Compositional searching of CpG islands in the human genome

Pedro Luis Luque-Escamilla
Department of Engineering and Mining Mechanics, University of Jaén, Escuela Politécnica Superior, Campus Las Lagunillas s/n, 23071 Jaén, Spain

José Martínez-Aroza*
Department of Applied Mathematics, University of Granada, Facultad de Ciencias, Avenida Fuentenueva s/n, 18071 Granada, Spain

José L. Oliver
Department of Genetics, University of Granada, Facultad de Ciencias, Avenida Fuentenueva s/n, 18071 Granada, Spain

Juan Francisco Gómez-Lopera and Ramón Román-Roldán
Department of Applied Physics, University of Granada, Facultad de Ciencias, Avenida Fuentenueva s/n, 18071 Granada, Spain
(Received 21 October 2004; revised manuscript received 31 January 2005; published 29 June 2005)

We report on an entropic edge detector based on the local calculation of the Jensen-Shannon divergence with application to the search for CpG islands. CpG islands are pieces of the genome related to gene expression and cell differentiation, and thus to cancer formation. Searching for these CpG islands is a major task in genetics and bioinformatics. Some algorithms have been proposed in the literature, based on moving statistics in a sliding window, but its size may greatly influence the results. The local use of Jensen-Shannon divergence is a completely different strategy: the nucleotide composition inside the islands is different from that in their environment, so a statistical distance—the Jensen-Shannon divergence—between the composition of two adjacent windows may be used as a measure of their dissimilarity. Sliding this double window over the entire sequence allows us to segment it compositionally. The fusion of those segments into greater ones that satisfy certain identification criteria must be achieved in order to obtain the definitive results. We find that the local use of Jensen-Shannon divergence is very suitable in processing DNA sequences for searching for compositionally different structures such as CpG islands, as compared to other algorithms in literature.

DOI: 10.1103/PhysRevE.71.061925 PACS number(s): 87.15.Cc, 87.15.Aa

I. INTRODUCTION

The CpG dinucleotide (two adjacent nucleotides, cytosine and guanine, joined by a phosphodiester bond) is generally very deficient in mammalian genomes possibly due to methylation, but it may appear clustered in "CpG islands" dispersed along the genome, especially close to or within the genes. These CpG islands seem to play an important role in gene expression regulation and cell differentiation [1]. In addition, the fact that they are mainly placed near the promoter makes them useful to predict promoters and first exons in the human genome [2,3]. The lack of definition up to now, and the absence of a characteristic length of these pieces of genome information, together with their high functional genomic role, make the search for CpG islands one of the present challenges to beat in bioinformatics and genomics. Different methods to find these CpG islands along any DNA sequence have been proposed, most of them based on moving statistics in a window sliding over the entire sequence. The detected islands must satisfy some biological identification criteria [3–6]. However, in these approaches the size of the window is a very sensitive parameter.

Here we propose a totally different, window size independent, strategy for searching for CpG islands. The central point of the technique is that the CpG island is a piece of DNA sequence compositionally different from the background. Thus, we propose the local usage of the Jensen-Shannon divergence (δ_{JS} hereafter), an information-theory function which has shown to be a good measure of the compositional difference between two (or several) probability distributions [7]. The adequacy of δ_{JS} is highlighted by the fact that it is presently claiming attention from different fields, becoming a central element within the set of statistical measures of distance. Recent theoretical studies about generalized δ_{JS} functions and their metric properties have being achieved [8–12]. The authors and collaborators have developed methods based on δ_{JS}, namely, in basic and applied image processing [13–16] and in sequence analysis [17–21]. Applications to new and diverse fields are currently being proposed [22]. However, it is noteworthy that the use of δ_{JS} in sequence analysis [17–21] has been made globally to the whole sequence.

In this paper, δ_{JS} is used in a local way to detect edge positions associated with changes of composition in a DNA sequence, by looking for the maximal δ_{JS} values between two adjacent windows sliding along the sequence. CpG islands are then identified as the regions between two such edge points that satisfy the appropriate requirements.

*Corresponding author. Mailing address: Departamento de Matemática Aplicada, Facultad de Ciencias, 18071 Granada, Spain. FAX: +34 58 24 29 40. Electronic address: jmaroza@ugr.es

II. METHOD

A. δ_{JS} as a measure of the divergence between probability distributions

The basic objective, common to many of the applications mentioned in the Introduction, is to divide a sequence into adjacent segments with different compositional structure by using δ_{JS} as the discriminant function to measure the compositional distance between dinucleotide relative frequencies taken as probability distributions. The properties of the δ_{JS} function and the advantages of using it for these objectives are shown in the literature [13,18,23]. For the sake of self-containment, the most important rationale follows.

The δ_{JS} function, as an information-theory measure of probabilistic divergence between the probability distributions $\mathbf{P}_1$ and $\mathbf{P}_2$, is defined as

$$\delta_{JS}(\mathbf{P}_1,\mathbf{P}_2) = H\left(\frac{1}{2}\mathbf{P}_1 + \frac{1}{2}\mathbf{P}_2\right) - \frac{1}{2}H(\mathbf{P}_1) - \frac{1}{2}H(\mathbf{P}_2), \quad (1)$$

where $H(\{p_1,p_2,\ldots,p_n\}) = -\Sigma_{j=1}^{n} p_j \log p_j$ is the Shannon entropy. The δ_{JS} function has several interesting properties: it is nonnegative, bounded, and symmetric with respect to its arguments. δ_{JS} satisfies the branching property and can be generalized to any number N of distributions $\mathbf{P}_1,\mathbf{P}_2,\ldots,\mathbf{P}_N$ which are not required to be absolutely continuous. Another interesting property of δ_{JS} is that the distributions can be weighted, which has been used with advantage in segmentation by taking the weights as the relative sizes of the segments involved [16,17]:

$$\delta_{JS}(\pi_1,\mathbf{P}_1;\pi_2,\mathbf{P}_2) = H(\pi_1\mathbf{P}_1 + \pi_2\mathbf{P}_2) - \pi_1 H(\mathbf{P}_1) - \pi_2 H(\mathbf{P}_2), \quad (2)$$

where $\pi_1 = L_1/(L_1+L_2)$, $\pi_2 = L_2/(L_1+L_2)$. In addition, it is noteworthy that the square root of the nonweighted δ_{JS} is a metric [11].

The weighted divergence δ_{JS} of a sequence segmented is related to other information-theory functions. In particular, it is proved that δ_{JS} is the mutual information between the probability distribution of the basis and that of the N segments [18]. That is to say, by randomly choosing a letter in the sequence, the first means the probability distribution about the basis, $\mathbf{P}_b$; the second is about what segment the letter belongs to, $\mathbf{P}_s$ [18]:

$$\delta_{JS}(\mathbf{P}_1,\mathbf{P}_2) = I(\mathbf{P}_b,\mathbf{P}_s) = H(\mathbf{P}_b) + H(\mathbf{P}_s) - H(\mathbf{P}_b \otimes \mathbf{P}_s), \quad (3)$$

where $\mathbf{P}_b = \{p_A,p_T,p_C,p_G\}$, $\mathbf{P}_s = \{\ell_1/L,\ell_2/L,\cdots,\ell_N/L\}$, $L = \ell_1+\ell_2+\cdots+\ell_N$, and $H(\mathbf{P}_b \otimes \mathbf{P}_s)$ is the corresponding bivariate probability distribution. Thus, δ_{JS} quantifies the information provided by saying "what symbol a letter is" about "what segment may it belong to," or equivalently, "what segment a letter belongs to" about "what symbol may it be."

On the other hand, the weighted δ_{JS} is equal to the average—with the same weights—of the Kullback-Leibler relative entropies K of the distributions of symbols in the segments with respect to the overall symbol distribution in the sequence [12]:

$$\delta_{JS}(\pi_1,\mathbf{P}_1;\pi_2,\mathbf{P}_2) = \pi_1 K(\mathbf{P}_1,\mathbf{P}) + \pi_2 K(\mathbf{P}_2,\mathbf{P}), \quad (4)$$

with $K(\mathbf{P}_i,\mathbf{P}) = \Sigma_j p_{ij} \log(p_{ij}/p_j)$, $\mathbf{P} = \pi_1\mathbf{P}_1 + \pi_2\mathbf{P}_2$.

In this sense, δ_{JS} can be seen as the average gain of information obtained when the sequence is being described by the weighted segment distributions, instead of by the overall sequence distribution. All these relationships reinforce the meaning of δ_{JS} as a measure of the divergence between probability distributions.

B. Symbolic sequence entropic segmentation

In the literature [17–21] δ_{JS} has been applied to symbolic sequence segmentation in a global way by comparing the two sides of the entire sequence currently in process. However, when the pieces of sequence to segment are of very different sizes, and commonly much shorter than the whole sequence (typically less than 1 kilobase pairs), it is better to apply δ_{JS} locally by comparing two adjacent regions of the genome. Thus the borders between regions of different composition may be detected. It is noteworthy that this approach is better than traditional methods for searching for CpG islands which use moving statistics over a sliding window along the sequence [3–6]. This is because the absence of a characteristic length in CpG islands makes them difficult to detect with a single window of fixed size. In fact, these approaches prove to be very sensitive to window size [19]. Therefore, the local use of δ_{JS} proves to be a powerful task to detect not only CpG islands but also short compositionally different regions with no characteristic length along the genome.

C. CpG island detection method

As pointed out above, the local application of δ_{JS} to the DNA sequence may be used to detect regions of different composition. In particular, it may be used with benefit in the search for CpG islands. The whole procedure is described in three stages: (1) the calculation of local Jensen-Shannon di-

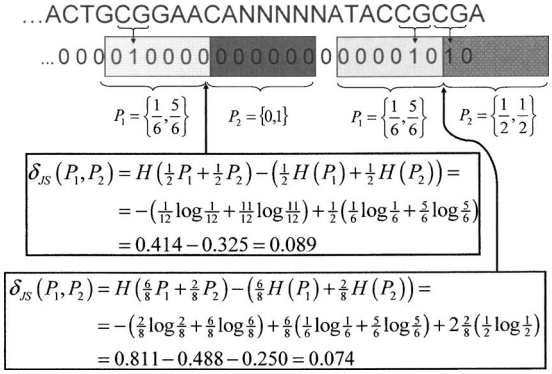

FIG. 1. The original DNA sequence is composed of adenine, guanine, cytosine, and thymine (A,G,C,T), and eventually N when masked or unknown. The algorithm first encodes the DNA sequence to binary, being 1 if a CG pair is found and 0 otherwise. A sliding double window then runs the binary sequence and the weighted δ_{JS} is calculated at each position, as shown. As is customary, base 2 is used hereinafter for logarithms.

TABLE I. Mean values for the training set using the proposed algorithm and CPGPROD for comparison. The window size is denoted by w. Four significant digits are shown.

Mean	T_P	T_N	F_P	F_N	Q_{sn}	Q_{sp}
$w=100$ bp	1.714×10^3	1.172×10^5	1.719×10^3	0.2858×10^3	0.8692	0.5270
$w=200$ bp	1.716×10^3	1.164×10^5	1.704×10^3	0.2887×10^3	0.8764	0.5324
$w=500$ bp	1.631×10^3	1.174×10^5	1.569×10^3	0.4454×10^3	0.8177	0.5520
$w=1000$ bp	1.690×10^3	1.170×10^5	1.506×10^3	0.4262×10^3	0.8462	0.5522
$w=2500$ bp	1.729×10^3	1.178×10^5	1.487×10^3	0.4999×10^3	0.8280	0.5556
CPGPROD	1.676×10^3	1.160×10^5	1.824×10^3	0.2658×10^3	0.8706	0.4877

vergences along the sequence; (2) the detection of the edges of the segments, as local maxima of the Jensen-Shannon divergence; and (3) the fusion of segments into the definitive ones, according to certain identification criteria. These steps are summarized in Fig. 1.

1. Jensen-Shannon divergence calculation

In order to avoid misdetections of CpG islands the sequence is analyzed for repetitive sequence content, and the repeats called *Alus* [5] are masked using a well-known algorithm (REPEATMASKER [24]). Then, the masked DNA sequence must be converted into a convenient binary sequence by assigning 1 to a *CG* pair and 0 to any other pair (overlapped) along the sequence (see Fig. 1). Next, δ_{JS} is calculated between the relative frequencies of symbols in both sides of the window for each position on the sequence. As the sequence is now binary, the probability distribution in each window is just $\mathbf{P}=\{p,1-p\}$, p being the probability of having a *CG* pair. Since the two sides of the window could eventually not have the same size due to lack of information, such as in the beginning and ending of the sequence, the weighted δ_{JS} is computed in general (see Fig. 1).

2. Edge detection

Once δ_{JS} is computed along the entire sequence, all local maxima might be identified with edges between compositionally different regions. However, not all local maxima have to be due to significant differences in composition; they may be due to statistical fluctuations. Thus, an efficient and fast method for identifying the reliable maxima must be performed based on the idea of adequately reducing the searching space. This is done here by the simultaneous use of two parameters: a *divergence threshold*, below which the maxima are not statistically significant, and a *sampling interval* which is the minimum distance expected between significant local maxima, which permits every local maximum to be roughly located first and then finely determined. The algorithm first looks for the maximum of δ_{JS} in the whole sequence, and then tries to find the next local maximum above the divergence threshold and using the sampling interval. This procedure is repeated at each side of the global maximum. Note that the located significant maxima have divided the entire sequence into adjacent segments of different dinucleotide composition.

The value of the divergence threshold may be chosen based on the fact that in any large enough segment of mam-

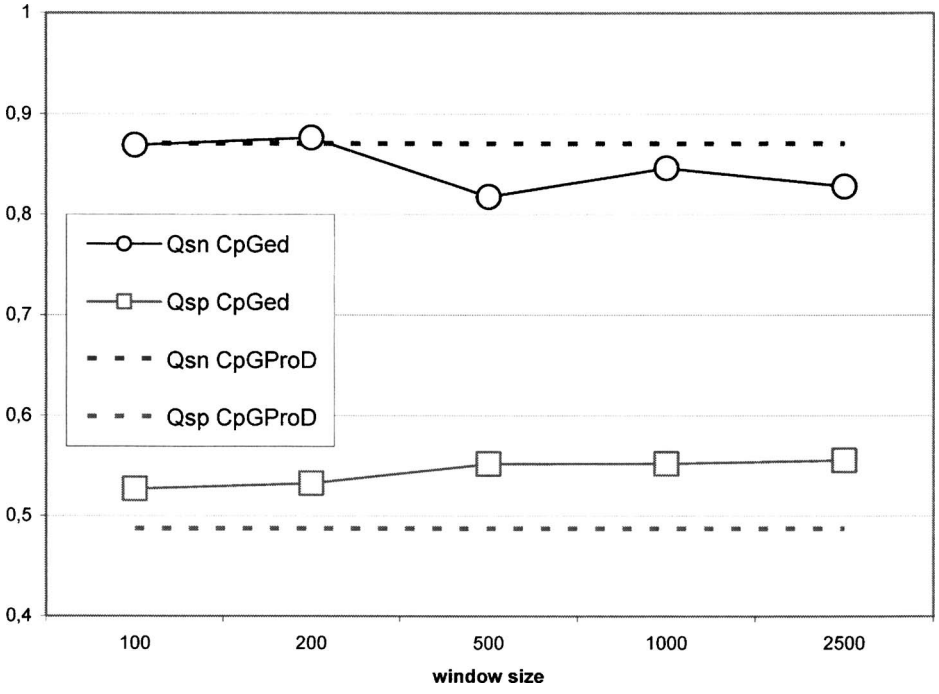

FIG. 2. Sensibility Q_{sn} and specificity Q_{sp} values when the window size is varied from 100 to 2500 bp. Sampling interval is fixed at 5 bp. CPGED is the proposed method.

TABLE II. Mean values for the training set using the proposed algorithm and CPGPROD for comparison. The sampling interval is denoted by s. Four significant digits are shown.

Mean	T_P	T_N	F_P	F_N	Q_{sn}	Q_{sp}
$s=5$ bp	1.727×10^3	1.167×10^5	1.730×10^3	0.2863×10^3	0.8806	0.5317
$s=10$ bp	1.716×10^3	1.164×10^5	1.704×10^3	0.2887×10^3	0.8765	0.5324
$s=50$ bp	1.730×10^3	1.170×10^5	1.642×10^3	0.3302×10^3	0.8626	0.5507
$s=100$ bp	1.709×10^3	1.172×10^5	1.582×10^3	0.3782×10^3	0.8464	0.5644
CPGPROD	1.676×10^3	1.160×10^5	1.824×10^3	0.2658×10^3	0.8706	0.4877

malian DNA the minimum proportion of *CG* pairs needed to justify a significant deviation from the statistical fluctuations is established as 7% [5]. Thus a threshold value of 0.001 was chosen, which is equivalent to a deviation less than 1.5% with respect to the 7% of *CG* pairs in the segment. This choice is compatible with the order of the observed size of random fluctuations in δ_{JS}. In addition, the final value of the threshold proves not to be critical in the results. On the other hand, the value of the sampling interval is empirically established, as will be discussed in Sec. III.

3. Fusion of segments

From all the segments in the previous step, we have to identify those that are CpG islands, or whose fusion gives rise to CpG islands. Thus, the method must check all possible fusions of one or more segments, giving priority to length—hence checking first the entire sequence. Any fusion is found to be a CpG island if it satisfies some statistic criteria, following biologists: length greater than 500 base pairs (bp), percentage of both nucleotides *C* and *G* greater than 55%; ratio of observed to expected (defined as the number of CpG divided by the product of the percentages of *C* and *G* nucleotides [4]) greater than 0.65 [3,5], and, as said before, percentage of *CG* pairs greater than 7% [5]. In such a case the CpG island is definitive, and no longer considered by the algorithm.

III. RESULTS

To check the prediction capacity of the proposed method we initially divided a set of human DNA sequences from GenBank into two sets (400 contigs each): one for training the procedure, and the other for testing it. The training is performed in order to optimize the values of the window size and the sampling interval, and to verify how robust the method is under a little variation in any of those parameters.

Our CpG island identification is related to the CpG island location annotated in the contig. Thus, if the proposed method correctly identifies a CpG island according to the annotations, our prediction is true positive (T_P) or negative (T_N), respectively. If the algorithm predicts a CpG island in a location that is not annotated in the contig, a false positive (F_P) occurs, and if the method does not predict an annotated CpG island, a false negative (F_N) takes place. Two quality measures may be built from these quantities [3]: the sensitivity $Q_{sn}=T_P/(T_P+F_N)$, the proportion of T_P predictions out

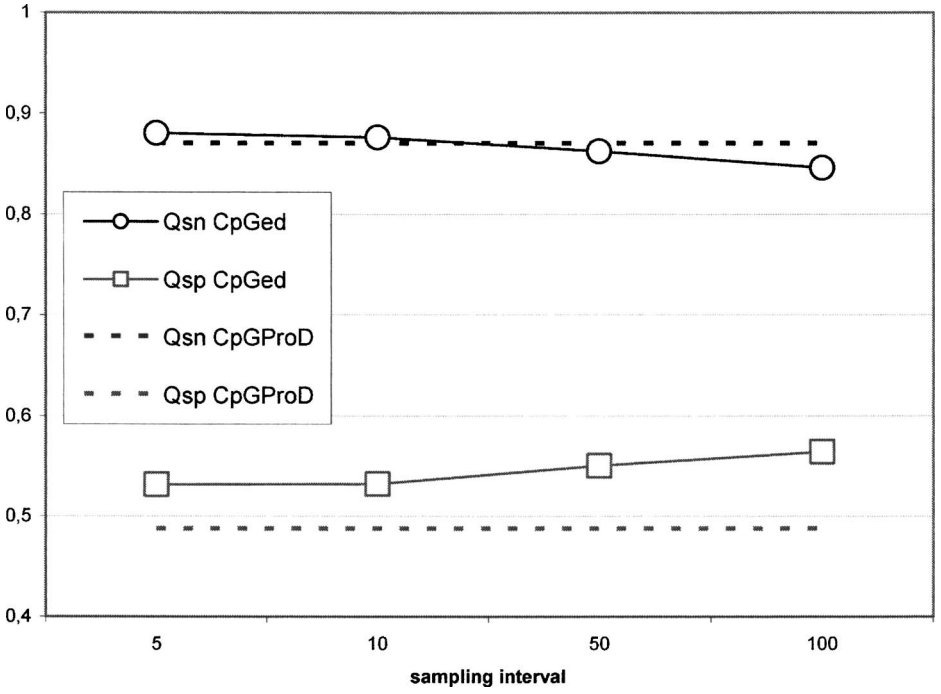

FIG. 3. Sensibility Q_{sn} and specificity Q_{sp} values when the sampling interval is varied from 5 to 100 bp. Window size is fixed at 200 bp. CPGED is the proposed method.

TABLE III. Mean values for the test set using the proposed algorithm and CPGPROD for comparison. Four significant digits are shown.

Mean	T_P	T_N	F_P	F_N	Q_{sn}	Q_{sp}
Proposed	2.191×10^3	1.075×10^5	2.262×10^3	0.4034×10^3	0.8538	0.5197
CPGPROD	2.380×10^3	1.037×10^5	2.285×10^3	0.5651×10^3	0.8242	0.4910

of the total number of actual positives, and the specificity $Q_{sp}=T_P/(T_P+F_P)$, the proportions of T_P's out of the total number of predicted positives. These two measures must be as high as possible for a good prediction result, but it should be taken into account that for our purposes values of F_P could be more acceptable than F_N ones. This is because the CpG islands associated with promoters are the only biologically interesting ones, and this association is more probable if the island is large. Thus, Q_{sp} may get relatively low values although this may be not such a bad result.

In the training set five window sizes were probed with a fixed sampling interval of 10 bp: 100, 200, 500, 1000, and 2500 bp. Results are shown in Table I and in Fig. 2, along with the results obtained with the recent and accurate CPGPROD algorithm of Ponger and Mouchiroud [3]. Two conclusions may be extracted: the best choice is the 200 bp window, and the algorithm is very robust against changes in the window size as expected. This is very important because the traditional window based algorithms were very sensitive to this parameter.

In order to find the best choice of the sampling interval value we run the algorithm with 5, 10, 50, and 100 bp, with fixed 200 bp window size. Results are shown in Table II and in Fig. 3, indicating that the best value for the significant interval is 5 bp, and that the algorithm is certainly not sensitive to variations in this parameter.

Once the parameters are trained, the test set is used to check the prediction ability of the method. CPGPROD is used for comparison. As may be seen in Table III, the performance of the proposed algorithm is better than CPGPROD, not only in Q_{sn} but even in Q_{sp}. It is noteworthy that this method, as previously trained, has been used in the detection of the CpG islands in the human chromosome 22 obtaining as a preliminary result that some island sizes are much larger that supposed up to now (up to 30000 base pairs).

IV. DISCUSSION

In this article an entropic edge detector (or cut points detector) based on the local use of the Jensen-Shannon divergence is presented, and some of the mathematical properties useful for the task are also outlined. The detector proves to be a powerful tool to analyze DNA sequences in order to segment compositionally different regions in the genome. In particular, this method turns out to be very useful in the search for CpG islands in the human genome. As can be seen in the shown results, the method is very robust against changes of any of its three eligible parameters. This means that there is no necessity of a complex training procedure to adjust the searching method parameters to our particular problem at hand. At the same time, the settings of the method do not depend critically on the island features to be detected.

Another remarkable feature is the general searching strategy used. Previous work [3–6] use a single window to look inside for an island. This strategy inherently lacks completeness since an island is primarily defined in relation to the background (ocean) in which it is immersed; therefore the edge island-ocean must be detected. That is why we propose a sliding double window with fixed size along with a fusion criterion and procedure to merge nearby islands. In addition, the method proposed to discriminate CpG islands from the background is completely different from previously published ones. While these algorithms are based on moving statistics, here the δ_{JS} function is used to segment the sequence into regions of different composition.

The proposed algorithm has been trained with a set of contigs and then applied to another set, both sets from GenBank and larger than 400 DNA chains. Numerical results show than our method is better than CPGPROBD [3] in both the sensibility Q_{sn} and the specificity Q_{sp} mean values.

The described results back the local use of δ_{JS} as an edge detector when applied to DNA sequences, and make it a worthy choice to be considered in other types of searches, with adequate codification.

ACKNOWLEDGMENT

This work was partially supported by Grant No. BIO2002-04014-C03-03 from the DGICYT of the Spanish Government.

[1] A. P. Bird, Genes Dev. **16**, 6 (2002).
[2] M. Scherf, A. Klingenhoff, and T. Werner, J. Mol. Biol. **297**, 599 (2000).
[3] L. Ponger and D. Mouchiroud, Bioinformatics **18**, 631 (2002).
[4] M. Gardiner-Garden and M. Frommer, J. Mol. Biol. **196**, 261 (1987).
[5] D. Takai and P. A. Jones, Proc. Natl. Acad. Sci. U.S.A. **99**, 3740 (2002).
[6] I. P. Ioshikhes and M. Q. Zhang, Nat. Genet. **26**, 61 (2000).
[7] J. Lin, IEEE Trans. Inf. Theory **37**, 145 (1991).
[8] W. K. Wootters, Phys. Rev. D **23**, 357 (1981).
[9] P. W. Lamberti, M. T. Martín, A. Plastino, and O. A. Rosso,

Physica A **334**, 119 (2004).
[10] P. W. Lamberti and A. P. Majtey, Physica A **329**, 1 (2003).
[11] D. M. Endress and J. E. Schindelin, IEEE Trans. Inf. Theory **49**, 7 (2003).
[12] F. Topsoe (unpublished).
[13] J. F. Gómez Lopera, J. A. Martínez Aroza, A. M. Robles Pérez, and R. Román Roldán, J. Math. Imaging Vision **13**, 35 (2000).
[14] C. Atae Allah, M. A. Cabrerizo Vílchez, J. A. Holgado Terriza, P. L. Luque Escamilla, J. F. Gómez Lopera, and R. Román Roldán, Meas. Sci. Technol. **12**, 288 (2001).
[15] R. Román Roldán, J. F. Gómez Lopera, J. A. Martínez Aroza, P. L. Luque Escamilla, and C. Atae Allah, Pattern Recogn. **34**, 969 (2001).
[16] D. Pozo Vázquez, P. L. Luque Escamilla, and C. Atae-Allah, J. Atmos. Ocean. Technol. **16**, 970 (1999).
[17] P. A. Bernaola Galván, R. Román Roldán, and J. Oliver, Phys. Rev. E **53**, 5181 (1996).
[18] I. Grosse, P. A. Bernaola Galván, P. J. Carpena Sánchez, R. Román Roldán, J. Oliver, and H. E. Stanley, Phys. Rev. E **65**, 041905 (2002).
[19] W. Li, P. A. Bernaola Galván, F. Haghighi, and I. Grosse, Comput. Chem. (Oxford) **26**, 491 (2002).
[20] J. Oliver, P. A. Bernaola Galván, P. J. Carpena Sánchez, and R. Román Roldán, Gene **276**, 47 (2001).
[21] P. A. Bernaola Galván, I. Grosse, P. J. Carpena Sánchez, R. Román Roldán, J. Oliver, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).
[22] C. J. Tessone, A. Plastino, and H. S. Wio, Physica A **326**, 37 (2003).
[23] R. Román Roldán, P. A. Bernaola Galván, and J. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).
[24] A. F. A. Smit and P. Green, computer code REPEATMASKER, available at http://ftp.genome.washington.edu/RM/