Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database: an expanded data model integrating natural language text and sequence analysis programs

Rami Kantor[1], Rhoderick Machekano[1], Mathew J. Gonzales[1], Kathryn Dupnik[1], Jonathan M. Schapiro[1,2] and Robert W. Shafer[1,]*

[1]Division of Infectious Diseases, Stanford University Medical Center, Stanford, CA 94305, USA and [2]National Hemophilia Center, Tel-Hashomer Hospital, Tel Aviv, Israel

Received October 3, 2000; Accepted October 11, 2000

ABSTRACT

The HIV Reverse Transcriptase and Protease Sequence Database is an on-line relational database that catalogs evolutionary and drug-related sequence variation in the human immunodeficiency virus (HIV) reverse transcriptase (RT) and protease enzymes, the molecular targets of anti-HIV therapy (http://hivdb.stanford.edu). The database contains a compilation of nearly all published HIV RT and protease sequences, including submissions from International Collaboration databases and sequences published in journal articles. Sequences are linked to data about the source of the sequence sample and the antiretroviral drug treatment history of the individual from whom the isolate was obtained. During the past year 3500 sequences have been added and the data model has been expanded to include drug susceptibility data on sequenced isolates. Database content has also been integrated with didactic text and the output of two sequence analysis programs.

DATABASE RATIONALE AND HISTORY

Medical and biological relevance

Human immunodeficiency virus (HIV) reverse transcriptase (RT) and protease enzymes are the molecular targets of the 15 currently licensed antiretroviral drugs. Sequence changes in the genes coding for these enzymes are directly responsible for phenotypic resistance to RT and protease inhibitors. Individuals infected with drug-susceptible HIV isolates experience reductions in morbidity and mortality with appropriate antiretroviral drug therapy. In contrast, individuals infected with drug-resistant isolates do not usually respond to drug therapy (1). Assays for sequencing HIV RT and protease are commercially available and are widely used in clinical settings. However, the optimal means of interpreting RT and protease sequences is not known, and the potential utility of such sequence results in clinical settings is an area of intense clinical investigation (2).

HIV RT and protease

HIV RT is a heterodimer composed of a p66 subunit, which contains the DNA-binding groove and the active site, and a p51 subunit, which appears to function as a scaffold for the enzymatically active p66 subunit. HIV RT is responsible for RNA-dependent DNA polymerization, RNase H activity and DNA-dependent DNA polymerization. Although it is 560 amino acids in length, almost all known drug resistance mutations are found in the 5′-polymerase coding region, which is the region sequenced by most clinical laboratories. HIV protease is responsible for the post-translational processing of the viral gag and gag-pol polyproteins to yield the structural proteins and enzymes of the virus. The enzyme is an aspartic protease composed of two non-covalently associated, structurally identical monomers 99 amino acids in length. The protease has a binding cleft that specifically recognizes and cleaves at least 10 different sequences on viral precursor polyproteins.

HIV genetic variation

Genetic analysis of HIV isolates has revealed 10 different group M (main) subtypes, differing from one another by 10–30%, and several highly divergent group O and group N outlier sequences (3). Sequences of HIV isolates collected from around the world provide evidence for naturally occurring genetic polymorphisms; sequences of isolates from persons receiving anti-HIV drug therapy indicate genetic changes selected under antiretroviral drug pressure that may be associated with drug resistance. Sequences within the HIV RT and Protease Sequence Database indicate that ~40% of the 99 amino acids in the protease are polymorphic in untreated individuals, and up to 67% are polymorphic in patients receiving antiretroviral therapy. In the RT, percentages of polymorphisms are ~25 and 40%, respectively.

Database history

The HIV RT and Protease Sequence Database is a relational database begun in 1998 as a catalog of sequences linked to data

*To whom correspondence should be addressed. Tel: +1 650 725 2946; Fax: +1 650 723 8596; Email: rshafer@cmgm.stanford.edu

Table 1. Summary of the HIV RT and Protease Sequence Database Web Site

Web page	Description	URL
Database documentation		
Background; primer	Database rationale for lay audience and novice users	http://hivdb.stanford.edu/hiv/rationale.html http://hivdb.stanford.edu/hiv/primer.html
Data model for understanding HIV drug resistance mutations	Description of the sources of knowledge of drug resistance mutations	http://hivdb.stanford.edu/hiv/drm.html
User guide	Description of database schema and explanation of specific web site features	http://hivdb.stanford.edu/hiv/userguide.html
Resistance notes	Overview of HIV drug resistance with links to relevant database entries	http://hivdb.stanford.edu/hiv/Notes.pl
Summary statistics	Dynamically updated summary of database content	http://hivdb.stanford.edu/hiv/summary.pl
Database query forms[a]		
Mutations	Users retrieve sequences containing a selected mutation or mutations	http://hivdb.stanford.edu/hiv/prmut.pl http://hivdb.stanford.edu/hiv/rtmut.pl
Drug therapy	Users retrieve sequences of isolates from patients receiving a selected drug or drug combination.	http://hivdb.stanford.edu/hiv/PI_form.pl http://hivdb.stanford.edu/hiv/RTI_form.pl
References	Users retrieve sequences and summaries of data from studies published by the selected author	http://hivdb.stanford.edu/hiv/Reference.pl
Drug susceptibility	Users retrieve published drug susceptibility data for isolates with selected mutations	http://hivdb.stanford.edu/hiv/Notes.pl
Sequence analysis programs		
HIV-SEQ	Compares new RT and protease sequences to previously published sequences with the same mutations	http://hivdb.stanford.edu/hiv/programs.htm
Drug resistance interpretation (beta test version)	Infers drug resistance to 15 available drugs using rules hyperlinked to data within the database	http://hiv-4.stanford.edu/hiv/cgi-test/hivtest-web.pl

[a] Additional database query forms include forms providing access to specific amino acid insertions and to nucleotide patterns at specific amino acid positions.

about the source of the sequence sample and the antiretroviral drug treatment history of the individual from whom the isolate was obtained. The database schema has been described in detail in earlier publications (4,5). During the past year, the data model has been expanded to include phenotypic drug susceptibility data and other data pertaining to the interpretation of RT and protease sequences obtained in clinical settings. In addition, two sequence analysis programs and didactic text have been integrated via hyperlinks to specific database queries. In the sections that follow, we will review key database features with an emphasis on those that were introduced within the past year.

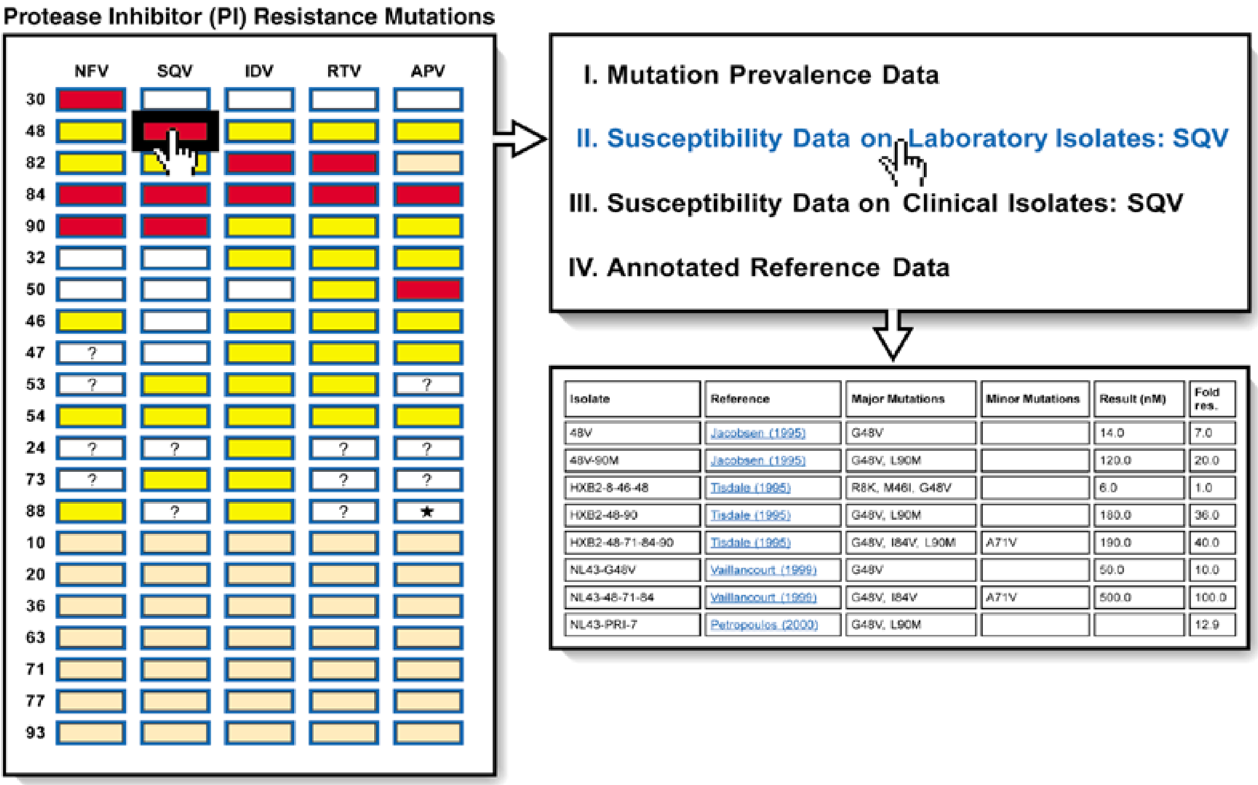
DATABASE OVERVIEW

Table 1 contains a summary of the key web pages in the database divided into three sections: documentation, queries and sequence analysis programs. As of September 15, 2000, the database contained 7909 sequences from 2215 individuals, including 4038 RT and 3871 protease sequences. These sequences include 5820 sequences published in GenBank and 2089 sequences from published journal articles. The database also contains >3500 drug susceptibility results.

Database documentation

The 'Background' is a brief bulleted description of the database rationale targeted towards a lay audience. The 'Primer' is a more detailed description of the database rationale, particularly with respect to the problem of HIV drug resistance; the 'Primer' also describes how the HIV RT and Protease Sequence Database differs from other databases containing HIV sequences, including the International Collaboration databases and Los Alamos HIV Sequence Database (6). The 'Data Model for Understanding HIV Drug Resistance' describes an ontology of HIV drug resistance based on four types of correlations between genetic sequences and other types of data, including drug treatment histories of patients from whom HIV isolates are sequenced, *in vitro* drug susceptibility data from laboratory HIV isolates, *in vitro* drug susceptibility data from clinical HIV isolates and clinical outcome data from patients receiving anti-HIV therapy.

'Summary Statistics' is a dynamically generated page based on SQL queries that provides the number of patients, isolates


Protease Inhibitor (PI) Resistance Mutations
NFV
SQV
IDV
RTV
APV
30
48
82
84
90
32
50
46
47
53
54
24
73
88
10
20
36
63
71
77
93
?
★
I. Mutation Prevalence Data
II. Susceptibility Data on Laboratory Isolates: SQV
III. Susceptibility Data on Clinical Isolates: SQV
IV. Annotated Reference Data

Isolate	Reference	Major Mutations	Minor Mutations	Result (nM)	Fold res.
48V	Jacobsen (1995)	G48V		14.0	7.0
48V-90M	Jacobsen (1995)	G48V, L90M		120.0	20.0
HXB2-8-46-48	Tisdale (1995)	R8K, M46I, G48V		6.0	1.0
HXB2-48-90	Tisdale (1995)	G48V, L90M		180.0	36.0
HXB2-48-71-84-90	Tisdale (1995)	G48V, I84V, L90M	A71V	190.0	40.0
NL43-G48V	Vaillancourt (1999)	G48V		50.0	10.0
NL43-48-71-84	Vaillancourt (1999)	G48V, I84V	A71V	500.0	100.0
NL43-PRI-7	Petropoulos (2000)	G48V, L90M			12.9

Figure 1. Protease mutations and their relationship to protease inhibitor drug susceptibilities. Drug names are shown at the top of the columns. Amino acid positions are shown at the left of each row. NFV, nelfinavir; SQV, saquinavir; IDV, indinavir; RTV, ritonavir; APV, amprenavir. Red rectangles denote primary resistance mutations. Yellow rectangles denote mutations that contribute to resistance when present with other mutations. Pale yellow rectangles represent mutations that are accessory resistance mutations which also occur as polymorphisms in untreated individuals. Rectangles containing a question mark imply that the relationship between the mutation and drug resistance has not been fully defined. Rectangles containing a star imply that the mutation causes hypersusceptibility to the drug shown at the top of the column. Each rectangle is hyperlinked to a page containing the four types of data shown in the top right box that link mutation and drug. An example of one of these types of data (e.g. SQV susceptibilities of laboratory isolates containing G48V) is shown in the bottom right box.

and sequences in the database meeting a variety of criteria. The 'Drug Resistance Notes' page contains graphical didactic summaries of the most common drug resistance mutations. The graphical summaries are image maps linked to data within the database relating the mutation to each of the clinically available anti-HIV compounds (Fig. 1).

Database queries

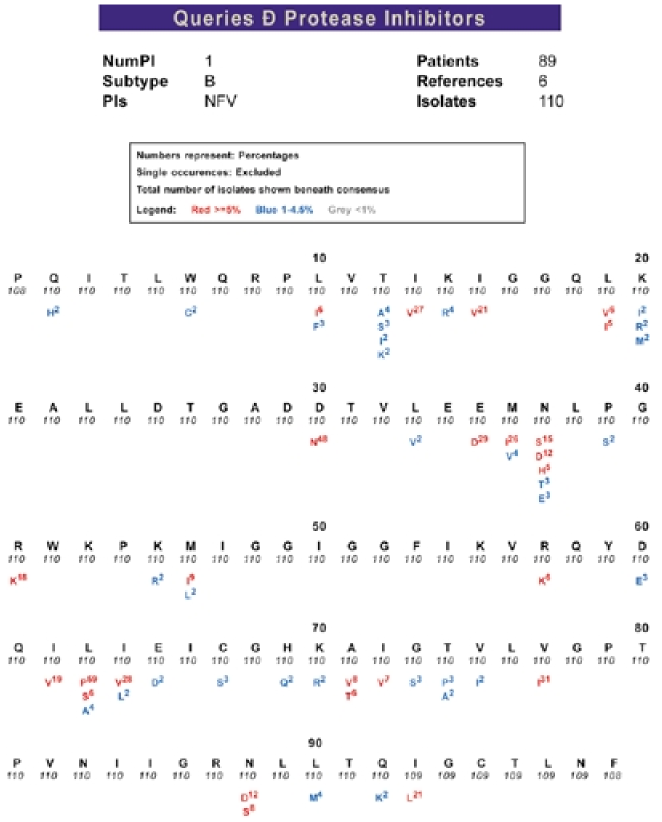
Users can retrieve theoretically limitless numbers of different sequence sets matching selection criteria based on specific references, drug treatments, RT and protease mutations, and drug susceptibility patterns (Table 1, Database query forms). Each query returns a new table and each record in the new table contains 8–12 columns of associated data selected by the user. The data returned may include: (i) hyperlinks to MEDLINE abstracts and GenBank records; (ii) complete nucleotide sequences and translations; (iii) classification of the sequences by individual and time point; (iv) data on the HIV species, group and subtype represented by each sequence; (iv) data on drug treatment including lists of drugs and complete summaries of drug regimens received by the patient from whom the isolates were obtained; and (v) technical data on the methods of virus isolation and sequencing. Following retrieval of a sequence set, users can view or download complete sequence alignments in a variety of formats, including a composite sequence alignment summarizing the frequency and type of mutation at each position in the sequence (Fig. 2).

Sequence analysis programs

HIV-SEQ (HIV RT and Protease Search Engine for Queries, formerly HRP-ASAP) accepts user-submitted RT and protease sequences, compares them to a reference sequence and uses the differences (mutations) as query parameters for interrogating the sequence database (7). The program allows users to discover associations between a submitted sequence and previously published sequences containing the same mutations. The web site also contains a beta test version of an anti-HIV drug resistance interpretation program, which accepts user-submitted RT and protease sequences and returns inferred levels of resistance to 15 clinically available anti-HIV drugs. Drug resistance is inferred using a comprehensive set of rules hyper-linked to the output of specific database queries.

FUTURE DIRECTIONS

The data model will be broadened to link sequence data with additional types of *in vitro* and clinical phenotypic data. The model will continue to utilize relational features to represent



Queries Ð Protease Inhibitors

NumPI	1	Patients	89
Subtype	B	References	6
PIs	NFV	Isolates	110

Numbers represent: Percentages
Single occurences: Excluded
Total number of isolates shown beneath consensus
Legend: Red >=5% Blue 1-4.5% Grey <1%

									10										20
P	Q	I	T	L	W	Q	R	P	L	V	T	I	K	I	G	G	Q	L	K
108	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110
	H^{2}				C^{2}				I^{6} F^{3}		A^{4} S^{3} I^{2} K^{2}	V^{27}	R^{4}	V^{21}				V^{6} I^{5}	I^{2} R^{2} M^{2}

									30										40
E	A	L	L	D	T	G	A	D	D	T	V	L	E	E	M	N	L	P	G
110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110
									N^{48}			V^{2}		D^{29}	I^{21} V^{4}	S^{15} D^{12} H^{5} T^{3} E^{3}		S^{2}	

									50										60
R	W	K	P	K	M	I	G	G	I	G	G	F	I	K	V	R	Q	Y	D
110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110
K^{58}				R^{2}	I^{6} L^{2}											K^{5}			E^{3}

									70										80
Q	I	L	I	E	I	C	G	H	K	A	I	G	T	V	L	V	G	P	T
110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110	110
	V^{19}	P^{69} S^{6} A^{4}	V^{28} L^{2}	D^{2}		S^{3}		Q^{2}	R^{2}	V^{8} T^{6}	V^{7}	S^{3}	P^{3} A^{2}	I^{2}		I^{31}			

									90									
P	V	N	I	I	G	R	N	L	L	T	Q	I	G	C	T	L	N	F
110	110	110	110	110	110	110	110	110	110	110	110	109	109	109	109	109	109	108
							D^{12} S^{8}		M^{4}		K^{2}	L^{21}						

Figure 2. Composite sequence alignment. This is one of the methods for viewing the results of queries. This query retrieved sequences from HIV-1 subtype B patients receiving nelfinavir as their sole protease inhibitor. The page header shows the query parameters and the numbers of references, patients and isolates returned by the query. The figure legend shows whether the user wished to view the absolute number of mutations at each position and whether the user wished to ignore rare occurrences that were observed in only one sequence. The first line in that summary shows the numbered consensus sequence. The second line contains the total number of isolates in the data set at each position. The remaining lines show the frequency of variation at each position in the sequence data set. If this composite alignment is compared to the composite alignment of protease sequences from untreated patients, one would observe that the changes at codons 30, 46, 73, 88 and 90 occurred only in isolates from patients who had received nelfinavir and that changes at codons 36 and 71 were present in both alignments but occurred more commonly in isolates from patients receiving nelfinavir.

data such as sequences and *in vitro* drug susceptibility results. However, object-oriented features will be incorporated to model higher-level data such as disease stage, complex treatment regimens and response to anti-HIV therapy. These changes in the data model will expand the usefulness of the database, increasing the user base from beyond a core group of sequence analysts to include clinical investigators studying other aspects of HIV drug resistance. Preliminary research on a third molecular target of HIV therapy, the gp41 envelope protein (8), has also begun and will be included in the database when appropriate.

CITING THE DATABASE

Please refer to this article, when citing the HIV RT and Protease Sequence Database.

REFERENCES

1. Carpenter,C.C., Cooper,D.A., Fischl,M.A., Gatell,J.M., Gazzard,B.G., Hammer,S.M., Hirsch,M.S., Jacobsen,D.M., Katzenstein,D.A., Montaner,J.A. *et al.* (2000) Antiretroviral therapy in adults: updated recommendations of the International AIDS Society-USA Panel. *JAMA*, **283**, 381–390.
2. Hirsch,M.S., Brun-Vezinet,F., D'Aquila,R.T., Hammer,S.M., Johnson,V.A., Kuritzkes,D.R., Loveday,C., Mellors,J.W., Clotet,B., Conway,B. *et al.* (2000) Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society-USA Panel. *JAMA*, **283**, 2417–2426.
3. Robertson,D.L., Anderson,J.P., Bradac,J.A., Carr,J.K., Foley,B., Funkhouser,R.K., Gao,F., Hahn,B.H., Kalish,M.L., Kuiken,C.L. *et al.* (2000) HIV-1 Nomenclature Proposal A Reference Guide to HIV-1 Classification. *Science*, **288**, 55–56.
4. Shafer,R.W., Stevenson,D. and Chan,B. (1999) Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database. *Nucleic Acids Res.*, **27**, 348–352.
5. Shafer,R.W., Jung,D.R., Betts,B.J., Xi,Y. and Gonzales,M.J. (2000) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.*, **28**, 346–348.
6. Kuiken,C.L., Foley,B., Hahn,B.H., Marx,P., McCutchan,F.E., Mellors,J., Mullins,J., Wolinsky,S. and Korber,B. (1999) *Human Retroviruses and AIDS. A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences*. Theoretical Biology and Biophysics, Los Alamos, NM.
7. Shafer,R.W., Jung,D.R. and Betts,B.J. (2000) Human immunodeficiency virus type 1 reverse transcriptase and protease search engine for queries. *Nature Med.*, **6**, 1290–1292.
8. Kilby,J.M., Hopkins,S., Venetta,T.M., DiMassimo,B., Cloud,G.A., Lee,J.Y., Alldredge,L., Hunter,E., Lambert,D., Bolognesi,D. *et al.* (1998) Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. *Nature Med.*, **4**, 1302–1307.