 203–209.
11. Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, et al. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res 61: 5974–5978.
12. Cover TM, Thomas JA (1991) Elements of information theory. New York: Wiley Interscience. p. 15.
13. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27: 379–423.
14. Boole G (1854) An investigation of the laws of thought, on which are founded the mathematical theories of logic and probabilities. London: Walton and Maberly. 424 p.

15. Mano MM (1979) Digital logic and computer design. Englewood Cliffs (New Jersey): Prentice-Hall. 612 p.
16. Brayton RK, Hachtel G, McMullen C, Sangiovanni-Vincentelli A (1984) Logic minimization algorithms for VLSI minimization. 7th edition. Boston: Kluwer Academics. 193 p.
17. Yang C, Ciesielski M (2002) BDS: A BDD-based logic optimization system. IEEE Trans CAD 21: 866–876.
18. Walpole RE, Mayers RH, Mayers SL, Ye K, Yee K, (2002) Probability and statistics for engineers and scientists. Upper Saddle River (New Jersey): Prentice Hall. 730 p.
19. Magee JA, Araki T, Patil S, Ehrig T, True L, et al. (2001) Expression profiling reveals hepsin overexpression in prostate cancer. Cancer Res 61: 5692–5696.
20. Kirchhofer D, Peek M, Lipari MT, Billeci K, Fan B, et al. (2005) Hepsin activates pro-hepatocyte growth factor and is inhibited by hepatocyte growth factor activator inhibitor-1B (HAI-1B) and HAI-2. FEBS Lett 579: 1945–1950.
21. Herter S, Piper DE, Aaron W, Gabriele T, Cutler G, et al. (2005) Hepatocyte growth factor is a preferred in vitro substrate for human hepsin, a membrane-anchored serine protease implicated in prostate and ovarian cancers. Biochem J 390: 125–136.
22. Kitta K, Day RM, Ikeda T, Suzuki YJ (2001) Hepatocyte growth factor protects cardiac myocytes against oxidative stress-induced apoptosis. Free Radic Biol Med 31: 902–910.
23. Okada M, Sugita K, Inukai T, Goi K, Kagami K, et al. (2004) Hepatocyte growth factor protects small airway epithelial cells from apoptosis induced by tumor necrosis factor-α or oxidative stress. Pediatr Res 56: 336–344.
24. Kanazawa K, Ashida H (1991) Relationship between oxidative stress and hepatic phosphoglucomutase activity in rats. Int J Tissue React 13: 225–231.
25. Lopes LR, Dagher MC, Gutierrez A, Young B, Bouin AP, et al. (2004) Phosphorylated p40PHOX as a negative regulator of NADPH oxidase. Biochemistry 43: 3723–3730.
26. Sathyamoorthy M, de Mendez I, Adams AG, Leto TL (1997) p40(phox) down-regulates NADPH oxidase activity through interactions with its SH3 domain. J Biol Chem 272: 9141–9146.
27. Vasioukhin V (2004) Hepsin paradox reveals unexpected complexity of metastatic process. Cell Cycle 3: 1394–1397.
28. Zhou Q, Ji X, Chen L, Greenberg HB, Lu SC, et al. (2005) Keratin mutation primes mouse liver to oxidative injury. Hepatology 41: 517–525.
29. Ouyang X, DeWeese TL, Nelson WG, Abate-Shen C. (2005) Loss-of-function of Nkx3.1 promotes increased oxidative damage in prostate carcinogenesis. Cancer Res 65: 6773–6779.
30. Mannervik B, Danielson UH (1988) Glutathione transferases—structure and catalytic activity. CRC Crit. Rev. Biochem. 23: 283–337.
31. Hatchey DL, Dawling S, Roodi N, Parl FF (2003) Sequential action of phase I and II enzymes cytochrome p450 1B1 and glutathione S-transferase P1 in mammary estrogen metabolism. Cancer Res 63: 8492–8499.
32. Tokizane T, Shiina H, Igawa M, Enokida H, Urakami S, et al. (2005) Cytochrome P450 1B1 is overexpressed and regulated by hypomethylation in prostate cancer. Clin Cancer Res 11: 5793–5801.
33. McFadyen MC, Melvin WT, Murray GI (2004) Cytochrome P450 enzymes: Novel options for cancer therapeutics. Mol Cancer Ther 3: 363–371.
34. Lee WH, Morton RA, Epstein JI, Brooks JD, Campbell PA, et al. (1994) Cytidine methylation of regulatory sequences near the π-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. Proc Natl Acad Sci U S A 91: 11733–11737.
35. Zhang YJ, Chen Y, Ahsan H, Lunn RM, Chen SY, et al. (2005) Silencing of glutathione S-transferase P1 by promoter hypermethylation and its relationship to environmental chemical carcinogens in hepatocellular carcinoma. Cancer Lett 221: 135–143.
36. Lo HW, Ali-Osman F (1997) Genomic cloning of hGSTP1*C, an allelic human pi class glutathione S-transferase gene variant and functional characterization of its retinoic acid response elements. J Biol Chem 272: 32743–32749.
37. Farias EF, Marzan C, Mira-y-Lopez R (2005) Cellular retinol-binding protein-I inhibits PI3K/Akt signaling through a retinoic acid receptor-dependent mechanism that regulates p85-p110 heterodimerization. Oncogene 24: 1598–1606.
38. Zou Z, Anisowicz A, Hendrix MJ, Thor A, Neveu M, et al. (1994) Maspin, a serpin with tumor-suppressing activity in human mammary epithelial cells. Science 263: 526–529.
39. Futscher BW, Oshiro MM, Wozniak RJ, Holtan N, Hanigan CL, et al. (2002) Role for DNA methylation in the control of cell type-specific maspin expression. Nat Genet 31: 175–179.
40. Yin S, Li X, Meng Y, Finley RL Jr, Sakr W, et al. (2005) Tumor-suppressive maspin regulates cell response to oxidative stress by direct interaction with glutathione S-transferase. J Biol Chem 280: 34985–34996.
41. Zhang W, Shi HY, Zhang M (2005) Maspin overexpression modulates tumor cell apoptosis through the regulation of Bcl-2 family proteins. BMC Cancer 5: 50.
42. Nishizawa J, Nakai A, Matsuda K, Komeda M, Ban T, et al. (1999) Reactive oxygen species play an important role in the activation of heat shock factor 1 in ischemic-reperfused heart. Circulation 99: 934–941.
43. Chen Z, Fan Z, McNeal JE, Nolley R, Caldwell MC, et al (2003) Hepsin and maspin are inversely expressed in laser capture microdissectioned prostate cancer. J Urol 169: 1316–1319.
44. Dehan P, Waltregny D, Beschin A, Noel A, Castronovo V, et al. (1997) Loss of type IV collagen alpha 5 and alpha 6 chains in human invasive prostate carcinomas. Am J Pathol 151: 1097–1104.
45. Sugimoto M, Oohashi T, Ninomiya Y (1994) The genes COL4A5 and COL4A6, coding for basement membrane collagen chains alpha 5(IV) and alpha 6(IV), are located head-to-head in close proximity on human chromosome Xq22 and COL4A6 is transcribed from two alternative promoters. Proc Natl Acad Sci U S A 91: 11679–11683.
46. Trinklein ND, Murray JI, Hartman SJ, Botstein D, Myers RM (2004) The role of heat shock transcription factor 1 in the genome-wide regulation of the mammalian heat shock response. Mol Biol Cell 15: 1254–1261.
47. Sanders MA, Basson MD (2004) Collagen IV regulates Caco-2 migration and ERK activation via alpha1beta1- and alpha2beta1-integrin-dependent Src kinase activation. Am J Physiol Gastrointest Liver Physiol 286: G547–G557.
48. Buckley S, Driscoll B, Barsky L, Weinberg K, Anderson K, et al. (1999) ERK activation protects against DNA damage and apoptosis in hyperoxic rat AEC2. Am J Physiol 277: L159–L166.
49. Rose A, Xu Y, Chen Z, Fan Z, Stamey TA, et al. (2005) Comparative gene and protein expression in primary cultures of epithelial cells from benign prostatic hyperplasia and prostate cancer. Cancer Lett 227: 213–222.
50. Abe M, Manola JB, Oh WK, Parslow DL, George DJ, et al (2004) Plasma levels of heat shock protein 70 in patients with prostate cancer: A potential biomarker for prostate cancer. Clin Prostate Cancer 3: 49–53.
51. Chang BL, Zheng SL, Isaacs SD, Turner A, Hawkins GA, et al. (2003) Polymorphisms in the CYP1B1 gene are associated with increased risk of prostate cancer. Br J Cancer 89: 1524–1529.
52. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array. Proc Natl Acad Sci U S A 96: 6745–6750.
53. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. Nat Genet 22: 281–285.
54. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.
55. Shamir R, Sharan R (2000) CLICK: A clustering algorithm for gene expression analysis. In: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology. Menlo Park (California): AAAI Press. pp. 307–316.
56. Li H, Zhang K, Jiang T (2004) Minimum Entropy Clustering and Applications to gene expression analysis. In: Proceedings of the 3rd IEEE Computational Systems Bioinformatics Conference, Stanford, California. Washington (D. C.): IEEE Computer Society. pp. 142–151.
57. Zhou X, Wang X, Dougherty ER, Russ D, Suh E (2004) Gene clustering based on clusterwide mutual information. J Comput Biol 11: 147–161.
58. Butte AJ, Kohane IS (2000) Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput: 418–429.
59. Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. Science 220: 671–680.