Application of Subset Theory towards Solution of Functional Diversity Paradox

Muhammad Naeem[1], Saira Gillani[2] and Sohail Asghar[3]

[1]*Department of Computer Science, Mohammad Ali Jinnah University, Islamabad, Pakistan*
[2]*Centre of Research in Networks & Telecom (CoReNeT), Mohammad Ali Jinnah University Islamabad, Pakistan*
[3]*University Institute of Information Technology PMAS-Arid Agriculture University Rawalpindi, Pakistan*
naeems.naeem@gmail.com, sairagilani@yahoo.com, sohail.asg@gmail.com

Abstract

Alternative splicing (AS) and Gene duplication are two well known evolutionary and ubiquitous mechanisms. Both have the common gist for conveying the functional diversity by means of escalating gene variegation. The objective of this study is to understand the established hidden relationship between both of these mechanisms to find the answer for the long standing puzzle circumventing the evolution of genome size while utilizing the computational power. We investigated that the alternative splicing can promise to unveil the nature as well as operation of cellular codes meticulously which play inevitable role of combinations of regulatory elements in pre-mRNA. This study also describe in detail of the cellular complements of splicing regulators, which collectively establish regulated splicing pathways. We have proposed a computational framework based on subset theory to address a part of the inexplicable functional diversity paradox which is also known as c-value enigma. The operational detail of this framework has been discussed to give an insight into the possible solution towards the problem of more proteins than their corresponding genetic material.

Keywords: *Alternative splicing, C-value enigma, Functional diversity paradox, Regulatory mechanism, Intron, Exon, Subset theory problem*

1. Introduction

Machine C-value may be defined as the amount of DNA in a haploid nucleus or one half the amounts in a diploid cell of a eukaryotic organism. Alternatively the concept of C-value may be used interchangeably to genome size. C-value is described in weight (picograms) or counting of base pairs (bp). A single sperm of *Drosophila melanogaster* for example contains $1.64x10^8$ nucleotide pairs or 0.18 picogram (pg) of DNA where C-value in pg is computed by assuming 9.13 X 10^8 nucleotide pairs per picogram [1]. The *C-value enigma* or *C-value paradox* is a term used to describe the complex puzzle encompassing the widespread variation in nuclear genome size among eukaryotic species. The C-value enigma undoubtedly goes beyond taxonomic boundaries leading towards a heightened demand to study the genome size evolution, whether in animals, plants or other organisms. The core idea underlying the C-value paradox is spurred by the fact that no serious and straight forward correlation found between morphological or organismal complexity and genome size. This reality can be highlighted by the fact that some single-celled protists have been observed with genomes much larger as compared to humans though human's genome is far much complex.

Such observation may negate the popular evolutionary theory. Briefly but important and independent component question of the C-value enigma can be summarized as:

What types of non-coding DNA are found in different eukaryotic genomes, and in what proportions?

Gregory [2] illustrated some interesting statistical values related to C-value enigma for haploid nuclear DNA over the data summarized. The graph is plotted between C-value in picogram versus various species. Vertical line in each bar is portraying the average of C-value. The graph is depicting that most C-value in animals as well as in plants are small with a few exceptions of some groups. Gregory [2] showed in this graph that there is no correlation or connection ever observed so far between the complexity of different species and their related genome size. Gregory in quest of investigation for C-value paradox reported that variation in genome size is principally due to direct selection on the amount of bulk DNA through its underlying effects on cell volume and organismal cellular characteristics [2, 3, 4].

Second part of this puzzle narrates that "*From where does this non-coding DNA come, and how is it spread and/or lost from genomes over time?*"

C-value puzzle has been a renowned problem in the field of molecular biology. It can be concluded that species related specific characteristics emerged as a surprise to the early researchers where several orders of magnitude of genome materials was revealed among various eukaryotes. It was stated that "*the lowly liverwort has 18 times as much DNA as we, and the slimy, dull salamander known as Amphiuma has 26 times our complement of DNA"* [5].

Third part of the puzzle describes: "*What effects, or perhaps even functions, does this non-coding DNA have for chromosomes, nuclei, cells, and organisms?*"

A possible answer to solve this puzzle leads to develop two broad kinds of theories. These include *mutation pressure* and *optimal DNA* theory. In *mutation pressure* theory, major portion of non-coding DNA in eukaryotic genomes is considered as *junk* or *selfish* DNA. The role of non coding DNA is limited to only evolution of secondary DNA through transcriptional and translational regulation of protein-coding sequences. On the other hand optimal DNA theory is focused on highlighting the strong connection between DNA content, cell and nuclear volumes. Most of the available evidences go in favor of later theory. Gregory (2001) developed a model of nucleotypic influence which falls under the category of *optimal DNA* theory. They by means of the result obtained from this model, concluded that large amounts of DNA exhibit large and slowly dividing cells.

Fourth part of the C-value enigma argues: "*Why do some species exhibit remarkably streamlined chromosomes, while others possess massive amounts of non-coding DNA?*"

Analysis of genome size is directly related to the understanding of the functions of genetic material. Such analysis leads to the scrutinization of mRNA diversity which acts as a signature of genome functionality. The foundation of novel genes and its associated new functionalities is an important step towards evolution of organisms [6]. It was reported that more than seventy percent of whole human genetic material can generate multiple transcripts via alternative splicing [7, 8]. Exon shuffling, gene duplication, lateral gene transfer, retro position, *de novo* origination and gene fission/fusion are considered essential means in novel genes generation [8, 9, 10]. Alternative splicing can be considered as an essential step towards proteome diversity and transcriptome. This makes alternative splicing an exciting and appealing area of interest during the evolutionary course. Although new genes typically

evolve swiftly in structure, sequences and expression, the evolutionary pattern of alternative splicing in new genes and the molecular mechanisms involved in this alternation still needs to be investigated in depth [6].

In the domain of molecular biology, splicing can be described as a posttranscriptional alteration of RNA. The discovery of RNA splicing was made during 1970s, an era considered as overturning years in the domain of gene expression. In the same period it was also discovered that some of the RNA molecules have the ability to splice themselves; this discovery of self-splicing ability in the protozoan *Tetrahymena thermophila* was awarded with the Nobel Prize in 1989. Splicing is characterized by the removal of intron followed by fusion of remaining exons. Splicing is a mandatory requirement for the classic eukaryotic mRNA prior to its translation into proteins. Splicing in majority of eukaryotic introns is performed in a chain of reactions. This chain of reactions includes catalyzation by the spliceosome, a complex of small nuclear ribonucleoproteins, and self-splicing introns. The exon composition of the same mRNA can be rearranged in diversified patterns yielding a variety of unique proteins. In molecular biology, such phenomenon can be described as alternative splicing. Alternative splicing can take place in variety of ways including extending exons, skipping exons or even retaining one or more introns.

The last but not least part of the quiz give forth the question: "*Why does the number of proteins exceed the number of protein coding genes?*"

In this paper, the study investigated is explicitly not an examination of all of the questions formulated in notable C-value enigma, rather we have targeted towards the last question of this paradox. We have proposed a computational approach towards the solution of the C-value enigma. Our study pointed out that a part of this problem can be resolved by means of mathematical set theory problem. Our approach is focused on converting a big set into smaller sub set. The cross intersection of these entire sub-sets may give rise to all of the possible candidate protein out of DNA/mRNA genetic material.

Wakamatsu, *et al.*, [11] examined the mRNA diversity of genes after inducing neuronal differentiation in human NT2 teratocarcinoma cells using all-transcriptional retinoic acid (RA). Their examination brought the results that 274 genes generated multiple protein-coding transcripts by means of alternative splicing. They also identified that gene exhibiting different RNA-induced alters in the expression of their protein-coding transcripts. Zhan, *et al.*, [6] identified new genes in the Drosophila melanogaster lineage while investigating alternative splicing and possible functional consequences of these genes. They also argued that new genes tend towards exhibiting low degree of alternative splicing. They highlighted that loss of introns in retroposed new genes is responsible for one third of the low-level alternative splicing, whereas partial gene duplication without alternative splicing exons and mutations in the duplicated alternative splicing exons/introns collectively have resulted in two-third alternative splicing losses in new genes. They concluded that reducing the degree of alternative splicing is a broader inclination towards all categories of new genes. They described that alternative splicing with tissue expression pattern of new genes had somewhat less expression level. Zhan, *et al.*, [6] concluded that these new genes are likely to have gained diverged structures and expression patterns from their parental genes after alternative splicing. Matuda, *et al.*, [12] investigated alternative splicing for dihydrolipoamide succinyltransferase (DLST) gene which is a mitochondrial protein. They identified an uncharacterized protein found reacting with an anti-DLST antibody in the I bands of myofibrils in rat skeletal muscle. Their results implied that a pyrophosphate concentration N10 mM was mandatory to elicit the protein from myofibrils in the presence of salt with a higher concentration than 0.6 M, at an alkaline pH of 7.5–8.0.

Jin, *et al.*, [13] described that gene duplication and alternative splicing are two main evolutionary mechanisms responsible for the functional variation by means of enhancing gene diversification. They implied that the proportion of alternative splicing loci and the average number of alternative splicing isoforms per locus are observed larger in duplicated genes in comparison to those in singleton genes while establishing the evolutionary relationship between gene duplication and alternative splicing. They also showed that small gene families have larger proportion of alternative splicing while larger average number of alternative isoforms per locus than large gene families.

Lareau, *et al.*, [14] argued that human genome has far fewer genes than expected. They described that alternative splicing may give rise to proteome expansion while bridging a perceived complexity gap. They also argued that alternative splicing may be commonly believed to affect more than half of human genes. It was concluded that alternative splicing may lead to multiple alternatively spliced mRNA isoforms, resulting generation of mature mRNAs and ultimately polypeptides which can be remarkably related or highly different while originating from the same locus [15, 16]. Black (2003) implied that alternative splicing is a core model in genetic regulation. They described that diversity in splicing patterns is a crucial source of protein diversity from the genome. They discussed about the combination of tissue specific splicing complex system using the Drosophila sex determination pathways.

It was reported this fact that there exist a violation of the concept stating ‘one gene, one polypeptide’ axiom [16]. They argued that alternative splicing permits individual genes to express multiple protein isoforms. Such role of alternative splicing eventually results in generating complex proteomes. It was also described that in the area of quantitative gene control, alternative splicing may exhibit many concealed functionalities. Exhibitions of these hidden functionalities can be achieved by targeting RNAs for nonsense-mediated decay. They described that conventional gene-by-gene examination of alternative splicing mechanisms can be accomplished by global approaches.

Marcus [17] proposed a partial solution to the C-value enigma. They described that during half century, no explicit statistical comparisons between C-value size to its corresponding organism complication have been presented. However, performing their experimentation of sequenced genomes on 139 prokaryotic and eukaryotic organisms they highlighted that there are noteworthy positive correlations may likely exist between measures of genome size to its complexity in comparison to measures of non-hierarchical morphological complexity. They observed that these correlations are non-trivial to correction for phylogenetic history by means of independent contrasts. They reported that the C-value enigma is likely to be more apparent in smaller genomes organisms. They also reasoned out that morphological complexity and genome complexity correlate positively with one another considerably in fact.

Koren, *et al.*, [18] described that Alternative 3′ and 5′ splice site events comprise a non trivial role among all of the alternative splicing events. They showed the correlation of these events to various splicing diseases. They also described that a limited distinguishing features have been identified so far which are related to alternative cassette exons. They identified important features of constitutive exons, alternative cassette exons, alternative 3′ss and 5′ss exons. Their results discovered that alternative 3′ss and 5′ss exons are observed in between constitutive and alternative cassette exons, while the constitutive side and alternative side look a lot like constitutive exons, and alternative cassette exons respectively. The other interesting discovery was related to symmertry level feature of both altervative 3′ss and 5′ss exons showing low frame-preservation effect while the sequence between both of them exhibit high symmetry levels.

2. Methodology

Based on the literature review investigated, we concluded that though a partial answer to C-value enigma appeared when non-coding DNA was explored. However, C-value enigma related to one protein per polypeptide hypothesis still not clearly enlightened. We introduced a computational heuristic to solve the issue of inconsistency between the numbers of proteins to their relevant protein coding genes issue. The input of the proposed technique in this methodology will be DNA material in *fasta* format. The final outcome of the technique will be a complete list of all proteins which can be expressed out of any given nucleotide sequence of DNA or RNA .

2.1 Statement of Alternating Splicing Problem

Alternative splicing problem is to identify the genomic sequence. A gene is a set of exons, intron and poly-A tails. Each of them is a substring of genomic sequence G such that $G_i \in Exon \mid Intron \mid Poly - A$ where $\Gamma = \{g_1, g_2, ...g_n\}$ is of substring with $g_1 \prec g_n$.

Algorithm 1 describes the possible implementation of alternative splicing technique. It starts by selecting the genes nucleotide sequences. The next step is to initialize these sequences. The algorithm selects each of the genomic sequence comprising of a series of nucleotide sequence. Each genomic sequence is subjected to all of the steps as described in the Figure 2 of proposed.

$G = \{g_1, g_2, g_3 .... g_n\}$

$GeneSet\, G = \{g_1, g_2, g_3 ... g_n\}$

$\forall g \in G$

$if\ \ ValidateNucleotideBase(g) = false$

$G = G - g$

$if\ \ startSignal(g) = false$

$G = G - g$

$if\ \ stopSignal(g) = false$

$G = G - g$

$if\ countBaseNucleotide(g) <> threshold$

$G = G - g$

$g = \bigcup_{b_i \in B} \because B \subseteq (Adenine \vee Uracine \vee Cytocine \vee Guanine)\}$

$\forall b, b \in g$

$if\ (b = Mar\ker[i])$

$B \leftarrow B \cup b$

$sp \leftarrow Identification(g, Codon, B)$

$Splicesite \leftarrow Splicesite \cup sp$

$in \leftarrow Identification(g, B)$

$Intron \leftarrow Intron \cup in$

$en \leftarrow Identification(g, B)$

$Exon \leftarrow Exon \cup en$

$$\begin{aligned}
&\text{if } g \in \{SingleExon, SingleIntron \\
&\qquad\qquad , SinglePolyA, SingleUTR\} \\
&\quad Stop \\
&\forall in \in \text{int}\, ron \wedge \forall en \in exon \wedge \forall pl \in PolyA \\
&possible_proteinSet \leftarrow Combine(in, ex, pl) \\
&\forall pt \in possible_proteinSet \\
&\quad proteinSet \leftarrow UniprotAPI(pt)
\end{aligned}$$

Figure 1. Algorithm for Computational Solution

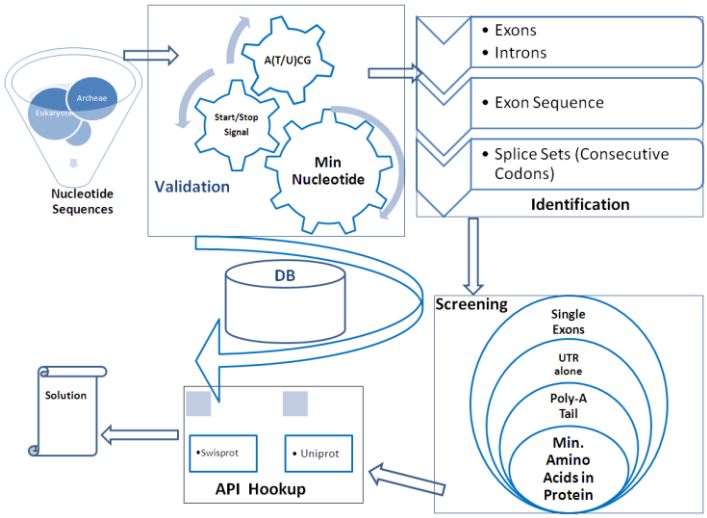

Figure 2. Proposed Methodology for Computational Solution of c-Value Enigma

In the Figure 2, we have shown our proposed methodology. This methodology consists of following steps.

Step-I. DNA Genes Acquisition: This step deals with the preprocessing of the raw data obtained from various biological data source web sites. An important feature of all of these steps is that from step one to last step, all data is also being stored in a database so that an archive can be generated. Such archival repository may be helpful for in depth analysis or data mining of genes, proteins and RNA data.

Step-II. Validation of DNA Genes: Validity of the DNA genes is an important step in any methodology. We focused on three aspects of validity which include validity of nucleotide bases as it must contain adenine, cytosine, guanine and thymine or uracil. If the gene is containing any other letter then such gene would be termed as invalid gene. The other two validations are related to checking its start and stop signal and conformation of minimum number of nucleotide base.

Step-III. Identification: Identification of various parts of a gene is the most critical part of our methodology. This step is most versatile and dynamic due to its intrinsic nature of biological material's diversity. Spliceosomal introns are usually found in eukaryotic protein-coding genes. For splicing, within the intron, a 3' splice site, 5' splice site, and branch site are entailed. The 5' splice site also known as splice donor site includes an almost invariant sequence of

Guanine and Uracil (GU) at the 5' end of the intron, within a larger, but less densely preserved consensus region. The 3' splice site also known as splice acceptor site terminates the intron with an invariant Adenine and Guanine (AG) sequence. Upstream (5'-ward) from the AG, a region high in pyrimidines (C and U) or polypyrimidine tract is found. Upstream from the polypyrimidine tract can be termed as the stem point, which includes an adenine nucleotide. Clancy [19] and Black, *et al.*, [15] described that *cryptic splice site* may be observed due to point mutations in the underlying DNA or errors during transcription resulting in part of the transcript normally goes unspliced. This eventually results in a mature mRNA with a missing part of an exon. This leads to the point that mutation can cause a deletion in the final protein though it normally affects only a single amino acid. Considering into these fact in this step exons, introns and poly-A tail are identified and marked. This helps us in marking gene splicing site location so that we can cut down the gene into its various respective parts. The computational complexity of this step is critically important. The conventional technique might lead to exponential growth of complexity. However we have adopted the technique in which a big set is divided into smaller sub sets. Such heuristic will eventually reduce the complexity of this step.

Step-IV. Screening: The outcome of the previous section will deliver a sequence of nucleotide base. However this is not sufficient for inferring proteins out of it. The possible reason behind this is buried in the fact that at the end of the identification, we may come up with single exon, single intron or single poly-A tail etc. Obviously such small parts are not enough to express genes. So any such outcome is simply discarded in this step.

Step-V. Cross Multiplication: The crux of our technique lies in this step. As depicted by Figure 3 this step is related to mathematical sub set theory problem. Before this stage, we have successfully identified introns, exons, poly-A tails etc. These materials can now be treated as the input for the cross multiplication of data items. All of these genetic materials are available to us in form of sequence of nucleotide bases. Our goal is to generate all possible candidates out of combining these subsets. All of the possible outcome would be termed as possible protein which will be subjected to step-5. Figure 3 is showing first scan of the heuristic on the data set, in which six datasets 5′ and 3′ terminus sides, two introns, two exons are included. Cross section of all of these datasets would yield a candidate protein say, P-1. Other possible candidate proteins P-2 to P-7 are generated by skipping any one member of the data set. This is the detail of first scan of cross multiplication. In the other scan, two members out of this set would be left over while joining rest of the data set values. At the completion of this step, we will have 2^6 candidate proteins.

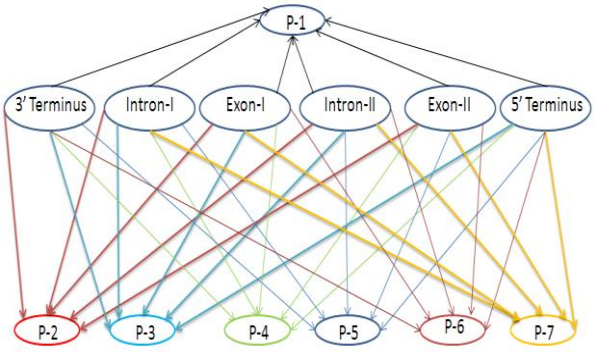

Figure 3. Seven Possible Proteins out of six spliced parts, stage-I

Step-VI. API Hookup: There are various online tools available which can identify protein if some appropriate input is provided to them. At the end of the step-v, our technique will have a large number of candidate proteins. In this step, these candidate proteins will be examined by online tools like uniprot or swisprot. To speed up the process of automation, this technique will be hooked up to these online tools. Output of this step will list down all of the possible proteins which can be generated from a given DNA material.

3. Results and Discussion

We deliver in this section theoretical result achieved in this study. We have developed a computational heuristic providing an insight into how to solve one part of the famous C-value enigma. The algorithmic heuristic is based upon the various alternative splicing strategies. These strategies need to be elaborated in context of the result in this section for the benefit of the reader. Strategies includes Wild Type splicing, Exon Skipping, Intron retention mode, Alternative 3′ site (3′ splice site switching) and Alternative 5′ site (5′ splice site switching), Alternative selection of promoters, Alternative Poly-A sites. Alternative selection of cleavage or polyadenylation sites, exon cassette mode. Alternative selection of promoters include myosin primary transcript. Alternative selections of cleavage include tropomyosin. Transposase primary transcript and troponin primary transcript are examples for intron retention mode and exon cassette mode respectively.

We shall explain our technique by an example. The first strategy is related to exon skipping. In male *drosophila dsx* gene there have been observed six exons. Among these exon located at 4^{th} position is skipped while the others are merged to form the mRNA. This mRNA is responsible for encoding a transcriptional regulatory protein. Another example is related to female drosophila dsx. In this case the last two exons situated at fifth and sixth position are skipped while the other exons are combied to transcript mRNA needed for female development. Moreover *polyadenylation* signal in exon four also emboss a cleavage at that stipulated location. The intron upstream from exon 4 carries a polypyrimidine tract which doesn't match the consensus sequence, so that U2AF proteins bind poorly to it without assistance from splicing activators. This 3' splice acceptor site is therefore not used in males. However on the female side, it produce the splicing activator Transformer (Tra). The SR protein Tra2 is produced in both sexes and binds to an ESE in exon 4. If Tra is present, it binds to Tra2 and also to another SR protein, develops a complex that assists U2AF proteins in binding to the weak *polypyrimidine* tract. U2 is recruited to the associated branch point, and this leads to inclusion of exon 4 in the mRNA.

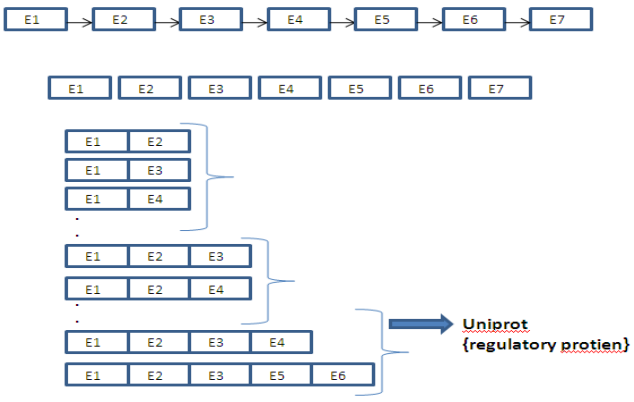

Figure 4. Exon Skipping in Alternative Splicing

In the Figure 4, one strategy, we have considered as a possible solution for the C-value enigma has been proposed. In *drosophila dsx* there are seven exons out of which many possible groups of exons can be formed. Among all of these only two mentioned in the last are valid proteins while the others are discard able.

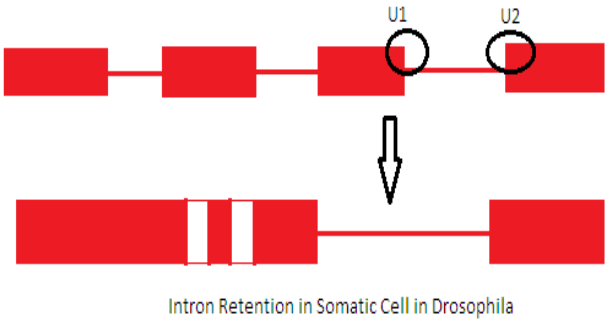

Figure 5. Intron Retention in Alternative Splicing

Another strategy discussed in the proposed algorithm is related to Intron retention. Figure 5 describe the regulation of drosophila P element splicing. In drosophila P element transcription is a good example to portray this application. The element P can be defined as the transposon which is restricted to the movement of in the genome of drosophila. The Intron retention has a distinction in both of the cases such that in somatic cells, third Intron is retained in translation which forbids transposase enzyme restricting somatic transposition. However in sex cells, all of the P element introns are skipped during expression of transposase. The exon upstream counted from the 3rd Intron bears a regulatory sequence inhibiting its splicing in somatic cells. P element somatic inhibitor binds in the exon upstream from Intron 3 with U1 snRNP which binds to two pseudo 5 terminus sites. Moreover the regulatory region also bears binding sites for proteins including P element somatic inhibitor.

References

[1] R. J. Britten and E. H. Davidson, "Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty", Quarterly Review of Biology, **(1971)**, pp. 111-138**.**
[2] T. R. Gregory, "The C-value enigma in plants and animals: a review of parallels and an appeal for partnership", Annals of Botany, vol. 95, no. 1, **(2005)**, pp. 133-146**.**
[3] T. R. Gregory, "Genome size and developmental complexity", Genetica, vol. 115, no. 1, **(2002)**, pp. 131-146.
[4] T. R. Gregory, "Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma", Biological Reviews, vol. 76, no. 1, **(2001)**, pp. 65-101**.**
[5] A. Wolffe, "Chromatin: structure and function", Academic Press, **(1998).**
[6] Z. Zhan, J. Ren, Y. Zhang, R. Zhao, S. Yang and W. Wang, "Evolution of alternative splicing in newly evolved genes of Drosophila", Gene, vol. 470, no. 1, **(2011)**, pp. 1-6**.**
[7] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin and D. Grafham, "Initial sequencing and analysis of the human genome", Nature, vol. 409, no. 6822, **(2001)**, pp. 860-921**.**
[8] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes", Nature, vol. 456, no. 7221, **(2008)**, pp. 470-476**.**
[9] M. Long, E. Betrán, K. Thornton and W. Wang, "The origin of new genes: glimpses from the young and old", Nature Reviews Genetics, vol. 4, no. 11, **(2003)**, pp. 865-875**.**
[10] Q. Zhou and W. Wang, "On the origin and evolution of new genes—a genomic and experimental perspective", Journal of Genetics and Genomics, vol. 35, no. 11, **(2008)**, pp. 639-648**.**
[11] A. Wakamatsu, J. I. Imai, S. Watanabe and T. Isogai, "Alternative splicing of genes during neuronal differentiation of NT2 pluripotential human embryonal carcinoma cells", FEBS letters, vol. 584, no. 18, **(2010)**, pp. 4041-4047.

[12] S. Matuda, T. Arimura, A. Kimura, H. Takekura, S. Ohta and K. Nakano, "A novel protein found in the I bands of myofibrils is produced by alternative splicing of the DLST gene", Biochimica et Biophysica Acta (BBA)-General Subjects, vol. 1800, no. 1, **(2010)**, pp. 31-39**.**
[13] L. Jin, K. Kryukov, J. C. Clemente, T. Komiyama, Y. Suzuki, T. Imanishi and T. Gojobori, "The evolutionary relationship between gene duplication and alternative splicing", Gene, vol. 427, no. 1-2, **(2008)**, pp. 19-31**.**
[14] L. F. Lareau, R. E. Green, R. S. Bhatnagar and S. E. Brenner, "The evolving roles of alternative splicing", Current opinion in structural biology, vol. 14, no. 3, **(2004)**, pp. 273-282**.**
[15] D. L. Black, "Mechanisms of alternative pre-messenger RNA splicing", Annual review of biochemistry, vol. 72, no. 1, **(2003)**, pp. 291-336**.**
[16] D. Holste and U. Ohler, "Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events", PLoS Computational Biology, vol. 4, no. 1, **(2008)**, pp. e21**.**
[17] J. Marcus, "A partial solution to the c-value paradox", Comparative Genomics, **(2005)**, pp. 97-105.
[18] E. Koren, G. Lev-Maor and G. Ast, "The emergence of alternative 3′ and 5′ splice site exons from constitutive exons", PLoS computational biology, vol. 3, no. 5, **(2007)**, pp. e95.
[19] S. Clancy, "RNA splicing: introns, exons and spliceosome", Nature Education, vol. 1, no. 1, **(2008).**

Authors

Muhammad Naeem: Research scholar at department of computer science, M. A. Jinnah University Islamabad Pakistan. His research area includes machine learning, semantic computing, text retrieval, graph mining, classification and data mining.

Saira Gilani: She received her M.IT degree in Information Technology from Balochistan University, Quetta, Pakistan in 2004. She is an MS student at M.A. Jinnah University, Islamabad, Pakistan and is a member of Center of Research in Networks and Telecommunication (CoReNeT). Her current research activities include investigation of MAC layer schemes, network security for Vehicular Ad Hoc Networks and data mining and semantic computing.

Sohail Asghar: Dr. Sohail Asghar is Director/ Associate Professor at Arid-Agriculture University Rawalpindi Pakistan. He earns PhD in Computer Science from Monash University, Melbourne, Australia in 2006. Earlier he did his Bachelor of Computer Science (Hons) from University of Wales, United Kingdom in 1994. His research interest includes data mining, decision support system and machine learning.