Computational Study of Viral Segments Inserted within the Regions of Human Genome

P. Pandarinath*, M. Shashi and A. Appa Rao

Department of Computer Sciences, College of Engineering (A) Andhra University, Visakhapatnam–530 003, Andhra Pradesh, India

Abstract

Humans have been carrying unwanted viral gene segments since many years and reports suggests that approximately 3-8% of the human genome has been comprised of viral DNA. In this point of view, various viral sequences were downloaded from NCBI Tax Browser and scanned against complete genome of *Homo sapiens* for the presence of possible viral inserts in human genome. The results from the computational analysis revealed that dengue virus resulted in viral segments inserted in the intron regions of human genome and exon region insertions were observed with polio and simian enterovirus. The alignments which show > 25-30 residues, 90-100% identities and the sequences located in the exon regions were considered.

Keywords: Viral segments; *Homo sapiens*; Insertions; Intron; Extron; Computational

Introduction

The human genome is the genetic blue print for all cellular structures and activities of the human body. Genes direct the manufacture of cellular proteins which allow cells to carry out normal mechanisms. Defective or mutated genes can direct the cell to make aberrant proteins which in turn cause diseases. There is a growing public demand for genetic information; a heightening curiosity and bewilderment about one's own genetic predisposition (Teri, 2001).

Every human contains a significant amount of DNA that is not actually human; and contains dormant fossil viruses that have infiltrated the genome. Estimates range from 3-8% of the human genome as being comprised of sections of viral DNA. During the course of a viral infection, some viruses insert their DNA into the host's genome (proviruses) and direct the host cellular machinery to make the proteins and genetic material needed to make/assemble more viruses. If this gene insertion takes place in a cell that will become an egg or sperm, the host's offspring's will have a copy of the virus in every single cell. These and other parasite, self-replicating pieces of nucleic acids have evolved with us over millions of years after being inserted into the human DNA by the virus that infected our ancestors.

According to Cullen, an ancient family of viruses known as HERV-K (for human endogenous retrovirus K), took up permanent residence in the genetic material of Old World monkeys shortly after they diverged from New World monkeys. The viruses then traveled with their simian and pre-human hosts as these species moved along the evolutionary path that led to *Homo sapiens* (Jin et al., 1999).

Because of these viral gene insertion events, genetic material from inactive viruses accounts for roughly 3 percent of the human genome. Nearly 30-50 copies of HERV-K exist in the human genome, and that some of the copies appear to be active at a low level in normal testicular and placental tissue (Jin et al., 1999). Earlier Japanese researchers found copies of the Bornavirus N (for nucleoproteins) gene inserted an atleast four separate locations in the human genome and clearly provided a fossil record of bornavirus (Tina, 2010).

It has been since decades that many viruses have been infecting vertebrates, humans in particular. Experimental evidence suggests that HIV, Epstein-Barr viral segments were inserted in few regions of the human genome. As there are many infectious viruses such as polio, enterovirus, etc, an attempt has been made to identify the insertions of various viral gene segments in either exon/intron region of human genome. To the best of the knowledge, till date there is no evidence of computational analysis to determine the presence of viral inserts in human genome. Keeping this aspect into consideration, various viral sequences were downloaded from NCBI Tax Browser and scanned against complete genome of *Homo sapiens* for the presence of any inserted segments.

Material and Methods

The genomic sequences of *polio virus*, simian entero virus and dengue virus were downloaded from the NCBI Genbank database. *Blastn* algorithm was used to perform simlarity search using default parameters and scanned against the complete genome of *Homo sapiens*. The option, somewhat similar sequences were selected in order to obtain medium length matches. The program compares the query sequence with the human genome sequence and calculates the statistical significance of matches (Altschul et al., 1990).

Sequence tools are used for basic and advanced analysis of nucleotide and protein sequence. The hydropathy plot was used to find cluster of hydrophobic amino acids, which indicate that the polypeptide in question is a transmembrane protein. Molecular Genetic Evolutionary Analysis (MEGA 4) was readily useful for making phylogenic trees and aligning sequences. Discovery studio software was used to perform sequence comparison and structure visulaization. The software compares a query sequence with our own sequence and produces a graphical dot plot using dot plot function and displays maps, sites and enzymes.

Results

Polio virus

The blast result of the *polio virus* 1 strain sabin nucleotide sequence and the human genome shows that the sequence was present in the exon region of sialidase-4 gene (chromosome 2) sequence of *Homo sapiens* with a match of 32 residues (Supplementary Figures 1 (a) and 1 (b)). The sequence in fasta format of the sialidase 4 gene (subject sequence region) and human *polio virus* 1 strain sabin (query sequence region) were aligned using bl2Seq. The encoded amino acid matching for both the sequences i.e., PPQSPTWLLYS was visible in the sequence view of the graphics page (Supplementary Figures 2 (a) and 2 (b)). The result of SEQ tools shows that the codon usage and the GC content of the whole genomic sequence was 3446 (46%) (Supplementary Figure 3). The protein sequence shows the properties and the hydropathy plot i.e., protein length was 522 aa and the molecular weight was 55,351 (Supplementary Figure 4). WebLab Viewer generates and display surfaces and features of sialidase 4 (Supplementary Figure 5). The identical pairs (32), transitional pairs (0) and transversional pairs (2) were also obtained (Supplementary Figures 6 (a) and 6 (b)) using MEGA4 software.

Simian entero virus

The blast result of the simian entero virus nucleotide sequence and the human genome shows that the sequence was present in the exon region of cell division cycle 40 homolog gene (chromosome 6) sequence of *Homo sapiens* with a match of 34 residues (Supplementary Figure 7) and the match of amino acid coding of both the sequences i.e., DWLKILGDWGC was shown graphically (Supplementary Figure 8). The codon usage and the GC content of the whole genomic sequence was 2930 (44.6%) (Supplementary Figure 9). The length of the protein was 616 aa and the molecular weight was 69,548 (Supplementary Figure 10). The surface properties and features of cell division cycle 40 homolog were presented (Supplementary Figure 11). The identical pairs (34), transitional pairs (2) and transversional pairs (3) were also obtained (Supplementary Figures 12 (a) and 12 (b)).

Dengue virus

The blast result of the dengue virus nucleotide sequence and the human genome shows that the sequence was present in the exon region of inositol polyphosphate-5-phosphatase A (chromosome 10) sequence and ccctc-binding factor (chromosome 16) sequence of *Homo sapiens* with a match of 34 and 39 residues, respectively (Supplementary Figures 13 (a) and 13 (b)) but the graphics has not shown the residues in the presence of the exon region, where as it was proved that the residues were present in the intron region (Supplementary Figure 14). The result of SEQ tools shows that the codon usage and the GC content of the whole genomic was 7439 (48.7%) (Supplementary Figure 15).

Conclusions

From computational analysis, it was observed that Dengue virus resulted in viral segments inserted in the intron regions of human genome, however, exon region insertions were observed with polio and simian enterovirus. Nearly 33 nucleotide residue segment (91% identities) of *Polio virus* 1 strain sabin was found to be inserted in two different exon region of Sialidase-4 gene (chromosome 2) sequence of *Homo sapiens*. Comparing the genome of simian enterovirus SV6 against human genome resulted in 85% match (40 residues) with cell division cycle 40 homolog gene sequence of chromosome 6. The work indicates that with few computational efforts viral inserts in human genome can be identified and represents that further analysis needs to be carried out to study the influence of such inserts in the structural and functional features of respective genes.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

2. Jin Y, Hal PB, Sheila P, Heather W, Ray T, et al. (1999) An ancient family of human endogenous retroviruses encodes a functional homolog of the HIV-1 Rev protein. Proc Natl Acad Sci USA 96: 13404-13408.

3. Teri C (2001) Gene Therapy: Pharmaceutical Alteration of the Human Genome. MedSurg Nursing

4. Tina HS (2010) Bornavirus genes found in human DNA Science News 177: 10.