Redundancy Attributes of a Complex System: Application to Bioinformatics

Perambur S. Neelakanta[*]
Tomás V. Arredondo
Dolores DeGroff
Department of Electrical Engineering,
Florida Atlantic University,
Boca Raton, Florida 33431

In general, a complex system consists of a large number of interacting units, which when viewed in an information-theoretic perspective, could be seen to possess gross redundant features. Further, a complex system is inherently stochastical in its extensive spatiotemporal universe and hence, some of its statistical features could manifest as patterns occurring more frequently than others; and, likewise some other patterns could form rare elements of the same set. In information-theoretic perspectives (in Shannon's sense) such more (or less) probabilistic occurrences of features specify the redundant (or nonredundant) aspects of the complex system.

This attribute of information redundancy in a complex system could be adopted to devise a useful complexity metric as described in this paper. The use of such a complexity metric is demonstrated in differentiating codon–noncodon domains in human and bacterial genomes (which constitute a complex system when seen at the scale of observing a DNA sequence made of vast domains of protein molecules).

1. Introduction

This paper refers to emphasizing the information-theoretic based redundant characteristics of a complex system and elucidating thereof a complexity metric in terms of the associated information redundancy. Further studied is a potential application of such a metric in ascertaining the border between the constituent subsets of a complex system by delineating the information profile across the border.

To meet the stated objectives, considered here is a random (statistical) mixture of two constituents to represent a complex system. That is, a large-sized, binary mixture is regarded as a complex system where each constituent subset (of the mixture) depicts a conglomeration of elements occurring in a proportion as decided by some probabilistic distribution. Relevant to such a complex set, an algorithm is developed to specify

[*] Electronic mail address: neelakan@fau.edu.

a metric that quantitatively evaluates the complexity (in terms of the associated information redundancy).

A typical example of a complex system depicting a binary mixture is the set of codon–noncodon constituents in a DNA structure. It is well known [1] that a strand of DNA is made of a chain constituted by four building molecules known as *nucleotides* that are linked covalently. These four nucleotides are nucleic acid bases (or *side-chains*); namely, Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). The order of these bases along a DNA strand is known as the *sequence*.

The nucleotide bases of the set {A, T, C, G} can form triplets and a DNA sequence is essentially made of two compositional domains: (*i*) The coding DNA part where the triplets (made of nucleotide bases) constitute the so-called *codons* and the codon usage is directed at encoding for a protein; and, (*ii*) noncoding (or "*junk*" codon part), which is not involved in such protein encoding functions. In species like *escherichia coli* about 90 percent of the DNA is coding by virtue of an arranged set of nucleotide bases. However, such arrangement is uncommon in *homo sapiens* because the coding content is less than 5 percent. The noncodons are considered to have no defined functions, except of some genetic relics [1], and many of their functions remain unknown. There are many regulatory functions (e.g., promoters), which are known to be located in this junk DNA part, mainly in regions that flank coding DNA. Regardless of their functional attributes, these codon and noncodon parts can cohesively be regarded as a complex system.

The occurrence frequencies of triplet contents in the coding and noncoding domains of a genomic DNA (representing a large scale of sequencing bases), constitutes the so-called *coding statistics*. The embodiments of codons and noncodons and their random occurrences conform to a statistical (binary) mixture description of the DNA structure as indicated earlier. Further, the large-sized population of codon–noncodon constituents and their interdependence characteristics with stochastical attributes render such a mixture to be aptly described as a complex system. Such complex system profiling of a DNA sequence is commensurate with the inherent stochastical attributes and the modular nature of proteins constituting the DNA sequences. More so, the complex functional attributes of a DNA structure are exhibited through their structural units or domains and lead to the consensus of classifying DNA structures as adaptive complex systems as portrayed in [2]. The justifiable considerations presented thereof can be summarized as follows.

- Existence of a large number and gross features of the DNA constituents.
- Dynamic interaction of the elements involved.
- Richly interconnected attributes of underlying units.
- Collective properties of interconnected units arising from nonlinear interactions.

- Short-range based spatial interactions between the units.
- Existence of recurrency in the interaction pathways.
- Hysteresis-specified behavior of the units.
- Stochastical nature of the system as a whole and the associated probabilistic considerations.

Inspired by these characteristics and relevant considerations of statistical mixture theory (to be elaborated later) as applied to a DNA complex, indicated in this paper is a strategy to develop a *complexity metric* that can be adopted to locate transitions between coding and noncoding regions of a DNA structure. This complexity metric is derived under the premise of statistical mixture theory and considerations of information-theoretics [3]. The approach pursued models the complex system, in general, as a binary (statistical) mixture and relevant heuristics on the associated information-theoretic considerations are elucidated in the following section. Hence, an *information redundancy* (R) factor is deduced and related to the complexity metric being defined. Use of this metric (R) is then demonstrated in delineating the boundary between the codons and noncodons of a DNA structure. The corresponding simulated results are compared against those obtained by the so-called *entropy segmentation technique* [4] (which is a traditional method used in delineation strategies of bioinformatics), and relative merits of the proposed metric are then indicated.

2. Information-theoretic framework of a complex system

Consider a complex system specified by a domain $\mathbf{X}$ as illustrated in Figure 1. Suppose two constituent (interactive) subsystems $\{x(\mu); i = 1, 2, \ldots, \lambda, \ldots, n_1\}$ and $\{x(\nu); j = 1, 2, \ldots, \lambda, \ldots, n_2\}$, are respectively characterized by two sets of attributes $\{\mu\}$ and $\{\nu\}$, where $x \in \mathbf{X}$ and $(n_1 + n_2) = N$ depicts the cardinality of the total universe of the compositional domains. Further, the occurrence probabilities of the sets $\{x(\mu); i = 1, 2, \ldots, \lambda, \ldots, n_1\}$ and $\{x(\nu); j = 1, 2, \ldots, \lambda, \ldots, n_2\}$ are $\{P_{1\lambda}\}_{\lambda=i}$ and $\{P_{2\lambda}\}_{\lambda=j}$ with the subscripts 1 and 2 depicting the attribute sets $\{\mu\}$ and $\{\nu\}$ respectively.

Suppose the randomness associated with the subsets of Figure 1 is expressed in terms of occurrence probabilities $P_{1\lambda}(\mu; i : n_1)$ and $P_{2\lambda}(\nu; j : n_2)$, corresponding to the attribute sets $\{\mu\}$ and $\{\nu\}$, respectively. Now, the maximum entropy concept [5–7] applied to each group in the domain $\mathbf{X}$ leads to the following entropy functionals:

$$H(s_i) = \ln(n_1 + 1) \approx \ln(n_1) \qquad \text{with } n_1 \gg 1 \tag{1a}$$

$$H(s_j) = \ln(n_2 + 1) \approx \ln(n_2) \qquad \text{with } n_2 \gg 1 \tag{1b}$$

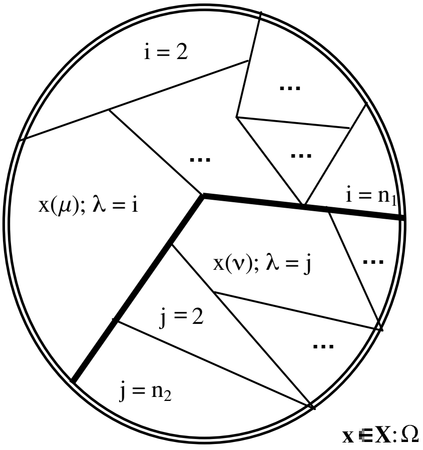

Figure 1. The test complex system $\mathbf{X} : \boldsymbol{\Omega}$ depicting a mosaic of statistical mixture with constituent binary subsystems (compositional domains).

where s_μ and s_ν refer to some metrics of gross complexity corresponding to the extensiveness of the populations of the sets $\{x(\mu); i = 1, 2, \ldots, n_1\}$ and $\{x(\nu); j = 1, 2, \ldots, n_2\}$ respectively.

With reference to a complex system viewed in an entropy-based framework, the global complexity (depicting s_μ and s_ν) has been described in [5–7] by a complexity metric (s). It is defined in terms of the associated disordered sets of constituents and corresponds to a solution equal to $\exp(-\beta)$ where β is a lagrangian that maximizes the entropy functional of the complex system. Further, considering a large set of disordered entities (constituting a complex system), s defines a dichotomy of two regimes [5]: (*i*) $0 \leq s \leq 1$ and (*ii*) $1 < s < \infty$. When s is very small ($s \to 0$), the system is regarded as "simple;" and, as $s \to \infty$, the system becomes totally complex. (The value of $s = 1$ is a transition that bifurcates the system of being simple or complex when viewed in terms of the entropy involved.)

Equation (1) is consistent with the so-called Jaynes' principle of maximum entropy or maximum uncertainty and a class of distribution corresponding to the maximum entropy formalism has been identified in [8] to exist. Further, equation (1) concurrently leads to the following Shannon information formulations [3]:

$$I_1 = \{x(\mu); i = 1, 2, \ldots n_1\} = -\sum_{x \in \mathbf{X}; \lambda = i} P_{1\lambda} \ln(P_{1\lambda}) \tag{1c}$$

$$I_2 = \{x(\nu); j = 1, 2, \ldots n_2\} = -\sum_{x \in \mathbf{X}; \lambda = j} P_{2\lambda} \ln(P_{2\lambda}) \tag{1d}$$

Equations (1c) and (1d) can be regarded as implicit representations of gross complexity pertinent to the sets $\{x(\mu); i = 1, 2, \ldots, n_1\}$ and $\{x(v); j = 1, 2, \ldots, n_2\}$ respectively, *in lieu* of the relations specified by equations (1a) and (1b). While equation (1) depicts the maximum entropy measuring the gross complexity (s) of the set $\{x(\mu); i\}$ or $\{x(v); j\}$, an alternative metric can also be specified to measure the relative complexity between these sets. It refers to a pair of cross-entropy functionals, which can be written in the following forms [7]:

$$H(s_\mu || s_v) = D(s_\mu || s_v) = \sum_{\lambda=i} P_{1\lambda} \ln\left(\frac{P_{1\lambda}}{P_{2\lambda}}\right) \tag{2}$$

$$H(s_v || s_\mu) = D(s_v || s_\mu = \sum_{\lambda=j} P_{2\lambda} \ln\left(\frac{P_{2\lambda}}{P_{1\lambda}}\right). \tag{3}$$

The cross-entropy functionals of equations (2) and (3) denote synonymously the "statistical divergence" $D(s_\mu || s_v)$ between the random attributes of $\{x(i) : \mu\}$ *versus* $\{x(j) : v\}$, or *vice versa*. This cross-entropy measure also refers to relative or mutual information content in Shannon's sense [3]. Further, the measure specified *via* equations (2) and (3) follows Kullback's minimum (directed) divergence or minimum cross-entropy principle [9].

The cross-entropy concept of depicting the relative complexity as above, also implicitly implies an expected logarithm of the likelihood ratio (L), namely,

$$L = \frac{[(P_{1\lambda})_{\lambda=1}]_{\mathrm{I}}}{[(P_{2\lambda})_{\lambda=j}]_{\mathrm{II}}} \tag{4}$$

where $[P_{1\lambda}]_{\mathrm{I}}$ and $[P_{2\lambda}]_{\mathrm{II}}$ are respective probabilities of observations of the attributes $\{\mu\}$ and $\{v\}$ in the complex system when a certain hypothesis $(h_{\mathrm{I}}, h_{\mathrm{II}})$ is true. Corresponding to L, the log-likelihood ratio function (LLR) given by $\ln(L)$, can be regarded as a "discrimination measure" that provides a choice, whether to choose $\{\mu\}$ in preference to $\{v\}$ or *vice versa*. The LLR is well known [7] as a useful metric in decision-making efforts and can be considered identically to depict a measure of contrast between the constituents involved.

Designated as the Jensen–Shannon measure (JS-measure) [10], is a variation of the Kullback–Leibler divergence formulation that is explicitly given by the following expression:

$$\mathrm{JS}_\pi(P_{1\lambda}, P_{2\lambda}) = H(\pi_1 P_{1\lambda} + \pi_2 P_{2\lambda}) - \pi_1 H(P_{1\lambda}) - \pi_2 H(P_{2\lambda}) \tag{5a}$$

where $(\pi_1, \pi_2) \geq 0$ and $(\pi_1 + \pi_2) = 1$; and,

$$H(P_{1\lambda}) = -\sum_{\lambda=i} P_{1\lambda} \ln(P_{1\lambda}) \tag{5b}$$

$$H(P_{2\lambda}) = -\sum_{\lambda=j} P_{21\lambda} \ln(P_{2\lambda}) \tag{5c}$$

$$H(\pi_1 P_{1\lambda} + \pi_2 P_{2\lambda}) = -\sum_{\mu=i} \sum_{\nu=j} (\pi_1 P_{1\mu} + \pi_2 P_{2\mu}) \ln(\pi_1 P_{1\nu} + \pi_2 P_{2\nu}). \tag{5d}$$

The weights π_1 and π_2 for example, can be taken respectively as $\theta = n_1/(n_1 + n_2)$ and $(1 - \theta) = n_2/(n_1 + n_2)$ in the context of a mixture complex.

3. Statistical mixture complex: Measure of complexity

Exclusive to a statistical mixture under discussion, a measure of global complexity can be specified in terms of the maximum entropy associated with the disordered constituent entities (assuming each has a large population namely, (n_1 and n_2) → ∞, and are mixed in a specified proportion). Relevant considerations are discussed below.

Following the concept of statistical mixture theory due to Lichtenecker and Rother [11], the underlying heuristics specifies a weighted probability r that describes the effective statistical attribute of the mixture proportioned by the attributes $\{\mu\}$ and $\{\nu\}$. This weighted probability is given by:

$$r(\theta) = P_{1\lambda}^{\theta} P_{2\lambda}^{1-\theta} \tag{6}$$

which is valid, however, within the statistical upper and lower bounds, namely, $r_{\min} \leq r \leq r_{\max}$. Explicitly, $r_{\min}$ and $r_{\max}$ are given by:

$$r_{\min} = \left[\frac{\theta}{P_{1\lambda}} + \frac{(1-\theta)}{P_{2\lambda}} \right]^{-1} \tag{7a}$$

$$r_{\max} = \theta P_{1\lambda} + (1-\theta) P_{2\lambda}. \tag{7b}$$

With reference to the set $\{r : r_{\min}$ and $r_{\max}\}$, the corresponding Shannon measure of entropy (negentropy) can be written as follows:

$$H(r) = -r \ln(r) \tag{8a}$$

$$H(r_{\min}) = -r_{\min} \ln(r_{\min}) \tag{8b}$$

$$H(r_{\max}) = -r_{\max} \ln(r_{\max}) \tag{8c}$$

Suppose one of the constituent entities of the statistical mixture; say, the one with a population n_2, has a uniform distribution implying that the occurrences of its elements (in the statistical mixture space) are equally likely. That is, ($P_{21} = P_{22} = P_{23} = \cdots = P_{2n_2} = 1/n_2$); and, $[P_{21} + P_{22} + P_{23} + \cdots + P_{2n_2} = 1]$. In contrast, the other constitutive entity (with a population n_1) is presumed to be of elements each bearing a distinct probability of occurrence. That is, ($P_{11} \neq P_{12} \neq P_{13} \neq \cdots \neq P_{1n_1}$); and, $[P_{11} + P_{12} + P_{13} + \cdots + P_{1n_1} = 1]$.

Considerations on equally-likely occurrences as above are true, for example, in the case of noncodons coexisting with codons in a mixed state within a DNA system [4]. Inasmuch as the functions of the noncodons in the DNA structure are not defined, the presence of noncodons denotes a state of maximum entropy resulting from a uniformly distributed statistics. On the contrary, pertinent to each protein-encoding codon, prevails a distinct (unequal) occurrence probability (as decided by the designated encoding function). Hence, codons are an informative (negentropy) part of the DNA sequence.

From an information theory point of view, it is known that equally-likely occurrences of entities or events of a set of random variates mean a degree of certainty, whereas random (unequal) chances of occurrences imply an associated uncertainty of the set. The certainty consideration will bring down the negative entropy (or information content) of the set while any uncertainty involved will augment the negative entropy.

Thus, the existence of equally-likely probabilities associated with the elements (such as junk codons) of a set amounts to an efficiency of the associated information content of the whole set; and, relevant consideration leads to a redundant information-theoretic attribute [3] to the set under discussion (in Shannon's sense). Such a measure of redundancy (R) can be specified with reference to a statistical mixture as indicated below (relevant details are furnished in the Appendix):

$$R = 1 - \frac{H(r)}{[H(r)]_{\mathrm{M}}} \tag{9a}$$

where $[H(r)]_{\mathrm{M}}$ refers to the maximum value of $H(r)$ over the fraction $0 \leq \theta \leq 1$ (or $1 \geq (1-\theta) \geq 0$) of the binary mixture constituents. Referring to the upper and lower bounds on r specified by equation (7), the corresponding range of R can be deduced as follows:

$$R_{\min} = 1 - \frac{H(r_{\max})}{[H(r_{\max})]_{\mathrm{M}}} \tag{9b}$$

$$R_{\max} = 1 - \frac{H(r_{\min})}{[H(r_{\min})]_{\mathrm{M}}}. \tag{9c}$$

The complexity metric of a statistical measure evaluated in terms of the redundancy measure (R) as above can be used to delineate the binary constituents of such mixtures. That is, considering a heterogeneous DNA sequence of codons and noncodons, the borders between the compositional (codon–noncodon) domains can be distinctly identified using the parameter R. For this purpose, additional details on codon–noncodon population mix and the related coding statistics considerations are furnished in the following section.

4. DNA coding statistics and codon–noncodon delineation

As mentioned earlier, the concept of DNA coding statistics envisaged in bioinformatics allows assaying the likelihood that a given DNA sequence is coding for a protein. Considering the set of nucleotide bases {A, T, C, G} forming a set of 64 triplets, the coding statistics ascertains whether a portion of the DNA sequence is a codon (namely, an identifiable triplet that is part of the coding for a protein) or a noncodon (namely, the junk parts of the triplets not involved in coding for a protein). In essence, the models developed towards such coding statistics enable discriminating a coding from a noncoding DNA sequence. Knowledge of relevant coding statistics is important to develop gene identification programs of bioinformatics and interpret their predictions [1].

The models of coding DNA are based on stochastical considerations depicting the occurrence population of codon and noncodon parts. Traditional measures to score the coding statistics (in a query sequence) are based on metrics such as the LLR mentioned before. Alternatively, as described in [4], entropy-based considerations can also be used to segment and discriminate the codon and noncodon regions in a query sequence. Essentially, the entropy segmentation technique uses the Jensen–Shannon form of cross-entropy measure of equation (5) for the purpose of ascertaining the border between codon and noncodon regions (consistent with the relative proportions of the compositional domains in the DNA sequence). In the state-of-the-art advances concerning gene recognition efforts, especially in prokaryotes, relevant solutions have been obtained *via* a comprehensive set of statistical measures yielding fair results. Nevertheless, contemporary methods and new algorithms are constantly being proposed as indicated recently [4] for finding the borders between coding and noncoding DNA regions by an entropic segmentation method. Such methods are aimed at developing algorithms depicting better and "sensitive scoring schemes," especially when one of the entities (such as codons in a mixed population of coding and noncoding parts) constitute a sparse percentage. The present study leads to an algorithm which is more sensitive than that of [4] as will be demonstrated *via* computed results.

Hence, *in lieu* of the traditional codon–noncodon delineation technique of [4], proposed here is the information-redundancy based complexity metric (R) of equation (6) to score the statistics of codon–noncodon occurrences. And, the efficacy of this R-measure as a discrimination function (for codon–noncodon delineation) is analyzed relative to other measures such as the entropy segmentation method [4].

5. Complexity-based metric for codon–noncodon delineation

The existing method for ascertaining the borders across coding–noncoding DNA regions; namely, the entropic segmentation technique

[4], introduces the concept of a *contrast function* (*via* the JS-measure) for estimating the difference in composition in question between the two regions. That is, a comparative function based on the JS-measure is indicated in [4] that reaches low values when the DNA regions being compared (or contrasted) are similar, namely codons–codons or noncodons-noncodons; and, this comparative function attains a large value when the compared entities are dissimilar, namely codon–noncodon or noncodon–codon.

The JS-measure adopted in [4] is based on the Kullback–Leibler (KL-measure) concept of cross-entropy (or mutual information) arising from divergence in the statistical attributes of the two regions being compared [9]. The KL-measure essentially compares two vector spaces $\mathbf{V}_{c,\lambda}$ and $\mathbf{V}_{nc,\lambda}$ corresponding to the codon and noncodon regions respectively, so as to elucidate the divergence in the associated stochastical characteristics. Each of these vector spaces are constituted by the triplets composed of the nucleotide set $\lambda \in \{A, T, C, G\}$.

In general, the codon and noncodon regions being compared can be regarded as parts of a total sequence length N as considered earlier. Designating the codon region by subscript 1 and the noncodon region by subscript 2, the fractional regions being compared are proportioned as n_1/N and n_2/N respectively; so that, $(n_1 + n_2) = N$ constitutes the universe of statistical mixture of codons and noncodons. Each of the triplets or phases of the nucleotide set $\lambda \in \{A, T, C, G\}$ occurs with a particular probability in a given genome type. For the regions 1 and 2, these probabilities can be specified as $P_{1,\lambda}$ and $P_{2,\lambda}$ respectively. For example, with reference to human genes, the set of probabilities $P_{1,\lambda}$ are specified in terms of relative frequencies of occurrence of the codon triplets as listed in Table 1 [12].

Codon triplet	Relative frequency	Codon triplet	Relative frequency	Codon triplet	Relative frequency	Codon triplet	Relative frequency
GGG	.01645	AGG	.01162	TGG	.01296	CGG	.01175
GGA	.01636	AGA	.01166	TGA	.00137	CGA	.00631
GGT	.01081	AGT	.01201	TGT	.01012	CGT	.00468
GGC	.02260	AGC	.01937	TGC	.01236	CGC	.01081
GAG	.04033	AAG	.03255	TAG	.00059	CAG	.03444
GAA	.02895	AAA	.02399	TAA	.00077	CAA	.01190
GAT	.02208	AAT	.01666	TAT	.01207	CAT	.01057
GAC	.02571	AAC	.01927	TAC	.01548	CAC	.01503
GTG	.02873	ATG	.02231	TTG	.01267	CTG	.04013
GTA	.00703	ATA	.00718	TTA	.00733	CTA	.00701
GTT	.01093	ATT	.01579	TTT	.01707	CTT	.01293
GTC	.01460	ATC	.02128	TTC	.02043	CTC	.01954
GCG	.00756	ACG	.00617	TCG	.00448	CCG	.00702
GCA	.01598	ACA	.01485	TCA	.01187	CCA	.01675
GCT	.01865	ACT	.01294	TCT	.01475	CCT	.01734
GCC	.02839	ACC	.01912	TCC	.01753	CCC	.02003

Table 1. Relative usage frequencies of triplets in human codons [12].

With reference to Table 1, it is implied that the statistical feature of coding regions refers to a nonuniform codon usage. That is, inside the coding regions, not all triplets of nucleotides (namely, the codons) occur with the same probability; also, the probability of the appearance of a nucleotide is different in each of the three positions of the triplets. The reason for this has been identified as possible restrictions imposed by the genetic code and also probably due to preferential (synonymous) codon usages. Regardless of the causative mechanism involved, this attribute of nonuniform probability distribution of codons does not prevail in the so-called noncoding or junk codon part of the DNA. This distinguishing feature (of probabilities of occurrence) leads to identifying codon and noncodon regions in a DNA sequence; and, the entropic segmentation method of [4], in essence, prescribes a strategy of delineating such codon and noncodon regions using the aforesaid statistical distinction expressed in terms of entropy considerations. Further, the entropy segmentation efforts as mentioned before essentially conform to divergence measures such as the JS-measure. It partitions a heterogeneous DNA sequence into homogenous subsequences (or compositional domains); and, it is shown that the relevant approach could lead to predicting accurately the borders between coding and noncoding regions without any *a priori* training details on known sets.

As an alternative to the entropy segmentation algorithm, proposed here to discern the codon–noncodon borders is the R-measure, which accounts for the informative profile of the constituents of a DNA structure. Such constituents are vast when viewed in terms of their attributes, variety, stochasticity, and interactions between the associated units; and hence, they represent the gamut of informative complex systems deliberated earlier; and the R-metric assesses the demarcation (between the codon and noncodon regions). That is, the R-measure is an alternative extensive measure towards entropic segmentation viewed in the information-theoretic framework. The R-metric is also closely related to entropy, but in terms of the information content of the compositional domains identified through the redundant features in the complex structure of the DNA system. In the following section, computations using this redundancy measure (R) (as well as the JS-measure depicting the entropic segmentation metric [4]) in delineating the codon and noncodon compositional domains of DNA sequences are presented and compared.

6. Computed results

The R- or JS-measure provides a contrast function to distinguish coding and noncoding DNA composition. The procedure for the delineation as adopted in [4], refers to a controlled experiment in which, first a known set of coding and noncoding DNA sequences is taken and these sequences are then concatenated. Next, a pointer is used along the

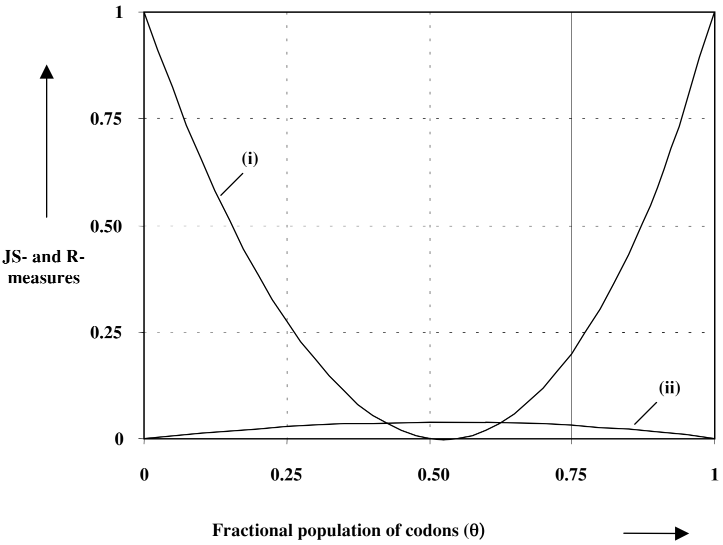

Figure 2(a). The JS- and R-measure *versus* fractional populations of *human* codons (statistical data is pertinent to human codon usage of Table 1). (*i*) JS-measure of equation (5) specified by $[\mathrm{JS}(r_1, r_2)]$ with $\pi_1 = \theta$ and $\pi_2 = (1 - \theta)$, $r_1 = \theta P_{1\lambda}$ and $r_2 = (1-\theta)P_{2\lambda}$. (*ii*) Corresponding R-measure ascertained in terms of equation (9a) with $[H(r_1, r_2)]_{\max} \equiv [\mathrm{JS}(r_1, r_2)]$.

concatenated sequence to bifurcate (the sequence) into two sections; and the resulting (bifurcated) subsequences are contrasted *via* the JS-measure in terms of the statistics of codons and noncodons. That is, in computing the JS-measure, the probabilities $P_{1\lambda}$ (of Table 1) are used for codons and the probabilities $P_{2\lambda}$ (each equal to 1/64) are used for noncodons. A similar procedure can also be adopted with the R-measure.

The resulting graphs of the R- and JS-measures *versus* the fractional population ($0 \leq \theta \leq 1$) of human codons used in a concatenated test sequence are illustrated in Figure 2(a) where the redundancy measure (R) is computed as per equation (9c), namely,

$$R = 1 - \frac{H(r_1, r_2)_{\max}}{[H(r_1, r_2)_{\max}]_{\mathrm{M}}}$$

with

$$\begin{aligned} \pi_1 &= \theta, \\ \pi_2 &= (1 - \theta), \\ r_1 &= \theta P_{1\lambda}, \\ r_2 &= (1 - \theta)P_{2\lambda} \\ H(r_1, r_2)]_{\max} &\equiv [\mathrm{JS}(r_1, r_2)]. \end{aligned}$$

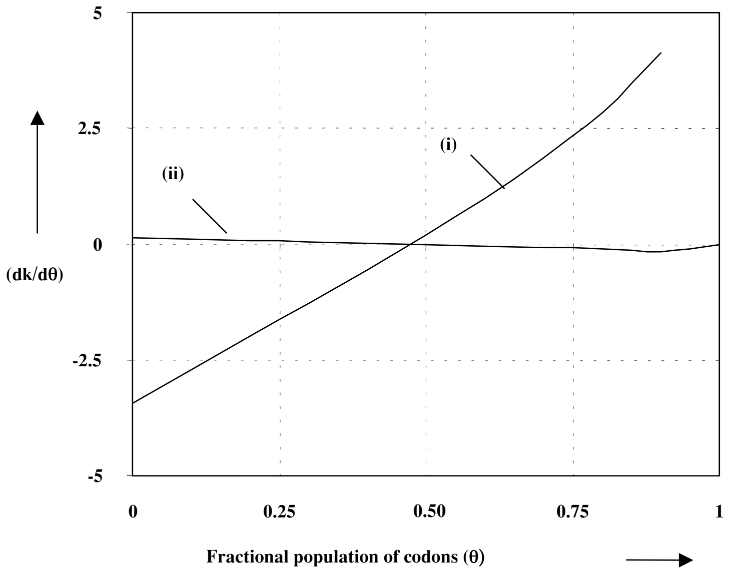

Figure 2(b). The slopes $dk/d\theta$ (k: JS- or R-measure) of Figure 2(a) *versus* θ. (i): $[dk/d\theta]_{k=\text{JS-measure}}$ and (ii): $[dk/d\theta]_{k=R\text{-measure}}$.

Using the computed results on JS- and R-measures as above, the corresponding slopes of ($dk/d\theta$) *versus* θ (where k represents JS- or R-values) are computed for *human* codons. They are plotted in Figure 2(b).

The computational details as above are also extended to three bacterial genomes, namely: *escherichia coli*, *rickettsia prowzekii*, and *methanococcus jannaschii* [12]. The results on k *versus* θ and ($dk/d\theta$) *versus* θ on the codon data pertinent to these bacterial genomes are presented in Figures 3 to 5. The data on relative frequency of occurrences of the codons of the three bacteria types indicated are available in [12].

7. Discussions on results

In the computational recognition of codon to noncodon borders, it is necessary to consider how effectively the algorithm used enables such recognitions or differentiation, even when there is a subtle statistical difference in their relative populations. In prokaryotic genomes, for example, the coding regions may be separated by a very small noncoding region, sometimes too small for distinct identification of the borders on a statistical basis.

It is therefore preferable to have a metric that yields a significant measure of contrast even for small fractions of codon populations (relative to noncodon populations). In other words, the efficacy of the algorithm used for distinguishing codon–noncodon regions can be specified by the

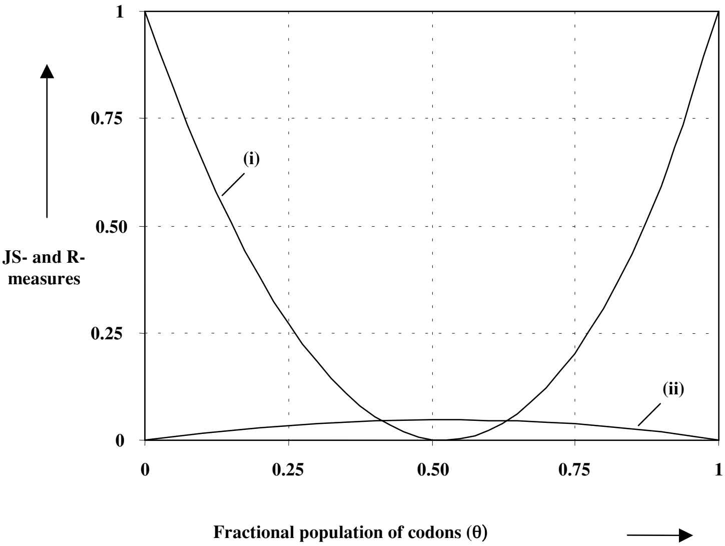

Figure 3(a). The JS- and R-measure *versus* fractional populations of *escherichia coli* codons. (*i*) JS-measure of equation (5) specified by $[\mathrm{JS}(r_1, r_2)]$ with $\pi_1 = \theta$ and $\pi_2 = (1 - \theta)$, $r_1 = \theta P_{1\lambda}$ and $r_2 = (1 - \theta)P_{2\lambda}$. (*ii*) Corresponding R-measure ascertained in terms of equation (9a) with $[H(r_1, r_2)]_{\max} \equiv [\mathrm{JS}(r_1, r_2)]$.

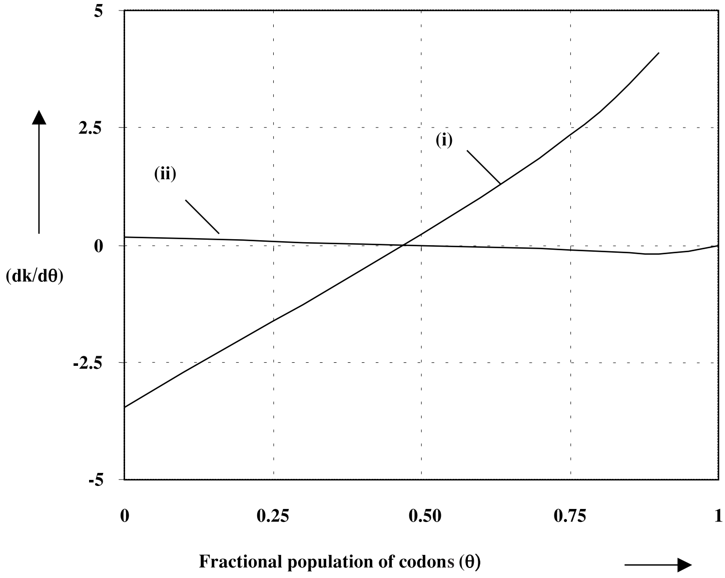

Figure 3(b). The slopes, $dk/d\theta$ (k: JS- or R-measure) of Figure 3(a) *versus* θ. (*i*): $[dk/d\theta]_{k=\text{JS-measure}}$ and (*ii*): $[dk/d\theta]_{k=R\text{-measure}}$.

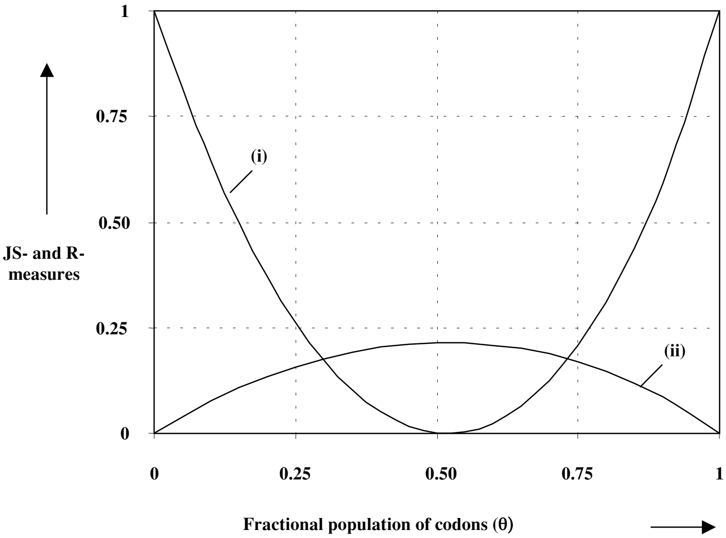

Figure 4(a). The JS- and R-measure *versus* fractional populations of *rickettsia prowzekii* codons. (*i*) JS-measure of equation (5) specified by $[\mathrm{JS}(r_1, r_2)]$ with $\pi_1 = \theta$ and $\pi_2 = (1 - \theta)$, $r_1 = \theta P_{1\lambda}$ and $r_2 = (1 - \theta)P_{2\lambda}$. (*ii*) Corresponding R-measure ascertained in terms of equation (9a) with $[H(r_1, r_2)]_{\max} \equiv [\mathrm{JS}(r_1, r_2)]$.

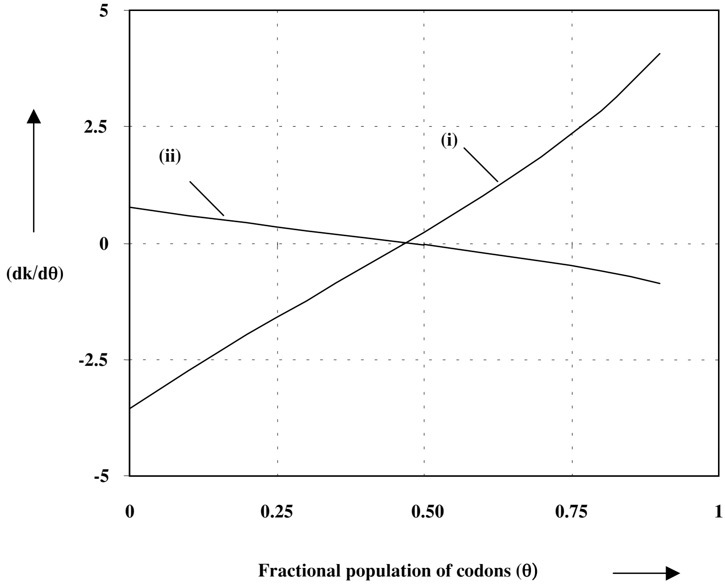

Figure 4(b). The slopes, $dk/d\theta$ (k: JS- or R-measure) of Figure 4(a) *versus* θ. (*i*): $[dk/d\theta]_{k=\text{JS-measure}}$ and (*ii*): $[dk/d\theta]_{k=R\text{-measure}}$.

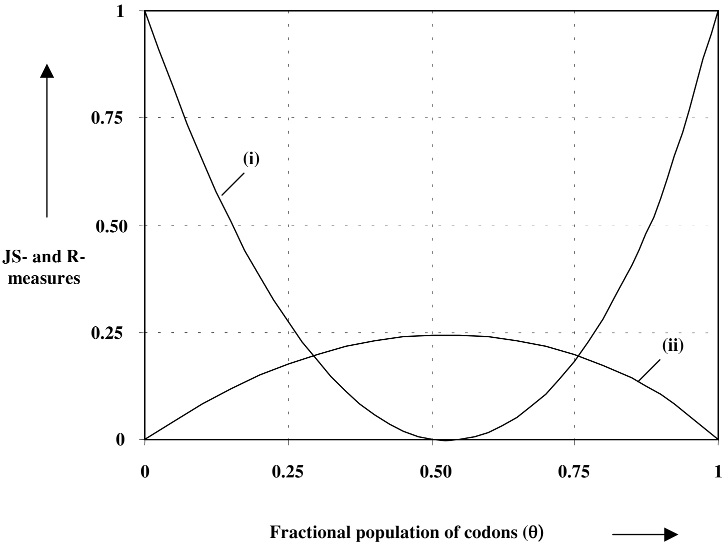

Figure 5(a). The JS- and R-measure *versus* fractional populations of *methanococcus jannaschii* codons. (i) JS-measure of equation (5) specified by $[\mathrm{JS}(r_1, r_2)]$ with $\pi_1 = \theta$ and $\pi_2 = (1 - \theta)$, $r_1 = \theta P_{1\lambda}$ and $r_2 = (1 - \theta)P_{2\lambda}$. ($ii$) Corresponding R-measure ascertained in terms of equation (9a) with $[H(r_1, r_2)]_{\max} \equiv [\mathrm{JS}(r_1, r_2)]$.

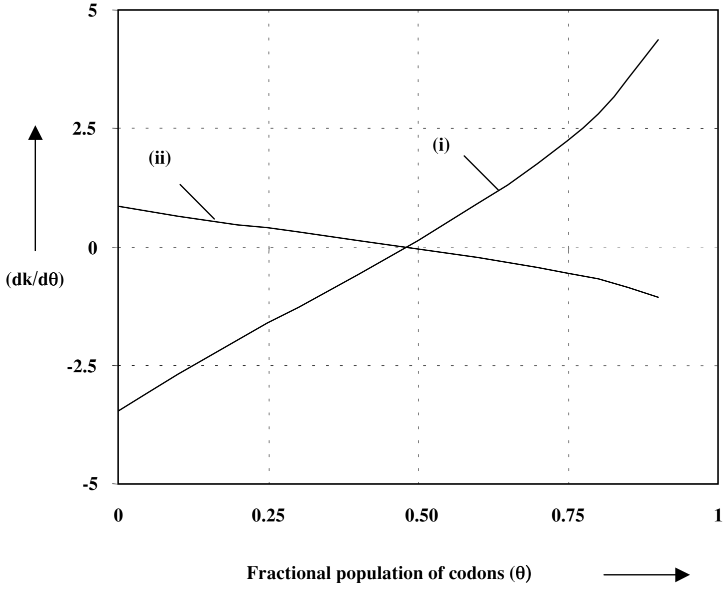

Figure 5(b). The slopes, $dk/d\theta$ (k: JS- or R-measure) of Figure 4(a) *versus* θ. (i): $[dk/d\theta]_{k=\text{JS-measure}}$ and (ii): $[dk/d\theta]_{k=R\text{-measure}}$.

slope ($dk/d\theta$) where k denotes the computed value (JS- or R-value in Figures 2–5) of the metric used.

From Figures 2–5, it is obvious that the initial slope $\{dk/d\theta|_{k=R;\theta\to 0}\}$ is much larger than the initial slope $\{dk/d\theta|_{k=\mathrm{JS},\theta\to 0}\}$. That is, the R-measure deliberated in this study offers a better sensitivity in discerning the codon–noncodon compositional domains at low values of θ, (i.e., when $\theta \to 0$) as compared to the entropy segmentation method (using the JS-measure) proposed in [4].

8. Closure

In conclusion, the present study indicates that a complex system can be modeled in terms of the associated information redundancy. Hence an information-theoretic metric (R-measure) can be elucidated to specify the complexity involved. It is further shown that this R-measure can be determined in terms of the statistical divergence, such as the Jensen–Shannon measure (JS-measure). An application of the R-measure is demonstrated in delineating the codon–noncodon regimes of a DNA structure. The efficacy of the R-measure discerning such a border across the codon-to-noncodon domains is compared with the results obtained *via* the entropy segmentation technique [4], especially when low population fractions of codons are present in the DNA complex.

Instead of the JS-measure specified for the entropy segmentation technique in [4] and adopted for the R-measure in the present study, a host of other divergence measures (called Csiszár measures [6, 13, 14]) can also be used as metrics for the redundancy measure (and adopted to contrast the compositional domains of codons and noncodons). Similarly, the family of so-called *distance measures* [15–17] can be probed for the same purpose. Relevant investigations, however, are still open questions.

Appendix

A note on information redundancy in a DNA sequence

In a DNA sequence, the inherent mapping of codons corresponds to a set of encoded messages constituted by the codons composed of triplets from the set $\{A, T, C, G\}$. Each such codon occurs at a specified probability in the sequence in conformance with the formation of the encoded structure. Pertinent to the statistics of the codon part, one can associate an efficiency factor, which should enable an implicit optimization of a cost function in constituting the DNA chain. The underlying considerations follow.

Suppose a constant c_i is assigned as a cost figure to each codon whose occurrence probability is $P_{1\lambda}$. Then, the average cost per codon can be written as follows:

$$C_{\text{av}} = \sum_{i=1,N=64} P_{1i} c_i. \tag{A.1}$$

The optimization of the cost function (in the constitution of the codons in the layout in the DNA sequence) would refer to a value of C_{av} equal to $\overline{\text{C}}$ and subject to certain restrictions on the encoding rule, the lowest bound (*infimum*) of $\overline{\text{C}}$ is given by,

$$\overline{\text{C}}|_{\text{inf}} = \frac{H(x)}{\ln(N = 64)} \tag{A.2}$$

where $H(x)$ is the entropy of the ensemble of codons and $N = 64$ is the cardinality of the encoding codons.

The constrained optimization exercised on the DNA sequence relevant to the codon layouts as above can be expanded in its scope by taking into account the junk codons (namely, the noncodon part does not contribute toward encoding for the proteins). That is, inasmuch as the codons and noncodons prevail in a DNA sequence complex as mixture constituents, the entropy of the ensemble of the mixture should be viewed in terms of redundancy arising from the noncodon population (that contributes negentropy rather than posentropy to the system).

Hence, it is possible to define an *information efficiency* (η) factor of the encoded structure of a DNA structure using the classical concepts of information theory. It is the ratio of the average information (per codon) of the encoded ensemble to the maximum possible (average) information (per codon). That is,

$$\eta = \frac{H(x)}{\overline{\text{C}}\ln(N = 64)}. \tag{A.3}$$

And concurrently, $(1 - \eta)$ can be regarded as the *redundancy factor* (R) indicated in equation (9a). It is the reduction in information content of an ensemble from the maximum possible and is specified as:

$$R = 1 - \frac{H}{H_{\text{max}}} \tag{A.4}$$

where H is an entropy functional such as the Jensen–Shannon measure (JS-measure). In [4], the JS-measure is adapted (*in lieu* of the R-measure proposed in this study) to ascertain the delineation information on random codon–noncodon assembly across the DNA complex. The JS-measure essentially assays the statistical divergence between the codon and noncodon parts using the cross-entropy (or mutual information) considerations. The ratio $\text{JS}/\text{JS}_{\text{max}}$ used in equation (A.4) identically represents a redundancy based information efficiency factor η defined above. Explicitly, the JS-measure is given by equation (5).

References

[1] J. M. Claverie and C. Notredame, *Bioinformatics for Dummies* (Wiley Publishing, Inc., New York, NY, 2003).

[2] S. B. Nagl, "Protein Evolution as a Parallel-distributed Process: A Novel Approach to Evolutionary Modeling and Protein Design," *Complex Systems*, **12** (2000) 261–280.

[3] T. M. Cover and J. Thomas, *Elements of Information Theory* (John Wiley and Sons Inc., New York, NY, 1991).

[4] P. B. Galvan, I. Grosse, P. Carpena, J. L. Oliver, R. R. Roldán, and H. E. Stanley, "Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method," *Physical Review Letters*, **85** (2000) 1342–1345.

[5] A. E. Ferdinand, "The Theory of System Complexity," *International Journal of General Systems*, **1** (1994) 19–23.

[6] P. S. Neelakanta, *Information Theoretic Aspects of Neural Networks* (CRC Press, Boca Raton, FL, 1999).

[7] R. M. Bendett and P. S. Neelakanta, "A Relative Complexity Metric for Decision-theoretic Applications in Complex Systems," *Complex Systems*, **12** (2000) 281–295.

[8] E. Jaynes, "On the Rationale of Maximum Entropy Methods," *Proceedings of IEEE*, **70** (1982) 939–952.

[9] S. Kullback, *Information Theory and Statistics* (Wiley Interscience Publications, New York, NY, 1959).

[10] J. Lin, "Divergence Measures Based on the Shannon Entropy," *IEEE Transactions on Information Theory*, **37** (1991) 145–151.

[11] P. S. Neelakanta, "Permittivity of Dielectric-conductor Mixture: Application of Logarithmic Law of Mixing to Electric Susceptibility," *Electronics Letters*, **25** (1989) 800–802.

[12] Website: http://www.kazusa.or.jp/codon/.

[13] P. S. Neelakanta, S. Abusalah, D. De Groff, R. Sudhakar, and J. C. Park, "Csiszár's Generalized Error Measures for Gradient-descent-based Optimizations in Neural Networks Using the Backpropagation Algorithm," *Connection Science*, **8** (1996) 79–114.

[14] I. Csiszár, "A Class of Measures of Informativity of Observation Channels," *Periodica Mathematica Hungarica*, **2** (1972) 191–213.

[15] S. M. Ali and S. D. Silvey, "A General Class of Coefficients of Divergence of One Distribution from Another," *Journal of Royal Statistical Society* (Series B), **28** (1996) 131–142.

[16] T. Kailath, "The Divergence of Bhattacharyya Distance Measures in Signal Detection," *IEEE Transactions on Communication Technology*, **COM-15** (1967) 52–60.

[17] P. C. Mahalanobis, "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Science* (India), **12** (1936) 49–55.