10

Feature Selection in Microarray Data Using Entropy Information

Ali Reza Soltanian[1] • Niloofar Rabiei[2] • Fatemeh Bahreini[3]

[1]Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran; [2]Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran; [3]Department of Molecular Medicine and Genetics, Faculty of Medicine, Hamadan University of Medical Sciences, Hamadan, Iran

Author for correspondence: Ali Reza Soltanian, Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran.
Email: soltanian@umsha.ac.ir

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch10

Abstract: Researchers in biological sciences and genetics are faced with high-dimensional data, such as the microarray data, and the analysis and proper interpretation of these data are very important in bioinformatics and systems biological sciences. In such types of data, the number of variables, for example, the genes, is many times greater than the number of samples. Therefore, the dimension of the data must be reduced at the primary point. Then, the analysis, for example, clustering, is performed on the compacted data. This process is called data summarization. There are various ways to summarize high-dimensional data, which depends on the nature of the data. The aim of data summarization is to remove unnecessary features so that the data are classified more accurately. Shannon's entropy information is a common method for clustering genes in microarray data and selecting a set of disease-related genes. This chapter introduces and illustrates statistical inference concepts of entropy in microarray data clustering to select a set of the most important genes associated with a disease.

Keywords: data mining; entropy; genetics; microarray; system biology.

In: Computational Biology. Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

To analyze high-dimensional data, many mathematical and statistical models have been developed. Most of these models focus on eliminating the unnecessary and unimportant features of a dataset, so that the clustered data are accurate. A popular source of clustering and modeling of high-dimensional data is the microarray data. The concept of systems biology has become more prominent in biological sciences (1). Systems biology is the science of summarizing data and detecting patterns among datasets. In other words, systems biology is the computational modeling of biosystems to interpret high and complex genetic data and other complex biological systems (2–4). Systems biology incorporates computational science, mathematics and statistics in the modeling of genetic and biological data (Figure 1). Entropy is one of the mathematical concepts that can be used in the modeling of systems biology data. In entropy, there are two concepts: entropy and information. Researchers usually do not distinguish between the two concepts. Entropy represents the irregularity (i.e., uncertainty) of a system, while information represents the difference between the maximum and the actual value of entropy of a system. In other words, information shows the correlation between two systems (e.g., two genes), which is derived from the entropy of the two systems and their subscription (5). Entropy application is a kind of mathematical challenge in analyzing biological data that can be important in determining relationships and clustering of results. Researchers have used entropy techniques to model cellular systems and study changes in gene expression patterns. In this chapter, the role of entropy to model the expression of genes in microarray data is discussed with emphasis on clustering, refinement and Shannon's entropy theory.

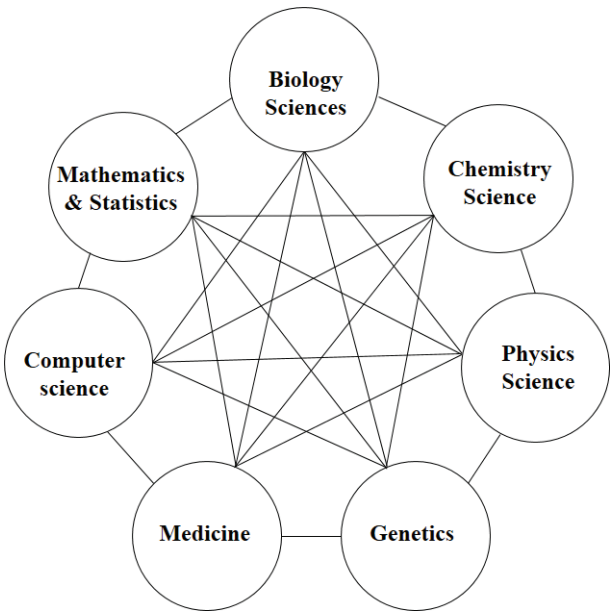

Figure 1 Schematic diagram for the concept of communication in systems biology.

DATA NORMALISATION METHOD

Data refinement is very important in the analysis of complex systems such as the microarray data. The calculation of gene expression in the microarray technique is based on the coloration of the genes, and problems associated with coloration are not uncommon. The occurrence of such problems leads to an unreasonable or artificial increase or decrease in the expression level of genes. A simple method to avoid outliers is to use *mean* ± *3SD* and, occasionally, *mean* ± *2SD* intervals. This approach eliminates the values outside of these ranges. Another approach to avoid staining errors in the microarray data is the fold-change criterion. Usually, this criterion is obtained based on the expression of a gene in healthy and diseased samples as follows:

$$Fold-change = 2^{\left|\log_2\left(\frac{Ave(C)}{Ave(N)}\right)\right|},$$

where the mean of gene expression levels in healthy and patient samples is indicated by *Ave*(*C*) and *Ave*(*N*), respectively. A cut-off is considered for the obtained fold-changes, so that fold-changes less than the cut-off are usually left out of the analysis process.

SHANNON'S ENTROPY

In the analysis of high-dimensional data, there are two approaches to estimating parameters and effects: the classical approach (i.e., frequentist) and the Bayesian approach. Entropy is a classical approach, and it indicates the degree of irregularity or uncertainty in a system. Uncertainty exists in many of the learning stages of high-dimensional data (6). Although the concept of entropy is used and defined in physics and mathematical sciences, we have attempted to determine a set of coordinates with the least irregularity in a signaling complex using the concept of entropy. Entropy is based on the concept of uncertainty, which means one is unconfident about the occurrence of a process. Therefore, increasing the uncertainty of a system means reducing the entropy of that system. In fact, evaluation, measurement and modeling of uncertainty that affects the whole process of data analysis have a significant impact on the learning performance of high-dimensional data. Without considering this uncertainty, the performance of learning strategies is sharply reduced. Claude E. Shannon, an American mathematician, introduced Shannon's entropy and information theory in 1948 under the title "A mathematical theory of communication" (7). In the concept of entropy, Shannon refers to the degree of uncertainty in the received information and expresses it with probability theory. Shannon's entropy in information theory is the criteria for measuring the uncertainty expressed by a probability distribution.

Note, information theory is the expectation value of information (i.e., mean) contained in each variable which can also be a gene. In other words, the entropy of each variable is the amount of its uncertainty. To calculate the uncertainty of a system, we must be able to formulate the probability of events in that system. Let us consider a

random experiment X (e.g., microarray data) with m events $x_1, x_2,\ldots, x_m$, with the occurrence probabilities $p(x_1), p(x_2),\ldots, p(x_m)$, respectively. In this case, we can consider x_i as the i^{th} gene, in a microarray dataset. The uncertainty of X (i.e., entropy) is represented by $H(X)$, and the function must depend only on the $p(x_i)$, $i = 1, 2,\ldots, m$. The formulation of $H(X)$ function should be the following properties.

Property 1: The desired function should not be dependent on the sequence of events (e.g., genes), hence:

$$H(p_1, p_2,\ldots, p_i, p_{i+1},\ldots, p_m) = H(p_1, p_2,\ldots, p_{i+1}, p_i,\ldots, p_m).$$

Note, the $H(X)$ function on any $p_i = p(x_i)$, $i = 1,2,\ldots,m$ must be continuous.

Property 2: Since the entropy function is continuous, so with a slight change in the probabilities $p(x_1), p(x_2),\ldots, p(x_m)$, the amount of uncertainty (i.e., entropy value) will also change.

Property 3: If an event divides into two events, the original $H(X)$ function must be the sum of the weighed $H(X)$ functions.

Property 4: The entropy function $H(p_1, p_2,\ldots,p_i, p_{i+1},\ldots, p_m)$ should be established in the following equation:

$$H(p_1, p_2,..., p_m) = H(p_1 + p_2, p_3,..., p_m) + (p_1 + p_2)H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

Property 5: Let, two experiments X and Y with m and n events ($n < m$), respectively. If occurrence probability of the events in the two experiments is $\frac{1}{m}$ and $\frac{1}{n}$, then:

$$H\left(\frac{1}{m},...,\frac{1}{m}\right) \geq H\left(\frac{1}{n},...,\frac{1}{n}\right).$$

It can easily be shown that the entropy function has the above properties. Shannon's entropy can be defined for a random variable with a discrete or continuous distribution (7). In this section, we try to mention both together and illustrate the concept of entropy by several examples. Let a discrete random variable such as $X = \{x_1, x_2,\ldots, x_m\}$ with a probability mass function $p(x)$. The entropy of X is:

$$H(X) = \Sigma_{i=1}^{n}\ p(x_i) log_2\left(\frac{1}{p(x_i)}\right) = -\Sigma_{i=1}^{n}\ p(x_i) log_2 p(x_i) = -E[log(p_x)],$$

where $p(x_i) = \Pr(X = x_i)$ and is the probability of the i^{th} value of the random variable X.

Now, let a continuous random variable X. Usually, entropy for the continuous random variables is called the differential entropy. The entropy value for the continuous random variable X with the probability density function $f(X)$ is:

$$H(X) = H(f(x)) = E[-log(f(x))] = -\int f(x) log(f(x)) dx,$$

where $0log0 = 0$.

The entropy may have a logarithmic base 2, 10 or Euler's number *e*. If the logarithmic base is 2 or *e*, then the entropy unit is "bit" or "nat," respectively. Here we should note that some physicists and mathematicians such as Lazare Carnot, Ludwig Eduard Boltzmann, and Rudolf Clausius have tried to introduce the concept of entropy theory, and others such as Claude Elwood Shannon were leading the introduction of entropy information theory (4, 7, 8).

Let, two random variables *X* and *Y* with probability density functions *f*(*X*) and *f*(*Y*) from the support regions *S* and *T*, respectively. The three entropies (i.e., *H*(*X*), *H*(*Y*), and *H*(*X, Y*)), the mutual information *I*(*X; Y*), and the entropy of *X* conditioned on *Y* and vice versa (i.e., *H*(*X*|*Y*) and *H*(*Y*|*X*)), are shown graphically in Figure 2.

The above features will be described below. The entropy function is fundamentally different from the maximum likelihood function. To further understand the entropy concept compared to maximum likelihood, let a Bernoulli random variable X with parameter p. The entropy value of the Bernoulli random variable is:

$$H(X) = -\Sigma_{X=0}^{1} p^X (1-p)^{1-X} log_b \left(p^X (1-p)^{1-X} \right) = -(1-p) log_b (1-p) - p log_2 (p),$$

where $b = 2,10$ *or* e; then,

$$H(X) = log_b \left(\frac{(1-p)^{p-1}}{p^p} \right).$$

The Bernoulli entropy function has the highest value when $p = \frac{1}{2}$. Here are some simple examples for understanding entropy and calculating it.

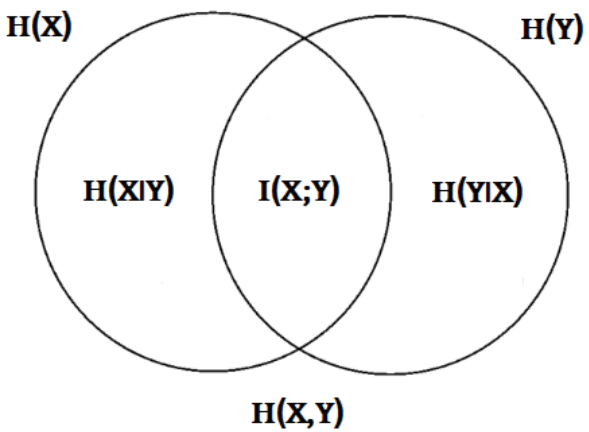

Figure 2 Features of Shannon's entropy function.

Example 1

Let us assume X is the presence or absence of a *G* allele in the CAPN10 gene, which associate with type 2 diabetes mellitus. The *G* and *A* alleles are detected with the probability *p* and (1–*p*), respectively, that is,

$$X = \begin{cases} 1 & \text{with prob } p; \text{ for "G" allele} \\ 0 & \text{with prob } (1-p); \text{ for "A" allele} \end{cases}$$

For various values of *p*, that is, *G* allele frequency, the entropy value for the random variable X is shown in Table 1. In fact, when *p* is closer to 0.5, the uncertainty level over the *G* allele is increased, and thus, the amount of information about the test will be increased. In this example, we obtain an entropy *G* allele with $p = 0.25$.

$$H(X) = -\sum_{x=0}^{1} p^x (1-p)^{1-x} ln\left(p^x (1-p)^{1-x}\right)$$

$$= (0.25^1 \ (1-0.25)^{1-1} \ ln[0.25^1(1–0.25)^{1-1}])+$$

$$(0.25^0)(1-0.25)^{1-0} \ ln[0.25^0(1\text{-}0.25)^{1\text{-}0|}]) = 0.56 \ nat \ .$$

Example 2

In this example, we will show how to calculate the Shannon's entropy information, which is a kind of dependency between variables using discrete expression profile. Now, suppose the discrete expression profile for two genes *A* and *B* is [1, 1, 0,-1, 0] and [1, -1, 0, 1, 1], respectively. The occurrence probability of each mode for the genes is presented in Table 2.

Therefore, the amount of entropy for the gene *A* and gene *B* is:

$$H(A) = -\Sigma_{x=0}^{3} P_x ln(p_x) = -\left[\left(\frac{2}{5}\times ln\left(\frac{2}{5}\right)\right)+\left(\frac{2}{5}\times ln\left(\frac{2}{5}\right)\right)+\left(\frac{1}{5}\times ln\left(\frac{1}{5}\right)\right)\right] = 1.05 \ nat,$$

$$H(B) = -\Sigma_{x=0}^{3} P_x ln(p_x) = -\left[\left(\frac{3}{5}\times ln\left(\frac{3}{5}\right)\right)+\left(\frac{1}{5}\times ln\left(\frac{1}{5}\right)\right)+\left(\frac{1}{5}\times ln\left(\frac{1}{5}\right)\right)\right] = 0.95 \ nat.$$

TABLE 1 **The five various values and entropies of G allele of CAPN10 gene**

p:	0	0.25	0.50	0.75	1
H(*X*):	0	0.56	0.69	0.56	0

TABLE 2 Frequency distribution of the expression profile of A and B genes

Gene:	$P(1)$	$P(0)$	$P(-1)$
"A"	2/5	2/5	1/5
"B"	3/5	1/5	1/5

To calculate $H(A,B)$, the nine possible combinations with respect to the joint probabilities $P(A,B)_s$ should be considered as follows:

$$P(1,1)=\frac{1}{5};\quad P(1,0)=0;\quad P(1,-1)=\frac{1}{5},$$

$$P(0,1)=\frac{1}{5};\quad P(0,0)=\frac{1}{5};\quad P(0,-1)=0,$$

$$P(-1,1)=\frac{1}{5};\quad P(-1,0)=0;\quad P(-1,-1)=0,$$

then $H(A, B) = 1.61$. Finally, the mutual information between the two expression profiles A and B is:

$$I(A, B) = H(A) + H(B) - H(A, B) = 1.05 + 0.95 - 1.61 = 0.39\ nat.$$

Note, high levels of mutual information suggest similarity between two expression profiles.

In addition to the mentioned concepts, one of the important entropy rules for random variables (*iid*) is the *asymptotic equipartition property* (AEP) theorem, which points out that the joint probability of a sequence of random variables, that is, $p(X_1, X_2,\ldots, X_n)$, is very close to $2^{-nH(X)}$.

Let $X_1, X_2,\ldots, X_n$ be a sequence of *iid* random variables with a probability of density function $f(X)$, then:

$$-\frac{1}{n}\log\left(f\left(X_1,X_2,\ldots,X_n\right)\right)\xrightarrow{p}E\left(-\log\left(f(X)\right)\right)=H(X).$$

The above definition leads to the definition of a typical set $A_{\epsilon}^{(n)}$; so that $\epsilon > 0$ and $\forall n$, the usual set $A_{\epsilon}^{(n)}$ is defined as follows:

$$A_{\epsilon}^{(n)}=\left\{\left(x_1,x_2,\ldots x_n\right)\in S^n:\left|-\frac{1}{n}\log\left(f\left(x_1,x_2,\ldots,x_n\right)\right)-H(X)\right|\le\epsilon\right\},$$

where

$$f\left(x_1,x_2,\ldots,x_n\right)=\Pi_{i=1}^{n}f\left(x_i\right).$$

In the context of entropy, we encounter another concept called joint and conditional differential entropy. In other words, the differential entropy for a set of

n random variables $X_1, X_2, \ldots, X_n$ with the density function $f(x_1, x_2, \ldots, x_n)$ is defined as follows:

$$H(X_1, X_2, \ldots, X_n) = -\int f(x_1, x_2, \ldots, x_n)\ log(f(x_1, x_2, \ldots, x_n))\ dx_1\, dx_2 \ldots dx_n.$$

For example, suppose that n random variables $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with mean vector $\mu_{n\times 1}$ and a variance–covariance matrix Σ, then the entropy of a multivariate normal distribution is:

$$H\left(X_1, X_2, \ldots X_n\right) = \frac{1}{2}\log\left(2\pi e\right)^n |\Sigma|,$$

where $|\Sigma|$ is the determinant of variance–covariance matrix Σ.

On the other hand, if X and Y are two random variables (e.g., two genes) with a joint density function $f(X, Y)$, then their conditional differential entropy indicated by $H(X, Y)$ is defined as follows:

$$H(X|Y) = -\int f(x,y)\log(f(x|y))\ dxdy.$$

Since

$$f\left(x|y\right) = \frac{f\left(x,y\right)}{f\left(y\right)}$$

can be written as

$$H(X|Y) = H(X, Y) - H(Y).$$

In choosing a set of random variables (e.g., a set of related genes), we must use two concepts of relative entropy and mutual information, which are referred to next. The relative entropy for continuous random variables X and Y with probability density functions $f(X)$ and $g(Y)$ is equal to:

$$D\left(f||g\right) = E_f\left[\log\left(\frac{f(x)}{g(x)}\right)\right] = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx.$$

Note that relative entropy is always non-negative, that is,

$$D(f||g) \geq 0.$$

The mutual information for the two continuous random variables X and Y with the joint probability density function $f(X, Y)$ is:

$$I\left(X;Y\right) = E\left(log\left(\frac{f\left(X,Y\right)}{f\left(X\right)f\left(Y\right)}\right)\right) = \int f\left(x,y\right)log\left(\frac{f\left(x,y\right)}{f\left(x\right)f\left(y\right)}\right)dxdy.$$

For simplicity, the mutual information can be written as follows:

$$I(X; Y) = H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X),$$

hence:

$$I(X; Y) = D(f(x, y)||f(x)f(y)).$$

Note,

$$I(X; Y) \geq 0,$$

$$H(X|Y) \leq H(X).$$

In $H(X|Y) \leq H(X)$, equality will be achieved if and only if X and Y are independent.

Example 3

Now, let two gene expressions corresponding to A and B genes as two random variables, which they have bivariate normal distribution as follows:

$$\begin{pmatrix} A \\ B \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma\right), \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Suppose that the gene expressions of two genes A and B for three tumor tissues are:

Then, the measures of entropies $H(A)$, $H(B)$ and $H(A,B)$ are:

$$H(A) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(A-\mu_1)2}{2\sigma_1^2}} ln\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(A-\mu_1)^2}{2\sigma_1^2}}\right) dA$$

$$= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{(A-\mu_1)^2}{2\sigma_1^2}} \left(-ln\left(\sqrt{2\pi\sigma_1^2}\right) - \frac{(A-\mu_1)^2}{2\sigma_1^2}\right) dx$$

$$= \ln\left(\sqrt{2\pi\sigma_1^2}\right) + \frac{1}{2\sigma_1^2}\int_{-\infty}^{\infty} \frac{(A-\mu_1)^2}{2\sigma_1^2} e^{-\frac{(A-\mu_1)^2}{2\sigma_1^2}} dx$$

$$= \frac{1}{2} ln\left(2\pi\sigma_1^2\right) + \frac{\sigma_1^2}{2\sigma_1^2} = \frac{1}{2}\left(2\pi e\sigma_1^2\right)$$

$$= \frac{1}{2} + \frac{1}{2} ln(2\pi) + ln\left(\sigma_1^2\right)$$

Note, the above equation shows that the high variance increases the measure of entropy or uncertainty.

Consider the gene expression levels in Table 3 for the two genes and three tissues. Therefore, $\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}=\begin{pmatrix}2.48\\ 2.40\end{pmatrix}$, and

$$\Sigma=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}=\begin{pmatrix}1.66 & (0.98)(1.29)(1.49)\\ (0.98)(1.29)(1.49) & 2.22\end{pmatrix}=\begin{pmatrix}1.66 & 1.88\\ 1.88 & 2.22\end{pmatrix}$$

In this case, $H(A) = 1.93$, $H(B) = 2.22$ *nat*. In addition, $H(A, B)$ calculate as follows:

$$H(A,B)=\frac{1}{2}ln\left[(2\pi e)^2|\Sigma|\right]=1+ln(2\pi)+ln(\sigma_1\sigma_2)+\frac{1}{2}ln(1-\rho^2)=1.88\ nat.$$

Gel'fand and Yaglorn (9) showed that an exact relationship between entropy information, $I(A,B)$, and the correlation coefficient for A and B gene, r is:

$$I(A,B)=H(A)+H(B)-H(A,B)=-\frac{1}{2}ln(1-\rho^2)=1.61.\ nat.$$

The important limitation of entropy information is that its upper limit is unknown, that is, $I(X, Y) \in (0, +\infty)$ Therefore, an index to measure the correlation of two random variables based on entropy information should be introduced, which does not have this limitation. The normalized mutual information, $U(X,Y)$, has such property. The normalized mutual information concept, $U(X,Y)$, is used to choose a set of correlated variables using the uncertainty function, which is shown for two random variables (e.g., two genes) X and Y as follows:

$$U(X,Y)=\frac{2I(X,Y)}{H(X)+H(Y)},$$

where $0 \le U(X,Y) \le 1$. The value $U(X,Y)$ close to zero means that the two random variables X and Y have a high mutual relevance, that is, relation, while the value $U(X,Y)$ close to 1 means that the two random variables have a low mutual

TABLE 3 The gene expression levels for three tumor tissues

	Tissue 1	Tissue 2	Tissue 3
Gene "*A*"	1.12	2.65	3.68
Gene "*B*"	0.98	2.28	3.95

relevance, that is, independence (4, 10). Therefore, $U(A,B)$ with respect to data in example 3 is 0.78 *nat*, which it is a low mutual relevance. For further study on entropy and its properties, we suggest two books: *Handbook of Statistical System of Biology* and Elements of Information Theory (4, 8).

APPLICATION

In this section, we use the results of Bahreini et al. (11) which extracted the information (i.e., gene expression) from the study of Notterman et al. (12). In their study, 18 adenocarcinoma colon and 18 normal tissue samples from the Cooperative Human Tissue Network were evaluated. In that research, the mean (±SD) age of the patients was 67.56 (±14.09) years. Of the total 7465 available cDNAs, only 3,228 genes had fold change more than one and they were selected for analysis. Shannon's entropy method was used to select an appropriate set of genes associated with colon cancer, and 29 genes with the highest amount of information were finally selected.

Before using entropy to select a gene set associated with colon cancer based on gene expressions, the hierarchical method was used for clustering of the genes (Figure 3). The figure shows that 3128 selected genes are shared in three clusters. However, the hierarchical cluster analysis dendrogram shows that the frequency

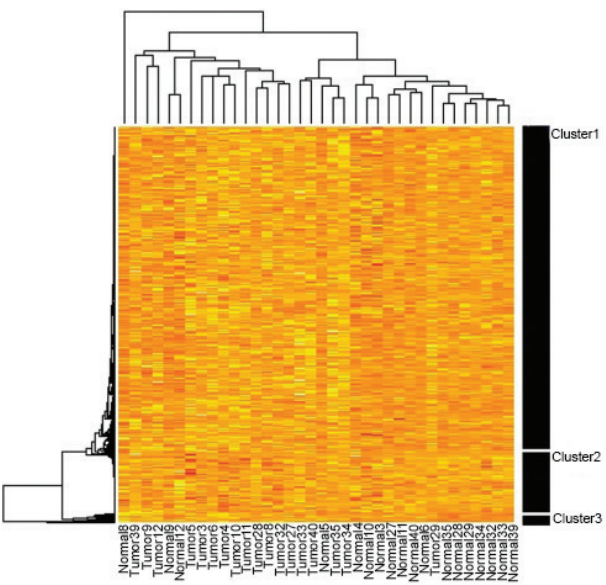

Figure 3 Cluster map derived from a two-way cluster analysis by the hierarchical method. Approximately 3000 common genes in tumor tissues and paired normal tissues were combined in a matrix. Clustering was performed on this matrix. Each color patch on the cluster map indicates the expression intensity level of the associated gene in that tumour and normal tissue samples. The color patches on the cluster map have continuity on expression levels from yellow (highest) to red (lowest) (11).

of the yellow points (i.e., high gene expression) in tumor samples is higher than normal samples, but it is simply not possible to identify a set of most relevant genes with colon cancer recognition. In other words, in Figure 3 there is no specific visible pattern in the color spectrum. Usually, clustering is appropriate when a specific spectrum of colors can be found in normal and tumor samples. Therefore, although the genes are shared into three clusters in Figure 3, the obtained result is not accurate. One of the problems may be the lack of refinement of the levels of gene expression. In studies on gene expression analysis, data refinement process and the removal of outlier values are very important.

Figure 4 demonstrates the importance of refining the data. The data refinement methods are numerous and varying. For example, in analyzing microarray data, the gene expression levels obtained may be very large. In this case, fold change can be used to refine the data. Due to the choice of a suitable cut-off point in the fold-change index, we can omit the outside domain data from the analysis to yield more accurate results. It should be emphasized that we were not able to find a proper and accurate statistical method for choosing the fold-change critical point. In Bahreini et al.'s study (11), 29 genes were selected from 3128 genes after performing Shannon's entropy information to determine a collection of the most relevant genes associated with colon cancer. Usually, for graphical representation of the gene expression levels, a dendrogram plot was used. Figure 4 shows that the 29 selected

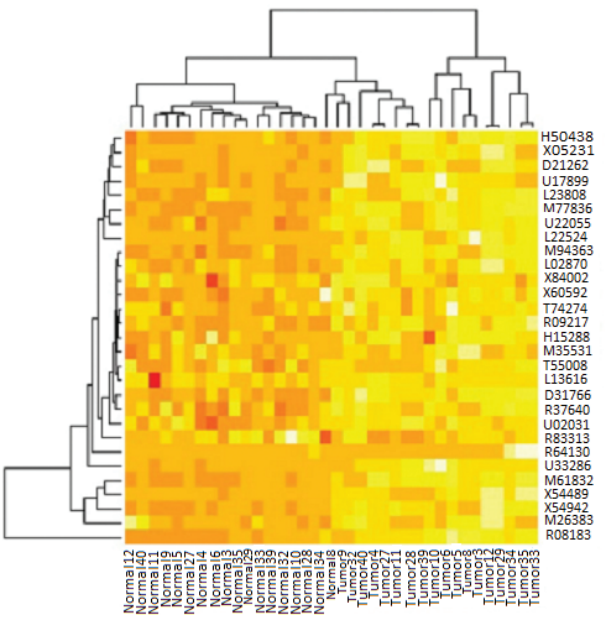

Figure 4 Cluster map derived from two-way cluster analysis with the hierarchical method. We combined 29 common genes in tumor and normal tissues in a matrix. Clustering was performed on this matrix. Each color patch on the cluster map indicates the expression intensity level of the associated gene in that tumor and normal tissue samples. The color patches on the cluster map have continuity on expression levels from yellow, that is, highest, to red, that is, lowest (11).

genes are shared into two clusters by Shannon's entropy method. By comparing two dendrograms (Figures 3 and 4), it can be seen easily that in the second dendrogram (Figure 4), the gene expression in tumor samples is far more than in normal samples, while such a difference was not obvious in the first dendrogram (Figure 3).

CONCLUSION

To reduce dimension in the microarray data and to prevent common errors in statistical modeling, many methods have been introduced, and entropy is one of the most widely used concept in medical and genetic sciences. Entropy was introduced by Nicholas Georgescu-Roegen in 1971 and later developed by scientists based on the principles established by Shannon. Shannon had a major role in introducing entropy information, which has been widely used in high-dimensional studies. One of the advantages of entropy is that calculation of values is based on theoretical forms, not the empirical and personal concepts. These values give small or large weights, proportional to the small or large actual values. Where researchers seek to estimate the risk from an agent, the level of uncertainty is the basis of the computational form of the risk value (Risk = Uncertainty + Damage, where "Damage" in the equation shows the measure of the loss). The function of conventional feature selection algorithms is based more on the choice of the ones that have the most connection with the target class and the least redundancy among the selected features. The major disadvantage of these algorithms is that they ignore the dependencies between the candidate and the unselected feature. However, based on Shannon's entropy information, we can introduce a theoretical algorithm that does not display such disadvantages. Although entropy is often used as a feature of the information concept, it is crucially dependent on the probability model.

Conflict of Interest: The authors declare no potential conflict of interest with respect to research, authorship, and/or publication of this chapter.

REFERENCES

1. Stadtländer CTK-H. Systems biology: Mathematical modeling and model analysis. J Biol Dyn. 2018;12(1):11–15. http://dx.doi.org/10.1080/17513758.2017.1400121
2. Westerhoff HV, Winder C, Messiha H, Simeonidis E, Adamczyk M, Verma M, et al. Systems biology: The elements and principles of life. FEBS Lett. 2009;583:3882–90. http://dx.doi.org/10.1016/j.febslet.2009.11.018
3. Breitling R. What is systems biology? Front Physiol. 2010;1:9. http://dx.doi.org/10.3389/fphys.2010.00009
4. Stumpf MPH, Balding DJ, Girolami M. Handbook of statistical systems biology. 1st ed. Chichester: Johan Wiley & Sons, Ltd., The Atrium; 2011.

5. Adami C. Information theory in molecular biology. Phys Life Rev. 2004;1:3–22. http://dx.doi.org/10.1016/j.plrev.2004.01.002
6. Wang X-Z, Xing H-J, Li Y, Hua Q, Dong C-R, Pedrycz W. A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning. IEEE Trans Fuzzy Syst. 2015;23(5):1638–54. http://dx.doi.org/10.1109/TFUZZ.2014.2371479
7. Shannon CE. A Mathematical theory of communication. Bell Syst Tech J. 1948;27:379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x
8. Cover TM, Thomas JA. Elements of information theory. New York: John Wiley & Sons; 2012.
9. Gel'fand IM, Yaglom AM. Calculation of amount of information about a random function contained in another such function. Am Math Soc Transl. 1957;2(12):199–246. English translation of original in Uspekhi Matematicheskikh Nauk 12(1):3–52.
10. Cover TM, Thomas JA. Elements of information theory. New York: John Wiley & Sons, Inc.; 1991. http://dx.doi.org/10.1002/0471200611
11. Bahreini F, Soltanian AR. Identification of a gene set associated with colorectal cancer in microarray data using the entropy method. Cell J. 2019;20(4):569–75.
12. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res. 2001;61(7):3124–30.