Artificial Intelligence in Medicine (2005) **35**, 87—105

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Fuzzy attributes of a DNA complex: Development of a fuzzy inference engine for codon-"junk" codon delineation

Tomás V. Arredondo [a], **Perambur S. Neelakanta** [b,*], **Dolores De Groff** [b]

[a] *Departamento de Electrónica, Universidad Técnica Federico Santa María, Valparaíso, Chile*
[b] *Department of Electrical Engineering, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA*

Received 29 October 2004; received in revised form 21 January 2005; accepted 22 February 2005

KEYWORDS
Fuzzy inference engine;
Fuzzy DNA complex;
Codon—noncodon delineation;
Information-theoretic metrics;
Scoring in fuzzy domains

Summary

Objective: The present study is concerned with the need that exists in bioinformatics to identify and delineate overlapping codon and noncodon structures in a deoxyribonucleic acid (DNA) complex so as to ascertain the boundary of separation between them. Codons refer to those parts in a DNA complex encoded towards forming a desired set of proteins. Also coexist in the DNA structure noncodons (or "junk" codons), whose functions are not so well defined. Such codon and noncodon parts (at least over some sections of a DNA chain) may conform to diffused (overlapping) states exhibiting sharpless boundaries with indistinctive statistics of occurrence of their constituents. Such overlapping mix of codon and noncodon entities constitutes a (fuzzy) universe with information constituent having a fuzzy structure, which can only be identified in descriptive norms with characteristic membership of belongingness to certain attributes. Hence, this work is directed to develop a fuzzy inference engine (FIE), which delineates the fuzzy codon—noncodon parts.
Methods and material: Relevant algorithms developed for the fuzzy inference in question are based on information-theoretic (IT) considerations applied to symbolic as well as binary sequence data representing the DNA. Pseudocodes, as needed are furnished.
Results: Simulated studies using human and other bacterial codon statistics are presented to illustrate the efficacy of the approach pursued. The outcome of the study is illustrated via tabulated results and graphs depicting the delineation sought.
Conclusion: The results signify the success of IT-approach pursued in delineating imprecise codon/noncodon boundaries. The FIE applies both for human and bacterial codon statistics.

* Corresponding author. Tel.: +1 561 297 3469; fax: +1 561 297 2336.
E-mail address: neelakan@fau.edu (P.S. Neelakanta).

0933-3657/$ — see front matter
doi:10.1016/j.artmed.2005.02.008

1. Introduction

The deoxyribonucleic acid (DNA) utters the language of life systems profiled by the essence biochemistry as elaborated in the topics of molecular biology. 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, namely, adenine (A), guanine (G), cytosine (C) and thymine (T). The hereditary instructions are written in this set of four alphabets {A, G, C, T}. A DNA in essence, represents a chain of these bases in the form of two-stranded double helix conforming to a chemical fitting of A pairing with T and G with C. The order of the bases along a DNA strand is known as the sequence [1–3].

The nucleotide bases of the set {A, T, C, G} form triplets, also known as tri-nucleotides. A DNA sequence is essentially made of two compositional set of such triplets: (i) The coding DNA part where the triplets constitute the so-called codons and the codon usage is directed at encoding for a protein. These proteins are responsible in driving the enzymatic machinery of living organisms. (ii) The noncoding (or ''junk'' codon part) in the DNA is not involved in such protein encoding functions. Although considered to have no defined functions except of some genetic relics [4–7], many of noncodon functions still remain unknown. (There are however, many regulatory functions (e.g. promoters), which are known to be located in this ''junk'' DNA part mainly in regions flanking coding DNA).

The occurrence frequencies of triplet contents in the coding and noncoding domains of a genomic DNA (representing a large scale of sequencing bases), constitute the so-called coding statistics [8]. Further, the embodiment of codons and noncodons and their random occurrences conform to a statistical (binary) mixture description of a DNA structure; and, the massive population-size of codon/noncodon constituents in living systems (regardless of their functional attributes) and their interdependence characteristics render such a mixture to be aptly described as a complex system [9].

Associated with the complexity of molecular biology is a rich profile of information in vivo; and, the art of bioinformatics in general, refers to collecting, organizing, retrieving and analyzing such information-bearing data set. However, rigorous use of information-theoretics (IT) and related measures to handle and assay bioinformatic entities is still in infantile stage [10]. The heuristics of information in vivo form the core of bioinformatic efforts [5,6,11].

Viewed in its complex system profile, a DNA sequence can be mostly described only in an ''approximate way'' in terms of its constituents and their characteristics. This ''approximate nature'' and the associated subjectiveness as well as the imprecise queries on the data sequence characteristics and the expression profiles make the associated data-mining efforts in bioinformatics (dealing with massive DNA information) a difficult task vis-à-vis the underlying fuzzy (mining) space. Concurrently, when observed in the IT-plane, the DNA complex reflects a profile of information complexity with fuzzy attributes [12–14].

The fuzzy consideration in mining a large sequence data in general, has real-world implications. Specific to bioinformatics [6,11,15], – a discipline that applies computer technology to bioengineering and helps managing massive streams of data of molecular sequence information for diagnostic and therapeutic applications – the data-mining applications involving fuzzy analysis space [12–14] warrant computationally flexible as well viably tractable efforts that are fast in identifying the consensus patterns in a vast DNA sequence [3].

Fuzzy logic methods have been only sparingly used in bioinformatics. For example, in [13] Chang and Halgamuge present a technique to extract the protein motif (defined as a signature or consensus pattern) from sequences of the same family using neuro-fuzzy optimization. Another study by Tomida et al. [14] offers an analysis of expression profile using fuzzy adaptive resonance theory applied towards clustering of expression data. An algorithm for fuzzy sequence pattern-matching (in zinc finger domain) proteins is described in [12]. But considering codon–noncodon delineation, the methods available as in [16]) mostly concern with the crisp border of separation and no fuzzy details are addressed.

The scope of the present study is therefore, focused on relevant issues of dealing with such fuzzy aspects of DNA structures involving a large database. This study presents algorithms for codon–noncodon delineation in a DNA sequence by applying the concepts of fuzzy inference engine (FIE). Simulations using codon statistics of human-beings and other bacterial species (namely, *Escherichia coli* (*E. coli*), *Rickettsia prowzekii* (*R. prowzekii*) and *Methanococcus jannaschii* (*M. jannaschii*)) are done to illustrate the efficacy of delineation approach pursued. Content-wise, the paper is organized as follows: In the following section (Section 2), the fuzzy attributes of codon and noncodon parts in a DNA complex are described. In Section 3, considerations on DNA sequence analysis and relevant scoring metrics for statistical discrimination between the contents of the sequence is elaborated. Section 4 is devoted to illustrate the use of if-then rule-based

logical inference towards a coarse-search for codon, noncodon and overlapping (fuzzy) subspaces along a DNA sequence. In Section 5, an FIE that uses certain (IT-based) scoring metrics for content discrimination (via feature detection) across a fuzzy subspace having an overlap of codons and noncodons is explained. Presented in Section 6 are details on the information-theoretics of a fuzzy DNA domain and the related algorithmic considerations. Section 7 provides particulars on the simulations performed and the results obtained thereof. Lastly, inferential remarks are presented along with a concluding note in the closure part of Section 8.

2. The nucleotide structure and fuzzy DNA complex

As mentioned earlier, in the field of molecular biology, real-world sequence patterns of a DNA sequence are largely described only in subjective notions and approximate norms. That is, sufficient information about a particular sequence (or a part of it) may not be available in certain situations involving DNA studies and some information pertinent to a data set could be missing altogether.

A ''spatial event'' in a DNA sequence pattern depicts an occurrence set of triplet symbols phased as a three-letter permutation of A, T, C and G. This depiction translates into modelling the nucleotide that conforms to a codon set of 12-dimensions (=3 phases $\times$ 4 bases) fuzzy code. That is, considering a triplet codon set {Bs0, Bs1, Bs2} made of four base chemicals Bs $\in$ {A, T, C, G} with three phases, namely {0, 1, 2}, a fuzzy code space, $\mathcal{J}^r$ {r = 1, 2, ..., 12} can be specified with a cardinality (of the fuzzy code set) equal to (3 $\times$ 4) = 12. Further, each spatial event under consideration can be attributed with an ''event-length'' (or segment length or subsequence), which is characterized by a ''value'' that depicts a quantitative measure or ''score'' illustrating the statistical feature of the event-space.

The portion of a DNA that bears meaningful coding information is known as exon. In between the event-lengths exists an ''interval'' that has insignificant or ''junk'' features presenting almost a null-score on the associated event characteristics. Such noncoding parts of a sequence that occur within the active gene are called introns. Both event-lengths of exons as well as the intervals due to introns can be regarded as random variables and they constitute a stochastical domain representing the DNA complex. Junk or not, the entire DNA sequence is faithfully copied in protein synthesizing processes of a DNA code transcripted into a molecular language, which eventually translates into an amino acid language of the proteins as described in [1,8]. Specifically, presented in [8], is an article by Guigó on the DNA composition wherein the usage of codons towards protein-coding is highlighted in terms of uneven probability or frequency of occurrence of tri-nucleotides (triplets).

Thus, considering the stochastical universe of a DNA sequence, the triplets constituting a protein-coding part are those that can be characterized by uneven probabilities of occurrence in the chain specific to a given living system. Denoting this set as $\mathbf{X}$:{x_i}, its elements (triplets) being purely protein-making codons are assumed to have non-uniform probabilities of occurrence; and, the subscript i on the element x_i denotes the location of this codon set as a subspace (subsequence) in the

Table 1 An example of codon statistics: human codon frequency usage (f_c) (http://www.kazusa.or.jp/codon/ [17]).

Codon triplet	f_c	Codon triplet	f_c	Codon triplet	f_c	Codon triplet	f_c
GGG	0.01708	AGG	0.01209	TGG	0.01474	CGG	0.01040
GGA	0.01931	AGA	0.01173	TGA	0.00264	CGA	0.00563
GGT	0.01366	AGT	0.01018	TGT	0.00999	CGT	0.00516
GGC	0.02494	AGC	0.01854	TGC	0.01386	CGC	0.01082
GAG	0.03882	AAG	0.03379	TAG	0.00073	CAG	0.03295
GAA	0.02751	AAA	0.02232	TAA	0.00095	CAA	0.001194
GAT	0.02145	AAT	0.01643	TAT	0.01180	CAT	0.00956
GAC	0.02706	AAC	0.02130	TAC	0.01648	CAC	0.01400
GTG	0.02860	ATG	0.02186	TTG	0.01143	CTG	0.03993
GTA	0.00609	ATA	0.00605	TTA	0.00555	CTA	0.00642
GTT	0.01030	ATT	0.01503	TTT	0.01536	CTT	0.01124
GCC	0.01501	ATC	0.02247	TTC	0.02072	CTC	0.01914
GCG	0.00727	ACG	0.00680	TCG	0.00438	CCG	0.00702
GCA	0.01550	ACA	0.01504	TCA	0.01096	CCA	0.01711
GCT	0.02023	ACT	0.01324	TCT	0.01351	CCT	0.01803
GCC	0.02843	ACC	0.02152	TCC	0.01737	CCC	0.02051

universe of DNA sequence domain. The distribution statistics of codons is decided by the species (such as human-being, bacteria, etc.) to which the DNA belongs to, and relevant probability set is denoted by $[\{P_{1x}\}_{\alpha=1,2,\ldots,64}]_i$ with $x_i \in \mathbf{X}:\{x_i\}$ and $\left[\sum_{\alpha=1}^{64}(P_{1x})_\alpha = 1\right]_i$ where α depicts the $4^3 = 64$ permuted triplets (or possible three-letter words of the genetic language) formed by the base chemicals A, T, C, or G. The set of $\{P_{1x}\}_\alpha$ of the human DNA, for example, is listed in Table 1. (Similar details on bacterial species etc. are available in the GenBank web site cited as [17].)

The noncodons (or the so-called ''junk'' codons) that do not carry any distinct functional attributes can be modelled with equally likely random occurrences of the 64 triplets. That is, denoting the noncodons by a set $\mathbf{Y}:\{y_k\}$, the occurrence probabilities of the constituent triplets are assumed to have a uniform distribution, namely $[\{P_{2y}\}_{\alpha=1,2,\ldots,64}]_k = 1/64$ and $\left[\sum_{\alpha=1}^{64}(P_{2y})\alpha = 1\right]_k$. These subspaces of noncodon set (junk codon subsequences) are presumed to exist as ''interval'' events (identified by the index k) along a DNA sequence. The **X** and **Y** subspaces are illustrated in Fig. 1.

Apart from the set of **X** and **Y** subsequences (depicting codon-only and noncodon-only parts respectively), there is also a possibility of a third set of subsequence domains (**Z**) along the DNA chain that contain overlapping codons and noncodons as shown in Fig. 1. This fuzzy database relation representing imprecise codon/noncodon attributes of the contents exists in a jth subspace $[\mathbf{Z}:\{z_j\}]_r$ (with $r \in \mathcal{J}^{12}$) as detailed in Fig. 2.

Thus, the DNA chain under consideration contains a set of differential segments (or blocks) $\{\Delta\ell\}_{i \text{ or } k}$ belonging to **X** or **Y** respectively; and, corresponding to each such differential segment, certain ''tuples'' can be prescribed in linguistic norms in order to describe the extent to which the differential segment in question belongs to codon or noncodon type. For instance, $(\Delta\ell)_i \in \mathbf{X} \Rightarrow$ (Excellent, very good), $(\Delta\ell)_k \in \mathbf{Y} \Rightarrow$ (Good, fair), etc. are examples of such tuple formats [18]. Eventually, these qualitative codon/noncodon attributes (in linguistic descriptions) have to be quantitatively specified via metrics depicting some degree of confidence of belongingness towards codon or noncodon attributes.

In summary, considering codon, noncodon and fuzzy subsequences of Fig. 1, these are denoted by the sets: $\mathbf{X}:\{x_i\}$, $\mathbf{Y}:\{y_k\}$ and $\mathbf{Z}:\{z_j\}$ respectively with event-lengths $(\ell_c)_i$, $(\ell_{nc})_k$ and $(\ell_f)_j$ as illustrated where $\{i = 1, 2, \ldots, I\}$, $\{k = 1, 2, \ldots, K\}$ and $\{j = 1,$

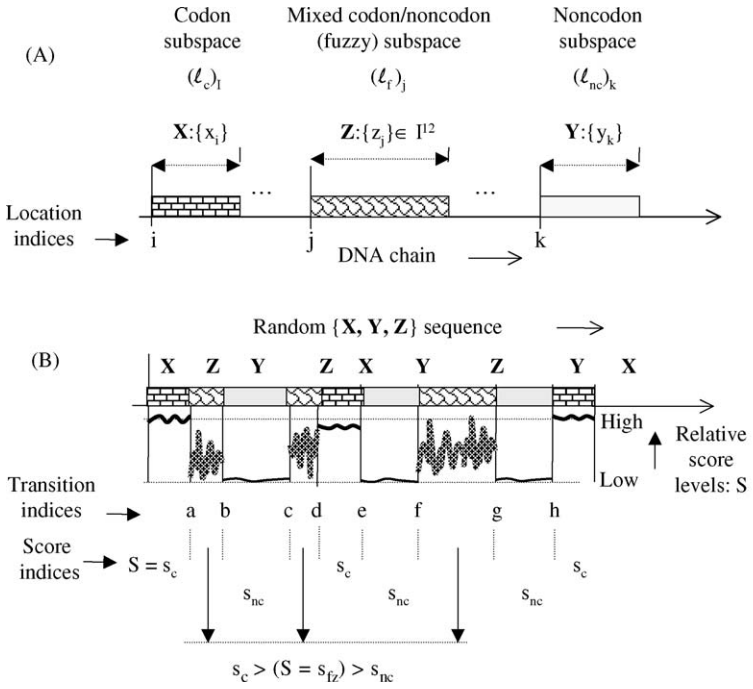

Figure 1 (A) A DNA chain divided into subsequence domains (or subspaces) containing codon-only, noncodon-only and mixed codon—noncodon occurrences. These subsequences are identified by location indices i, k and j respectively and designated as follows: codon subsequence set $\mathbf{X}:\{x_i\}$, noncodon subsequence set $\mathbf{Y}:\{y_k\}$, and, mixed/fuzzy subsequence set $\mathbf{Z}:\{z_j\}$. They correspond to lengths of event-spaces $(\ell_c)_i$, $(\ell_{nc})_k$ and $(\ell_f)_j$ respectively. (B) A random sequence of **X**, **Y** and **Z**. Illustrated are transition indices (a, b, ..., h) and the scale of scores (S) that distinguishes the codon **X**, noncodon **Y** and fuzzy **Z** regimes.

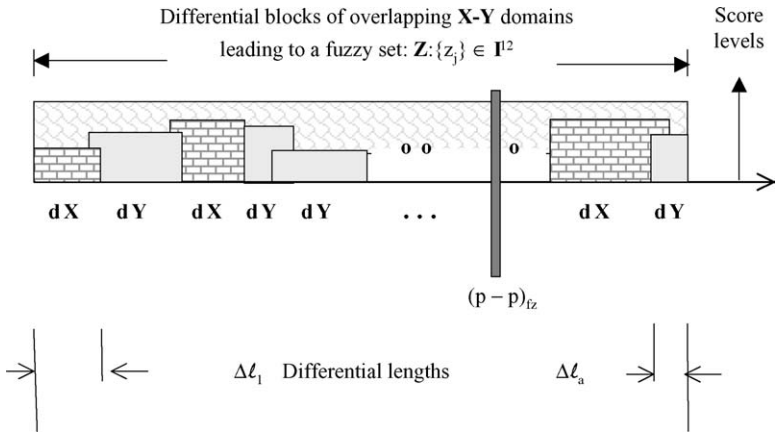

Figure 2 Illustration of the overlaps randomly placed as d**X** and d**Y** differential blocks constituting a fuzzy set **Z**:{z_j}. A moving-window based scoring is performed across these differential-lengths (blocks). Defuzzification of the scored entities leads to a centre-of-moment value for the codon—noncodon delineating boundary of the fuzzy subspace. It is denoted as $(p—p)_{fz}$ in the illustration.

2, ..., *J*} are sets depicting the locations (sites or subspaces) of **X**, **Y** and **Z** respectively as they prevail along a DNA sequence. The codon parts depicted by **X** contain elements $x_i \in \mathbf{X}$ only to some degree of membership {0, 1}. Likewise, the noncodon parts denoted by **Y**, contain elements $y_k \in \mathbf{Y}$ only to some degree of membership {0, 1}. The universe $\mathbf{\Omega} \in \{\mathbf{X}, \mathbf{Y}\}$ could also map as patches of **X** and **Y** that are smeared (overlapped) in the space $\mathbf{\Omega}$ leading to a fuzzy part (**Z**) that can only be vaguely identified as per some fuzzy relational rules. This region of overlapping codons and noncodons depicting a fuzzy domain, $z_j \in \mathbf{Z}$ can be visualized as a fuzzy space, $\mathcal{J}^{12}$ with 64 codons residing at $2^{64} = 4096$ corners of a 12-dimensional unit hypercube.

The procedure to construct a data sequence containing **X**, **Y** and **Z** subspaces that emulate a DNA sequence is as follows: A test DNA data sequence is required to contain crisp sets **X** and **Y** and fuzzy parts **Z** as shown in Fig. 1. The total stretch of such a DNA chain (of a given species) is a massive set of triplets of base chemicals {A, T, C, G} located along its length. Such a DNA chain containing a large number of *N* triplet locations can be emulated by random occurrences of subsequences of three categories, namely, **X**, **Y** and **Z**. The subspaces (**X**, **Y** and **Z**) are identified in the simulations by pointer-positions traversed along the stretch of *N* locations at *i*th, *k*th and *j*th classes respectively.

A task involved in DNA analysis and related mining process refer to both identifying locations of subspaces (**X**, **Y** and **Z**) as well as analyzing exclusively the fuzzy vector space representing **Z**:{z_j}. Applying fuzzy set theory in the context of such DNA analysis warrants decision-theoretics based on if-then rules. These rules are first applied to identify the locales of subspaces, namely (**X**, **Y** and **Z**); and, as a next step, the concepts of fuzzy theory are invoked with reference to fuzzy subspaces depicting the $\mathcal{J}^{12}$-dimensional vector space of **Z**:{z_j}, in order to deal with the associated imprecise information. Hence, subjective (expert) opinions are exercised on the entities that are specified as linguistic terms across the elements of the set {z_j}. In this context, there could be vague queries such as, ''Which part of the DNA chain or segment is domineering in coding for the protein?'' or ''Which clusters of triplets along the DNA chain belong significantly to the ''junk set?'' that are posed more often to DNA database users than precise queries. Therefore, resorting to fuzzy concepts is essential to address such imprecise or fuzzy queries.

3. DNA sequence analysis

The DNA analyses (such as codon—noncodon discerning) are bioinformatic data-mining efforts imminently required in the feature detection efforts pertinent to DNA structures. That is, a practical need prevails in DNA analysis to detect regions of shared similarities in the polynucleotide complex and distinguish them from regions that are distinctly dissimilar from them.

The analysis concerning codon/noncodon delineation in the universe of a DNA chain has essentially two steps: (i) the first task is to distinctly segregate similar subsequences and group them as **X**, **Y** and **Z** categories; and (ii) the second exercise refers to determining the border of separation of overlapping (fuzzy) codons and noncodons within the subsequence of **Z**:{z_j} category.

Aggregating similar subsequences in a DNA chain will reduce the search-space of analysis into three

conglomerated sets, **X**, **Y** and **Z**. Otherwise, multiple constituents of subsequences representing codons, no ncodons and fuzzy parts in a DNA complex, the total data sequence in real-world situations could be very large. As such, simple applications of if-then rules to analyze and discern the constituent subsequences may become explosive due to inevitable curse of dimensionality. By segregating the subspaces (as **X**, **Y** and **Z**), the regions will then be distinctly identified in three smaller parts, so that each part would become more amenable for a restricted search-space analysis.

Thus, isolating a the fuzzy subsequence **Z** for example, it becomes available as an exclusive search-space and can be subjected to limited fuzzy queries pertinent to its tuple contents; and, appropriate metrics can be used to quantify the membership of belongingness of the tuples (in each differential block) of this fuzzy domain. The associated effort corresponds to ''scoring'' shared similarity features. In other words, the scoring method quantifies the extent of codon and/or noncodon attributes of a test segment (denoting a differential block) of a subspace in a DNA chain. For the purpose of such scoring, some molecular-biology concepts have been adopted in practice. Relevant considerations are briefly indicated below [19]:

(1) *Signature scoring*: A ''signature'' in the context of a DNA sequence refers to an amino acid distribution resulting from the expression of the DNA. Scoring on the signature within a sequence token involves assessing the probability of occurrence of the amino acid.
(2) *Length scoring:* In this effort, a target length is specified for a token sequence of a DNA; and, any observed length in the scoring effort is assumed as a random variable around the target length within a specified set of upper and lower bounds.
(3) *Charge and hydrophobicity scoring*: This refers to scoring based on the abstract notions on the extents (high or low) of charge-content and hydrophobicity associated with a DNA sequence. Correspondingly, a measure of charge is associated with an amino acid in the sequence and a hydrophobicity index is specified for the sequence as scoring metrics.
(4) *Amphipalic alpha helix and beta sheet scoring*: These are another set of scoring functions similar to charge and hydrophobicity scoring and they correspond to hydrophobic and hydrophilic weightings that can be attributed to a given DNA sequence.

Thus, there are four characteristics of a DNA sequence that can be subjected to scoring towards assaying shared similarities and feature detection in DNA populations.

As mentioned before, an example of feature distinguishing task in bioinformatics corresponds to identifying and delineating codon (**X**) and noncodon (**Y**) entities with a boundary of separation either when the regions of codons and noncodons are adjunct to each other (with a distinct or crisp boundary of separation) or when the codon and noncodon constituents exist within a subsequence as overlapping differential blocks posing an imprecise demarcation of separation.

The present study offers a method of analyzing a DNA complex in the perspective of the following two efforts: implementing first a coarse-search to identify the **X**, **Y** and **Z** regions using codon statistics pertinent to a test DNA sequence of a given species. Concurrently, this coarse-search would enable ascertaining borders between the three categories of regions {**X**, **Y** and **Z**} involved. The procedure adopted thereof is based on if-then or else conditional logic. It uses the Euclidean distance measure for the if-then decision applied to distinguish a codon from a noncodon field.

The second task refers to performing exclusively a fine-search on the region **Z** that has fuzzy attributes. Inasmuch as the codons and noncodons are overlapping in this fuzzy domain **Z**:$\{z_j\}$, a precise or crisp boundary of separation for the codon—noncodon contents cannot be specified (within **Z**). Therefore, the concepts of fuzzy characterization are applied to prescribe tuples for the differential blocks of events in the fuzzy subsequence in qualitative norms that describe (subjectively) the extents of codon and/or noncodon content (across each differential length). A subsequent effort will be the conversion of these qualitative descriptions into classes of membership function so that a quantitative value can be prescribed for the contents of each differential block being tested; hence, a centroidal location of the border of codon—noncodon delineation for the subsequence being analyzed can be elucidated via defuzzification.

Relevant to scoring evaluated in each search process (namely, the coarse- and fine-search), a set of IT-measures [20—24] are considered here as outlined below:

(I) Euclidean distance (ED) metric—a metric that indicates greater similarities between a pair of vector spaces (representing, for example, codons and noncodons) when the Euclidean distance between them is smaller.

In general, the Euclidean distance (d_E) specifies scoring (S_E) of similarities/dissimilarities

between a pair of two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ by a measure given by [21]:

$$S_E = d_E(\mathbf{v}_1, \mathbf{v}_2) = ||\mathbf{v}_1 - \mathbf{v}_2|| \quad (1)$$

(II) To estimate the shared similarities/dissimilarities across the differential blocks in a fine-search applied to a fuzzy domain representing the subsequence **Z**:{z_j}, the scoring measures adopted are as follows.

Fisher linear discriminant metric [25]. This is a discriminant function depicting some measured scores elucidated (for example, on a set of differential blocks of overlapping codons and noncodons) enables a way by which the blocks are ''best discriminated''.

Cross-entropy measure: This metric facilitates the assessment of relative information between the differential blocks in the fuzzy subspace, **Z**:{z_j} using the statistical divergence concept of Kullback and Leibler (KL) [23,26,27].

Statistical distance measure: This measure assesses the statistical parameters (like variances and covariances) of distributions belonging to a pair of populations (such as codons and noncodons) in a test space so as to distinguish the similarities/dissimilarities between them. Example of this metric are: Mahalanobis and Bhattacharyya distances [21,24].

Correlation measure: It denotes for example, Hamming distance applied to a pair of sequences (representing for example, the codon and noncodon segments) taken in their binary formats. It refers to the outcome of an XOR (modulo-2) operation performed between the binary chains [28].

4. If-then rule-based decision logic for a coarse-search of X, Y and Z regions

Referring to Fig. 1, various transition boundaries across the subsequences of codon, noncodon and the set of fuzzy regimes along an arbitrary DNA sequence can be identified as follows:

$$\begin{aligned} &\{a, c : \rightarrow (x_i)\text{-to-}(z_j)\}; \quad \{b, d : \rightarrow (z_j)\text{-to-}(x_i)\}; \\ &\{e : \rightarrow (x_i)\text{-to-}(y_k)\}; \quad \{f : \rightarrow (y_k)\text{-to-}(z_j)\}; \\ &\{g : \rightarrow (z_j)\text{-to-}(y_k)\}; \quad \{h : \rightarrow (y_k)\text{-to-}(x_i)\} \end{aligned} \quad (2)$$

The transitions marked as *a*, *b*, *c*, *d*, *f* and *g* in Fig. 1 exist between crisp and fuzzy subspaces. On the other hand, the transitions *e* and *h* exist between nonfuzzy regimes of **X**:(x_i) and **Y**:(y_k) or vice versa and as such, they are crisp transitions.

To identify the set of transitions {*a*, *b*, ..., *h*} in Fig. 1, a coarse-search can be performed using the following considerations: suppose a metric is assigned to score the extents of codon and noncodon characteristics in the event-spaces of the DNA sequence. Hence, let ($S = s_c$) be the score measured in a codon-only domain (**X**). Likewise, considering an all-noncodon region (**Y**), let the score value be ($S = s_{nc}$).

Now, with reference to a **Z**:(z_j) subspace, suppose the score ($S = s_{fz}$) depicts the mean value of the measurements on the differential blocks across the subsequence (as per the adopted metric). This score will be distinctly smaller than s_c but larger than s_{nc}, that is, $s_{nc} < s_{fz} < s_c$. Thus, using an if-then rule-based logical algorithm (described below), the **X**, **Y** and **Z** regions can be distinguished in terms of their relative scores. Concurrently, the border set {*a*, *b*, ..., *h*} that delineates the subsequences **X**, **Y**, and **Z** is ascertained.

As mentioned before, the score values as above are determined on the basis of a statistical ED-measure ($\Theta = S_E$). Currently, it is specified in a scale from 0 to 100. The zero value of Θ means that the probability distribution of the population of the content (in the region being scanned) is uniform with a probability of occurrence (P_{2y}) equal to 1/64 for each of 64 possible triplet sets (made of the base set A, T, G and C). This would confirm that the scanned subsequence is a noncodon region (**Y**). When $\Theta = 100$, it means that the region being tested fully conforms to an uneven occurrence probability distribution of the triplets, namely, P_{1x} (of $x_i \in$ **X**) as specified in the GenBank database [17] for the codons of the DNA (of the biological species) being investigated.

In order to compute Θ-values, first a reference set of ED-values pertinent to noncodon occurrence frequency of $f_{nc} = (P_{2y}) = 1/64$ versus codon occurrence frequency $f_c = (P_{1x})$ is determined for all 12 base-phase combinations (A0, T0, C0, G0; A1, T1, C1, G1; A2, T2, C2, G2). Denoting these reference values as D_{Ref}A0, etc., they are explicitly given by: $D_{Ref}A0 = |f_cA0 - f_{nc}A0|$, $D_{Ref}T0 = |f_cT0 - f_{nc}T0|$, ..., $D_{Ref}C2 = |f_cC2 - f_{nc}C2|$ and $D_{Ref}G2 = |f_cG2 - f_{nc}G2|$.

Next, considering a window of sample length along the test sequence, the occurrence frequency (f_s) of the sample (test) population of bases versus the occurrence frequency of bases in a pure codon set (of an aligned length/window), namely, f_c is determined for all base-phase combinations (A0, ..., G2). Denoting these computed values as D_{Sample}A0, etc., the following set is obtained: $D_{Sample}A0 = |f_cA0 - f_sA0|$, $D_{Sample}T0 = |f_cT0 - f_sT0|$, ..., $D_{Sample}C2 = |f_cC2 - f_sC2|$ and $D_{Sample}G2 = |f_cG2 - f_sG2|$.

By definition, Θ is specified as a percentage of the ratio between the reference and sample distance. As such, the values of Θ are prorated to be between 0 and 100 by appropriately weighting the ratio in question. An example of such prorating is as follows:

$$[\Theta_{\text{Sample}}\text{A0}] = [(D_{\text{Sample}}\text{A0}/D_{\text{Ref}}\text{A0}) \times 100],$$
$$\text{if } (D_{\text{Sample}}\text{A0} < D_{\text{Ref}}\text{A0}); \quad \text{or},$$
$$[\Theta_{\text{Sample}}\text{A0}] = [(D_{\text{Ref}}\text{A0}/D_{\text{Sample}}\text{A0}) \times 100],$$
$$\text{if } (D_{\text{Sample}}\text{A0} \geq D_{\text{Ref}}\text{A0}); \ldots; \text{etc.} \quad (3)$$

The aforesaid calculations on Θ-values are done for the entire base-phase combinations (Θ_{Sample}A0, ..., Θ_{Sample}G2) across the entire DNA sequence and are used to construct a decision logic so as to infer whether the sample length in question belongs to codons (if, $\Theta \to 100$), or noncodons (if, $\Theta \to 0$). In the event of overlaps of sequence contents posing an ambiguity on codon/noncodon classification, the scoring statistics will invoke a decision logic in order to arrive at the required classification based on the value, $0 < \Theta < 100$.

The if-then or else decision logic for the sample set of polynucleotide bases considered above uses appropriately the computed set of 12 scores, namely (Θ_{Sample}A0, ..., Θ_{Sample}G2) in the Mandani's algorithm [29] where each possible combination of Θ-values is subjected to the if-then or else conditions. This leads to the decision whether the output function specifies a codon, a noncodon or corresponds to a fuzzy subsequence. Since there are 12 Θ-scores, there will be $2^{12} = 4096$ sets of if-then or else statements as listed in Appendix A.

With reference to the if-then or else statements indicated above, there are two input activation functions used. They refer to noncodon- and codon-scoring levels abbreviated as s_c and s_{nc} respectively. Correspondingly, there are three output activation functions to declare the output as codons, noncodons or fuzzy (abbreviated as s_c, s_{nc} and s_{fz} respectively). The resulting set of discrete outputs is aggregated over the entire 4096 decisions to produce a hard decision for each sample length (window) of pointer-positions traversed across the test DNA sequence. This process leads to ascertaining the codon/noncodon delineating transitions, namely $\{a, b, \ldots, h\}$ prevailing across the DNA chain. The above method is pursued in a sequential codon-first and then noncodon-next search across a search-space by traversing only an appropriate (limited) DNA chain of interest. The decision-logic algorithm described above for coarse segregation of codon/noncodon regions is illustrated with a pseudocode in Appendix B.

5. A fine-search for delineation of imprecise boundaries within a Z:{z_i} subspace

Having identified a fuzzy subsequence $\mathbf{Z}\{z_j\}$ via the coarse-search as above, the next step is to determine the boundary of separation (marked as $(p{-}p)_{fz}$ in Fig. 2) across the differential blocks containing codon and noncodon constituents (whose extents of the presence are imprecisely known due to the overlaps of the blocks). The task in hand is therefore, to ascertain this boundary using the associated codon–noncodon statistics. It refers to computing the coordinates of the boundary location, $(p{-}p)_{fz}$ done via defuzzification using the centre-of-area (CoA) or moment procedure [29]. The relevant considerations are as follows.

The delineation of codon–noncodon boundary as addressed in the existing works (such as in [16]), considers only those cases wherein the parts being contrasted are crisp sets. Hence, the studies pursued elucidate borders between codon and noncodon domains assuming that the delineating boundaries are sharp or unambiguous so that they can be dichotomized or bifurcated distinctly with a specified clarity and preciseness. For example, the technique envisaged in [16] uses thereof an entropy segmentation method. It involves computing the mutual entropy on codon occurrence statistics in the adjacent subsequences of codon and noncodon parts. This mutual entropy between them is scored in terms of a divergence measure (known as Jensen–Shannon (JS) measure [22], which is related to the general class of KL-measure [2] as will be described later). This metric estimates the relative entropy (or mutual information) between the data sets using the associated codon statistics. The corresponding change in the measure (score) computed across the tested parts enables ascertaining the required crisp demarcation.

Contrary to such precise demarcation evaluated in [16], as emphasized earlier, the real-world mixture of codon and noncodon parts is rather a random domain with imprecise overlaps that exhibit fuzzy characteristic across their transition boundaries. With due consideration of such dissonance in clarity of mixture constituents in a DNA subsequence such as $\mathbf{Z}\{z_j\}$, attempted here is a method to determine the borders between overlapping blocks (within $\mathbf{Z}$) using information-theoretic scores based on a set of metrics imposed on the membership functions. To illustrate the underlying considerations, the details on IT-considerations pertinent to a fuzzy codon–noncodon structure are presented in the following section.

6. Information-theoretics of fuzzy codon—noncodon mix

The heuristics of information-theoretics that can be applied to the fuzzy sequence model depicting a subspace such as $\mathbf{Z} \in \{z_j{:}(x_i, y_k)\}$, in essence follows the principles of relative entropy (in KL sense [2]). In terms of relevant statistical divergence, the mutual or relative entropy of a fuzzy set $\mathbf{Z} \in \{z_j\}$ under consideration can be described as follows: Suppose $\mathbf{Z}^c$ denotes the complement of $\mathbf{Z}$. The fuzzy mutual entropy implies a logistic map of relative entropy of the sum J-components belonging to the fuzzy set $\mathbf{Z}{:}\{z_{j=1,2,\ldots,J}\} \equiv \{x_1, x_2, \ldots, x_I{:}y_1, y_2, \ldots, y_K\}_{J=(I+K)}$, relative to a complement set $\mathbf{Z}^c$. In other words, for the set $\mathbf{Z}{:}\{z_j\}$ specified in a unit hypercube $[0,1]^J$, the fuzzy mutual entropy is given by [24]:

$$I = [H(\mathbf{Z}/\mathbf{Z}_c) - H(\mathbf{Z}_c/\mathbf{Z})] \in \mathbf{I}^J \tag{4}$$

It is shown in [23] that the fuzzy mutual entropy of Eq. (4) is equal to a corresponding divergence of Shannon entropy in the fuzzy domain implying that the information field in the fuzzy cube $\mathbf{I}^J$ refers to an inward or outward flow of entropy (information) flux in the hypercube. Further, the associated fuzzy cubes map smoothly onto extended real-spaces ($\mathbf{R}^J$) of the same dimension and vice versa. That is, $\mathbf{R}^J \to \mathbf{I}^J$ indicate an one-to-one ratio on a differentiable map $\mathbf{D}$: $\mathbf{Z}(z_j)$ in $\mathbf{I}^J$ with a differentiable inverse $\mathbf{D}'(d_j)$. This differential map of $\mathbf{D}(d_j)$ or $\mathbf{D}'(d_j)$ refers to a fuzzy system; and, it specifies a conversion of unbounded real inputs z_j to a set of bound values (or "fit values"), d_i as follows [23]:

$$d_j = 1/[1 + \exp(-z_j)] \tag{5}$$

so that iff $z_j \to \infty$, $d_j \to 1$, and iff $z_j \to -\infty$, $d_j \to 0$.

As well known, by associating a probability of occurrence p_j with z_j, the concept of Shannon entropy is governed by the following chain of probabilities (and entropies) in accordance with the traditional concepts of Shannon information-theoretics:

$$p_j \to (1/p_j) \to \log(1/p_j) \to \Sigma_j p_j \log(1/p_j) \tag{6}$$

$$\begin{aligned}(1-p_j) &\to [1/(1-p_j)] \to \log[1/(1-p_j)] \\ &\to \Sigma_j (1-p_j)\log[1/(1-p_j)]\end{aligned} \tag{7}$$

Likewise, a fuzzy chain can be formulated in terms of d_j as follows:

$$\begin{aligned}d_j &\to [1/(1-d_j)] \to \log[d_j/(1-d_j)] \\ &\to \Sigma_{j=1,2,\ldots,n}\{d_j \log[d_j/(1-d_j)]\} = H(\mathbf{Z}/\mathbf{Z}_c)\end{aligned} \tag{8}$$

Similarly,

$$\begin{aligned}(1-d_j) &\to 1/d_j \to \log[(1-d_j)/d_j] \\ &\to S_{j=1,2,\ldots,J}\{(1-d_j)\log[(1-d_j)/d_j]\} \\ &= H(\mathbf{Z}_c/\mathbf{Z})\end{aligned} \tag{9}$$

Hence, by defining a relative fuzzy information unit as $\log[d_j/(1-d_j)]$, the following relation can be deduced for the explicit representation of Eq. (4) from the considerations of fuzzy mutual entropy indicated above:

$$\begin{aligned}&\Sigma_{j=1,2,\ldots,J}\{\log[d_j/(1-d_j)]\} \\ &\quad = H(\mathbf{Z}/\mathbf{Z}_c) - H(\mathbf{Z}_c/\mathbf{Z})\end{aligned} \tag{10}$$

The characteristics of fuzzy information flux discussed above are elaborated by Neelakanta and Abusalah in [23] and summarized as follows:

- Fuzzy mutual entropy is flux-like and is specified by an information field in a fuzzy cube;
- The points on a fuzzy cube correspond to fuzzy uncertainty descriptions;
- Fuzzy mutual entropy is equal to the negative of the divergence of Shannon entropy;
- Shannon entropy and fuzzy mutual entropy define vector fields on the fuzzy cube;
- Shannon entropy is similar to a potential of the conservative mutual entropy vector field. Thus, the dynamical aspects of information flux flows on the fuzzy cube correspond to governance by the second law of thermodynamics.

Commensurate with the IT-aspects of a fuzzy domain deliberated above, a set of (information-theoretic) metrics can be formulated for scoring purposes in fuzzy domains akin to similar measures specified for crisp sets [2,23,26]. In general, the IT-measures required for contrast evaluations (in crisp as well as in fuzzy domains) can be specified in terms of the associated statistical divergence between the entities being contrasted or by determining the statistical distance between them. Generally known as statistical discriminants and distance measures, a host of such formulations have been developed in the past [21–23,26] and applied to various disciplines of science, engineering, economics, etc.

Currently, four versions of such metrics are chosen and adopted toward scoring for codons—noncodons discrimination in a fuzzy block ($\mathbf{Z}$) of a DNA sequence. These measures as identified earlier are as follows: With reference to a symbolic representation of the DNA sequence, (i) the Fisher linear discriminant metric; (ii) the statistical divergence based on cross-entropy considerations; (iii) the statistical distance depicting the measure of stochas-

tical dissimilarity between a pair of statistical data sets are adopted; and (iv) a correlation measure ascertained from the modulo-2 operation (yielding the Hamming distance) between a pair of binary sequences is used to score the sequence taken in the binary format. (However, without any loss of generality, other related measures as reported in [21–23,26] can as well be used in lieu of the four metrics mentioned above.) Presented in the following subsections are brief notes on the four selected metrics that are subsequently deployed in the computations of the delineations in test fuzzy data sequence domains.

6.1. Fisher linear discriminant measure

The Fisher linear discriminant (F) metric is an IT-measure that can be built to prescribe a linear discrimination function with coefficients optimized on the basis of statistical or entropy features of a data set. It can be used as a scoring metric to contrast two statistical sets possessing a relative uncertainty.

Relevant to the present study, this contrasting ability of F-metric is shown to determine whether a set of triplets constituted by chemical bases {A, T, G, C} that are situated over a differential length (or ''block'') of a DNA subsequence is coding (protein-making) or noncoding (''junk'') type. That is, the F-metric adopted provides a scoring mechanism that helps distinguishing similarities or dissimilarities across the differential blocks of **Z**.

The concept of discriminant function under discussion was originally developed by Fisher in 1936 [25] with the objective of elucidating a method that classifies or distinguishes a pair of data sets. In other words, relevant effort is concerned with finding out the extent to which two sets of data are statistically similar or dissimilar. The classical effort of Fisher refers to taxonomic problems and involves prescribing a linear function (F) with unknown coefficients $\{\lambda_i\}$ for a set of measurements $\{\phi_i\}$; and, this function in effect, is optimized with a choice of $\{\lambda_i\}$ so as to provide the largest scoring that distinguishes the two test data sets. That is, the linear discriminant of the measurements with optimized coefficients enable an algorithm by which the ''populations are best discriminated'' [25].

For example, the question addressed by Fisher in [25] is pertinent to four measurements (ϕ_1, ϕ_2, ϕ_3, ϕ_4) on four characteristics of a flower (namely, sepal length, sepal width, petal length and petal width); and, the discriminant function $F = (\phi_1\lambda_1 + \phi_2\lambda_2 + \phi_3\lambda_3 + \phi_4\lambda_4)$ is optimized with appropriate choice of the set $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ so that the ratio of the difference between the specific means to the standard deviations within the species of flowers studied. This maximized value indicates the extent of statistical separation between the species in terms of their similarities or dissimilarities (assessed via floral characteristics obtained in the experiments due to Fisher).

6.2. Statistical divergence metrics

Another set of metrics advocated here for contrast evaluation in a fuzzy data sequence is based on statistical divergence concept. A number of such metrics have been formulated in the existing literatures [2,21–23,26] and they are in general, based on the so-called KL-concept of cross-entropy or mutual information arising from the difference in the statistical attributes of the two statistical entities being compared. More generally, they all fall under the scope of so-called Csiszár metrics described in [23,26,30] and briefly addressed later in this section.

The statistical divergence concept of contrasting can be applied to both crisp as well as fuzzy features and can be used for data sequence discrimination. For example, the entropy-based metric suggested and used in [31] refers to the so-called Jensen–Shannon (JS) measure (indicated before) and it essentially belongs to the class of KL-measures [22]. It compares two vector spaces, $\mathbf{v}_c$ and $\mathbf{v}_{nc}$ corresponding to codon and noncodon regions respectively in terms of the associated cross-entropy. It is shown in [16] that the entropic segmentation approach using the JS-measure could lead to predicting the borders between coding and noncoding regions without any a priori training details on known sets. Relevant procedure is also shown to be more precise than a simple moving-window technique that can be used in discerning a coding DNA from a noncoding DNA. However, the method detailed in [16] addresses only the crisp sets of codons and noncodons, which do not have any ambiguous overlaps.

The concept of divergence metric (such as JS-measure adopted in [16]) can be extended in terms of the associated entropy considerations to model a fuzzy DNA composition. As such, addressed in the present study are the feasibility and usage of entropic segmentation technique with the infusion of fuzzy considerations. Also indicated is the feasibility of using other statistical divergence/discriminant measures (in lieu of JS-measure used in [16]) for data sequence discrimination applications under discussion).

The entropy-based discriminant measures representing the divergence between vector-spaces in question are essentially entropy-specific metrics based on the expected value of the likelihood-ratio

of statistics of two entities (such as codon—noncodon populations). These expected values of likelihood-ratio are general representations of cross-entropy (or mutual information) between the statistics of the two regions under consideration. The cross-entropy/mutual information functions between the statistics stipulated via P_{1x} and P_{2y} are given by KL-metric pair, namely, KL1 = $KL_{1,2}$ and KL2 = $KL_{2,1}$:

$$KL_{1,2} = I_{1,2} = -\sum_n P_{1x,n} \log(P_{1x,n}/P_{2y,n}) \tag{11}$$

$$KL_{2,1} = I_{2,1} = -\sum_n P_{2y,n} \log(P_{2y,n}/P_{1x,n}) \tag{12}$$

Suppose a DNA sequence is specified by a vector space $\mathbf{v}_{1,2}$ with l,2 depicting a pair of accountable sets of entities (constituted by subsequences of codons and noncodons in the present case) having the probabilities of occurrence P_{1x} and P_{2y}. Suppose the minimum discernible "distortion" (D_L) depicting the similarity and dissimilarity features between the subsequence #1 and #2) is assessed in terms of minimum cross-entropy given by the relations of Eqs. (11) and (12); that is, $D_L(\delta) \equiv$ information content $[I_{1,2 \text{ or } 2,1}(\delta)]$ where, δ is a measurable entity of dissimilarity between the sequences seen as a "distortion" feature (D_L) or information loss in one data sequence with respect to the other.

Further, Eqs. (11) and (12) representing the KL-measures of divergence are expectations under two hypotheses, namely $[h]_u$ with $u = 1, 2$; and, $P_{1x}(x)$ is the probability of the observations on x_i when the hypothesis $[h]_u$ is true. Hence P_{1x}/P_{2y} or P_{2y}/P_{1x} depicts the log-likelihood-ratio, $L(x)$ and the relations of Eqs. (11) and (12) denotes, in essence, the $E[\log\{L(x)\}]$, where $E[.]$ is the expectation operator.

An associated measure called symmetrized KL-measure is referred to as Jeffery's measure (J). It is defined as, $J = \pi_1 KL1 + \pi_2 KL2 = \pi_1 I_{1,2} + \pi_2 I_{2,1}$ where the weighting coefficients π_1 and π_2 are such that $(\pi_1, \pi_2) < 1$ and $(\pi_1 + \pi_2) = 1$. In a symmetric consideration (between $1 \Leftrightarrow 2$ vector spaces), $\pi_1 = \pi_2 = 0.5$.

Further, as mentioned before, the JS-measure adopted in [16] is also implicitly related to the KL-measure. It is a generalized divergence given by [22]:

$$JS(P_{1x}, P_{2y}) = H(\pi_1 P_{1x} + \pi_2 P_{2y}) - \pi_1 H(P_{1x}) - \pi_2 H(P_{2y}) \tag{13}$$

where $H(\gamma)$ is equal to $-\sum P(\gamma) \log[P(\gamma)]$ denoting the Shannon information.

In addition to the KL-, J- and JS-measures indicated above, there are other divergence measures that can also be considered to distinguish the statistics of two entities (such as the codons and noncodons). As mentioned before, they can be grouped under the so-called Csiszár family of entropy measures [23,26,30]. These are based on Csiszar's f-divergence concept described in [23,26]; and a generalized representation of these measures is given by:

$$D_R(P_{1x} : P_{2y}) = \sum_\ell (P_{1x})_\ell \Phi\{[(P_{1x})_\ell]/[(P_{2y})_\ell]\} \tag{14}$$

$$D_R(P_{2y} : P_{1x}) = \sum_\ell (P_{2y})_\ell \Phi\{[(P_{2y})_\ell]/[(P_{1x})_\ell]\} \tag{15}$$

where $\Phi(\cdot)$ is a twice differentiable convex function for which $\Phi(1) \equiv 0$ and the discriminant function D_R satisfies certain essential and desirable characteristics elaborated in [23,26]. The KL-, J- and/or JS measures are special cases of Csiszár measure.

6.3. Statistical distance measures

These are statistical distance metrics that can also be used as discriminant functions to determine the "similarities" or "distances" between two sets of entities denoted by the subscripts 1 and 2. A commonly used measure of such distance or similarities is known as Mahalanobis measure (M) [21,24]. It is based on the concept of Euclidean distance $(d_E)_{12}$. When the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ representing two different populations (pools) of data, the ED-measure $(d_E)_{12}$ can be specified in terms of μ_1 and μ_2 depicting the mean values of the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ respectively. That is, by defining $\mu_1 = E[\mathbf{v}_1]$ and $\mu_2 = E[\mathbf{v}_2]$ with $E[\cdot]$ again representing the expectation operator, the squared value of the ED from $\mathbf{v}_1$ and $\mathbf{v}_2$ is given by:

$$[(d_E)_{12}]^2 = (\mathbf{v}_1 - \mu_1)^T \Sigma^{-1} (\mathbf{v}_2 - \mu_2) \tag{16a}$$

where T is the transpose operation and Σ^{-1} the inverse covariance matrix Σ. The covariance matrix is given by:

$$\Sigma = E[(\mathbf{v}_1 - \mu_1)(\mathbf{v}_1 - \mu_1)^T] = E[(\mathbf{v}_2 - \mu_2)(\mathbf{v}_2 - \mu_2)^T] \tag{16b}$$

The ED-measure depicted via Eq. (16) is the explicit form of Mahalanobis measure and it is a "coefficient of likeness" classically specified as a D^2-statistical distance norm. It accounts for both variances and covariances between the frequencies of "n-words", such as the 64 triplets of a DNA sequence [24].

Another statistical distance measure developed on the basis of Mahalanobis measure is known as the Bhattacharyya distance (B) and it is defined with

reference to a pair of probabilities P_{1x} and P_{2y} as follows [21]:

$$B = -\sum_{\ell} \ln(\rho\ell), \quad 0 < B < \infty \tag{17a}$$

where ρ is known as the Bhattacharyya coefficient given by:

$$\rho_{\ell} = \sum_{\ell} [P_{1x} \times P_{2y\ell}]^{1/2}, \quad 0 < \rho_{\ell} < \infty \tag{17b}$$

The *B*-measure is a suitable metric for a select property of average divergence between the two statistics being compared. Hence, it is taken as an example in the present study to illustrate the use of distance-measure concept to distinguish the codon and noncodon regimes in the test data pertinent to DNA compositions.

Another measure closely related to *B*-measure, is known as Kolmogorov variational distance (*K*) [21]. Likewise, there are other hosts of distance measures identified as Ali—Silvey distance metrics that have been comprehensively used in practice, especially in communication systems [21,32]. Though, the present study uses only the *B*-measure as a representative scoring metric, without any loss of generality, the gamut of entire distance measures can be tried in the codon—noncodon delineation efforts.

6.4. Correlation measure

Yet another scoring strategy that can be applied to quantify the linguistic tuples of a fuzzy domain (such as **Z**), is a simple correlation measure (*R*) that can be evaluated by modelling a data sequence set in a binary form. For example, suppose the set {A, T, C, G} is identically represented by a binary set {00, 11, 01, 10}. Hence, a sequence/subsequence containing triplets of {A, T, C, G} is first converted into a binary string. And, a test DNA sequence made of the triplets is correspondingly constructed in terms of zeros and ones. Then the *R*-measure under consideration would refer to the difference between the level of approval (meaning number of 0s) and the level of disapproval (meaning number of 1s) counted in the outcomes of a modulo-2 operation performed on a test binary fuzzy segment (of $\mathbf{Z}_{bs}$ being scored) with respect to an aligned and totally noncodon subsequence ($\mathbf{Y}_{bs}$) of the same length. (Here, the subscript "bs" is introduced to denote explicitly that the subsequences being considered are binary sets.) The outcome of XOR operation, namely, $\mathbf{Z}_{bs} \oplus \mathbf{Y}_{bs}$ is the Hamming distance [28] depicting the correlation measure (*R*) under discussion.

6.5. Fine-search with Fisher discriminant metric

A fine (local) search on an identified fuzzy subspace (or block) $\mathbf{Z}$:$\{z_j\}$ is required in order to locate codon—noncodon transition boundaries located within the fuzzy block and depicted as $(p—p)_{fz}$. This search is based on fuzzy evaluations following the heuristics of "search and score" applied appropriately to assign membership values for the qualitative descriptions of overlapping and ambiguous codon—noncodon locales across the fuzzy site. It is done by using a set of finely tuned metrics (such as the Fisher discriminant, the statistical divergence, statistical distance or the correlation measure described above).

The linear discriminant procedure due to Fisher adopted here refers to scoring the extent to which the contents of a given differential region within a fuzzy block of a DNA subsequence described qualitatively (in terms of codon-like or noncodon-like tuple attributes). This scoring is done by comparing the contents of a test block (d**Z**) with those of a reference block that has codon-only (d**X**) or noncodon-only (d**Y**) constituents. The procedure is as follows: suppose two neighbouring differential event-lengths d**X**: $(\ell_c)_i$, d**Y**: $(\ell_{nc})_k$ or d**Z**: $(\ell_f)_j$ within a test subspace of **Z** (in Figs. 1 and 2) are assessed. The procedure under consideration should enable identifying the delineation of separation, namely, the border between the differential units in question. The score for the set of DNA triplets that designates the codon or noncodon attributes to each differential length (of the block being tested), is a metric that refers to a numerical count of bases in that block. With reference to the four bases (A, T, C, G), suppose the corresponding three phases identifying the triplets are designated by a set {0, 1, 2}. Hence, there are 12 measured scores for each test differential length depicting to the following base-phase possibilities: {A0, T0, C0, G0, A1, T1, C1, G1, A2, T2, C2, G2}.

The next step is to determine the coefficients of a linear discriminant function of the fuzzy domain **Z**. For this purpose, two reference subspaces of equal lengths XR and YR containing respectively, crisp data on purely-codons and purely-noncodons are used. Relevant steps pursued are indicated in the pseudocode of Appendix C.

In terms of known coefficients obtained following the procedure indicated in Appendix C, the Fisher discriminant function (described earlier) is applied to the test DNA subsequence Z versus the reference subspace YR. The values of scores then generated in each differential-window accounts for the extents of codons and noncodons in the fuzzy test block.

Corresponding to the score obtained for each differential block, tuple characterization is specified to the pointer-position that defines the differential block in question. That is, mapping is done to reflect the scored value of a differential block to a corresponding descriptive tuple for the pointer-position of that differential window. Then, these tuples or descriptive values of pointer-positions so gathered are then subjected to defuzzification. This process would lead to a centroidal position (of the pointer) depicting the delineating boundary, $(p\text{—}p)_{fz}$ in the test fuzzy subspace.

6.6. Fine-search with statistical divergence and distance metrics

The procedure indicated for the fine-search of the delineating boundary within a fuzzy subsequence $\mathbf{Z}\{z_j\}$ using the Fisher metric can be adopted per se, with alternative metrics like the KL-, JS- or *B*-measure. The only difference will be that each metric will yield a different scale of measure that scores the codon—noncodon attributes of differential blocks. (In the example simulation performed in this study, the symmetrized form of the KL-measure, namely the *J*-measure is used to elucidate the scores across the differential blocks in a fuzzy test subsequence scanned by the pointer-position).

Regardless of the type of scoring metric used (the *F*-, KL-, *J*-, JS-, or *B*-measure), the defuzzification of the set of evaluated scores when mapped equivalently upon the pointer-positions (as described before with reference to Fisher metric) would provide the centroidal fix, $(p\text{—}p)_{fz}$ of the delineating boundary within the test fuzzy subsequence. The steps involved in defuzzification are summarized below.

6.7. Centre-of-area (CoA) or moment method of defuzzification

This procedure applies to the fine-search procedure on codon—noncodon delineation within a fuzzy subspace, **Z** (when the DNA sequence is represented in its symbolic form), using the *F*-, *J*- or *B*-measure as a scoring algorithm. The scored values assessed across the differential blocks in the test fuzzy region (**Z**) and mapped into corresponding tuples of pointer-positions are specified as entities belonging to a specified membership function. The resulting (mapped) fuzzy pointer-positions are then subjected to defuzzification using a centre-of-area (CoA) method. Relevant computations are summarized via a pseudocode presented in Appendix D.

6.8. Fine-search with the correlation metric

While the fine-search methods discussed above are indicated for a DNA sequence represented in its symbolic contents of triplets made of A, T, C and G, the sequence analysis in general, and the codon—noncodon delineation in particular, can also be addressed by representing the sequence in question in a binary format. In such a case, the correlation measure (described earlier) presenting the Hamming distance across a pair of subsequence sets being compared, ascertains the similarity or dissimilarity features across a fuzzy subspace (**Z**) with reference to a subspace of crisply defined contents such as purely-codons (**X**) or purely-noncodons (**Y**).

To implement the method under discussion, three versions of subspaces along the DNA chain are considered in their binary counterparts, $\mathbf{X} \rightarrow \mathbf{X}_{bs}$, $\mathbf{Y} \rightarrow \mathbf{Y}_{bs}$, and $\mathbf{Z} \rightarrow \mathbf{Z}_{bs}$. Then the correlation-metric-based assessment of codon—noncodon delineation in the fuzzy subspace, $\mathbf{Z}_{bs}$ is done via steps in the pseudocode of Appendix E.

7. Computation, results and discussion

The simulations performed in the present study in delineating codon—noncodon parts of a DNA sequence can be listed as below:

- Construction of a DNA sequence (see Figs. 1 and 2).
- Coarse delineation/identification of codon-only, noncodon-only and overlapping codon—noncodon (fuzzy) subspaces along a DNA sequence using if-then decision logic: The results illustrated in Fig. 3 show distinct levels of the computed scores in each region indicating the discerned subspaces with clarity.
- Implementation of fine-search on a fuzzy subspace (**Z**) using the scoring metrics: Fisher discriminant, symmetrized KL-measure (*J*-metric) and Bhattacharyya distance and application of CoA method to defuzzify the scored data in the subspace **Z**: relevant presentations of results in Figs. 4—6 show codon and noncodon subspaces (**X** and **Y**) distinguished by the distinct levels of the scores. In the fuzzy subspace (**Z**), the measured scores are jagged indicating the variations in the scores as the pointer traverses across the overlapping codon—noncodon differential blocks.
- Implementation of fine-search using the correlation measure (*R*) on a fuzzy subspace (**Z**) taken in binary format: shown in Fig. 7 is a fuzzy subspace

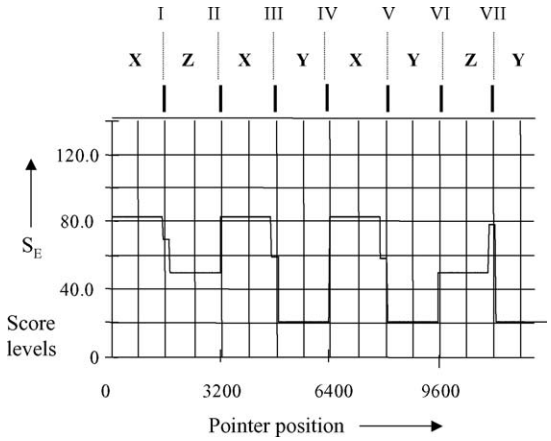

Figure 3 Computed results to illustrate how the scores using ED-measure (S_E) vs. the pointer-positions along a test DNA string classify/demarcate the codon (**X**), fuzzy (**Z**) and noncodon (**Y**) subspaces. Positions I—VII are transition locales of the codon, noncodon, and fuzzy parts. The values of S_E plotted correspond to an ensemble average of 100 runs. (The test DNA sequence refers to *M. jannaschii*.)

(**Z**) subjected to scoring across its differential blocks and Fig. 8 depicts computed results with the *R*-measure and exercising appropriate defuzzification procedure.

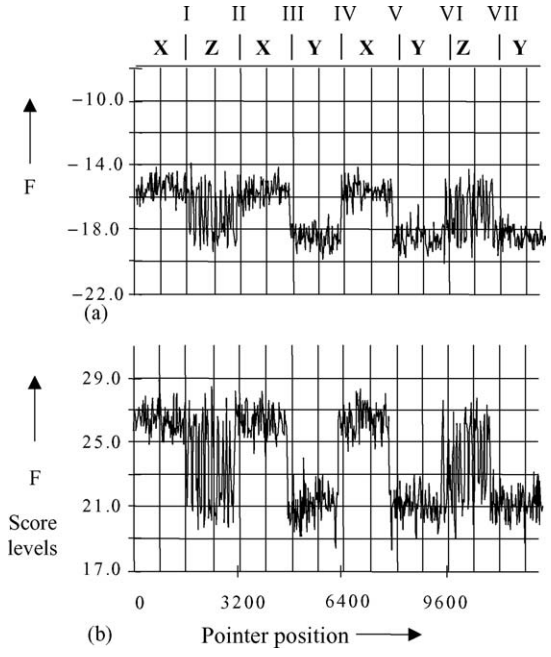

Figure 4 Computed scores using Fisher metric (*F*) versus the pointer-positions along a test DNA string made of codon (**X**), fuzzy (**Z**) and noncodon (**Y**) subspaces illustrate the associated demarcations. Positions I—VII are transition locales of the codon, noncodon, and fuzzy parts. The values of *F* plotted correspond to an ensemble average of 100 runs. (Test DNA string: (a) human-being and (b) *E. coli*.)

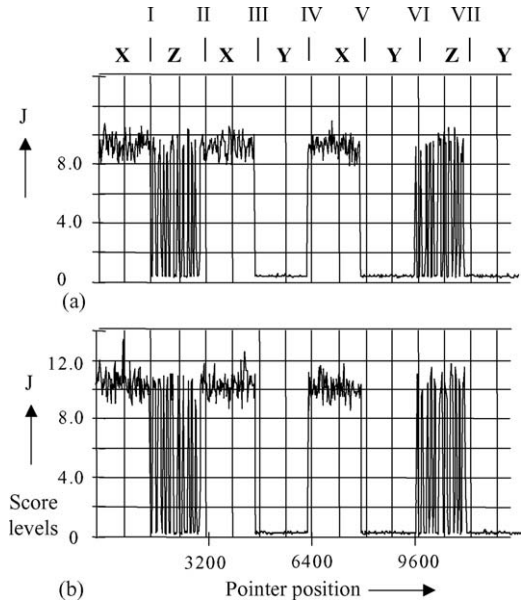

Figure 5 Computed scores using Jeffery's measure (*J*) vs. the pointer-positions along a test DNA string made of codon (**X**), fuzzy (**Z**) and noncodon (**Y**) subspaces illustrate the associated demarcations. Positions I—VII are transition locales of the codon, noncodon, and fuzzy parts. The values of *J* plotted correspond to an ensemble average of 100 runs. (Test DNA string: (a) *R. prowzekii* and (b) *M. jannaschii*.)

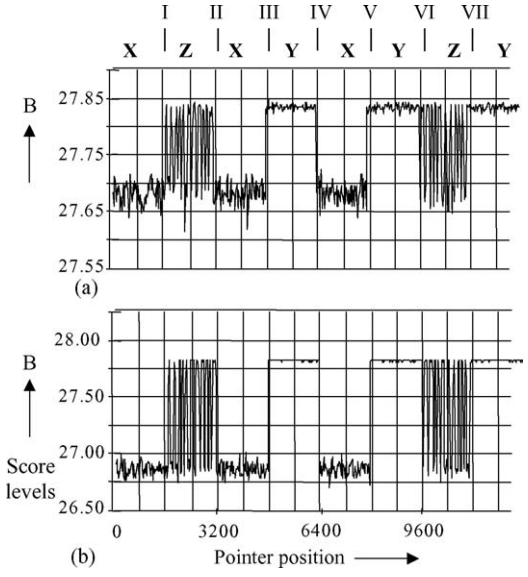

Figure 6 Computed scores using Bhattacharyya distance measure (*B*) versus the pointer-positions along a test DNA string made of codon (**X**), fuzzy (**Z**) and noncodon (**Y**) subspaces illustrate the associated demarcations. Positions I—VII are transition locales of the codon, noncodon, and fuzzy parts. The values of *B* plotted correspond to an ensemble average of 100 runs. (Test DNA string: (a) *Rickettsia prowzekii* and (b) *M. jannaschii*.)

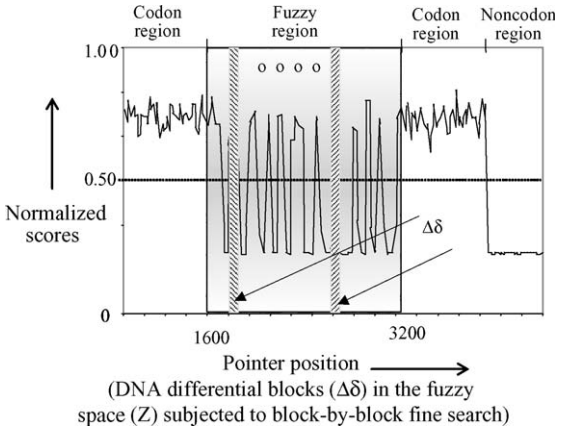

Figure 7 Illustration of a fuzzy subspace (**Z**) subjected to a fine search and scored for codon—noncodon characteristics across the differential blocks ($\Delta\delta$) within the subspace **Z**. The set of scores obtained and normalized between 0 and 1 are mapped as tuples (and are defuzzified as illustrated in Fig. 8 so as to ascertain the centre of moment location of codon—noncodon delineation boundary (within the subsequence **Z**).

All the simulations indicated above were performed using the algorithms and pseudocodes indicated in this study and applied to real-world DNA (codon statistics) data pertinent to human-being as well as a set of three bacterial species (*E. coli*, *R. prowzekii* and *M. jannaschii*), the codon statistics of which are available in [17]. Computations on symbolic DNA representations were done using relevant codes written in C (ANSI) with some C^{++} object-oriented elements. These computations were mostly executed on a Sun Solaris 8 System compiled on a Sun Workshop 64-bit C^{++} compiler. The correlation measure based scoring on a binary sequence was done on a PC platform using MatLab^{TM}/C compiler. The simulated results as presented in Figs. 3—6 and 8 (with relevant legends and figure titles) explicitly give details on the data used and the computational outcomes. The running times ranged from 20 to 30 s for a sequence of 6000 codons. Presented in Table 2 is a summary of results extracted from the computed results pertinent to the symbolic DNA representations. The simulations were performed over 100 ensemble runs in obtaining scores in the differential blocks using different seeds in emulating the relevant codon statistics. (The listed defuzzified scores in Table 2 correspond to mean values of such ensemble runs.)

Scoring computation using correlation measure: this simulation is done on a DNA sequence taken in a binary format. In a representative simulation, the parameters used are as follows: DNA type — human-being; length of total DNA sequence (containing *X*, *Y*, and *Z* strands): 6400 bases; number of subspaces: 100 (each containing 64 bases); and, three locations of X_{bs} (Csub): 0—1600, Y_{bs} (NCsub): 1600—4800, and Z_{bs} (fSub): 4800—6400 subspaces along the sequence are constructed. (The subscript "bs" explicitly denotes the binary format of the sequence.)

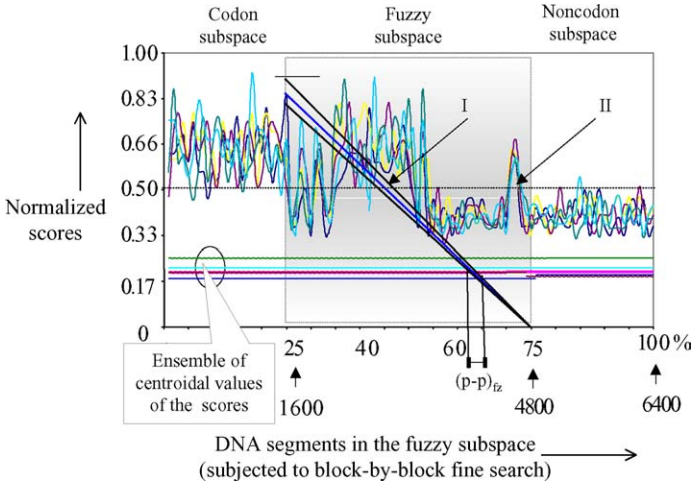

Figure 8 Simulated results on fine-search scores obtained on differential blocks across the fuzzy region (**Z**) of the human DNA sequence (constructed in binary form). The scores correspond to correlation measure for a binary sequence as defined in the text. The scored values are attributed with tuples and the results are defuzzified as follows: (I) Ensemble of slant-lines constructed based on linear membership function: for a given simulation-run, this slant-line joins the maximum score intercept on the ordinate to the minimum score (zero value) across the subspace; (II) Ensemble of scores obtained over repeated simulation runs (p—$p)_{fz}$: defuzzified value of the pointer-positions specifying the codon—noncodon demarcation. (It is shown as an error-bar obtained from ensemble results.)

Table 2 Computed results pertinent to human and bacterial DNAs on the percentage deviations (Δ%) of delineation boundaries in the test fuzzy subsequences (the results correspond to defuzzified centroid values, $(p-p)_{fz}$ determined as per *F*-, *J*- and *B*-metrics and presented relative to the mid-pointer-position $(p-p)_{mid}$ in each subsequence: $\Delta = [((p-p)_{mid} - (p-p)_{fz})/(p-p)_{mid}] \times 100\%$).

Metrics	Human-being		*E. coli*		*R. prowzekii*		*M. jannaschii*	
	Centroidal position $(p-p)_{fz}$	Δ%	Centroidal position $(p-p)_{fz}$	Δ%	Centroidal position $(p-p)_{fz}$	Δ%	Centroidal position $(p-p)_{fz}$	Δ%
Fuzzy subsequences—fSub1: 1600—3200 and fSub2: 9000—10600								
fSub1: mid-position, $(p-p)_{mid} = 2400$								
F	2256	+6.00	2243	+6.50	2247	+6.40	2259	+5.90
J	2197	+8.50	2180	+9.10	2175	+9.40	2187	+8.90
B	2250	+6.30	2250	+6.30	2252	+6.20	2251	+6.20
fSub2: mid-position, $(p-p)_{mid} = 9800$								
F	9744	+0.57	9758	+0.43	9753	+0.48	9823	−0.23
J	9814	−0.14	9807	−0.07	9824	−0.24	9820	−0.20
B	9750	+0.51	9750	+0.51	9749	+0.52	9749	+0.52

(1) Scored and defuzzified values using, *F*: Fisher metric, *J*: Jeffery's (symmetrized KL) measure and *B*: Bhattacharyya distance. (2) The fuzzy scores computed correspond to average of 100 ensemble runs using different seeds on the codon statistics simulated.

8. Concluding remarks

The study addressed here was motivated to seek a strategy that enables identifying the delineation or border separation between codon and noncodon regions in a massive stretch of DNA chain. Specifically, the work considers the situation in which the delineating boundary in question (denoted as $(p-p)_{fz}$ of pointer-positions along the test sequence) is submerged in a subspace of a DNA sequence; and, in that subspace the codon and noncodons exist as overlapped and ambiguous (fuzzy) entities. Such fuzzy considerations for the inference strategy pursued are novel and unexplored hitherto. Ascertaining such a boundary is a practical need in the state-of-the-art bioinformatics. Relevant effort refers to a feature classification problem (and assessing the accuracy of such prediction algorithms for classification purposes is elaborated in [33]).

Commensurate with the scope of IT-considerations vis-à-vis molecular biology indicated in the introductory section, the need for information-theory (in Shannon's sense) and using it in fuzzy domains of bioinformatics is the driving impetus behind this paper. That is, considered here is an effort to fuse the concepts of IT and fuzzy logic so as to evolve appropriate metrics that are useful in bioinformatic efforts. The efficacy of the proposed methods and the success of using the metrics developed can be observed from the simulation results obtained. For example, delineation of boundaries across the subspaces is distinctly made feasible by the proposed metrics and the algorithms pursued as could be evinced from the graphical details presented in Figs. 3—6. Further, referring to Table 2, the different measures adopted consistently yield results on the delineation boundary on test subspaces that are closely located (with respect to a reference pointer-position) within a deviation less than 10%. The ensemble of runs performed has also given consistent results affirming the procedure adopted.

The procedure to apply IT concepts and fuzzy considerations upon a DNA sequence taken in binary format as indicated in this paper is again a novel approach. Analyzing a fuzzy binary sequence depicting a DNA chain is rather new and unexplored (as far as the authors know of). Corresponding application of a correlation measure, such as the Hamming distance, is again an approach perceived in the IT framework. This contemporary pursuit allows a similarity/dissimilarity comparison of DNA sequences taken in binary format (in lieu of traditional symbolic format of such sequences). The efficacy of this new strategy is evident in yielding results within the span of a short error-bar on $(p-p)_{fz}$ over an ensemble of simulation runs as shown in Fig. 8. The scope of this study still has open-questions on applying exhaustively all other IT-metrics (such as the ones explicitly used in this paper as well as those indicated in the relevant literatures [23,26]) for bioinformatic sequence data analysis problems.

Appendix A

If-then or else statements on $2^{12} = 4096$ sets of Θ-values (ED-measures) for the base-phase combinations across the entire DNA sequence

1. If ($\Theta_{Sample}A0 = s_{nc}$) and ($\Theta_{Sample}T0 = s_{nc}$) and ($\Theta_{Sample}C0 = s_{nc}$) and ($\Theta_{Sample}G0 = s_{nc}$), ..., and ($\Theta_{Sample}G2 = s_{nc}$), then output $\Rightarrow s_{nc}$
2. If ($\Theta_{Sample}A0 = s_{c}$) and ($\Theta_{Sample}T0 = s_{nc}$) and ($\Theta_{Sample}C0 = s_{nc}$) and ($\Theta_{Sample}G0 = s_{nc}$), ..., and ($\Theta_{Sample}G2 = s_{nc}$) then, output $\Rightarrow s_{nc}$
3. ...
4. ...

.
.
.

64. If ($\Theta_{Sample}A0 = s_{c}$) and ($\Theta_{Sample}T0 = s_{c}$) and ($\Theta_{Sample}C0 = s_{c}$) and ($\Theta_{Sample}G0 = s_{c}$), ..., and $\Theta_{Sample}G2 = s_{nc}$ then, output = s_{fz}

.
.
.

4095. ...
4096. If ($\Theta_{Sample}A0 = s_{c}$) and ($\Theta_{Sample}T0 = s_{c}$) and ($\Theta_{Sample}C0 = s_{c}$) and ($\Theta_{Sample}G0 = s_{c}$), ..., and ($\Theta_{Sample}G2 = s_{c}$) then, output = s_{c}

Appendix B

Pseudocode on decision-logic algorithm for coarse delineation of codon/noncodon sections across the test DNA sequence

```
Inputs
TOT_MIXED_CODON_SEQ_SIZE
    ←   generation of random sequence/
    string according to the codon statistics
    depicting the probability of occurrence
    of each nucleotide for the DNA of the test
    organism
    ←   creation of test sequence for
    analysis
Initialise:        pointer position ← 0
Compute:  f_c = P_1x (codon occurrence
  frequency)
    ←   for pointer positions (0 to TOT_
    MIXED_CODON_SEQ_SIZE), count the bases
    and calculate probabilities (f_c) for each
    sequence block for (A0, T0, C0, G0, ...,
    A2, T2, C2, G2)
    ←   calculate probabilities for the
    entire sequence
Compute:    Θ-values
    ←   using the average probabilities,
    compute theta (Θ) for all runs for A0, T0,
    C0, G0, ..., A2, T2, C2, G2 on a per block
    basis of the sequence
    ←   Euclidian distance measures
    (Θ-values) taken in a normalised scale of
    0 to 100% correspond to the following:
    Θ = 0 refers to the probability
    distribution of the population (scanned
    with pointer positions) is uniform with
    f_nc = P_2y (noncodon occurrence frequency =
    1/64). Θ = 100% refers to the probability
    distribution of the population (scanned
    with pointer positions) refers to codon
    statistics with f_c = P_1x (that is, codon
    occurrence frequency computed)
If-then ... or else rule:    ←    (Table 2)
    ←   for each block compute the outputs
    of 4096 groups of if...else rules:
                If majority of score of
    if...else rules is codon, then output
    function should be codon rule
                If majority of score of
    if...else rule is noncodon, then output
    function should be noncodon
                Or else, output member
    function should be fuzzy rule
Write:
    ←   display fuzzy decision values for
    each block scanned by the pointer
    ←   establish X, Y and Z domains
End
```

Appendix C

Fine-search of codon/noncodon delineation in the fuzzy space (Z) using Fisher discriminant metric

```
Input:    ←   fuzzy domain (Z) identified and
    extracted as per the decision algorithm
    of Appendix B
Initialise:      pointer position in the
fuzzy
    domain (Z) ← 0
Compute:    ←   step 1
          ←  reference subspaces XR and YR
    in the fuzzy domain are scanned across
    their entire set of respective
    differential event-spaces; and, for each
    event-space (differential window), the
    score of numerical count of bases
    corresponding to the base-phase
    combinations is computed
Compute:    ←   step 2
          ←  mean of all twelve measurements
    (of the base-phase set) is determined
    for each event-space (of the differential
    window subjected to scoring) across the
    aligned pair of reference blocks; and,
    the difference in each measurement
    pertinent to the two blocks is computed.
    The results are then stored as a
    twelve-element difference vector set
Compute:    ←   step 3
          ←   a covariance matrix for all
    the twelve measurements is obtained and
    inverted
```

Compute: ← step 4
← set of coefficients (λ_r) of the linear discriminant function is determined by the product of the inverted covariance matrix times the twelve element difference vector indicated in step (2).
End

Appendix D

Centre-of-area (CoA) or moment method of defuzzification

Initialise: pointer position ← 0
Write LS: = $\sum L_j$
← sum of all pointer-positions across all the differential windows of the test fuzzy region $(\mathbf{Z})_j$
Write SS: = $\sum S_j$
← sum of all scored values (output values obtained with the test-metric based computations) of the windows mapped into corresponding value 0 attributes of the pointer-positions in the fuzzy region $(\mathbf{Z})_j$
Write SP: = $\sum_j (L_j) \times (S_j)$
← sum of the product of L_j and S_j for all pointer-positions of the fuzzy region $(\mathbf{Z})_j$
Write YC: = (SP/LS)
← y-coordinate of the centroid of the fuzzy region, $(\mathbf{Z})_j$
Write XC: = (SP/SS)
← x-coordinate of the centroid of the fuzzy region, $(\mathbf{Z})_j$
End

Appendix E

Correlation-metric-based assessment of codon–noncodon delineation in the fuzzy subspace

Input: ← invoke the subspace strings $\mathbf{X}_{bs}$, $\mathbf{Y}_{bs}$ and $\mathbf{Z}_{bs}$
Initialise:
← The binary subspace strands $\mathbf{X}_{bs}$, $\mathbf{Y}_{bs}$ are divided into blocks of differential windows and aligned with a fuzzy subspace $\mathbf{Z}_{bs}$. (The size of the differential block decides the resolution and accuracy of the final result)
Compute:
← computation performed refers to finding the score that informs the relative profile of the block being a codon or a noncodon in each differential window. The scoring strategy, namely, assessing the correlation measure R corresponds to the Hamming distance equal to the difference between the level of approval (meaning number of 0s) and the level of disapproval (meaning number of 1s) counted in the outcomes of a modulo-2 operation, $\mathbf{Z}_{bs} \oplus \mathbf{Y}_{bs}$ [13]
Compute:
← scores obtained for each differential window are normalized with respect to the maximum value along the strand. The normalized score (in the scale of 0–1) is represented in terms of tuples corresponding to some membership levels, (based on subjective reasoning expressed in linguistic terms). For example, the tuples can be specified as three levels such as: (0–0.6): Low; (0.6–0.8): Medium; and, (0.8–1): High. (Here, the linguistic descriptions, namely, {low, medium, high} are such that, the relative score of the highest value (equal to 1) conforms to ''all-codon'' conditions and the lowest value of the score (equal to 0) refers to an ''all-noncodon'' condition)
Compute:
← Next, the scores specified in their tuples across the differential windows are subjected to defuzzification procedure using the centroid method; and, a horizontal line is drawn through the centroid value of the score. Assuming a linear membership function across $\mathbf{Z}_{bs}$, a slant line is then drawn from the highest score value to the lowest (zero) score value; The intersection of the horizontal line and the slant line defines the delineation point, $(p–p)_{fz}$
End

References

[1] Claverie JM, Notredame C. Bioinformatics for dummies. New York: Wiley, 2003.
[2] Kullback S, Leibler RA. On information and sufficiency. Ann Math Statist 1951;22:79–86.
[3] Pevzner PA. Computational molecular biology: an algorithmic approach. Cambridge, MA: MIT Press, 2001.
[4] Chela-Flores J, Raulin F, editors. Chemical evolutions: physics of origin and evolution of life. Dorchrecht, The Netherlands: Kluwer Academic Press, 1996.
[5] Schneider TD. Sequence logos, machine/channel capacity, Maxwell demon, and molecular computers: a review of the theory of molecular machines. Nanotechnology 1994;5:1–18.
[6] Schneider TD. Evolution of biological information. Nucl Acids Res 2000;28:2794–9.

[7] Standish TG. Structure and function in noncoding or ''junk'' DNA. In: Kennedy ME, editor. Proceedings of the BRISCO Meeting. Loma Linda, CA, USA: Geoscience Research Institute; 1995. p. 4—5.
[8] Bishop M, editor. Genetics databases. San Diego, CA: Academic Press, 1999.
[9] Neelakanta PS, Arredondo TV, De Groff D. Redundancy attributes of a complex system: application to bioinformatics. Complex Syst 2003;14:215—33.
[10] Schneider TD, Mastronarde DN. Fast multiple alignment of ungapped DNA sequences using information theory and relaxation method. Discrete Appl Math 1996;71:259—68.
[11] Nishikawa K. Information concept in biology. Bioinformatics 2002;18:649—51.
[12] Chang BCH, Halgamuge SK. Fuzzy sequence pattern matching in zinc finger domain proteins. In: Hall L, Smith M, Gruver B, editors. Proceedings of the Joint Ninth IFSA World Congress and 20th NAFIPS International Conference. New York: IEEE Press; 2001. p. 1116—20.
[13] Chang BCH, Halgamuge SK. Protein motif extraction with neurofuzzy optimization. Bioinformatics 2002;18:1084—90.
[14] Tomida S, Hanai T, Honda H, Kobayashi T. Analysis of expression profile using fuzzy adaptive resonance theory. Bioinformatics 2002;18:1073—83.
[15] Lusscombe NM, Greenbaum D, Gernstein M. What is bioinformatics? A proposed definition and overview of the field. Meth Inform Med 2001;40:346—58.
[16] Bernaola-Galván P, Grosse I, Carpena P, Oliver JL, Román-Roldán R, Stanley HE. Finding borders between coding and noncoding DNA regions by entropic segmentation method. Phys Rev Lett 2000;85:1342—5.
[17] http://www.kazusa.or.jp/codon/ (last accessed: 22 March 2005).
[18] Buckles BP, Petry FE. Information-theoretical characterization of fuzzy relational data bases. IEEE Trans Syst Man Cybernet 1983;SMC-13:74—7.
[19] Gonnet P, Lisacek S. Probabilistic alignment of motifs and sequences. Bioinformatics 2002;18:1091—101.
[20] Deco G, Obradovic D. An information-theoretic approach to neural computing. New York: Springer-Verlag, 1996.
[21] Kailath T. The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans Commun Technol 1967;Com-15:52—60.
[22] Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inform Theory 1991;37:145—9.
[23] Neelakanta PS, editor. Information-theoretic aspects of neural networks. Boca Raton, FL: CRC Press, 1999.
[24] Wu TJ, Burke JP, Davison DB. A measure of DNA sequence based Mahalanobis distance between frequencies of words. Bioinformatics 1997;53:1431—9.
[25] Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugencis 1936;7:179—88.
[26] Kapur JN, Kesavan HK. Entropy optimization principles with applications. Boston, MA: Academic Press, 1992.
[27] Klug WS, Cummings MR. Concepts of genetics. New York: Macmillan, 2002.
[28] Golomb SW. Digital communications with space applications. Los Altos, CA: Peninsula Publishing, 1981.
[29] Yager RR, Filev DP. Essentials of fuzzy modeling and control. New York: Wiley, 1994.
[30] Csiszár I. Information-type measures of difference of probability distributions and indirect observations. Stud Sci Math Hungar 1967;2:299—318.
[31] Hertz GZ, Stormo GD, Identifying. DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 1999;15:563—77.
[32] Ali SM, Silvey SD. A general class of coefficients of divergence of one distribution from another. J Roy Statist Soc Ser B 1966;28:131—42.
[33] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412—24.