Alternate measure of information useful for DNA sequences

Ranjan Bose[1] and Sonali Chouhan[2]

[1]*Department of Electrical Engineering, IIT Delhi, Hauz Khas, New Delhi, India*

[2]*Department of Electronics and Electrical Engineering, IIT Guwahati, Guwahati, India*

(Received 27 August 2010; revised manuscript received 7 March 2011; published 20 May 2011)

We propose an alternate measure of information, called superinformation, which has been found to be very effective for analyzing the coding and noncoding regions of the DNA. This superinformation is actually a measure of the "randomness of randomness." It has been found to be highly accurate in classifying coding and noncoding regions of human DNA. In the proposed method, no prior training is required. This technique exhibits higher accuracy than previously reported techniques in distinguishing between the coding and the noncoding portions of the DNA. Superinformation can also be used to analyze the untranslated regions in various genes.

DOI: 10.1103/PhysRevE.83.051918 PACS number(s): 87.14.gk, 87.14.gf, 87.15.Cc, 89.70.Cf

I. INTRODUCTION

A gene is a sequence of deoxyribonucleic acid (DNA) that contains genetic information, and specifies parts of protein [1]. In eukaryotes (organisms that have cells with complex structures), genes consist of exons and introns. The exon, sometimes termed as the coding section, is responsible for coding proteins. These exons are interspersed with introns, sometimes termed as the noncoding section of the DNA. Most of the introns have no known biological function and are sometimes termed as "junk DNA" [2]. There are also instances of exons that do not direct the production of a peptide sequence [3].

Several varied techniques have been explored in the past to analyze the information content of genes [4–8]. In [9] the authors have estimated the entropy of DNA sequences using the standard Shannon entropy measure. It is concluded in the paper that the DNA sequence has more randomness than human language or computer language. Long range correlation between nucleotide sequences have been investigated in [10]. The method of statistical linguistics has been used to evaluate the coding and noncoding regions of the DNA in [11]. In [12] it is reported that natural DNA sequences have significantly lower entropy estimates. A well-known statistical feature of coding regions is nonuniform codon usage. In coding regions, not all codons (triplets of nucleotides) occur with the same probability [13]. It has been shown in [14] that the block entropies from frequencies observed in different realizations of the same length and with the same underlying probability distribution are highly concentrated around the mean. An entropy-based segmentation method for finding the borders between coding and noncoding DNA regions is proposed in [15]. In this paper, the DNA sequences are described by a 12-letter alphabet that captures the differential base composition at each codon position. In [16] the authors focus on sequence and structure features of single protein residues and analyze how they describe a residue's contributions to the DNA-binding event using mutual information.

This paper is organized as follows. Sec. II provides the basic motivation behind defining superinformation. Sec. III gives the mathematical definition of superinformation. Segmentation of coding and noncoding regions of the human DNA using superinformation is discussed in Sec. IV. Sec. V deals with the accuracy of the proposed technique in discriminating coding and noncoding regions. This mathematical tool is also effective in analyzing the untranslated regions (UTR) of the DNA. Some examples are provided in Sec. VI to corroborate the theory of superinformation. Finally, we conclude the paper in Sec. VII.

II. MOTIVATION FOR SUPERINFORMATION

In all the previous attempts to analyze the information content of genetic data, Shannon's entropy (or some variation) has been used. Shannon's entropy, also called average self-information [17], is a measure for order (or disorder), and is defined as

$$H(X) = -\sum_{i=1}^{N} p(x_i)\ln p(x_i), \tag{1}$$

where X is a discrete random variable with possible values $\{x_1,x_2,\ldots,x_N\}$ and $p(x_i)$ represents the probability of x_i. When the base of the logarithm is 2, $H(X)$ is measured in bits. $H(X)$, also called the average self-information [18], describes the uncertainty or randomness in X. We observe here that more randomness (disorder) implies a higher information content. $H(X)$ has been widely used to study the information content of DNA sequences earlier [8,12,16,19,20]. However, as the name implies, it gives the information content in the *average* sense, and may not provide the true picture in many cases, as illustrated by the following example. Consider the following three sequences constructed from the symbols $\{A, T, G, C\}$:

$$\begin{aligned} S_1 &: AAAAGGGGTTTTCCCC, \\ S_2 &: AGTCAGTCAGTCAGTC, \\ S_3 &: ATTGACCCTGTCGAGA. \end{aligned} \tag{2}$$

While S_1 and S_2 have repeating patterns, S_3 is more disordered. However, the average self-information, as defined by (1), is the *same* for all the three sequences. The probability of occurrence of a particular symbol has been derived from its relative frequency. In (2), $H(S_1) = H(S_2) = H(S_3) = 2$ bits, which is the maximum possible for four equiprobable symbols. Clearly, $H(X)$ is not a good measure for analyzing sequences of symbols with repeating patterns, as is common in DNA sequences [21,22]. This motivates us to first subdivide the sequence into blocks, and then determine the information content of each block. Depending on the particular sequence under evaluation, the different blocks will have different information content (i.e., measure for randomness). Thus there is randomness in the measure for randomness, or in other words, *uncertainty of uncertainty*.

III. SUPERINFORMATION

Let us divide the entire sequence of symbols into N blocks of length B each. Let the ith block be represented by X_i and the corresponding entropy be $H(X_i)$. By definition, $H(X_i)$ is a non-negative quantity. Let us define a probability measure as follows:

(i) Construct the histogram of $H(X_i)$, i.e.,

$$\{H_j(X_i,M)\} = \text{histogram}\,(\{H(X_i)\})\,, \quad i = 1,2,\ldots,N, \quad (3)$$

where the histogram function collects the elements of vector $\{H(X_i)\}$ into M equally spaced bins and returns the number of elements in each bin. $\{H_j(X_i,M)\}$ is a vector of length M. The histogram function thus gives the distribution of data values corresponding to$\{H(X_i)\}$.

(ii) Normalize the histogram to form the probability measure:

$$p_j(X_i,M) = \frac{H_j(X_i,M)}{\sum_{k=1}^{M} H_k(X_i,M)}, \quad j = 1,2,\ldots,M. \quad (4)$$

Here, $p_j(X_i,M)$ represents the probability of $H(X_i)$ being in the jth bin.

Now, define *superinformation* as follows:

$$H_s(X_i,B,M) = -\sum_{j=1}^{M} p_j(X_i,M) \ln p_j(X_i,M). \quad (5)$$

$H_s(X_i,B,M)$ is a measure of the “entropy of entropy” and B defines the resolution at which this superinformation is calculated. For the three sequences given in (2), the values for $H_s(X_i,B,M)$ are given in Table I. The value of $M = 10$. From the table we can see that the superinformation for the three sequences are different (as opposed to the average self-information, which was the same), and depends on the value of B.

This example is purely illustrative. Typically, much longer sequences are used to get a good estimate of entropy. From the table, the value of the superinformation $H_s(X_i,B,M) = 0$ for $B = 2$ and 4 for the first two sequences S_1 and S_2. This is because there is a distinct repetitive pattern at this resolution. However, no such pattern is present for S_3, and hence the superinformation is nonzero. superinformation is non-negative, by definition, as is clear from (5).

Intuitively, superinformation represents the variation in the entropy content of different portions of a given sequence. If there is a large variation in the entropy of different segments, the superinformation will be higher. Thus superinformation is a measure of “randomness of randomness” of the sequence. If all portions of a given sequence are highly random, the superinformation will come out to be low. The parameter “B” defines the resolution at which the sequence is being analyzed. The distribution of entropy estimates from finite sequences can typically be characterized by a Gaussian distribution [23,24]. Consequently, superinformation may be interpreted as roughly the logarithm of the standard deviation of finite-sample estimates of entropies.

TABLE I. Values for $H_s(X_i,B,M)$ for different values of B.

Symbols/block (B)	Number of blocks (N)	$H_s(S_1)$	$H_s(S_2)$	$H_s(S_3)$
2	8	0	0	0.543
3	5	0.971	0	0.971
4	4	0	0	1.500
5	3	0.918	0	1.585

IV. ANALYSIS OF CODING AND NONCODING REGIONS

We have used superinformation as a mathematical tool to study the coding and noncoding sections of DNA. The data consist of the human chromosomes taken from [25].

First we consider gene TTC34 from human chromosome 1. This gene corresponds to tetratricopeptide, repeat domain 34 (HGNC:34297) and is known to code for protein. The first transcript of this gene has nine exons (coding portions), interspersed with eight introns (noncoding portion). The nine exons are concatenated to form a contiguous coding portion of the gene. Similarly, the eight introns are concatenated to form a contiguous noncoding portion of the gene. The superinformation is then determined for these concatenated coding and the noncoding sections. Figure 1 shows the plot of the superinformation, $H_s(S,B)$ of the coding segment (S_c) and the noncoding sections (S_nc) versus B. As discussed in the earlier section, the entire sequence of symbols is divided into blocks of length B, which essentially defines the resolution at which the sequence is being analyzed. The value of B has been varied in steps of 3 in order to maintain codon structure. We observe a clear demarcation between these two regions when we analyze them from the perspective of superinformation. This result is different from that obtained for the analysis of coding and noncoding regions on the basis of order index ϕ, which gives a quantitative measure of randomness for genomic sequences [26]. The paper reports that both the coding and noncoding regions have similar ϕ’s.

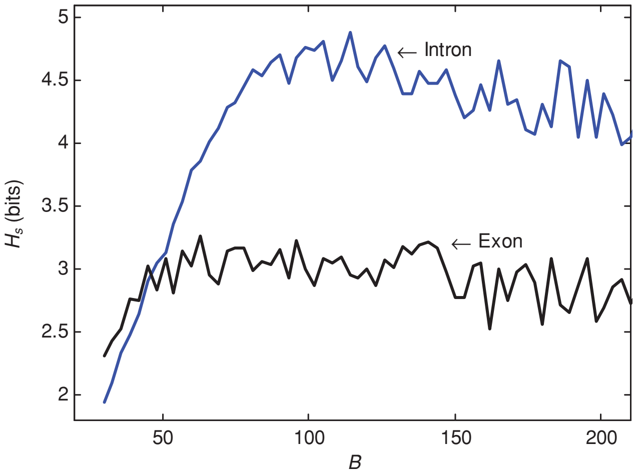

FIG. 1. (Color online) Plot of the superinformation of all introns and all exons of gene TTC34 of chromosome1 for different block sizes B.

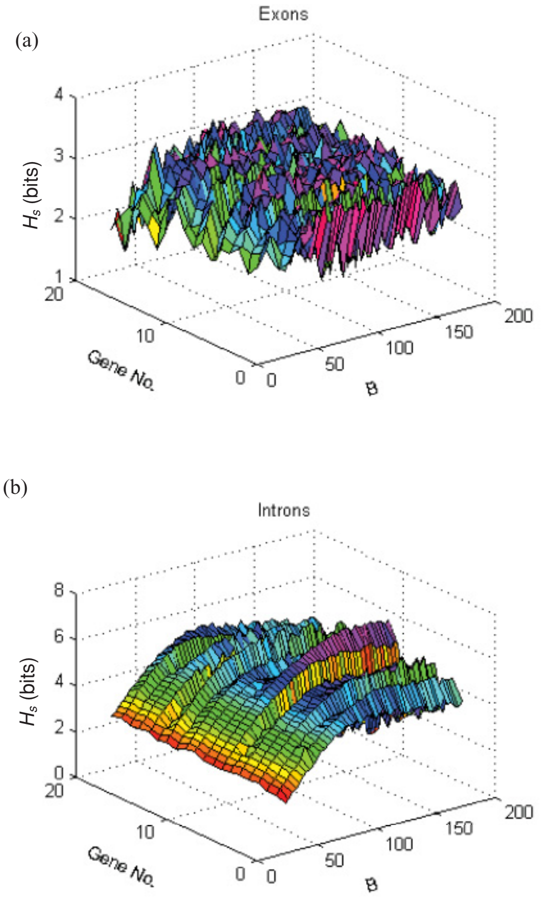

FIG. 2. (Color online) Plot of the superinformation of (a) all exons and (b) all introns for 20 genes belonging to chromosome1 for different block sizes B.

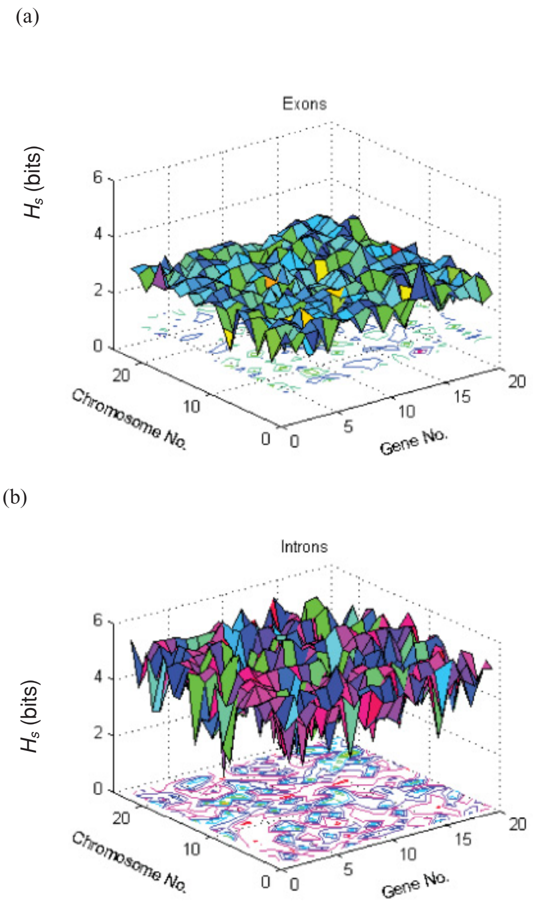

FIG. 3. (Color online) Plot of the superinformation for (a) all exons and (b) all introns present in 20 randomly selected genes in each of the 24 chromosomes (22, X and Y) for $B = 90$.

We further observe that $H_s(S,B)$ has a maximum for a certain value of B. It can be explained as follows. B defines the block size which is used to calculate superinformation. For smaller values of B, a smaller number of entropy values are possible. So, when calculating the entropy of entropy, $H_s(S,B)$ comes out to be small. At large values of B, the spread of the entropy function is less, leading to a smaller $H_s(S,B)$.

Next, we consider 20 genes, picked randomly, from chromosome 1. The number of exons for each gene is different, varying from 6 to 34, interspersed with introns. Within each gene, the coding portions (exons) and the noncoding portion (introns) are concatenated as discussed earlier. The superinformation is then determined for these concatenated coding and the noncoding sections for each gene. The superinformation for the coding section for the 20 genes of chromosome 1 is plotted in Fig. 2(a). The corresponding superinformation for the noncoding section for the 20 genes is plotted in Fig. 2(b). It is interesting to note that there appears to be a pattern for the noncoding region, whereas no such trend is observed for the coding region. Similar results were observed for all 23 human chromosomes.

Since $B \in (80,100)$ results in the maximum values of $H_s(S,B)$ for most cases, we now fix $B = 90$. Figures 3(a) and (b) show the plots of $H_s(S_c)$ and $H_s(S_{nc})$ for 20 randomly selected genes in each of the 24 chromosomes (22, X and Y). Thus we have used 480 randomly selected genes for our studies. There is a clear separation between $H_s(S_c)$ and $H_s(S_{nc})$. Only in approximately 5% of the cases the $H_s(S_c)$ found to exceed $H_s(S_{nc})$. The average value and standard deviation for $H_s(S_c)$ are 2.73 bits and 0.097 bits, respectively, while that of $H_s(S_{nc})$ are 4.510 bits and 0.150 bits, respectively. It is clear from the figure that superinformation is a very effective tool for analyzing the coding and noncoding portion of the DNA.

Figure 3 may have an interesting intuitive (though, speculative) explanation for higher standard deviation for $H_s(S_{nc})$ as compared to that for $H_s(S_c)$. One possible purpose of introns is that they serve as “bricks” or building blocks for exons, and are useful for adaptive evolution [27]. However, these bricks are of different sizes, probably suitable for different purposes. These bricks are chiseled into the desired shape and size (exon) during the process of adaptive evolution. Figure 3(a) is a depiction of the finished building (protein) made out of chiseled bricks (exons). Figure 3(b) can be compared to a pile of bricks of different sizes, ready to be used for further enhancements of the building (evolution). As observed from Fig. 3(b), the pile of bricks of different sizes has a larger variation (standard deviation) than the finished building, i.e., Fig. 3(a). We note that this is just a conjecture and there is a need to find functional evidence to substantiate this claim.

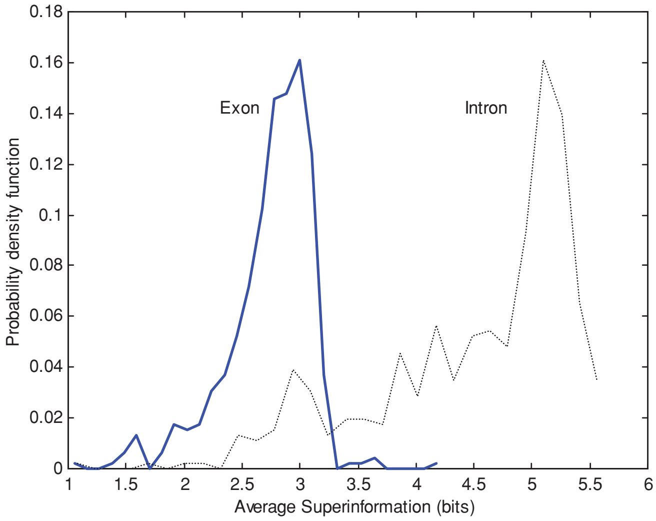

FIG. 4. (Color online) Plot of the probability density functions of the superinformation for all exons and all introns present in 480 randomly selected genes from 24 chromosomes (22, X and Y) for $B = 90$.

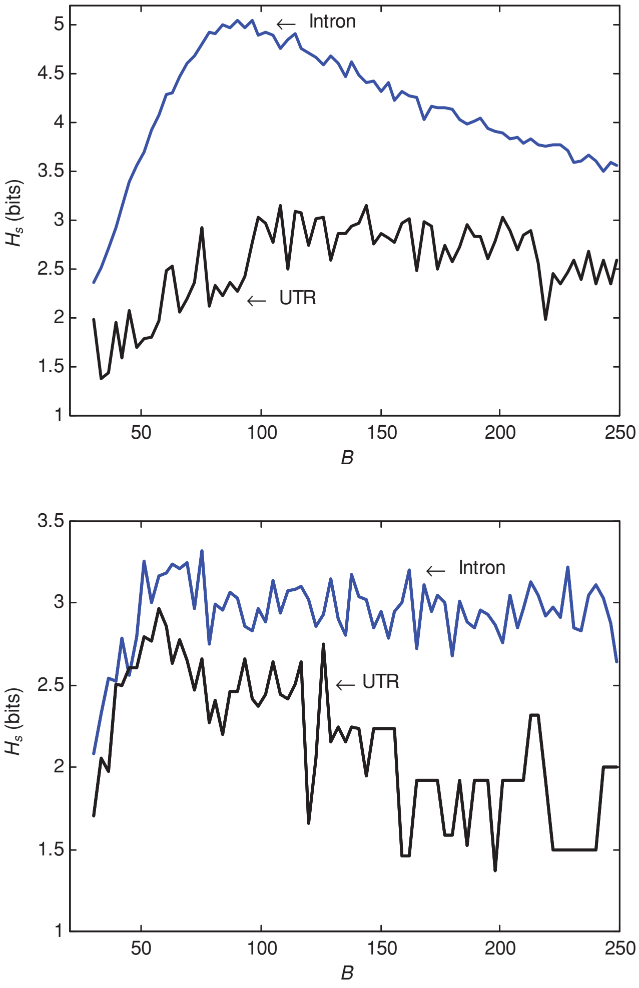

FIG. 5. (Color online) Plot of the superinformation for noncoding portions (introns) and the untranslated region (UTR) for (a) gene RP1-14N1.2 (transcript 1) and (b) gene BX004987.5 (transcript 1).

V. ACCURACY OF SUPERINFORMATION

We next investigate how accurately superinformation can distinguish between the coding and the noncoding portions of DNA. Consider Fig. 4, which shows the probability density function (pdf) of the superinformation for all exons and all introns present in 20 randomly selected genes in each of the 24 chromosomes (22, X and Y). Thus we have used 480 randomly selected genes for our studies. The pdf has been obtained by first normalizing the histogram of the superinformation (with $B = 90$) and then averaging it across the different chromosomes. Let us denote the pdfs for the coding portions (exons) and the noncoding portion (introns) by ρ_c and by ρ_{nc}, respectively. As observed from the figure, there is a clear separation between the two distributions, with marginal overlap. In order to determine the accuracy A, we define the overlap integral as follows [20]:

$$O(H_s) = \int M(H_s) dH_s, \tag{6}$$

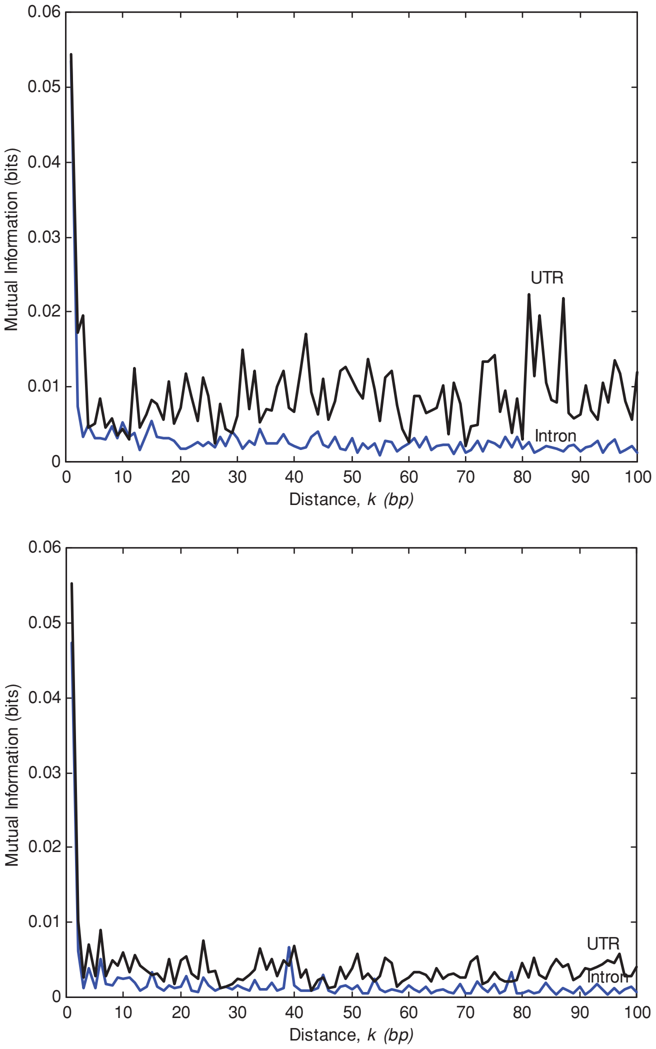

FIG. 6. (Color online) Plot of mutual information function $I(k)$ for noncoding portions (introns) and the untranslated region (UTR) for (a) gene RP1-14N1.2 (transcript 1) and (b) gene BX004987.5 (transcript 1).

where $M(H_s)$ denotes the maximum of $\{\rho_c(H_s),\ \rho_{nc}(H_s)\}$ at position H_s. The accuracy A, defined as

$$A\,(H_s) = O\,(H_s)\,/2, \tag{7}$$

varies from 0.5 (no discrimination) to 1.0 (perfect discrimination). It was found that the accuracy of superinformation $A(H_s) = 0.92$ for the random data set taken from 480 genes across the 24 chromosomes. This is better than the reported accuracy of average mutual information [20] and the accuracy evaluated in [28]. It should be noted that the accuracy for superinformation has been calculated using concatenated sequences of coding and noncoding regions separately.

VI. ANALYSIS OF UNTRANSLATED REGIONS (UTR)

Superinformation can also provide an insight regarding the untranslated regions (UTRs) in certain genes. For example, Fig. 5(a) gives the superinformation plot of the noncoding portions (introns) and the superinformation plot corresponding to the UTR for gene RP1-14N1.2 (transcript 1). This UTR is of length 1109 base pairs. There is a clear separation between the two curves. On the other hand, Fig. 5(b) gives the superinformation plot of the noncoding portions and the superinformation plot corresponding to the UTR for gene BX004987.5 (transcript 1). This UTR is of length 2995 base pairs. Here, the separation between the two curves for smaller values of B is lesser. In fact, for smaller values of B, there is a partial overlap. One may conjecture the following:

(i) From the perspective of superinformation, the UTR for gene RP1-14N1.2 (transcript 1) and gene BX004987.5 (transcript 1) behave more like exons than introns.

(ii) Currently, the sequences are classified as UTRs because no known protein formation is linked to them. However, it is possible that some portion of these UTRs may act as exon (at present or in future) and code a protein.

It should be noted that there is a need to obtain functional evidence and experimental results to corroborate these conjectures.

For comparison, we carry out the analysis of UTRs using the mutual information function defined in [20]:

$$I(k) = \sum_{i,j=1}^{4} P_{ij}(k)\ln_2 \frac{P_{ij}(k)}{p_i q_j}, \tag{8}$$

where $P_i\,j(k)$ denotes the joint probability of finding the pair of nucleotides n_i and n_j (n_i,$n_j \in \{A,G,C,T\}$) spaced by a gap of $k-1$ nucleotides, $p_i = \sum_j P_{ij}(k)$and $q_j = \sum_i P_{ij}(k)$. Figure 6(a) shows the plot of mutual information function $I(k)$ for noncoding portions (introns) and the untranslated region (UTR) for gene RP1-14N1.2 (transcript 1) and Fig. 6(b) is the corresponding plot for gene BX004987.5 (transcript 1). From the figures we observe that the mutual information function is barely able to differentiate between the intron and the UTR. In comparison, Figs. 5(a) and 5(b) based on superinformation show a clear separation between the intron and the UTR.

VII. CONCLUSIONS

In this paper, we have proposed an alternate measure of information called superinformation, which is very useful for analyzing the coding and noncoding regions of the DNA. We have provided the steps for determining this quantity. Superinformation has been found to be very effective for classifying coding and noncoding regions. This technique exhibits higher accuracy than previously reported techniques in distinguishing between the coding and the noncoding portions of the DNA. Superinformation can also be used to analyze the un-translated regions in various genes.

[1] B. Hayes, Am. Sci. **86**, 8 (1998).

[2] S. Ohno, Brookhaven Symp. Biol. **23**, 366 (1972).

[3] J. Wang, J. Kudoh, A. Shintani, S. Minoshima, and N. Shimizu, Biochem. Biophys. Res. Commun. **250**, 704 (1998).

[4] P. Grassberger, Phys. Lett. A **128**, 369 (1988).

[5] H. Herzel, Sys. Anal. Mod. Sim. **5**, 435 (1988).

[6] H. Herzel, W. Ebeling, and A.O. Schmitt, Phys. Rev. E **50**, 5061 (1994).

[7] M. Farach *et al.*, in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, (San Francisco, 1995), p. 48.

[8] J. K. Lanctot, M. Li, and E. Yang, in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, (San Francisco, 2000), p. 409.

[9] A. O. Schmitt and H. Herzel, J. Theor. Biol. **1888**, 369 (1987).

[10] C. K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[11] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **52**, 2939 (1995).

[12] D. Loewenstern and P. N. Yianilos, J. Comput. Biol. **6**, 125 (1999).

[13] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier, Nucleic Acids Res. **9**, R43 (1981).

[14] A. O. Schmitt, *Structural analysis of DNA sequences*, Ph.D. Thesis, Humboldt University of Berlin, Berlin, Verlag, 1995.

[15] P. Bernaola-Galvan, I. Grosse, P. Carpena, J. L. Oliver, R. Roman-Roldan, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).

[16] C. Kauffman and G. Karypis, Pacific Symposium on Biocomputing (PSB) (2008), pp. 477–488.

[17] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

[18] R. Bose, *Information Theory, Coding and Cryptography* (Tata, McGraw-Hill, New Delhi, India, 2002).

[19] S. Mereuta and V. Munteanu, International Symposium on Signals, Circuits and Systems (ISSCS), Vol. 2 (2007), pp.1–4.

[20] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, Phys. Rev. E **61**, 5624 (2000).

[21] A. Van Belkum, S. Scherer, L. van Alphen, and H. Verbrugh. Microbiol. Mol. Biol. Rev. **62**, 275 (1998).

[22] M. Leung, G. Marsh, and T. Speed, J. Comput. Biol. **3**, 345 (1996).

[23] G. P. Basharin, Theory Probab. Appl. **4**, 333 (1959).

[24] B. Harris, Topics Inf. Theory (Keszhtely) **16**, 323 (1975).

[25] [http://www.ensembl.org/Homo_sapiens]

[26] S.-G. Kong, W.-L. Fan, H.-D. Chen, J. Wigger, A. E. Torda, and H. C. Lee, Phys. Rev. E **79**, 061911 (2009).

[27] P. Andolfatto, Nature (London) **437**, 1149 (2005).

[28] J. W. Fickett and C.-S. Tung, Nucleic Acids Res. **20**, 6441 (1992).