SPECIAL ARTICLE

Hepatitis C Databases, Principles and Utility to Researchers

Carla Kuiken,[1] Masashi Mizokami,[2] Gilbert Deleage,[3] Karina Yusim,[1] Francois Penin[3] Tadasu Shin-I,[2] Céline Charavay,[3] Ning Tao[1] Daniel Crisan,[3] Delphine Grando,[3] Anita Dalwani,[1] Christophe Geourjon,[3] Ashish Agrawal,[1] and Christophe Combet[3]

Part of the effort to develop hepatitis C–specific drugs and vaccines is the study of genetic variability of all publicly available HCV sequences. Three HCV databases are currently available to aid this effort and to provide additional insight into the basic biology, immunology, and evolution of the virus. The Japanese HCV database (http://s2as02.genes.nig.ac.jp) gives access to a genomic mapping of sequences as well as their phylogenetic relationships. The European HCV database (http://euhcvdb.ibcp.fr) offers access to a computer-annotated set of sequences and molecular models of HCV proteins and focuses on protein sequence, structure and function analysis. The HCV database at the Los Alamos National Laboratory in the United States (http://hcv.lanl.gov) provides access to a manually annotated sequence database and a database of immunological epitopes which contains concise descriptions of experimental results. In this paper, we briefly describe each of these databases and their associated websites and tools, and give some examples of their use in furthering HCV research. (HEPATOLOGY 2006;43:1157-1165.)

The hepatitis C virus (HCV) has infected approximately 170 million people worldwide. HCV infection is cleared in about 25% of cases,[1,2] and in the rest results in chronic infection. Chronic HCV infection can lead to cirrhosis and liver cancer, and is the leading cause of liver transplantation in the United States. A recent Canadian study[3] estimated that lifetime HCV-associated mortality is around 1 in 8; a much larger number (an estimated 1 in 4) will develop cirrhosis of the liver. Most likely this number will be higher in less developed countries. With 170 million people infected worldwide, this means 20 million HCV-related deaths in the next few decades.

HCV is a positive-sense RNA virus with a genome of ≈10 kb, which encodes a single polyprotein of ≈3000 amino acids (aa) that is cleaved into three structural proteins (core, Envelope E1 and E2), the p7 protein whose function has not been determined, and six non-structural proteins (NS2, NS3, NS4A, NS4B, NS5A and NS5B). It has been classified as a hepacivirus, in of the Flaviviridae family, which also includes flaviviruses (West Nile, Japanese encephalitis and yellow fever viruses) and pestiviruses (bovin viral diarrhea and hog cholera virus). HCV shares some structural features with these viruses. However, the genetic distance between HCV and other flaviviruses is large enough that HCV cannot be meaningfully aligned to its flavivirus "relatives" over its entire genome[4] (also see http://hcv.lanl.gov/content/hcv-db/GET_ALIGNMENTS/flavi-align.html), and it also shares structural features with the pestivirus family.

HCV is subdivided into six genotypes and about 80 subtypes on the basis of nucleotide sequence identity.[5] In addition to genotypes, HCV exists within its hosts as a pool of genetically distinct but closely related variants referred to as quasispecies.[6] While there is limited knowledge about the immunogenicity of HCV, it is widely expected that both the generation of escape and resistance mutations and the high variability itself will create formidable problems for drug and vaccine design.[7]

This paper discusses three HCV databases available worldwide, in order of seniority: the Japanese HCV map and phylogeny database, the European HCV sequence and molecular models database and the Los Alamos HCV sequence and immunology databases. We will first describe the three databases, highlighting common features and differences. Next, the tools available to extract and

Abbreviations: HCV, hepatitis C virus; HVDB, the Hepatitis Virus Database; HBV, hepatitis B virus; HEV, hepatitis E virus; DDBJ, DNA Data Bank of Japan; euHCVdb, the European hepatitis C virus database.

From the [1]Theoretical Biology and Biophysics group, Los Alamos National Laboratory, Los Alamos, NM; [2]Department of Clinical Molecular Informative Medicine, Nagoya City University Graduate School of Medical Sciences — Kawasumi Mizuho Nagoya, Japan; [3]Institut de Biologie et Chimie des Protéines —UMR5086 CNRS/UCBL - IFR128 BioSciences Lyon-Gerland - 7, Lyon, France.

Received April 13, 2005; accepted February 10, 2006.

HVDB is supported by a Grant-in Aid for the Publication of Scientific Research Results (#168111), Grant-in Aid for Scientific Research, Japan Society for the Promotion of Science(JSPS). The European HCV database is funded by HepCVax Cluster (EU FP5 grant QLK2-2002-01329) and VIRGIL Network of Excellence (EU FP6 grant LSHM-CT-2004-503359). The Los Alamos HCV database is funded by the Division of Microbiology and Infectious Diseases, National Institute of Allergy and Infectious Diseases.

Address reprint requests to: Carla Kuiken, Los Alamos National Laboratory, Group T10, Mail Stop K710, Los Alamos, NM 87545. E-mail: kuiken@lanl.gov.

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI 10.1002/hep.21162

Potential conflict of interest: Nothing to report.

Table 1. Summary of Data Available in the Databases

	Japanese HCV Database	European HCV Database	LANL HCV Database
Source of sequences	DDBJ/Genbank/EMBL	EMBL/DDBJ/Genbank/	Genbank/EMBL/DDBJ
Entry annotation	Sequence mapping aginst reference and DDBJ entry	Monthly updated. Automatic using a set of reference complete genomes and parsing of EMBL entries	Parsed from Genbank entry; manual based on associated publication or communication with authors
Genotype		Deposited in EMBL, provisional, confirmed or computed	Parsed from Genbank entry, or manual based on phylogenetic analysis by the annotators
Clinical data			Patient data as reported by the authors
Immunology data			HLA type; references and links to immunology database
Protein 3D-models		NS3 protease/NS4A, NS3 helicase and NS5B models	
Bibliographic references	DDBJ references and Pubmed link	EMBL references and Pubmed link	Genbank sequences and Pubmed references

analyze the database content are introduced. The third part will give examples of how to use the databases. Finally, analyses made possible by the databases and their associated tools are described.

A. Database Content

All sequence databases are regularly updated using publicly available sequences obtained from DDBJ/EMBL/Genbank. Efforts have been made to standardize nomenclature and numbering (*i.e.,*region names, genotype). The data available are summarized in Table 1. The main differences are the annotation and the types of analyses provided, which reflect different emphasis.

I. The Japanese Genomic Map and Phylogeny Database

The Hepatitis Virus Database (HVDB) in Japan started as "HCV Database" in 1997. Recently,data were added of two other hepatitis related viruses, hepatitis B virus (HBV) and hepatitis E virus (HEV). It is routinely reconstructed from the newest release of DDBJ (DNA Data Bank of Japan), and updated four times a year. Currently, it contains approximately 30,000 HCV, 5,000 HBV, and 1,000 HEV entries.

In HVDB, all the HCV entries available in DDBJ are gathered and arranged by two aspects, genomic location and phylogenetic relationship. All the entries are aligned against the reference genome sequence, AF009606 (HCV-H77), which has been defined as a reference for consistent numbering of HCV sequences (manuscript in preparation), to compile information of genomic location. It is called "map information". The information of phylogenetic relationship is prepared for each nucleic and amino acids sequence for each locus, C, E1, E2, p7, NS2, NS3, NS4a, NS4b, NS5a, and NS5b. A data set of the information is called a "division", which contains all nucleic or amino acid sequences that belong to a specific locus, their annotations obtained from the DDBJ release, a multiple alignment of the sequences, a genetic distance matrix, and a phylogenetic tree. In addition, a full sequence division is prepared, which contains all the full-length HCV genome entries. The procedure to make a new release of HVDB is automated.

The Japanese HVDB is mainly oriented toward phylogenetic analyses of viruses, and provides some analytical services as well as data viewers.

II. The European Sequence and Structural Model Database

The development of the European hepatitis C virus database (euHCVdb) started in 1999 as the French HCV database.[8]

Great efforts have been made to develop a fully automatic annotation procedure thanks to a reference set of HCV complete annotated well-characterized genomes representing 18 subtypes. This automatic procedure ensures standardisation of nomenclature for all entries and provides annotation as genomic regions/proteins present in the entry, EMBL references and links to PubMed, genotypes/subtypes, interesting sites (*e.g.,* HVR1) or domains (*e.g.,* NS3 helicase), and the source of the sequence (*e.g.,* isolate, country). Patent and synthetic construct sequences are excluded from the database. The annotation produces an entry in an extended EMBL format. Three genotype qualifiers in the *source* feature have been added to handle the genotype and subtype classification levels (deposited in EMBL, provisional or confirmed as defined in Simmonds et al[5]). A qualifier *prod_ft* has been added to *mat_peptide* feature to handle protein annotations. The syntax of these *prod_ft* qualifier follows the UniProt feature field (FT lines). The entries contain several types of *prod_ft* qualifier (*e.g.,* site, domain, transmem). Structural

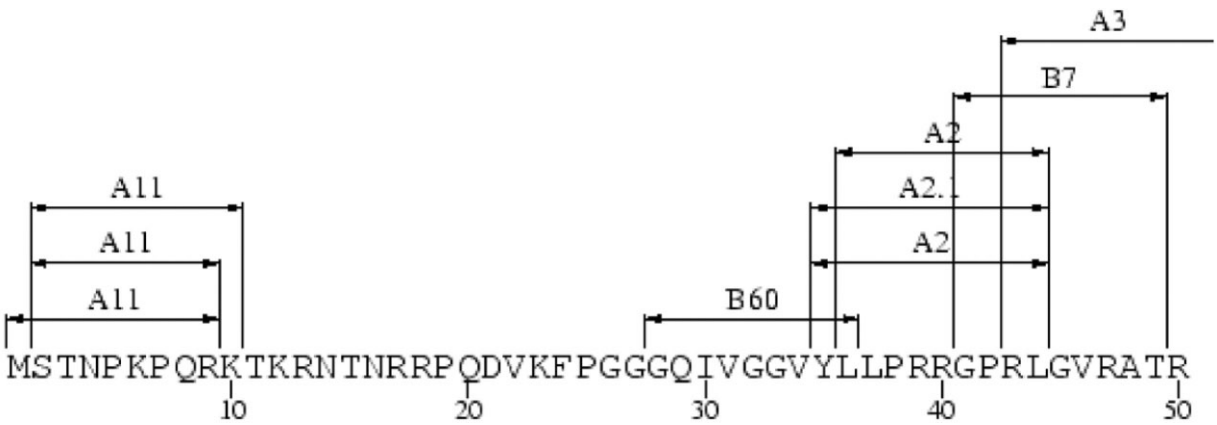

Fig. 1. Map showing the location, frequency and HLA restriction of currently known CTL epitopes in part of the HCV Core protein (Los Alamos HCV immunology database).

annotation of sequences is provided as a *prod_ft* of type *model3d* with links to protein 3D models computed with the Geno3D tool.[9] Moreover, for each of the 10 HCV proteins, pre-computed alignments of full-length sequences available in the database are provided.

The euHCVdb is mainly oriented toward protein sequence, structure and function analyses with the objective to understand resistance at the molecular level.

III. The Los Alamos Sequence and Immunology Databases

The Los Alamos sequence database, which is based on the model of the HIV database, has been available since the Fall of 2003.[10,11] It is updated weekly from GenBank entries. Background information such as genotype, sampling country, and sampling year is harvested and placed in searchable fields. The database staff adds additional annotation based on the literature, in-house phylogenetic analysis, and personal communication with the authors. Annotated fields in the database include:

Sequence information: Genotype, sampling country, -city, -date and –tissue

Host information: Health status, age, gender, ALT level, treatment information, co-infection with HIV and hepatitis B, infection date, country, city, route, and outcome, HLA type, and epidemiological relations to other patients

The HCV Immunology Database is an annotated, searchable collection of HCV immunological epitopes. This database provides an up-to-date, comprehensive listing of defined HCV epitopes that is updated continuously. It includes tables, maps, and associated references of HCV-specific epitopes. To make the epitope database as comprehensive as possible, the optimal boundaries of the epitope do not have to be defined, it only has to be mapped to a region of at most 30 amino acids. One epitope can have multiple database entries, and each entry represents a single publication. The database can be searched on many fields, including epitope location, sequence, genotype, patient HLA, and annotated notes. Figure 1 shows an example of a CTL epitope map.

The Los Alamos databases are mainly oriented toward DNA sequence analysis, molecular epidemiology and immunology.

B. Description of the Associated Tools

Each database website provides numerous analytical tools. While some basic algorithms are shared, the ways in which they are used differ significantly, reflecting the focus of the database. The analytical tools available on the database websites are summarized in Table 2. Examples of available tools, selected by each database on the basis of perceived importance and uniqueness, are discussed in more detail below.

I. The Japanese Database

The HVDB is mainly oriented toward phylogenetic analyses of viruses, and provides some analytical services as well as data viewers:

– **Map viewer:** This viewer displays a map information of HCV data. In this graphic, filled boxes showing the loci are on the top, the ruler of the base position on the genome follows, and horizontal lines showing the location of each entry are plotted under the ruler (Fig. 1). By using it, users can obtain a set of nucleic or amino acid sequences or annotations of entries which cover a specific locus or region on the genome.
– **Tree viewer:** This viewer displays a phylogenetic tree of a division. If bootstrap re-sampling analysis had been applied, the value of each node is also displayed. Users can modify the graphic by changing the root position of the tree, exchanging two branches under the specified node, or zoom in to the specified portion

Table 2. Summary of Analysis Tools Provided by the Databases

	Japanese Database	European Database	Los Alamos Databases
Similarity search	BLAST and FASTA, including searches on database subsets	BLAST, FASTA, SSERACH, HMMSEARCH, including searches on database subsets	BLAST including searches on database subsets
Creating, manipulating and editing multiple sequence alignments	Align Viewer	ClustalW, Multalin, Repertoire, EditAlignment	SynchAligns, Consensus, Primalign, Epilign, SeqPublish, Gapstrip, Sequence Locator, OmniRead, Seq-Convert, Translate,
Protein structure analysis and prediction		Sumo, PHD, GOR, MLRC, SOPM, SOPMA, HNN, DPM, DSC, SIMPA96, PREDATOR, Consensus, Geno3D	
Motifs, sites or signatures detection		InterProScan, PattinProt, ProScan	MotifScan, N-Glycosite, ELF, Vespa
Epitope prediction, immunology tools		PCProf	ELF, Peptgen
Phylogeny, evolution	Tree Viewer		TreeMaker, Findmodel, Branchlength, Syn-nonsyn, PCOORD
Annotation tools (genotyping, numbering, mapping)	Map viewer, Geno/sub-typing (private)	Number,	Gene Cutter, Sequence Locator, Primalign, Epilign
Sequence analysis and display	Map viewer		Geography, Distplot, PCOORD, Entropy, SeqPublish, N-Glycosite

of the tree. Depictions of OTU (leaves of the tree) names can be changed to DDBJ locus names or subtypes, although the accession numbers are used by default (Fig. 2). By using it, users can obtain a set of sequences or annotations of entries that belong to a specific branch or OTU.

– **Align viewer:** The viewer displays a multiple alignment of a division in interleaved format. The first line of the alignment block shows a consensus, followed by all the sequences in which only bases different from the consensus are identified as A, C, G, T, or "—"(gap).

A private service is also available. Before using this service, users need to obtain their access account through a world wide web page. The accounts are used to manage divisions (who owns which division) in the database, and are free of charge. As previously stated, a division is used as a unit of data analysis in this database. Here, users can make their own divisions for data analyses by extracting partial data by referring to map information (Map viewer) or phylogenetic tree (Tree viewer), by uploading their data files, by copying an existing division, and by merging two or more divisions. They can execute phylogenetic analyses using various parameters for multiple alignment, various methods to estimate genetic distances and phylogenetic trees. Once the analyses have finished, users can refer the results by using the Tree or Align viewer. Several tools are also provided in this private area:

– **Virus geno/subtyping:** This is a pipeline service for geno/subtyping of a query sequence. Once a user submit a query sequence with parameters, the system automatically executes mapping the query to the genome, extraction of reference sequences whose genotypes are well identified, multiple alignment, estimation of genetic distances, a phylogenetic tree construction and sometimes bootstrap re-sampling. The results can be referred by using Map, Tree or Align viewers.
– **Others:** Similarity searches by BLAST and FASTA are available. Both nucleotide and amino acid sequence set can be used as targets.

II. The European Database

The euHCVdb website is divided into a static part and a dynamic part.

The static part allows the virologists to access description of genomic regions or proteins (Fig. 3A). Pre-computed multiple sequence alignments of reference genomes or full-length protein sequences can be edited with the EditAlignment applet (Fig. 3B). If the experimental three-dimensional structures of the molecules are known, links to the Protein Data Bank (PDB) files are available and allow the users to view and analyze the structure with the Jmol applet (Fig. 3C). In this static part, the users could also find recommended nomenclature and links to other resources.

In the dynamic part, a query system can be used to build sets of sequences matching given criteria entered by the user through the query interface *e.g.,* extract all the full-length NS3 of genotype 1a. The query interface is divided in 5 main sections: *general information*, *source*, *references*, *cross-references* and *features*. In the *general infor-*

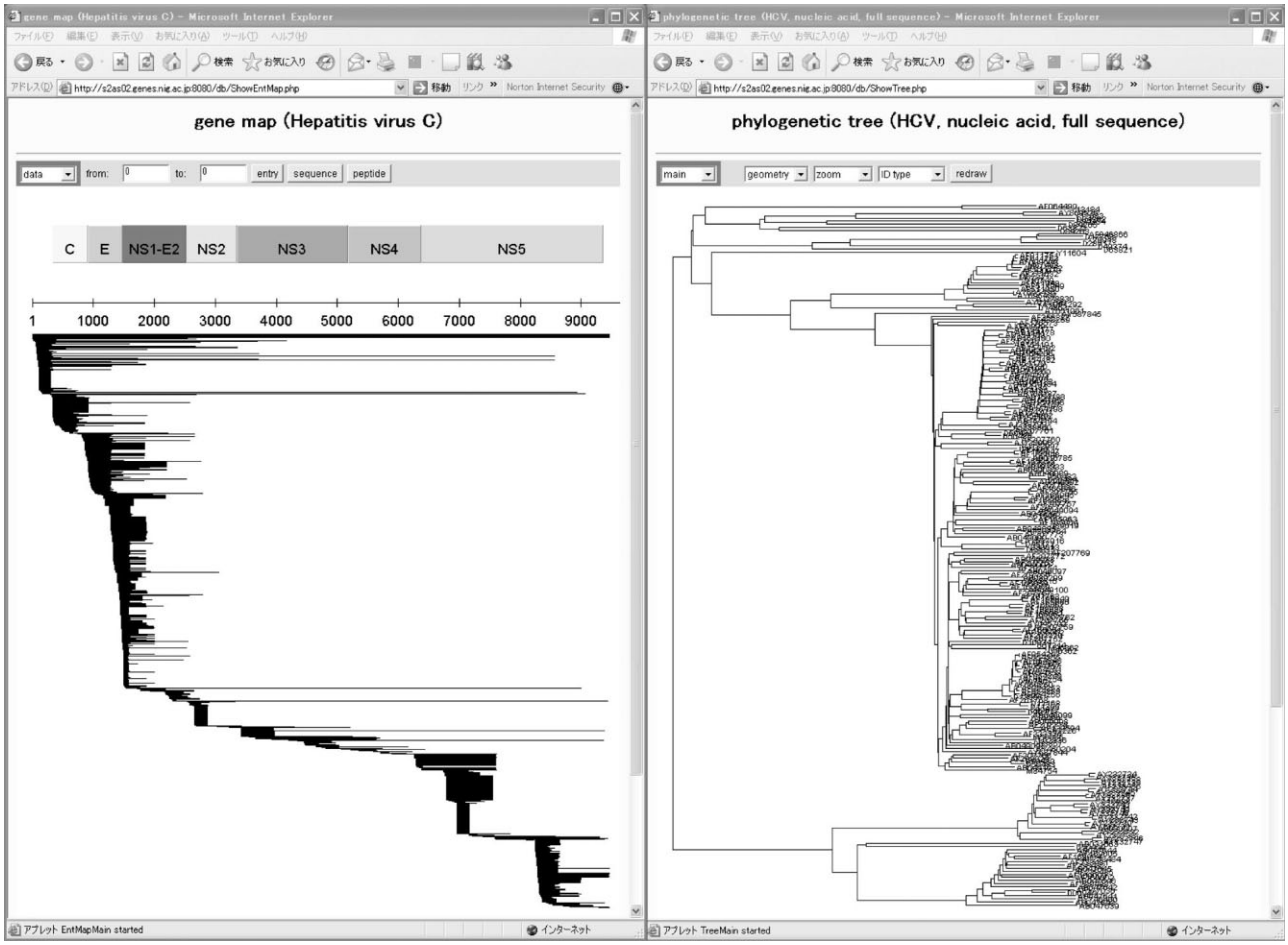

Fig. 2. HCV dataset in the Japanese database website. The left window is the "Map viewer", which shows genomic location of each entry (horizontal bar) as well as the position of each locus (colored box). The right window is the "Tree viewer". In this case, a phylogenetic tree calculated from all the full-length HCV genome entries is displayed.

mation, the *sequence type* parameter defined the type of sequence that will be extracted from the database (*e.g.,* genome will extract sequence as deposited in EMBL, single protein will extract sequence spanning only one protein). Overall, more than 30 parameters can be used to build the query.

The results (Fig. 3D) are displayed in a table where each row corresponds to an entry or a genomic region/protein that is described with a small set of data about it (columns). A hypertext link on the accession number of the entries displays the *Entry details* form containing all the data of the entry with hyperlinks to external resources such as EMBL database or to the Jmol viewer web page to analyze pre-computed 3D models when available. At the start of each row, a checkbox can be used to select results to be exported for further analysis (see below).

Above the table, two toolboxes are available to change display options and browse the results (*Display* box) or to export data for further analysis (*Tools* box). The *Tools* box allows exporting all/selected/unselected/page results as extended EMBL format entries file, FASTA formatted sequences files, tab-delimited file with/ without sequences. Another possibility is to transfer sequence data to the NPS@ web server for analysis. Identical sequences can be removed from the exported set.

The NPS@ system is an integrated sequence analysis Web server that provides in a simple interface 46 analytical methods and 12 databases including euHCVdb.[12] Among this great number of tools, the similarity search tools (FASTA, BLAST) can be used to find similar sequences in order to do sequence clustering or sequence genotyping or to extract sequences matching a given subregion or spanning several regions in the genome or the polyprotein. Other useful tools are the alignment tools (CLUSTALW, MULTALIN). They can be used to analyse variability thanks to the display options available to hide/show residues based on their conservation level or using a reference sequence. A consensus sequence is also provided.

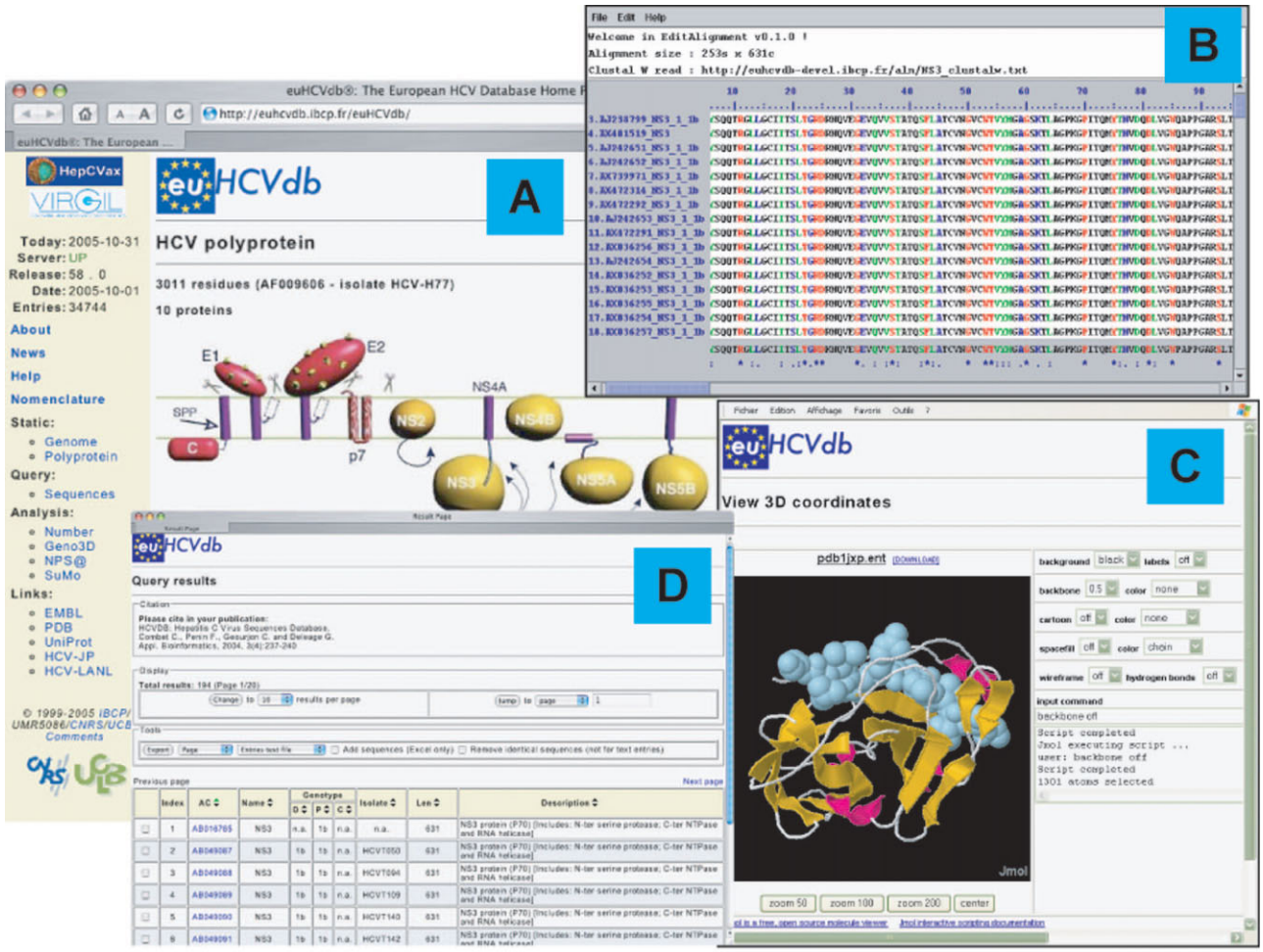

Fig. 3. (A) The European hepatitis C virus database Website home page with clickable proteins pictures. (B) Multiple sequence alignment viewed through the EditAlignment java applet. (C) NS3-NS4A HCV protease rendered through the Jmol applet (PDB code 1jxp chain A and C). (D) Query result page for complete individual NS3 proteins.

Moreover, protein 3D models for variants not yet available in the database can be build with the Geno3D Web server.[13] The protein models can be further analyzed to detect ligand-binding or active sites with the SuMo Web server.[14]

Some HCV-specific analytical tools are under development. The first available one is the *Number* tool that allows numbering of nucleotide or amino acids using a common reference sequence (H77, AF009606).

III. The Los Alamos Databases

The information in the database can be accessed via a versatile search interface that allows searches on some 30 different fields. It lets the user automatically exclude sequences that are (a) from non-human hosts, (b) from patent applications, (c) synthetic, (d) contain a large fraction of indeterminate residues or are judged to be possible contaminants, or (e) are epidemiologically related (either from one patient or from a cluster of linked infections). The search results can be sorted and selected in various ways. They include an icon that shows at a glance how long each sequence is and where in the genome it is located. A graphical overview showing which regions and which genotypes are included in the entire set of retrieved sequences can be generated at the touch of a button. This can help to determine which region is best represented in the sequences of interest, and therefore most suitable for further analysis.

An important feature is the ability to retrieve all sequences in a specific region (for example, E1 and E2). Retrieved sequences can be downloaded as an alignment. The search interface also allows the user to download the background data for the sequences as a tab-delimited file, optionally including the (unaligned) sequences. A recent addition allows users to include their own sequences in the search, and to generate phylogenetic trees based on any combination of database, reference, and user se-

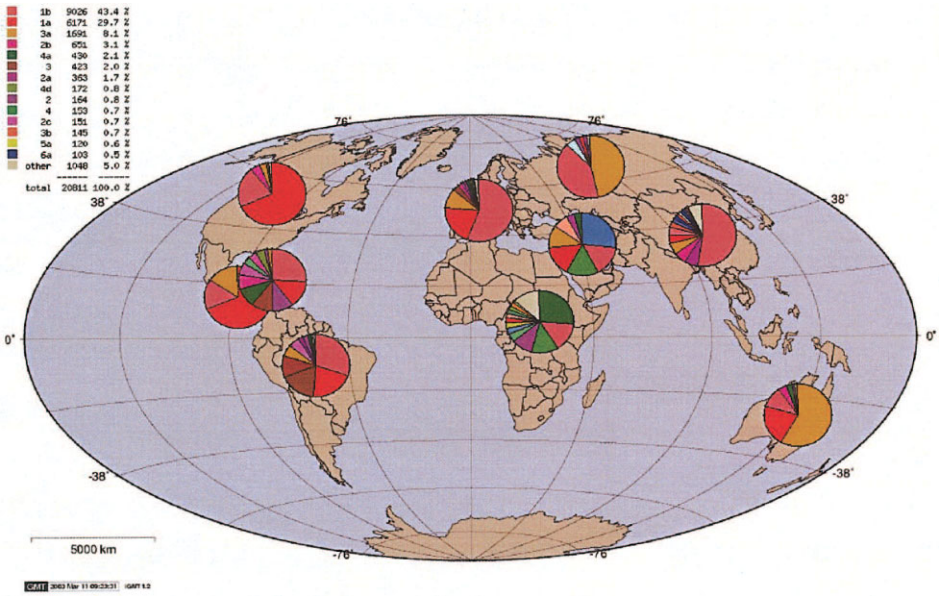

Fig. 4. Numbers of HCV sequences sampled in all geographic areas of the world, colored by genotype (Los Alamos sequence database).

quences. Finally, the website contains manually optimized alignments of all complete genomes and of each gene and protein. These alignments contain only one sequence per patient or transmission cluster, to avoid biasing the data with large sequence sets from a few sources; coding region alignments are codon-aligned, so they can be readily translated.

Many of the tools for this database were originally written for use by the HIV database, and programs and tools continue to be shared and jointly developed. Some of the programs are:

- **The Geography** tool (Fig. 4) draws pie charts of the numbers of sequences of each genotype on geographical maps. The geographical area shown can be determined by the user, down to single-country resolution. The tool also prints tables that can be imported into a different plotting program. This tool is useful to show which genotypes have been found in which countries and the density of sampling in different regions.
- **Glycosite** tallies, plots and compares N-links glycosylation sites in a protein.[15]
- Entropy calculates and plots the variability of each position in an alignment.[16] The variability is calculated as an entropy score.
- **Findmodel** is a web implementation of Modeltest,[17] a procedure that finds the evolutionary model that is most suitable for a given dataset.
- **PCOORD**[18] offers a principal coordinate analysis, a data-reduction technique similar to principal components analysis to identify and plot co-varying positions in groups of sequences.

In addition to these and other analytical programs, a large number of tools are provided to facilitate the work of virologists and immunologists. Many time-consuming tasks are included, such as (for virologists) isolating all the genes from a complete genome, finding the coordinates of a sequence, aligning sequences to all available complete genome sequences and highlighting the differences, joining two partially overlapping sequence alignments, and (for immunologists) finding the variability of an epitope in the database, designing overlapping peptide sets to probe an immune response, analyzing and predicting peptide CTL reactivity data.

C. Use of the HCV Databases

In this section, examples of use of each database are described.

I. The Japanese Database

Users may want to phylogenetically analyze data sets that contain both their own (private) data and publicly available data. In such case, the user can do it by the following steps: First, the private data is uploaded to make a subset. Then, another subset is made from public data by extracting sequences which cover a specific locus, or a region on the genome by using Map viewer or sequences which belong to a specific subtype or a phylogenetic classification by using Tree viewer. Both subsets are merged into one and the resulting data set is analyzed. Once the user choose the analytical methods and parameters, the process which contains multiple alignment, genetic distance estimation, and phylogenetic

tree construction is executed automatically. The results can be viewed using Tree viewer and Align viewer.

II. The European Database

Starting from the query web page, the user can extract subtype 1b NS3 protease domain of 180 amino-acids (*sequence type*=protein feature; feature protein *key*=domain *value*=protease *length min*=180; *genotype/subtype*=provisional *value*=1b). The result set is then transferred to NPS@ to build a multiple sequence alignment after removing identical sequences to reduce redundancy. The resulting color coded alignment highlights identical residues in all sequences in red. A second alignment with only 1a sequences can be constructed in the same way. Using these two alignments, subtype specific conserved position could be identified. As the structure of NS3 protease domain has been solved and using the static NS3 page of the website, the subtype specific position can be visualized on the structure.

Such analysis can be extended to other sub-genomic regions or protein sites, genotypes/subtypes or to other objectives (*e.g.,* PCR primer design, epitope analysis, resistance analysis).

III. The Los Alamos Databases

Some examples of manipulations and analyses that can be done using the Los Alamos databases are:

(a) Sequence analysis:
- Download aligned sequences, ready for analysis: the search interface can generate alignments of any set of database sequences, reference sequences, and user sequences
- Annotate sequences: the Gene Cutter program uses an internal alignment model to find all HCV genes present in a sequence, and their protein translation, to facilitate the annotation process.
- Check sequences for contamination: the HCV BLAST interface, the Treemaker interface to rapidly generate simple phylogenetic trees, and SeqPublish to print the alignments in easily legible format, the website offers the most important tools to check your sequences for contamination.

(b) Immunology:
- Design peptides for peptide scanning: given an amino acid sequence, PeptGen can generate peptides of user-specified length. Peptides are can be colored by hydrophobicity, and N- and C-terminal forbidden amino acids can be specified
- Generate an 'immunological overview' of a protein, restricted to certain HLA types; show potential epitopes that differ from known ones by a few amino acids. Align epitopes to others from the sequence database, and generate maps showing the concentration of epitopes in each protein

D. Examples of Research Performed Using the Databases

The *raison d'être* for the databases described here is the contribution they can make to the HCV research field. We cite a few examples here to illustrate some research that was based on the data they provide.

I. The Japanese Database

The database is mainly used as a resource of phylogenetic analysis results of HCV nucleic and amino acid sequences which are done by using all the publicly available data. For example, multiple alignments of nucleic acid sequences of core and NS5 loci, their genetic distance matrices, and phylogenetic trees prepared in the database help to design PCR primer sets which are to be used a new combination test for HCV genotype and viral load determination.[19] A multiple alignment of C protein sequences prepared in the database and amino acids substitution patterns derived from it was used get a preliminary aspect in studying differences in HCV C protein processing among genotypes 1 and 2.[20] The database is also used as a repository which contains a complete set of HCV sequences publicly available. Phylogenetic analysis of core and NS5b to explain a pattern of endemic infection[21] and an analysis of population genetic history of HCV 1b[22] have been done by use of sequences retrieved from HVDB.

II. The European Database

A typical research example using euHCVdb is the study of HVR1.[23] 460 HVR1 sequences were used from 6 patients and 1382 non-redundant HVR1 sequences of the database. Using multiple sequence alignment of these sequences and the repertoire tool, it was found that (i) despite strong amino acid variability, the physicochemical properties and conformation of HVR1 were highly conserved, and (ii) HVR1 is a globally basic stretch, with the basic residues located at specific positions. This conservation of positively charged residues indicates that HVR1 is involved in interactions with negatively charged molecules at the host cell surface and likely plays a role in host cell recognition and attachment. Other studies of various regions of the HCV genome have been done either in collaboration with euHCVdb (*e.g.,* Carrère-Kremer et al.[24]) or independently by others.[25]

III. The Los Alamos Databases Using only database resources, a recent paper discussed the distribution of glycosylation sites in HCV and HIV.[15] Authors from UCSD introduced a simplified model to describe substitution rate distributions, using data from the Los Alamos se-

quence database as an illustration.[26] Two studies by the group of Peter Simmonds in Edinburgh relied partly on the Los Alamos database: one analyzing RNA secondary structures in the HCV Core and NS5B genes,[27] the other reviewing current knowledge about HCV evolution.[28] Other examples are papers studying a potential role of CD81 as a cellular HCV receptor,[29] the influence of interferon on HCV evolution,[30] and the variability of genotype 4 sequences from South-Western France.[31]

Conclusion

Hepatitis C is a major cause of liver transplantation worldwide. The hepatitis C virus is the infectious agent responsible for this disease and shows a high genomic variability. In this paper, three databases are described that were designed to handle and analyze this variability. These databases are very different in the added-value data (phylogeny, 3D models, epidemiological and immunological data) they provide in addition to the sequence, the ways they are built and they provide access to the data. These differences reflect their focus: phylogeny for the Japanese database, protein structure for the European database and molecular epidemiology and immunology for the Los Alamos database. Efforts have been made to standardize nomenclature between databases. There is an amount of overlap between the three databases, but the differences in emphasis ensure that they are complementary. It is expected that the databases will each specialize more in their chosen area, and so will become more indispensable.

Acknowledgment: The Los Alamos HCV database project is deeply indebted to the Los Alamos HIV database group, led by Dr. Bette Korber, for much of its tools, infrastructure and philosophy; the two databases continue to collaborate closely and to share software and resources.

References

1. Alter MJ, Margolis HS, Krawczynski K, Judson FN, Mares A, Alexander WJ, et al. The natural history of community-acquired hepatitis C in the United States. The Sentinel Counties Chronic non-A, non-B Hepatitis Study Team. N Engl J Med1992;327:1899-1905.
2. Hoofnagle JH. Hepatitis C: the clinical spectrum of disease. HEPATOLOGY1997;26(Suppl):15S–20S.
3. Krahn M, Wong JB, Heathcote J, Scully L, Seeff L. Estimating the prognosis of hepatitis C patients infected by transfusion in Canada between 1986 and 1990. Med Decis Making 2004;24:20-29.
4. Simmonds P. Viral heterogeneity of the hepatitis C virus. J Hepatol 1999; 31(Suppl 1):54-60.
5. Simmonds P, Bukh J, Combet C, Deleage G, Enomoto N, Feinstone S, et al. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. HEPATOLOGY 2005;42:962-973.
6. Weiner AJ, Christopherson C, Hall JE, Bonino F, Saracco G, Brunetto MR, et al. Sequence variation in hepatitis C viral isolates. J Hepatol 1991; 13(Suppl 4):S6–1S4.
7. Farci P, Purcell RH. Clinical significance of hepatitis C virus genotypes and quasispecies. Semin Liver Dis 2000;20:103-126.
8. Combet C, Penin F, Geourjon C, Deleage G. HCVDB : hepatitis C virus sequences database. Appl Bioinformatics 2004;3:237-240.
9. Penin F, Brass V, Appel N, Ramboarina S, Montserret R, Ficheux D, et al. Structure and function of the membrane anchor domain of hepatitis C virus nonstructural protein 5A. J Biol Chem 2004;279:40835-40843.
10. Kuiken C, Yusim K, Boykin L, Richardson R. The Los Alamos hepatitis C sequence database. Bioinformatics 2005;21:379-384.
11. Yusim K, Richardson R, Tao N, Szinger JJ, Funkhouser R, Korber B, Kuiken CL. The Los Alamos Hepatitis C Immunology Database. Appl Bioinformatics 2005;4:217-225.
12. Combet C, Blanchet C, Geourjon C, Deléage G. NPS@: network protein sequence analysis. Trends Biochem Sci 2000;25:147-150.
13. Combet C, Jambon M, Deléage G, Geourjon C. Geno3D: automatic comparative molecular modelling of protein. Bioinformatics 2002;18:213-214.
14. Jambon M, Imberty A, Deléage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 2003; 52:137-145.
15. Zhang M, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C, et al. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes, and influenza hemagglutinin. Glycobiology 2004;14:1229-1246.
16. Korber BT, MacInnes K, Smith RF, Myers G. Mutational trends in V3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type 1. J Virol 1994;68:6730-6744.
17. Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics 1998;14:817-818.
18. Higgins DG. Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. Comput Appl Biosci 1992;8:15-22.
19. Mukaide M, Tanaka Y, Kakuda H, Fujiwara K, Kurbanov F, Orito E, et al. New combination test for hepatitis C virus genotype and viral load determination using Amplicor GT HCV MONITOR test v2.0. World J Gastroenterol 2005;11:469-475.
20. Kato T, Miyamoto M, Date T, Furusaka A, Hiramoto J, Nagayama K, et al. Differences in hepatitis C virus core protein processing among genotypes 1 and 2. Hepatol Res 2004;30:204-209.
21. Ndjomou J, Pybus OG, Matz B. Phylogenetic analysis of hepatitis C virus isolates indicates a unique pattern of endemic infection in Cameroon. J Gen Virol 2003;84:2333-2341.
22. Nakano T, Lu L, He Y, Fu Y, Robertson BH, Pybus OG. Population genetic history of hepatitis C virus 1b infection in China. J Gen Virol 2006;87:73-82.
23. Penin F, Combet C, Germanidis G, Frainais PO, Deléage G, Pawlotsky JM, et al. Conservation of the conformation and positive charges of hepatitis C virus E2 envelope glycoprotein hypervariable region 1 points to a role in cell attachment. J Virol 2001;75:5703-5710.
24. Carrère-Kremer S, Montpellier-Pala C, Cocquerel L, Wychowski C, Penin F, Dubuisson J. Subcellular localization and topology of the p7 polypeptide of hepatitis C virus. J Virol 2002;76:3720-3730.
25. Timm J, Lauer GM, Kavanagh DG, Sheridan I, Kim AY, Lucas M, Pillay T, et al. CD8 epitope escape and reversion in acute HCV infection. J Exp Med 2004;200:1593-1604.
26. Pond SLK, Frost SDW. A simple hierarchical approach to modeling distributions of substitution rates. Mol Biol Evol 2005;22:223-234.
27. Tuplin A, Evans DJ, Simmonds P. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. J Gen Virol 2004;85:3037-3047.
28. Simmonds P. Genetic diversity and evolution of hepatitis C virus—15 years on. J Gen Virol 2004;85:3173-3188.
29. McKeating JA, Zhang LQ, Logvinoff C, Flint M, Zhang J, Yu J, et al. Diverse hepatitis C virus glycoproteins mediate viral infection in a CD81-dependent manner. J Virol 2004;78:8496-8505.
30. Sumpter R, Wang C, Foy E, Loo Y-M, Gale M. Viral evolution and interferon resistance of hepatitis C virus RNA replication in a cell culture model. J Virol 2004;78:11591-11604.
31. Nicot F, Legrand-Abravanel F, Sandres-Saune K, Boulestin A, Dubois M, Alric L, et al. Heterogeneity of hepatitis C virus genotype 4 strains circulating in south-western France. J Gen Virol 2005;86:107-114.