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

Budhayash Gautam<sup>\*</sup>, 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<sup>1</sup>. 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<sup>2</sup>.

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<sup>3</sup>.

#### Genes and proteins prediction

Genes were predicted in both hepatitis B and hepatitis C using FGENESV tool (<u>http://linux1.softberry.com/berry.phtml</u>). 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<sup>4</sup>.

#### 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<sup>5</sup>.

**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.

gi 21326584 mJNC_003977.1 /1-3215 gi 22129792 mJNC_004102.1 /1-9646	1
gi 21326584 re1 NC_003977.1 /1-3215	48 CCTATATTTTCCTGCTGGTGGCTCCAGTTCCGGAACAGTAAACCCTGTTC.CGACTACTGCCTCACCC111
gi 22129792 re1 NC_004102.1 /1-9646	2269 CCTGCAACTGGACGCGGGGCGAACGCTGTGATCTGGAAGACAGGGACAGGTCCGAGCTCACCCATTGCTGCTGCCCACCA 2349
gi 21326584 ref NC_003977.1 /1-3215	112 CATATICSTCAA, TCTTCTCGAGGACCTGGGGACCCTGCAACATGGAGGGA, CAACATCAGGATTCCTAGGACCCCTG 180
gi 22129792 ref NC_004102.1 /1-9646	2350 CACAGTGGCAGGTCCTTCCGTGTTCTTTCACGACCCTGCCAGCCTTGTCCACCGGCCTCATCCACCTCCACCAGAACATTG 2430
gi 21326584 re1 NC_003977.1 /1-3215	191 CTCGTGTTACAGGCGGGGTTTTTTCTTGTTGACAAGAATC - CTCACAATACCACAGAGTCTAGACTCGTGGTGGTGGACTTCTCT 270
gi 22129792 re1 NC_004102.1 /1-9646	2431 TGGACGT - GCAGTACTTGTACGGGGTAGGGTCAAGCATCGCGTCCTGGGCCATTAAGTGGGAAGTACGTCGTTCTCCTGTTC 2510
gi 21326584 ref NC_003977.1//1-3215 gi 22129792 ref NC_004102.1//1-9646	271 CAATTTTCTA0000000CACCCCACGT0. TCCT00CCCAAAATTC0CC0TCCCCAACCTCCCCAATCACTCAC
gi 21326584 re1 NC_003977.1 /1-3215	350 CCTCCAACTTGTCCTGGC TATCGCTGGATGTGTCTGCGGCGTTTTATCATATTCCTCTTCATCCTGCTGCTATGCCTCA 428
gi 22129792 re1 NC_004102.1 /1-9646	2591 CCTCGTAATACTGAATGCGGCGATGCCTGGGCGGGGG. GCACGGTCTTGTGTCCTCCTCCTCGT.GTTCTTCTGCCTTGGGTGG 2889
gi 21326584 ref NC_003977.1 /1-3215	429 TCTTCTOTTOTTOCTCTTCTOGACTACCAAGGTATGTTOCCCGTT.TOTCCTCTACTTCCAGGAACATCAACTA 500
gi 22129792 ref NC_004102.1 /1-9646	2870 TATCTGAAGGGTAGGTGGGTGCCCGGAGGGGTCTACGCCTTCTACGGGATGTGGCCTCTCCTCCTCCTCCTGCTGCTGCTGC750
gi 21326584 mf NC_003977.1 /1-3215	501 CCAGGACGGGACCATGCA GAACCTGCACGATTCCTGCTGCTGAGGAACCTCTATGTTTCCCTCTTGTTGCTGCTGTAGAAAACC 579
gi 22129792 mf NC_004102.1 /1-9646	2751 CCTCAGCGGGCATACGCACTGGACACGGAGGTGGCCGCGTGTGTGCGGGGCGTGTTCTTGTCGGGTTAATG - GCGCTGAC 2830
gi 21326584 ref NC_003977.1 /1-3215	580 TTCGGACGGAAACTGCACTTGTATTCCCATCCCAT - CATCCTGGGCTTTCGCAAGATTCCTATGGGAGTGGGCCTCAGTCC 859
gi 22129792 ref NC_004102.1 /1-9646	2831 TCTGTCGCCCATATTACAAGCGCTACATCAGCTGGTGGATGGGCTGGGCTTCAGTATTTTCTGACCAGAGTAGAAGCGCAACT 2611
gi 21326584 mf NC_003977.1 /1-3215	860 G T T T C T C C T C G G T T A C T A G T C C A T T T G T T C A G T G G T T C G T A G G G C T T T C C C C C C T C T T C G C T T T C A G C T A T A T 740
gi 22129792 mf NC_004102.1 /1-9646	2912 G C A C G T G G G T T C C C C C C C C C C A A C G T C C C G G G G G G G G G G G G G G
gi 21326584 mf]NC_003977.1 /1-3215	741 GGATGATGTGGTATTGGGGGGCCAAGTCTGTACAACATCTTGAGTCCCTTTTTACCTCTATTA·CCAATTTTCTT····TTG 818
gi 22129792 mf]NC_004102.1 /1-9646	2001 CTGGTATTTGACATCACCAAACTACTCCTGGC··CATCTTCGGACCCCTTTGGATTCTTCAAGCCAGTTTGCTTAAAGTCC 3089
gi 21326584 mf NC_003977.1 /1-3215	B17 TCT - TTGGGTATACATTTGAACCCTAAT - · AAAACCAAACGTTGGGGCTAC - · · TCCCTTAACTTCATGGGATATGTAATT B91
gi 22129792 mf NC_004102.1 /1-9646	3070 CCTACTTCGTGCGCGGTTCAAGGCCTTCTCCGGATGTGCGCGCTAGCGCGGAAGATAGGCGGAGGTCATT - · ACGTGCAAAT 3148
gi 21326584 mf]NC_003977.1 /1-3215	892 66 · · · · · · · AAG TTG 666 TACTTTACC6 CAGG AACATAT · TG TACAAAAACTCAAG CAATG TTTTCGAAAATTG CC TG TA 963
gi 22129792 mf]NC_004102.1 /1-9646	3149 66 CCATCATCAAG TAG 666 C6 CTTACT6 · · · 6 CACCTATG TG TATAACCATCTCACC · · CCTCTTCGA6ACT660 C6 CAC 3224
gi 21326584 mJNC_003977.1 /1-3215	964 AATAGACCTATTGATTGGAAAGTATGTCAAAGAATTGTGGGTCTTTTGGGCTTTGCTGCCCCTTTTACACAATGTGGC 1041
gi 22129792 mJNC_004102.1 /1-9646	3225 AACGGCCTGCGAGATCTGGCCGTGGCTGTGGAACCAGTCGTCTTCTCCCGAATGGAGACCAAGCTCATCACGTGGGGGGCA 3305
gi 21326584 ne1 NC_003977.1 /1-3215	1042 TATCCTGCCTTGATGCCTTTATAIGCATGTATACAATCTAAGCAGGCTTTCACTTTCTCGCCAACT 1107
gi 22129792 ne1 NC_004102.1 /1-9646	3306 GATACCGCCGCGTGCGGTGACATCATCAACGGCTTGCCCGTCTCGCCCGTAGGGGCCAGGAGATACTGCTTGGGCCAGCC 3386
gi 21326584 ref NC_003977.1/1-3215	1108 TAC AAGGCCTTTTCTGTGTAAACAATATCTAAACCTTTACCCCGTTGCCCGGC - AAGGGTCAGGTCTCTGCCAAGTGTTT 1185
gi 22129792 ref NC_004102.1/1-9646	3387 GACGGAATGGTCTCCAAGGGGTGGAGGTTGCTGGCGCCCATCACGGCGTACGCCCAGCAGACGAGGCCTCCTAGGGTGT 3467
gi 21326584 me1 NC_003977.1 /1-3215	1188 GETGACCCAACCCCAACGGTTGGGGCTTGGCCATAGGCCATCGGCGCATGCGTGGAACCTTTGTGGCTCCTCTGCCGATC 1288
gi 22129792 me1 NC_004102.1 /1-9646	3488 ATAATCAACCAOCCTGACTGGCCGGGACAAAAACCAAGTGGAGGTGAGGT
gi 21326584 re1 NC_003977.1 /1-3215	1267 CATACTOCOGAACTCCTAOCAOCTTOTTTTOCTCOCAOCCOGTCTOGACCOAACTTATCOGAACCGACAACTCAOTT 1344
gi 22129792 re1 NC_004102.1 /1-9646	3540 CTGGCAACGTOCATCAATGGGGTATGCTGGACTGTCTACCACGGGGCCGGAACGAGGACCATCGCATCACCCAAGG 3824
gi 21326584 m/]NC_003977.1 /1-3215	1345 GTCCTGTC - · TCCGAAATACACCTCCTT - · · TCCATGGCTGCTAGGCTGTGCCAACT - · GGATCCTGCCCGG · GACGT 1417
gi 22129792 m/]NC_004102.1 /1-9646	3825 GTCCTGTCATCCGGTGTATACCAATGTGGACCAAGACCTTGTGGGCTGGCCGGCC
gi 21326584 re1 NC_003977.1 /1-3215	1418 CCTTTOTCTACGTCCCGTCGGCCGTGAATCCCGCGGACGACCCGTCTCGGGG1469
gi 22129792 re1 NC_004102.1 /1-9646	3708 CCTGCACCTGCGCCTCCTCGGACCTTTACCTGGTCACGAGGCACGCCGATGTCATTCCCCGTCGCGCGGCGAGGTGATAGCA 3788
gi 21326584 me1 NC_003977.1 /1-3215	1470 · · · · · · · CCGTTTGGGCCTCTACC · GTCCCCTTCTTCATCTCCCGTTCCGGCCGACCACCGC · GCGCACCTCTTTTACG 1540
gi 22129792 me1 NC_004102.1 /1-9646	3787 GGGGTAGCCTGCTTTCGCCCCGGCCCATTTCCTACTGAAAGGCTCCTCGGGGGGCCCGCTGTTGTGCCCCCGGGGACACG 3887
gi 21326584 re1 NC_003977.1 /1-3215 gi 22129792 re1 NC_004102.1 /1-9646	1541 CGGTCTCCCCGTCTGTGCCTTCTCATCTGCCGGACCGTGTGCACTTCGCTTCACCTCTGCACGTAGCAGACGAGAAACCAAGAAAAAAAA
gi 21326584 ref NC_003977.1/1-3215	1815 CCACCOTOACCOCCACCACOTOTTOCCCAAGOTOT. TACACAAO AGOACTOTTOGACTOTTOGACTOTCAGO CAATOTCAACO 1888
gi 22129792 ref NC_004102.1/1-9646	3949 CAACCATGACCCCGGTGTTCACGGACAACTCCTCTCCACCAGCAGTGCCCCAGAGCTTCCAGGTGCCCCCCCGCACCTGCATG 4029
gi 21326584 me1 NC_003977.1 /1-3215	1659 · · · · ACCCACCTTCAGCCATACTCCAAGACT · GTTTGTTTA · · · · · · AAGACTGGGAGGAGTTGGGGGGGGGG · · · 1751
gi 22129792 me1 NC_004102.1 /1-9646	4030 CTCCCACCGGCAGCGGTAGAGGCACCAAGGTCCCGGCTGCGTGCG
gi 21326584 ref NC_003977.1 /1-3215	1752 · · ATTAGGTTAA · · · AGGTCTTTGTACTAGGAGGCTGTAGGCATA · · · AATTGGTCTGTTCACCAGCACCATGC. · · · AA 1919
gi 22129792 ref NC_004102.1 /1-9646	4111 CTGTTGCTGCAACGCTGGGCTTTGGTGCTTACATGTCCAAGGCCCCATGGGGTTGATCCTAATATCAGGACCGGGGTGAAA 4191
gi 21326584 ref NC_003977.1 /1-3215 gi 22129792 ref NC_004102.1 /1-9646	1820 CTTTTTCCCCT CTGCCTAATCATCTCATGTTCATGTCCTACTGTTCAAGCCTCCAAGCTGTGCCTTGGCTGGC
gi 21326584 ref NC_003977.1 /1-3215	1899 GCATGGACATTGACCCGTATAAAGAATTTGGAGGTTCTGTGGAGTTACTCTCTTTTTTGCCTTCTG-ACTTCTTTCCTTCTTTT
gi 22129792 ref NC_004102.1 /1-9646	4273 ACATAATTAATTTGTGACGAGTGCCACTCCACGGATGCCACATCCATCTTGGCATCGCCATGGCCATGGCCATGGCCATGGCATGACATAAAA
gi 21326584 ref NC_003977.1 /1-3215	1979 ATTCAAGATCTCCTCGACACCCCCTCTCCTCTCTATCGCGAGGCCTTAGAGTCTCCCGAAGATTGTTC.ACCTCACCA 2055
gi 22129792 ref NC_004102.1 /1-9646	4348 AGCAGAGACTGCGGGGGGGGGGGGGGGGGGGGGTGGTCCCCCCCC
gi 21326584 re1 NC_003977.1 /1-3215 gi 22129792 re1 NC_004102.1 /1-9646	2056 TACAGCA CTCAGGC
gi 21326584 ref NC_003977.1/1-3215	2121 O TAATITTOGAAOACCCAO CATCCAOOGAATTAGT AO TCAOCTATOTCAATOTTAATAT. OO CCTAAAAATTA 2192
gi 22129792 ref NC_004102.1/1-9646	4507 GACATCTC.ATCTTCTGCCACTCAAAGAAGAAGAAGGGCGCACGAGCTGGCCGCGAAGCTGGTCGCATGGCCATCAATGCCGTG 4588
gi 21326584 mf NC_003977.1/1-3215	2193 GACAACTATTGTGGTTTC - ACATTTCCTGCCTTACTTTTTGGAAGAGAAACTGTCCTTGAGT - ATTTGGTGTCTTTTTGGAGT 2271
gi 22129792 mf NC_004102.1/1-9646	4587 GCCTACTACCGCGGTCTTGACGTGTCTGTCGTCCGCGCCGACGGCGGGGGGGG
gi 21326584 ref NC_003977.1/1-3215	2272 © T G G A T T C G C A C T C T C C C C C C T A C A G A C C A C C A A T G C C C C T A T C T T A T C A A C A C T T C C G G A A C T A C T C T 2341
gi 22129792 ref NC_004102.1/1-9646	4688 © G C T T T A C C G C G A C T T C A C T C T G A T A G A C T G C A C A C G T G T G T C A C A G T C A C A C T C C A C A C T C C A C A C
gi 21326584 re1/NC_003977.1/1-3215 gi 22129792 re1/NC_004102.1/1-9646	2342 6 T T G T T G A C G A C G A C G A C G C C C C C A G A A G A A G A A G A C T C C G T C
gi 21326584 m1 NC_003977.1/1-3215	2404 I CAATCECCCCCTCCCCCT CAAGATCTCAATCTCCGGAATCTCAAT GTTAGTATCCCTTGGACTC. ATAAGCTG 2474
gi 22129792 m1 NC_004102.1/1-9646	4830 OCCATCTACAGATTTGTGCCACCCGOGGAGCCCCCCTCCGGCATGTTCGACTCGTCCGTCCTCTGTGAGTGCTATGACCCG 4910
gi 21326584 re1/NC_003977.1/1-3215	2475 06 · AAACTITAC · TOGOCTITATICTICT · · · ACTOTACCT · OTCITITAATCCTGATTGGAAAACTCCCTCCTITCC · · · 2545
gi 22129792 re1/NC_004102.1/1-9646	4011 GGCTGTGCTTGGCTATGACCTCACGCCCCGCCGAGACTACAGTTAGGCTACGAGCCTACATGAACACCCCCGGGCCTTCCCCCT 4001
gi 21326584 mf]NC_003977.1/1-3215 gi 22129792 mf]NC_004102.1/1-9646	2548
gi 21326584 ref NC_003977.1/1-3215	2019 AGGAGATTAAAATTAATTATGCCGCTAGGTCTATCCTAACCTAACCAAATATTGCCCTTGGACA-AAGGCATTAAACC 2098
gi 22129792 ref NC_004102.1/1-9646	5071 AGCAGAGTGGGGAGAACTTTCCTTACCTGGTAGCGTACCAAGCCACCGTGTGCGCTAGGGCTCAAGCCCCCTCCCC 5148
gri 21326584 jre1/WC_003977.1/1-3215	2009 0 T ATTAICE IO AA TAID CAGINATCAINT ACTINCAACTAGOGATTAT TTACATACTCTCTGTGGAAGGCTGC272
gri 22129792 jre1/WC_004102.1/1-9646	5147 ATCGTGGGACCAGATGTGGAAGTGTTTGATCCGCCTTAAACCCACCC
gij 2 i 326584 jezijwC_003977. 1/1-3215 gij 22129792 jezijwC_004102. 1/1-9646	2//3 110 TATA TARAGA BAAACTA LA LO LA LO
gri 21326584 jren jwC_003977.1/1-3215 gri 22129792 jren jwC_004102.1/1-9646	2009 GO TEGETO TEGETO CONCERSION CONCERS
gij 212252684/m3/WC_003977.1/1-3215 gij 22129792/m3/WC_004102.1/1-9646	2031 - TOWACCT DE DE TRUGAS COMACTUMANTA TUNAGATI GOGO CITICACCC - CAACAAGATCACTAC TO CAAGAC A A A A A A A A A A A A A A A A
gri 21326584 jre1/wC_003977.1/1-3215 gri 22129792 jre1/wC_004102.1/1-9646	3013 A T C AGO T AGO AG COGO AGO AT 1 10 O T C LAGO O T C ACC C C C A C C C A G C AG C C T T T T C C G T G C A C C C A C C C C A C C C C A C
gi 22129792/mt/WC_003977.1/1-3215 gi 22129792/mt/WC_004102.1/1-9646	5038 CTCGGCCTCCTCGAGACGCGCCCCCCCCCAGAGAGAGTTATCACC.CCTGCTGCCAGACCAGCAGAAACTCGAGGT5017
cil 22129792/m0//C 004102 1/1-9646	



······································						
N	S		Start	End	Score	Protein length
Hepatitis <b>B</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						
Ν	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.

 Table 1: Representing Genes and Hypothetical Proteins

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  $1^{st}$  and  $4^{th}$  were present on + strand and genes  $2^{nd}$ , 3rd and  $5^{th}$  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  $2^{nd}$  in hepatitis B and gene 1st in hepatitis C) and all other genes in hepatitis B and hepatitis C were of similar sizes<sup>6</sup>.

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. TTTTTTTTTTTTTTTTTTCCTTC		
	1bp.	Т	
	7bp.	TTTTTTC	
	22bp.	ттттттттттстттссттс	
	21bp.	тттттсстттстттссттс	
	45bp.	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	

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 3<sup>rd</sup> and

4<sup>th</sup> three patterns in each, were identified and in protein 5<sup>th</sup> 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 2<sup>nd</sup>, 3<sup>rd</sup> and 4<sup>th</sup> 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<sup>5</sup>.

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 1<sup>st</sup> two patterns (one significant and one insignificant) were identified. In protein 2<sup>nd</sup> two patterns (two significant and zero insignificant) were identified. In protein 3<sup>rd</sup> two patterns (one significant and one insignificant) were identified. In protein 4<sup>th</sup> 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 2<sup>nd</sup>, 3<sup>rd</sup>, 4<sup>th</sup> and in protein 5<sup>th</sup> no pattern (zero significant and zero insignificant) were identified. Most of the patterns obtained in both strains were different<sup>7</sup>, except some e.g. Hepatitis core protein, putative zinc finger, Hepatitis core antigen etc. and these patterns were present in only in the 1<sup>st</sup> 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<sup>8</sup>.



	Total No. of Patterns	Name	Pattern
Hepatitis E	3		
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
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 (	;		
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 3: Representing details of the patterns found in different hypothetical proteins.



	Total No.	Name of Matched Patterns in Pfam-A			
of Patterns		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) DNA polymerase (viral) C-terminal domain	NO PATTERN FOUND		
Protein 3	2	Trans-activation protein X	F-box associated		
Protein 4	2	Hepatitis core protein, putative zinc finger Hepatitis core antigen	NO PATTERN FOUND		
Protein 5	0	NO PATTERN FOUND	NO PATTERN FOUND		
(B) Hepatitis (		•	•		
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		

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

# 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, positionspecific 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.
- 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.

\*\*\*\*\*\*\*\*

