**In Silico Analysis of Tandem Repeats in GIF of Gastric Parietal Cells**

Sim-Hui Tee
Multimedia University
Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
shtee@mmu.edu.my

**Abstract**
Tandem repeats are ubiquitous in the genome of organisms and their mutated forms play a vital role in pathogenesis. In this study, tandem repeats in Gastric Intrinsic Factor (GIF) of gastric parietal cells have been investigated using an in silico approach. Six types of the nucleotide tandem repeat motifs have been investigated, including mono-, di-, tri-, tetra-, penta- and hexanucleotide. The distribution of the repeat motifs in the gene was analyzed. The results of this study provide an insight into the biomolecular mechanisms and pathogenesis implicated by the GIF of gastric parietal cells. Based on the findings of the tandem repeats in GIF of gastric parietal cells, therapeutic strategies and disease markers may be developed accordingly by the biomedical scientists.

Keywords- Bioinformatics; Stomach epithelium cells, Parietal cells, tandem repeats, in silico, gastric intrinsic factor

I. INTRODUCTION

Parietal cells are the stomach epithelium cells that secrete hydrochloric acid and intrinsic factors in response to the receptor binding of histamine, gastrin, and acetylcholine [1-2]. Acid secretion by parietal cells was observed to be mediated by receptor binding of histamine, gastrin, and acetylcholine [1-2]. Hydrochloric acid and intrinsic factors in response to the findings of the tandem repeats in GIF of gastric parietal cells, implicated by the GIF of gastric parietal cells. Based on the insight into the biomolecular mechanisms and pathogenesis in the gene was analyzed. The results of this study provide an insight into the biomolecular mechanisms and pathogenesis implicated by the GIF of gastric parietal cells. Based on the finding of the tandem repeats has widespread application in various fields such as medicine, forensic science, and population genetics [13]. Notably, the identification of tandem repeats has special function in genetic studies, such as gene mapping [15].

In this study, tandem repeats in Gastric Intrinsic Factor (GIF) of gastric parietal cells have been investigated using an in silico approach. GIF is required for the absorption of vitamin B12, without which could result in megaloblastic anemia, atrophic gastritis, and pernicious anemia. In silico approach appears as an efficient approach that has been widely used in the investigation of biomedical phenomena. A wide range of in silico techniques and tools has been established, including algorithms [16-23], databases [24-29], web servers [30-34], and modellings [35-42], which make possible the efficient processing and analysis of large volume of biomedical data. The results of this study provide an insight into the biomolecular mechanisms and pathogenesis implicated by the gastric parietal cells. Based on the finding of the tandem repeats in GIF of gastric parietal cells, therapeutic strategies and disease markers may be developed accordingly by the biomedical practitioners.

II. METHODS

The nucleotide sequences of GIF of gastric parietal cells have been retrieved from NCBI GenBank. An alignment was done for all the perfect repeat units. A combinatorial-based mreps algorithm [43] was used to find the repeated sequence: (1) Given an error threshold k, the maximal k-mismatch runs of tandem repeats were sought. (2) The left and right edges of the tandem repeats were trimmed. (3) The best period of each repeat was computed. (4) The statistically expected repeats were identified. (5) The tandem repeats with the same period p overlapping by at least 2p were merged.

To identify spatial motifs in GIF (for interacting sequences), the following formula was used [44]:

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

Despite the observed pathogenesis of tandem repeat expansion, it was reported that tandem repeats are ubiquitous in all organisms, and they could occur on any locus on a gene [9]. Tandem repeats have the tendency to accumulate in organisms with larger genomes, such as that of mammals [14]. In addition, it is also known that variable tandem repeats in the genome of an organism can accelerate the evolution of coding and regulatory gene sequences [12]. The information obtained from the investigation of tandem repeats has widespread application in various fields such as medicine, forensic science, and population genetics [13]. Notably, the identification of tandem repeats has special function in genetic studies, such as gene mapping [15].
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 sequences; and \( IE[I(X,Y)] \) is the expected count.

Tandem repeat motif occurrence was determined using hierarchical gene-set genetics based algorithm [45]. Given \( S^j \) which is the subsequence of length \( W \) at position \( j \) in a sequence \( i \). Let \( a \) be the symbol that occurs at a position \( k \) of either motif or non-motif; let the position \( k \) be \( 1 \leq k \leq W \). Let \( \theta^M \) and \( \theta^B \) model the motif and non-motif positions in a sequence, respectively. The conditional probabilities that \( S^j \) is found using the motif model and background (non-motif) model are computed as such [45], respectively:

\[
P_M(S^j) = \prod_{k=1}^{W} \prod_{a=1}^{L} (\theta^M_{ak})^{I(S_{j+k-1}a)} \tag{2}
\]

\[
P_B(S^j) = \prod_{k=1}^{W} \prod_{a=1}^{L} (\theta^B_{ak})^{I(S_{j+k-1}a)} \tag{3}
\]

The motif model \( \Theta \) is given in the following equation [45]:

\[
\Theta = \{ \theta^B, \theta^M \} = \begin{bmatrix}
\theta^B_{a1,0} & \theta^M_{a1,1} & \cdots & \theta^M_{a1,w} \\
\theta^B_{a2,0} & \theta^M_{a2,1} & \cdots & \theta^M_{a2,w} \\
\vdots & \vdots & \ddots & \vdots \\
\theta^B_{aw,0} & \theta^M_{aw,1} & \cdots & \theta^M_{aw,w}
\end{bmatrix} \tag{4}
\]

Let \( \lambda \) be the prior probability of motif occurrence in the gene sequences. The motif occurrence probability \( Z \) at position \( j \) in sequence \( i \) is derived from [45]:

\[
Z_j = \frac{\lambda P_M(S^j)}{(\lambda P_M(S^j)) + (1-\lambda)P_B(S^j)} \tag{5}
\]

We take \( S^j \) to be a motif encounter when the following is fulfilled [45]:

\[
\log(P_M(S^j) / P_B(S^j)) \geq \log(1-\lambda)/\lambda \tag{6}
\]

In this research, six types of tandem repeat motifs were analyzed (mono-, di-, tri-, tetra-, penta-, and hexanucleotide). Relative frequency was used as a measure to analyze the total repeat per kilobase in the nucleotide sequence of GIF. A triplet classification system [46] was used to categorize the trinucleotide tandem repeats.

III. RESULTS AND DISCUSSION

GIF of gastric parietal cells is 94270 base pair (bp) in length, which is located on chromosome 11. The composition of GIF is: A = 27.02%, T = 26.66%, G = 23.47%, and C = 22.86%. A total of 4919 perfect tandem repeats were found and retrieved from GIF gene sequence. The distribution of tandem repeats of GIF is shown in Table 1.

| Repeat motif      | No. of occurrence | Relative frequency |
|-------------------|-------------------|--------------------|
| Mononucleotide:   |                   |                    |
| A                 | 100               | 1.06               |
| T                 | 84                | 0.89               |
| C                 | 15                | 0.16               |
| G                 | 13                | 0.14               |
| Total:            | 212               | 2.25               |
| Dinucleotide:     |                   |                    |
| AT/TA             | 548               | 5.81               |
| AC/CA             | 517               | 5.48               |
| AG/GA             | 712               | 7.55               |
| CG/GC             | 191               | 2.03               |
| GT/GT             | 555               | 5.89               |
| CT/TC             | 705               | 7.48               |
| Total:            | 3228              | 34.24              |
| Trinucleotide:    |                   |                    |
| Total:            | 1048              | 11.12              |
| Tetranucleotide:  |                   |                    |
| Total:            | 219               | 2.32               |
| Pentanucleotide:  |                   |                    |
| Total:            | 40                | 0.42               |
| Hexanucleotide:   |                   |                    |
| Total:            | 8                 | 0.08               |

From Table 1, it is clear that dinucleotide is the most abundant tandem repeat motifs (3228 occurrence; relative frequency=34.24) in GIF, following with trinucleotide tandem repeat motifs (1048 occurrence; relative frequency=11.12). The occurrence of tandem repeats has recorded a decline from tetranucleotide to hexanucleotide. This is expected because the probability of the occurrence of tandem repeats is reduced as the length of motif is increased. Our results are consistent with the findings of Astolfi et al. [47] in general, the work which demonstrated that eukaryotic genes (that of human chromosome 21+22, M. musculus, D. melanogaster, C. elegans, A. thaliana, and S. cerevisiae) are abundant with dinucleotide repeat motif in the comparison with trinucleotide and tetranucleotide repeat motif. Besides, the report of Ouyang et al. [48] was also consistent with our results in terms of the abundance of different tandem repeat motifs, with the exception that mononucleotide tandem repeats were observed being the most abundant in their study (relative frequency=4.36, genome-wide). The deviation between the findings of Ouyang et al. and ours in the abundance level of mononucleotide tandem repeat could be accounted by the fact that different organisms from different kingdoms have been used as the study subject, that is, human gene being the subject in this research whereas virus gene was the subject in Ouyang et al.’s study. However, by comparing the relative frequency of mononucleotide tandem repeats of our findings and that of Ouyang et al. [48], it was found that both our result (relative frequency=2.25) and Ouyang et al.’s (relative frequency=4.36) do not mark a very high value of relative frequency. This implies that the occurrence of mononucleotide tandem repeats in human and virus gene is relatively scarce in the complete gene sequence. Besides, a comparison between the relative
frequency of other types of tandem repeat motif shows that the human GIF has more tandem repeats than virus gene.

The number of repeat-in-tandem for the mononucleotide is given in Table 2.

Table 2. Number of repeat-in-tandem for mononucleotide

| Motif with the no. of repeat-in-tandem | No. of occurrence | Percentage |
|--------------------------------------|-------------------|------------|
| A                                    | 42                | 42%        |
| A                                    | 16                | 16%        |
| A                                    | 6                 | 6%         |
| A                                    | 10                | 10%        |
| A                                    | 3                 | 3%         |
| A                                    | 2                 | 2%         |
| A                                    | 3                 | 3%         |
| A                                    | 2                 | 2%         |
| A                                    | 3                 | 3%         |
| A                                    | 1                 | 1%         |
| A                                    | 2                 | 2%         |
| A                                    | 5                 | 5%         |
| A                                    | 1                 | 1%         |
| A                                    | 1                 | 1%         |
| A                                    | 1                 | 1%         |
| T                                    | 6                 | 6%         |
| T                                    | 7                 | 7%         |
| T                                    | 5                 | 5%         |
| T                                    | 2                 | 2%         |
| T                                    | 1                 | 1%         |
| T                                    | 1                 | 1%         |
| T                                    | 1                 | 1%         |
| T                                    | 1                 | 1%         |
| C                                    | 6                 | 6%         |
| C                                    | 1                 | 1%         |
| G                                    | 6                 | 6%         |
| G                                    | 7                 | 7%         |

In Table 2, the number of repeat-in-tandem is analyzed for each mononucleotide motif. G and C tandem repeats do not have a wide variety of repeat length as A and T tandem repeats do. Most of the tandem repeats of C and G, which are more than 90%, have the length of 6 for the repeat-in-tandem. Although a majority of A and T tandem repeats also demonstrates a length of 6 for the repeat-in-tandem, the percentage is apparently lower (mononucleotide A$_6$=42%; T$_6$=55.95%) than that of C (93.33%) and G (92.31%).

The analysis of the number of repeat-in-tandem for the dinucleotide is given in Table 3.

Table 3. Number of repeat-in-tandem for dinucleotide

| Motif with the no. of repeat-in-tandem | No. of occurrence | Percentage |
|--------------------------------------|-------------------|------------|
| (TC)$_2$                             | 326               | 91.32%     |
| (TC)$_3$                             | 27                | 7.56%      |
| (TC)$_4$                             | 3                 | 0.84%      |
| (CG)$_2$                             | 61                | 91.04%     |
| (CG)$_3$                             | 6                 | 8.96%      |
| (AG)$_2$                             | 338               | 93.11%     |
| (AG)$_3$                             | 20                | 5.51%      |
| (AG)$_4$                             | 5                 | 1.38%      |
| (GC)$_2$                             | 112               | 90.32%     |
| (GC)$_3$                             | 12                | 9.68%      |
| (GA)$_2$                             | 319               | 91.40%     |
| (GA)$_3$                             | 26                | 8.54%      |
| (GA)$_4$                             | 3                 | 0.86%      |
| (GA)$_5$                             | 1                 | 0.29%      |
| (AT)$_2$                             | 299               | 90.06%     |
| (AT)$_3$                             | 26                | 9.73%      |
| (AT)$_4$                             | 5                 | 0.60%      |
| (AT)$_5$                             | 2                 | 0.29%      |
| (AC)$_2$                             | 181               | 94.76%     |
| (AC)$_3$                             | 10                | 5.24%      |
| (CA)$_2$                             | 288               | 88.34%     |
| (CA)$_3$                             | 35                | 10.74%     |
| (CA)$_4$                             | 3                 | 0.92%      |
| (CT)$_2$                             | 316               | 90.81%     |
| (CT)$_3$                             | 29                | 8.33%      |
| (CT)$_4$                             | 3                 | 0.86%      |
| (TG)$_2$                             | 283               | 91.29%     |
| (TG)$_3$                             | 18                | 5.81%      |
| (TG)$_4$                             | 5                 | 1.61%      |
| (TG)$_5$                             | 1                 | 0.32%      |
| (TG)$_6$                             | 1                 | 0.32%      |
| (TG)$_7$                             | 1                 | 0.32%      |
| (GT)$_2$                             | 210               | 85.71%     |
| (GT)$_3$                             | 28                | 11.43%     |
| (GT)$_4$                             | 6                 | 2.45%      |
| (GT)$_5$                             | 1                 | 0.41%      |
| (TA)$_2$                             | 194               | 89.81%     |
| (TA)$_3$                             | 17                | 7.87%      |
| (TA)$_4$                             | 3                 | 1.39%      |
| (TA)$_5$                             | 1                 | 0.46%      |
| (TA)$_6$                             | 1                 | 0.46%      |

Of a total of 12 types of dinucleotide motifs, it is apparent that all of them dominantly exhibit 2 repeats-in-tandem (XY)$_2$. Higher order of repeat-in-tandem [(XY)$_n$, where $n \geq 10$] has been found in dinucleotide TG, GT, and TA. However, high order of repeat-in-tandem is rare. For the order $n>3$, all 12 types of dinucleotide motifs mark an occurrence of lesser than 3%. Besides, there is a drastic reduced in percentage from the order $n=2$ to $n=3$. For example, (AG)$_3$ marks 93.11% whereas (AG)$_4$ marks 5.51%, which is a gap of 87.6%.

We observed the similar phenomena of the distribution of repeat-in-tandem for tetra-, penta-, and hexanucleotide motifs.
Most of the motifs exhibit an order of 2 of repeat-in-tandem. Table 4 summarizes the repeat-in-tandem for tetra-, penta- and hexanucleotide motif.

Table 4. Summary of the repeat-in-tandem for tetra-, penta-, and hexanucleotide motifs

| Nucleotide motifs with the order of repeat-in-tandem, n | No. of occurrence | Percentage |
|--------------------------------------------------------|-------------------|------------|
| Tetranucleotide:                                       |                   |            |
| n=2                                                    | 208               | 94.98%     |
| n=3                                                    | 7                 | 3.20%      |
| n>3                                                    | 4                 | 1.83%      |
| Pentanucleotide:                                       |                   |            |
| n=2                                                    | 38                | 95%        |
| n=3                                                    | 2                 | 5%         |
| n>3                                                    | 0                 | 0%         |
| Hexanucleotide:                                        |                   |            |
| n=2                                                    | 8                 | 100%       |
| n=3                                                    | 0                 | 0%         |
| n>3                                                    | 0                 | 0%         |

From Table 4, it is striking that more than 90% of the nucleotide motifs exhibit n=2 for the order of repeat-in-tandem. This result is comparable and consistent with the patterns of dinucleotide motifs shown in Table 3. Pentanucleotide motifs and hexanucleotide motifs do not have any repeat-in-tandem that is larger than an order of 3. To take CCTCTC for example, this hexanucleotide motif occurs twice in GIF gene, both in the form of (CCTCTC)2, which are located at loci 943—954 and 67776—67787, respectively. It was found that these nucleotide motifs have lesser repeat-in-tandem of the order larger than 3. Our finding is partially supported by the research done by Yang et al. [49], who have shown that the genome of Shigella flexneri serotype 2a (strain Sf301) has a drastic reduction of tandem-in-repeat for the order which is higher than 3 in mono-, di-, and trinucleotide motifs. In their study, they found that the percentage of tandem-in-repeat for the order higher than 3 is only 6.32% in dinucleotide motifs, but 33.33% in hexanucleotide motifs. However, no comparison is possible between our results and theirs for tetra- and pentanucleotide because they did not disclose the detailed repeat-in-tandem for these motif types. Notably, certain tandem repeat motifs have been discovered in dynamic mutation, such as CCTG [50], TCAT [51], and ATTCT [52]. These tandem repeats which have a high mutation rate are disease inclined. In our study, it was found that CCTG occurs 6 times in the gene of GIF, which is accounted for 2.74% of tetranucleotide total number of occurrence. However, the disease-inclined TCAT and ATTCT were not found in the gene.

Because trinucleotide encodes for protein in the coding region, it has special importance in the study of tandem repeat. A triplet classification system [46] was used to categorize the trinucleotide tandem repeats, as shown in Table 5.

Table 5. Distribution of trinucleotide tandem repeats in GIF

| Type | Repeat motifs (with frequency) | Total |
|------|------------------------------|-------|
| T1   | AAT(18)                      | 120   |
|      | ATA(17)                      |       |
|      | TAA(17)                      |       |
|      | ATT(144)                     |       |
|      | TTA(8)                       |       |
|      | TAT(16)                      |       |
| T2   | AAG(43)                      | 164   |
|      | AGA(33)                      |       |
|      | GAA(35)                      |       |
|      | CTT(19)                      |       |
|      | TTC(15)                      |       |
|      | TCT(19)                      |       |
| T3   | AAC(9)                       | 91    |
|      | ACA(11)                      |       |
|      | CAA(25)                      |       |
|      | GTT(16)                      |       |
|      | TTCG(14)                     |       |
|      | TGC(16)                      |       |
| T4   | ATG(18)                      | 101   |
|      | TGA(20)                      |       |
|      | GAT(6)                       |       |
|      | CAT(18)                      |       |
|      | ATC(21)                      |       |
|      | TCA(18)                      |       |
| T5   | AGT(7)                       | 50    |
|      | GTA(11)                      |       |
|      | TAG(8)                       |       |
|      | ACT(17)                      |       |
|      | CTA(7)                       |       |
|      | TAC(0)                       |       |
| T6   | AGG(53)                      | 180   |
|      | GGA(33)                      |       |
|      | GAG(29)                      |       |
|      | CCT(16)                      |       |
|      | CTT(16)                      |       |
|      | TCC(24)                      |       |
| T7   | AGC(17)                      | 161   |
|      | GCA(29)                      |       |
|      | GAG(32)                      |       |
|      | GCT(22)                      |       |
|      | CTG(29)                      |       |
|      | TGC(32)                      |       |
| T8   | ACG(0)                       | 11    |
|      | CGA(4)                       |       |
|      | GAC(2)                       |       |
|      | CGT(0)                       |       |
|      | GTC(2)                       |       |
|      | TCG(3)                       |       |
| T9   | ACC(15)                      | 119   |
|      | CCA(25)                      |       |
|      | CAC(26)                      |       |
|      | GGT(8)                       |       |
|      | GTG(25)                      |       |
|      | TGG(20)                      |       |
| T10  | GGC(7)                       | 51    |
|      | GCG(10)                      |       |
|      | CGG(6)                       |       |
|      | GCC(18)                      |       |
Notably, the trinucleotide tandem repeat motifs in the type T8 are scarce, which contributes 11 occurrences (1.05%) out of a total of 1048 trinucleotide occurrences. However, Ouyang et al. [48] have found that the T8 type of trinucleotide tandem repeat motifs of Herpes simplex virus type 1 is at the moderate level of abundance (20.65% of the total trinucleotide occurrence). The deviation between our results and that of Ouyang et al.’s indicates that there is an enormous difference between the type of amino acids encoded by human and viruses.

In a research carried out by Astolfi et al. [47], they have found that the occurrences of ACG(T8), ACT(T5), and CCG(T10) were low in human chromosome 21 and 22, M. musculus, D. melanogaster, C. elegans, A. thaliana, and S. cerevisiae. However, our finding on the abundance level of ACG(0) and CCG(3) were low in their respective type, but ACT(17) is the most abundant trinucleotide tandem repeat motif in type T5.

In our study, AGG is the most abundant trinucleotide tandem repeat motif in T6 and across all types. The similar finding was reported by Astolfi et al. [47], where AGG was the most abundant motif in M. musculus.

From our results, we have identified several potential trinucleotide tandem repeat motifs that are associated with diseases. CAG has occurred 32 times, which is relatively high and it has the risk to induce spinocerebellar ataxia [11]. GAA which has occurred 35 times has been identified to exceed the normal range of repeat number (5-30), a case where it is exposed to Friedreich’s ataxia [11]. Lastly, CTG which has a frequency of 29 is exposed to the risk of Huntington’s disease-like 2, as the normal range of repeat number is 6-27. The analyses of the repeat consensus and the loci distribution of these three disease-predisposed trinucleotide tandem repeat motifs are given in Table 6.

| Repeat no | Repeat consensus | Start | End |
|-----------|------------------|-------|-----|
| 9         | (CAG)₁           | 141   | 146 |
| 66        | (CAG)₂           | 1190  | 1195|
| 274       | (CAG)₂           | 5348  | 5353|
| 284       | (CAG)₂           | 5573  | 5578|
| 296       | (CAG)₂           | 5901  | 5906|
| 457       | (CAG)₁           | 9105  | 9110|
| 467       | (CAG)₁           | 9330  | 9335|
| 479       | (CAG)₁           | 9658  | 9663|
| 598       | (CAG)₁           | 11958 | 11963|
| 623       | (CAG)₁           | 12459 | 12464|
| 643       | (CAG)₁           | 13036 | 13041|
| 676       | (CAG)₁           | 13680 | 13685|
| 710       | (CAG)₁           | 14240 | 14245|
| 776       | (CAG)₁           | 15481 | 15486|
| 1541      | (CAG)₁           | 30068 | 30073|
| 2121      | (CAG)₁           | 40507 | 40512|
| 2255      | (CAG)₁           | 43133 | 43138|
| 2256      | (CAG)₁           | 43142 | 43150|
| 2329      | (CAG)₁           | 44418 | 44423|
| 2552      | (CAG)₁           | 48613 | 48618|
| 2625      | (CAG)₁           | 50149 | 50154|
| 2696      | (CAG)₁           | 51699 | 51704|
| 2706      | (CAG)₁           | 51924 | 51929|

| Repeat no | Repeat consensus | Start | End |
|-----------|------------------|-------|-----|
| 154       | (GAA)₂           | 2767  | 2772|
| 251       | (GAA)₂           | 4887  | 4892|
| 262       | (GAA)₂           | 5103  | 5108|
| 282       | (GAA)₂           | 5541  | 5546|
| 294       | (GAA)₂           | 5766  | 5771|
| 431       | (GAA)₂           | 8680  | 8685|
| 442       | (GAA)₂           | 8896  | 8901|
| 465       | (GAA)₂           | 9298  | 9303|
| 474       | (GAA)₂           | 9523  | 9528|
| 732       | (GAA)₂           | 14696 | 14701|
| 940       | (GAA)₂           | 18706 | 18711|
| 1094      | (GAA)₂           | 21658 | 21663|
| 1223      | (GAA)₂           | 23986 | 23991|
| 1321      | (GAA)₂           | 25989 | 25994|
| 1329      | (GAA)₂           | 26128 | 26133|
| 1649      | (GAA)₂           | 31216 | 31212|
| 1657      | (GAA)₂           | 32336 | 32341|
| 1663      | (GAA)₂           | 32426 | 32431|
| 1687      | (GAA)₂           | 32851 | 32856|
| 1872      | (GAA)₂           | 35980 | 35985|
| 2117      | (GAA)₂           | 40413 | 40418|
| 2233      | (GAA)₂           | 42742 | 42747|
| 2567      | (GAA)₂           | 48907 | 48912|
| 2673      | (GAA)₂           | 51238 | 51243|
| 2684      | (GAA)₂           | 51454 | 51459|
| 2704      | (GAA)₂           | 51892 | 51897|
| 2713      | (GAA)₂           | 52117 | 52122|
| 2853      | (GAA)₂           | 55031 | 55036|
| 2864      | (GAA)₂           | 55247 | 55252|
| 2884      | (GAA)₂           | 55685 | 55690|
| 2893      | (GAA)₂           | 55910 | 55915|
| 3089      | (GAA)₂           | 59744 | 59749|
| 3368      | (GAA)₂           | 65229 | 65234|
| 4676      | (GAA)₂           | 89872 | 89877|
| 4710      | (GAA)₂           | 90353 | 90358|

Table 6. The repeat consensus and the loci distribution of the three disease-predisposed motifs CAG, GAA, and CTG.
A graph is provided in Figure 1 to depict the distribution of these disease-predisposed trinucleotide tandem repeat motifs (as shown in Table 6) in the gene.

Fig. 1 Distribution of the trinucleotide tandem repeat motifs

From Fig. 1, it is apparent that most of the tandem repeat motifs of these disease-predisposed trinucleotides are located on the first (<30k bp) and the second segment (30k bp – 60k bp) of the gene. For CAG, 14 tandem repeats were found on the first segment whereas 13 tandem repeats were located on the second segment. For GAA, 15 were found on the first segment while 17 motifs were found on the second segment. CTG displays a similar pattern of motif distribution, where 11 were found on the first segment and 10 on the second segment. Therefore, particular attention needs to be paid to these gene segments for therapeutic strategies and drug design.

IV. CONCLUSIONS

An in silico analysis has been carried out to study the tandem repeats in GIF. Six types of the nucleotide tandem repeat motifs have been investigated, including mono-, di-, tri-, tetra-, penta- and hexanucleotide. It was found that the relative frequency of dinucleotide tandem repeat motifs is highest, following with that of trinucleotide. The probability of the occurrence of tandem repeats is reduced as the length of motif is increased. Besides, from the analysis it was found that most of the motifs exhibit an order of 2 of repeat-in-tandem, and higher order of repeat-in-tandem is rare in GIF. The analysis of the trinucleotide tandem repeat motifs has identified three motifs which are disease-predisposed. The loci of these motifs have been identified and this information is important for the mutation analysis carried out by geneticists and biomedical scientists. Our analysis of the tandem repeats in GIF sheds light on the better understanding of the diseases implicated by this gene at the molecular level.

| 3013 | (CTG)$_2$ | 58384 | 58389 |
| 3163 | (CTG)$_2$ | 61415 | 61420 |
| 3251 | (CTG)$_3$ | 63036 | 63044 |
| 3599 | (CTG)$_2$ | 69662 | 69667 |
| 4103 | (CTG)$_2$ | 79070 | 79075 |
| 4156 | (CTG)$_2$ | 80023 | 80028 |
| 4343 | (CTG)$_2$ | 83521 | 83526 |
| 4371 | (CTG)$_2$ | 83985 | 83990 |
| 4572 | (CTG)$_2$ | 87928 | 87933 |
23] S. Tuna and M. Niranjam, “Reducing the algorithmic variability in transcriptome-based inference,” Bioinformatics, vol. 26, no. 9, pp. 1185-1191, 2010.

24] Y. Yamaki, Y. Maki, M.B. Swindells, and M. Suwa, “GENIUS II: a high-throughput database system for linking ORFs in complete genomes to known protein three-dimensional structures,” Bioinformatics, vol. 20, no. 4, pp. 596-598, 2004.

25] G.B. Fogel, D.G. Weekes, G. Varga, E.R. Dow, A.M. Craven, H.B. Harlow, E.W. Su, J.E. Ouyia, and C. Su, “A statistical analysis of the TRANSFAC database,” BioSystems, vol. 81, pp. 137-154, 2005.

26] A. Bhasi, P. Philip, V. Manikandan, and P. Senapathy, “ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes,” Nucleic Acids Research, vol. 37, pp. D703-D711, 2009.

27] C.I. Richardson, Q. Gao, C. Mitsopoulos, M. Zvelebil, L.H. Pearl, and P.M.G. Pearl, “MOKCa database—mutations of kinases in cancer,” Nucleic Acids Research, vol. 37, pp. D824-D831, 2009.

28] C. Llorens, R. Futami, L. Covelli, L. Domínguez-Escribá, J.M. Viu, D. Tamarit, J. Aguilar-Rodríguez, M. Vicente-Ripolles, G. Fuster, G.P. Bernet, F. Maunus, A. Munoz-Pomer, J.M. Sempere, A. Latorre, and A. Moya, “The Gypsy Database (GyDB) of mobile genetic elements: release 2.0,” Nucleic Acids Research, vol. 39, pp. D70-D74, 2011.

29] C.H. Wu, S. Zhao, and H-L. Chen, “A protein class database organized with ProSite protein groups and PIR superfamilies,” Journal of Computational Biology, vol. 5, no. 4, pp. 547-561, 1996.

30] C. Lushbough, M.K. Bergman, C.J. Lawrence, D. Jennewein, and V. C. Lushbough, “BioExtract server—an integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data,” Nucleic Acids on Computational Biology and Bioinformatics, vol. 7, no. 1, pp. 12-24, 2010.

31] 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, pp. 1524-1529, 2008.

32] M. Marjalevičius, M. Laganekas, and Č. Venclovas, “COMA server for protein distant homology search,” Bioinformatics, vol. 26, no. 15, pp. 1905-1906, 2010.

33] C. Hemmerich, A. Buechlein, R. Podicheti, K.V. Revanna, and Q. Dong, “An Ergatis-based prokaryotic genome annotation web server,” Bioinformatics, vol. 26, no. 8, pp. 1122-1124, 2010.

34] L. Falquet, L. Bordoli, V. Ioannidis, M. Pagni, and C.V. Jongeneel, “Swiss-Model: Ab initio protein structure homology modeling,” Bioinformatics, vol. 26, no. 3, pp. 302-309, 2009.

35] J. Jurka and C. Pethiyagoda, “Simple repetitive DNA sequences from primates: compilation and analysis,” Journal of Molecular Evolution, vol. 40, pp. 120-126, 1995.

36] F. Astolfi, D. Bellizzi, and V. Sgaramella, “Frequency and coverage of trinucleotide repeats in eukaryotes,” Gene, vol. 317, pp. 117-125, 2003.

37] 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, pp. 37-40, 2012.

38] J. Yang, J. Wang, L. Chen, J. Yu, J. Dong, Z-J. Yao, Y. Shen, Q. Jin, and R. Chen, “Identification and characterization of simple sequence repeats in the genomes of Shigella species,” Gene, vol. 322, pp. 85-92, 2003.

39] 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, pp. 524-536, 2010.

40] T. Le, T. Altman, and K. Gardiner, “HIGEDA: a hierarchical gene-set genetics based algorithm for finding subtle motifs in biological sequences,” Bioinformatics, vol. 26, no. 3, pp. 302-309, 2010.

41] F.M.G. Pearl, “MoKCa database—mutations of kinases in cancer,” C.J. Richardson, Q. Gao, C. Mitsopoulos, M. Zvelebil, L.H. Pearl, and P.M.G. Pearl, “MOKCa database—mutations of kinases in cancer,” Nucleic Acids Research, vol. 37, pp. D824-D831, 2009.

42] C. Llorens, R. Futami, L. Covelli, L. Domínguez-Escribá, J.M. Viu, D. Tamarit, J. Aguilar-Rodríguez, M. Vicente-Ripolles, G. Fuster, G.P. Bernet, F. Maunus, A. Munoz-Pomer, J.M. Sempere, A. Latorre, and A. Moya, “The Gypsy Database (GyDB) of mobile genetic elements: release 2.0,” Nucleic Acids Research, vol. 39, pp. D70-D74, 2011.

43] C.H. Wu, S. Zhao, and H-L. Chen, “A protein class database organized with ProSite protein groups and PIR superfamilies,” Journal of Computational Biology, vol. 5, no. 4, pp. 547-561, 1996.

44] C. Lushbough, M.K. Bergman, C.J. Lawrence, D. Jennewein, and V. C. Lushbough, “BioExtract server—an integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data,” Nucleic Acids on Computational Biology and Bioinformatics, vol. 7, no. 1, pp. 12-24, 2010.

45] 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, pp. 1524-1529, 2008.

46] M. Marjalevičius, M. Laganekas, and Č. Venclovas, “COMA server for protein distant homology search,” Bioinformatics, vol. 26, no. 15, pp. 1905-1906, 2010.

47] C. Hemmerich, A. Buechlein, R. Podicheti, K.V. Revanna, and Q. Dong, “An Ergatis-based prokaryotic genome annotation web server,” Bioinformatics, vol. 26, no. 8, pp. 1122-1124, 2010.

48] L. Falquet, L. Bordoli, V. Ioannidis, M. Pagni, and C.V. Jongeneel, “Swiss EMBlnet node web server,” Nucleic Acids Research, vol. 31, no. 13, pp. 3782-3783, 2003.

49] G. Hu and D. Li, “Three-dimensional modeling of transport of nutrients for multicellular tumor spheroid culture in a microchannel,” Biomedical Microdevices, vol. 9, pp. 315-323, 2007.

50] H. Li and R. Luo, “Modeling and characterization of glucose-sensitive hydrogel: effect of Young’s modulus,” BioSensors and Bioelectronics, vol. 24, pp. 3630-3636, 2009.

51] C. Xiao and J.H.T. Luong, “A simple mathematical model for electric cell-substrate impedance sensing with extended applications,” Biosensors and Bioelectronics, vol. 25, pp. 1774-1780, 2010.

52] H. Fatehi, M. Meyer-Hermann, and M.T. Figge, “Modelling cellular aggregation induced by chemotaxis and phototaxis,” Mathematical Medicine and Biology, vol. 27, pp. 373-384, 2010.

53] B. West, F.L.H. Brown, and F. Schmid, “Membrane-protein interactions in a generic coarse-grained model for lipid bilayers,” Biophysical Journal, vol. 96, pp. 101-115, 2009.

54] J.H. Choi, B.C.H. May, H. Wille, and F.E. Cohen, “Molecular modeling of the misfolded insulin subunit and amyloid fibril,” Biophysical Journal, vol. 97, pp. 3187-3195, 2009.

55] S. Inada, J.C. Hancox, H. Zhang, and M.R. Boyett, “One-dimensional mathematical model of the atrioventricular node including atrio-nodal, nodal, and nodal-his-cells,” Biophysical Journal, vol. 97, pp. 2117-2127, 2009.

56] S-Y. Huang, D. Bolser, H-Y. Liu, T-C. Hwang, and X. Zou, “Molecular modeling of the heterodimer of human CFTR’s nucleotide-binding domains using a protein-protein docking approach,” Journal of Molecular Graphics and Modelling, vol. 27, pp. 822-828, 2009.

57] R. Kolpakov, G. Bana, and G. Kucherov, “mcreps: efficient and flexible detection of tandem repeats in DNA,” Nucleic Acids Research, vol. 31, no. 13, pp. 3672-3678, 2003.