AIDS Rev 2003;5:52-61

HIV Sequence Databases

Carla Kuiken[1], Bette Korber[1] and Robert W. Shafer[2]

Los Alamos National Laboratory, Los Alamos, New Mexico, USA[1], Division of Infectious Diseases, Department of Medicine, Stanford University, Stanford, CA, USA[2]

Abstract

Two important databases are often used in HIV genetic research, the HIV Sequence Database in Los Alamos, which collects all sequences and focuses on annotation and data analysis, and the HIV RT/Protease Sequence Database in Stanford, which collects sequences associated with the development of viral resistance against anti-retroviral drugs and focuses on analysis of those sequences. The types of data and services these two databases offer, the tools they provide, and the way they are set up and operated are described in detail.

Key words

HIV. Database. Analysis. Resistance. Genetic sequences. Evolution.

Background

UNAIDS (http://www.unaids.org/) estimates 40 million people are living with HIV, and 5 million new HIV infections and 3 million AIDS-related deaths are expected to occur in the year 2002. The extraordinary variability of the virus has been one of the many major obstacles for vaccine and drug development. Expensive and complicated regimens consisting of combinations of three or more drugs are essential for effective treatment, as drug resistant variants rapidly render single drugs useless. Candidate vaccines that induce broadly cross-reactive immunogenic responses have yet to be developed, and an effective vaccine is not expected for at least another five to seven years.

Two databases have been created to help researchers cope with the enormous amount of data generated by studies of HIV variation, which not only aid drug and vaccine development but also provide insight into the basic biology, immunology and evolution of other pathogens. The Los Alamos HIV Sequence Database provides easy access to all published sequences and their accompanying annotations and offers a multitude of tools and programs for their analysis. The Stanford Protease and RT Database allows users to screen their own sequences for resistance-conferring mutations and compare them to stored sequences. Both databases are described in greater detail in this paper.

The Los Alamos HIV sequence database

It became clear in the early 1980s that HIV was an unprecedented pathogen with great potential for variation. At that time Dr. Gerald Myers conceived of an easily accessible centralized database to store HIV sequence data. In its earliest form, the database was simply a yearly compendium of HIV sequence data containing entire Genbank entries. As the database evolved, summaries, alignments and analyses of the data were added, and a relational database of HIV/SIV sequences and annotation was developed. The goal has been to format available data in such a way that new results could be readily integrated with those from earlier studies.

In the last five years an extensive, interactive website has been developed that provides researchers with flexible retrieval tools for sequenc-

Correspondence to:
Carla Kuiken
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
E-mail: kuiken@lanl.gov

es and background data, as well as online facilities for data analysis (http://www.hiv.lanl.gov/). The ready-made alignments provided by the database have long been used for the development of PCR primers and diagnostic kits, molecular epidemiology studies and vaccine design.

Since its inception, the database staff has responded to new research developments, advancing web technologies, and the evolving needs of the research community by creating tools and providing information helpful to the understanding of the implications of HIV-1 variation. For example, the realization in the mid-1990s that HIV recombinants are common[1] elicited a series of responses from the group: they devised general strategies to identify contamination; worked with the research community to increase awareness of the issue[2]; developed the first web-interactive rapid recombination screening tool, RIP[3]; and created a standardized numbering system for nucleotides and proteins[4]. The group also facilitated the development and implementation on the web site of a uniform nomenclature system[5]. With the ongoing help of Dr. John Mellors the group also tracks drug resistance mutations[6] and links sequences to information about HIV co-receptors used for viral entry[7]. Recently, the group published a study to assist in the selection of representative strains that could be used as a basis for vaccine development[8].

In 1995 a web-accessible immunology database was created to summarize and track variability in cytotoxic T-cell, helper T-cell, and antibody epitopes. Automated interfaces between the sequence and immunology databases map the epitopes onto proteins. A third database containing detailed information extracted from the literature about anti-HIV compounds and resistance-related mutations in the HIV genome was created in 1997. This database is distinct from the Stanford Protease and RT database, which contains sequences of RT and protease. A fourth database that tracks animal vaccine trials was added in 2001. The first publicly available version is based on a database created for NIAID that tracked primate vaccine trials conducted in the 1990s[9]. The ultimate goal is to integrate these four databases and cross-link the available information in the framework of vaccine trials, drug development, and evolutionary studies.

The complexity of HIV sequences

HIV belongs to the lentivirus family, whose members infect many species of African primates. Understanding the genetic relationships between the primate viruses[10] will provide insight into the origin of HIV-1 from chimpanzee[11] and HIV-2 from sooty mangabey[12]. The Los Alamos sequence database maintains and annually updates basic alignments of representative sequences of the vastly divergent SIVs, as well as alignments of HIV-1 and related chimpanzee viruses.

HIV-1 comprises three major groups –M, N, and O– which can be as much as 40% divergent. The M group is divided into subtypes A-K[5], and inter-subtype recombination has been shown to be very common in geographic regions with a high prevalence of more than one subtype. Fifteen circulating recombinant forms, or CRFs, have been identified to date. CRFs 01 and 02 have been responsible for major regional epidemics in Asia and Africa, respectively[13,14], and these recombinant lineages are as important for vaccine and molecular epidemiology purposes as the known HIV-1 subtypes. The Los Alamos sequence database tracks the geographic origins of the viral subtypes for molecular epidemiology purposes and Dr. Brian Foley maintains a full-length genome alignment of representative examples of each subtype and CRF.

Tasks of the Los Alamos HIV sequence database

Compiling and annotating sequences

The Los Alamos database is fed by biweekly downloads of HIV-1, HIV-2 and SIV sequences from Genbank, which (with the exception of HIV sequences generated during drug resistance studies) contains most published HIV sequences. Currently there are 79,327 HIV and SIV sequences in the Los Alamos sequence database.

The HIV database group is committed to the use of Open Source software whenever practical and feasible; the website runs on a Linux-based Apache web server, the database on a Linux-based Sybase Adaptive Server Enterprise database server. Website development is primarily done using Mason, an open source solution that allows Perl code to be embedded inside the HTML and encourages the construction of web pages with reusable components.

For security reasons there are actually two sets of database servers and websites: one publicly available and one behind a firewall, accessible only from inside the LANL network and annotated and updated continuously by the database staff. Changes to the website and database are copied from the internal site to the publicly accessible site every evening. This arrangement assures the reliability of the external data and would allow the external site to be re-created from scratch if necessary.

The database is updated from GenBank entries with a series of parsing scripts that extract available information such as subtype, country of isolation, year of isolation and isolate number. The database staff then annotates the sequences with supplemental information from the literature and from direct communication with the authors, who can frequently supply additional critical information not included in the GenBank entry from the initial and subsequent publications.

Not all sequences are annotated; priorities are set based on the length of the sequences (>280 nucleotides), whether they span a complete gene, and whether they represent an unusual subtype or a rarely sampled geographical region. Particular emphasis is placed on complete annotation of full-length viral genome sequences, as these are often used as key reference strains.

Annotations frequently include sampling date, sampling country, and subtype information. Subtyping information has been confirmed or assigned for over 45,000 sequences by in-house testing against our reference sets using neighbor joining trees. Annotations relating to sibling sequences; possible contamination; patient information; viral phenotype, isolation and passage history; and other features have also been added to hundreds of important sequences in the database. Recently the HIV sequence database has also begun to provide access to the patient information, including the patient's sex, risk group or route of infection, epidemiological linkage to other patients in the database, and time and place of infection. Sequence sets from the same patient are identified so they can easily be retrieved or excluded. As in the primary publications, extraordinary care is taken to avoid including any information that could be used to identify individual patients.

Occasionally, authors will submit sequences prior to publication for the express purpose of making them rapidly available to the public. For example, the early full-length HIV-1 genomic sequences representing the new circulating recombinant forms CRF07 and CRF08 from China were provided before publication by Dr. Francine McCutchan for inclusion in our subtype reference set. Early submission also allows the database staff to assist with classification and to prevent naming conflicts in new subtypes and CRFs.

Annotation of related data

The database staff annually compiles and annotates related data that may have relevance to the pathology and immunological recognition of

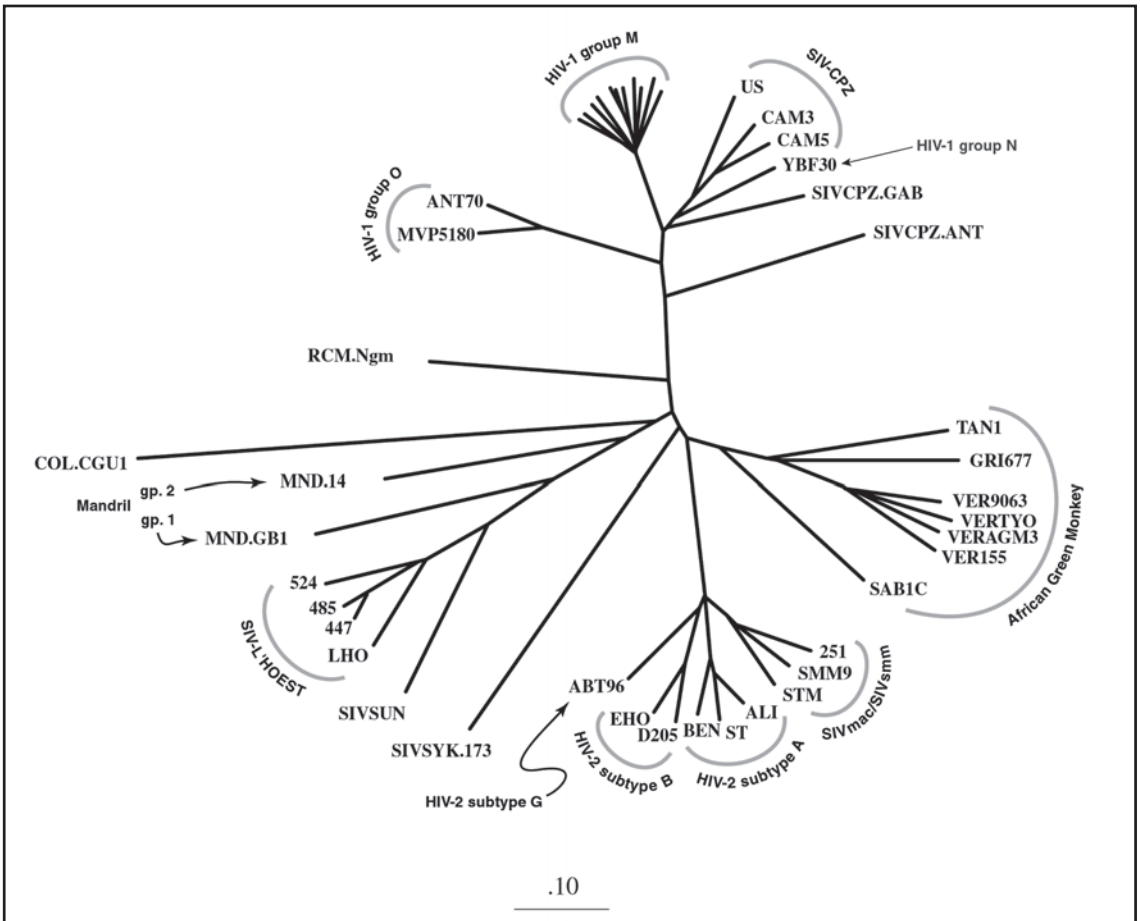

Figure 1. *Primate lentivirus tree. This phylogenetic tree illustrates the basic relationships of different primate lentiviruses, and is based on a pol gene alignment. The human HIV-1s and HIV-2s are the most closely related to SIVs found in chimpanzee and sooty mangabey. The HIV M, N, and O groups are noted, and representative sequences from different HIV-1 M group subtypes are included, each subtype is a set of genetically associated viruses. The map on the left shows the region where chimpanzees carrying SIVs have primarily been found, although this search is being extended through non-invasive methods to test wild chimpanzee populations.*

the isolate. The sequence database itself presently contains several fields relevant to the biological properties of the virus isolate:

Phenotype: NSI/SI
Coreceptor usage: CCR5, CXCR4, CCR2, CCR3, etc.
Culture method: none, PBMC, T-cell lines
Culture duration: field isolate, primary culture, expanded stock
Sequencing method: PCR or biological clones, direct sequencing
Molecule type: DNA, RNA
Source material: PBMC, serum, tissue type

When available, the age, sex, risk factor or route of infection, and health status of the patient at the moment of sampling are also stored, and since early 2002 most of this information can be accessed via the new search interface.

HIV reference alignments

The website contains whole genome alignments and background information for hundreds of sequences of HIV-1, HIV-2 and SIV. Alignments for each protein are also available. The alignments, which are updated annually, are annotated and printed in the annual sequence compendium. Only one sequence per patient is included in the alignments to avoid biasing the data with large sequence sets from a few sources.

Although frameshifts, insertions and deletions are relatively common, sequences are codon-aligned as much as possible. Because sequences with inactivating frameshifts may result from PCR error or may represent a virus that is only one replication away from a viable ancestor, they are still considered to contain valuable sequence information and are included in the alignments when possible. Reference alignments are also generated, consisting of small subsets of sequences that are representative of all subtypes and circulating recombinant forms (the subtype reference sets).

In addition to these ready-made alignments, the search interface of the HIV sequence database now allows users to download alignments of specified regions in addition to unaligned sequence sets. This feature is based on HIV-MAP, which allows users to create alignments of pre-defined and user-defined gene regions[15]. The user can retrieve sequences based on country, continent, subtype and gene region (for example, all complete A subtype envelope sequences from Kenya) –an important feature that allows the researcher to locate the region of interest in a longer stretch of sequence. The tool accomplishes this by using pre-calculated coordinates from a pairwise alignment of all sequences against a model sequence using a Hidden Markov Method approach. The coordinates can then be used to calculate the location of any specific stretch, and this stretch can be cut out and aligned to other sequences. The alignments are not always optimal, but they can be hand-edited further. Presently this feature is only offered for HIV-1.

Researchers working on HIV vaccines have requested a set of subtype "model" sequences for the HIV-1 M group consisting of consensus sequences and reconstructed most recent common ancestor sequences. These sequences, which were recently added to the database, can be used for as models for vaccine preparation as well as "centers" of the theoretical subtype distribution.

Analysis programs at the HIV sequence database web site

Through the years, the HIV database has developed a number of interactive programs to facilitate basic research, responding to the needs in the field as they arose. All these programs can be accessed via interfaces on the web site.

The SNAP interface allows rapid analysis of synonymous/non-synonymous mutation patterns. It produces graphical as well as tabulated output, allows the generation of phylogenetic trees on the basis of the synonymous or non-synonymous distances obtained, and calculates statistics –including the confidence intervals of the observed synonymous or non-synonymous substitution frequencies. This is a simple strategy for tracking regional differences in synonymous and non-synonymous substitutions to look for positive selection; more rigorous methods have recently been developed and applied to HIV-1[16].

The Recombination Identification Program (RIP) was developed to identify genetic sequences with regions similar to more than one phylogenetic clade. Known as mosaics, these sequences are most likely recombinants between different viruses that represent an important source of variation in HIV-1. RIP was first designed to detect recombinants of sequences with regions belonging to distinct subtypes of HIV-1, but it can also be used for other applications, including the analysis of non-HIV sequences or of any sequence set with distinctive phylogenetic clades that could be used as reference strains[3]. An overhaul of RIP to make it more user-friendly and less prone to problems is ongoing; coming improvements include a more intuitive user interface, an increase in the number of useable different distance measures (in addition to the standard, same-or-different "p"), better quality graphics, and the optional use of error bars. In addition, two excellent and frequently used programs available on the web are SimPlot by Stuart Ray (http://www.med.jhu.edu/deptmed/sray/download/) which also creates a "bootscan" plot of a query sequence of interest, and a rapid BLAST based program (http://www.ncbi.nlm.nih.gov/retroviruses/subtype/subtype.html) written by Uwe Plikat and Colombe Chappey that can be used to screen new sequences for subtype and recombination information prior to aligning them.

HYPERMUT is designed to highlight hypermutational changes against the background of other base mutations. It documents the nature and context of nucleotide substitutions in a nucleotide

alignment relative to a reference sequence. HYPERMUT contains 4 parts: a data sheet summarizing the hypermutations, a graph providing an illustrated overview of all the sequences and their nucleotide changes, a graph depicting all mutations in a selected sequence, and a table for quick analysis of stop codon mutations[17].

SUDI can be used to help determine if a newly defined clade of related sequences qualifies as a new subtype or sub-subtype, or if the clade is simply part of a previously defined subtype. It calculates tree-based genetic distances among a new cluster of sequences and known subtypes and compares them to typical distances found among pre-existing subtypes. This tool is designed according to specifications of the HIV nomenclature committee[5] to be used in conjunction with other tools that will identify possible new subtypes and sub-subtypes, as well as to clarify phylogenetic relationships.

ADRA can be used to scan inputted sequences or alignments and list all the codons that contain a mutation that might be associated with resistance to an anti-HIV compound. Links to the resistance database are provided for more information about the mutations that are found. The program is embedded in warnings against use as a basis for clinical decisions.

ADRA also provides access to an interface to RASMOL, a protein 3D structure viewer that can be used to view and rotate proteins. The interface also allows users to identify the positions of amino acids associated with drug resistance.

The Sequence Locator Tool uses a reference alignment to quickly assign position numbers to a HIV or SIV sequence. The tool accepts both amino acid and nucleotide sequences and can usually automatically recognize if the sequence is from HIV-1, HIV-2 or SIV. The tool was designed to standardize numbering for primers, peptides and proteins, which has been a source of great confusion in HIV and SIV research in the past[18].

HIV-BLAST performs a rapid BLAST search on the HIV sequence database alone. The output is organized to facilitate the detection of contamination[2].

Primalign automatically aligns a primer or sequence fragment to the HIV-1 complete genome alignment and returns the coordinates and the alignment of the fragment to the user.

Geography plots an overview of the numbers of sequences with different subtypes found by region onto geographical maps; the resolution of the maps and the type of variation plotted can be set by the user. The program also prints tables of these data that can be imported into a different plotting program. The current version of the pro-

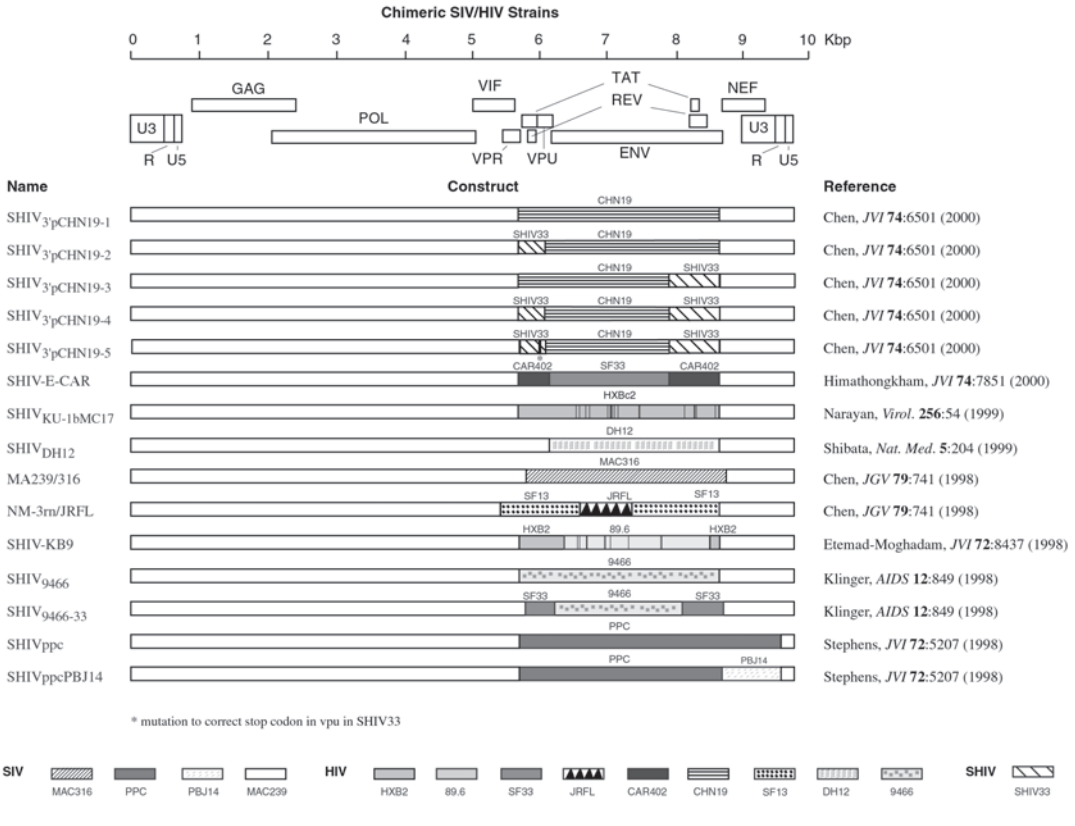

Figure 2. *Vaccine constructs. A subset of the SIV and SHIV (combined SIV-HIV) constructs that are being used for SIV vaccine trials, which are often complex products of molecular engineering, are shown. Each different color or pattern represents a different clone that has contributed. A more complete listing is available at the database.*

gram is not dynamic, but an upgrade that will pull data from the database every time it is invoked is expected to come online in the spring of 2003.

TreeMaker is a very simple interface to some of the PHYLIP (http://evolution.genetics.washington.edu/phylip.html) programs. It can be used to generate a tree from an alignment. It rectifies certain commonly made mistakes, such as forgetting to gapstrip. At this time the interface does not allow bootstrapping, which is too computationally intensive and would overburden the server.

SeqPublish formats user alignments into versions suitable for publication: identical columns are replaced by dashes, and the sequences are printed in blocks of user-determined length.

The website provides an easy-to-use interface to PCOORD, a program for finding multidimensional patterns or groups in sequence data[19].

Hdent is a program that can quantify heteroduplex mobility shift (HMA and HTA) data for statistical comparisons of genetic diversity[20].

Web tutorials

The HIV Database and Analysis Project staff has used its combined experience in HIV sequence analysis to create a series of short tutorials on the subject. Three tutorials have been published so far: one on HIV nomenclature, one on the use of interactive tools and web examples to recognize and deal with contamination, and one that utilizes a tree-making web interface to teach the basics of building phylogenetic trees. Additional tutorials on alignments, consensus sequences, and genetic distance calculations are planned.

Future directions

Advances in sequencing technology and immunological methods now make it feasible to acquire full-length viral genome sequences in conjunction with detailed immunological data for large cohorts and complex host genetic data at a population level. This progress will probably lead to a continued rapid expansion of the number of sequences, along with the accompanying background information. Although the use of automated processes to download new sequences from Genbank has made the acquisition of sequence data almost trivial, entering the epidemiological data is still very labor intensive, and will become a major burden as the number of new sequences expands. The Los Alamos HIV database project is initiating collaborations with groups worldwide that are planning to generate large amounts of sequence data to enable direct computerized links to any accompanying background information and enhance the efficiency of data acquisition.

These groups will be offered the opportunity to store their own sequences in a private "corner" of the database before publication. Upon entry of a username and password, the group's sequences will appear to as a part of the regular database and they will be available for analysis as a subset or in conjunction with the rest of the sequences. However these sequences will remain hidden from all other users. This set-up will also enable groups of collaborators in different locations to store their sequences in a central location and ensure that the sequences they use can be updated and corrected by all the participants. When the sequences are ready to be made public, an internal parameter in the database will be reset and the sequences will immediately become part of the public database.

The Los Alamos HIV sequence database staff is also hoping to add patient's HLA type to other patient data in the near future. Analyzing sequences for CTL variation is much more accurate and meaningful if only the epitopes from patients with the appropriate HLA types can be selected. While HLA type is presently recorded for only a small percentage of the patients, the increased realization of the importance of CTL and CTL epitope variation will most likely result in a much greater availability of this data.

HIV RT and Protease Sequence Database

Database rationale

HIV drug resistance is caused by mutations in the molecular targets of HIV therapy: protease, reverse transcriptase (RT) and gp41. Understanding the genetic basis of drug resistance has therefore become essential for the detection of drug resistant viruses, the use of antiretroviral drugs in clinical settings, and the design of new antiviral drugs that are less likely to trigger resistance or are effective against previously resistant HIV strains.

A large number of retrospective and prospective studies have demonstrated that the presence of drug resistance before starting a treatment regimen is an independent predictor of success of that regimen. Several expert panels have recommended that HIV RT and protease sequencing be done to help physicians select antiretroviral drugs for their patients and, in several countries, genotypic resistance testing has been part of routine clinical care[21].

There have been several comprehensive reviews of HIV drug resistance testing and of the biological and clinical significance of HIV drug resistance mutations[22-27]. In addition, the Los Alamos HIV Sequence Database maintains a comprehensive listing of all published HIV drug resistance mutations[6]. This listing contains also contains data on mutations conferring resistance to experimental compounds (e.g. fusion and integrase inhibitors).

The HIV RT and Protease Sequence Database (HIVRT & PrDB; http://hivdb.stanford.edu) was designed to represent, store, and analyze the

diverse forms of data underlying drug resistance knowledge and to make these data available to researchers studying HIV drug resistance and to clinicians using HIV drug resistance tests. The database catalogs data linking sequence changes in the molecular targets of HIV therapy to other forms of data including treatment history, phenotypic (drug susceptibility) data, and clinical outcome (plasma HIV RNA levels and CD4+ cell counts) data[28]. These data are necessary to create the four major correlations that provide the basis for interpreting genotypic tests in clinical settings (Table 1).

Database content

The HIVRT & PrDB contains data from > 420 published papers. RT and protease sequence data are available on "clinical" isolates from > 6,000 individuals at > 8,000 time points. The database also contains sequences of > 500 "laboratory" isolates. Laboratory isolates contain mutations generated by virus passage or site-directed mutagenesis. About 15,000 drug susceptibility results are available on about 2,000 isolates. Clinical isolates can be linked to the treatment of the person from whom the isolates were obtained or to drug susceptibility results. Laboratory isolates can be linked only to drug susceptibility results. Data are added approximately every two weeks. A web page has recently been added which describes recently added data and documents other changes to the database and web site.

The goals and structure of the HIVRT & PrDB allow the active recruitment of data. In addition to obtaining sequence data from GenBank, the database staff solicits data from those publishing important papers on drug resistance. The solicited data, which are often available only as amino acid sequences, contain much of the most clinically relevant data in the database. These amino acid sequences, as well as those translated from nucleic acid data are stored in a virtual alignment with the subtype B consensus sequence, allowing them to be represented as a list of amino differences from the consensus. Insertions are stripped out of the sequence and stored in a separate table.

Database design

The HIVRT & PrDB is a relational database with 19 normalized (nonredundant) core tables, 10 look-up tables, and about 20 derived tables. The database is implemented using MySQL on a Linux platform. Data from the database are made available over the web using an Apache server running Perl CGI scripts. To increase performance, the scripts obtain data predominantly from the derived tables. The database can be accessed using Netscape versions 4.0 and above and Internet Explorer versions 5.0 and above; it is likely that most HTML 2.0 and javascript compliant browsers can also be used.

There are several major hierarchical relationships linking key entities in the database: (i) patient → treatment history (list of drug regimens and their start and stop dates); (ii) patient → isolate (clinical) → sequence → drug susceptibility result; (iii) isolate (laboratory) → drug susceptibility result. The database design also includes tables for plasma HIV RNA levels and CD4+ cell counts but such data have begun to be added only recently. The patient is represented by an anonymous code name and is not linked to any form of identifying information.

Table 1. *Major correlations linking mutations to HIV drug resistance*

Correlation	Contribution to knowledge
Genotype - phenotype (laboratory isolates)	(a) Assess the impact of mutations in an isogenic background (b) Isolates derived from virus passage or site-directed mutagenesis
Genotype - phenotype (clinical isolates)	(a) Clinical isolates contain more relevant combinations of drug resistance mutations (b) The complexity of sequences obtained on clinical isolates often precludes site-directed mutagenesis
Genotype - treatment history	(a) Sequences of HIV-1 isolates from patients failing antiretroviral therapy are crucial observations of HIV evolution that show which virus mutations are most significant *in vivo* (b) Such data are also essential for elucidating the genetic mechanisms of resistance to drugs for which susceptibility testing is not fully reliable (e.g. stavudine, didanosine)
Genotype - clinical outcome	(a) Possibly the most relevant information for clinicians, however, few data of this type are publicly available (b) May provide insight into those mutants which are most fit *in vivo* (c) May provide insight into resistance to drugs for which susceptibility testing is not fully reliable (d) Clinical outcome data (plasma HIV RNA levels and CD4+ cell counts) are now being added to the HIV RT and Protease Sequence Database

Database search interface

There are seven web pages that allow users to retrieve sets of sequences meeting specific criteria. On the "Protease inhibitor" and "RT inhibitor" pages, users specify the number, names and combinations of drugs received and whether the results should be constrained to isolates of a specific subtype. On the "Protease mutations" and "RT mutations" pages, users can specify one or more protease or RT mutations and optional subtype requirements. On the "Protease inhibitor susceptibilities" and "RT inhibitor susceptibilities" pages, users can specify one or more mutations, one or more drugs, and one or more drug susceptibility assays. On the "References" pages users can select isolates from more than 400 published studies by specifying author names.

Each query initially returns data in the form of a table of records containing 8 or more columns of data. The data returned include (i) hyperlinks to the MEDLINE abstract and GenBank record of the sequence, (ii) a list of mutations in the sequence, (iii) a classification of the sequence by patient and time point, (iv) the patient's drug treatment history, and (v) additional data depending upon the query. The Protease and RT inhibitor susceptibility queries return a written summary of drug suscep-

Table 2. *Types of queries possible using the HIV RT and Protease Sequence Database*

Page	Criteria (examples)	Results
Protease inhibitors	No treatment + subtype B No treatment + subtype C	Natural variability of HIV-1 subtype B protease Natural variability of HIV-1 subtype C protease[39]. Similar queries are also possible for RT and each of the group M subtypes
	NFV + one PI + subtype B NFV + one PI + subtype C	Mutations in patients with subtype B isolates receiving NFV as their sole PI. Similar queries are possible for each of the other PIs and each of the other subtypes. However > 90% of published treated isolates are from persons with subtype B virus
	RTV + SQV (any subtype)	Mutations associated with the combination of RTV + SQV
Protease inhibitors or RT inhibitors	≥3 PIs ≥4 NRTIs	Mutations observed in patients failing therapy with multiple PIs or NRTIs. Will identify many mutations not described in the literature, because much of what is in the literature is based on sequences from monotherapy studies
Protease mutations	V82I	Occurs predominantly in isolates from untreated patients particularly in non-subtype-B isolates. This information together with data from the phenotypic query page suggests that this mutation, although it is in the substrate cleft, does not cause drug resistance
	I54T	Nearly all isolates with I54T have been found using HIVRT & PrDB to have G48V and I84V[40]
RT mutations	K103N, K103R	K103N occurs nearly always in persons receiving NNRTIs. K103R occurs predominantly in individuals who have not received NNRTIs indicating that it probably does not cause drug resistance. This can be confirmed entering K103R into the susceptibility query page
	T69D, T69N	Isolates with these mutations have been found using HIVRT&PrDB to have been obtained from persons receiving many different NRTIs, not just ddC[41], the drug with wich the mutation was first associated
	V75I + F77L + F116Y + Q151M	Returns all the published with this combination of mutations (the "multinucleoside resistance" pattern)
Mutation ("PI susceptibility" or "RT inhibitor susceptibility" page)	RT: M41L + L210W + T215Y (drug: d4T, assay: ViroLogic)	Returns 54 d4T susceptibility results on isolates having at least 41L, 210W, and 215Y (28 results are obtained if the assay checked is Virco)
	RT: M41L + M184V + L210W + T215Y (drug d4T, assay: ViroLogic)	Returns 46 d4T susceptibility results on isolates having at least 41L, 184V, 210W, and 215Y (5 results are obtained if the option "No other major mutations" is checked)
	PR: M46I + I84V + L90M (all PLs, all assays)	Determine which PR inhibitors have the least antiretroviral activity (have the highest fold-resistance compared with wild type)

PI: protease inhibitor; **NRTI**: nucleoside RT inhibitor; **NNRTI**: non-nucleoside RT inhibitor; **NFV**: nelfinavir; **RTV**: ritonavir; **ddC**: zalcitibine; **d4T**: stavudine; **SQV**: saquinavir

Data in this table are based on the August 2002 release of HIVRT&PrDB.

tibility results on isolates containing the specified mutations, as well as a graphic representation of the results for each drug and for each drug susceptibility test method. In addition to the tabular results, users have the option of downloading or viewing the raw sequence data in a variety of formats. Table 2 shows examples of the results returned by specific queries.

Sequence interpretation programs

The database website contains three sequence interpretation programs and a program for calculating synonymous and nonsynonymous mutation rates between sequences containing nucleotide mixtures[29]. The first sequence interpretation program, HIVseq, accepts user-submitted RT and protease sequences, compares them to a reference sequence, and uses the differences (mutations) as query parameters for interrogating the database[30]. HIVseq allows users to examine new sequences in the context of previously published sequences, providing two main advantages. First, unusual sequence results can be detected and immediately rechecked. Second, unexpected associations between sequences or isolates can be discovered when the program retrieves data on isolates sharing one or more mutations with the new sequence.

The second program, a drug resistance interpretation program (HIVdb), accepts user-submitted protease and RT sequences and returns inferred levels of resistance to the 16 FDA-approved antiretroviral drugs. Each drug resistance mutation is assigned a drug penalty score; the total score for a drug is derived by adding the scores associated with each mutation. Using the total drug score, the program reports one of the following levels of inferred drug resistance: susceptible, potential low-level resistance, low-level resistance, intermediate resistance, and high-level resistance.

The third program (HIValg), allows researchers to compare the output of different publicly available drug-resistance algorithms on the same sequence or set of sequences. The algorithms used by this program are encoded using a programming platform or Algorithm Specification Interface (ASI) developed to facilitate the comparison of HIV genotypic resistance algorithms. ASI consists of an XML format for specifying an algorithm and a compiler that transforms the XML into executable code.

New additions planned for the HIVRT & PrDB

Two additions to the database are planned: (i) links to clinical outcome data, and (ii) gp41 sequences and data on resistance to fusion inhibitors. The HIVRT & PrDB is now collecting serial plasma HIV RNA levels and CD4+ cell counts in persons from whom sequence data and treatment histories are available. A browsable form of this data will be available by 2003. The first fusion inhibitor, enfuvirtide (T-20) has been shown to have potent antiretroviral activity in both phase I/II[31] and phase III[32,33] clinical trials. This drug is likely to be approved in 2003. A wide range of mutations in gp41 contributing to T-20 resistance, most occurring between residues 36-45, have been reported, but mutations outside of this region also appear to contribute to drug resistance.

Most sequences in the database currently belong to HIV-1 subtype B, the most prevalent subtype in the United States and Europe. However, because subtype B makes up only a small proportion of worldwide HIV-1 infections the WHO and International AIDS Society are establishing international collaborative networks to compile resistance data on non-B viruses. HIVRT & PrDB staff are also collaborating with other investigators to enrich the database with sequences of non-B isolates from treated and untreated persons[34]. It is currently not known whether drug resistance interpretation programs should be modified for use with non-B sequences.

Important drug resistance data not included in HIVRT & PrDB

The HIVRT & PrDB accepts sequences only from patients developing virologic failure and as such does not provide data on the rate at which resistance occurs with different treatment regimens or the frequency with which resistant variants are transmitted. These two types of data each require a different type of database design. The determination of resistance rates would require the collection of information from all persons receiving treatment –not just those developing virologic failure (and undergoing virus sequencing). Transmitted resistance is a growing problem in the United States and Europe[35-38]. Ascertaining the frequency of transmitted resistance requires the capture of additional information such as whether a patient is undergoing seroconversion, has had a positive detuned serological assay (indicating recent infection), or knows the year in which HIV was acquired.

References

1. Robertson D, Gao F, Hahn HB, Sharp P. Intersubtype recombinant HIV-1 sequences. In: Korber B, Hahn B, Foley B, et al. (eds). Human Retroviruses and AIDS 1997. Los Alamos: Los Alamos National Laboratory 1997.
2. Kuiken C, Korber B. Sequence quality control. In: Korber B, Kuiken CL, Foley B, et al. (eds). Human Retroviruses and AIDS 1998. Los Alamos: Los Alamos National Laboratory 1998.
3. Siepel A, Halpern A, Macken C, Korber B. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. AIDS Res Hum Retroviruses 1995;11:1413-6.
4. Korber B, Foley B, Kuiken C, Pillai S, Sodroski J. Numbering Positions in HIV Relative to HXB2CG. In: Korber B, Kuiken C, Foley B, et al. (eds). Human Retroviruses and AIDS. Los Alamos: Los Alamos National Laboratory 1998.

5. Robertson D, Bradac J, Foley B, et al. HIV-1 nomenclature proposal. Science 2000;5463.
6. Parikh U, Hammond J, Calef C, Larder B, Schinazi R, Mellors J. Mutations in retroviral genes associated with drug resistance. In: Kuiken C, Foley B, Hahn B, et al. (eds). HIV Sequence Compendium. Los Alamos, NM: Los Alamos National Laboratory 2001.
7. Cormier E, Dragic T. An overview of HIV-1 coreceptor function and its inhibitors. In: Kuiken C, Foley B, Hahn B, et al. (eds). HIV Sequence Compendium. Los Alamos, NM: Los Alamos National Laboratory 2000.
8. Gaschen B, Taylor J, Yusim K, et al. Diversity considerations in HIV-1 vaccine selection. Science 2002;296:2354-60.
9. Warren J, Levinson M. Fourth annual survey of worldwide HIV, SIV, and SHIV challenge studies in vaccinated nonhuman primates. J Med Primatol 1995;24:150-68.
10. Sharp P, Bailes E, Robertson D, Gao F, Hahn B. Origins and evolution of AIDS viruses. Biol Bull 1999;196:338-42.
11. Hahn B, Shaw G, De Cock K, Sharp P. AIDS as a zoonosis: scientific and public health implications. Science 2000;287:607-14.
12. Chen Z, Luckay A, Sodora D, et al. HIV-2 seroprevalence and characterization of a distinct HIV-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys. J Virol 1997;71:3953-60.
13. Gao F, Robertson D, Morrison S, et al. The heterosexual HIV type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J Virol 1996;70:7013-29.
14. Carr J, Salminen M, Koch C, et al. Full-length sequence and mosaic structure of a HIV type 1 isolate from Thailand. J Virol 1996;70:5935-43.
15. Gaschen B, Kuiken C, Korber B, Foley B. Retrieval and on-the-fly alignment of sequence fragments from the HIV database. Bioinformatics 2001;17:415-8.
16. Yang Z, Nielsen R, Hasegawa M. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 1998;15:1600-11.
17. Rose P, Korber B. Detecting hypermutations in viral sequences with an emphasis on G —> A hypermutation. Bioinformatics 2000;16:400-1.
18. Calef C. Numbering positions in SIV relative to SIVMM239. In: Kuiken C, Foley B, Hahn B, et al. (eds). Human Retroviruses and AIDS. Los Alamos: Los Alamos National Laboratory 2001.
19. Higgins D. Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. Comput Appl Biosci 1992;8:15-22.
20. Delwart E, Gordon C. Tracking changes in HIV-1 envelope quasispecies using DNA heteroduplex analysis. Methods 1997;12:348-54.
21. Hirsch M, Brun-Vezinet F, D'Aquila R, et al. Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society-USA Panel. JAMA 2000;283:2417-26.
22. Demeter L, Haubrich R. Phenotypic and genotypic resistance assays: methodology, reliability, and interpretations. J Acquir Immune Defic Syndr 2001;26(Suppl 1):3-9.
23. Haubrich R, Demeter L. Clinical utility of resistance testing: retrospective and prospective data supporting use and current recommendations. J Acquir Immune Defic Syndr 2001;26(Suppl 1):51-9.
24. Miller V, Larder B. Mutational patterns in the HIV genome and cross-resistance following nucleoside and nucleotide analogue drug exposure. Antivir Ther 2001;6(Suppl 3):25-44.
25. Miller V. Resistance to protease inhibitors. J Acquir Immune Defic Syndr 2001;26(Suppl 1):34-50.
26. Hanna G, D'Aquila R. Clinical use of genotypic and phenotypic drug resistance testing to monitor antiretroviral chemotherapy. Clin Infect Dis 2001;32:774-82.
27. Shafer R. Genotypic testing for HIV type 1 drug resistance. Clin Microbiol Rev 2002;15:247-77.
28. Kantor R, Machekano R, Gonzales M, Dupnik B, Schapiro J, Shafer R. HIV reverse transcriptase and protease sequence database: An expanded model integrating natural language text and sequence analysis. Nucleic Acids Res 2001;29:296-9.
29. Gonzales M, Dugan J, Shafer R. Synonymous-non-synonymous mutation rates between sequences containing ambiguous nucleotides (Syn-SCAN). Bioinformatics 2002;18:886-7.
30. Shafer R, Jung D, Betts B. HIV type 1 reverse transcriptase and protease mutation search engine for queries. Nat Med 2000;6:1290-2.
31. Kilby J, Hopkins S, Venetta T, et al. Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. Nat Med 1998;4:1302-7.
32. Clotet B, Lazzarin A, Cooper D, et al. Enfuvirtide (T-20) in combination with an optimized background (OB) regimen vs. OB alone in patients with prior experience or resistance to each of the three classes of approved antiretrovirals in Europe and Australia. World AIDS Conference. Barcelona 2002 [LbOr19A].
33. Henry K, Lalezari J, O'Hearn M, et al. Enfuvirtide (T-20) in combination with an optimized background (OB) regimen vs. OB alone in patients with prior experience or resistance to each of the three drug classes of approved antiretrovirals in North America and Brazil. World AIDS Conference. Barcelona 2002 [LbOr19B].
34. Kantor R, Katzenstein D, Gonzales M, et al. Influence of subtype and treatment on genetic profiles for HIV-1 reverse transcriptase and protease: do they act independently in predicting mutation specific probabilities in non-subtype B sequences. Antivir Ther 2002;7(Suppl):142.
35. UK Collaborative Group on Monitoring the Transmission of HIV Drug Resistance. Analysis of prevalence of HIV-1 drug resistance in primary infections in the United Kingdom. BMJ 2001;322:1087-8.
36. Simon V, Vanderhoeven J, Hurley A, et al. Evolving patterns of HIV-1 resistance to antiretroviral agents in newly infected individuals. AIDS 2002;16:1511-9.
37. Grant R, Hecht F, Warmerdam M, et al. Time trends in primary HIV-1 drug resistance among recently infected persons. JAMA 2002;288:181-8.
38. Little S, Holte S, Routy J, et al. Antiretroviral drug resistance among patients recently infected with HIV. N Engl J Med 2002;347:385-94.
39. Gonzales MJ, Machekano R, Shafer R. HIV type 1 reverse transcriptase and protease subtypes: classification, amino acid mutation patterns, and prevalence in a northern California clinic-based population. J Infect Dis 2001;184:998-1006.
40. Schiffer C, Scott W, Stewart F, King M, Kempf D. The uncommon HIV-1 protease mutation I54T is highly associated with G48V and may affect cleavage dynamics by stabilizing the structure of the flaps. Antivir Ther 2001;6(Suppl 1):52.
41. Winters M, Merigan T. Variants other than aspartic acid at codon 69 of the HIV type 1 reverse transcriptase gene affect susceptibility to nucleoside analogs. Antimicrob Agents Chemother 2001;45:2276-9.