BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 4 2011, pages 556–563
doi:10.1093/bioinformatics/btq704

Systems biology
Advance Access publication December 20, 2010

Statistical analysis of the cancer cell's molecular entropy using high-throughput data

Wessel N. van Wieringen[1,2,*] and Aad W. van der Vaart[2]

[1]Department of Epidemiology and Biostatistics, VU University Medical Center, PO Box 7075, 1007 MB Amsterdam and [2]Department of Mathematics, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Associate Editor: Trey Ideker

ABSTRACT

Motivation: As cancer progresses, DNA copy number aberrations accumulate and the genomic entropy (chromosomal disorganization) increases. For this surge to have any oncogenetic effect, it should (to some extent) be reflected at other molecular levels of the cancer cell, in particular that of the transcriptome. Such a coincidence of cancer progression and the propagation of an entropy increase through the molecular levels of the cancer cell would enhance the understanding of cancer evolution.

Results: A statistical argument reveals that (under some assumptions) an entropy increase in one random variable (DNA copy number) leads to an entropy increase in another (gene expression). Statistical methodology is provided to investigate the relation between the genomic and transcriptomic entropy using high-throughput data. Analyses of multiple high-throughput datasets using this methodology show a close, concordant relation among the genomic and transcriptomic entropy. Hence, as cancer evolves, and the genomic entropy increases, the transcriptomic entropy is also expected to surge.

Contact: wvanwie@few.vu.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 14, 2010; revised on December 2, 2010; accepted on December 15, 2010

1 INTRODUCTION

DNA copy number aberrations are one of the key characteristics of cancer (Stratton *et al.*, 2009). In fact, the accumulation of DNA copy number aberrations is the most consistent feature of cancer progression (Fearon and Vogelstein, 1990). The entropy at the genomic level (chromosomal disorganization) of cancer cells thus exceeds that of healthy cells and tends to surge as cancer progresses. DNA copy number aberrations affect mRNA expression levels as the central dogma of molecular biology suggests and numerous high-throughput studies have shown (e.g. Pollack *et al.*, 2002). Aberrations need not only affect the expression of its driver gene, but may also alter expression levels of the other genes that map to the aberrated segment (Kauraniemi and Kallioniemi, 2006; Van Wieringen *et al.*, 2010). In fact, genomic aberrations also affect expression levels of other transcripts like microRNAs (e.g. Zhang *et al.*, 2006). The close relation between the genome and transcriptome suggests that the entropy increase spreads to other molecular levels of the cell's regulatory system, and is expected to manifest itself most prominently in the transcriptome. Indeed, for this surge in genomic entropy to have any phenotypic (oncogenetic) effect, it needs (to some extent) to propagate to the transcriptomic level and beyond.

Cancer may be considered an evolutionary process, driven by random variation and natural selection (Merlo *et al.*, 2006; Stratton *et al.*, 2009). During its life a cell may undergo heritable genetic alterations (e.g. DNA copy number aberrations). Such alterations may be neutral but may also affect the cell's phenotype. Irrespective of the type of alteration, any cell is subject to natural selection: it has to survive in the environment of the organism's tissue. Within this framework, a cancer cell can be thought of as having acquired alterations that resulted in beneficial traits to survive, proliferate and metastasize (Hanahan and Weinberg, 2000).

Evolution explores different paths via random variation, and the path that leads to a faster entropy increase is naturally selected (Kaila and Annila, 2008). As cancer evolves/progresses, the entropy at the genomic level increases (Castro *et al.*, 2006; Höglund *et al.*, 2005). Here we investigate, using high-throughput studies, whether this is reflected at the transcriptomic level (as suggested by Tsafrir *et al.*, 2006). If so, this may shed light on the path of evolution of the cancer cell.

Before we facilitate the study of the propagation of increased entropy of genome to transcriptome in the cancer cell, we first provide a statistical argument that suggests this indeed seems plausible. We then present statistical methodology to analyze high-throughput genomic experiments in order to answer the following related questions:

- Is the entropy of a cancer sample's transcriptome higher than that of a normal sample?
- Is a cancer sample's genomic entropy associated to that of its transcriptome?

These questions are portrayed schematically in Figure 1. We illustrate how the discussed methodology may be utilized by application to multiple datasets.

2 METHODS

2.1 Motivation

We provide statistical motivations for the hypothesis that an increase of the entropy at the genomic level leads to an increase of the entropy at the

*To whom correspondence should be addressed.

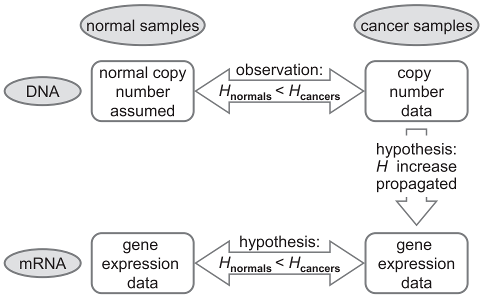
normal samples
cancer samples
DNA
normal copy number assumed
observation: $H_{\text{normals}} < H_{\text{cancers}}$
copy number data
hypothesis: H increase propagated
mRNA
gene expression data
hypothesis: $H_{\text{normals}} < H_{\text{cancers}}$
gene expression data

Fig. 1. Schemata of entropy relations of genome and transcriptome within and between normal and cancer samples.

transcriptomic level. To do so, we need the definition of the (*differential*) entropy of a continuous random variable X with density f_X as (Cover and Thomas, 2006):

$$H(X) := -\int_{-\infty}^{\infty} f_X(x)\log[f_X(x)]dx.$$

For our first motivation, we assume that DNA copy number and gene expression are both measured without error and their relation at any loci may be described by a strict monotone increasing function $g(\cdot)$: $Y=g(X)$. In addition, we note that, according to Theorem 1 of Ramsay (1998), any smooth monotone function $g(x)$ may be represented as:

$$g(x) = c_0+c_1\int_{-\infty}^{x}\exp\left[\int_{-\infty}^{t} w(s)ds\right]dt, \qquad (1)$$

where c_0 and c_1 are constants and w is a coefficient function that is Lebesque square integrable. By construction, w satisfies the second-order differential equation $d^2g(x)/dx^2=wdg(x)/dx$. For $w(s)=0$, $g(\cdot)$ is linear; $w(s)=\alpha$ implies $g(\cdot)$ is an exponential function.

Proposition 1. *Let X_1 and X_2 be random variables with symmetric and zero-centered densities f_{X_1} and f_{X_2}, respectively, and $g(\cdot)$ a strict monotone function generated by Representation (1) using an even coefficient function $w(s)$, i.e. $w(s)=w(-s)$. Then, $H(X_1)\leq H(X_2)$ implies $H(g(X_1))\leq H(g(X_2))$*

To show that $H(g(X_1))\leq H(g(X_2))$, we first note that the entropies of X_1 and $g(X_1)$ are linked via:

$$H(g(X_1)) = H(X_1)+\int_{-\infty}^{\infty} f_{X_1}(x)\log\frac{dg(x)}{dx}dx.$$

It thus suffices to show that:

$$\int_{-\infty}^{\infty} f_{X_1}(x)\log\frac{dg(x)}{dx}dx \leq \int_{-\infty}^{\infty} f_{X_2}(x)\log\frac{dg(x)}{dx}dx.$$

In case $g(\cdot)$ is a linear map, $g(x)=ax$, this is immediate. To prove the desired inequality for other choices of $g(\cdot)$, we note that, appealing to Representation (1), the original inequality follows if we prove:

$$\int_{-\infty}^{\infty} f_{X_1}(x)\int_{-\infty}^{x} w(s)dsdx \leq \int_{-\infty}^{\infty} f_{X_2}(x)\int_{-\infty}^{x} w(s)dsdx. \qquad (2)$$

Integration by parts yields:

$$\int_{-\infty}^{\infty} f_{X_1}(x)\int_{-\infty}^{x} w(s)dsdx = \int_{-\infty}^{\infty} w(s)ds-\int_{-\infty}^{\infty} F_{X_1}(x)w(x)dx.$$

From the assumption that X_1 and X_2 are symmetrically distributed around zero and the symmetry of $w(s)$, $w(s)=w(-s)$, the second term on the right-hand side equals $\int_0^{\infty} w(s)ds$ and equality in Equation (2) follows. Hence, under the assumptions of Proposition 1, we have shown that $H(X_1)\leq H(X_2)$ implies $H(g(X_1))\leq H(g(X_2))$.

If one is not willing to assume that DNA copy number and gene expression are known without error, we provide an alternative motivation. Hereto we introduce the concept of dispersive ordering between two random variables (Shaked and Shanthikumar, 2007). Let X_1 and X_2 be two random variables with distributions F_{X_1} and F_{X_2}, respectively. Let $F_{X_1}^{-1}$ and $F_{X_2}^{-1}$ be the right continuous inverses of F_{X_1} and F_{X_2}, and assume that:

$$F_{X_1}^{-1}(b)-F_{X_1}^{-1}(a)\leq F_{X_2}^{-1}(b)-F_{X_2}^{-1}(a) \qquad \text{for all } 0\leq a\leq b\leq 1.$$

Then X_1 is said to be smaller than X_2 in the *dispersive order*, denoted by $X_1\leq_{\text{disp}} X_2$.

Corollary 1. *Let X_1, X_2 and ε be independent random variables with log-concave densities f_{X_1}, f_{X_2} and f_ε, respectively. Then, if $X_1\leq_{disp} X_2$ and $\beta\geq 0$:*

$$\beta X_1+\varepsilon\leq_{disp}\beta X_2+\varepsilon.$$

Corollary 1 follows directly from Theorems 3.B.4 and 3.B.9 of Shaked and Shanthikumar (2007).

To interpret Corollary 1, let X_1 and X_2 be random variables representing the DNA copy number changes at two different loci and Y_1 and Y_2 the expression levels of two genes that map to these loci. Furthermore, assume that the relation between DNA copy number changes and gene expression at both loci may be described by $Y=\beta X+\varepsilon$ with X and ε independent, $E(Y\,|\,X)=\beta X$, and $\beta\geq 0$ to reflect the empirical observation that a(n) increase/decrease in DNA copy number leads to a(n) increase/decrease in gene expression. Theorem 1 then tells us that if locus 2 is more prone to be aberrated at the genomic level than locus 1 (statistically operationalized as $X_1\leq_{\text{disp}} X_2$), this is reflected in the transcriptome (under the assumption of a simple linear model) and locus 2 will exhibit more abnormal expression levels than locus 1.

Corollary 1 is formulated in terms of the dispersive ordering, whereas our interest is in the entropy ordering. The relevance of Theorem 1 stems from the fact that dispersive ordering implies entropy ordering, i.e. $X_1\leq_{\text{disp}} X_2 \Longrightarrow H(X_1)\leq H(X_2)$ (Oja, 1981). Corollary 1 is illustrated for the entropy ordering by two examples. In the first, let $X_1\sim\mathcal{N}(0,\sigma_{X_1}^2)$, $X_2\sim\mathcal{N}(0,\sigma_{X_2}^2)$, $\varepsilon_1\sim\mathcal{N}(0,\sigma_\varepsilon^2)$ and $\epsilon_2\sim\mathcal{N}(0,\sigma_\varepsilon^2)$, all are independent. Then: $H(\beta X_2+\varepsilon_2)-H(\beta X_1+\varepsilon_1)=\frac{1}{2}\log(\beta^2\sigma_{X_2}^2+\sigma_\varepsilon^2)-\frac{1}{2}\log(\beta^2\sigma_{X_1}^2+\sigma_\varepsilon^2)$. In case $H(X_2)\geq H(X_1)$, and thus $\sigma_{X_2}^2\geq\sigma_{X_1}^2$, this difference is non-negative. In the second example, let $X_1\sim\text{Cauchy}(0,\gamma_{X_1})$, $X_2\sim\text{Cauchy}(0,\gamma_{X_2})$, $\varepsilon_1\sim\text{Cauchy}(0,\gamma_\varepsilon)$ and $\varepsilon_2\sim\text{Cauchy}(0,\gamma_\varepsilon)$, all are independent. Then, using a result from Blyth (1986): $H(\beta X_2+\varepsilon_2)-H(\beta X_1+\varepsilon_1)=\log(\beta\gamma_{X_2}+\gamma_\varepsilon)-\log(\beta\gamma_{X_1}+\gamma_\varepsilon)$. In case $H(X_2)\geq H(X_1)$, and thus $\gamma_{X_2}\geq\gamma_{X_1}$, this difference is non-negative.

2.2 Experiments

Two types of experiments are considered. The first (referred to as experiment of Type I) involves n_C samples from cancerous tissue of a particular same type. For all samples, both a DNA copy number and a gene expression profile are assumed to be available. The random variables X_{ij} and Y_{ij} represent the DNA copy number and the expression level of gene j, $j=1,\ldots,p$, of sample i, $i=1,\ldots,n_C$, respectively. Together the DNA copy number profiles of all samples make up $\mathbf{X}=(X_{ij})_{i=1,\ldots,n_C;j=1,\ldots,p}$, the $n\times p$ genomic aberration matrix, which we assume to contain no missing values. The gene expression matrix $\mathbf{Y}$ is defined similarly, also without missing values. The DNA copy number and gene expressions profiles of sample i are thus the i-th rows of matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively, and will be denoted $\mathbf{X}_i$ and $\mathbf{Y}_i$.

The second type of experiment (Type II) involves $n=n_N+n_C$ samples, of which n_N originate from normal, healthy tissue and n_C from cancerous tissue of the same type. For all samples, only a gene expression profile is available. The gene expression data of experiments of Type II is denoted as above by $\mathbf{Y}=(Y_{ij})_{i=1,\ldots,n;j=1,\ldots,p}$.

2.3 Entropy

In the present context, entropy measures the diversity (at the molecular level) of the samples under study. Here diversity (entropy) at the transcriptomic level is compared between normal and cancer samples. At the genomic level, the entropy of normal samples reaches a minimum (the DNA copy number of its autosomale genome equals two), whereas that of cancer samples can be anything as DNA copy number aberrations may be abound. From Section 2.1 we thus expect that this propagates to the expression levels, resulting in a higher entropy of the cancer transcriptome. To test this, we use the entropy difference between the normal and cancer samples as test statistic, and generate its null distribution by resampling. In order to calculate the test statistic, we first point out how the entropy of a set of samples may be calculated from high-dimensional data ($p > n$), then discuss the test.

The entropy of a multivariate normally distributed random variable, $\mathbf{Y}_i^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is given by $\frac{1}{2}\log[\det(\boldsymbol{\Sigma})] + \frac{1}{2}p[1+\log(2\pi)]$. For low-dimensional situations ($p < n$) the determinant of $\boldsymbol{\Sigma}$, and thus the entropy of $\mathbf{Y}_i$, is straightforwardly obtained by estimating $\boldsymbol{\Sigma}$ by $\mathbf{S} = \frac{1}{n}(\mathbf{Y} - \hat{\boldsymbol{\mu}})^T(\mathbf{Y} - \hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\mu}} = (\frac{1}{n}\sum_{i=1}^n Y_{i1}, \ldots, \frac{1}{n}\sum_{i=1}^n Y_{ip})$, and computing the product of its eigenvalues. However, in case $p > n$ the estimate $\mathbf{S}$ is singular and consequently $\det(\mathbf{S}) = 0$. This problem can be overcome when using a shrunken estimate of $\boldsymbol{\Sigma}$ (Schäfer and Strimmer, 2005). The shrunken estimate of the covariance matrix is given by: $\hat{\boldsymbol{\Sigma}} = (1-\lambda)\mathbf{S} + \lambda\mathbf{T}$, where $\mathbf{S}$ as before, $\mathbf{T}$ a diagonal matrix with $\mathrm{diag}(\mathbf{T}) = \mathrm{diag}(\mathbf{S})$ and the optimal (in some sense) λ equals $[\sum_{i \neq j} \widehat{\mathrm{Var}}(r_{ij})]/[\sum_{i \neq j} r_{ij}^2]$ with $r_{ij} = (\mathbf{S})_{ij}/\sqrt{(\mathbf{S})_{ii}(\mathbf{S})_{jj}}$ (confer Schäfer and Strimmer, 2005). The determinant of the shrunken covariance estimate is given by:

$$\begin{aligned}
\det(\hat{\boldsymbol{\Sigma}}) &= \det(\mathbf{T}^{1/2}\mathbf{T}^{-1/2}\hat{\boldsymbol{\Sigma}}\mathbf{T}^{-1/2}\mathbf{T}^{1/2}) \\
&= \det(\mathbf{T}^{1/2})^2 \det[(1-\lambda)\mathbf{T}^{-1/2}\mathbf{S}\mathbf{T}^{-1/2} + \lambda\mathbf{I}] \\
&= \det(\mathbf{T})\det[(1-\lambda)\tilde{\mathbf{S}} + \lambda\mathbf{I}] \\
&= \prod_{j=1}^{p} s_{jj}^2 \prod_{j=1}^{p} \big[(1-\lambda)\nu_j + \lambda\big],
\end{aligned}$$

where $\nu_p \geq \ldots \geq \nu_{p-n} \geq 0 = \nu_{p-n-1} = \ldots = \nu_1$ are the eigenvalues of $\tilde{\mathbf{S}} = \mathbf{T}^{-1/2}\mathbf{S}\mathbf{T}^{-1/2}$. The non-zero νs coincide with the square of the singular values of $(\mathbf{Y} - \hat{\boldsymbol{\mu}})\mathbf{T}^{-1/2}/\sqrt{n}$. Hence, we are able to estimate the entropy when $p > n$.

To contrast the multivariate normality-based entropy, we also use a non-parametric motivated entropy estimate (Kozachenko and Leonenko, 1987; Leonenko *et al.*, 2008). Hereto we need the k-th nearest neighbor probability density estimate of $f(\cdot)$ given by:

$$\hat{f}_k(\mathbf{Y}_i) = \frac{k}{n-1} \frac{\Gamma(p/2+1)}{\pi^{p/2}} \frac{1}{[d_k(\mathbf{Y}_i)]^p},$$

where $\pi^{p/2}/\Gamma(p/2+1)$ the volume of the unit ball in $\mathbb{R}^p$, and $d_k(\mathbf{Y}_i)$ the Euclidean distance between $\mathbf{Y}_i$ and its k nearest neighbor. The entropy is then estimated by:

$$\hat{H}_k(\mathbf{Y}) = -\frac{1}{n}\sum_{i=1}^{n} \log\Big[\hat{f}_k(\mathbf{Y}_i)\Big], \tag{3}$$

the average entropy of the observations (as determined within the sample).

To test for a difference in entropy, we use the following test statistics: $T = \hat{H}^{(C)} - \hat{H}^{(N)}$, where $\hat{H}^{(C)}$ and $\hat{H}^{(N)}$ are the estimated entropy in cancer and normal sample, respectively. To obtain the null distribution of this test statistic, we permute[1] the sample labels and recalculate the test statistic. This process is repeated L times. The significance level of the tests is now calculated by $\{\#\ell \,|\, T_{obs} \leq T_\ell \text{ for } \ell = 1, \ldots, L\}/L$, the proportion of the null distribution that exceeds the observed test statistic.

2.4 Mutual information

To investigate the genomic–transcriptomic entropy association, we use the concept of mutual information, a general measure of dependence between two random variables. Here, it measures the amount of information shared by genome and transcriptome. Or, loosely, how much knowledge of the genomic entropy tells us about the transcriptomic entropy. Formally, mutual information is defined as (Cover and Thomas, 2006):

$$I(Y;X) := \int\!\!\int f_{(Y,X)}(y,x) \log\left(\frac{f_{(Y,X)}(y,x)}{f_Y(y) f_X(x)}\right) dy\, dx. \tag{4}$$

Mutual information and entropy are related via $I(Y;X) = H(Y) - H(Y|X) = H(Y) + H(X) - H(Y,X)$, where $H(Y|X)$ is the conditional entropy of Y given X and $H(Y,X)$ the joint entropy of Y and X. Hence, by studying the mutual information between $\mathbf{Y}$ and $\mathbf{X}$, we compare the unconditional entropy of the gene expression to its conditional counterpart, conditional on DNA copy number. In case, gene expression is independent of DNA copy number: $H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{Y})$ and $I(\mathbf{Y};\mathbf{X}) = 0$. Mutual information can thus be used to test whether the transcriptomic entropy is associated with the genomic entropy, using resampling to assess whether mutual information deviates significantly from zero. We describe how to estimate $I(\mathbf{Y};\mathbf{X})$ for the distributions considered in Section 2.3.

In case of multivariate normality, $\mathbf{Z} := (\mathbf{X}, \mathbf{Y})^T$:

$$\mathbf{Z} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{YY} \end{pmatrix}\right) =: \mathcal{N}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_{ZZ}),$$

the mutual information in the high-dimensional setting of $p > n$ is estimated by:

$$\begin{aligned}
&\frac{1}{2}\sum_{j=1}^{p} \log\Big[(1-\lambda_Z)\nu_j^Y + \lambda_Z\Big] + \frac{1}{2}\sum_{j=1}^{p} \log\Big[(1-\lambda_Z)\nu_j^X + \lambda_Z\Big] \\
&\qquad\qquad - \frac{1}{2}\sum_{j=1}^{2p} \log\Big[(1-\lambda_Z)\nu_j^Z + \lambda_Z\Big],
\end{aligned}$$

where λ_Z the shrinkage parameter of shrunken estimate of $\boldsymbol{\Sigma}_{ZZ}$, ν_j^Z the eigenvalues of $\mathbf{T}_{ZZ}^{-1/2}\mathbf{S}_{ZZ}\mathbf{T}_{ZZ}^{-1/2}$ with $\mathbf{S}_{ZZ} = \frac{1}{n}(\mathbf{Z} - \hat{\boldsymbol{\mu}}_Z)(\mathbf{Z} - \hat{\boldsymbol{\mu}}_Z)^T$ and $\mathbf{T}_{ZZ}$ a diagonal matrix with $\mathrm{diag}(\mathbf{T}_{ZZ}) = \mathrm{diag}(\mathbf{S}_{ZZ})$, and ν_j^Y and ν_j^X defined similarly.

Again to contrast the multivariate normality assumption, we also estimate the mutual information under the assumption of the k-th nearest neighbor distribution. We follow the approach described in Kraskov *et al.* (2004). The joint entropy, the entropy of $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})^T$, can be estimated using (3). To obtain the mutual information we need to subtract if from the entropies of $\mathbf{X}$ and $\mathbf{Y}$, using the same k. However, this results in a biased mutual information estimate (Kraskov *et al.*, 2004). To resolve this Kraskov *et al.* (2004) propose, in the estimation of the marginal entropies, to fix the nearest neighbor distance instead of k. Let $d_k(\mathbf{Z}_i)$ be the distance (with respect to the uniform norm) of $\mathbf{Z}_i$ to its k-th nearest neighbor, and $k_x(i)$ and $k_y(i)$ the number of samples that fall marginally in the balls $B(\mathbf{X}_i, d_k(\mathbf{Z}_i))$ and $B(\mathbf{Y}_i, d_k(\mathbf{Z}_i))$. Now, following Wang *et al.* (2009), the marginal entropies are estimated, by, e.g.:

$$\begin{aligned}
\tilde{H}_{k_z}(\mathbf{X}) &= -\frac{1}{n}\sum_{i=1}^{n} \psi[k_x(i)] + \log(n-1) \qquad (5) \\
&\quad + \log\left[\frac{\pi^{p/2}}{\Gamma(p/2+1)}\right] + \frac{1}{n}\sum_{i=1}^{n} \log[d_{k_x(i)}(\mathbf{X}_i)]^p],
\end{aligned}$$

where ψ is the digamma function and k_Z refers to the k chosen for $\mathbf{Z}$. The mutual information is then estimated by $\tilde{H}_{k_z}(\mathbf{X}) + \tilde{H}_{k_z}(\mathbf{Y}) - \hat{H}_k(\mathbf{Z})$:

$$\hat{I}(\mathbf{Y};\mathbf{X}) = \psi(k_z) - \frac{1}{n}\sum_{i=1}^{n} [\psi(k_x(i)) + \psi(k_y(i))] + \psi(n), \tag{6}$$

where last term of entropy estimates (5) cancel out as the same nearest neighbor distances are used in the calculation of joint and marginal entropies. Estimator (6) is less biased than a mutual information estimate using the same k for the estimation of $H(\mathbf{X}), H(\mathbf{Y})$ and $H(\mathbf{Z})$ (Kraskov *et al.*, 2004; Wang *et al.*, 2009). Note that the estimator (6) is not scale invariant, and Kraskov

[1] The non-parametric bootstrap seems inappropriate here (and in Section 2.4) as it draws datasets of the same dimensions with identical replicate samples, which inflates the compactness (decreases entropy) of the dataset.

et al. (2004) recommend to scale **X** and **Y** to comparable scales before the estimation of mutual information.

The null distribution of the test statistic is estimated by permuting the columns of the gene expression matrix **Y**, while the DNA copy number matrix **X** is left unchanged (as is done in Van Wieringen and Van de Wiel, 2009). Under the null hypothesis, there is no association between copy number and expression, consequently the permutation only yields the random behavior of the test statistic. For each permutation, the test statistic is recalculated. After L permutations, the P-value is calculated as in Section 2.3.

3 SIMULATION

Here we report a simulation study that serves three ends: (i) to make an informed choice on the nearest neighbor parameter k for entropy and mutual information estimation, (ii) to investigate the behavior of the entropy and mutual information estimators under increased high dimensionality and (iii) to study the relation between the two entropy estimators (similarly for the two mutual information estimators). The simulation study involves artificial data, which facilitates insight on all three ends, as they are sampled from a known distribution with known entropy and mutual information. The setup is described next.

Artificial datasets involve $n = 20$ samples from a p-variate, $p = \alpha n$ with α varying from 0.5 to 50, normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Prior to the generation of each dataset, the covariance matrix $\mathbf{\Sigma}$ is itself randomly drawn from an inverse Wishart distribution $\mathcal{W}^{-1}(d, \mathbf{\Omega})$, with $d = p+2$ (allowing $\mathbf{\Sigma}$ to vary around $\mathbf{\Omega}$) and $\mathbf{\Omega} = \mathbf{V}^T \mathbf{\Lambda} \mathbf{V}$. The $p \times p$ dimensional matrix $\mathbf{V}$ is diagonal with diagonal elements sampled randomly from the empirical distribution of MADs of gene expression data from individual genes of the Chitale dataset. The $p \times p$ dimensional matrix Λ is the correlation matrix estimated from the expression data of the same p genes of the Chitale dataset (see Section 4), with correlations estimated by Kendall's τ, which ensures the positive definiteness of $\mathbf{\Lambda}$. As, for $d = p+2$, $E(\mathbf{\Sigma}) = \mathbf{\Omega}$, the sampled covariance structure mimics that found in real datasets. For each α, multiple datasets are drawn, for which the entropy is estimated (assuming normality and k-NN) and the true entropy using the corresponding drawn $\mathbf{\Sigma}$ is calculated. For the investigation of mutual information datasets with twice the number of genes are drawn, which is randomly split in two equal sized parts. In the estimation of the k-nearest neighbor entropy and mutual information $k = \beta n$, where β ranges from 0.05 to 0.95.

The simulation results are first used to find an optimal β for entropy and mutual information estimation (when assuming k-NN). Hereto the correlation between $\hat{H}_{\text{knn}}$ and H_{true} and between $\hat{I}_{\text{knn}}$ and I_{true} are calculated for all combinations of α and β. Ideally, these correlations are high, and an optimal choice of β results in the highest correlations over the whole range of α. The correlations are displayed as a contour plot in Figure 2. From Figure 2, it is clear that both entropy and mutual information estimation are served best by choosing a small β (irrespective of the high-dimensionality parameter α).

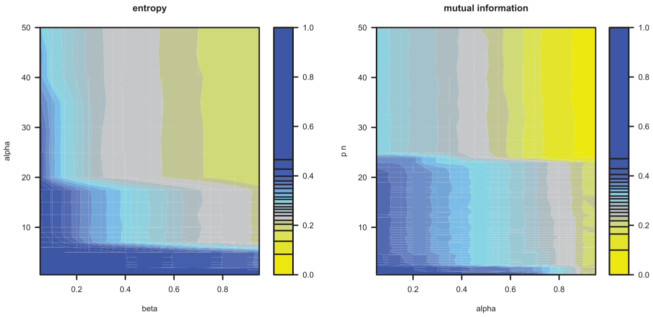

Fig. 2. Contour plots of the correlations between $\hat{H}_{\text{knn}}$ and H_{true} and between $\hat{I}_{\text{knn}}$ and I_{true} in relation to high dimensionality ($\alpha = p/n$) and the number of nearest neighbors ($\beta = k/n$).

To investigate the effect of increased high dimensionality on entropy and mutual information, we again calculate the correlations between $\hat{H}$ and H_{true} and between $\hat{I}$ and I_{true} (assuming both normality and k-NN), which are denoted $\rho^{\text{norm}}_{\hat{H},H}$, $\rho^{\text{knn}}_{\hat{H},H}$, $\rho^{\text{norm}}_{\hat{I},I}$ and $\rho^{\text{norm}}_{\hat{I},I}$. These correlations are plotted against α (plots not shown, but can be deduced from Figure 2 for the k-NN assumption). This reveals that the correlations at first decay rapidly as α is increased, but that this decay levels off to settle at $\rho^{\text{norm}}_{\hat{H},H} \approx 0.30$ and $\rho^{\text{norm}}_{\hat{I},I} \approx 0.40$ after $\alpha = 20$ and $\rho^{\text{knn}}_{\hat{H},H} \approx 0.30$ and $\rho^{\text{knn}}_{\hat{I},I} = 0.30$ after $\alpha = 20$. The decay in itself is not surprising, but the leveling off is encouraging, as the proposed estimators, although noisy, do provide information on the quantities of interest.

Finally, we investigate whether the estimates of the different operationalizations of entropy and mutual information induced by the different distributional assumptions yield concordant results. We now correlate $\hat{H}_{\text{norm}}$ with $\hat{H}_{\text{knn}}$ and $\hat{I}_{\text{norm}}$ with $\hat{I}_{\text{knn}}$. For the simulated data, Spearman's correlation between the entropy estimates is high (varying from 0.80 to 0.95, with highest values assumed for small k) and between the mutual information estimates is substantial (varying from 0.40 to 0.75, with highest values assumed for small k). The concordance of the results from both entropy and mutual information operationalizations indicates that they indeed measure the same quantity. By lack of information on the true entropy and mutual information, which is the case when analyzing real data, corroboration between the two operationalizations enhances the confidence in the results.

4 APPLICATION TO CANCER DATA

The aforementioned methodology is applied to publicly available datasets representing both experiment types described in Section 2.2. The first five columns of Table 1 give an overview of the datasets analyzed. Details on array platforms, preprocessing, etc. are provided in the Supplementary Material. Here we only point out that we have used normalized instead of segmented or called copy number data (refer to Van Wieringen *et al.*, 2007, for details on the differences between these data) in the analysis of Type I experiments, as beforehand the multivariate normality assumption is unlikely to hold for the latter two. See Section 4.4 for a discussion on normalization.

4.1 Analyses of Type II experiments

We analyzed the Type II experiments listed in Table 1 by calculating the entropy difference between cancer and normal groups under the normality and k-th nearest neighbor distributional assumptions discussed in Section 2.3. Permutation tests with $L = 1000$ were used to evaluate the null hypothesis of no entropy difference between the groups. The results are displayed in Table 1. All Type II datasets show a significant ($P < 0.10$ for all analyses, but more often than not

Table 1. Datasets used (more details are provided in the Supplementary Material) and their analysis results

Dataset name	Cancer type	Experiment type	n_N	n_C	Entropy *P*-value (normality)	Entropy *P*-value (knn, k = 1)	Mutual information *P*-value (normality)	Mutual information *P*-value (knn, k = 1)
Chandran	Prostate	II	54	54	0.012	0.026	–	–
D'Errico	Gastric	II	31	38	0.000	0.000	–	–
Kim	Prostate	II	15	32	0.008	0.002	–	–
Landi	Lung	II	49	58	0.000	0.000	–	–
Mougeot	Ovary	II	14	32	0.000	0.000	–	–
Scotto	Cervix	II	25	41	0.070	0.000	–	–
Singh	Prostate	II	50	52	0.000	0.046	–	–
Stirewalt	AML	II	18	26	0.000	0.000	–	–
Bergamaschi	Breast	I	–	85	–	–	0.000	0.000
Carvalho	Colon	I	–	62	–	–	0.000	0.002
Chitale	Lung	I	–	88	–	–	0.000	0.000
Kim	Prostate	I	–	17	–	–	0.664	0.184
Lenz	DLBCL	I	–	203	–	–	0.000	0.106
Oudejans	DLBCL	I	–	42	–	–	0.000	0.002
Pollack	Breast	I	–	41	–	–	0.000	0.020
Zhang	Breast	I	–	263	–	–	0.000	0.000

Note that the Kim dataset is a combination of a Type I and a Type II experiment and therefore appears twice in the table below. Its n_C differs, for the analysis of its Type I version the cancer samples are limited to those that have both DNA copy number and gene expression profiles available.

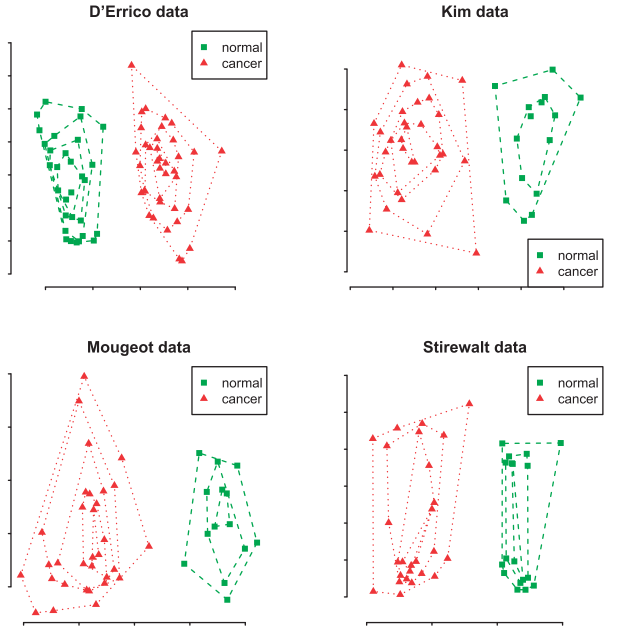

Fig. 3. The dotted and dashed lines result from convex hull peeling per group. For clarity, the groups have been artificially separated by adding a constant to the first principal component of the normal samples.

Table 2. Additional results for the Kim dataset: *P*-values under both distributional assumptions

Cancer stage	Distributional assumption	Cancer stage		
		PIN	Cancer	Metastasis
Normal	Normal	0.468	0.012	0.000
Normal	knn, $k = 1$	0.356	0.004	0.000
PIN	Normal		0.016	0.000
PIN	knn, $k = 1$		0.014	0.000
Cancer	Normal			0.008
Cancer	knn, $k = 1$			0.008

$P < 0.01$) difference in entropy: the cancer groups exhibit a higher entropy than the normal group, under normality and k-NN.

In order to visualize the entropy difference, we have projected the gene expression matrix $\mathbf{Y}$ on its first two right principal components $\nu_1 \mathbf{v}_1$ and $\nu_2 \mathbf{v}_2$, where κ_1 and κ_2 are the largest singular values of $\mathbf{Y}$ and $\mathbf{v}_1$ and $\mathbf{v}_2$ the corresponding right singular vectors. The projected data for four datasets have been plotted in Figure 3. From this figure, it is clear that the cancer samples are much more spread out, indeed indicating a higher entropy.

The Kim and Mougeot datasets are of particular interest as they comprise, next to the analyzed normal and cancer samples, samples from other cancer stages. The former also yields samples from an intermediate and a more advanced stage: 'pin' (13 samples) and 'metastatic' (20 samples), whereas the latter contains two intermediate categories: 'benign' (18 samples) and 'borderline malignant' (3 samples). This allows further investigation of the transcriptomic entropy increase with advanced cancer stage. All possible stage pairs are compared, testing for larger entropy in the more advanced cancer stage. Tables 2 and 3 contain the results. In the Mougeot dataset, not all entropy comparisons are significant, which is (to a large extent) due to the low sample size of the 'borderline malignant' group (consisting of only three samples). Nonetheless, the general picture that emerges from the pairwise analyses in both datasets is that the entropy of a cancer stage exceeds that of preceding stages.

Table 3. Additional results for the Mougeot dataset: P-values under both distributional assumptions

Cancer stage	Distributional assumption	Cancer stage		
		Benign	Borderline malignant	Malignant
Normal	Normal	0.104	0.019	0.000
Normal	knn, $k=1$	0.450	0.048	0.000
Benign	Normal		0.004	0.000
Benign	knn, $k=1$		0.082	0.004
Borderline malignant	Normal			0.446
Borderline malignant	knn, $k=1$			0.340

The entropy surge may be interpreted as an increased diversity among the cancer samples. This may suggest that cancers develop along many different routes, though all leading to unproliferated growth. Consequently, it may prove difficult to develop one therapy that will benefit all individuals with the same cancer. A large (genomic and transcriptomic) diversity increases chances of individuals being and becoming resistant against the therapy, as has been observed in many cases.

4.2 Analyses of Type I experiments

For all Type I experiments listed in Table 1, the mutual information $I(\mathbf{Y};\mathbf{X})$ between gene expression and DNA copy number is calculated. To evaluate whether $I(\mathbf{Y};\mathbf{X})$ deviates significantly from zero, a permutation test as described in Section 2.4 with $L=1000$ permutations is conducted. The results are presented in Table 1. Due to very different (and in these high-dimensional situations hard to verify) distributional assumptions, the P-values may differ between the two presented tests. Nonetheless, in almost all Type I datasets the mutual information deviates significantly from zero, under both distributional assumptions. Hence, the unconditional transcriptomic entropy $H(\mathbf{Y})$ significantly exceeds the conditional transcriptomic entropy $H(\mathbf{Y}|\mathbf{X})$, conditional on the DNA copy number. The Kim dataset, however, shows deviating behavior with clear non-significant P-values. This seems due to the fact that these cancer samples contain relatively few (compared to, e.g. the Bergamaschi dataset) genomic aberrations (as determined by the calling method CGHcall, Van de Wiel *et al.*, 2007). When analyzing the metastatic samples of the Kim dataset, that contain more genomic aberrations, the permutation test for $H_0: I(\mathbf{Y};\mathbf{X})=0$ yields P-values 0.076 (normality) and 0.064 (k-NN). This suggests that also in prostate cancer the genomic entropy surge is propagated to the transcriptome, although perhaps at a later stage in the progression of the disease.

To accompany the above test results, we propose a visualization. The k-NN entropy statistic (3) is composed of the entropies at each observation. Each sample's contribution to the k-th nearest neighbor genomic entropy estimate may then be plotted against its contribution to the k-th nearest neighbor transcriptomic entropy estimate. If indeed the entropies of the two molecular levels are closely related, we expect the 'marginal' entropies at each observation, $\log[\hat{f}_k(\mathbf{Y}_i)]$ and $\log[\hat{f}_k(\mathbf{X}_i)]$, to be positively associated. This is visualized for two datasets in the upper panels of Figure 4. Indeed, as the test results reported in Table 1 suggest, the 'marginal' entropies of the two molecular levels reveal a positive association.

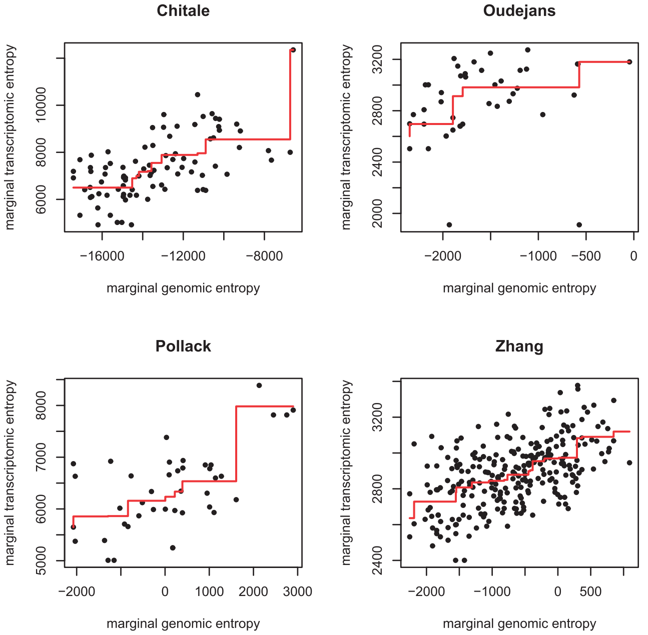

Fig. 4. In all panels, the marginal genomic and transcriptomic entropies are plotted against each other. Each graph contains the isotonic regression fit in red.

The concordant surge in entropy of the cancer cell's molecular levels may be interpreted as follows. Hereto we exploit the reciprocal link between entropy and information (Gatenby and Frieden, 2004). As cancer progresses, the information content of a cancer cell's genome declines until some minimum necessary for the cell to function and proliferate is reached. The above shows that this information decline is reflected at the transcriptomic level, which is a prerequisite for it to have any phenotypic/oncogenic effect. Together these observations suggest that in the evolution (which naturally selects the path of steepest entropy ascent, Kaila and Annila 2008) of the cancer cell, natural selection acts on those parts of the genome and transcriptome that contribute to the cancer cell's minimum information content that enable it to realize its small but focused agenda, making more copies of itself.

4.3 Potential

Besides providing insight into cancer progression, entropy may be used in clinical cancer research. We illustrate this by two examples, which can be the basis for further research. We shall refer to the observation that molecular entropy increases with cancer progression as the Entropy Assumption. In the first example, we aim to reconstruct (unsupervisedly) the order of the samples' cancer advancement using the Entropy Assumption. A simple unsupervised procedure to achieve this would be to start with all samples together, and remove them one by one in accordance with the largest entropy decrease of the expression levels caused by a sample's removal. If the Entropy Assumption has some value, the order of removal is, negatively related to the samples' cancer advancement:

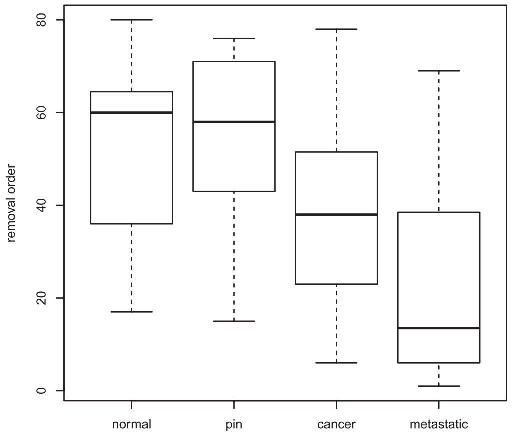

Fig. 5. Order in which samples are removed (by the unsupervised procedure) against cancer stage.

samples removed first cause the highest increase in entropy and are (according to the Entropy Assumption) furthest advanced, whereas samples removed at the end are least advanced. Application of this procedure to the Kim dataset (as used for generating Table 2) allows comparison of the reconstructed order to the sample's known cancer stages (a proxy for cancer progression). In Figure 5, the reconstructed order (removal rank) for the Kim dataset (using k-NN entropy; results are similar under the normality assumption) is plotted against the known cancer stage. Clearly, the metastatic samples tend to be removed earlier than the other samples; the cancer samples tend to be removed earlier than the normal and PIN samples. The samples from the pin stage do not seem to be removed earlier than the normal samples, which is in line with the non-significance of their comparison (Table 2). In addition, the Spearman's rank correlation coefficient between the reconstructed order and cancer stage equals -0.41, also indicating a strong relationship between the two.

In the second example, we interpret entropy increase as increased stochasticity. Higher variability of gene expression levels among cancer samples (compared to normal samples) may be an indication of different regulation. At the level of the pathway (to which presented techniques can be directly applied), such increased randomness in expression patterns may have functional implications. To test this idea, consider the *p53* signalling pathway, which is an ideal test case, as its main function is to guard the integrity of the genome (Weinberg, 2006). The *p53* gene, central to this pathway, is silenced in many cancers, causing loss of function and leading to genomic instability (Weinberg, 2006). This suggests that the *p53* pathway should exhibit an entropy increase as cancer progresses. Again the Kim dataset is used to investigate this. Hereto the (k-NN) entropy of the pathway's (defined by the KEGG repository, ID = 04115, Ogata *et al.*, 1999) expression levels at each cancer stage is estimated: $\hat{H}^{(\text{normal})} = -8.44$, $\hat{H}^{(\text{pin})} = -5.10$, $\hat{H}^{(\text{cancer})} = -2.78$ and $\hat{H}^{(\text{metastatic})} = 5.34$. The *p53* pathway entropy increases monotonely with the cancer's advancement. To assess whether this increase is not due to chance, we assess the significance of this monotonic trend. A simple measure of monotonicity would be the area under the entropy curve (when connecting the entropy of the cancer stages), using $\hat{H}^{(\text{normal})}$ as a base line. Using this as a test statistic and generating the null distribution by permutation of the cancer stage labels, yields a P-value of 0.009 (using 1000 permutations).

4.4 Discussion

Analysis results depend to some extent on the normalization method applied. It may therefore be argued that we should have preprocessed the data with more than one normalization method. In fact, we analyzed two different preprocessed version of the Singh dataset (Type II), one preprocessed by the RMA approach of Irizarry *et al.* (2003) and the other by the MAS approach of Affymetrix. This yielded comparable significant P-values. In addition, we re-analyzed the Chitale dataset (Type I) with segmented (instead of normalized) DNA copy number data, also leading to identical results. For other datasets (when preprocessed differently, we expect similarly consistent analysis results.

Perhaps more convincing, is the fact that, e.g. an entropy difference among normal and cancer is observed over many datasets, involving many different tissue types, generated on various platforms and preprocessed with a diverse array of normalization methods. This despite the fact that normalization aims to remove differences between hybridizations, and is thus likely to obscure (at least partially) the entropy signal. In addition, results are, within tissue type, consistent over datasets (for Type II experiments, the prostate cancer datasets Chandran, Kim and Singh reveal concordant results; for Type I experiments, the results of the breast cancer datasets Bergamaschi, Pollack and Zhang all agree).

In future, the data analyzed here (DNA copy number and gene expression data) will be generated by means of massive parallel sequencing (MPS). Apart from the improved resolution, such data will be less noisy and normalization is expected to be less of a nuisance for entropy comparison.

5 CONCLUSION

We provided a motivating statistical argument which suggests that an increase in genomic entropy is reflected in the transcriptome. Statistical methodology that facilitates the investigation of this hypothesis through analysis of high-throughput DNA copy number and gene expression data is presented. The methodology is illustrated on a multitude of datasets from cancers of various tissues. In addition, the results from analyses of these data suggest that the hypothesis of a related genomic and transcriptomic entropy in cancer has more than only face value.

Conflict of Interest: none declared.

REFERENCES

Blyth,C.R. (1986) Convolutions of Cauchy distributions. *Am. Math. Mon.*, **93**, 645–647.

Castro,M.A.A. *et al.* (2006) Chromosome aberrations in solid tumors have a stochastic nature. *Mutat. Res.*, **600**, 150–164.

Cover,T.M. and Thomas,J.A. (2006) *Elements of Information Theory*, 2nd edn. John Wiley, New York.

Fearon,E.R. and Vogelstein,B. (1990) A genetic model for colorectal tumorigenesis. *Cell*, **61**, 759–767.

Gatenby,R.A. and Frieden,B.R. (2004) Information dynamics in carcinogenesis and tumor growth. *Mutat. Res.*, **568**, 259–273.

Hanahan,D. and Weinberg,R. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Höglund,M. *et al.* (2005) Statistical behavior of complex cancer karyotypes. *Genes Chromosomes Cancer*, **42**, 327–341.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kaila,V.R.I. and Annila,A. (2008) Natural selection for least action. *Proc. R. Soc. A*, **464**, 3055–3070.

Kauraniemi,P. and Kallioniemi,A. (2006) Activation of multiple cancer-associated genes at the ERBB2 amplicon in breast cancer. *Endocr. Relat. Cancer*, **13**, 39–49.

Kozachenko,L.F. and Leonenko,N.N. (1987) On statistical estimation of entropy of a random vector. *Problems Inform. Transm.*, **23**, 95–101.

Kraskov,A. *et al.* (2004) Estimating mutual information. *Phys. Rev. E*, **69**, 066138.

Leonenko,N. *et al.* (2008) Estimation of entropies and divergences via nearest neighbors. *Tatra Mountains Math. Publications*, **39**, 265–273.

Merlo,L.M.F. *et al.* (2006) Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*, **6**, 924–935.

Ogata,H. *et al.* (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.

Oja,H. (1981) On location, scale, skewness and kurtosis of univariate distributions. *Scand. J. Stat.*, **8**, 154–168.

Pollack,J.R. *et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci.*, **99**, 12963–12968.

Ramsay,J.O. (1998) Estimating smooth monotone functions. *J. R. Stat. Soc. Ser. B*, **60**, 365–375.

Schäfer,J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article 32.

Shaked,M. and Shanthikumar,J.G. (2007) *Stochastic Orders and Their Applications*. Springer, New York.

Stratton,M.R. *et al.* (2009) The cancer genome. *Nature*, **458**, 719–724.

Tsafrir,D. *et al.* (2006) Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res.*, **66**, 2129–2137.

Van de Wiel,M.A. *et al.* (2007) CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, **23**, 892–894.

Van Wieringen,W.N. and Van de Wiel,M.A. (2009) Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, **65**, 19–29.

Van Wieringen,W.N. *et al.* (2007) Normalized, segmented or called aCGH data? *Cancer Inform.*, **3**, 331–337.

Van Wieringen,W.N. *et al.* (2010) A random coefficients model for regional co-expression associated with DNA copy number aberrations. *Stat. Appl. Genet. Mol. Biol.*, **9**, Article 25: 1–28.

Wang,Q. *et al.* (2009) Divergence estimation for multidimensional densities via *k*-nearest-neighbor distances. *IEEE Trans. Inform. Theory*, **55**, 2392–2405.

Weinberg,R.A. (2006) *The Biology of Cancer*. Garland Science, New York.

Zhang,L. *et al.* (2006) MicroRNAs exhibit high frequency genomic alterations in human cancer. *Proc. Natl Acad. Sci. USA*, **103**, 9136–9141.