Whites Science
A Scientific Erudition

Estımatıon of Purıne/Pyrımıdıne Infestatıon At Splıce-Junctıons In Human Bıosequences: Relevant Bıomarker-Based Metrıc for Cancer Dıagnostıcs

Perambur Neelakanta*, Inan Ahmed and Dolores De Groff

Department of Computer and Electrical Engineering & Computer Science Florida Atlantic University, Boca Raton, FL 33431, USA

ABSTRACT

This study explains the way of considering purine and pyrimidine residues (regarded as biomarkers) existing across splice-junctions in human genomic sequences; and, any associated artifact details observed thereof are assessed (*via* compatible statistical divergence metrics) so as to diagnose any prevailing pathogenic state (like cancer). That is, considering a domain of purine and pyrimidine infestations as above (identified as biomarkers), it is marked as a sample-space, **X** with no mutations on residues; otherwise, when the residues are mutated (posing artifact details in the biomarkers), relavant sample-space is identified as **Y.** The statistical divergence between **X** and **Y** is proposed here as a possible measure of pathogenic condition due to the presence of mutations (manifesting as artefacts in the biomarkers of **Y**). That is, the residues of **Y** depict a set of randomly comingled items of artefacts due to mutational changes. Hence, the statistical ensembles of artificial sample-spaces, **X** and **Y** simulated denote a pair of biomarker sets with a finite statistical divergence as a result of their common and/or differential signatures. The said statistical divergence is determined using a compatible (cross-entropy) measure. It is expressed in terms of the *ratio of purine and pyrimidine* residues (rP-P) parameter. In summary, the underlying efforts denote *in silico* simulations emulating real-world sample-spaces of **X** and **Y**; and, the computed results on statistical distance are interpreted as possible indications of pathogenic conditions due to the presence of mutations.

Keywords: Biomarkers, purine-pyramidine ratio, splice-junctions, cross-entropy, divergence measure, mutual information, cancer diagnostic metrics, in silico simulations

INTRODUCTION

As well known, certain characteristic molecular signatures possess identifiable details as regard to their level of activities towards genomic and/or proteomic functions in cells. Such unique features of cells are designated and identified as biomarkers. They refer to naturally occurring molecule, gene or characteristic by which a particular pathological or physiological process can be identified. Such biomarkers typically include a range of biochemical entities such as, nucleic acids, proteins, sugars, lipids, small metabolites, etc.; and, they are recognized as possible objective measures with proven diagnostic evaluations on normal, pathological and pharmacological processes enabling efficacious therapeutic interventions towards clinical management of diseases (like cancer) [1-6]. Ascertaining the presence of biomarkers in humans is a part of genome sequecing exercise, which was initially started to

* **Corresponding Author:**
neelakan[at]fau[dot]edu

Receiving Date: November 10, 2020
Acceptance Date: November 27, 2020
Publication Date: December 04, 2020

ISSN 2256-9912

solve the algorithms specific to the four-letters A, G, C, T (depicting the nucleotides adenine (A), guanine (G), cytosine (C) and thymine (T) respectively) of the genome so as to understand the complex nature of human metabolism [7]. Artifacts in biomarkers could possibly exist as comingled items, for example, as purine and pyrimidine residues across trans- splice-junctions of biosequences. Such purines and pyrimidines are building blocks of the nucleic acids (DNA and RNA). They serve as second messengers for signal transduction pathways and nucleotide sugar donors in metabolic pathways. The purines refer to the nucleotide pair: (A) and (G) and, pyrimidine correspond to the pair (C) and (T). These are found primarily in DNA; and, uracil (U) is seen in RNA replacing (T).

Suppose such pairs of biomarkers (depicting stretches of purine plus pyrimidine residues) exist across trans- splice-junctions of biosequences of human test subjects (identified with and without pathogenic states, (like cancer). The study proposed here envisages simulation experiments to assess objectively the relative infestations of purine and pyradimine residues present in the trans- splice-junction; and, hence find relevant *ratio of purine and pyrimidine* (rP-P) data correlated to two situations, one for the test subject having no pathogenic condition (say, cancer) and the other for an affirmed cancerous state. Such studies relating purine-pyridimine *vis-à-vis* development of tumors for example is addressed in [8]. In view of the considerations in correlating purine-pyridimine details *versus* presence of cancer, the effort pursued here refers to an *in silico* study of simulating ensembles of artificially-mutated purine plus pyrimidine stretches at trans-splice-junctions of test human biosequences. Hence, similar and/or dissimilar signatures of purine-pyridimine infestations at the focused site with and without mutational polymorphism are determined; and, in terms of observed statistics, pertinent, differential rP-P features are estimated. The level of rP-P implying similar and/or dissimilar signatures of purine-pyridimine infestations, is expressed in terms of a compatible informatic (negentropic) measure of statistical divergence; and, the statistical distance so estimated is then correlated to the feasibility of presence or absence of any observable pathogenic state.

Background Details

Studies in recent years on cancer biomarkers have increased the efficiency of detection and efficacy of treatment plus management of cancer. Advancements in such efforts include identifying potential biomarkers; and, specific biomarkers of cancer-related interests include a wide range of biochemical entities such as, nucleic acids, proteins, sugars, lipids and small metabolites, cytogenetic and cytokinetic parameters as well as, whole tumor cells found in body fluids. A comprehensive understanding of the relevance of each biomarker is necessary not only towards diagnosing the disease reliably, but also in finding compatible therapeutic regimens. Various biomarkers considered in practice for diagnosis, prognosis and therapeutic purposes and those already studied and identified in [1,8].

The scope of present study is to use the statistical profiles of a class of biomarkers (purines and pyridimines) as objective sample-space and the underlying statistical divergence measure is suggested towards making diagnostic decisions on the presence or absence of cancerous state. That is, surmised here is that the infestation of a distinct set of biomarkers at specific genomic sites can provide observable distinction between the signatures in the biomarker frameworks. Relevant characteristic distinction can be regarded as an indication of the presence or absence of pathogenic state. That is, proposed here is that the possible differences observed in the signature profiles (of specified regions of genomic and/or proteomic residues) could be due to

ISSN 2256-9912

normal and diseased states; and, the underlying theme of the study thereof refers to following: Distinguishing statistical profiles of biomarkers in the sequences of subjects can lead to a robust assertion of normal and diseased states. As stated earlier, for simulation purposes, the statistical details of purine and pyrimidine infestations at the splice-junctions in human genomes are considered. In parallel, another sample-space of purine and pyrimidine infestations at the same splice-junctions (in human genomes) is framed with artificially introduced artifacts in the profiles of purine-pyrimidine populations. Such deliberately introduced changes emulate “artificial mutations” and introduce (morphs) in the biomarker profile at the test sites implying implicit possibilities of pathogenic conditions.

MATERIALS AND METHODS

In view of central doma of microbiology, the non-coding (intron) segments present in the DNA strand (during transcription) get spliced out at splice-junctions (that delineate adjacent exon-intron or intron-exon segments in the sequence); as such, only exons possessing genetic information are retained in the resulting RNA strand (with the base residue T in the codons changed to another base called uracil, U). Typically, the stretches of codon sites on either side of trans- splice-junctions, are infested with purine and pyridimine residues. Consistent with the objectives of this study, relevant focus is on relative rate of occurrence of purines *versus* pyrimidines at a splice-junction. Measures based on relative extents of infestation of purines and pyrimidines across the splice-junction have been used in studies towards identifying the location of the splice-junction site in a genomic sequence [9].

Presently, the relative frequency of occurrence of purine and pyrimidines across the splice-junction regions is considered; and, such regions of infestations (depicting purine-pyrimidine biomarkers) are viewed as two possible sample-spaces: One without any coexisting mutations depicted as: **X**; and the other sample-space (indicated as **Y**) refers to the region of biomarkers having mutational changes on residues signifying pathogenic conditions. Hence, relevant *in silico* simulations pursued in this study involve: (i) Emulating the up- and down-stream regions at the splice-junction of a human genome with known percentages of purine and pyridimine residues prescribed as randomly mixed entities (depicting a stochastic mixture) [10] and it corresponds to **X**; (ii) constructing a similar framework with (similar percentages of purine and pyridimine residues) except that artificial mutations are randomly introduced as polymorphism on purine and pyridimine specific dinucleotide pairs in the stretches of purines and pyrimidines occurring bilaterally across the splice-junction regions; and, such changes made *via* artificial mutations enable variations in the *ratio of purine-to-pyrimidine* (rP-P) residues; that is, emulation of **Y** conforms to a sample-space of biomarkers with a dispersion of random artefacts that can be eventually correlated to corresponding diseased states. A statistical comparison of associated (relative) entropy profiles of the test regions **X** and **Y** can be regarded as a diagnostic suite for deciding the diseased-state in question. Relevantly, the statistical distance (divergence) between the characteristics of sample-spaces **X** and **Y** refers to the cross-entropy (in Shannon sense) that can be adopted to distinguish the the features of compared sample-spaces as indicated in [11]. In summary, the efforts described in this study primarily address *in silico* simulations performed on emulated test regions: Sample-space, **X** representing infestation of purine and pyrimidine residues at a trans- splice junction (of a genomic sequence of human-beings; and, sample-space, **Y** with designated “artificial mutations” introduced on purine and pyrimidine residues so as to resemble artefacts on biomarkers. Due to possible statistical variations that may inherently prevail between individual

ISSN 2256-9912

subjects *vis-à-vis* residue levels of nucleotides, amino acids and proteins etc., it is suggested in this work that a number of ensemble sample-spaces representing pseudoreplicates of **X** and **Y** are simulated and prediction exercise in elucidating the statistical divergence is done on the average feature of pseudoreplicated ensemble spaces.

Simulation Experiments

As indicated before, **X** depicts the sample-space containing biomarkers constituted by stretches of purine-plus-pyrimidine residues across trans- splice-junctions of a test biosequences; and, the test space **Y** is mutated to exhibit artifacts in the biomarkers (implying a pathogenic state like, cancer). In general, the mutations indicated denote disruptors of splicing region; for example, when they fall on either side of the splice-site, the consensus intronic dinucleotide splice donor, GT, or the splice acceptor, AG, such splice site mutations are presumed to be invariably deleterious because of their disruption of the conserved sequences that identify exon-intron boundaries [12]. Consistent with the aforesaid heuristics, simulation exercises pursued here involve first identifying locations of splice-junctions in the test DNA sequence (for example of human genome). Relevant details on the sites of splice-junctions are confirmed through NCBI database results. A splice-junction so identified and used in simulation experiments of the present study refers to, for example, that located on Chr8 of the human genome at a distinct base-pair (bp). With reference to this site, a stretch of 1000 base-pairs (bp) is considered and the relative extents of the residues of purines and pyrimidines per 100 bp window are listed in Table 1 for both up- and down- streams. That is, summarized in Table 1, are details on the composition of purines-to-pyrimidines per 100 bp (window) for the full stretch of (1000 + 1000) bp on either sides of the splice-junction. Relevant proportions of purines and pyrimidines are specified in terms of percentages (% Pu and % Py) occurring in 20 windows ($W_{i=1,2,\ldots 20}$), each window containing 100 bp of residues.

With reference to the sample-spaces **X** and **Y** defined earlier, the comparative pursuit refers to deciding the relative proportion of purine and pyridimine (expressed *via* ratio measure, rP-P mentioned earlier). As mentioned before, a more comprehensive metric can, however be specified so as to decide the statistical-divergence between **X** and **Y**. Relevant metric duly accounting for the cross-entropic features of the associated populations (of purine and pyrimidine residues) is based on probabilities of occurrence of purines and pyrimidines across the sample-spaces; and, it implies a mutual-information based decision on the distinguishability between statistical distributions of purines and pyrimidines in the test spaces. Essentially, it offers details on the underlying commonality or distinguishability in the information-theoretic sense. A number of such cross-entropy (or mutual information) metrics exists as detailed in [13,14] and, they have been adopted in bioinformatic studies concerning biosequences comparisons [15]. A popular measure thereof refers to the so-called Kullback-Leibler (KL) measure, which is adopted in the present study on the sample-spaces of purine-to-pyrimidine statistics. The underlying stochastic distinguishability is framed in the information-theoretic sense as detailed in [16] *vis-à-vis* biological complexity observed in bioinformatics.

ISSN 2256-9912

Table 1: Percentages of purines (Pu) and pyrimidines (Py) across 20 windows ($W_{i=1,2,\ldots20}$) each containing 100 bp residues with 1000 bp in up- and 1000 bp in down-streams at a splice-junction of a human genome

Down-stream windows	**% Pu**	**% Py**
W1: BP142839552 – 142839652	49%	51%
W2: BP 142839652- 142839752	38%	62%
W3: BP 142839752- 142839852	48%	52%
W4: BP 142839852- 142839952	52%	48%
W5: BP 142839952- 142834052	55%	45%
W6: BP 142840052- 142840152	57%	43%
W7: BP 142840152 - 142840252	57%	43%
W8: BP 142840252 - 142840352	57%	43%
W9: BP 142840352- 142840452	45%	55%
W10: BP 142840452-142840552	46%	54%
Up-stream windows		
W11: BP 142840552-142840652	54%	46%
W12: BP 142841652- 142841752	63%	37%
W13: BP 142841752- 142841852	57%	43%
W14: BP 142841852- 142841952	47%	53%
W15: BP 142841952- 142842052	67%	33%
W16:BP 142842052- 142842152	54%	46%
W17: BP 142842152- 142842252	43%	57%
W18: BP 142842252- 142842352	32%	68%
W19:BP142842352-142842452	44%	56%
W20:BP 142842452 -142842552	48%	52%

Outlined in the following section are computational details in using the aforesaid KL measure for the intended analysis of purine-pyridimine sample-spaces.

Computational Details

Considering a test segment of DNA sequence containing mixed residues of purine plus pyrimidine residues, (for example, as listed in Table 1), it is divided into 20 subsegments (or windows, $W_{i=1,2,\ldots20}$ with each window denoting a receptacle for 100 bp residues as mentioned earlier. Suppose the probability of occurrence of purine is specified as "p(Pu)" and that of pyrimidine as "q(Py)", they correspondingly refer to percentages % Pu and % Py respectively seen in each window. That is, the prorated values of purine and pyrimidine contents per window as in Table 1, concomitantly refer to p(Pu) and q(Py) denoting respectively pertinent probabilities of occurrence of purine and pyridimine in each window of the test sample-space. Further, the residues in each i^{th} window of the test sample-space depict a statistical mixture of purine and pyramidine populations; and, based on relative proportion of such purine and pyramidine residues in the mixture, the resultant mixture property (denoted as Q_i) in each (i^{th}) window can be ascertained *via* Lichtenecker-Rother algorithm [10] (based on logarithmic law of mixing). That is, for any arbitrary fractional level of purine, ($0 \leq \Theta_i \leq 1$) and corresponding fractional level of pyridimine ($1 - \Theta_i$), in the binary mixture content (of i^{th} window), the mixture-theoretic model for Q_i is given by the following Lichtenecker-Rother algorithm:

ISSN 2256-9912

$Q_i = [p(Pu)^{\Theta i}]_i \times [q(Py)^{(1- \Theta i)}]_i$ ………………………….(1)

Thus, the parameter Q_i in equation (1) implicitly denotes the resultant probabilistic attribute of the inclusions (namely, purine and pyramidine contents) present in the statistical mixture-state of the window.

Further, with reference to purine *versus* pyrimidine population (in each window of the sample-space), the statistical divergence measure can be specified *via* Kullback-Leibler ($KL_i = KL_{1i} + KL_{2i}$) formulation [13-15] as follows:

$KL_{1i} = [p(Pu) \times \log_e\{p(Pu)/q(Py)\}]_I$ nats …………………….(2a)
$KL_{2i} = [q(Pu) \times \log_e\{q(Py)/p(Pu)\}]_I$ nats …………………….(2b)

In the present study, the simulation exercise as stated earlier involves constructing artificial sample-spaces (**X** and **Y**) for each window representing a receptacle containing a mixture of appropriately prescribed and prorated extents of purines and pyridimines (with or without mutations). Hence, each simulated window has a corresponding value of Q_i as decided by equation (1). Pertinent simulation steps in constructing the artificial sample-spaces as above are outlined below in the pseudocode.

A pseudocode on simulating an artififial sample-spaces (depicting a set of windows, each representing mixture-medium of prorated contents of random extents of 100 bp of purines (Pu) and pyrimidines (Py).

The stretch of $W_{i=1,2,\ldots 20}$ windows is constituted by 1000 bp of residues in up-stream plus 1000 bp of residues in down-stream at a splice-junction in the human genome. There are two sample-spaces being simulated, **X** (without any mutations enforced on purines and pyridimine contents in each window) and, **Y** having ± 20 % mutations randomly introduced in the contents of purines and pyridimines across the windows.

```
---------------------------------------------------------------
%  The pseudocode outlines two parts of simulations:
   Part I refers to constructing the sample-space, X
   Part II refers to details on simulating sample-
   space, Y.
---------------------------------------------------------------------
%% Part I Constructing the sample-space,  X

Initialization

→  Step I.1: A stretch of twenty windows ($W_{i=1,2,\ldots 20}$ ) is framed with each window
   accommodating 100 bp of residues across ten up-stream and ten down-stream windows
   symmetrically specified at a splice-junction in a human genome.

Assigning a Set of N = 10 Random Values of the Fraction, ($0 \le \Theta_i \le 1$) for Each Window
```

→ Step I.1: A set of $\{(\Theta_i)\}_{n=1, 2, \ldots, N=10}$ is generated, each with ($0 \leq \Theta_i \leq 1$) depicting a uniformly-distributed random numbers and posted in ith window, W_i of the set, $\{W_i\}_{i=1, 2, \ldots, 20.}$

Framing an Ensemble of Ten Pseudoreplicates (*via* Bootstrapping) of each of the sets$\{(\Theta_i)\}_{n=1, 2, \ldots, N=10}$ to Fill Each Window

→ Step I.2: Considering the prescribed set of $\{(\Theta_i)\}_{n=1, 2, \ldots, N=10}$ (as decided above) in each window of the set $\{W_i\}_{i=1, 2, \ldots, 20,}$ the contents of each set $\{(\Theta_i)\}_n$ is shuffled *via* bootstrapping procedure [17][18] so as to get M = 10 pseudoreplicates; that is, each window now has M = 10 peudoreplicated sets $\{[(\Theta_i)]_n]\}_{m=1, 2, \ldots, M=10.}$

→ In summary, for any ith window, a set of M = 10 pseudoreplicated sets are formed each consisting of N =10 shuffled-values of Θ_i as listed below:

$\{(\Theta_i)_{n=1,} (\Theta_i)_{n=2, \ldots,} (\Theta_i)_{n=N=10}\}_{m=1}$
$\{(\Theta_i)_{n=1,} (\Theta_i)_{n=2, \ldots,} (\Theta_i)_{n=N=10}\}_{m=2}$
....
$\{(\Theta_i)_{n=1,} (\Theta_i)_{n=2, \ldots,} (\Theta_i)_{n=N=10}\}_{m=M=10}$

→ That is, each ith window holds a total of ($N \times M = 100$) Θ_i-values; and, corresponding mean fractions Θ_{iA} and ($1 - \Theta_{iA}$) are calculated.

Finding in each ith Window, the Ensemble Set of: $\{Q_i\}_{M\times N}$ and Its Mean Value (Q_{Ai})

→ Step I.3: This step involves finding possible ($N \times M = 100$) values of Q_i using equation (1) for each ith window with ($0 \leq \Theta_i \leq 1$), corresponding ($1 - \Theta_i$) values and the pair $[p(Pu)]_i$ and $[q(Py)]_I$ taken from Table 1. For example, with reference to W_1 : BP142839552 – 142839652, p(Pu) is equal to 0.49 (49 %) and q(Py) refers to 0.51 (51 %).

→ That is, for any ith window and the associated pair of values $[p(Pu)]_i$ and $[q(Py)]_i$ availed from Table1, the mean fractions Θ_{iA} and ($1 - \Theta_{iA}$) ascertained earlier are applied in the relation of equation (1) to get:

$$Q_{iA}= ([p(Pu)]_i)^{\Theta iA} \times ([q(Py)]_i)^{(1-\Theta iA)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Estimating KL Values: $(KL_i)_{MN}$ for Each Window

→ Step I.4: The ratio of mean fractions $[\Theta_{iA} /(1 - \Theta_{iA})]$ is determined for each window and it denotes the relative extents of purine-to-pyrimidine population in the artificial sample-space being simulated.

→ In terms of $[\Theta_{iA}/(1 - \Theta_{iA})]$, the associated rP-P can be implicitly specified *via* mean-value formulation of statistical divergence, ($KL_{iA} = KL_{1iA} + KL_{2iA}$) deduced from equation (2) as follows:

$$KL_{1iA} = [\Theta_{iA} \times \log_e\{ \Theta_{iA} /(1 - \Theta_{iA})\}]_i \quad \text{nats} \quad (4a)$$
$$KL_{2iA} = [(1 - \Theta_{iA})\times \log_e\{(1 - \Theta_{iA})/\Theta_{iA} \}]_i \quad \text{nats} \quad (4b)$$

ISSN 2256-9912

%% Part II Constructing the sample-space, Y

Initialization

% The procedures of Steps I.1 through I.4 of Part I simulations are repeated with a fresh set of random variables of ($0 \leq \Theta_i \leq 1$) and framing corresponding pseudoreplicates; and, a change is included to introduce mutations so that the emulated sample-space corresponds to, **Y** as detailed below:

→ Step II.1: The set of percentages of purines (Pu) and pyrimidines (Py) across 20 windows ($W_{i=1,2,\ldots 20}$) is considered (as per details in Table 1) and modified with a random change (say, to a maximum extent of ± 20 %) to depict the mutational changes. Shown below in Table 2, is an exemplar of such modified data set

Table 2: Percentages of purines and pyrimidines across 20 windows ($W_{i=1,2,\ldots 20}$) each containing 100 bp residues with 1000 bp in up- and 1000 bp down-streams at a splice-junction of a human genome with and without random changes for artificially imposing mutations

Without random mutations (Table 1 values)				With random mutations (of ± 20 % on Table 1 values)	
Down-stream windows					
	% Pu	% Py		% Pu	% Py
W1:	49%	51%		54%	56%
W2:	38%	62%		34%	60%
W3:	48%	52%		53 %	47%
W4:	52%	48%		47%	54%
W5:	55%	45%		60%	49%
W6:	57%	43%		59%	48%
W7:	57%	43%		58%	38%
W8:	57%	43%		62%	48%
W9:	45%	55%		43%	57%
W10:	46%	54%		50%	58%
Up-stream windows					
W11:	54%	46%		51%	49%
W12:	63%	37%		59%	42%
W13:	57%	43%		54%	44%
W14:	47%	53%		41%	55%
W15:	67%	33%		70%	36%
W16:	54%	46%		56%	49%
W17:	43%	57%		49%	59%
W18:	32%	68%		35%	63%
W19:	44%	56%		40%	50%
W20:	48%	52%		42%	55%

Next

→ Step II.2: With the values of $[p(Pu)]_i$ and $[q(Py)]_i$ in Table 3 modified to include random mutations to an extent of ± 20 % changes (on Table 1 values) are considered.

→ Then Steps I.3 through I.4 are exercised with mutated data on $[p(Pu)]_i$ and $[q(Py)]_i$ plus the fresh set of random variables of ($0 \leq \Theta_i \leq 1$) and corresponding pseudoreplicates

→ Hence, details as listed in Table 3 are obtained *via* relevant computions of KL_{iA}-values for the sample spaces **X** and **Y**

RESULTS AND DISCUSSIONS

In all, a statistically-implied artificial sample-space of purine-pyridimine infestation across up- and down streams at a splice-junction in a human genomic sequence is simulated. This simulation conforms to real-world data on percentages of purines and pyrdimines as in Table 1; and, the artificial sample-space constructed is rendered statistically robust inasmuch as, the residue infestations prescribed therein duly assume required probabilistic attributes *via* random variables, Θ_i and ($1 - \Theta_i$). These random variables are uniformly-distributed in conformance with Lapalace's hypothesis on unspecified/unknown probabilistic distributions. Further, by simulating a set of 100 pseudoreplicated ensembles of Θ_i and ($1 - \Theta_i$), their mean values of Θ_{iA} and ($1 - \Theta_{iA}$) are ascertained; and, a relevant estimation of statistical divergence namely, the KL-measure, KL_{iA} is done for each ith window.

Table 3: Computed KL_{iA}-value for each window of 100 bp length across up- and down-stream stretches of (1000 + 1000) bp bilateraly placed at the splice-junction located on Chr8 of the human genome at base-pair (bp) 142839552

W_i	KL_{iA} value (in nats)	
	Without random mutations (Table 1 values)	With random mutations (of ± 20 % on Table 1 values)
Upstream		
W1	0.0008	0.0007
W2	0.1175	0.1477
W3	0.0032	0.0072
W4	0.0032	0.0097
W5	0.0201	0.0222
W6	0.0395	0.0227
W7	0.0395	0.0846
W8	0.0395	0.0358
W9	0.0200	0.0396
W10	0.0128	0.0119
Downstream		
W11	0.0128	0.0008
W12	0.0138	0.0578

ISSN 2256-9912

W13	0.0395	0.0205
W14	0.0072	0.0411
W15	0.2407	0.2261
W16	0.0128	0.0093
W17	0.0395	0.0186
W18	0.2714	0.1646
W19	0.0289	0.0223
W20	0.0066	0.0351

Relevant domain simulated to a sample space-space **X**; and, a corresponding sample-splace (**Y**) is also constructed with changes to include specified extents of mutations in purine-pyridimine populations. Hence the KL-values for the spaces X and Y are determined as listed in Table 3 and plotted in Figure 1, against window numbers.

In summary, the simulation and computational efforts and results obtained as outlined above correspond to: (i) Constructing an artificially-simulated (statistical) sample-space depicting the mixture property of test residues namely, purine and pyridimine; (ii) relevant property is rendered to conform a set of pseudoreplicated Q-values depicting the mixture-theoretic profile of a unbiased blend of test residues; (iii) using average of pseudoreplicated Q-values (in each test window), the associated statistical divergence measure is ascertained and expressed as, KL_{iA} depicting the Kullback-leibler measure. It specifies the statistical distance between the infestation statistics of purine *versus* pyridimine; and, it is implicitly tied to the rP-P of purine-pyridimine residues in the sample-space. (iv) The artificially-generated test sample-space at the splice-site (representing a mixture with random inclusions of purines and pyridimine contents), is further mutated (artificially) by altering the infestation levels of associated purine-pyridimine contents by randomly introducing, ± 20 % changes; and, corresponding KL_{iA} values are obtained as listed in Table 3. (v) The deduced KL-measures *versus* window-sites at the splice-junction are illustrated in Figure 1, where the curves denote the mean KL_{iA} values as listed in Table 3. The demarcation line of splice-junction is also indicated in Figure 1.

Inferential Remarks and Closure

The simulation exercises addressed in this study and results obtained thereof conform to establishing required statistical profiles of purine-pyridimine infestations at the splice junction in a huan gnomic sequence asper the details in Table 1. Similar trans- splice junction features due to associated residue profiles in genomic sequences are also addressed, for example, in [16,19-22]. Pertinent studies have focused on finding precisely, the delineating exon-intron (or intron-exon) segments. For example, the Shapiro–Senapathy (S & S) algorithm of [21,22] enables predicting splice-sites, exons and genes in animals and plants and they are indicated towards discovering disease-causing mutations (existing at the splice-junctions). Relevant platforms are compatible in modern clinical bioinformatics in terms of diagnostic and therapeutic tasks of Next Generation Sequencing (NGS) technology [15].

ISSN 2256-9912

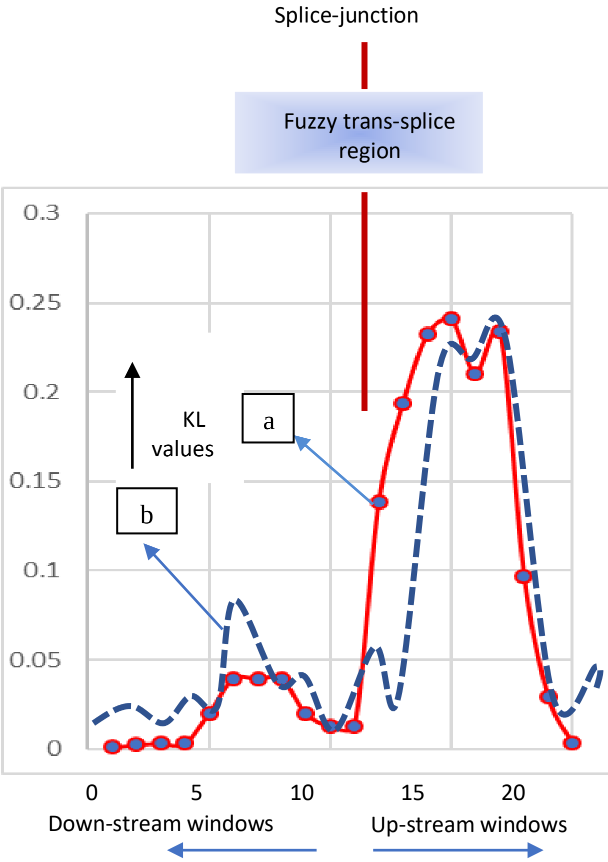

Figure 1: Computed values of average KL_{iA} divergence (in nats) of Table 3 *versus* window segment numbers plotted across up- and down-streams at the splice junction located on Chr8 of the human genome at base-pair (bp) 142839552. (a): Curve regressed on KL-values calculated having no mutational changes in the percentages of purines and pyrimidines present; and, (b): curve regressed on KL-values calculated with ± 20 % mutational changes (as in Table 3) in the percentages of purines and pyrimidines present.

Further, considering the results of the present study, they enable predicting the site of genomic splice-junction in terms of observable, abrupt changes in the composition of residues present in the vicinity of the splice-site. Suppose a normal infestation of residues (say, purines and pyridimine) exists, the location site of the splice-junction distinctly delineates up- and down-stream regions of windows, (each representing a segmented, 100 bp width of residues) in concurrence with the details of [16,20-22]. Further, with reference to the scope of the present study, it is surmised that the presence of any mutational changes in the residues (expressed in terms of the ratio of purines-to-pyridimine popupation) at trans-splice regions, could show characteristic morphs across delineating features (at the splice-site).

The noticeable changes (as in Figure 1) can be specified as follows: (i) With imposed mutational changes, the splice-site is no longer abrupt and precisely demarceted; but, it is seen smeared and ‘fuzzy” [23,25]. (ii) Such fuzzy-splicing in genomic sequences could be indicative of aberrant splice-junctions reflecting the contexts of diseased conditions. (iii) Hence, the mutated purine-

ISSN 2256-9912

pyridimine residues at the splice-site is proposed here as biomarkers towards specifying a pathogenic state (say, for example, cancer). (iv) That is, any observed artefacts in the location (being abrupt, fuzzy or shited) of the splice-site could refer to a diseased state with purine-pyridimine specified as biomarkers; that is, disease-causing mutations in the infested purine-pyridimine residues (at the splice-junctions) can be identified as proposed here as a relevant suite viably adopted in clinical/translational bioinformatics towards underlying diagnostic and therapeutic endeavors.

Fundamental considerations on splice-junction specific site residue characterization in terms of cross-entropy properties can be seen in [26].

Conclusion

Though the present study refers to *in silico* schemes illustrated with artificially-simulated sample-space of biomarkers, the underlying approach can be pursued in NGS exercises of translational base-pair to bed-side efforts of clinical diagonosis. Such studies can lead to discovery of genes causing inherited disorders in terms of associated biomarkers. Specific mutations in different splice sites present in various genes that could cause inherited disorders for example, Type 1 diabetes, hypertension, marfane syndrome, cardiac diseases, eye disorders etc. can be analyzed by the proposed method. Understanding how real-world mutations affect splice-junction composition of biomarkers can lead to promising researching in the future.

REFERENCES

1. K. A. Calzone, KA, Genetic biomarkers of cancer risk. Seminars in Oncology Nursing. 2012 28 (2): 122–128.
2. Herceg, Z and Hainaut P, Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. Molecular Oncology. 2007 1(1): 26–41
3. Sawyers, CL, The cancer biomarker problem. Nature. 2008 452:548-552
4. Calin, GA and Croce, CM, MicroRNA signatures in human cancers. Nature Reviews - Cancer. 2006 6: 857-866
5. Bartels, CL, Tsongalis, GJ, MicroRNAs: Novel biomarkers for human cancer, Clinical Chemistry. 2009 55(4): 623-31.
6. Weber, G, Burt, ME, Jackson, RC, Prajda, N, Lui, MS, and Takeda, E, Purine and pyrimidine enzymic programs and nucleotide pattern in sarcoma. Cancer Research. 1983 43(3): 1019-1023
7. Tutar, L, Tutar, E, Tutar Y, miRNA and cancer: An overview. Current Pharmaceutical Biotechnology. 2014 15(5):430-437.
8. Danielyan, KE, Yeghiazaryan, TA, Chailyan, SG, Harutyunyan, LR, Harutyunyan, RL, Petrosyan, GS, Purine and pyridimine-linked enzymes and genes are strongly responsible for the development of tumors, particularly glioblastoma multiforme. Computational Molecular Bioscience. 2020 10:73-80.
9. Foster, E, Acceptor Splice Site Prediction. Thesis 2007, Rochester Institute of Technology, (Rochester, NY, USA).
10. Lichtenecker, K and Rother, K, Die Herleitung des logarithmischen Mischungsgesetzes aus allgemeinen Prinzipien der statioenaeren Stroemung. Physikalische Zeitschrift. 1938 32: 255-260.
11. Das, JK, Choudhury, PP, Chaudhuri, A, Hassan, Sk. S., Basu, P, Analysis of purines and pyrimidines distribution over miRNAs of human, gorilla, chimpanzee, mouse and rat. Scientific Reports -Nature. 2018 8(9974): 1-19.

ISSN 2256-9912

12. Jayasinghe, RG, Cao, S, Gao, Q, Wendl, MC, Vo, NS, et al., Systematic analysis of splice-site-creating mutations in cancer, Cell Reports. 2018 23(1): 270–281
13. Neelakanta, PS (1999) Information-Theoretic Aspects of Neural Networks, CRC Press, Boca Raton, FL:USA)
14. Kapur, JN., Kesavan, HK (1992) Entropy Optimization Principles and Their Applications. Springer Netherlands.
15. Neelakanta, PS (2020) Text Book of Bioinformatics -Information-theoretic Perspectives of Bioengineering and Biological Complexes, World Scientific, Singapore
16. Neelakanta PS, Arredondo TV, and De Groff D, Redundancy attributes of a complex system: Applications in bioinformatics. Complex System. 2003 14:215-233.
17. Efron, B, Tibshirani, R (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC, Boca Raton, FL:USA
18. Varian, H Bootstrap Tutorial. Mathematica Journal. 2005 9:768–775.
19. Targonski, CA Courtney A. Shearer, BT, Shealy, T, Smith, MC, Feltus, FA, Uncovering biomarker
20. Watakabe A, Comparative molecular neuroanatomy of mammalian neocortex: what can gene expression tell us about areas and layers? Development, Growth & Differentiation. 2009 51(3): 343-354.
21. Shapiro, MB, Senapathy, *P,* RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Research. 1987 15(17):7155–7174.
22. Senapathy, P, Shapiro, MB, Harris, NL, Splice junctions, branch point sites, and exons: Sequence statistics, identification, and applications to genome project. Methods in Enzymology. 1990 183:252–278.
23. Neelakanta, PS, Chatterjee, S, Pavlovic, Pandya, M, De Groff, D, Fuzzy splicing in precursor-mRNA Sequences: Prediction of aberrant splice-junctions in viral DNA context, Journal of Biomedical Science and Engineering. 2011 4:270-279
24. Arredondo, TV, Neelakanta, PS, De Groff, D, Fuzzy attributes of a DNA complex: Development of a fuzzy Inference engine for codon-“junk” codon delineation. Artificial Intelligence in Medicine, 2005 35: 87-105
25. B. C. H. Chang, BCH, Halgamuge, SK, Protein motif extraction with neurofuzzy optimization. Bioinformatics, 2002 18:1084-1090
26. P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, H. E. Stanley, Finding borders between coding and noncoding DNA regions by entropic segmentation method, Physical Review Letters. 2000 85:1342-1345.

ISSN 2256-9912