Appl Bioinformatics 2004; 3 (4): 237-240
1175-5636/04/0004-0237/$31.00/0

HCVDB

Hepatitis C Virus Sequences Database

Christophe Combet, *François Penin*, *Christophe Geourjon* and *Gilbert Deléage*

Institut de Biologie et Chimie des Protéines, Lyon, France

Abstract

Abstract: To date, more than 30 000 hepatitis C virus (HCV) sequences have been deposited in the generalist databases DNA Data Bank of Japan (DDBJ), EMBL Nucleotide Sequence Database (EMBL) and GenBank®. The main difficulties with HCV sequences in these databases are their retrieval, annotation and analyses. To help HCV researchers face the increasing needs of HCV sequence analyses, we developed a specialised database of computer-annotated HCV sequences, called HCVDB. HCVDB is re-built every month from an up-to-date EMBL database by an automated process. HCVDB provides key data about the HCV sequences (e.g. genotype, genomic region, protein names and functions, known 3-dimensional structures) and ensures consistency of the annotations, which enables reliable keyword queries. The database is highly integrated with sequence and structure analysis tools and the SRS (LION bioscience) keywords query system. Thus, any user can extract subsets of sequences matching particular criteria or enter their own sequences and analyse them with various bioinformatics programs available on the same server.

Availability: HCVDB is available from http://hepatitis.ibcp.fr

Contact: Gilbert Deléage (g.deleage@ibcp.fr)

Hepatitis C virus (HCV) infection is a major cause of chronic hepatitis, liver cirrhosis and hepatocellular carcinoma worldwide. The HCV genome is approximately 9600 nucleotides in length and carries a single, long open reading frame (ORF) flanked by 5′ and 3′ untranslated regions. The ORF encodes a polyprotein of about 3000 amino acids that is processed by cellular and viral proteases to yield at least ten mature proteins: C, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A and NS5B.[1]

The sequence dissimilarity among HCV genomes leads to the definition of a large number of genotypes.[2] The genotype nomenclature follows the format HCV-1a, HCV-1b, HCV-2a etc., with 1−11 indicating the type and a, b etc. indicating the subtype. The various genotypes are distributed into six clades: 1, 2, 3 (including type 10), 4, 5 and 6 (including types 7, 8, 9 and 11). It is now well established that the genotype is a crucial predictive factor of the response to interferon therapy.[3] Consequently, intensive sequencing and sequence analyses of HCV genomes are currently conducted, and >30 000 sequences that make up 187 complete genomes of various genotypes have been deposited to date into DNA Data Bank of Japan (DDBJ), EMBL Nucleotide Sequence Database (EMBL) and GenBank® databases.

In order to manage such large and growing collections of sequences and to facilitate their analysis, we developed a new HCV sequences database: HCVDB (for a review of virus databases see Kellam and Mar Albà[4]). The HCVDB database contains computer-annotated HCV sequences integrated with analysis tools.

Building the Database

HCVDB is re-built every month by an automated process using the EMBL database that is updated weekly.[5] The annotation process uses a reference database (HCVREF) of 33 manually annotated complete genomes representing 17 well characterised genotypes.[6] The process is divided into three main steps.

First, all the EMBL entries with the OrganiSm (OS) field matching the keyword 'hepatitis C virus' are collected into the HCVEMBL database. Second, each sequence of HCVEMBL is compared against HCVREF using the FASTA3 program.[7] Default parameters are used, except for the penalty for the first residue in a gap (set to −24) and the statistics estimation model (set to 3) to identify the closest reference genome (command options: fasta –b3 –d3 –f–24 –g–4 –z3). The annotations (e.g. protein names, interesting sites) of the closest reference genome are transferred to the nucleotide sequence examined and to its translation. The sequence is then genotyped. The sequence genotyping algorithm uses the full-length nucleotide sequence of three regions of the genome: C, E1 and NS5B. Each of these subsequences serves as a query for a sequence similarity search against HCVREF. If the sequence identity between the reference genome and the query subsequence is higher than a 'type' threshold (table I), the subsequence is annotated with the virus type of the reference genome. If the sequence identity is higher than a 'subtype' threshold, the subsequence is also annotated with the virus subtype of the reference genome; thus, the genotype (type with subtype) is defined.

The sequence identity thresholds (table I) were from data from the phylogenetic study carried out by Tokita et al.[6] on the 11 clusters classification, with an additional security margin of 2% for C and E1 and 1% for NS5B. These thresholds therefore represent the minimum percentage of sequence identity, plus the security margins, for grouping together all sequences in the correct type/subtype and for discriminating between types and subtypes – at best. The HCV clade is then inferred from the type. When two or three subsequences are present in the same sequence, the consistency of the genotype determination results is checked. If the genotypes are not the same for the three subsequences, clade and genotype are set to 0. If the sequence identity is below the 'type' threshold, the 'not available' (n.a.) value is given for the genotype and clade. A leave-one-out test was used to check the robustness of the genotyping algorithm with the reference genomes. For 18 genomes the genotype was correctly predicted, for 2 the type/clade was correctly predicted, 7 were predicted as n.a. (each one is the unique representative of its genotype in the set) and the last 6 were predicted as 0 genotype (for all of them there was inconsistency between the right type and the n.a. value). In summary, the algorithm never attributes a wrong genotype.

For the third step in the annotation process, the nucleotide and polypeptide sequences are split according to the HCV genome structure. This automated annotation process creates four different databases as flat text files named HCVUFN, HCVUFP, HCVUSN and HCVUSP, where U indicates public sequences, F and S indicate the split state of the sequences (F = full length of the sequence as deposited in EMBL and S = split sequences according to genome structure) and N and P indicate the sequence type (nucleotide or protein). The format of HCVDB entries is described in the user manual (in PDF format), available from URL http://hcvpub.ibcp.fr/cgi-bin/hcvpub_automat.pl?page=/HCVPUB/hcvpub_help.html or from the home page by following the links Public→Help→HCVDB help→User manual. The numbers of entries in the different databases for the current release of HCVDB (number 48, 21 November 2004) are HCVEMB, 30816; HCVUFN, 30816; HCVUFP, 26028; HCVUSN, 46464; and HCVUSP, 37302.

Table I. Sequence identity thresholds used for hepatitis C virus type and subtype assignment[6]

Genomic region	Type threshold (%)	Subtype threshold (%)
C	90	95
E1	72	90
NS5B	81	93

Accessing HCVDB

Users can access HCVDB data through HCVSRS and HCVSA modules (figure 1). HCVSRS is a dedicated interface to HCVDB based on the widely used SRS system (LION bioscience). It allows the fast retrieval of sequences using keywords (e.g. sequence accession number or author names) and enables the building of a subset of sequences that match a particular criteria (e.g. a specified protein domain of a specific HCV genotype; see figure 1a to figure 1c). The SRS layout files allow the display of the entries in a more human-readable view (called 'nice view'; see figure 1d) than the text view. A quick tour of SRS is available from the home page by following the links Public→Help→HCVSRS help.

HCVSA is a light version of the NPS@ server[8] developed in our team at IBCP and implemented in Perl programming language. The main feature of HCVSA is the interconnection of 33 analysis methods and 7 biological databases within a simple, user-friendly web interface (table II). Thus, it provides an easy method for the analysis of HCV nucleotide and protein sequences and structures, and avoids the tedious and time-consuming cut-and-paste intermediate operations between different servers. Typically, the user can (i) search for specific homologous sequences; (ii) extract a subset; (iii) perform multiple alignments; (iv) make secondary structure predictions, add them to the multiple alignment and generate a consensus prediction; (v) plot physicochemical profiles (e.g. hydrophobicity, antigenicity, potential membranous regions, predicted solvent accessibility); and (vi) detect functional sites or signatures specific to a protein family. The input data required for HCVSA is one of the following: a single sequence (personal data); a personal sequence databank (built from experimental sequencing or from a keywords query with HCVSRS); or a user-defined pattern using the PROSITE syntax. The network protein sequence analysis (NPSA) link is another key feature of HCVSA and is mainly available in the similarity search result pages. It provides sequence and structure analysis tools for sequences that are similar to the query sequence. It also offers a link to our automatic,

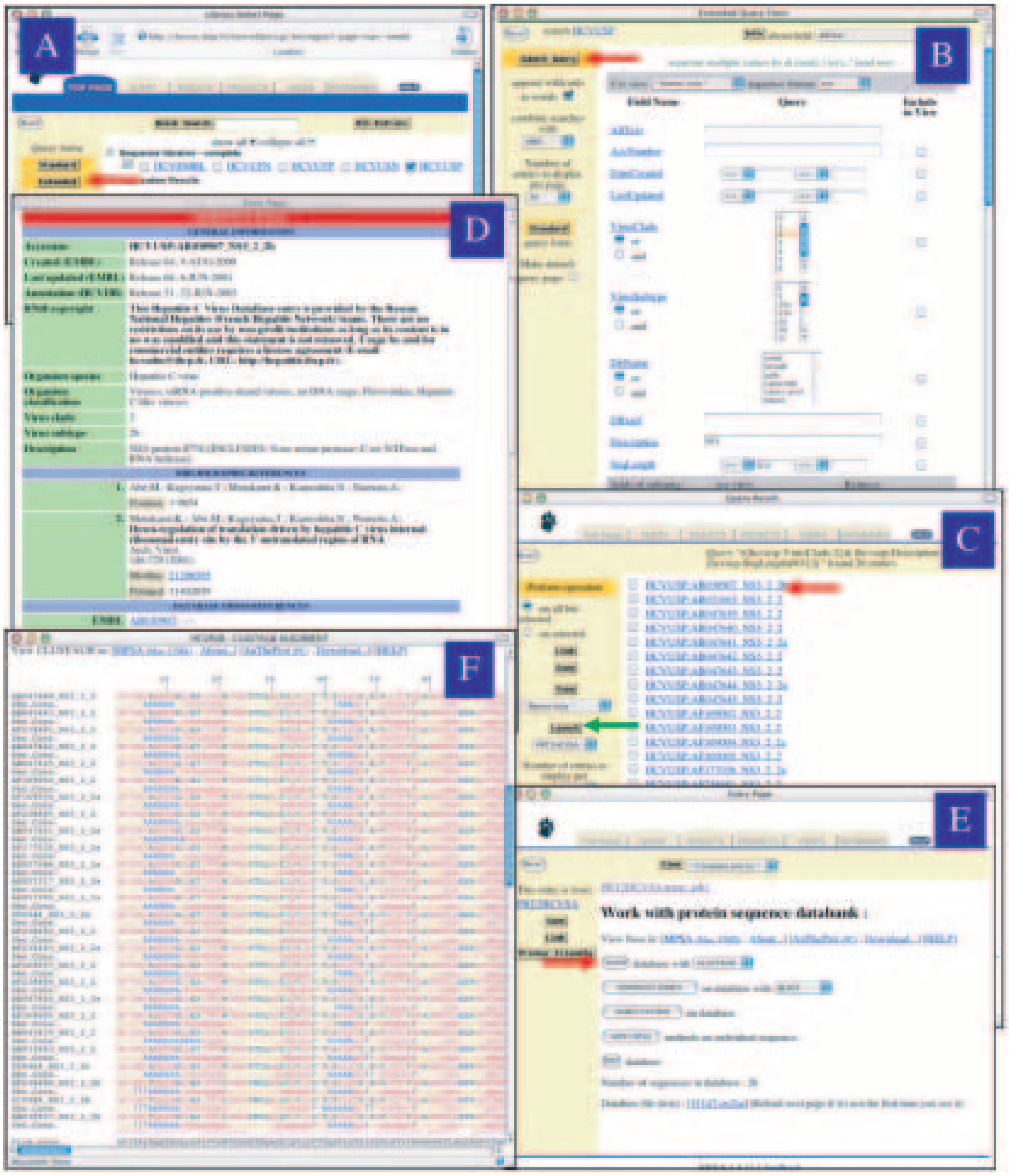

Fig. 1. Example of analysis with HCVDB. (**a**) In HCVSRS, the user selects a database to query (e.g. HCVUSP) and the form to use (e.g. extended, see red arrow). (**b**) The user enters the query criteria (e.g. NS3 sequences with a length of 631 amino acids and belonging to clade 2) and submits the query (red arrow). (**c**) List of entries matching the criteria. The user can access entry data through the entry 'nice view' (**d**) by clicking on the hyperlink in the list (red arrow) and, using the SRS launch button (green arrow), can transfer the sequences to the HCVSA module (**e**). (**f**) ClustalW alignment with predicted secondary structure consensus as viewed in HCVSA.

3-dimensional, protein structure-modelling server: Geno3D.[9] A help page describing HCVSA methods is available from the home page by following the links Public→Help→HCVSA help.

HCVSRS and HCVSA modules are interconnected, i.e. data are exchanged automatically between them on user request. HCVSA offers links to the 'nice view' of the entries in the similarity search result pages. HCVSRS can send sequences to HCVSA for further analyses (e.g. alignment; see figure 1e and figure 1f), thanks to a Perl script call by the SRS launch button.

The full- and split-sequence databases (HCVUFN, HCVUFP, HCVUSN, HCVUSP) were built to facilitate sequence analyses with HCVSA after their retrieval with HCVSRS. For example,

Table II. Bioinformatic analysis methods available in the HCVSA module of HCVDB

Category	Methods
Homology search	BLAST®, PSI-BLAST, FASTA, SSEARCH
Functional sites or signatures detection	PATTINPROT, PROSCAN
Multiple alignment	ClustalW, MultAlin
Protein secondary structure prediction	PHD, GOR (I, II and IV), MLRC, SOPM, SOPMA, HNN, DPM, DSC, SIMPA96, PREDATOR, Consensus
Miscellaneous tools	Composition, ColorSeq, COILS, HTH, PCProf, PHDhtm

starting from the home page, by clicking on the Public button→HCVSRS button→SRS start button, then selecting HCVUSP database and entering in the extended form the criteria 'Description = NS3' and 'SeqLength> = 631', the sequences extracted by HCVSRS (after two consecutive clicks on the Launch button) and sent to HCVSA correspond to only full-length NS3 sequences, i.e. without mixing with partial NS3 sequences or complete genomes deposited in EMBL (figure 1). Consequently, this means that multiple-sequence alignment programs can compute more reliable alignments.

Conclusions

We developed a database of annotated HCV sequences, integrated with a keywords sequence retrieval system and bioinformatics programs, for the analyses of nucleotide and protein sequences. The automatic annotation process in HCVDB allows for consistency of annotations, which is a guarantee of the efficiency of the keywords search system and of the relevance of the sequence and structure analyses. HCVDB entry identifiers are built from the EMBL primary access number to avoid having to give new references for sequences. HCVDB receives about 200 visits per day. By offering the ability to perform complex searches and analyses in the HCV field, HCVDB is a powerful tool for all HCV researchers.

Acknowledgements

This work was supported by Réseau National Hépatites of 'Programme national de recherche fondamentale en microbiologie et maladies infectieuses et parasitaires du Ministère de la Recherche' and the EU contract QLK2-CT-2002-01329. C. Combet received a doctoral fellowship from Agence National de Recherches sur le Sida (ANRS). The authors gratefully acknowledge Christian Bréchot, Jean Dubuisson, Geneviève Inchauspé and Jean-Michel Pawlotsky for stimulating discussions and their support for HCVDB creation. The authors have no conflicts of interest that are directly relevant to the content of this article.

References

1. Lindenbach BD, Rice CM. Flaviviridae: the viruses and their replication. In: Knipe DM, Howley PM, editors. Fields virology. Philadelphia (PA): Lippincott Williams & Wilkins, 2001: 991-1042
2. Robertson B, Myers G, Howard C, et al. Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. International Committee on Virus Taxonomy. Arch Virol 1998; 143: 2493-503
3. Pawlotsky JM. Hepatitis C virus genetic variability: pathogenic and clinical implications. Clin Liver Dis 2003; 7: 45-66
4. Kellam P, Mar Albà M. Virus bioinformatics: databases and recent applications. Appl Bioinformatics 2002; 1: 37-42
5. Stoesser G, Baker W, van den Broek A, et al. The EMBL Nucleotide Sequence Database: major new developments. Nucleic Acids Res 2003; 31: 17-22
6. Tokita H, Okamoto H, Iizuka H, et al. The entire nucleotide sequences of three hepatitis C virus isolates in genetic groups 7–9 and comparison with those in the other eight genetic groups. J Gen Virol 1998; 79: 1847-57
7. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988; 85: 2444-8
8. Combet C, Blanchet C, Geourjon C, et al. NPS@: network protein sequence analysis. Trends Biochem Sci 2000; 25: 147-50
9. Combet C, Jambon M, Deleage G, et al. Geno3D: automatic comparative molecular modelling of protein. Bioinformatics 2002; 18: 213-4

Correspondence and offprints: Professor *Gilbert Deléage*, Institut de Biologie et Chimie des Protéines, UMR 5086, CNRS/UCBL, IFR128 Biosciences Lyon-Gerland, 7 passage du Vercors, 69367 Lyon Cedex 07, France.
E-mail: g.deleage@ibcp.fr