Available online at www.sciencedirect.com
ScienceDirect
Antiviral Research 75 (2007) 64–74
Antiviral Research
www.elsevier.com/locate/antiviral

SEQHEPB: A sequence analysis program and relational database system for chronic hepatitis B

Lilly K.W. Yuen [a,*], Anna Ayres [a], Margaret Littlejohn [a], Danielle Colledge [a], Andrew Edgely [b], William J. Maskill [a], Stephen A. Locarnini [a], Angeline Bartholomeusz [a]

[a] *Victorian Infectious Diseases Reference Laboratory, 10 Wreckyn Street, North Melbourne, 3051 Vic., Australia*
[b] *Last Resort Support Pty Ltd., 2 Hunter Court, Cranbourne, 3977 Vic., Australia*

Received 10 August 2006; accepted 27 November 2006

Abstract

SeqHepB is a combination of a HBV genome sequence analysis program and a relational database that houses data collected from multiple data sources. Registered users can access the sequence analysis component of SeqHepB online for rapid and detailed interrogation of HBV genomic sequences. Its main function is to determine the HBV genotype, identify key mutations associated with antiviral resistance, and identify clinically important HBV mutants. All information generated is uploaded into a database and integrated with patient medical records, pathology laboratory tests, and supplemental virology results such as *in vitro* drug cross-resistance values. Combined with structured query language (SQL) queries developed in the database, it is possible to extract and correlate clinical, virological, and *in vitro* phenotypic data rapidly and efficiently. An important component of SeqHepB is its ability to integrate mutations detected within the reverse transcriptase (RT) and locate them onto a three-dimensional (3D) model of the HBV RT that can be viewed at any angle with known antiviral drug molecules in the catalytic pocket of the enzyme. SeqHepB will enable virologists and physicians to individualise patient management, cope with the explosion of antiviral associated HBV mutations, and to conduct cross-sectional retrospective or prospective studies on HBV-infected individuals during therapy.

Keywords: Hepatitis B virus; Drug resistance; Relational database; Data mining

1. Introduction

Chronic hepatitis B (CHB) is a life-long disease that typically requires long-term clinical management of patients and places a substantial burden on a national health care system (Margolis et al., 1991), especially if diagnostics and therapeutic management of patients is subsidised. A wide spectrum of disease outcomes is possible in people chronically infected with HBV, and this ranges from mild to chronic hepatitis to the more advanced disease manifestations that include cirrhosis with hepatic decompensation and hepatocellular carcinoma (HCC). There are still a number of unknown factors relating to the outcome of CHB in people at different stages of the disease and its treatment, and there are few reliable risk predictors for the development of the serious consequences of persistent infection or outcome of therapy. To identify risk predictors, it is often necessary to cross-examine and correlate large volumes of patient clinical and routine pathology test data, as well as the specific viral genomic mutations, and published *in vitro* antiviral drug cross-resistance values relating to common clinical isolates of HBV against existing and recently licensed nucleos(t)ide analogues, and then search for novel associations between them. Such data correlation would be labour intensive and time consuming. There is therefore a pressing need to overcome this problem so that the effects of the growing number of approved and investigational antiviral agents against HBV on the natural history of CHB can be characterised efficiently, and to optimise patient management.

There are at least eight different genotypes of HBV (A to H) and a number of recombinant genotypes (Norder et al., 1994). The genomes for each of the HBV genotypes vary in length with specific insertions and/or deletions. The HBV genome is a partially double-stranded circular deoxyribonucleic acid (DNA) molecule that is approximately 3200 nucleotides in length. It is composed of at least four open reading frames (ORF) that overlap with one another so that the total coding capacity is around

* Corresponding author. Tel.: +61 3 9342 3923; fax: +61 3 9342 2666.
E-mail address: lilly.yuen@mh.org.au (L.K.W. Yuen).

0166-3542/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.antiviral.2006.11.014

one-and-a-half times the length of the genome. The longest ORF encodes for the HBV polymerase (POL) and it is composed of at least four functional regions that include terminal protein, spacer region, reverse transcriptase (rt), and RNaseH. This ORF completely overlaps the envelope ORF (PreS1, PreS2 and S), and partially overlaps the precore/core and X ORFs. The use of frameshifted and overlapping reading frames in the HBV genome is an additional layer of complexity for sequence analysis. The occurrence of a single nucleotide mutation within the genome can result in amino acid mutations within two overlapping ORF as well as regulatory regions and/or epitopic sites. An example is the selection of the lamivudine (LMV)-resistant mutant (rtV173L, rtL180M and rtM204V) while a patient is on monotherapy (Delaney et al., 2003). This HBV isolate will also have corresponding mutations within the surface antigen (sE164D and sI195M) as a consequence of the mutations in POL (rtV173L and rtM204V), thus resulting in the selection of a virus in the patient with reduced anti-HBs binding (Torresi et al., 2002).

The only licensed treatments available for CHB in Australia are interferon-alpha and two nucleos(t)ide analogues, LMV and adefovir dipivoxil (ADV). Entecavir (ETV) and telbivudine (L-dT) are nucleoside analogues that have recently been approved by the Food and Drug Administration (FDA) of the United States of America (USA) for use as an anti-HBV agent, and there are several other HBV antiviral agents in various phases of clinical trials. Unfortunately, the efficacy of nucleos(t)ide analogues is limited by the selection of resistant viruses during the course of therapy. In the case of LMV-resistance, rates are estimated to reach over 70% in monoinfected individuals and up to 90% in HIV-HBV coinfected individuals after 4 years of treatment (Benhamou et al., 1999). The main mutation associated with LMV resistance has been well characterised, and it occurs within the YMDD motif of the rt region of POL, rtM204V/I/S (±rtL180M—outside the YMDD motif) (Ling et al., 1996; Niesters et al., 2002). More recently, it has been recognised that there are at least six major genotypic patterns associated with LMV-resistance (Locarnini, 2003), including a triple mutation pattern (rtV173L + rtL180M + rtM204V). Two patterns of ETV-resistance have been characterised, and they are composed of the LMV-resistant mutation at rtM204V/I plus additional mutations at either rtT184G + rtS202I or rtM250V + rtI169T (Tenney et al., 2004). Similarly, the mechanism for ADV-resistance has been characterised with the major resistance mutation located at rtN236T (Angus et al., 2003), and a number of other mutations located in three clusters within the rt region of the HBV POL (Bartholomeusz et al., 2004a). Furthermore, in addition to the major resistance mutations identified for all the antiviral agents, there are a large number of extra mutations that may have a role as compensatory mutations and/or associated with drug resistance. Although there are not as many drugs used for treatment of CHB as for HIV-AIDS, the complex pattern of mutations that can accumulate over time may affect the efficacy of subsequent treatments, thus suggesting mutation identification will quickly become an important component for the management of patients with CHB.

The process of identifying mutations that are associated with antiviral drug resistance occurs in a number of stages. Firstly, patients must meet the clinical definition of resistance, and this is defined either as failure to reduce HBV viral load by $\geq 1 \times \log_{10}$ (IU/ml) within 3 months of following initiation of therapy for primary antiviral treatment failure, or as a rebound of HBV replication of $\geq 1 \times \log_{10}$ (IU/ml) from nadir in patients who have initially responded to the antiviral agent for secondary antiviral treatment failure (Locarnini et al., 2004). Secondly, genomic sequence analysis is required to be performed to identify any novel mutation(s) within the target gene. Lastly, *in vitro* phenotypic confirmation is required to designate specific mutation(s) with reduced antiviral sensitivity. During each of these stages, vast amounts of information would be generated, and would need to be correlated in a rapid and efficient manner.

The ability to find associations between viral genomic data and potential resistance markers for new antiviral agents rapidly can be cost effective in many ways. Such an approach has already been used successfully for HIV treatment and there are a number of public databases available to aid in patient management (Gaschen et al., 2001; Rhee et al., 2003). Nonetheless, databases for different infectious diseases need to be individually tailored. The main aims of SeqHepB are to prevent antiviral drugs that are active against HBV from being incorrectly prescribed, to improve efficiency in the monitoring of therapy and outbreaks of vaccine-escape HBV mutants among the general vaccinated population, to prevent associated disease progression that may eventuate in a liver transplant, and to improve professional practice guidelines for the treatment of patients with CHB.

2. Methods

2.1. Sequence analysis program

The sequence analysis component of SeqHepB V 9.2 (Last Resort Support Pty Ltd., Vic., Australia) is an optimized C program that can be accessed via the web (http://www.seqvirology.com/genome7/index.htm) by registered users with their username and password.

The program is designed for genotype and mutational analysis of HBV genomic sequences that can be entered either as a nucleotide sequence or translated amino acid sequence. The data submission screen is simple and user-friendly, and has fields to capture a patient identifier, sequence identifier, sample date, and the sequence for analysis. Currently, the program allows for data entry by copying and pasting the sequence from text-based programs.

SeqHepB utilises a complete genome sequence for each of the genotypes from GenBank (Table 1) to enable differentiation of HBV genotypes. Selection criteria for the GenBank sequences include: (i) patients from whom the sequences were obtained must have been anti-HBV drug naïve, and (ii) the sequence must be an earlier representative for that HBV genotype. These GenBank sequences functions as the set of reference standards to which all the submitted HBV genomic sequences are compared. All the ORF of the reference standards are tagged to facilitate

Table 1
Complete genome sequences of the eight HBV genotypes used as reference standards in SeqHepB

HBV genotype	GenBank accession number
A	X02763
B	D00330
C	AB033556
D	X02496
E	X75657
F	X75663
G	AF160501
H	AY090460

detailed mutational analysis on all the overlapping genes within the HBV genome.

The genotype of the HBV sequence submitted for analysis is determined by comparison with each of the reference standards in turn using an incremental string pattern matching algorithm that analyses nine characters at a time, and a similarity score is calculated for each iteration. The genotype of the reference standard that resulted in the highest total similarity score is assigned to the submitted sequence. Subsequent mutational analysis on the submitted sequence is performed by comparing with the reference standard of the HBV genotype determined.

2.2. Relational database management system

The database component of SeqHepB is not accessible over the internet, and is designed to enable integration of data from multiple disciplines into a single relational database. Application tools required to access the relational database and for data manipulations are written in MS Access 2000, and they are linked to a MS Access 2000 database where all the HBV-related data are housed. The relational database is currently composed of 34 data tables (each containing a specific data type), and supplementary tables.

2.2.1. Patient demographics

The majority of patients in the database resided in Australia (88%), and the remainder were data collected from overseas collaborators (Hong Kong, South Africa, New Zealand, Greece, United States, and Germany). There were approximately 76% males and 24% females amongst patients who resided in Australia, and they were infected with viruses of HBV genotypes A (16%), B (30%), C (32%), D (20%), E (1%) and G (1%). Country of origin was only known for approximately 10% of these patients, and there were approximately 36% born in Australia, 1% from New Zealand, 45% from Asia, 2% from The Americas, and 16% were from European-Mediterranean countries.

2.2.2. Data tables

Data tables can be classified into four categories that include patient clinical histories, routine pathology test results, viral genomic sequence data, and *in vitro* antiviral cross-resistance phenotypic data. A diagrammatic representation of the database schema for the data tables is shown in Fig. 1. Clinical data include patient demographics, infections with other pathogens (to determine co-infection status), clinical histories (relevant past and current medical notes), treatment histories, and liver biopsy details. Routine pathology test results include serum HBV DNA viral load levels, routine serological tests for hepatitis viruses B, C, and D, and liver function tests (such as serum alanine aminotransferase, ALT levels). Viral genomic sequence data types include all the submitted sequence data and the nucleotide and amino acid variations to reference sequences detected by the sequence analysis program. Type of data included in the *in vitro* phenotypic tables depends on the study from which the data was extracted, and this can include comments associated with a particular mutant or their replication efficiency and antiviral susceptibility details.

Unique patient, specimen, and sequence identifying (ID) numbers that are generated automatically by the database are used to interlink all data tables, and to ensure data integrity. The three main tables in the database are Patient Details, Specimen Details, and Sequence Details, and they are linked to all the remaining data tables via primary and foreign keys with one-to-many relationships. Tables that contain *in vitro* phenotype and reference data are not directly linked to any of the above tables. One-to-many relationships between these data types and mutation records are generated via structured query language (SQL) queries during data analysis.

2.2.3. Supplementary tables

The majority of supplementary tables within the database serve as lists for dropdown boxes in data-entry screens to restrict the type of data that can be entered into specific data fields. Nonetheless, there are a number of supplementary tables that play important roles in data analysis. Of these, the most important is the Consensus table.

The Consensus table is used to filter out all the mutations identified at gene positions that are considered to be polymorphic during the generation of cumulative sequence analysis reports. Polymorphic changes are defined as amino acid variations or substitutions that can occur naturally within one HBV genotype or across all the genotypes, and they were identified via a number of steps: (i) nucleotide sequences covering the majority of the HBV genotypes from patients with CHB were downloaded from GenBank (Table 2), (ii) all nucleotide sequences or translated amino acid sequences were aligned using MacVector (Accelrys Inc., CA, USA) to generate a consensus sequence, and (iii) for each gene or promoter region position, all the nucleotide(s) or amino acid(s) observed in the consensus sequence were noted. Each record in the Consensus table is composed of a list all the nucleotide(s) or amino acid(s) observed in the consensus sequence at a particular gene or promoter region position for each of the HBV genotypes. All nucleotide or amino acid mutations detected by the sequence analysis component of SeqHepB that are present in the Consensus table are considered to be polymorphic.

3. Results and discussion

To enable correlation of large volumes of inter-disciplinary HBV-related data, a software system (SeqHepB) dedicated to

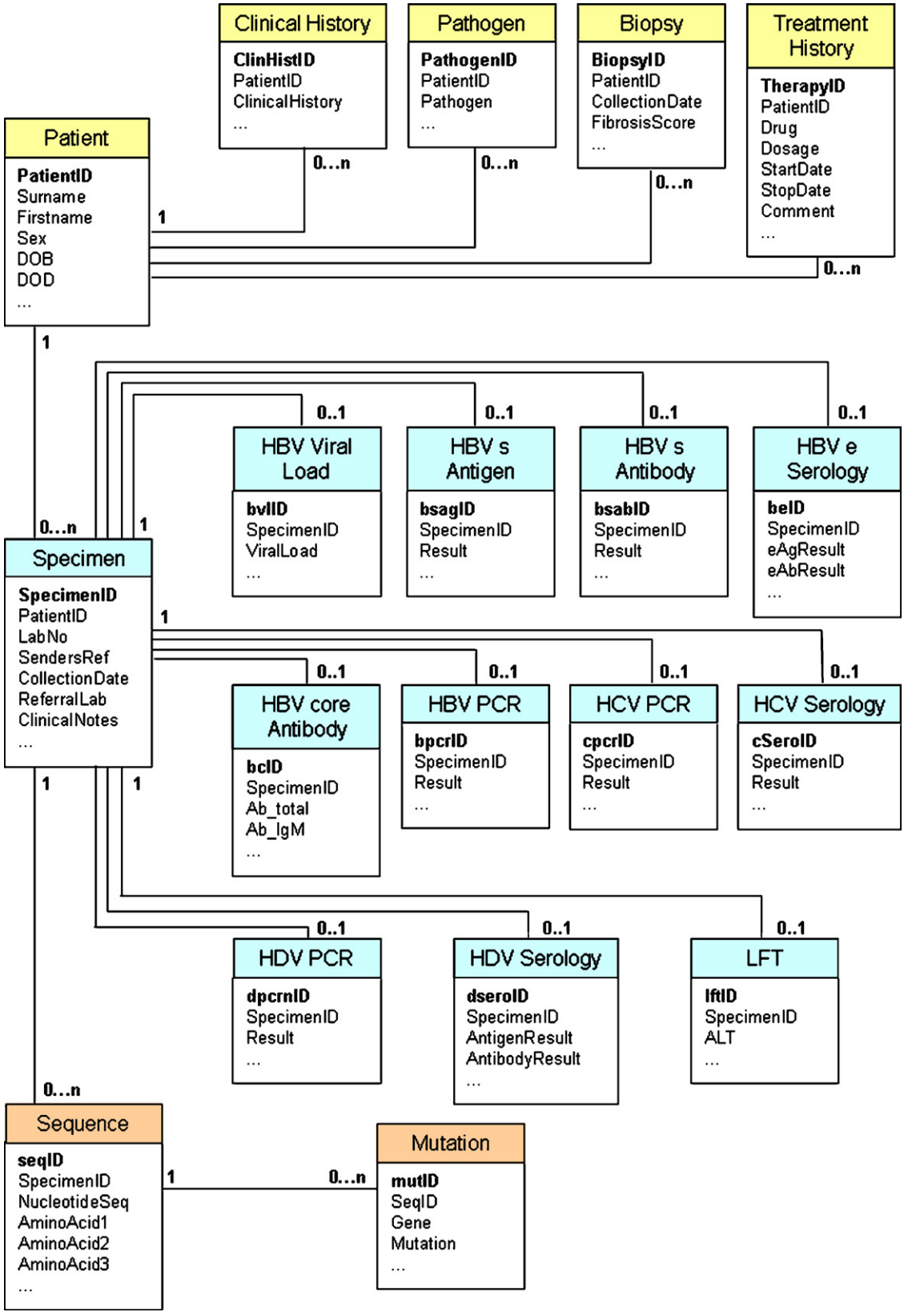


Fig. 1. A diagrammatic representation of the schema for the data tables in the relational database component of SeqHepB. Data tables that house patient clinical data are shaded in yellow, those that house the routine pathology test results are shaded in blue, and those that house the viral genomic sequence data are shaded in brown. Only the main data fields in each of the data table are listed.

Table 2
GenBank DNA sequences used for the generation of a consensus sequence, and this was in turn used for the identification of polymorphic mutations among the HBV genotypes

HBV genotype	GenBank accession number
A	M57663 X70185 X02763 V00866
B	M54923 D00329 D00330 D00331
C	X75656 X75665 D12980 X01587 V00867 M38454
D	V01460 X72702 X02496 X68292 X65257
E	X75664 X75657
F	X75658 X75663

the analysis of HBV-associated data was established. SeqHepB is composed of: (i) a web-based HBV sequence analysis program that can be accessed online by registered users and (ii) a relational database that cannot be accessed over the internet and it houses the data collected from collaborating physicians (patient clinical data), diagnostic laboratories (relevant pathology test results including HBV DNA viral load and HBV genome sequence data generated by the analysis program), and *in vitro* phenotypic data from published papers as well as other unpublished studies (S. Locarnini and T. Shaw, personal communication). An overview of the steps involved in data processing by SeqHepB is shown in Fig. 2. The database can be mined periodically for data patterns to aid in the identification of key virological factors that are potentially associated with severe disease, antiviral resistance, and causes for treatment failure.

3.1. Sequence analysis program

Since HBV was one of the first viruses sequenced, nomenclatures used to describe mutations varied considerably among different published literatures. A new nomenclature was developed by Stuyver et al. (2001) to overcome this problem, especially for describing antiviral resistant mutations within POL. In this new nomenclature, each functional region of POL is numbered individually. Thus, all amino acid variations detected by the SeqHepB analysis program are numbered relative to the start of each of the respective POL functional region. For nucleotide variations detected within the BCP, precore or core genes, the old nomenclature which numbers the nucleotides relative to an *Eco*R1 endonuclease restriction enzyme site is used (Galibert et al., 1979).

The SeqHepB sequence analysis program performs analysis in three phases (Fig. 2). Three tasks are performed in the first phase. Firstly, the program confirms that the sequence submitted does indeed belong to HBV; that is, the submitted sequence is $\geq$3% similar to at least one of the reference HBV genotype standards. The success rate in differentiating HBV and closely related non-HBV was found to be 100% using nucleotide sequences of 23 HBV, 15 woodchuck hepatitis virus (WHV) and 20 duck hepatitis B virus (DHBV) obtained from GenBank (data not shown). Secondly, this phase determines whether the submitted sequence is a nucleotide or amino acid sequence, and lastly, it determines the HBV genotype. Comparison of results determined via SeqHepB and INNO-LiPA HBV Genotyping assay (Innogenetics, Belgium) for 73 HBV sequences obtained from patients revealed a concordance level of 98.6% (data not shown). Nonetheless, the reliability of the genotype determined using individual genes or limited gene sequence is highly dependent on the size of the sequence submitted as well as where it is located in the genome. In general, if an individual gene is to be used, the envelope gene S that overlaps the POL ORF will provide the most accurate analysis for genotyping.

Once the genotype of the virus is determined, the nucleotide sequence is translated into the three possible reading frames to enable identification and analysis of the deduced POL, envelope, precore and core sequences. The analysis of the nucleotide sequence occurs in the second phase, while the analysis for the three translated amino acid sequences occur in the third phase. The second and third phases occur almost simultaneously by two different modules of the program. The main function of these two modules is to directly compare the test sequence with the pre-stored reference sequence of the same genotype, identify all the relevant HBV genes and regions, identify all differences between the two sequences, and flag the clinically important mutations detected. For example, rtM204V/I $\pm$ rtL180M in the rt gene of POL is associated with LMV-resistance, sG145R in the envelope surface (S) gene is associated with vaccine or hepatitis B immunoglobulin (HBIG) escape, G1896A in the precore gene is associated with loss of HBeAg synthesis, A1762T and G1764A in the basal core promoter (BCP) are associated with a reduction in HBeAg synthesis. Examples of key clinically important mutations that are reported by SeqHepB are shown in Table 3.

There are two limitations to this sequence analysis program. The first is that the program does not discriminate between polymorphic and non-polymorphic mutations. Nevertheless, it should be noted that this limitation was implemented on purpose since it is possible for mutations occurring at polymorphic sites to have a compensatory role in drug-resistant mutants that would otherwise have reduced replicative fitness. Therefore all nucleotide and amino acid differences found between the submitted sequence and the reference sequence should be noted and recorded into the relational database. Discrimination between polymorphic and non-polymorphic mutations is handled in the

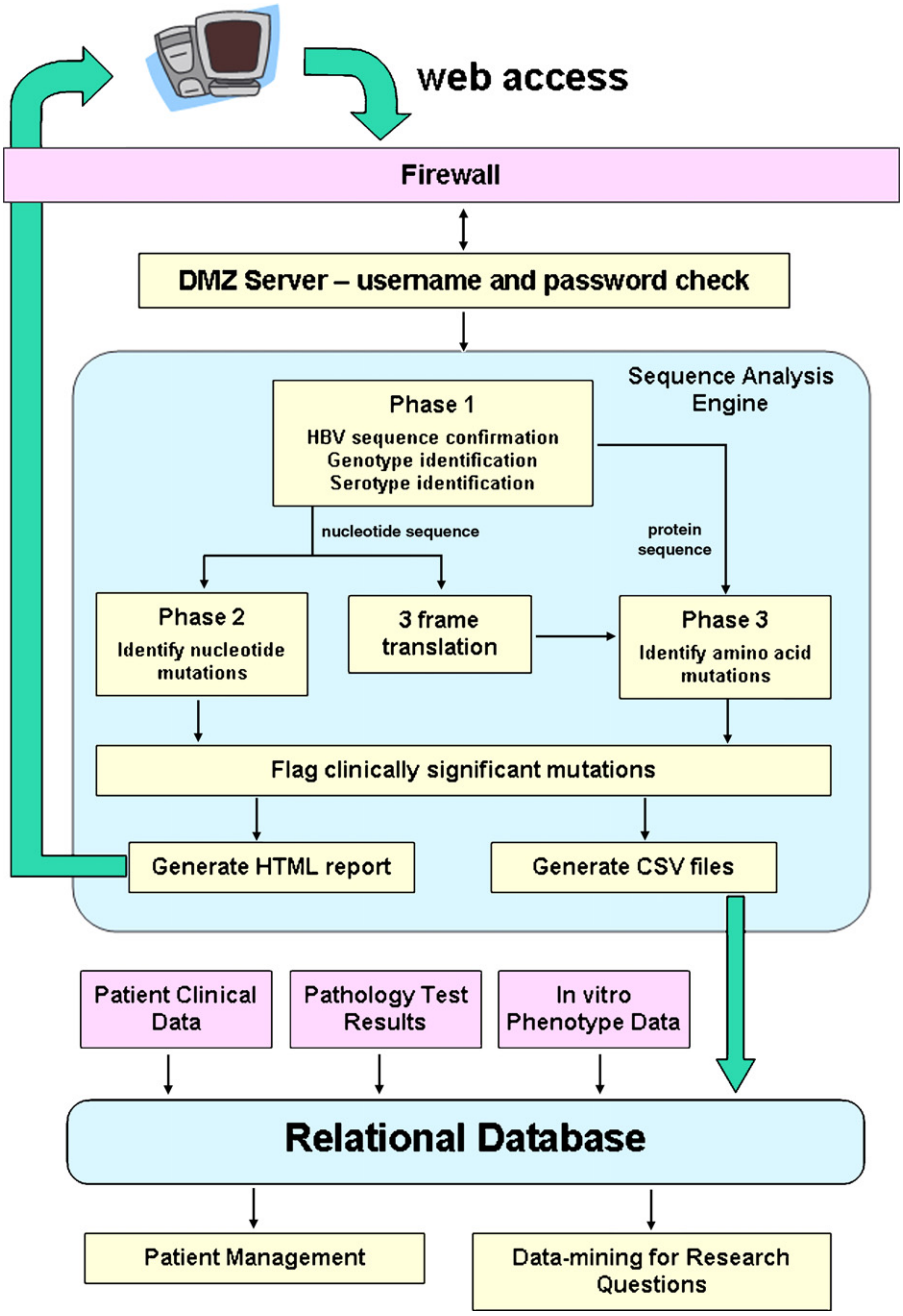


Fig. 2. Schematic diagram showing an overview of the steps performed by SeqHepB, starting from when a user logs in to the sequence analysis component of the software system via the web to the end usage of the data housed in the relational database component.

database component of SeqHepB. The second limitation to the sequence analysis program is that the current version does not support the identification of recombinant viruses, but this will be addressed in the next version of the program.

Information generated by the sequence analysis program is reported in two ways (Fig. 2). Following analysis, the program will generate a HTML document that shows the details of registered user, unique patient and sample identifiers, the HBV genotype and deduced serotype (if the submitted sequence includes the envelope S gene) determined. Other details include relevant statements for clinically significant mutations, a list of all the nucleotide and amino acid variations, insertions, and deletions detected relative to the reference sequence of same HBV genotype, the sequence submitted, and the translated products of all three reading frames if the data submitted is a nucleotide sequence. Registered users can print and save this version controlled and time-stamped report to facilitate easy and accurate preparation of reports and record keeping. An example of such a report is shown in Fig. 3. Concurrently, all data generated for this document are captured and recorded into the database component of SeqHepB for record keeping and subsequent analysis and data mining.

Home | About Us | Contacts | Articles | Articles

SeqHepB SeqHepB Virtual Virology System

SeqHepB RESULT at: 15:26:35 28 Mar 2006 rev:200405-01 ver:9.2 Message ID:HB90003212

Response to: yuenl - Lilly Yuen - SeqHepB Management Group

Patient: Patient 1 - Unknown Patient,

Sample Date: 28/03/2006 Sample ID: 12345678

The sequence data entered was compared with the SeqHepB database.

The sequence matched **HBV Genotype C**, HBV Serotype adw

Your base-pair raw data has been translated into amino acid data for consistent reporting.

Significant Mutation Findings:

Sequence analysis of the HBV polymerase region has detected a RT_L180M (L526M) and RT_M204V (M550V) change consistent with lamivudine resistance and also famciclovir resistance.

HBV Gene Analysis:

The amino acid data entered matched with the amino acids in the following genes:

Gene Name	Start in Gene
POL	421
S	67

Insertions, Deletions and Mutations:

The following nucleotide and amino acid mutations were detected:

```
 AA Mutation:L[106]M ( L[526]M ) RT_L180M (Reverse Transcriptase)
 AA Mutation:M[130]V ( M[550]V ) RT_M204V (Reverse Transcriptase)
NT RT_C538A (Reverse Transcriptase)
NT RT_A610G (Reverse Transcriptase)
AA PreS1_I369M   |   PreS2_I250M   |   HBsAg_I195M (Hep B Surface Antigen)
End of Indel Table.
The RAW data used for this report are:
TCCAATTTGTCCTGGCTATCGTTGGATGTGTCTGCGGCGTTTTATCATATTCCTCTTCATCCTGCTGCTA
TGCCTCATCTTCTTGTTGGTTCTTCTGGACTACCAAGGTATGTTGCCCGTTTGTCCTCTACTTCCAGGAA
CATCAACTACCAGCACGGGACCATGCAAGACCTGCACGATTCCTGCTCAAGGAACCTCTATGTTTCCCTC
TTGTTGCTGTACAAAACCTTCGGACGGAAACTGCACTTGTATTCCCATCCCATCATCCTGGGCTTTCGCA
AAATTCCTATGGGAGTGGGCCTCAGTCCGTTTCTCATGGCTCAGTTTACTAGTGCCATTTGTTCAGTGGT
TCGCAGGGCTTTCCCCCACTGTTTGGCTTTCAGTTATGTGGATGATGTGGTATTGGGGGCCAAGTCTGTA
CAACATCTTGAGTCCCTTTTTACCTCTATTACCAATTTTCTTTTATCTTTGGGTATACATTTAAACCCTA
ATAAAACCAAACGTTGGGGTTACTCCCTTAACTTCATGGGATATGTAATTGGAAGTTGGGGTACTTTACC
ACAGGAACATATTGTACTAAAACTCAAGCAATGTTTTCGAAAACTGCCTGTAAATAGA
The translated data used for this report are:
Frame 0:
SNLSWLSLDVSAAFYHIPLHPAAMPHLLVGSSGLPRYVARLSSTSRNINYQHGTMQDLHDSCSRNLYVSL
LLLYKTFGRKLHLYSHPIILGFRKIPMGVGLSPFLMAQFTSAICSVVRRAFPHCLAFSYVDDVVLGAKSV
QHLESLFTSITNFLLSLGIHLNPNKTKRWGYSLNFMGYVIGSWGTLPQEHIVLKLKQCFRKLPVNR
Frame 1:
PICPGYRWMCLRRFIIFLFILLLCLIFLLVLLDYQGMLPVCPLLPGTSTTSTGPCKTCTIPAQGTSMFPS
CCCTKPSDGNCTCIPIPSSWAFAKFLWEWASVRFSWLSLLVPFVQWFAGLSPTVWLSVMWMMWYWGPSLY
NILSPFLPLLPIFFYLWVYI*TLIKPNVGVTPLTSWDM*LEVGVLYHRNILY*NSSNVFENCL*I
Frame 2:
QFVLAIVGCVCGVLSYSSSSCCYASSSCWFFWTTKVCCPFVLYFQEHQLPARDHARPARFLLKEPLCFPL
VAVQNLRTETALVFPSHHPGLSQNSYGSGPQSVSHGSVY*CHLFSGSQGFPPLFGFQLCG*CGIGGQVCT
TS*VPFYLYYQFSFIFGYTFKP**NQTLGLLP*LHGICNWKLGYFTTGTYCTKTQAMFSKTACK*
```

Email Us

Victorian Infectious Diseases Reference Laboratory
10 Wreckyn Street, North Melbourne, Victoria 3051, Australia
Postal Address: Locked Bag 815, Carlton Sth, 3053 Australia
Tel: 61 3 9342.2602, Fax: 61 3 9342.2665
Email: seqhepb.coordinator@mh.org.au

Copyright: Melbourne Health, 2002

Analysis program and website provided by:
Last Resort Support Pty Ltd
support@lrsupport.com.au

Fig. 3. An example of a report generated by the sequence analysis component of SeqHepB.

Table 3
A sample of clinically significant mutations that are reported by SeqHepB (Ayres et al., 2004)

Regulatory Region or Gene[a]	Mutation	Frequency (%)[b]	Comment
BCP	A1762T	26	Reduction in HBeAg antigen expression. HNF4 binding site altered. Associated with fulminant hepatitis
BCP	G1764A	28	Reduction in HBeAg antigen expression. HNF4 binding site altered. Associated with fulminant hepatitis
Precore	G1896A	23	Stop at codon 28 results in truncation and loss of HBeAg
HBsAg	sG145R	1	Consistent with HBIG or vaccine escape mutants
HBsAg	sP120T	2	Consistent with HBIG or vaccine escape mutants
RT	rtM204I/V ± rtL180M	27	Associated with Lamivudine resistance
RT	rtN236T ± rtA181T/V	3	Associated with Adefovir resistance
RT	rtI169T + rtM250V + (rtM204I/V ± rtL180M)	0.1	May be associated with Lamivudine and Entecavir resistance
RT	rtT184G + rtS202I + (rtM204I/V ± rtL180M)	0.5	May be associated with Lamivudine and Entecavir resistance

[a] BCP: basal core promoter; RT: reverse transcriptase.
[b] Total frequency (%) of samples with specific individual and combined mutations in the database.

3.2. Relational database

The database component of SeqHepB currently contains routine diagnostic pathology and virology data for 1598 patients. During May 2006, associated with these patients are 329 clinical histories, 1465 treatment histories, 168 biopsy results, and 19,081 specimen records. In terms of routine pathology tests performed on the samples, there are 26,248 records in the database, and these include HBV, hepatitis C virus (HCV), and hepatitis D virus (HDV) related pathology test results, as well as liver function and haematology test results. Samples within the database are also associated with 3391 HBV sequence information corresponding to 104,246 nucleotide or amino acid variation data points.

The *in vitro* phenotype data components of the relational database are new additions to the system. There are at present approximately 323 records in the database, and unique identification numbers are used to link them with journal references in the reference data table via one-to-many relationships. All records in the reference table were imported from an EndNote library (Thomson ResearchSoft, Philadelphia), and the table is regularly updated from the library using pre-established SQL queries.

The relational database was designed for two main purposes—to function as a patient management tool, and to function as a data-mining instrument that utilizes pre-established SQL queries, sub-routine programs, and/or integration with other specialised softwares for data analysis. Access to personal details of patients is restricted only to managing physicians and laboratory staff that are associated with the management of patient data. For ease of use, self-intuitive graphical user interfaces (GUI) have been developed for the execution of all routinely used functions within the relational database system. To correlate and extract data for specific research studies, SQL queries and program codes can be tailor designed and created to target specific patient populations.

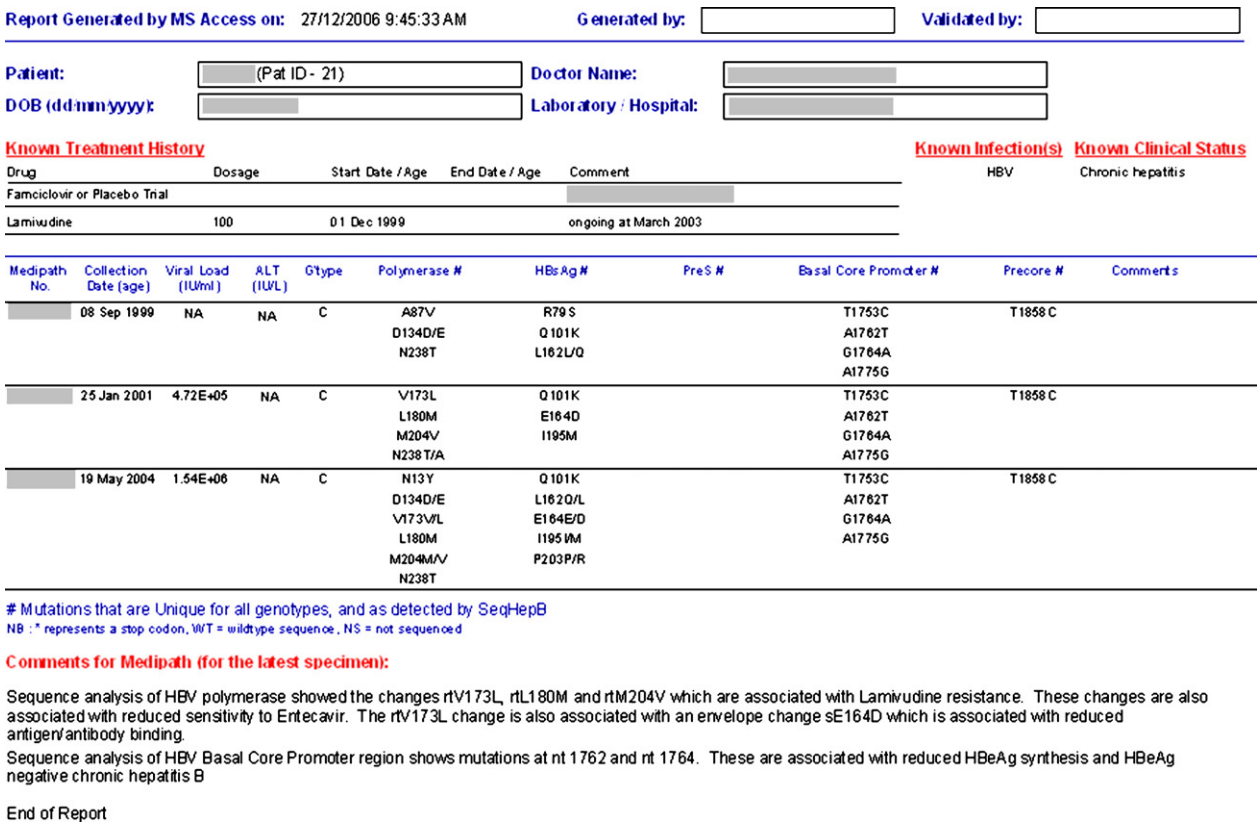
3.2.1. Patient management functions

All patient and sample records are searchable and viewable by a number of means. Record screens open in View-Mode only (except those for new record entries), and users would have to perform a number of tasks before records can be manipulated to avoid accidental modification or deletion of data.

3.2.1.1. General database summaries. Numerous queries and report templates have been established to enable easy generation of patient or sample summaries that are printable. These include a listing of all patients in the database ordered by physicians, a listing of all samples and HBV-related pathology test results for a patient, and a listing of all patients who have co-infections with another pathogen ordered by physicians. All summary reports generated are date and time stamped.

Queries and summary report templates are also available for determining the number of instances of all the nucleotide and amino acid variations (based on the latest sample of patients submitted for sequence analysis) that are in the relational database. Three types of summaries can be generated: (i) frequency of amino acids for the structural and non-structural genes such as PreS1, PreS2, HBsAg, terminal protein, rt, spacer, RNaseH and X; (ii) frequency of nucleotides for the regulatory genes or regions such as BCP, precore, and of amino acids for precore and core, and (iii) frequency of all HBV rt amino acid variations detected and their effects on the envelop ORF, and vice versa. These summaries also include insertions and deletions of nucleotide and amino acids detected by the HBV sequence analysis program.

3.2.1.2. Sequence analysis reports. For viral genomic sequence analysis, there are queries and templates available for three types of reports. The first and second types are cumulative sequence analysis reports that can be generated for a single patient and for all the patients in a single cohort respectively. Such reports will list all the samples that have been sequence analysed (ordered by patients and sample collection dates), all the non-polymorphic nucleotide and amino acid variations (within the rt, BCP and precore regions of the HBV genome) that were detected by SeqHepB for each of the corresponding sample, and the HBV viral load (IU/ml) and ALT level (IU/ml) determined for each serum sample if the pathology tests were performed. Cumulative sequence analysis reports generated for individual patients



Report Generated by MS Access on: 27/12/2006 9:45:33 AM Generated by: Validated by:

Patient: (Pat ID - 21) Doctor Name:

DOB (dd/mm/yyyy): Laboratory / Hospital:

Known Treatment History

Drug	Dosage	Start Date / Age	End Date / Age	Comment
Famciclovir or Placebo Trial				
Lamivudine	100	01 Dec 1999		ongoing at March 2003

Known Infection(s): HBV

Known Clinical Status: Chronic hepatitis

Medipath No.	Collection Date (age)	Viral Load (IU/ml)	ALT (IU/L)	G'type	Polymerase #	HBsAg #	PreS #	Basal Core Promoter #	Precore #	Comments
	08 Sep 1999	NA	NA	C	A87V D134D/E N238T	R79S Q101K L162L/Q		T1753C A1762T G1764A A1775G	T1858C	
	25 Jan 2001	4.72E+05	NA	C	V173L L180M M204V N238T/A	Q101K E164D I195M		T1753C A1762T G1764A A1775G	T1858C	
	19 May 2004	1.54E+06	NA	C	N13Y D134D/E V173V/L L180M M204M/V N238T	Q101K L162Q/L E164E/D I195I/M P203P/R		T1753C A1762T G1764A A1775G	T1858C	

Mutations that are Unique for all genotypes, and as detected by SeqHepB

NB : * represents a stop codon, WT = wildtype sequence, NS = not sequenced

Comments for Medipath (for the latest specimen):

Sequence analysis of HBV polymerase showed the changes rtV173L, rtL180M and rtM204V which are associated with Lamivudine resistance. These changes are also associated with reduced sensitivity to Entecavir. The rtV173L change is also associated with an envelope change sE164D which is associated with reduced antigen/antibody binding.

Sequence analysis of HBV Basal Core Promoter region shows mutations at nt 1762 and nt 1764. These are associated with reduced HBeAg synthesis and HBeAg negative chronic hepatitis B

End of Report

Fig. 4. An example of a cumulative sequence analysis report showing the treatment history, co-infection status, and disease status of the patient, as well as all the non-polymorphic mutations detected by SeqHepB in the patient's HBV sequences analysed over time. In addition, the report is date and time stamped, and validation fields are provided for audit tracking purposes.

also show treatment histories, co-infected pathogens, and HBV-related liver disease stages that are available for that patient, as well as the self-generated comments describing the significance of any clinically important mutations identified for the most recent serum sample analysed (Fig. 4).

The third type of reports that can be generated is for the HBV from patients who have had full-length HBV genome sequence analysis performed on the respective serum samples. The format of the report is similar to the one describe above; that is, it lists all the samples that had full-length genome analysis performed and are listed in the order of sample collection date. The full-length analysis is more detailed and comprehensive, and is not restricted to the four main ORF of the HBV genome. Analyses also include the regulatory elements, promoter sites and epitope sites. In addition, a list of references associated with all the clinically significant mutations detected for each sample is also shown in the report for further readings.

3.2.2. *Data mining functions*

Functions have been established in the database to enable searches on patients and their samples that possess specific combination of nucleotide or amino acid variations. Three types of report summaries can be generated, and they include: (i) lists all patients alphabetically who have those combination of variations, (ii) summaries on the frequency (number of instances) of those variations within the database, and (iii) lists that are similar to the first type of reports except that in addition to showing all the other nucleotide and amino acid variations detected for the respective samples, corresponding drug histories of the patient at time of sample collection would also be listed.

The combined use of these functions and the cumulative sequence analysis reports generated for patients who have been treated with antiviral agents would facilitate the identification of associations between mutation clusters found within POL and antiviral-resistance. Such an approach has enable the identification of associations between rtN236T and adefovir resistance (Angus et al., 2003; Tenney et al., 2004).

Other functions established within the database were predominately designed and created to ensure consistency in which clinical, virological and viral genomic data are correlated and extracted for specific research studies. For example, a series of subroutine functions were developed to facilitate the discovery of novel associations between resistance to an antiviral agent, deduced amino acid changes within the rt, HBV viral load levels, and time on treatment. Generation of two datasets (pre- and during-treatment) by these functions is simply by entering the antiviral agent of interest as a parameter. Briefly, these functions perform the following tasks: (i) identify a subset of patients within the dataset who have been treated with the antiviral agent of interest; (ii) determine for each patient when therapy started and ended; (iii) identify for each patient all the samples that have been sequences analysed and were collected pre- and

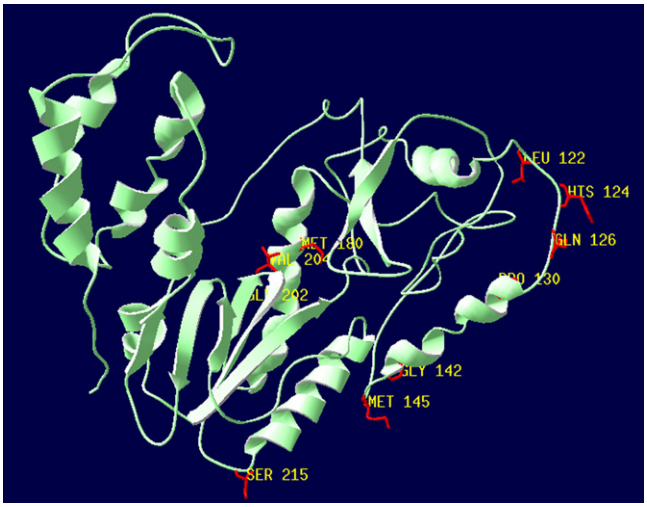


Fig. 5. A graphical representation of the 3D HBV reverse transcriptase model showing the integration of all the amino acid variations that were detected by SeqHepB for a HBV genome sequence. The model was viewed with DeepView (Swiss-Pdb Viewer V3.7, Swiss Institute of Bioinformatics, Geneva). The backbone and side-chains of the amino acids variations found by SeqHepB are shown in red and their locations within the protein are shown in yellow.

during-therapy, and calculate time on therapy for each sample; (iv) identify for each sample the corresponding HBV viral loads and deduced amino acid variations detected by SeqHepB within the rt; (v) generate a spreadsheet that separately lists the pre- and during-treatment datasets. Data extracted via this method can then be analysed with statistical and or specialised data exploratory programs to identify novel relationships between those factors that are statistically significant.

3.2.3. Protein modelling function

Embedded within the system is a script that links the relational database to a macromolecule analysis application DeepView (Swiss-PdbViewer V3.7, Swiss Institute of Bioinformatics, Geneva). This application enables the viewing of a three-dimensional (3D) structure of the HBV polymerase model that was deduced from a low resolution crystal structure of the rt of HIV at 3.2 Angstrom (Bartholomeusz et al., 2004b). This integration facilitates the localisation of all amino acid variations found by SeqHepB for a sample onto the 3D HBV polymerase model (Fig. 5). DeepView has functions that enable the macromolecular model to be rotated and viewed at any angle, thus providing the ability to develop greater understanding of the effects that the amino acid variations can have on the HBV polymerase.

3.2.4. Future development

Development of the relational database system is ongoing. Functions and SQL queries are regularly created, and if necessary updated, to enable better methods for analysing and mining the data within the database. For example, a suite of reports is currently being developed for simple statistical analysis on patients involved in specific study cohorts, and GUI interfaces are being developed to enable users to perform motif and epitope searches on all the sequences analysed by SeqHepB.

4. Conclusion

CHB is a disease with a complex natural history, and this complexity is increasing with the use of antiviral agents mainly due to the development of resistance. Thus, it is necessary to develop software systems to aid patient management. The main application of a relational database is for it to be mined effectively, and to correlate data between clinical, routine pathology tests and detailed viral sequence and drug sensitivity values. With the amount of data housed in the SeqHepB database, this can be accomplished as a result of advances made on machine learning technologies. Knowledge discovered with these technologies should lead to individualized patient management and improve the treatment of patients on antiviral therapy.

Acknowledgements

The authors would like to thank Geoff Thompson and Nadia Warner for their contributions on the processing of HBV genome sequence data.

References

Angus, P., Vaughan, R., Xiong, S., Yang, H., Delaney, W., Gibbs, C., Brosgart, C., Colledge, D., Edwards, R., Ayres, A., Bartholomeusz, A., Locarnini, S., 2003. Resistance to adefovir dipivoxil therapy associated with the selection of a novel mutation in the HBV polymerase. Gastroenterology 125, 292–297.

Ayres, A., Locarnini, S., Bartholomeusz, A., 2004. HBV genotyping and analysis for unique mutations. In: Hamatake, R.K., Lau, J.Y.N. (Eds.), Hepatitis B and D protocols. Humana Press, Totowa, New Jersey, pp. 125–149.

Bartholomeusz, A., Locarnini, S.A., Ayres, A., Thompson, G., Sozzi, V., 2004a. Molecular modelling of hepatitis b virus polymerase and adefovir resistance identifies three clusters of mutations. Hepatology 40 (4 Suppl. 1), 246A.

Bartholomeusz, A., Tehan, B.G., Chalmers, D.K., 2004b. Comparisons of the HBV and HIV polymerase, and antiviral resistance mutations. Antivirus Ther. 9, 149–160.

Benhamou, Y., Bochet, M., Thibault, V., Di Martino, V., Caumes, E., Bricaire, F., Opolon, P., Katlama, C., Poynard, T., 1999. Long-term incidence of hepatitis B virus resistance to lamivudine in human immunodeficiency virus-infected patients. Hepatology 30, 1302–1306 (see comments).

Delaney, W.E., Yang, H., Westland, C.E., Das, K., Arnold, E., Gibbs, C.S., Miller, M.D., Xiong, S., 2003. The hepatitis B virus polymerase mutation rtV173L is selected during lamivudine therapy and enhances viral replication in vitro. J. Virol. 77, 11833–11841.

Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P., Charnay, P., 1979. Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in *E. coli*. Nature 281, 646–650.

Gaschen, B., Kuiken, C., Korber, B., Foley, B., 2001. Retrieval and on-the-fly alignment of sequence fragments from the HIV database. Bioinformatics 17, 415–418.

Ling, R., Mutimer, D., Ahmed, M., Boxall, E.H., Elias, E., Dusheiko, G.M., Harrison, T.J., 1996. Selection of mutations in the hepatitis B virus polymerase during therapy of transplant recipients with lamivudine. Hepatology 24, 711–713.

Locarnini, S., 2003. Hepatitis B viral resistance: mechanisms and diagnosis. J. Hepatol. 39 (Suppl. 1), S124–S132.

Locarnini, S., Hatzakis, A., Heathcote, J., Keeffe, E.B., Liang, T.J., Mutimer, D., Pawlotsky, J.M., Zoulim, F., 2004. Management of antiviral resistance in patients with chronic hepatitis B. Antivirus Ther. 9, 679–693.

Margolis, H.S., Alter, M.J., Hadler, S.C., 1991. Hepatitis B: evolving epidemiology and implications for control. Semin. Liver Dis. 11, 84–92.

Niesters, H.G., De Man, R.A., Pas, S.D., Fries, E., Osterhaus, A.D., 2002. Identification of a new variant in the YMDD motif of the hepatitis B virus

polymerase gene selected during lamivudine therapy. J. Med. Microbiol. 51, 695–699.

Norder, H., Courouce, A.M., Magnius, L.O., 1994. Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. Virology 198, 489–503.

Rhee, S.Y., Gonzales, M.J., Kantor, R., Betts, B.J., Ravela, J., Shafer, R.W., 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res. 31, 298–303.

Stuyver, L.J., Locarnini, S.A., Lok, A., Richman, D.D., Carman, W.F., Dienstag, J.L., Schinazi, R.F., 2001. Nomenclature for antiviral-resistant human hepatitis B virus mutations in the polymerase region. Hepatology 33 (751–757), 757.

Tenney, D.J., Levine, S.M., Rose, R.E., Walsh, A.W., Weinheimer, S.P., Discotto, L., Plym, M., Pokornowski, K., Yu, C.F., Angus, P., Ayres, A., Bartholomeusz, A., Sievert, W., Thompson, G., Warner, N., Locarnini, S., Colonno, R.J., 2004. Clinical emergence of entecavir-resistant hepatitis B virus requires additional substitutions in virus already resistant to Lamivudine. Antimicrob. Agents Chemother. 48, 3498–3507.

Torresi, J., Earnest-Silveira, L., Deliyannis, G., Edgtton, K., Zhuang, H., Locarnini, S.A., Fyfe, J., Sozzi, T., Jackson, D.C., 2002. Reduced antigenicity of the hepatitis B virus HBsAg protein arising as a consequence of sequence changes in the overlapping polymerase gene that are selected by lamivudine therapy. Virology 293, 305–313.