Journal of Theoretical and Applied Information Technology
10th January 2013. Vol. 47 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPUTATIONAL ANALYSIS OF THE SIMPLE SEQUENCE REPEATS IN MUCIN-6 OF FOVEOLAR CELLS

SIM-HUI TEE
Multimedia University, Cyberjaya, Malaysia
E-mail: shtee@mmu.edu.my

ABSTRACT

In this study, a computational analysis has been carried out to investigate the patterns of the simple sequence repeats (SSRs) in mucin-6 mRNA gene of human foveolar cells. The computational analysis of SSRs is important because it can provide insight into the diseases generated by SSRs. Our results show that mucin-6 mRNA has a high relative frequency of dinucleotide and trinucleotide SSR motifs. The trinucleotide SSRs have special significance because they constitute the amino acid in the coding region of the gene. The analysis of trinucleotide SSRs shows that the motifs from the category T9 is highly repetitive. This research may provide insights into the biomedical research in the stomach pathogenesis that is implicated by SSRs in mucin-6 gene.

Keywords: *Bioinformatics, Simple Sequence Repeats, Mucin-6, Foveolar Cells, Genetic Analysis*

1. INTRODUCTION

Computational approaches and tools such as algorithms [1-3], databases [4-7], models [8-10] and web servers [11-14] have been widely used in computational biology and bioinformatics for genome analysis, protein analysis and the analysis of various biological data. The powerful computing capabilities of these computational tools allow computer scientists and bioinformaticians to analyze large volume of data in an efficient way. Rapid growth of bioinformatics not only depends on the development of theoretical computer science and mathematics; conversely, new insights from bioinformatics also bring about new problems to computer scientists, such as the optimization and complexity issue. It has been recognized that computing tools have a larger role to play in the field of biology in future [15].

In this study, a computational analysis has been carried out to investigate the simple sequence repeats in mucin-6 in human foveolar cells. Foveolar cells are crucial in stomach because of their essential roles in producing mucus [16] to cover the inner stomach. Mucus is formed from the dissolved mucin, which is a complex glycoprotein. Without this active mucus-producing role played by foveolar cells, stomach will digest itself due to its acidic and corrosive nature. Besides, it was reported that foveolar cells suppress the growth of tumor at the molecular level, via the transcription of peptides of trefoil factor family [17]. It is of essential importance to investigate the expression mechanisms of mucin-6 in foveolar cells in order to understand the malfunctions and diseases that are caused by this cell type. Computational analysis is one of the approaches to understanding the expression mechanism of mucin-6 in foveolar cells, which may provide insights into the therapeutic strategies and drug designs in curing foveolar cell-related diseases.

Simple sequence repeats (SSRs) are DNA sequences consisting of short motifs that exhibit a repeating pattern in tandem. It has been found that the extensive phenomenon of SSRs in human would lead to about 20 severe neurological disorders [18]. The early diagnosis of these SSRs-related diseases can be made by the understanding of the sequence composition and the location of the repeats in the gene [19]. Hence, computational approach serves as an efficient way to serve the purpose. Despite the fact that the occurrence of extensive SSRs may be an indicator of aberration in cells, sequence repeats are ubiquitous in all organisms. SSRs can occur in any location on a gene, such as in exon, intron and intergenic regions [20]. However, certain types of SSRs are prevalent among disease-related genes, such as the implications of trinucleotide repeats in cancer [21]. SSRs are unstable in the genome and their mutation rate is 10 to 100,000 times higher than the mean rate of mutation [22]. As such, it is important to understand the tandem repeat patterns of SSRs. The study of SSRs in mucin-6 of foveolar cells may unveil the pathologic mechanism of the mucus-producing foveolar cells.

Journal of Theoretical and Applied Information Technology
10th January 2013. Vol. 47 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. METHODS

The nucleotide sequences of mucin-6 messenger RNA (mRNA) were retrieved from GenBank, as following the standard practice in the field. A microsatellite extractor software [23] has been used to identify the SSRs in mucin-6 mRNA of foveolar cells. A sequence will be defined as a SSR if it is expressed as a series of tandem repeats with the motifs of 1 to 6 base pairs [23]. A simple string searching algorithm with sliding window approach was employed [23]. To identify spatial motifs in mucin-6 (for interacting sequences), the following formula was used [29]:

$$P(X,Y) = \frac{f_{obs}(X,Y)}{IE[f(X,Y)]} \qquad (1)$$

where $P(X,Y)$ is the propensity of the residue interacting pair X-Y; $f_{obs}(X,Y)$ is the actual count of X-Y contacts in the interacting sequence; and $IE[f(X,Y)]$ is the expected count.

The perfect repeat of the mucin-6 sequence rather than the imperfect repeat was investigated. A minimum of six mononucleotides for the SSRs were set. In addition, at least two minimum repeats for the di-, tri-, tetra-, penta-, and hexanucleotides of mucin-6 of foveolar cells were analyzed. Relative frequency was used to analyze the total repeat per kilo base in the nucleotide sequence of mucin-6. A triplet classification system [24] was used to categorize the trinucleotide SSRs.

3. RESULTS AND DISCUSSION

The nucleotide sequences of mucin-6 mRNA are 8003 base pair (bp) in length, which is GC-rich (61.6%). Notably, the composition of cytosine is among the most abundant nucleotide (3092 nucleotides; 38.6%). In this study, a total of 3130 SSRs were extracted and analyzed. The occurrence of SSRs in mucin-6 is high in view of the small gene size. The distribution of SSRs is shown in Table 1.

Table 1. Distribution Of Ssrs In Mrna Of Mucin-6

Repeat motif	# occurrence	Relative frequency
Mononucleotide:		
A	27	3.37
T	9	1.12
C	12	1.50
G	5	0.62
Total	**53**	**6.62**
Dinucleotide:		
AT/TA	191	23.87
AC/CA	616	76.97
AG/GA	327	40.86
CG/GC	105	13.12
GT/TG	318	39.74
CT/TC	307	38.36
Total	**1864**	**232.91**
Trinucleotide	**922**	**115.21**
Tetranucleotide	**107**	**13.37**
Pentanucleotide	**20**	**2.50**
Hexanucleotide	**118**	**14.74**

From Table 1, it is apparent that the occurrence of SSR mononucleotide is quite low, with the relative frequency of 6.62, implying that there are only approximately 7 occurrences per 1000 nucleotides. The low relative frequency of SSR mononucleotide is comparative to that of tetranucleotide (13.37), pentanucleotide (2.5), and hexanucleotide (14.74). Among these SSRs, the motif ACCCCA contributes largely (12.5; with 100 repeats) to the relative frequency of hexanucleotide. All of the motifs ACCCCA exist in the form of $(ACCCCA)_2$. This could be the consequences of viral genome integration or other unknown reasons that need further experimentation.

The relative frequency of dinucleotides and trinucleotides are high as compared to other repeat motifs. As for dinucleotide, a total of 1864 occurrences of SSRs were obtained in the analysis, which is tantamount to a relative frequency of 232.91. Among six types of dinucleotide motifs, AC/CA is the most prevalent in the gene of mucin-6, with the occurrence of 616 (relative frequency=76.97). Among the total of 290 AC repeat motif, most of the SSRs exist in the form of $(AC)_2$, with a few in the form of $(AC)_3$. Interestingly, of the total of 326 CA repeat motif, the number of $(CA)_3$ motif is as extensive as $(CA)_2$. Comparing the dinucleotide SSRs with

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

mononucleotide SSRs, it was found that the latter has a higher number of repeat in tandem. For example, mononucleotide C motif, which has an occurrence of 12 times in the gene of mucin-6, was found to exist in the form of $(C)_6$. Similar finding was encountered in mononucleotide A motif, which was found to exist in the form of $(A)_6$, $(A)_7$, $(A)_{10}$, $(A)_{19}$, $(A)_{34}$, and $(A)_{43}$. Mononucleotide G motif was found to have 4 occurrences of $(G)_6$ and 1 occurrence of $(G)_7$. Lastly, mononucleotide T motif exhibits 8 occurrences of $(T)_6$ and 1 occurrence of $(T)_9$. The motif tandem pattern of dinucleotide is summarized in Table 2. The percentage is relative to each motif.

Table 2. The Motif Tandem Pattern Of Dinucleotide

Dinucleotide motif	# Occurrence	Percentage
$(CA)_2$	202	62%
$(CA)_3$	122	37.4%
$(CA)_4$	2	0.6%
$(TG)_2$	196	95.1%
$(TG)_3$	8	3.9%
$(TG)_4$	2	1%
$(CT)_2$	173	95.1%
$(CT)_3$	9	4.9%
$(AC)_2$	280	96.6%
$(AC)_3$	10	3.4%
$(GA)_2$	164	95.3%
$(GA)_3$	8	4.7%
$(AG)_2$	147	94.8%
$(AG)_3$	8	5.2%
$(CG)_2$	25	92.6%
$(CG)_3$	1	3.7%
$(CG)_4$	1	3.7%
$(GT)_2$	97	86.6%
$(GT)_3$	14	12.5%
$(GT)_4$	1	0.9%
$(GC)_2$	74	94.9%
$(GC)_3$	4	5.1%
$(TC)_2$	109	87.2%
$(TC)_3$	14	11.2%
$(TC)_5$	2	1.6%
$(TA)_2$	79	95.2%
$(TA)_3$	4	4.8%
$(AT)_2$	100	92.6%
$(AT)_3$	4	3.7%
$(AT)_4$	4	3.7%

From Table 2, it is clear that most of the abundant nucleotide motif $(XY)_n$ exhibits n=2. Among 12 motif types, there are 9 dinucleotide motif types demonstrate at least 90% of n=2 tandem repeat. There are 2 occurrences of tandem repeat n=5 (TC), whereas none of the motif exhibits tandem repeat n=6 or higher. This is expected for the mammalian gene because the tandem repeat is not as highly prevalent as in the viral genomes. For example, the tandem repeats of the motif TAATATTAC is prevalent in Tomato leaf curl Patna virus, especially within the intergenic region [25]. The *Bam*HI Y fragment of herpes simplex virus type 2 was reported to have a lengthy repeat motif of AGGGGCGGCTGGGGC [26]. However, the number of SSRs in viruses was highly variable (15-180), as shown in the telomeric repeat sequences of the human herpesvirus 6 genome [27].

There are 30 tetranucleotide motifs with 107 SSRs. There is no single motif which is the most prevalent among the tetranucleotide. The highest number of SSR is AAAG (8 occurrences) and the lowest number of SSRs (2 occurrences) include CTCC, CAGG, ACCT, CCAG, AGGG, GTCA, GCTG, GCTC, GTCT. There are 7 pentanucleotide motifs which are GC-rich in SSRs, and 6 hexanucleotide motifs which are rich in AC.

Trinucleotide SSRs have special significance in cells because they are closely related to the amino acid repeats, which may impact the function of the proteins [28]. Thus, special analysis was carried out for this type of SSRs. We categorized trinucleotide SSRs into ten types (T1-T10) according to [24], as shown in Table 3.

Table 3. Distribution Of Trinucleotide Ssrs In Mucin-6 Mrna

Type	Repeat motifs (with frequency)	Total
T1	AAT(9) ATA(12) TAA(2) ATT(12) TTA(12) TAT(4)	51
T2	AAG(15) AGA(21)	91

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	GAA(33) CTT(10) TTC(10) TCT(2)	
T3	AAC(12) ACA(7) CAA(17) GTT(5) TTG(10) TGT(4)	55
T4	ATG(8) TGA(22) GAT(4) CAT(18) ATC(2) TCA(20)	74
T5	AGT(0) GTA(0) TAG(0) ACT(8) CTA(12) TAC(0)	20
T6	AGG(11) GGA(20) GAG(15) CCT(23) CTC(14) TCC(25)	108
T7	AGC(25) GCA(13) CAG(33) GCT(29) CTG(21) TGC(14)	135
T8	ACG(0) CGA(2) GAC(8) CGT(3) GTC(4) TCG(0)	17
T9	ACC(79) CCA(54) CAC(128) GGT(9) GTG(5) TGG(36)	311
T10	GGC(11) GCG(4) CGG(8) GCC(21) CCG(12) CGC(4)	60

From Table 3, the trinucleotides of type T5 and T8 are scarce, which are 20 and 17, respectively. It is primarily due to the non-existence of several SSR motifs in these types, such as AGT, GTA, TAG, TAC, ACG, and TCG. Our results for T5 are consistent with the findings of Ouyang et al. [30], but they have demonstrated that the genome of Herpes simplex virus type 1 (HSV-1) has a moderate level of T8 type of trinucleotide. However, they have demonstrated that HSV-1 SSR motifs are abundant in T10 type of trinucleotide, whereas our results for mucin-6 of foveolar cells do not exhibit very high level of T10 motifs. In this study, it was found that T9 motifs are the most abundant in mucin-6, whereas in the research of Ouyang et al. [30] it was found that T9 motifs were moderately abundant. In a study carried out by Behura and Severson [31] on the SSRs among 25 insect species, they have shown that none of the T9 motifs are abundant in insect. Strikingly, these insect species are either lacking of most of the T9 motifs or having a very low level of T9 motifs. Nonetheless, their findings on the T5 motifs are consistent with our findings, albeit the difference between our research and theirs is that they focused on the simple sequence repeat patterns for the coding region of the genes. The deviation between the results of our findings and that of Ouyang et al. [30] and Behura and Severson [31] is an indicator of the genetic variation between virus genes, insect genes and human genes.

4. CONCLUSION

The results of this study provide insights into the distribution of SSRs in mucin-6 mRNA of human foveolar cells. It was found that the relative frequency of dinucleotide and trinucleotide SSR motifs is higher than other type of motif. The abundance of trinucleotide SSRs has significant impacts on the encoded proteins for foveolar cells. As foveolar cell is implicated in the mucus production in the stomach, understanding in the trinucleotide SSR distribution patterns (and other motif types) may guide the biomedical research in the prevention and diagnosis of the foveolar cell-implicated stomach pathogenesis.

REFERENCES

[1] Y-F. Chen, R-C. Chen, Y-K. Chan, R-H. Pan, Y-C. Hseu, and E. Lin, “Design of multiplex PCR primers using heuristic algorithm for sequential deletion applications”, *Computational Biology and Chemistry*, Vol. 33, 2009, pp. 181-188.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[2] K. Takahashi, K. Kaizu, B. Hu, and M. Tomita, "A multi-algorithm, multi-timescale method for cell simulation", *Bioinformatics*, Vol. 20, 2004, pp. 538-546.

[3] K.S. Kim, J.H. Seo, S.H. Ryu, M.H. Kim, and C.G. Song, "Estimation algorithm of the bowel motility based on regression analysis of the jitter and shimmer of bowel sounds", *Computer Methods and Programs in Biomedicine*, Vol. 104, 2011, pp. 426-434.

[4] I-S. Jeong, K-W. Park, S-H. Kang, and H-S. Lim, "An efficient similarity search based on indexing in large DNA databases", *Computational Biology and Chemistry*, Vol. 34, 2010, pp. 131-136.

[5] O. Arnaiz, A. Malinowska, C. Klotz, L. Sperling, M. Dadlez, F. Koll, and J. Cohen, "Cildb: a knowledgebase for centrosomes and cilia", *Database*, Vol. 2009, 2009, doi:10.1093/database/bap022.

[6] T. Hilbel, D. Lossnitzer, R. Tesarczyk, H.A. Katus, and E. Giannitsis, "Long-time experience with a dedicated database for a chest pain observation unit", *Computers in Cardiology*, Vol. 36, 2009, pp. 249-252.

[7] E. Chautard, M. Fatoux-Ardore, L. Ballut, N. Thierry-Mieg, and S. Ricard-Blum, "MatrixDB, the extracellular matrix interaction database", *Nucleic Acids Research*, Vol. 39, 2011, pp. D235-D240.

[8] F. Ciocchetta and J. Hillston, "Bio-PEPA: A framework for the modelling and analysis of biological systems", *Theoretical Computer Science*, Vol. 410, 2009, pp. 3065-3084.

[9] E. Bartocci, F. Corradini, E. Merelli, and L. Tesei, "Detecting synchronisation of biological oscillators by model checking", *Theoretical Computer Science*, Vol. 411, 2010, pp. 1999-2018.

[10] S.G. Lee, J.U. Hur, and Y.S. Kim, "A graph-theoretic modeling on GO space for biological interpretation of gene clusters", *Bioinformatics*, Vol. 20, No. 3, 2004, pp. 381-388.

[11] J. Oberto, "BAGET: a web server for the effortless retrieval of prokaryotic gene context and sequence", *Bioinformatics*, Vol. 24, No. 3, 2008, pp. 424-425.

[12] A.B. Diallo, V. Makarenkov, and M. Blanchette, "Ancestors 1.0: a web server for ancestral sequence reconstruction", *Bioinformatics*, Vol. 26, No. 1, 2010, pp. 130-131.

[13] C. Lushbough, M.K. Bergman, C.J. Lawrence, D. Jennewein, and V.Brendel, "BioExtract Server—An integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 7, No. 1, 2010, pp. 12-24.

[14] F.P. Casey, N.E. Davey, I. Baran, R.S. Varekova, and D.C. Shields, "Web server to identify similarity of amino acid motifs to compounds (SAAMCO)", *Journal of Chemical information and Modeling*, Vol. 48, 2008, pp. 1524-1529.

[15] T.G. Lilbum, S.H. Harrison, J.R. Cole, and G.M. Garrity, "Computational aspects of systematic biology", *Briefings in Bioinformatics*, Vol. 7, No. 2, 2006, pp. 186-195.

[16] A. Lüdeking, P. Fegert, N. Blin, P. Gött, "Osmotic changes and ethanol modify TFF gene expression in gastrointestinal cell lines", *FEBS Letters*, Vol. 439, 1998, pp. 180-184.

[17] P. Azarschab, E-d. Al-Azzeh, W. Kornberger, P. Gött, "Aspirin promotes TFF2 gene activation in human gastric cancer cell lines", *FEBS Letters*, Vol. 488, 2001, pp. 206-210.

[18] C.T. McMurray, "Mechanisms of trinucleotide repeat instability during human development", *Nature Reviews Genetics*, Vol. 11, 2010, pp. 786-799.

[19] A.R. La Spada and J.P. Taylor, "Repeat expansion disease: progress and puzzles in disease pathogenesis", *Nature Reviews Genetics*, Vol. 11, 2010, pp. 247-258.

[20] A.J. Hannan, "Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'", *Trends in Genetics*, Vol. 26, No. 2, 2009, pp. 59-65.

[21] Y. Haberman, N. Amariglio, G. Rechavi, and E. Eisenberg, "Trinucleotide repeats are prevalent among cancer-related genes", *Trends in Genetics*, Vol. 24, No. 1, 2007, pp. 14-18.

[22] R. Gemayel, M.D. Vinces, M. Legendre, and K.J. Verstrepen, "Variable tandem repeats accelerate evolution of coding and regulatory sequences", *Annual Review of Genetics*, Vol. 44, 2010, pp. 445-477.

[23] S.B. Mudunuri and H.A. Nagarajaram, "IMEx: Imperfect microsatellite extractor", *Bioinformatics*, Vol. 23, No. 10, 2007, pp. 1181-1187.

[24] J. Jurka and C. Pethiyagoda, "Simple repetitive DNA sequences from primates: compilation and analysis", *Journal of Molecular Evolution*, Vol. 40, 1995, pp. 120-126.

[25] P. Kumari, A.K. Singh, B. Chattopadhyay, and S. Chakraborty, "Molecular characterization of a new species of *Begomovirus* and betasatellite

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

causing leaf curl disease of tomato in India", *Virus Research*, Vol. 152, 2010, pp. 19-29.

[26] T. Yamaguchi, Y. Yamashita, K. Kasamo, M. Inoue, H. Sakaoka, and K. Fujinaga, "Genomic heterogeneity maps to tandem repeat sequences in the herpes simplex virus type 2 U_L region", *Virus Research*, Vol. 55, 1998, pp. 221-231.

[27] A. Achour, I. Malet, C. Deback, P. Bonnafous, D. Boutolleau, A. Gautheret-Dejean, and H. Agut, "Length variability of telomeric repeat sequences of human herpesvirus 6 DNA", *Journal of Virological Methods*, Vol. 159, 2009, pp. 127-130.

[28] M.J. Lawson and L.Q. Zhang, "Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes", *Genome Biology*, Vol. 7, 2006, pp. R14.

[29] R. Jackups and J. Liang, "Combinatorial analysis for sequence and spatial motif discovery in short sequence fragments", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 7, No. 3, 2010, pp. 524-536.

[30] Q. Ouyang, X. Zhao, H. Feng, Y. Tian, D. Li, M. Li, and Z. Tan, "High GC content of simple sequence repeats in *Herpes simplex virus type* 1 genome", *Gene*, Vol. 499, 2012, pp. 37-40.

[31] S.K. Behura and D.W. Severson, "Genome-wide comparative analysis of simple sequence coding repeats among 25 insect species", *Gene*, Vol. 504, 2012, pp. 226-232.