




## Article

# Cancer Segmentation by Entropic Analysis of Ordered Gene Expression Profiles

Ania Mesa-Rodríguez <sup>1,2,\*</sup> , Augusto Gonzalez <sup>1,3</sup> , Ernesto Estevez-Rams <sup>4,\*</sup>  and Pedro A. Valdes-Sosa <sup>1,5</sup> 

<sup>1</sup> The Clinical Hospital of Chengdu Brain Science Institute, University of Electronic Sciences and Technology of China, Chengdu 610054, China

<sup>2</sup> Facultad de Matemática, Universidad de La Habana, San Lazaro y L, La Habana 10400, Cuba

<sup>3</sup> Instituto de Cibernética, Matemática y Física, La Habana 10400, Cuba

<sup>4</sup> Facultad de Física, Instituto de Ciencias y Tecnología de Materiales (IMRE), Universidad de La Habana, San Lazaro y L, La Habana 10400, Cuba

<sup>5</sup> Centro de Neurociencias, BioCubaFarma, La Habana 10400, Cuba

\* Correspondence: ania@neuroinformatics-collaboratory.org (A.M.-R.); estevez@fisica.uh.cu (E.E.-R.); Tel.: +53-5498-3743 (A.M.-R.); +53-5392-9459 (E.E.-R.)

**Abstract:** The availability of massive gene expression data has been challenging in terms of how to cure, process, and extract useful information. Here, we describe the use of entropic measures as discriminating criteria in cancer using the whole data set of gene expression levels. These methods were applied in classifying samples between tumor and normal type for 13 types of tumors with a high success ratio. Using gene expression, ordered by pathways, results in complexity–entropy diagrams. The map allows the clustering of the tumor and normal types samples, with a high success rate for nine of the thirteen, studied cancer types. Further analysis using information distance also shows good discriminating behavior, but, more importantly, allows for discriminating between cancer types. Together, our results allow the classification of tissues without the need to identify relevant genes or impose a particular cancer model. The used procedure can be extended to classification problems beyond the reported results.

**Keywords:** tumor discrimination; gene expression; Shannon entropy; information distance



**Citation:** Mesa-Rodríguez, A.; Gonzalez, A.; Estevez-Rams, E.; Valdes-Sosa, P.A. Cancer Segmentation by Entropic Analysis of Ordered Gene Expression Profiles. *Entropy* **2022**, *24*, 1744. <https://doi.org/10.3390/e24121744>

Academic Editor: Alessandro Giuliani

Received: 22 October 2022

Accepted: 24 November 2022

Published: 29 November 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Entropic magnitudes, as defined in information theory and related areas, have been used in a wide number of areas [1] beyond communication and physics, including literary analysis [2–4], painting [5], and music [6], among others. The advantage of entropic variables is the general framework on which they are founded, allowing them to be used in any system that conveys information in a broad sense. At the same time, it can be used to explore the emergence of patterns as opposed to noise as a random ordering of symbols without, in principle, resorting to a specific model of the data.

DNA is all about information, so there is no surprise in the use of information entropy to characterize the sequence of nucleotides [7–10] despite earlier works considering such analysis limited [11,12]. In this work, entropic magnitudes will be used as a tool for the classification of tissues between normal and tumor samples. Beyond its specific application in this contribution, the procedure presented shows that its generalization to other sequence analyses can be made straightforwardly.

More than two decades of experimental development have resulted in the availability of techniques to simultaneously probe thousands of gene expression levels from a single sample [13–17]. The access to such a huge data set has posed the challenge of meaningful harvesting of the information contained while discriminating, for a given purpose such as cancer diagnosis, the useful from the noise. Data analysis methods have been used over the years to feature-select and classify gene expression data [18–21]. Among others, this

includes the use of statistical methods [22,23], neural networks and machine learning [24,25], and singular value decomposition-based methods [26,27]. In almost all cases, data analysis must effectively reduce the large feature set to identify a subset of distinctive genes that carry enough relevant information for the study at hand [28–30]. Without an underlying model, such reduction is still a challenge not robustly solved.

It is known that the identification of subtle patterns in gene expression can be used as the basis for classifying tissues, differentiating cancerous from non-cancerous samples [31–36]. This has been used not only for exploratory-diagnosis tasks but also to understand what transformation in cancerous cells is relevant and identify the genes responsible for it [15]. Recent studies have looked into the relationship between gene expressions and pathways in both coding and non-coding mutations of genes for a large number of cancer genomes [37]. The relevance of pathways can be essential to explore the cell circuitry and the cause of cancer. It can also prove helpful in designing new experiments with ad hoc perturbed cell conditions to test different hypotheses.

Entropic measures derive from several sources, starting with information theory [38] and including other fields such as Kolmogorov complexity [39], dynamical systems, and ergodic theory, among others. We applied entropic analysis to several samples from the Cancer Genome Atlas (TCGA) database for thirteen cancers. Two main results will be reported: our analysis, using complexity maps and information distance, shows strong discriminating behavior between tumor and normal tissue samples without the need for gene number reduction, avoiding sophisticated feature extraction procedures; the second result is clear evidence that ordering of the expression data along generic reaction pathways proves to be relevant to the cancer characterization in terms of organization and pattern production, a behavior that can not be explained on the sole basis of relative abundance of the different gene expressions.

## 2. Materials and Methods

### 2.1. Gene Expression Level and Gene Expression Coding

Gene expression data and their corresponding normal tissues are taken from the TCGA portal (<https://portal.gdc.cancer.gov/>) (accessed on 10 September 2020). RNA-Seq data in the FPKM format are used. The data contain expressions for 60,483 protein-coding, RNA genes, and pseudogenes. Gene expression levels are coded into three classes with a three-value alphabet  $\chi \in \{-1, 0, 1\}$  as follows. The gene expression level is taken, and a value of 0.1 is added to each level allowing for geometric averaging, which is chosen due to the long tails in the gene expression distribution function. Geometric averaging is performed over the normal tissue set of samples, which allows the definition of a reference value for each gene expression,  $e_r$ . The differential expression is thus defined as  $e/e_r$ , and the base 2 logarithm is used to define a fold variation:  $e_f = \log_2(e/e_r)$ . We take  $e_f < 1$  as downregulated and assign a value of  $-1$  to the gene expression coding;  $e_f > 1$  is taken as upregulated, and the value  $1$  is assigned to the gene coding;  $0$  is assigned otherwise.

### 2.2. Reactome Pathways and Ordered Sequences

For the coding of the biological pathways, the Reactome database was used. A list of pathways and the identified genes participating in them (<https://reactome.org/download/current/Ensembl2Reactome.txt>) (accessed on 5 February 2021) was ordered lexicographically. An ordered sequence of the gene codes was built by parsing the Reactome list and assigning to each entry the corresponding code (See Figure A1 in the Appendix A.1).

### 2.3. Entropic Measures

Consider a bi-infinite sequence  $s$  of values  $s_i$  at position  $i$  ( $-\infty < i < \infty$ ), taken from a discrete alphabet  $\chi$ . The (Shannon) block entropy of length  $L$  of the sequence is defined as  $H(L) = \sum_{s(L) \in \chi^L} p[s(L)] \log p[s(L)]$ , where  $p[s(L)]$  is the probability of finding a finite subsequence  $s(L)$  of length  $L$  in the bi-infinite string  $s$ . There are  $|\chi|^L$  possible finite sequences of length  $L$ . If in the definition of block entropy, the logarithm is taken in base

2, the units are bits. The entropy density is then  $h = \lim_{L \rightarrow \infty} H(L)/L$ ; it measures the entropy per symbol when an infinite number of symbols have been observed, and therefore the irreducible randomness per symbol of the sequence. The entropy density is zero for a constant, a periodic, and a quasi-periodic string, while it attains its maximum value  $\log |\chi|$  for a completely random sequence.

Consider  $h(L) = H(L) - H(L - 1)$ , the effective measure complexity (E) [40], also known as excess entropy [41], can be defined as  $E = \sum_{L=1}^{\infty} [h(L) - h]$ .  $E$  is the mutual information between the two halves of the bi-infinite string [1,42]. It measures the memory or amount of information of one half of the sequence carried to the other half. We will be using plots of  $E$  vs.  $h$ , which is an example of what is called a complexity–entropy map. Complexity–entropy maps show the trade-off between randomness or pattern production as measured by  $h$  and pattern recurrence or memory as measured by  $E$ ; it has been used (under different names) in studies of an extensive set of systems [1,4,43].

Finally, information distance between two sequences comes from Kolmogorov complexity theory [44], and the Kolmogorov complexity  $K(s)$  of a string  $s$  is the length of the shortest program capable of reproducing the string when run in a Universal Turing Machine [39]. A pattern recurrence-driven sequence will have a Kolmogorov randomness that grows slowly (e.g., as  $\log(L)$ ) with the sequence length, such that its normalized value  $K(L)/L \rightarrow 0$  as  $L \rightarrow \infty$ . The same ratio goes to 1 for a random sequence. Information distance between two sequences  $s$  and  $q$  is defined as  $d(s, p) = [K(sq) - \min\{K(s), K(q)\}] / \max\{K(q), K(s)\}$ , where  $sp$  represents the concatenation of both strings. Information distance measures how innovative a string is with respect to the other from an algorithmic perspective: it measures the shortest program length that can transform one string into the other. Information distance has been used to study the relationship between instances of a system in a wide set of areas, for example, to extract phylogenetic evolutionary trees from the genes of mammals [39].

All entropic measures are estimated using the Lempel–Ziv approximation.

#### 2.4. Lempel–Ziv Estimates

When finite data are considered, the entropy rate has to be estimated [42,45] as the asymptotic and limiting values are unreachable. The Lempel–Ziv factorization [46] procedure has been used for the estimation of the entropy density.

Consider the following factorization of a sequence  $s = s_1s_2 \dots s_N$ :

$$F(s) = s(1, p_1)s(p_1 + 1, p_2) \dots s(p_{j-1} + 1, p_j) \dots s(p_{m-1} + 1, N),$$

where  $p_k = 1, 2, \dots$  is a natural number, and  $s(p, q)$  is the substring  $s_p s_{p+1} \dots s_q$ . Each symbol  $s_p$  is drawn from a finite alphabet  $\chi$  of cardinality  $|\chi|$  (e.g.,  $\chi = \{\bar{1}, 0, 1\}$ ,  $|\chi| = 3$ ).

$F(s)$  is called an exhaustive history of the sequence  $s$ , if any factor  $s(p_{j-1} + 1, p_j)$  is not a substring of the string  $s(1, p_j - 1)$ , while  $s(p_{j-1} + 1, p_j - 1)$  is, except perhaps for the last factor.

For example, the sequence  $01\bar{1}01\bar{1}100\bar{1}01010011111\bar{1}\bar{1}\bar{1}00\bar{1}0101111$ , taken from an alphabet  $\chi = \{\bar{1}, 0, 1\}$ , has the following factorization, where a dot separates each factor:

$$0 \cdot 1 \cdot \bar{1} \cdot 01\bar{1}1 \cdot 00 \cdot \bar{1}\bar{1} \cdot 010 \cdot 1001 \cdot 111 \cdot 1\bar{1}\bar{1} \cdot \bar{1}\bar{1}00 \cdot \bar{1}1 \cdot 01011 \cdot 11$$

The Lempel–Ziv complexity  $C(s)$  is then the cardinality (number of factors) of the exhaustive history  $F(s)$  (In the above example,  $C(s) = 14$ ).

In general,  $C(s)$  for a length  $N$  string is bounded by [46]

$$C(s) < \frac{N}{(1 - \epsilon_N) \log_{|\chi|} N} \tag{1}$$

where

$$\epsilon_N = 2 \frac{1 + \log \log (|\chi|N)}{\log N} \tag{2}$$

We used  $\log x \equiv \log_{|\chi|} x$  to simplify the notation.  $\varepsilon_N$  is a slowly decaying function of  $N$ , leading to an asymptotic value

$$C(s) < \frac{N}{\log N}, \quad (3)$$

for large enough  $N$ .

Ziv [47] proved that, if  $s$  is the infinite length output from an ergodic source with entropy rate  $h$ , then

$$\limsup_{N \rightarrow \infty} \frac{C[s(1, N)]}{N / \log N} = h. \quad (4)$$

almost surely. The use of the Lempel–Ziv factorization for estimating the entropy density for finite-size sequences has proved robust even for short-length strings [45]. For a  $10^4$  length sequence, which will be used in this study, the order of magnitude for the error bound is around  $10^{-2}$  [48]. The Lempel–Ziv factorization procedure was implemented in an in-house software (written in C++ and with run time below one minute for each data set) and has been used in previous studies [4,49,50].

The effective complexity measure  $E$  is estimated using a random shuffle procedure given by [51]

$$E(s) = \sum_{M=1}^{M_{max}} [h(s_{(M)}) - h(s)]. \quad (5)$$

$s_{(M)}$  is a surrogate string obtained by partitioning the string  $s$  in non overlapping blocks of length  $M$  and performing a random shuffling of the blocks. The shuffling for a given block length  $M$  destroys all correlations between symbols for lengths larger than  $M$  while keeping the same symbol frequency.  $M_{max}$  is chosen appropriately given the sequence length to avoid fluctuations. In spite of the fact that Equation (5) is not strictly equivalent to the  $E$ , it is expected to behave in a similar manner [51].

As already explained, information distance  $d(s, p)$  comes from the use of algorithmic randomness [52]  $K(s) = |s^*|$  of a string  $s$ , or the length of the shortest algorithm  $s^*$  capable of producing the string  $s$  when run in a Universal Turing Machine.

It is known [39] that the entropy density  $h$  is also given by

$$h(s) = \lim_{|s| \rightarrow \infty} \frac{K(s)}{|s|}. \quad (6)$$

From this result, it follows that [49]:

$$d(s, p) = \frac{h(sp) - \min\{h(s), h(p)\}}{\max\{h(s), h(p)\}}. \quad (7)$$

Again, we estimate the entropy density via Lempel–Ziv factorization and, from there,  $d(s, p)$ .

### 3. Results

Gene expression data for 13 tumors were studied (a table with the TCGA nomenclature for the cancers studied is published as Appendix A Table A1). The data contain expressions for 60,483 protein-coding, RNA genes, and pseudogenes. Gene expression levels are coded into three classes with a three-value alphabet  $\chi \in \{\bar{1}, 0, 1\}$  corresponding to under-, normal-, and over-expressed, respectively. The list of gene expression classes is ordered using a list of pathways and the identified genes participating in them. Such a list can be considered a data sequence.

### 3.1. Complexity–Entropy Maps

For every cancer type, the entropy and  $E$  were estimated from a randomly ordered sequence resulting from the unordered nature of the gene expression data. The corresponding complexity–entropy map was plotted, as shown for the COAD data in Figure 1a. The cancer samples tend to have higher values of entropy density. While segmentation is possible using the entropy density value, no relation was found with the  $E$  values. Upon sorting the gene expression classes with the pathway list, a different picture emerges as  $E$  strongly correlates with the  $h$  value. Figure 1b shows that segmentation is now possible, considering both the  $h$  and the  $E$  value. For the samples of cancer tissue, there is an order of magnitude increase of  $E$  compared to the unsorted strings, which points to the increasing appearance of structuring in the sorted data set.

From now on, all results are referred to the sorted data.

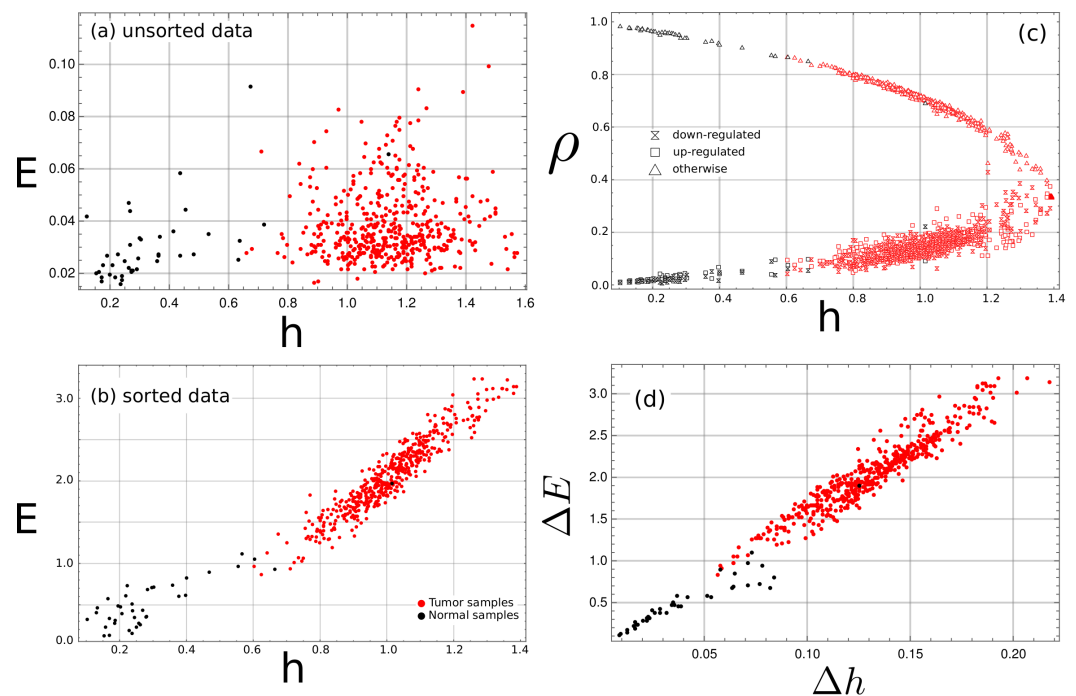
Normal tissue exhibits a mean (median)  $h$  value of 0.304 (0.251) (see Table 1) while, for the tumor sample, the mean (median) value is 1.012 (1.013), a 3.32 (4.04) fold increase. For  $E$ , the increase for the mean (median) is 3.85 (4.43) times from 0.537 (0.460) to 2.067 (2.037). The Euclidean distance in the complexity–entropy map between the mean(median) value of the normal sample and the tumor sample was calculated to be 1.686(1.751). Similar results were found in other cancer types.

**Table 1. Mean and median values of the entropic measures for tumor and cancer tissue samples.**  $h$  stands for the entropy density,  $E$  for the effective measure complexity, and  $d$  is the Euclidean distance between the normal and tumor sample’s mean (median) values. The table includes the values for all cancer types studied.

	Tumor		Mean Normal		$d$	Tumor		Median Normal		$d$
	$h$	$E$	$h$	$E$		$h$	$E$	$h$	$E$	
<b>BLCA</b>	1.073	2.222	0.683	1.192	1.101	1.098	2.260	0.730	1.178	1.143
<b>BRCA</b>	1.017	2.097	0.456	0.785	1.426	1.045	2.169	0.408	0.722	1.581
<b>COAD</b>	1.012	2.067	0.304	0.537	1.686	1.013	2.037	0.251	0.460	1.751
<b>HNSC</b>	0.975	1.891	0.626	1.119	0.847	0.985	1.864	0.583	1.016	0.939
<b>KIRC</b>	0.967	1.754	0.395	0.733	1.170	0.941	1.706	0.371	0.705	1.152
<b>KIRP</b>	0.973	1.772	0.372	0.797	1.145	0.950	1.682	0.384	0.790	1.057
<b>LIHC</b>	1.018	2.060	0.295	0.480	1.738	1.057	2.138	0.249	0.469	1.854
<b>LUAD</b>	0.999	1.964	0.301	0.581	1.549	1.017	1.946	0.259	0.584	1.559
<b>LUSC</b>	1.137	2.382	0.308	0.611	1.955	1.172	2.439	0.291	0.620	2.021
<b>PRAD</b>	0.712	1.481	0.454	0.896	0.639	0.700	1.435	0.394	0.796	0.708
<b>STAD</b>	1.000	1.989	0.757	1.486	0.558	1.031	2.010	0.730	1.449	0.636
<b>THCA</b>	0.710	1.296	0.346	0.780	0.632	0.691	1.246	0.301	0.678	0.689
<b>UCEC</b>	1.138	2.382	0.585	1.002	1.486	1.144	2.378	0.589	1.012	1.474

Structuring may come from two processes: change in the symbols fractions, in this case, the symbols represent one of three classes for up- and down- and normal-regulated genes, so this process refers to the change in the symbol fraction in each class; on the other hand, structuring can be the results of the ordering of the symbols. Structuring by symbol fraction change usually dominates when one of the symbols becomes dominant in the data.

Tumor samples show a larger number of up- and downregulated genes at the other genes’ expense than the normal samples. Figure 1c shows a change in symbols fraction when the normal samples with the tumor samples are compared. However, when we plot the complexity–entropy map of the difference between the unsorted set values and the sorted set values ( $\Delta E$  vs.  $\Delta h$ , where  $\Delta E = E_{sorted} - E_{unsorted}$  and similar for  $\Delta h$ ), as shown in Figure 1d, it becomes evident that the increase in  $E$  is mainly the result of symbol rearrangement, as any change coming from symbol fraction changes is canceled. The same can be said for the increase of  $h$  values.



**Figure 1. Complexity–entropy map for the COAD tumor type tissue samples.** Each point corresponds to a tissue sample. Red points are tumor samples, while black points correspond to normal samples. In (a,b,d), the vertical axis represents the  $E$  values, while the horizontal axis represents the entropy density. Three classes are considered up-, down- and normally expressed genes. In (c), the vertical axis represents the fraction of samples in each class. The entropy map  $E$  vs.  $h$  of the unsorted data (a) fails to show any correlation between the entropy rate  $h$  and the  $E$  values; upon sorting by pathways (b), a clear trend emerges in the map where larger values of  $h$  imply larger values of  $E$ . (c) The fraction  $\rho$  of the number of tissues in a given class, with respect to the total number of samples, shows that tumor samples have a larger number of down- and up-expressed genes compared to the normal samples, but the difference between the unsorted and sorted complexity–entropy map  $\Delta E (= E_{sorted} - E_{unsorted})$  vs.  $\Delta h (= h_{sorted} - h_{unsorted})$  (d) demonstrate that the increase of  $E$  values of the tumor tissue can not be explained by a change in symbol fraction only, but also by pattern formation and ordering.

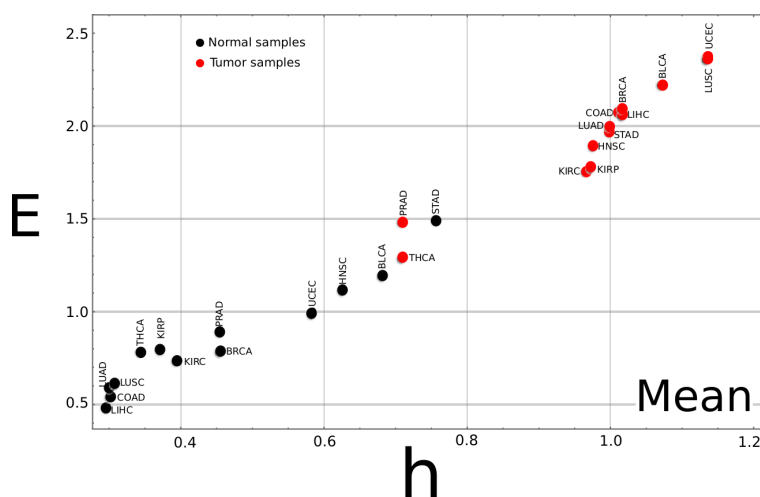
A straightforward discrimination procedure was used based on the complexity–entropy map. The Euclidean distance to the mean (median) values of the normal and tumor points was calculated for every sample. The class to which a sample belongs was decided by taking the one with the shortest Euclidean distance from its mean (median) value to the sample point. With such criteria, the success rate of detecting normal tissue was 0.976 (0.976) and of detecting tumor tissue was 0.979 (0.983).

Similar analyses were performed for the rest of the cancer types, and the complexity–entropy map for each is published as Appendix A. Table 2 summarizes the results. While good discriminating results are robust, above 0.9 for most cancer types, the present procedure had a lower performance for the PRAD, STAD, and THCA cancer types. The worst discriminating fraction was that of the PRAD samples with values around 0.7.

**Table 2. Success fraction using Euclidean distance as discriminating criteria in the complexity–entropy map.** The mean and median values for the normal and tumor points were calculated. A given sample is ascribed to the class whose  $(h, E)$  point has the smaller Euclidean distance to its mean (median) value. Each sample from a given tumor type was used as a test sample to compute the success ratio. The fraction corresponds to the ratio between the number of correctly classified samples and the total number of samples in a given class.

	Mean		Median	
	Cancer	Normal	Cancer	Normal
BLCA	0.917	0.947	0.903	0.947
BRCA	0.912	0.937	0.912	0.937
COAD	0.979	0.976	0.983	0.976
HNSC	0.880	0.750	0.934	0.750
KIRC	0.983	0.958	0.991	0.958
KIRP	0.979	1.000	0.990	1.000
LIHC	0.933	1.000	0.930	1.000
LUAD	0.933	0.983	0.938	0.983
LUSC	0.972	1.000	0.966	1.000
PRAD	0.681	0.731	0.735	0.731
STAD	0.723	0.656	0.723	0.656
THCA	0.872	0.914	0.916	0.879
UCEC	0.987	1.000	0.985	1.000

If the complexity–entropy map for the mean value of all cancer types is plotted (Figure 2), two groups can be identified. One on the upper right of the plot corresponds to the tumor tissue data, exhibiting larger  $h$  and  $E$  values. The second group corresponds to the normal tissues at the lower-left corner with smaller  $h$  and  $E$  values. The poor discriminating performance of the PRAD, STAD, and THCA samples comes from the fact that their mean values for cancer (PRAD, THCA) and normal (STAD) tissue data can not be included in their corresponding groups.

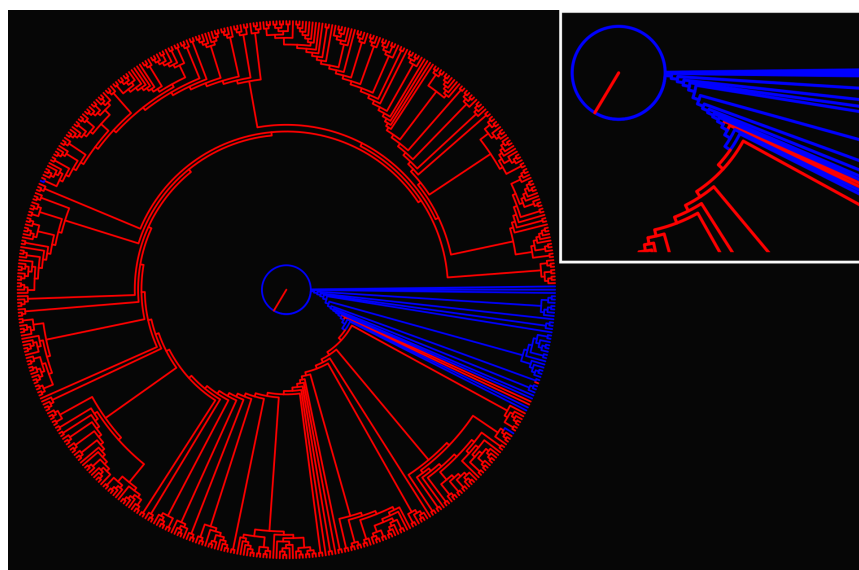


**Figure 2. Complexity–entropy map for the mean values of all cancer types.** The mean value for the two classes, tumor and normal, was calculated for each of the 13 tumor types. As a general trend, the plot shows that tumor samples exhibit larger  $h$  and  $E$  values than normal tissues. It can be noticed that, for the STAD type cancer, the  $(h, E)$  point for the normal sample is unusually high compared to the other samples type, while the points for tumor samples of the PRAD and THCA samples are much lower than the other cancer types. These three cancer types exhibit the worse success ratios (a similar plot for the median value can be found as Appendix A).

### 3.2. Discrimination by Information Distance Measure

The information distance matrix for all samples within a cancer type was computed; we show in Figure 3 the dendrogram corresponding to the COAD tumor type (the dendrogram for all cancer types is published as Appendix A).

From the dendrogram of Figure 3, it is clear that the normal and tumor tissue are clustered in the plot hierarchy. Interestingly enough, all the branches of the cancer tissue show deeper levels of derivation from the root. The result could point to further exploring the structure of the dendrogram in terms of the cancer gene expression in the individual tissues. Discrimination of normal and tumor tissues can be made for the dendrogram alone, yet we used a different approach.

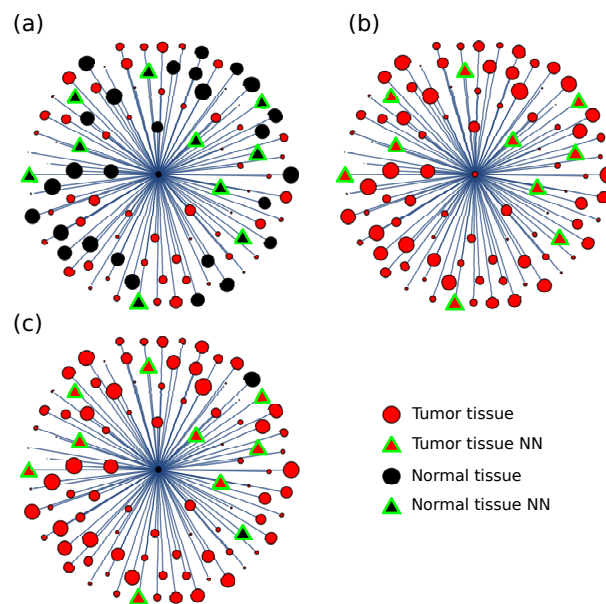


**Figure 3. Dendrogram for COAD type cancer.** Information distance  $d(s, p)$  between all samples for a given cancer type was calculated, resulting in a distance matrix. The dendrogram is built from the distance matrix using the Phylip suit of programs [53]. Blue color corresponds to normal, and red to tumor tissue data. The inset shows the details at the root of the dendrogram.

For the distance matrix, a discriminating procedure was designed based on a majority rule. The information distance to the other tissue samples was considered for each sample, and the nearest neighbor type was used as a majority discriminating criterion. The sample was assigned to the type, tumor, or normal, based on the majority of neighbor types (Figure 4). For each cancer type, the number of neighbors to consider was taken between 3 and 8 and optimized for performance.

Consider a neighborhood formed by the eight nearest samples. For the COAD type cancer, if we assigned a value of  $-1$  for each cancer neighbor and a value of  $1$  for each normal neighbor, the average number of neighbors for a random sample is  $-6.94$ , showing that there are many more cancer samples than normal samples. However, the average neighbor number for the normal sample is  $5.64$ , pointing to the fact that normal samples surround normal samples; and the average number for the cancer sample is  $-7.98$ , showing that cancer samples surround cancer samples.

A second discriminating procedure was also used, where the distance weights each neighbor, and no significant difference was found.



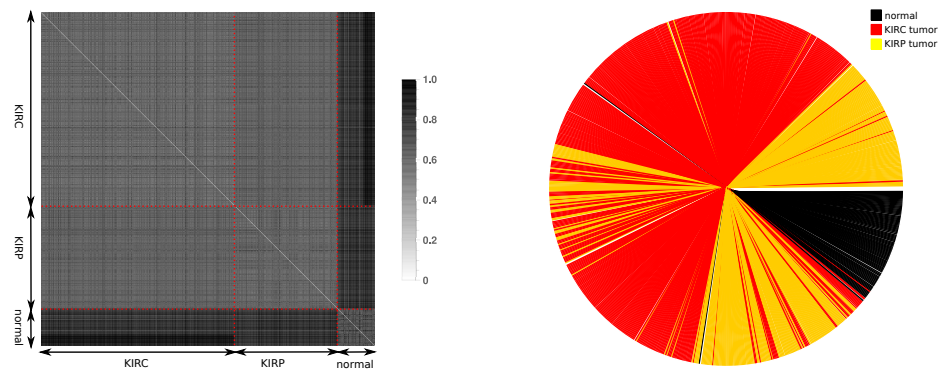
**Figure 4. Information distance discriminating criteria.** Given a sample (at the center), the distance  $d(s, p)$  to any other sample is calculated, and the nearest neighbors (NN) are considered for discriminating purposes. The sample is classified in the (a) normal or (b) tumor class, depending on the class with majority NN (distances in figure not at scale). For the COAD type cancer, in seldom cases, the used criteria (c) failed to ascribe the correct class to the sample.

The success fraction for all cancer types is reported in Table 3. While the success rate for the normal tissue is lower in the PRAD case, compared to the one using Euclidean distance in the complexity–entropy map, the success rate for the cancer tissue is much higher, up to 0.994, similar values to those obtained for the higher ranking success rates in Table 2. The smaller success fraction for the cancer tissue was as high as 0.973 for the KIRP sample. Care must be taken, though, with such a high success fraction, as the number of the normal tissue samples and tumor samples is biased towards the tumor sample, which results, statistically, in a larger number of tumor neighbors for the random case. The same comment can be made on the low success rate for the normal tissue, as, statistically, there are fewer normal neighbors than tumor type. Improvement should be expected when more normal tissues are included in the sample set.

**Table 3. Success fraction using the information distance as discriminating criteria.** From the distance matrix, the nearest neighbors for a given sample are taken, and a majority rule is used to classify the sample as a tumor or normal type. The fraction corresponds to the ratio between the number of correctly classified samples and the total number of samples in a given class.

	Cancer	Normal
BLCA	0.997	0.210
BRCA	0.995	0.821
COAD	1.000	0.872
HNSC	0.996	0.454
KIRC	0.998	0.972
KIRP	1.000	0.969
LIHC	0.973	0.760
LUAD	0.994	0.847
LUSC	0.998	0.959
PRAD	0.994	0.404
STAD	0.987	0.719
THCA	0.986	0.845
UCEC	0.996	0.609

Finally, it was studied whether using the distance matrix allows for discriminating between types of cancer. The KIRC and KIRP kidney cancer were chosen for two reasons. On the one hand, they are two types of tumors from the same organ; therefore, discrimination from the genetic data extracted from the sample tissue could make clinical sense. On the other hand, both mean value points are too close to consider discrimination from such analysis in the complexity–entropy map. Expression reference values were calculated using the set of normal tissue values for both the KIRC and KIRP datasets. The expression classes were determined as already explained using these common reference values. The distance matrix was calculated between all samples from both tumors, KIRP and KIRC, and the normal tissues. The distance matrix and dendrogram are shown in Figure 5, where each cancer type and the normal sample are grouped in mostly separate clusters. Using the discrimination procedure described above, 0.998 of cancer tissues were identified as a tumor, and from there, 0.939 were correctly identified for KIRC type tumors and 0.913 as KIRP type cancer. In addition, 0.981 of normal tissues were correctly identified as such. The same analysis was carried out for the LUSC and LUAD cancer types. The success fraction for cancer tissue was 0.994. The 0.839 and 0.984 fractions of LUAD and LUSC cancer were correctly discriminated, respectively. The dendrogram is shown in Appendix A.5.



**Figure 5.** (left) Distance matrix and (right) information distance dendrogram for the joint KIRC and KIRP tumor types together with the normal tissues. For the fold variation of each gene, the  $e_{ref}$  values were calculated from the joint set of normal tissue samples from both the KIRC and KIRP datasets.

#### 4. Discussion

This work reports the application of entropic measures that proved useful in analyzing an extensive gene expression data set. The measures seem to allow, at least for the scope of this study, the segmentation of tumor tissue samples from normal tissue samples for various cancer types. The sorting of the data along pathways enhanced the segmentation abilities of the applied techniques. One may ask if this points to features beyond cancer discrimination. Gene expression data studies from pathway analysis have proven that, even in the case where no significant variation of a single gene expression level is involved, studying groups of genes by their functional role identifies the difference between samples of healthy subjects and subjects with some types of diseases or abnormal condition, such as diabetes and smoking epithelial tissues [27,54]. Such studies emphasize that the analysis of gene expression levels alone may be insufficient to recognize or characterize a disease condition. In the hallmark paradigm of cancer, [55,56], changes in gene expression in tumor cells results in the reprogramming of the cell circuitry to sustain the so-called hallmarks of cancer. The intracellular integrated circuitry is divided into distinct subcircuitry, each of which performs some specialized function in normal cells, which is modified when changed to cancer cells. In this picture, the effect of varying gene expression levels is relevant as it modifies specific pathways supported by the cell circuitry. Our studies emphasize such a

picture, as the pathway analysis of gene expression profiles, via entropic measures can be interpreted as grasping some of the relevant features of this circuitry modification.

Entropic measures have been used to study the pattern gain at different levels of grammar hierarchy in written language [4]. In our case, the ordered sequence has two possible organization levels, one within the pathways and the other from the pathways position in the list. In both levels, there is not a priori “natural” ordering, and lexicographic order, as the one used, is, in principle, as good as any other ordering scheme. However, the ordered data immediately show the emergence of a clear trend between entropy density and  $E$ . The simultaneous increase of entropy density and  $E$  is a fingerprint of increased complexity. As measured by  $h$ , the given disorder can still accommodate enough pattern recurrence measured by  $E$ . In this sense, the analysis of the tumor samples shows a more complex pathway-ordered gene expression profile. This complexity does not arise mainly by changes in the fraction of samples in each symbol class, although it is clear that up- and downregulated genes increase at the expense of the non-regulated ones. The role of gene expression levels in connection with cancer’s appearance is usually discussed, but what is pointed out here is that this does not describe the whole story. There is no real physically meaningful order for the pathways, so randomization breaks any pattern at the intra-pathway ordering level, leading immediately to the disappearance of any relation between pattern production and pattern recurrence as  $E$  drops an order of magnitude.

## 5. Conclusions

We have proven that, in several cases, and within the scope of the present study, entropy measures allow the discrimination of cancer samples from normal samples by a class analysis of gene expression levels. In both complexity–entropy maps and information distance, classification was possible with a high success rate within a cancer type, above 98% for several cancer types. Furthermore, information distance allows for better discrimination when more than one cancer type is involved. The studies in two types of kidney tumors allowed the correct identification of the tumor type with a very high success rate. The study also suggests that increasing the number of samples, as seen in the joint analysis of KIRC and KIRP data, enhances the classification procedure’s robustness. The emergence of a clear trend in the complexity–entropy map upon sorting the gene expression level through an ordered pathways list allowed us to identify that, as crucial as upregulated and downregulated genes are, the whole context needs to look into the pathways that are activated or turned off in cancer samples.

Further experimental work is needed to evaluate if our findings are relevant in practical tumor discrimination, and current work is planned along this line. In addition, the reported results point to expanding the method to explore if it can be used to study different stages in cancer development and the occurring changes along its evolution.

The developed procedure can be extended to other data types as its generalization is straightforward.

**Author Contributions:** A.G. conceived the idea; E.E.-R. prepared the original draft; and all authors reviewed and edited the draft; A.M.-R. and E.E.-R. contributed with implementing calculations; A.G. contributed with data curing; A.G. and P.A.V.-S. contributed to making sense of the results within the context of cancer research. A.G. and E.E.-R. co-directed the project. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partially funded by the National Nature and Science Foundation of China (NSFC) Grant No. 61871105 and the CNS Program of the University of Electronic Science and Technology of China Grant No. Y0301902610100201.

**Data Availability Statement:** Gene expression data were extracted from the TCGA portal 2020[2020-09-10](<https://portal.gdc.cancer.gov/>). Processed data can be downloaded at Science Data Bank, 2022[2022-11-18]. <https://cstr.cn/31253.11.sciencedb.06470.CSTR:31253.11.sciencedb.06470>. An in-house program was used for Lempel–Ziv estimation. E.E.-R. can be contacted for availability. Plots were made using Wolfram Mathematica (R) framework.

**Acknowledgments:** We thank Rolando Pérez Rodríguez for the fruitful discussion and for pointing us to interesting connections with current cancer models. A.G. acknowledges the Cuban Program for Basic Sciences, the Office of External Activities of the Abdus Salam Center for Theoretical Physics. E.E-R. acknowledges the continued support of the AvH foundation and the support from the Max Plank Institute of Physics of Complex System for a visiting grant. The research is carried out under the Platform for Bioinformatics project of BioCubaFarma, Cuba.

**Conflicts of Interest:** The authors declare no conflict of interest.

### Abbreviations

The following abbreviations are used in this manuscript:

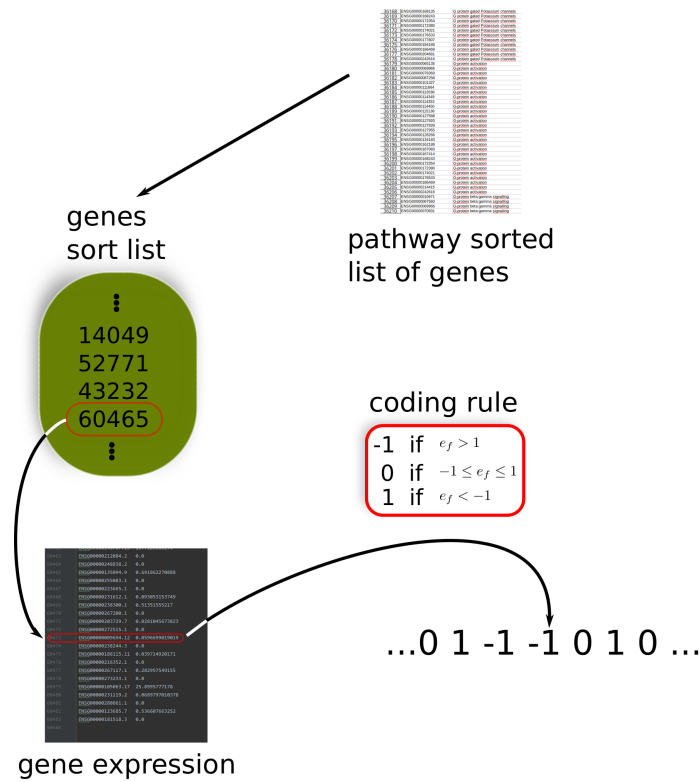
TCGA	The Cancer Genome Atlas
RNA	Ribonucleic Acid
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
H	Shannon Block Entropy
h	Entropy Density
E	Effective Measure Complexity
E	Effective Measure Complexity
NN	Nearest Neighbors
K	Kolmogorov Complexity

### Appendix A

#### Appendix A.1. Cancer Coding and Number of Samples

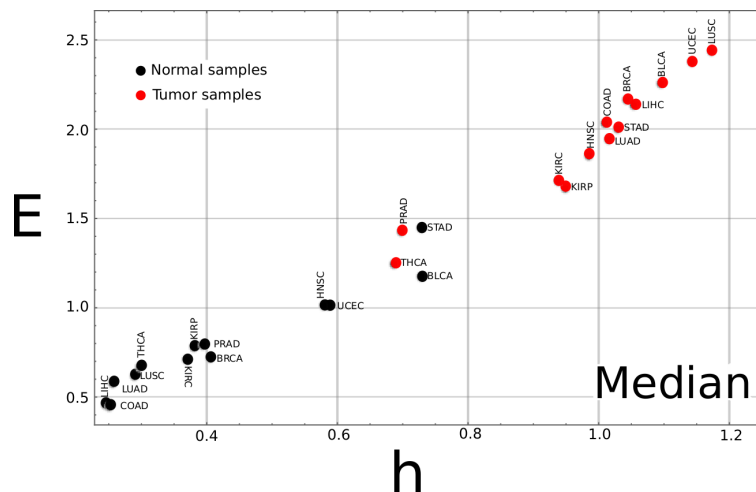
**Table A1.** Studied cancer types and number of samples in the dataset. Data were taken from the publicly available Cancer Genome Atlas.

TCGA	Tumor	Number of Samples	
		Normal	Tumor
BLCA	Bladder Urothelial Carcinoma	19	421
BRCA	Breast Invasive Carcinoma	112	1096
COAD	Colon Adenocarcinoma	41	474
HNSC	Head and Neck Squamous Cell Carcinoma	44	502
KIRC	Kidney Renal Clear Cell Carcinoma	72	539
KIRP	Kidney Renal Papillary Cell Carcinoma	32	289
LIHC	Liver Hepatocellular Carcinoma	50	374
LUAD	Lung Adenocarcinoma	59	535
LUSC	Lung Squamous Cell Carcinoma	49	502
PRAD	Prostate Adenocarcinoma	52	499
STAD	Stomach Adenocarcinoma	32	375
THCA	Thyroid Carcinoma	58	510
UCEC	Uterine Corpus Endometrial Carcinoma	23	552



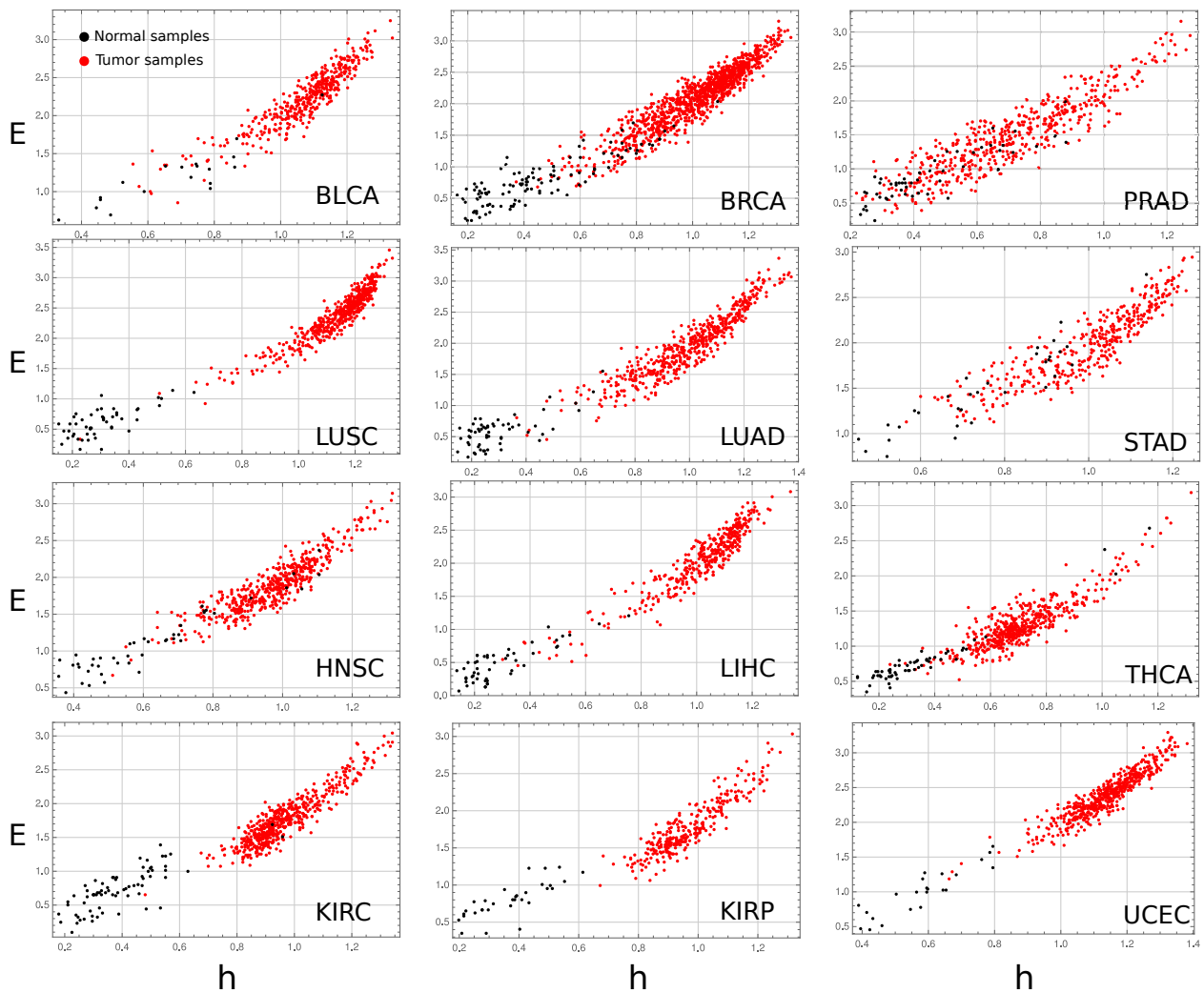
**Figure A1.** Coding and sorting the gene expression levels. (1) The list of genes and pathways they are involved in are lexicographically sorted with respect to pathways from which a gene list is extracted. (2) The list of sorted genes is used to code the expression levels in a sorted manner. The final string is a three-letter alphabet  $\chi = \{-1, 0, 1\}$  coded sequence, where each entry corresponds to the gene expression class in the original pathway sorted list. It should be noted that, in the string coding, a gene repeats itself as much as it appears in the pathway list.

Appendix A.2. Complexity–Entropy Map for the Median Values of All Studied Cancer Types



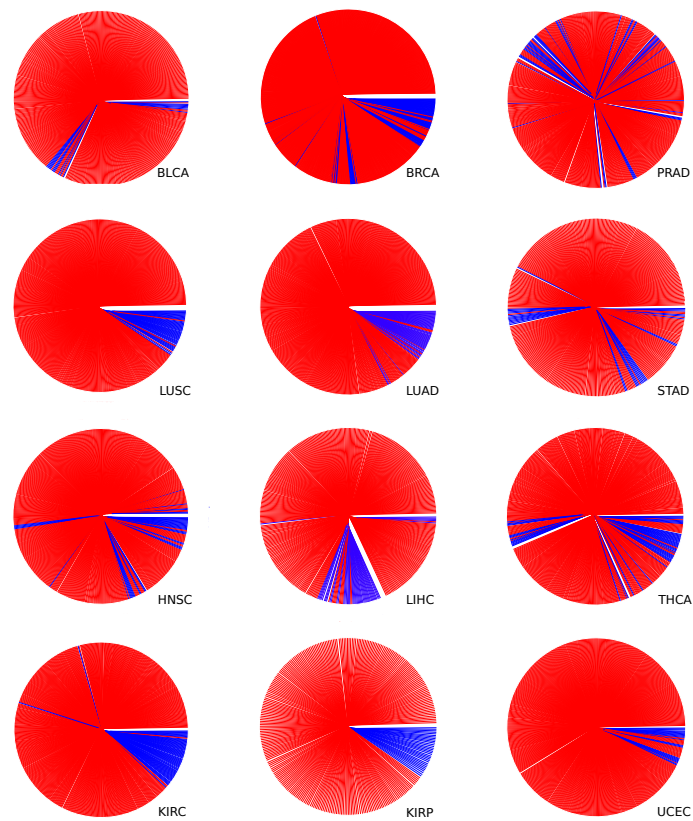
**Figure A2.** Complexity–entropy map for the median values of all studied cancer types. The median value for the two classes, tumor and normal, was calculated for each of the 13 tumor types. As a general trend, the plot shows that tumor samples exhibit larger  $h$  and  $E$  values than normal tissue. Observe that, for the STAD and BLCA type cancer, the  $(h, E)$  point for the normal sample is unusually high compared to the other  $\chi$  samples type, while the points for tumor samples of the PRAD and THCA samples are much lower than the other cancer types.

Appendix A.3. Complexity–Entropy map for All Studied Cancer Types



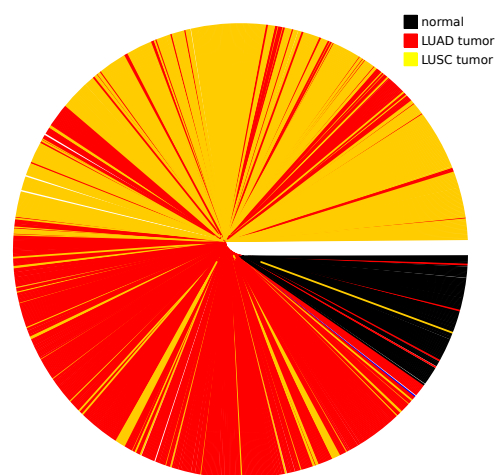
**Figure A3.** Complexity–entropy map for all studied cancer types except the COAD, shown in the article’s main text.  $E$  stands for effective measure complexity, and  $h$  for the entropy density.

#### Appendix A.4. Dendrogram for All Studied Cancer Types



**Figure A4.** Dendrogram for all studied cancer types except the COAD, shown in the article's main text. The blue color corresponds to the normal tissue samples and the red to the tumor tissue samples.

#### Appendix A.5. Dendrogram for LUAD and LUSC Cancer Types



**Figure A5.** Information distance dendrogram for the joint LUAD and LUSC tumor types together with the normal tissues. For the fold variation of each gene, the  $e_{ref}$  values were calculated from the joint set of normal tissue samples from both the LUAD and LUSC datasets.

## References

1. Crutchfield, J.P. Between order and chaos. *Nat. Phys.* **2012**, *8*, 17–24. [[CrossRef](#)]
2. Montemuro, M.A.; Zanette, D. Towards the quantification of the semantic information encoded in written language. *Adv. Complex. Syst.* **2010**, *13*, 135–153. [[CrossRef](#)]
3. Amancio, D.R.; Atmann, E.G.; Rybski, D.; Oliveira, O.N.; da Costa, F.L. Probing the statistical properties of unknown texts: Application to the Voynich manuscripts. *PLoS ONE* **2013**, *8*, e67310. [[CrossRef](#)] [[PubMed](#)]
4. Estevez-Rams, E.; Mesa-Rodriguez, A.; Estevez-Moya, D. Complexity–entropy analysis at different levels of organisation in written language. *PLoS ONE* **2019**, *14*, e0214863. [[CrossRef](#)]
5. Sigaki, H.Y.D.; Perc, M.; Ribeiro, H.V. History of art painting through the lens of entropy and complexity. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, e8585–e8594. [[CrossRef](#)]
6. Daikoku, T. Neurophysiological markers of statistical learning in music and language: Hierarchy, entropy, and uncertainty. *Brain Sci.* **2018**, *8*, 114. [[CrossRef](#)]
7. Farach, M.; Noordewier, M.; Savari, S.; Shepp, L.; Wyner, A. On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, USA, 14 July 1995; pp. 48–57.
8. Schmitt, A.O.; Herzel, H. Estimating the Entropy of DNA Sequences. *J. Theor. Biol.* **1997**, *188*, 369–377. [[CrossRef](#)]
9. Weiss, O.; Jiménez-Montaño, M.A.; Herzel, H. Information content of protein sequences. *J. Theor. Biol.* **2000**, *206*, 379–386. [[CrossRef](#)]
10. Vinga, S.; Almeida, J.S. Rényi continuous entropy of DNA sequences. *J. Theor. Biol.* **2004**, *231*, 377–388. [[CrossRef](#)] [[PubMed](#)]
11. Hariri, A.; Weber, B.; Olmstead, J. On the validity of Shannon-information calculations for molecular biological sequence. *J. Theor. Biol.* **1990**, *147*, 235–254. [[CrossRef](#)] [[PubMed](#)]
12. Shekin, P.S.; Erman, B.; Mastrandea, L.D. Information-theoretical entropy as a measure of sequence variability. *Proteins* **1991**, *11*, 297. [[CrossRef](#)] [[PubMed](#)]
13. Lipshutz, R.J.; Morris, D.; Chee, M.; Hubbell, E.; Kozal, M.J.; Shah, N.; Fodor, S.P. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* **1995**, *19*, 442–447. [[PubMed](#)]
14. Schena, M.; Shalon, D.; Davis, R.W.; Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**, *270*, 467–470. [[CrossRef](#)] [[PubMed](#)]
15. Lockhart, D.J.; Winzler, E.A. Genomics, gene expression and DNA arrays. *Nature* **2000**, *405*, 827–836. [[CrossRef](#)] [[PubMed](#)]
16. Mortazavi, A.; Williams, B.A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **2008**, *5*, 621–628. [[CrossRef](#)] [[PubMed](#)]
17. Lowe, R.; Shirley, N.; Bleackley, M.; Dolan, S.; Shafee, T. Transcriptomics technologies. *PLoS Comput. Biol.* **2017**, *13*, e1005457. [[CrossRef](#)]
18. Sherlock, G. Analysis of large-scale gene expression data. *Curr. Opin. Immunol.* **2000**, *12*, 201–205. [[CrossRef](#)]
19. Jiang, D.; Tang, C.; Zhang, A. Cluster analysis for gene expression data: A survey. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 1370–1386. [[CrossRef](#)]
20. Madeira, S.C.; Oliveira, A.L. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2004**, *1*, 24–45. [[CrossRef](#)]
21. Almgren, N.; Alshamlan, H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access* **2019**, *7*, 78533–78548. [[CrossRef](#)]
22. Dudoit, S.; Yang, Y.H.; Callow, M.J.; Speed, T.P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.* **2002**, *12*, 111–139.
23. Mar, J.C. The rise of the distributions: Why non-normality is important for understanding the transcriptome and beyond. *Biophys. Rev.* **2019**, *11*, 89–94. [[CrossRef](#)] [[PubMed](#)]
24. Daoud, M.; Mayo, M. A survey of neural network-based cancer prediction models from microarray data. *Artif. Intell. Med.* **2019**, *97*, 204–214. [[CrossRef](#)]
25. Danaee, P.; Ghaeini, R.; Hendrix, D.A. A deep learning approach for cancer detection and relevant gene identification. In Proceedings of the Pacific Symposium on Biocomputing 2017, Kohala Coast, HI, USA, 4–8 January 2017; pp. 219–229.
26. Alter, O.; Brown, P.O.; Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 10101–10106. [[CrossRef](#)]
27. Tomfohr, J.; Lu, J.; Kepler, T.B. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinform.* **2005**, *6*, 1–11. [[CrossRef](#)]
28. Aziz, R.; Verma, C.K.; Srivastava, N. A novel approach for dimension reduction of microarray. *Comput. Biol. Chem.* **2017**, *71*, 161–169. [[CrossRef](#)]
29. Li, Z.; Xie, W.; Liu, T. Efficient feature selection and classification for microarray data. *PLoS ONE* **2018**, *13*, e0202167. [[CrossRef](#)]
30. Cilia, N.D.; De Stefano, C.; Fontanella, F.; Raimondo, S.; Scotto di Freca, A. An experimental comparison of feature-selection and classification methods for microarray datasets. *Information* **2019**, *10*, 109. [[CrossRef](#)]
31. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6745–6750. [[CrossRef](#)]

32. Alizadeh, A.A.; Eisen, M.B.; Davis, R.E.; Ma, C.; Lossos, I.S.; Rosenwald, A.; Staudt, L.M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **2000**, *403*, 503–511. [[CrossRef](#)]
33. Bhattacharjee, A.; Richards, W.G.; Staunton, J.; Li, C.; Monti, S.; Vasa, P.; Meyerson, M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 13790–13795. [[CrossRef](#)]
34. Dudoit, S.; Fridlyand, J.; Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **2002**, *97*, 77–87. [[CrossRef](#)]
35. Dettling, M.; Bühlmann, P. Boosting for tumor classification with gene expression data. *Bioinformatics* **2003**, *19*, 1061–1069. [[CrossRef](#)] [[PubMed](#)]
36. Quackenbush, J. Microarray analysis and tumor classification. *N. Engl. J. Med.* **2006**, *354*, 2463–2472. [[CrossRef](#)] [[PubMed](#)]
37. Reyna, M.A.; Haan, D.; Paczkowska, M.; Verbeke, L.P.; Vazquez, M.; Kahraman, A.; Raphael, B.J. Pathway and network analysis of more than 2500 whole cancer genomes. *Nat. Commun.* **2020**, *11*, 729. [[CrossRef](#)] [[PubMed](#)]
38. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley Interscience: Hoboken, NJ, USA, 2006.
39. Li, M.; Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications*; Springer: Berlin/Heidelberg, Germany, 1993.
40. Grassberger, P. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* **1986**, *25*, 907–938. [[CrossRef](#)]
41. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos Interdiscip. J. Nonlinear Sci.* **2003**, *13*, 25–54. [[CrossRef](#)]
42. Schürmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos Interdiscip. J. Nonlinear Sci.* **1996**, *6*, 414–427. [[CrossRef](#)]
43. Feldman, D.P.; McTague, C.S.; Crutchfield, J.P. The organization of intrinsic computation: Complexity–entropy diagrams and the diversity of natural information processing. *Chaos Interdiscip. J. Nonlinear Sci.* **2008**, *18*, 043106. [[CrossRef](#)]
44. Li, M.; Chen, X.; Li, X.; Ma, B.; Vitányi, P.M. The similarity metric. *IEEE Trans. Inf. Theory* **2004**, *50*, 3250–3264. [[CrossRef](#)]
45. Lesne, A.; Blanc, J.L.; Pezard, L. Entropy estimation of very short symbolic sequences. *Phys. Rev. E* **2009**, *79*, 046208. [[CrossRef](#)]
46. Lempel, A.; Ziv, J. On the complexity of finite sequences. *IEEE Trans. Inf. Theory* **1976**, *22*, 75–81. [[CrossRef](#)]
47. Ziv, J. Coding theorems for individual sequences. *IEEE Trans. Inf. Theory* **1978**, *24*, 405–412. [[CrossRef](#)]
48. Amigó, J.M.; Kennel, M.B. Variance estimators for the Lempel–Ziv entropy rate estimator. *Chaos Interdiscip. J. Nonlinear Sci.* **2006**, *16*, 043102. [[CrossRef](#)]
49. Estevez-Rams, E.; Lora-Serrano, R.; Nunes, C.A.J.; Aragón-Fernández, B. Lempel–Ziv complexity analysis of one-dimensional, cellular automata. *Chaos Interdiscip. J. Nonlinear Sci.* **2015**, *25*, 123106. [[CrossRef](#)]
50. Estevez-Rams, E.; Estevez-Moya, D.; Garcia-Medina, K.; Lora-Serrano, R. Computational capabilities at the edge of chaos for one-dimensional, system undergoing continuous transitions. *Chaos Interdiscip. J. Nonlinear Sci.* **2019**, *29*, 043105. [[CrossRef](#)]
51. Melchert, O.; Hartmann, A.K. Analysis of the phase transition in the two-dimensional Ising ferromagnet using a Lempel–Ziv string-parsing scheme and black-box data-compression utilities. *Phys. Rev. E* **2015**, *91*, 023306. [[CrossRef](#)]
52. Kolmogorov, A.N. Three approaches to the quantitative definition of information. *Probl. Inf. Transm.* **1965**, *1*, 1–7. [[CrossRef](#)]
53. Felsenstein, J. Phylogenetic inference package (PHYLIP), version 3.2. *Cladistics* **1989**, *5*, 164–166.
54. Mootha, V.K.; Lindgren, C.M.; Eriksson, K.F.; Subramanian, A.; Sihag, S.; Lehar, J.; Groop, L.C. PGC-1 $\alpha$ -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **2003**, *34*, 267–273. [[CrossRef](#)]
55. Weinberg, R.A.; Hanahan, D. The hallmarks of cancer. *Cell* **2000**, *100*, 57–70.
56. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [[CrossRef](#)] [[PubMed](#)]