IN SILICO COMPARATIVE GENOME ANALYSIS OF HEPATITIS B AND HEPATITIS C VIRUS

Budhayash Gautam*, Shashi Rani, Satendra Singh and Rohit Farmer
Department of Computational Biology and Bioinformatics,
Sam Higginbottom Institute of Agricultural, Technology and Sciences, Allahabad-211007, U.P., INDIA
***Email:** budhayashgautam@gmail.com

ABSTRACT

In the present study, comparative genome analysis of Hepatitis B and C is done. The similarity and conservation of sequences were analyzed at the genome level by *In silico* approaches. The study revealed that both the sequences have identical conservation at the sequence level with each other. Both the genomes contain same numbers of the genes and sizes of the genes are almost similar. Most of the sequence patterns of both strains are identical. Thus, although the viruses possessed different size of the genome and slightly different positions and numbers of repeats, they were containing almost similar information at the genome level. Also, it may be possible that hepatitis C has added some genetic information to its viral genome and it may be evolved from hepatitis B.

Keywords: Comparative Genomics, Hepatitis, Patterns, Tandem Repeats.

INTRODUCTION

Genome analysis entails the prediction of genes in uncharacterized genomic sequences. The objective is to be able to take a newly sequenced uncharacterized genome and break it up into introns, exons, repetitive DNA sequences, transposons etc. and other elements. Several genetic disorders like Huntington's disease, Parkinson's disease, sickle cell anemia etc. are caused due to mutations in the genes or a set of genes inherited from one generation to another. There is a need to understand the cause for such disorders. An understanding of the genome organization can lead to concomitant progresses in drug target identification. Comparative genomics has become a very important emerging branch with tremendous scope, for the above mentioned reasons. If the genome for humans and a pathogen, a virus causing harm is identified, comparative genomics can predict possible drug targets for the invader without causing side effects to humans[1]. Comparative genomics is an exciting new field of biological research in which the genome sequences of different species of human, mouse and a wide variety of other organisms from yeast to chimpanzees are compared. By comparing the finished reference sequence of the human genome with genomes of other organisms, researchers can identify regions of similarity and difference. This information can help scientists better understand the structure and function of human genes and thereby develop new strategies to combat human disease. Comparative genomics also provides a powerful tool for studying evolutionary changes among organisms, helping to identify genes that are conserved among species, as well as genes that give each organism its unique characteristics[2].

The main objectives of the present study were to find out the sequence similarity and sequence conservation between Hepatitis B and C.

MATERIALS AND METHODS

Sequence retrieval The genomic sequences of hepatitis B and hepatitis C were retrieved from the "National Centre for Biotechnology Information", (NCBI) (http://www.ncbi.nlm.nih.gov), genome database using hepatitis B and hepatitis C as keywords in the fasta file format. There accession i.d., are NC_003977 and NC_004102 respectively. Sequence of Hepatitis B virus is complete genome sequence, dsDNA; circular; having length of 3,215 nucleotides and its replicon type is viral segment. Sequence of Hepatitis C virus is complete genome sequence, ssRNA; linear; having length of 9,646 nucleotides and its replicon type is viral segment.

Sequence alignment

Pairwise sequence alignment of Hepatitis B and Hepatitis C genomic sequences was done using ClustalW[3].

Genes and proteins prediction

Genes were predicted in both hepatitis B and hepatitis C using FGENESV tool (http://linux1.softberry.com/berry.phtml). Hypothetical proteins coded by these genes were also predicted in both hepatitis B and hepatitis C using same tool.

Tandem repeats identification

Tandem repeats were identified within the genomic sequences of hepatitis B and hepatitis C with the help of Tandem Repeat Finder tool[4].

Pattern identification

Conserve sequences or patterns were predicted in the hypothetical proteins of both hepatitis B and hepatitis C by using PROSCAN tool (http://npsapbil.ibcp.fr/cgibin/npsa_automat), and Pfam Search tool (http://pfam.sanger.ac.uk/search).

Comparative analysis

In the final step of this work, all the results obtained by above mentioned tools and steps, for hepatitis B and hepatitis C were compared to each other.

RESULTS AND DISCUSSION

Pairwise sequence alignment of Hepatitis B and Hepatitis C genomic sequences was done using ClustalW. It was clearly seen that there is very large difference in the size of the genomic sequences of the two most pathogenic strains of hepatitis i.e. Hepatitis B and Hepatitis C. Hepatitis B was completely aligned up to its whole length with a region of hepatitis C (Fig. 1), with good amount of sequence similarity and it was seen as conserve region in the two genome[5].

Figure 1: Pairwaise sequence alignment of Hepatitis B and Hepatitis C genomes showing sequence similarity between nucleotide position 29 – 3215 and 2200 – 5697 of Hepatitis B and Hepatitis C respectively.

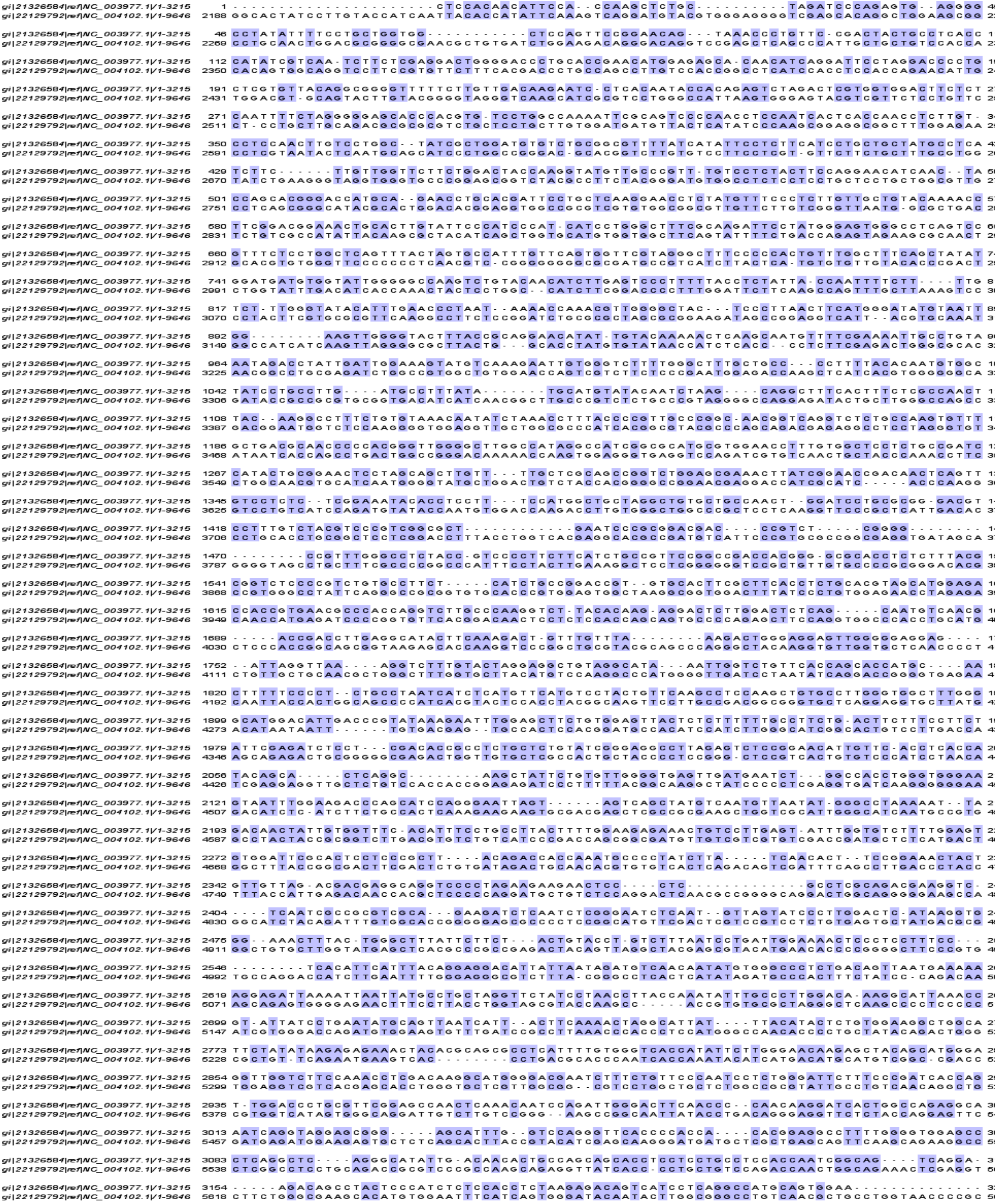

Table 1: Representing Genes and Hypothetical Proteins

N	S		Start	End	Score	Protein length
Hepatitis B						
1	+	CDS	155	835	3243	226 a.a.
2	+	CDS	421	1623	2510	400 a.a.
3	+	CDS	1374	1838	924	154 a.a.
4	+	CDS	1814	2452	1408	212 a.a.
5	+	CDS	2446	2604	378	52 a.a.
Hepatitis C						
N	S		Start	End	Score	Protein length
1	+	CDS	342	9377	6813	3011 a.a.
2	-	CDS	5477	5998	168	173 a.a.
3	-	CDS	6641	7063	261	140 a.a.
4	+	CDS	7276	7680	121	134 a.a.
5	-	CDS	9356	9604	129	82 a.a.

Total five genes and five related hypothetical proteins were predicted in the each of the two hepatitis strains (Table 1 A and B). It was revealed from the tables that genes present in hepatitis B were only present on + strand. While in hepatitis C genes 1st and 4th were present on + strand and genes 2nd, 3rd and 5th were present on – strand. Also genes were varying in lengths but only one gene in each of the strain was having larger size than others (i.e. gene 2nd in hepatitis B and gene 1st in hepatitis C) and all other genes in hepatitis B and hepatitis C were of similar sizes[6].

One tandem repeat was found in the genome of the hepatitis B, while 6 repeats were found in the genome of hepatitis C (Table 2).

Table 2: Tandem repeats/Patterns found in hepatitis B and C

	Consensus pattern	
	Size	**Pattern**
Hepatitis B	12bp.	AGGTCTTACACA
Hepatitis C	24bp.	TTTTTTTTTTTTTTTTTTTCCTTC
	1bp.	T
	7bp.	TTTTTTC
	22bp.	TTTTTTTTTTTTTCTTTCCTTC
	21bp.	TTTTTTCCTTTCTTTTCCTTC
	45bp.	TTTTTTTTTTTTTTTTTTTCTTTCCTTTTTTTT CCTTTCTTTCCC

Repeats obtained in hepatitis B and hepatitis C, were totally different in the size as well as in the patterns. Some of the repeats were too short as having only one base pair and some of them were too large having size of 45 base pairs. As Prosite and Pfam databases are based on the patterns of protein sequences, hypothetical proteins were submitted to prosite and Pfam database as query sequences. On submission of the hypothetical protein sequences of hepatitis B to the prosite database, different regular expressions were obtained as outputs (Table 3 A) which were differing in sizes as well as in the patterns. In protein 1st four patterns were identified. In protein 2nd six patterns were identified. In protein 3rd and 4th three patterns in each, were identified and in protein 5th no pattern was identified.

On submission of the hypothetical protein sequences of hepatitis C to the prosite database, different regular expressions were obtained as outputs (Table 3 B) which differ in sizes as well as in the patterns. In protein 1st nine patterns were identified. In protein 2nd, 3rd and 4th four patterns in each, were identified and in protein 5th two patterns were identified. It was clearly seen that in patterns generated in the case of hepatitis C (total 23) were more in numbers than hepatitis B (total 16). But most of the patterns were same in both hepatitis strains e.g. Nglycosylation site, Amidation site, cAMP and cGMP dependent protein kinase phosphorylation site, Nmyristoylation site, Protein kinase C phosphorylation site and Casein kinase II phosphorylation site etc. Thus, there is sequence as well as functional conservation in both the sequences[5].

On submission of the hypothetical protein sequences of hepatitis B to the Pfam database, different pfam matches/patterns were obtained as outputs (Table 4 A). In protein 1st two patterns (one significant and one insignificant) were identified. In protein 2nd two patterns (two significant and zero insignificant) were identified. In protein 3rd two patterns (one significant and one insignificant) were identified. In protein 4th two patterns (two significant and zero insignificant) and in protein 5th no pattern (zero significant and zero insignificant) was identified. On submission of the hypothetical protein sequences of hepatitis C to the Pfam database, different Pfam matches/patterns were obtained as outputs (Table 4 B). In protein 1st twenty seven patterns (twelve significant and fifteen insignificant) were identified. In protein 2nd, 3rd, 4th and in protein 5th no pattern (zero significant and zero insignificant) were identified. Most of the patterns obtained in both strains were different[7], except some e.g. Hepatitis core protein, putative zinc finger, Hepatitis core antigen etc. and these patterns were present in only in the 1st hypothetical protein of hepatitis C. While on the other hand in hepatitis B all the patterns were uniformly distributed, thus, there was some additional information present in hepatitis C[8].

Table 3: Representing details of the patterns found in different hypothetical proteins.

	Total No. of Patterns	Name	Pattern
Hepatitis B			
Protein 1	4	N-glycosylation site	N-{P}-[ST]-{P}
		Protein kinase C phosphorylation site	[ST]-x-[RK]
		N-myristoylation site	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}
		Leucine zipper pattern	L-x(6)-L-x(6)-L-x(6)-L
Protein 2	6	N-glycosylation site	N-{P}-[ST]-{P}
		cAMP- and cGMP-dependent protein kinase phosphorylation site	[RK](2)-x-[ST]
		Protein kinase C phosphorylation site	[ST]-x-[RK]
		Casein kinase II phosphorylation site	[ST]-x(2)-[DE]
		N-myristoylation site	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}
		Amidation site	x-G-[RK]-[RK]
Protein 3	3	Protein kinase C phosphorylation site	[ST]-x-[RK]
		Casein kinase II phosphorylation site	[ST]-x(2)-[DE]
		N-myristoylation site	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}
Protein 4	3	cAMP- and cGMP-dependent protein kinase phosphorylation site	[RK](2)-x-[ST]
		Protein kinase C phosphorylation site	[ST]-x-[RK]
		Casein kinase II phosphorylation site	[ST]-x(2)-[DE]
Protein 5	0	NO PATTERN FOUND	
Hepatitis C			
Protein 1	9	N-glycosylation site	N-{P}-[ST]-{P}
		cAMP- and cGMP-dependent protein kinase phosphorylation site	[RK](2)-x-[ST]
		Protein kinase C phosphorylation site	[ST]-x-[RK]
		Casein kinase II phosphorylation site	[ST]-x(2)-[DE]
		Tyrosine kinase phosphorylation site	[RK]-x(2,3)-[DE]-x(2,3)-Y
		N-myristoylation site	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}
		Amidation site	x-G-[RK]-[RK]
		Cell attachment sequence	R-G-D
		ATP/GTP-binding site motif A (P-loop)	[AG]-x(4)-G-K-[ST]
Protein 2	4	N-glycosylation site	N-{P}-[ST]-{P}
		Protein kinase C phosphorylation site	[ST]-x-[RK]
		Casein kinase II phosphorylation site	[ST]-x(2)-[DE]
		N-myristoylation site	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}
Protein 3	4	N-glycosylation site	N-{P}-[ST]-{P}
		Protein kinase C phosphorylation site	[ST]-x-[RK]
		Casein kinase II phosphorylation site	[ST]-x(2)-[DE]
		N-myristoylation site	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}
Protein 4	4	cAMP- and cGMP-dependent protein kinase phosphorylation site	[RK](2)-x-[ST]
		Protein kinase C phosphorylation site	[ST]-x-[RK]
		N-myristoylation site	G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}
		Amidation site	x-G-[RK]-[RK]
Protein 5	2	cAMP- and cGMP-dependent protein kinase phosphorylation site	[RK](2)-x-[ST]
		Amidation site	x-G-[RK]-[RK]

Table 4: Representing details of the patterns found in different hypothetical proteins.

	Total No. of Patterns	Name of Matched Patterns in Pfam-A	
		Significant	Insignificant
(A)Hepatitis B			
Protein 1	2	Major surface antigen from hepadnavirus	TB domain [Transforming growth factor beta binding protein (TB) domain]
Protein 2	2	Reverse transcriptase (RNA-dependent DNA polymerase)	NO PATTERN FOUND
		DNA polymerase (viral) C-terminal domain	
Protein 3	2	Trans-activation protein X	F-box associated
Protein 4	2	Hepatitis core protein, putative zinc finger	NO PATTERN FOUND
		Hepatitis core antigen	
Protein 5	0	NO PATTERN FOUND	NO PATTERN FOUND
(B) Hepatitis C			
Protein 1	27	Hepatitis C virus capsid protein	POPLD (NUC188) domain
		Hepatitis C virus core protein	ADP-ribosylation factor family
		Hepatitis C virus envelope glycoprotein E1	Phosphoribosylglycinamide synthetase, C domain
		Hepatitis C virus non-structural protein E2/NS1	Flavivirus DEAD domain
		Hepatitis C virus non-structural protein NS2	Helicase conserved C-terminal domain
		Hepatitis C virus NS3 protease	Glucose inhibited division protein A
		Hepatitis C virus non-structural protein NS4a	AIR synthase related protein, N-terminal domain
		Hepatitis C virus non-structural protein NS4b	Anemonia sulcata toxin III family
		Hepatitis C virus non-structural 5a protein membrane anchor	Protein of unknown function (DUF1668)
		Hepatitis C virus non-structural 5a zinc finger domain	Exopolysaccharide synthesis, ExoD
		Hepatitis C virus non-structural 5a domain 1b	Protein of unknown function, DUF482
		Viral RNA dependent RNA polymerase	Cobalamin-5-phosphate synthase
			Exo-polysaccharide synthesis, ExoD
			Protein of unknown function (DUF679)
			Probable cobalt transporter subunit (CbtA)
Protein 2	0	NO PATTERN FOUND	NO PATTERN FOUND
Protein 3	0	NO PATTERN FOUND	NO PATTERN FOUND
Protein 4	0	NO PATTERN FOUND	NO PATTERN FOUND
Protein 5	0	NO PATTERN FOUND	NO PATTERN FOUND

CONCLUSION

The complete genomic sequences of hepatitis B and hepatitis C have been compared. The similarity and conservation of sequences were analyzed at the genome level by *In Silico* approaches. Following conclusion were made on the basis of the results obtained in the present study: Both the sequences have identical conservation at the sequence level with each other, as genomic sequence of hepatitis B have very good amount of similarity to the sequence of hepatitis C. Both the genomes contained same numbers of the genes and sizes of the genes were almost similar. Thus probably their genetic contents were same. Most of the Patterns of both strains were identical. Thus, although the viruses possessed different size of the genome and slightly different positions and numbers of repeats, they were containing almost similar information at the genome level. Also, it may be possible that hepatitis C has added some genetic information to its viral genome and it may be evolved from hepatitis B.

Future Work

Present work can be extended on the structural as well as on the functional aspects especially of proteins found within hepatitis B and hepatitis C as three dimensional structures of proteins of both hepatitis B and hepatitis C can be predicted and on the basis of these structures probable functions can be hypothesized.

Acknowledgments: The authors would like to acknowledge the support and facilities provided by the Department of Computational Biology and Bioinformatics, Sam Higginbottom Institute of Agricultural, Technology and Sciences, Allahabad, U.P., India.

REFERENCES

1. Delcher, AL., Kasif, S., Fleischmann, RD., Peterson, J., White, O., and Salzberg, SL. 1999. Alignment of whole genomes, Nucleic Acids Res. Jun 1; 27(11):2369-76.
2. Fitch W.M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19: 99–113
3. Higgins D., Thompson J., Gibson T., Thompson J.D., Higgins D.G., Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673-4680.
4. Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27, No. 2, pp. 573-580.
5. Force A., Lynch M., Pickett F.B., Amores A., Yan Y.L., and Postlethwait J. 1999. Nucleotide preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.
6. Orito E, Mizokami M, Sakugawa H, Michitaka K, Ishikawa K, Ichida T, Okanoue T, Yotsuyanagi H, Iino S. 2001. A case-control study for clinical and molecular biological differences between hepatitis B viruses of genotypes B and C. Japan HBV Genotype Research Group. Hepatology; 33: 218-223.
7. Hino, O., K. Ohtake, and C. E. Rogler. 1989. Features of two hepatitis B virus (HBV) DNA integrations suggest mechanisms of HBV integration. J. Virol. 63:2638–2643.
8. Snel B., Bork P., and Huynen M. 2000. Genome evolution: Gene fusion versus gene fission. Trends Genet. 16: 9–11.

****