HCDP: HEPATITIS C DATA BANK OF PAKISTAN

Muhammad Shahzad[1], Arif Iqbal Umar[1], Syed Hamad Shirazi[1], Muhammad Tariq Pervez[2*], Zakir Khan[1], Waqas Yousaf[1]

ABSTRACT

Hepatitis C virus (HCV) is a blood born, positive, single stranded RNA strand and circular in shape. The hepatitis C virus is substantial threat to the public health and its frequency is increasing rapidly all over the world. Approximately 5 online databases are available related to Hepatitis C virus. All these databases mainly concerned their national level/specific demographic region strains for analysis. No HCV database available to find the HCV prevalence and distribution in Pakistan. We proposed a Hepatitis C virus database for Pakistani community, Hepatitis C Data Bank of Pakistan (HCDP). HCDP will be the first database that will holds HCV sequences obtained from Pakistani strains and HCV research publication by Pakistani researchers. We proposed a Hepatitis C Data Bank of Pakistan (HCDP) with an online interface. In addition to provision of annotated HCV sequences of Pakistani strains, HCDP allows the user to submit HCV sequences, find out N/O-linked glycosylation, Methylation, Ser/Tyr/Thr-phosphorylation, Methylation and ubiquitination sites in the protein sequences, motif/signature sub-sequences pattern, visual appearance of protein/nucleotide sequences for analysis of different sites, visual representation of multiple sequence alignments using colour code along with motif finding/conserved region in the sequence and analysing of graphical structure of phylogenetic tree. With the help of Format converter/Fasta generator tool user can convert sequence with formats of PhyLip, NEXUS, MSF, CLUSTAL and PIR into standard FASTA. It is also observed that genotype 3a (76%) is more prevalent followed by genotype 3 (13%). Geographic distribution reveals that rate of occurrence of HCV in Sindh and KPK is high with respect to other provinces. An annotated database of HCV genome sequences allows the researcher to investigate the structural and genetic variability of the sequences efficiently and effectively. HCDP is a specialized database that mainly focuses on the HCV strain from Pakistani community. It helps virologists in drug designing and vaccine development.

KEY WORDS: HCDP (Hepatitis C Data Bank of Pakistan), PTM (Post Translational Modification), Genotype, Hepatitis C virus (HCV).

INTRODUCTION

Hepatitis C virus has seven phylogenetically different genotypes with the division of related subtypes (Minosse, et al. 2016). HCV belongs to the genus *Hepacivirus* and family *Flaviviridae* with positive single strand of RNA (Moradpour, et al. 2007). Genome of HCV is ~9600nt in length with single, long open reading frame (ORF) flanked by 5/ to 3/ non-translated regions (Combet, et al. 2006). The diversity of HCV whole genome sequence is 30% over genotype and in-between 15-30% for sub-types (Smith, et al. 2014). This virus has high level of heterogeneity. The diversity of genotype is also variable worldwide. Genotype 1-3 are distributed globally, while 4 and 5 are restricted to Middle East and Africa and South East Asian Countries are predominantly affected by genotype 6 (Hajarizadeh, et al. 2013, McOmish, et al. 1994). Approximately 3-4 million peoples annually are affected by HCV (Berenguer1, et al. 2001). Four types of vaccines are available in clinical trials, naming DNA vaccine, vector vaccine, recombinant protein vaccine and peptide vaccine (Chiang, Yu and Bor-Luen. 2010). Previously developed HCV databases (Combet, et al. 2006), (Carla., et al. 2004), (Kwofie, et al. 2011), (Combet, et al. 2004), and (mmizokami. 2005) mainly focus on their national level strains and entries from DDBJ/EMBL and GenBank for HCV genome analysis. The European HCV database focus on the computer annotated set of sequences and molecular model of HCV protein (Combet, et al. 2006). Some of the tools are no longer functional for this database and last update release was on January 2011 (Floden, et al. 2016). The Hepatitis C database of Las Alamos Laboratory (Carla., et al. 2004) gives access to the immunological epitopes and manually annotated sequences. While Japanese HCV database (mmizokami. 2005) gives access to the phylogenetic

1 Department of Information Technology, Hazara University Mansehra, Pakistan,
2 Department of Bioinformatics and Computational Biology, Virtual University of Pakistan
**Corresponding author: m.tariq@vu.edu.pk*

relationship and genomic mapping of sequences. We proposed a Hepatitis C virus database for Pakistani community, Hepatitis C Data Bank of Pakistan (HCDP). So HCDP will be the first HCV database that will holds HCV sequences from Pakistani strains. Approximately 5 online databases are available about Hepatitis C virus. All databases concerned their national level strains for analysis and also take data from GenBank or PubMed. No one contain HCV sequences from Pakistan except Las Alamos HCV database. The key contribution of this research work are as follows:

We have collected total of 2056 HCV sequences from different regions of Pakistan with respect to prevalence of HCV. It helps the researchers for sequence variation analysis and SNPs analysis.

An interface that helps the user to upload the HCV genome sequence via sequence submission interface. None of the previously developed HCV databases provide information about Post Translational Modification (PTM) and Glycosylation of HCV virus but HCDP provide this information to their users.

HCDP also contain graphs and charts about distribution/prevalence of HCV virus across the Pakistani region.

This online repository provides tools to perform analysis on sequences like, Phylogenetic Tree construction, finding the post translational modification sites like Glycosylation, Methylation, Ubiqutination etc.

In the Protein sequences, Motif/Signature tool help to find any specific pattern define by the user from query sequence, format converter tool convert the files with PhyLip, NEXUS, MSF, CLUSTAL and PIR format into FASTA format.

It is very useful for researchers that are working in Hepatitis C research organizations/industries like biology, life sciences, biotechnology, computational biology, research institutions and also particularly students that work on the HCV. It also helps to understand the molecular mechanism of HCV growth, phylogenetic relationship with other protein sequences. This leads to find out the effective and economical vaccines and medication for Hepatitis patients. The algorithm of format converter tool is on the basis of IVMseq Converter which is the modified version of IVisTMSA (Pervez, et al. 2015). All these tools and information about genotypes/sub-genotypes of HCV sequences provide to the researcher through web base user friendly interface.

BIOLOGICAL IMPORTANCE OF TOOLS

Glycosylation is most complex Post Translational modification in the protein molecule (Wong and Chi-Huey. 2005). More than 50% of the human proteins are glycosylated but this modification is not found in the bacteria. Glycosylation is very essential process for proper functioning of the proteins. It is an enzymatic process that attaches glycan to protein molecule (Glycosylation 2019). Glycosylation is very necessary for correct folding of protein molecule (Varki and Ajit 2009). Glycosylation is also essential for stability of protein structure through the linkage of oligosaccharide at amide nitrogen of certain asparagine. The extremely solvable glycans may have a direct physicochemical stabilization effect while N-linked glycan mediate a critical quality control check point in glycoprotein folding in the endoplasmic reticulum (Dalziel, et al. 2014). There are two types of glycosylation namely N-linked and O-linked glycosylation. N-linked glycosylation is the most prominent modification in the protein molecule. During N-linked glycosylation, oligosaccharides is attach to the nitrogen atom (Consortium 2019). Usually it is attached with the N4 of asparagine residue. It is very essential for the folding of many eukaryotic glycoproteins and for cell-cell and cell-extracellular matrix attachment. The N-linked glycosylation process occurs in eukaryotes of endoplasmic reticulum and extensively in archaea, but infrequently in bacteria. O-linked glycosylation is a type of glycosylation that occurs in Golgi apparatus of eukaryotes. It is also occurs in archaea and bacteria (Glycosylation 2019). O-linked glycosylation refers to the attachment of glycans to serine and threonine. O-linked glycans play important roles in protein localization/trafficking, protein solubility, antigenicity and cell-cell interactions (Consortium 2019, Steentoft, et al. 2013). The consensus sequence for N-linked glycosylation is Asn-Xxx-Ser/Thr (Francois, et al. 2005, Kornfeld and Kornfeld 1985, Kasturi, et al. 1995), where Xxx is any amino acid except proline (Elliott, et al. 2004). Some possible motifs for O-linked glycosylation are [T-A-P-P], [T-V-X-P]. [S/T-P-X-P], [T-S-A-P] (Christlet and Veluraja 2001). Protein Methylation is a biological process that

play key role in many biological processes during post translational modification (Yinan, et al. 2015). It is the addition of methyl group with arginine or glysine amino acids residue in protein sequence. Methylation commonly occurs on carboxyl group of glutamate, leucine and isoprenylated cysteine or on the side-chain nitrogen atoms of lysine, arginine, and histidine residues. It occur on nitrogen atom in N-terminal and irreversible (Heidi, Robert and XiaodongCheng. 2006). Methylated rich region of protein molecule usually referred as "GAR motifs" (McBride and Silver 2001). It also play an important role in regulating cellular functions in an epigenetic manner. Studies reveal that lysine and arginine-specific methylation of histones may collaborate with other types of post-translational histone modification to regulate chromatin structure and gene transcription (Stallcup. 2001). Glysine methylation of histones is linked with transcriptionally active nuclei, controls other types of histone modifications, and is necessary for proper mitotic cell divisions. Identification of methylation sites is a very basic step to find out methylation in the protein molecules. Methylation occurs at either arginine (Arg) or glysine (gly) residues. The Arg-methylation consensus motifs are RGG/RXG/RGX (Wooderchak, et al. 2008) or GXXR (Kenneth, et al. 2005) or RGR/GRG/RG (Jaerang, et al. 2001). It is noticed that QRRGRTGRG motif in arginine rich region of NS3 helicase domain of HCV are more conserved. Here G & R are arginine and glycine respectively while XX mean any amino acids.

Phosphorylation is another type of PTM occurs in protein molecules and also observed in HCV core protein (Jana, et al. 2013). It plays key role for many intra-cellular processes. It is the key molecular mechanism through which protein function is regulated in response to extracellular stimuli both inside and outside the nervous system. Almost all types of extracellular signals, including hormones, light, neurotransmitters, neurotrophic factors and cytokines, produce most of their diverse physiological effects by regulating phosphorylation of specific phosphoproteins in their target cells. Phosphorylation type of PTM in protein molecule activates approximately half of the enzymes for regulating their function (Oliveira and Sauer 2012). The occurrence of phosphorylation type modification in protein molecule is more as compared to the other PTM types. It is estimated that 156,000, 40,000 and 230,000 phosphorylation sites occurs in mouse, yeast and human respectively (Vlastaridis, et al. 2017). Most of the enzymes and receptors are switched "on" or "off" by phosphorylation and de-phosphorylation. It also plays an important regulatory role in p53 tumour suppressor protein. P53 contain 18 different phosphorylation sites for regulatory functions (Ashcroft, et al. 1999). Phosphorylation of core protein by protein kinase A at Ser^{53} and Ser^{116} and by protein kinase C at Ser^{53} and Ser^{99} was reported in both vitro and HepG2 cells. It is also involve in modulating nuclear localization of core protein. Protein phosphorylation occurs at three amino acids named Serine (S), threonine (T) or tyrosine (Y) (Nikolaj, et al. 1999). A consensus motif **[N-P-X-Y]** is a Tyrosine phosphorylation motif. Here X is any amino acid. The consensus sequence for Serine (S) phosphorylation sites is [Ser/Thr-X-X-Glu] or [S/T-X-X-E] (Jana, et al. 2013, Franck, et al. 2005). It is highly conserved in all genotypes of HCV. While threonine phosphorylation motif site is (T-X-X-V[V/A] T-X-X-Y-R[A/S] P-E) (Nikolaj, et al. 1999). The consensus sequence of ubiquitination residues found in P53 protein is [S/T-X-X-X-L-L-G] (Banks, et al. 1998).

CONSTRUCTION AND CONTENT OF DATABASE

The main purpose of HCDP database development is to provide information about the HCV variability in the Pakistan. The architecture of the HCDP designed to be proficient and user-friendly to exploit utilization of the data and their application by users. We provide annotated sequences of Hepatitis c virus that are taken from different geographical areas of Pakistani community. These sequences enable the researchers, biologists, biotechnologists and bioinformatists to extract information about the HCV protein structure on the basis of different regions.

At present, total of 2056 sequences are collected. Out of which 1056 HCV annotated sequences are stored in the database that is collected from different areas of Pakistan. Most of the HCV sequence entries were collected from NCBI and Los Alamos hepatitis C sequence database (Carla, et al. 2004). 28 Hepatitis C virus partial genome, complete 5'UTR, complete CDS (Contains: C, E1, E2, P7, NS2, NS3, NS4A, NS4B, NS5A, NS5B), partial 3'UTR are also the part of HCDP repository. Fig. 1 shows the distribution of HCV on the basis of genotype prevalence in Pakistan. It also described the

HCV prevalence rate in Pakistan with respect to the year and geographic distribution respectively.

First line chart shows the HCV distribution on the basis of genotype/sub-type. Genotype 3a is more prevalent with more that 79% rate of occurrence followed by genotype 3 with 14%. Column chart explain the year wise distribution of HCV while bar chart shows that KPK and Sindh regions are more affected by HCV. Information from the database can be accessed through a versatile user friendly search interface. Sequences organized in such a way, so that user can find desired sequence on the basis of sampling year, sequence type or genotype/subtype. Basic search form that permits the

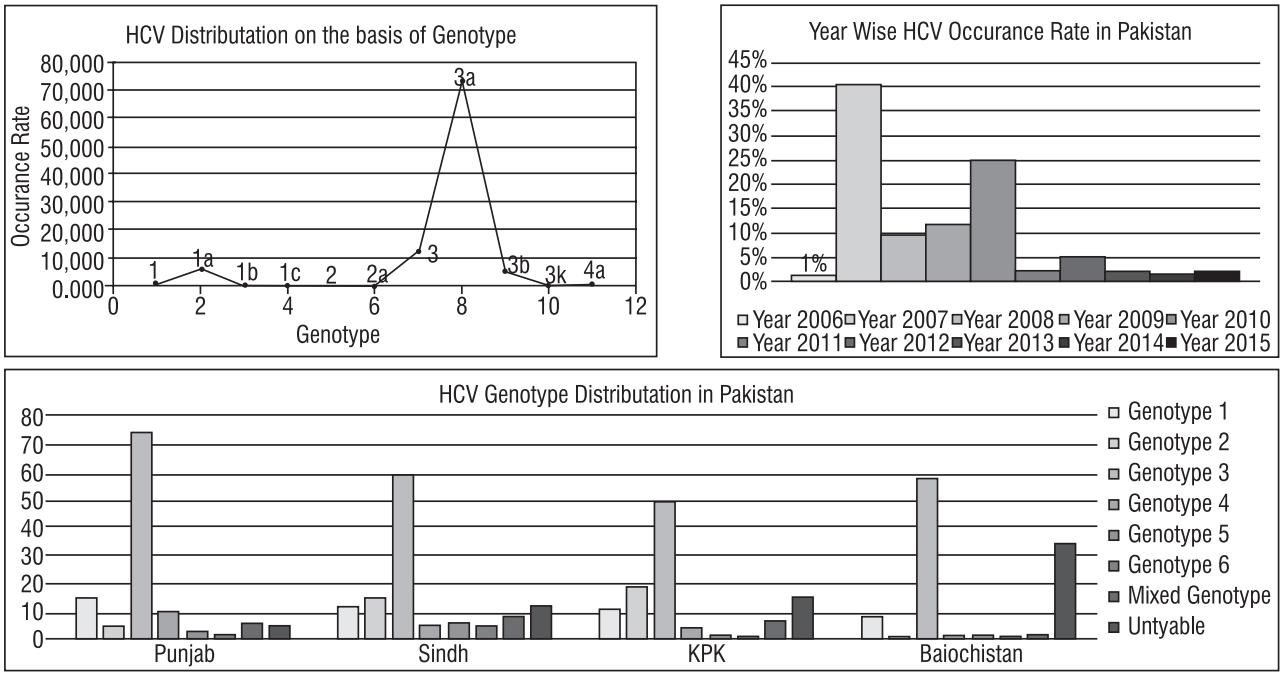

Fig. 1: Distribution of HCV in Pakistan

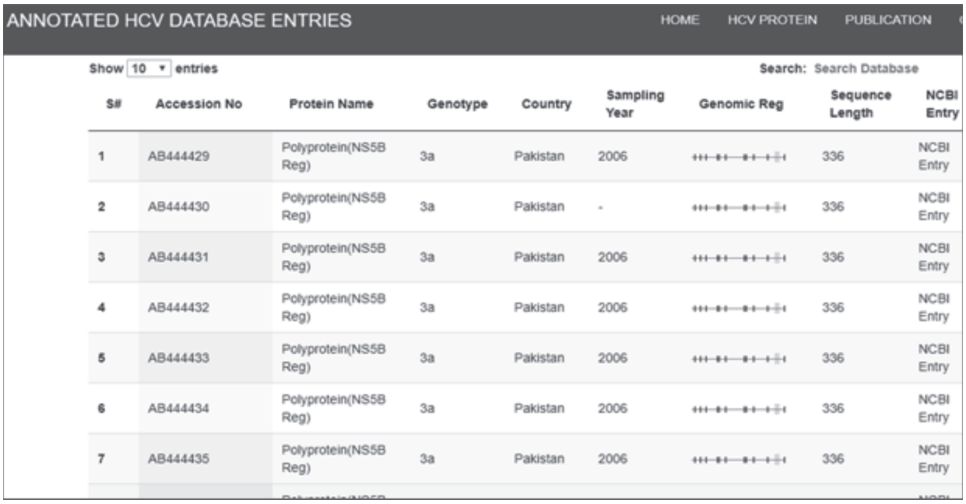

Fig. 2: Search interface of HCDP

user to search HCDP by using different parameters are also available Fig. 2. Web interface of HCDP initially divided into two main categories: 1) Dynamic Resources (Table 1) and 2) Static Resources/other features (Fig. 3).

User can search the database by using this page. Multiple options are available for searching database, either by entering the accession No of the sequences, or on the basis of sampling year in which a particular sequence obtain from the host, by entering the protein name or by selecting genotype. All the tools that are available on this web interface are listed in Table 1. These tools help the user for analysis on PTM sites in protein sequence, Motif/signature pattern finding and other important functionalities.

Dynamic Resources/Tools on HCDP

Table 1: List of Dynamic resources/tools available at HCDP

Tools type	Tools List
General Purposes tool	Format Converter Tool
	Sequence Analyzer
	Signature/Motif Sequence
	Multiple Sequence Alignment
	Phylogenetic Tree Viewer
Tools For Finding Post Translational Modification	N-linked Glycosylation
	O-linked Glycosylation
	Methylation Sites
	Ubiquitination Sites
	Serine (Ser) Phosphorylation
	Tyrosine (Tyr) Phosphorylation
	Threonine (Thr) Phosphorylation

On HCDP, a query system allows the users to conduct different analysis on protein/nucleotide sequences using different tools. Tools available as dynamic resources on HCDP are listed in Table 1. Fig. 1 illustrates the HCV distribution with respect to genotype, year wise and geographical location based. It is observed that genotype 3a is most prevalent in the Pakistan followed by 2a genotype. Geographic distribution chart showed that HCV virus in Sindh and KPK are more prevalent as compare to the other two provinces. The search interface of HCDP is show in Fig. 2. Multiple options available for searching database. First user selects the type of sequence i.e. Protein or Nucleotide. After that user can search the database either by entering the accession No of the sequences, or on the basis of sampling year in which a particular sequence obtain from the host or by selecting genotype. When user click on the search button, the results related to the query will display on the next page.

All links related to HCV virus distribution, HCV partial genomes, aligned & un-aligned sequences and annotated entries of HCV sequences are listed in Table 2. Fig. 3 explain the control flow of getting static information about HCV. Actual aligned and un-aligned sequences are also shown in Fig. 3. This pathway is used to extract all of the HCV proteins information.

Static part of the website provides information about

Table 2: List of all links regarding HCV protein information

Tools type	Link List
Hepatitis C protein Description Information links	HCV Molecular Model:C,P7,E1,E2,NS2,NS3,NS4A,Ns-4b,NS5A,NS5B
	HCV Genome Structure/Sequence (Aligned):
	HCV Genome Nucleotide/Protein Sequence (Un-Aligned):
	HCV Complete Nucleotide Sequence (Un-Aligned):
	28 different Partial Genome of HCV:
	Genotype Distribution of HCV:
	Year Wise Distribution of HCV:
	HCV Geographic Distribution:
	HCV Annotated Entries At HCDP :
	Nomenclature of HCV Genotype/ Subtype :

HCV Genome structure/Genome sequence, individual proteins/nucleotide of hepatitis c and its Pre-aligned/ Un-aligned sequences. List of all link that contain information regarding HCV are listed in Part A of the figure. Part B shows the genomic structure of all protein related to HCV. Part C contain aligned & un-aligned sequence of core protein. Fig. 4A, B & C explain the procedure of Motif/signature sequence tool. Researchers can find any type of motif/signature in the underlying sequences on

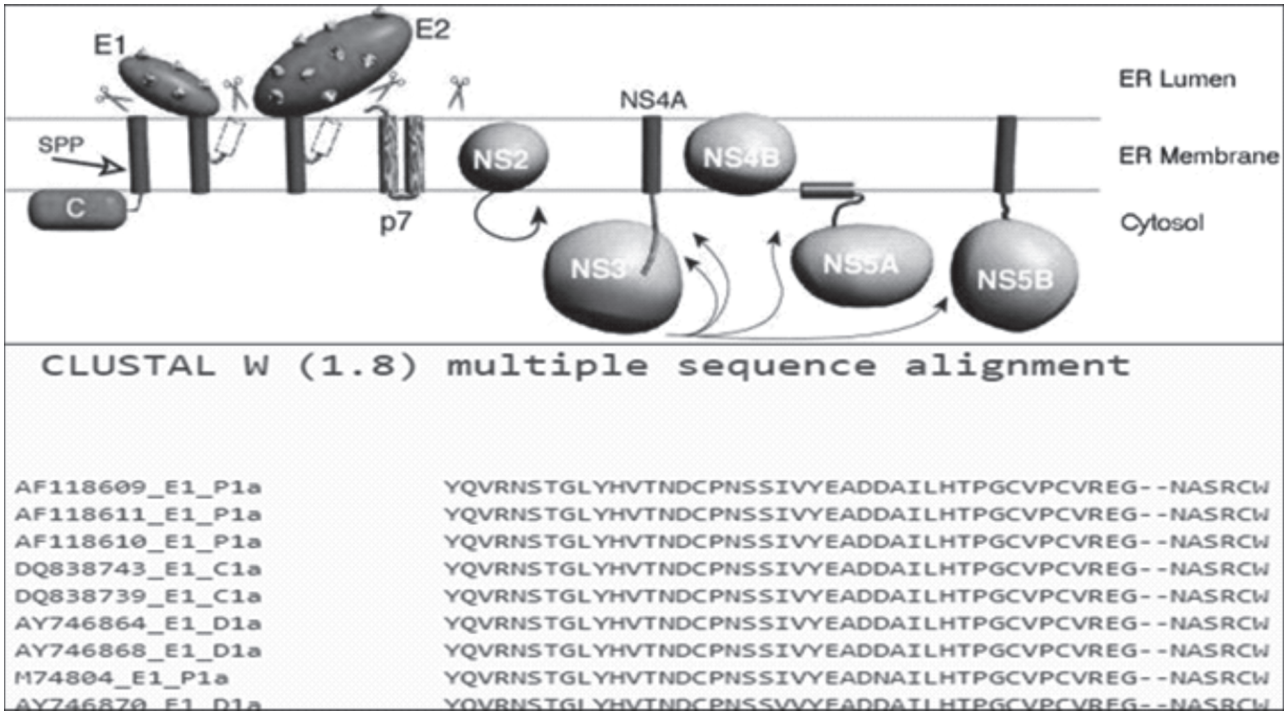

Fig. 3: Information about HCV proteins

their own will. Initially user put the desire sequence in the text field shown in Fig. 4B, followed by pasting Fasta sequence in text area and click "Find Motif/Signature" button. The control directs the user on result page Fig. 4C. Result page shows the entire matching signature present in the query sequence. Sequence analyzer tool allows the user to analyse the sequence with graphical view Fig. 5. On providing Accession No of any gene,

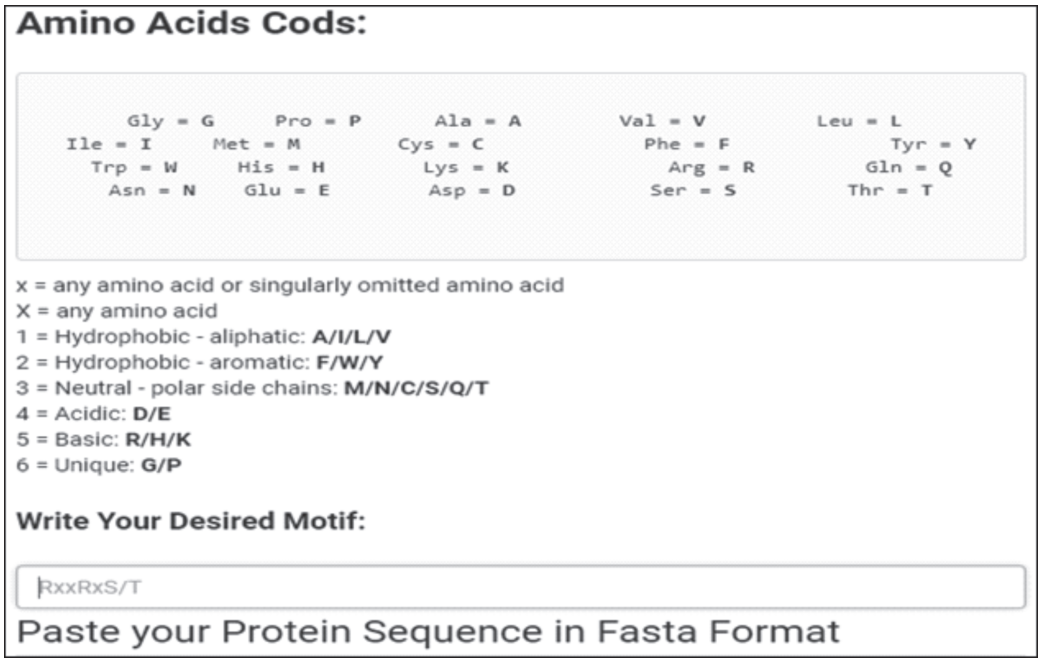

Fig. 4: Motif finding tool

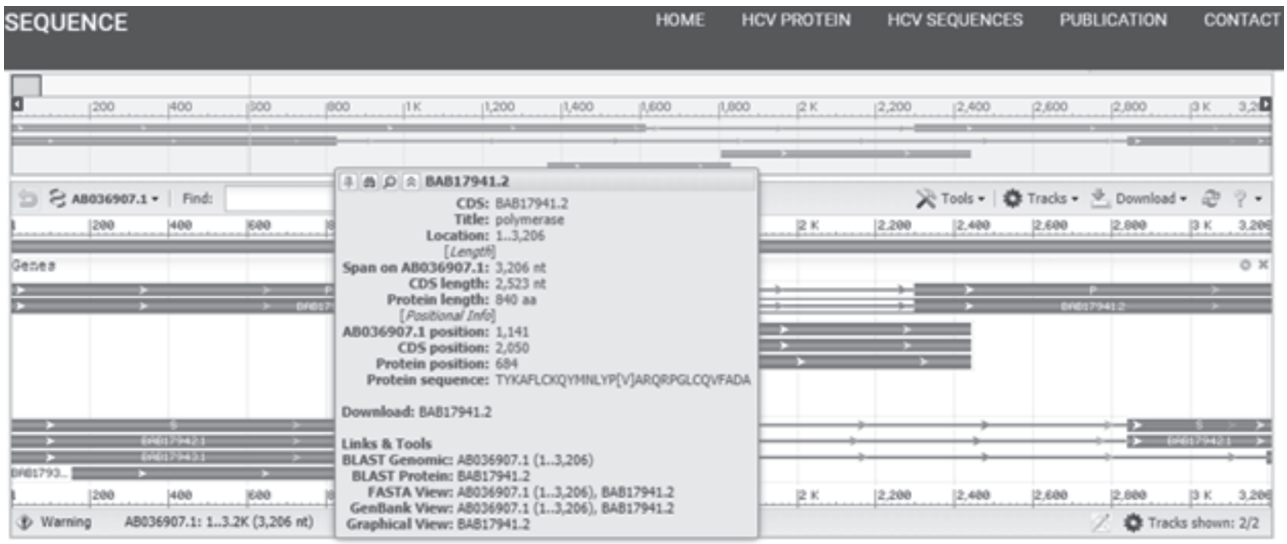

Fig. 5: Sequence Analyzer

sequence analyser tool extract all of the information about that gene from NCBI database and display visual graph on the screen.

Fig. 4 shows the procedure of motif finding. First of all user write the desired motif pattern in text filed then enter query sequence in the text area and click the button to submit query. If system find any matched pattern, it list down in the next page. Multiple sequence alignment visual tool give the MSA view of the sequence with colour coding Fig. 6. User can provide input sequence through URL, from local machine or Drag & Drop method. User can perform different analysis on the Protein/Nucleotide sequence, like motif finding, finding conserved region in the sequences, sorting of alignment using different parameter like gaps, ID, Label, Consensus to top etc. This tool supported Fasta, Clustal, GFF, jalview feature and Newick formats as an input sequence.

On providing Accession No of any gene, sequence analyser tool extract all of the information about that gene from NCBI database and display visual graph on the screen shown at Fig. 5. On pointing any location of gene strand with mouse pointer, information about that location will be display in the drop down menu i.e. CDS, location no, BLAST genomic, BLAST protein no etc. as shown in the Fig. 5.

Fig. 6 shows the MSA viewer tool with colour coding that assists the user to analyse their sequence on the basis of different parameters like showing specific motif with user define colour, arrange the sequence on the

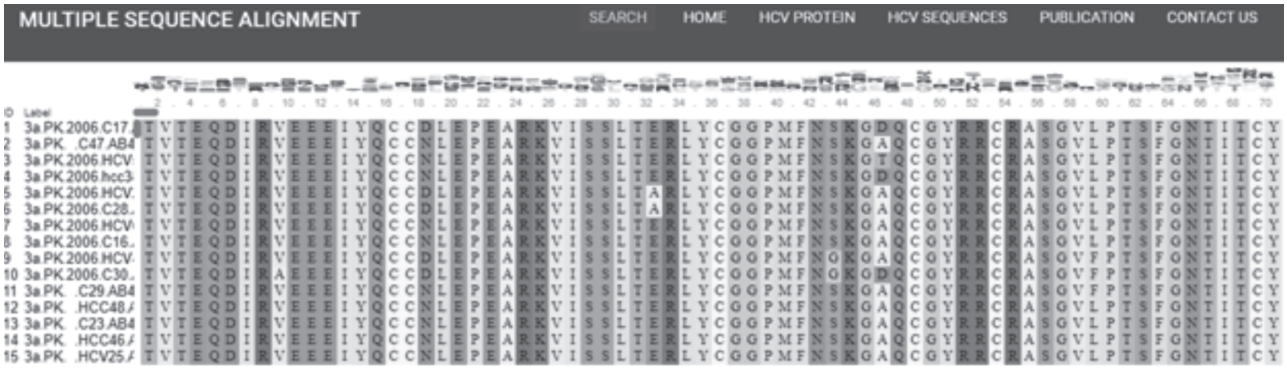

Fig. 6: Sequence viewer tools

basis of gaps, consensus sequence etc. it support fasta, clustal, jalview sequence formats. Colour scheme feature allows the user to change the colour of an alignment on their own will. User can also export their alignment in FASTA format.

PTM Sites Finding Tools

The interfaces for finding Post translational Modification sites in the protein sequences are mostly similar to each other. Fig. 7 explain the process of finding N-linked Glycosylation, Methylation and Ser-phosphorylation sites. On providing sequence with Fasta format in the text area and then clicked on the button. The control directs the user to the next page, where all sites related to Post translational Modification in protein are shown. The results are displayed in 4 different rows. Name of the sequence is in 1st row followed by total length of sequence. In 3rd row total no match records are shown. In the last section PTM sites are individually.

Fig. 7 illustrates the complete procedure of finding N-linked Glycosylation sites. First user enter the query sequence in the text area of the form (query sequence must be in fasta format) and click the button below. The control will direct the user to the next page (right side of Fig. 7). This page list down all the glycosylation sites present in the sequence including total length of query sequence.

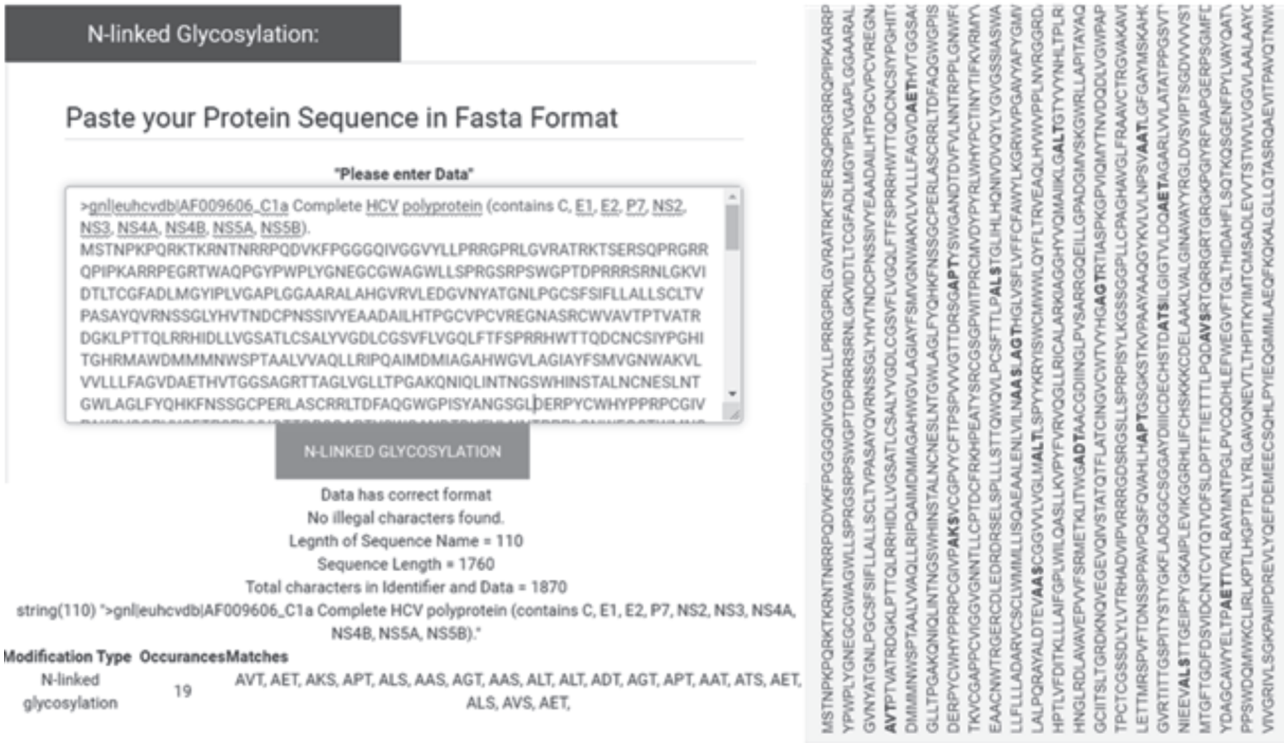

Fig. 7: N-linked Glycosylation sites finding tools

Format Converter Tool

Format converter tool/ Fasta generator tool can convert the sequences formats like CLUSTAL, Phylip, PIR, MSF and Nexus into standard FASTA format. We have converted successfully an alignment of CLUSTAL format consisting 150 sequences into FASTA format in less than 2 second. User can upload alignment in different format. Converter tool identify the sequence format automatically and convert it into FASTA sequence format. Interface and work processing of this tool shown in the Fig. 8A & B.

Fig. 8 shows the Fasta generator tool that convert the sequence file with 5 different sequence formats (CLUSTAL, PIR, MSF, PhyLip and Nexus) into standard FASTA format. User select the file from choose file option as shown above and click on the "CONVERT" button. Control direct the user to the next page, where FASTA converted file link available (Shown in part B of the Fig. 8).

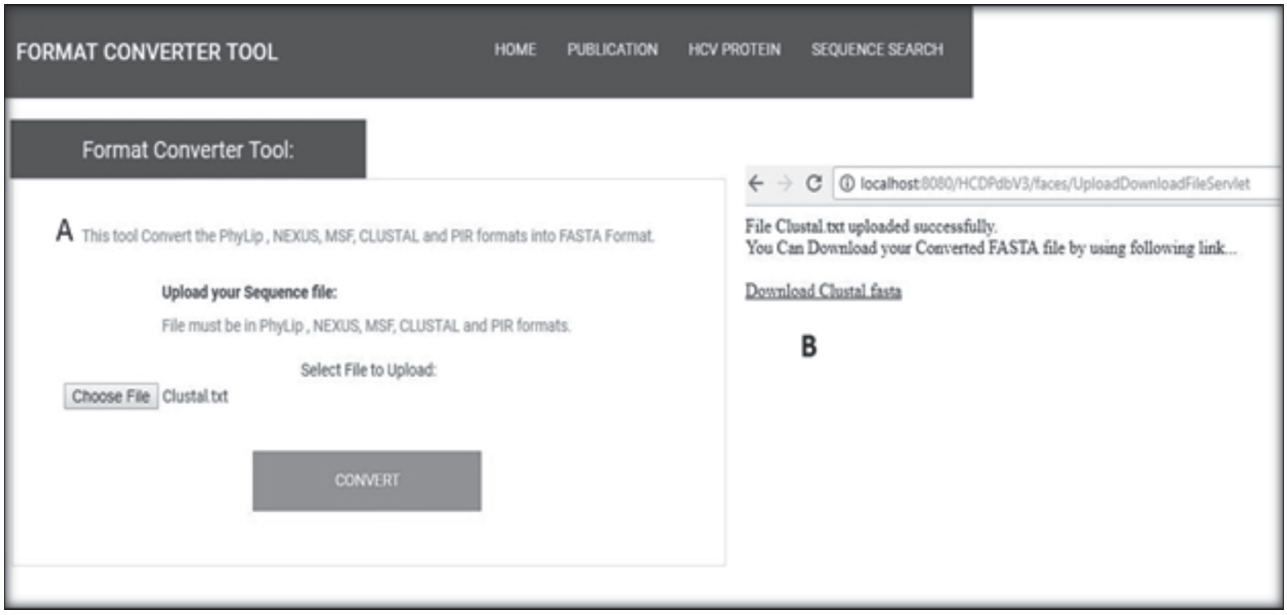

Fig. 8A & B: Format converter tool

OTHER FEATURES

Static part of the website provides information about HCV Genome structure/Genome sequence, individual proteins/nucleotide of hepatitis c and its Pre-aligned/ Un-aligned sequences Fig. 3A, B & C respectively. HCV distribution with respect to Genotype, geographic and sampling year are also included in static part of the website. Complete and annotated protein/ nucleotide sequences of 28 different partial genome of HCV are also available. Complete database entries of Protein/Nucleotides sequences are also accessed through "Nucleotides/Protein Sequences at HCDP/ HCV Annotated Entries at HCDP:" http://hcdp.huairg.com/ dbEntryAnnotated.html.

CONCLUSION

HCDP is a specialized database that mainly focuses on the HCV strain from Pakistani community. This database will also provide tools for finding Post translational Modification of protein sequences (including Glycosylation (O&N-Linked), Phosphorylation, Methylation, Ubiqutination etc), signature/Motif pattern in sequence, MSA colour viewer, sequence analyser, annotated HCV sequence strains, Tree viewer etc. some statistical analysis have also performed to explore the distribution of HCV in Pakistan with respect to genotype, sampling year and geographic distribution. It also provides well organized tools that help the user to perform specific analysis on the sequences like glycosylation, Phosphorylation, Methylation, Ubiquitination, motif

finding, sequence conversion, Multiple Sequence alignment, and construction of phylogenetic trees. The design of the database permits the users to access, analyze and download relevant information through the sophisticated but user-friendly graphical user interface. The HCDP is a resourceful and helping web interface for researcher/ students who working in the field of health sciences, bioinformatics and specifically for those researchers working on hepatitis C.

AVAILABILITY OF DATA AND MATERIAL

HCDP is freely available at www.hcdp.huairg.com. Researchers are encouraged to submit their data regarding Hepatitis C either through the dedicated user interfaces within the HCDP web site or directly to the authors through email. Release of v1.1 of HCDP will contain some additional features and function, like BLAST search interface on HCDP database entries, finding of Acetylation sites in the protein sequences, phylogenetic tree construction tool that can generate the phylogeny tree on inputting multiple sequence alignment sequence.

REFERENCES

1. *Ashcroft M., Kubbutat M. H. and Vousden K. H. (1999), "Regulation of p53 Function and Stability by Phosphorylation", , Molecular and Cellular Biology, vol. 19, no. 3, pp. 1751-1758.*

2. *Berenguer M., Lopez-Labrador F. X. and Wright T. L. (2001), "Hepatitis C and liver transplantation", Journal of Hepatology, vol. 35, no. 5, pp. 666–678.*

3. *Blom, N., Gammeltoft S. and Brunak S. (1999), "Sequence and Structure-based Prediction of Eukaryotic Protein Phosphorylation Sites", Journal of Molecular Biology, vol. 294, no. 5, pp. 12.*

4. *Christlet T. H. T. and Veluraja K. (2001), "Database Analysis of O-Glycosylation Sites in Proteins", Biophysical Journal, vol. 80, no. 2, pp. 952–960.*

5. *Chun I. Y. and Chiang Bor-Luen. (2010), "A New Insight into Hepatitis C Vaccine Development", BioMed Research International, vol. 10.*

6. *Combet C. et al. (2007), "euHCVdb: The European hepatitis C virus database", Nucleic Acids Research.*

7. *Combet C., Penin F., Geourjon C. and Deléage G. (2004), "HCVDB: hepatitis C virus sequences database", Applied Bioinformatics, vol. 3, no. 4, pp. 237-240.*

8. *Dalziel M., Crispin M., Scanlan C. N., Zitzmann N. and Dwek R. A. (2014), "Emerging principles for the therapeutic exploitation of glycosylation", Science, vol. 343 no. 6166.*

9. *Elliott S., Chang D., Delorme E., Eris T. and Lorenzini T. (2004), "Structural Requirements for Additional N-Linked Carbohydrate on Recombinant Human Erythropoietin", Journal of Biological Chemistry, pp. 16854-16862.*

10. *Floden, E. W., Khawaja A., Vopálensky V. and Pospíšek M. (2016), "HCVIVdb: The hepatitis-C IRES variation database", BMC Microbiology, vol. 16, no. 1.*

11. *Franck, N., Le Seyec J., Guguen-Guillouzo C. and Erdtmann L. (2005), "Hepatitis C Virus NS2 Protein Is Phosphorylated by the Protein Kinase CK2 and Targeted for Degradation to the Proteasome", Journal of Virology, vol. 79, no. 5.*

12. *Goffard A., Callens N., Bartosch B., Wychowski C., Cosset F. L., Montpellier C. and Dubuisson J. (2005), "Role of N-linked glycans in the functions of hepatitis C virus envelope glycoproteins", Journal of Virology, vol. 79, no. 13.*

13. *Hajarizadeh B., Grebely J. and Dore G. J. (2013), "Epidemiology and natural history of HCV infection", Gastroenterol Hepatol, vol. 10, no. 9.*

14. *Hundt J., Li Z. and Liu Q. (2013), "Post-translational modifications of hepatitis C viral proteins and their biological significance", World Journal of Gastroenterology, vol. 19, no. 47, pp. 8929-8939.*

15. *Jaerang Rho, Seeyoung Choi, Young Rim Seong, Joonho Choi and Im Dong-Soo (2001), "The Arginine-1493 Residue in QRRGRTGR1493G Motif IV of the Hepatitis C Virus NS3 Helicase Domain Is*

Essential for NS3 Protein Methylation by the Protein Arginine Methyltransferase 1", Journal of Virology, vol. 75, no. 17, pp. 8031–8044.

16. *Kasturi L., Eshleman J. R., Wunner W. H. and Shakin-Eshleman S. H. (1995), "The hydroxy amino acid in an Asn-X-Ser/Thr sequon can influence N-linked core glycosylation efficiency and the level of expression of a cell surface glycoprotein", The Journal of Biological Chemistry, vol. 270, no. 24, pp. 14756-14761.*

17. *Daily K. M., Radivojac P. and Dunker A. K. (2005), "Intrinsic Disorder and Prote in Modifications: Building an SVM Predictor for Methylation", IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 15 Nov. 2005, La Jolla, CA, USA.*

18. *Kornfeld, R., and Kornfeld S. (1985), "Assembly of asparagine-linked oligosaccharides", Annual Review of Biochemistry, pp. 631-664.*

19. *Kühne C., and Banks L. (1998) "E3-Ubiquitin Ligase/E6-AP Links Multicopy Maintenance Protein 7 to the Ubiquitination Pathway by a Novel Motif, the L2G Box*", The Journal of Biological Chemistry, pp. 34302-34309.*

20. *Kuiken C., Yusim K., Boykin L. and Richardson R. (2005), "The Los Alamos hepatitis C sequence database", Bionformatics, vol. 21, no. 3, pp. 379–384.*

21. *Kwofie S. K., Schaefer U., Sundararajan V. S., Bajic V. B., and Christoffels A. (2011), "HCVpro: Hepatitis C virus protein interaction database", Infection, Genetics and Evolution, vol. 11, no. 8, no. 1971-1977.*

22. *McBride A. E. and Silver P. A. (2001) "State of the arg: protein methylation at arginine comes of age", Cell Press, vol. 106, no. 1, pp. 5-8.*

23. *McOmish F., Yap P. L., Dow B. C., Follett E. A., Seed C., Keller A. J., Cobain T. J., Krusius T., Kolho E. and Naukkarinen R. (1994), "Geographical distribution of hepatitis C virus genotypes in blood donors: an international collaborative survey", Journal of Clinical Microbiology, vol. 32, no. 4, pp. 884-892.*

24. *Minosse C., Giombini E., Bartolini B., Capobianchi M. R. and Garbuglia A. R. (2016), "Ultra-Deep Sequencing Characterization of HCV Samples with Equivocal Typing Results Determined with a Commercial Assay", International Journal of Molecular Science, vol. 17, no. 10.*

25. *Moradpour D., Penin F. and Rice C. M. (2007), "Replication of hepatitis C virus", Nature Reviews Microbiology, vol. 5, no. 6, pp. 453-463.*

26. *National Institute of Genetics, Hepatitis Virus DataBase Server, Accessed: 11 Dec. 2019. Online: http://s2as02.genes.nig.ac.jp/index.html.*

27. *Oliveira A. P. and Sauer U. (2012), "The importance of post-translational modifications in regulating Saccharomyces cerevisiae metabolism", FEMS Yeast Research, vol. 12, no. 2.*

28. *Pervez M. T. et al. (2015), "IVisTMSA: Interactive Visual Tools for Multiple Sequence Alignments", Evolutionary Bioinformatics Online, pp. 35-42.*

29. *Schubert H. L., Blumenthal R. M. and Xiaodong Cheng. (2006), "1 Protein Methyltransferases: Their Distribution Among the Five Structural Classes of AdoMet-Dependent Methyltransferases", The Enzymes, vol. 24, pp. 3-28.*

30. *Shi Y., Guo Y., Hu Y. and Li M. (2015), "Position-specific prediction of methylation sites from sequence conservation based on information theory", Scientific Reports.*

31. *Smith D. B., Bukh J., Kuiken C., Muerhoff A. S., Rice C. M., Stapleton J. T. and Simmonds P. (2014), "Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: Updated criteria and genotype assignment web resource", Hepatology, vol. 59, no. 1.*

32. *Stallcup M. R. (2001), "Role of protein methylation in chromatin remodeling and transcriptional regulation", Oncogene, pp. 3014–3020.*

33. Steentoft C. et al. (2013), "Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology", The EMBO Journal, pp. 1478-1488.

34. UniProt, Glycosylation. UniProt EMBL-EBI, Accessed: 13 Dec. 2019, Online: http://www.uniprot.org/help/carbohyd.

35. Varki A. (2009), "Essentials of Glycobiology" 2nd Ed., Cold Spring Harbor Laboratory Press, USA.

36. Vlastaridis P., Kyriakidou P., Chaliotis A., Van de Peer Y., Oliver S. G. and Amoutzias G. D. (2017), "Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes", Gigascience, vol. 6, no. 2.

37. Wong Chi-Huey (2005), "Protein Glycosylation: New Challenges and Opportunities", Journal of Organic Chemistry, pp. 4219-4225.

38. Wooderchak W. L., Tianzhu Zang, Zhaohui Sunny Zhou, Acuña M., Tahara S. M. and Hevel J. M. (2008), "Substrate profiling of PRMT1 reveals amino acid sequences that extend beyond the "RGG" paradigm", Biochemistry, pp. 9456-9466.