348–352 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database

Robert W. Shafer*, Derek Stevenson and Bryan Chan

Division of Infectious Diseases, Stanford University Medical Center, Stanford, CA 94305, USA

Received September 15, 1998; Revised and Accepted October 16, 1998

ABSTRACT

The HIV RT and Protease Sequence Database is an on-line relational database that catalogues evolutionary and drug-related human immunodeficiency virus reverse transcriptase (RT) and protease sequence variation (http://hivdb.stanford.edu). The database contains a compilation of nearly all published HIV RT and protease sequences including International Collaboration database submissions (e.g., GenBank) and sequences published in journal articles. Sequences are linked to data about the source of the sequence sample and the anti-HIV drug treatment history of the individual from whom the isolate was obtained. The database is curated and sequences are annotated with data from 180 literature references. Users can retrieve additional data and view alignments of sequences sets meeting specific criteria (e.g., treatment history, subtype, presence of a particular mutation).

HIV RT AND PROTEASE

HIV RT is a heterodimer composed of p51 and p66 subunits. It is responsible for RNA-dependent DNA polymerization, RNase H activity, and DNA-dependent DNA polymerization. The p51 subunit is composed of the first 450 amino acids of the RT gene. The p66 subunit is composed of all 560 amino acids of the RT gene. Although the p51 and p66 subunits share 450 amino acids, their relative arrangements are significantly different. The p66 subunit contains the DNA-binding groove and the active site; the p51 subunit appears to function as a scaffold for the enzymatically active p66 subunit. The three dimensional structure of HIV-1 RT, bound to both a double-stranded nucleic acid, to a non-nucleoside RT inhibitor, and unbound have been determined by X-ray crystallography (Fig. 1A; 1).

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 the viral precursor polyproteins. The three dimensional structure of wild type HIV-1 protease and of several drug-resistant mutant forms bound to various inhibitors have been determined crystallographically (Fig. 1B; 2).

MEDICAL RELEVANCE

Recent studies show that HIV replication can be dramatically curtailed, if not completely arrested, with highly active antiretroviral drug combinations. The benefits of combination therapy, however, are greatly diminished in patients who have received previous anti-HIV therapy. Although 12 anti-HIV drugs are approved by the US Food and Drug Administration (FDA) (five nucleoside analog RT inhibitors, four protease inhibitors and three non-nucleoside analog RT inhibitors), there is considerable cross-resistance within each class of inhibitors. Several new anti-HIV drugs will be approved by the FDA within the next 1–2 years, but preliminary data suggest that many current drug-resistant HIV isolates will also be resistant to these new drugs.

HIV RT and protease sequencing, and drug susceptibility testing, have been used extensively in clinical trials. In these settings, drug resistance is often a confounding variable that influences the relative effectiveness of the drug regimens under study. Assays for sequencing RT and protease are also commercially available and are now being studied for their potential role in clinical practice (3). As the functional and clinical correlates of HIV RT and protease sequence changes become better understood, sequence results obtained in clinical settings will become more meaningful.

HIV SEQUENCE VARIATION

Factors contributing to HIV genetic variation include: (i) the lack of proofreading capability by HIV RT; (ii) the high *in vivo* rate of HIV replication; (iii) the accumulation of proviral HIV variants during the course of infection; and (iv) recombination. The likelihood of developing drug resistance depends on the size and heterogeneity of the HIV population within an individual, the ease of acquisition of a particular mutation (or set of mutations), and the effect of drug-resistance mutations on changes in drug susceptibility and virus fitness (4). Some mutations selected during drug therapy confer measurable resistance by themselves, other mutations produce measurable resistance only when present in combination.

Genetic analysis of HIV-1 isolates has revealed at least 10 distinct group M (main) subtypes (A–J) as well as several highly divergent group O (outlier) isolates. Differences between group M subtypes are based on the ~30% intersubtype genetic divergence in the *env* region and 14% intersubtype divergence in the *gag* region (5,6). The world-wide HIV pandemic is caused by

*To whom correspondence should be addressed. Tel: +1 650 725 2946; Fax: +1 650 725 2395; Email: rshafer@cmgm.stanford.edu

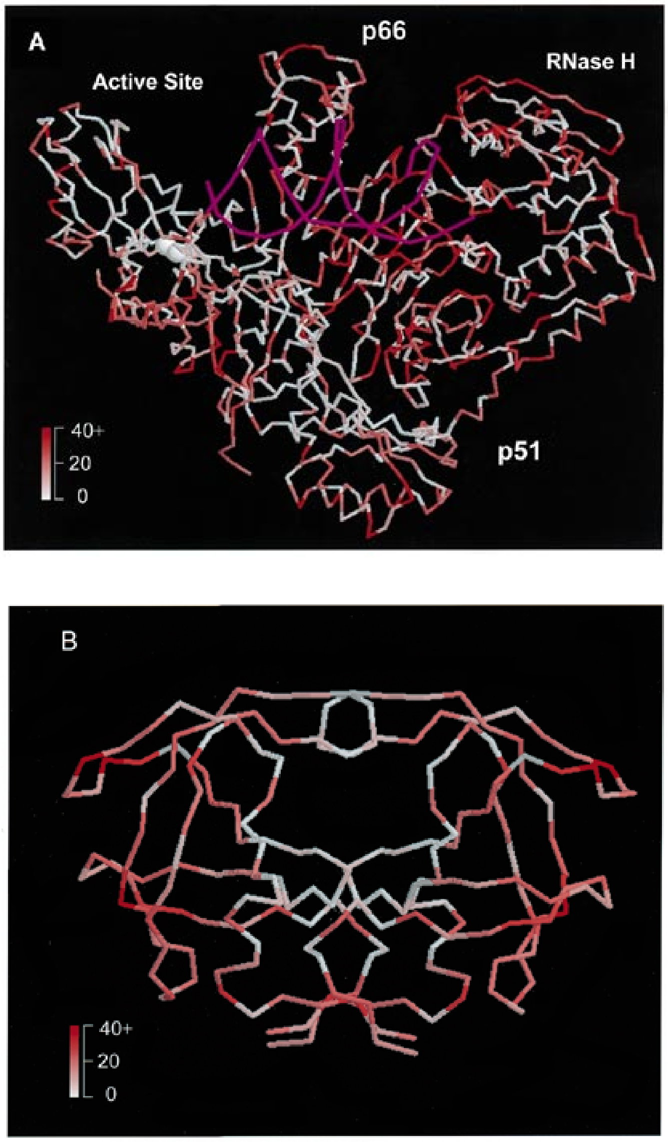

Figure 1. Three-dimensional structure of HIV-1 reverse transcriptase (**A**) and protease (**B**) as determined by X-ray crystallography. An alignment of HIV-1, HIV-2 and SIV-agm sequences were used to identify highly conserved residues (backbone is white) and polymorphic residues (backbone is red) with respect to the HIV-1 subtype B consensus sequence.

group M HIV-1 virus. In North America, Europe and Australia, most HIV-1 isolates belong to subtype B and the available anti-HIV drugs have been developed by drug screening and susceptibility testing using subtype B isolates. However, subtype B accounts for only a small proportion of HIV-1 isolates worldwide and, even in industrialized countries, non-B isolates are being identified with increasing frequency.

IDENTIFYING DRUG RESISTANCE MUTATIONS

Drug resistance mutations have traditionally been identified during the pre-clinical and early clinical evaluation of a new anti-HIV drug. During these studies, drug-resistant HIV-1 isolates are identified, sequenced and tested for drug susceptibility. Site-directed mutagenesis experiments are then done to confirm the role of specific mutations introduced into a wild type virus. Drug-resistance mutations identified by this method acquire widespread acceptance and are referred to as ‘canonical’ resistance mutations.

This experimental approach, however, has limitations because many different combinations of mutations are associated with HIV-1 drug resistance and because the effect of a mutation often depends on the genetic context in which it develops. In many individuals, particularly those receiving drug combinations, complex patterns of ‘non-canonical’ mutations develop in drug-resistant isolates (7,8). The HIV RT and Protease Sequence Database was developed on the premise that sequences of clinical HIV-1 isolates are experiments of nature that should be cataloged and examined methodically to help prioritize clinical investigations and further *in vitro* experimental work.

DATABASE SCHEMA

The web site is built around a relational database containing the text of each sequence, data about the person from whom the sequence was obtained (e.g., country, treatment history) and data about the methods of sequencing and sample isolation (e.g., year of isolation, body source, cloning method) (Fig. 2). Sequences are stored in a virtual alignment with the subtype B consensus sequence (5); thus, amino acid sequences are also represented as lists of amino acid differences from the consensus sequence (MUTATIONS table). The number of the start and stop residue of each sequence is maintained along with a table containing insertions, thereby avoiding the need for a multiple sequence alignment algorithm when a set of sequences is retrieved.

There is a hierarchical relationship linking four of the entities in the database: patient, patient-date, isolate, and sequence (Fig. 2). An individual (PATIENT table) may have isolates obtained at different times (PATIENT-DATE table). A patient-date will have more than one isolate (ISOLATE table) if samples are obtained from more than one source (e.g., peripheral blood mononuclear cells, plasma, lymph node). Finally, if multiple clones are sequenced from an individual isolate then each clone is considered a different sequence (SEQUENCE table).

It is often not desirable to have the sequences of multiple clones from a single sample (many sequences per sample) included with the results of direct PCR (population-based) sequencing (one sequence per sample) because the sequence set will be biased by the samples with multiple clones. Therefore, the database includes a consensus sequence for each set of clones from a single isolate and users are offered a choice as to whether the consensus of multiple clones will be retrieved or whether the sequences of each clone will be retrieved.

DATABASE USE

The database allows users to retrieve a set of sequences meeting specific criteria. The principal selection criteria include treatment history, subtype and the presence of a particular mutation. There are six pre-defined queries and two user user-defined queries. The pre-defined queries include: (i) ‘RT isolates from patients not receiving RT inhibitors’; (ii) ‘RT isolates from patients receiving RT inhibitors’; (iii) ‘Protease isolates from patients not receiving protease inhibitors’; (iv) ‘Protease isolates from patients receiving protease inhibitors’; (v) ‘Global isolates of known HIV-1 subtype’; and (vi) ‘Isolates according to reference’. The user defined queries include: (i) ‘Select isolates having a specific

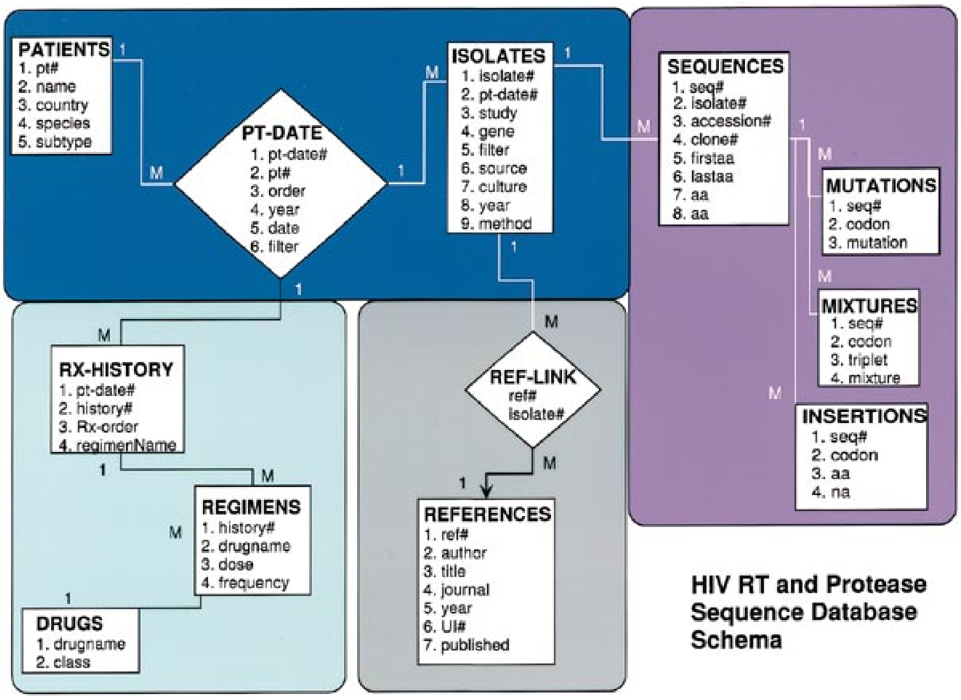

Figure 2. Schema of HIV RT and Protease Sequence Database. The 10 rectangles (entities) and two diamonds (relationships) depict 12 of the base tables in the database. Within the tables, the attributes (fields) are listed. The multiplicities of the linkages are demonstrated by either a 1 or an M (many). The schema demonstrates the hierarchical relationship between patients, patient-dates, isolates and sequences.

mutation'; and (ii) 'Select isolates obtained from a patient receiving a specific drug or drug combination'.

Each query returns a new table and each record in the new table contains 8–12 columns of data. The data returned include: (i) hyperlinks to the MEDLINE abstract, the GenBank record, and the amino acid sequence (or sequences in the case of multiple clones); (ii) a classification of the sequence by patient, patient-date and isolate; (iii) data on HIV-1 subtype and drug treatment history; and (iv) miscellaneous additional data depending on the query. Following retrieval of a sequence set, users have the option of viewing and downloading an alignment of the sequences.

DATABASE CONTENT

As of October 1, 1998 the database contained sequences from 825 individuals at 1306 time points. There were 3732 sequences, and nearly 40 000 mutations (differences from the consensus B sequence). The database contained 1794 RT and 1938 protease sequences. Nearly 3400 sequences had GenBank accession numbers; ~300 sequences were from published journal articles and were not in GenBank.

Figure 3A and B show neighbor-joining trees created from published primate lentivirus RT and protease sequences. Of the 1684 isolates, 1636 were HIV-1, 38 were HIV-2 (or SIV sooty mangabey) and 10 were SIV-agm (African green monkey). Table 1 summarizes the published HIV-1 isolates of known subtype. The HIV-1 group M sequences demonstrate the extent of evolutionary genetic variation that has taken place since entry of this virus into the human population estimated to have occurred sometime within the past 100 years (10). There is currently only a single non-subtype B sequence from an individual receiving antiretroviral therapy. This is likely to change if the prevalence of non-subtype B virus continues to increase within industrialized

Table 1. Published protease and reverse transcriptase sequences of global HIV-1 isolates

Subtype	# Protease	# RT	Countries
A	65	28	Africa: Ivory Coast, Uganda, Rwanda, Ethiopia, Nigeria, Congo, Kenya, Zambia, Zaire Asia: Thailand, Lebanon
B	89	45	Africa: Gabon Asia: Lebanon, China Australia Europe: U.K., France, Germany, Netherlands North America: U.S., Haiti, South America: Brazil
C	27	27	Africa: Zimbabwe, Rwanda, Ethiopia Asia: India, Lebanon Europe: U.K. South America: Brazil
D	17	14	Africa: Uganda, Zaire Asia: Lebanon Europe: U.K. South America: Brazil
F	4	15	Europe: Romania South America: Brazil
G	6	6	Africa: Nigeria, Dijibouti, Congo, Kenya
H	1	1	Africa: Congo
Group O	16	14	Africa: Cameroon Europe: France, Spain, Norway North America: U.S.
Other:	2	1	Africa: Cameroon Asia: Lebanon

The subtype classification is based on *env* and/or *gag* gene sequences from the same virus isolate.

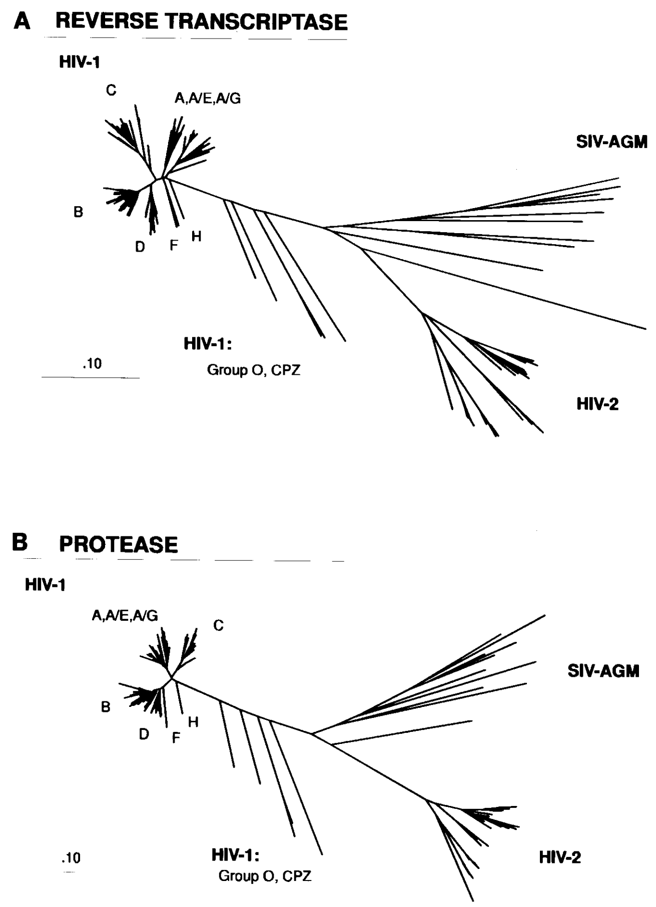

Figure 3. Neighbor-joining tree of primate lentiviruses viruses based on published reverse transcriptase (**A**) and protease (**B**) genes.

countries or if antiretroviral therapy is introduced into developing countries.

Figure 4A and B illustrate the polymorphisms of HIV-1 subtype B RT and protease from untreated individuals. Table 2 contains a summary of the numbers of published sequences from individuals receiving anti-HIV drug therapy. This table highlights some of the gaps in the body of published HIV-1 sequence data. By highlighting these gaps, the database aims to encourage submission of needed sequence data to GenBank, as well as the subsequent addition of the sequences to the HIV RT and Protease Sequence Database.

FUTURE DIRECTIONS

During the next year the database will expand in three principal directions. First, attempts will be made to fill many of the significant gaps in the body of published HIV RT and protease sequences. Second, an increased number of user-defined queries and additional options for retrieving sequence results will be added. Third, the database schema will be expanded to include drug susceptibility data and the database will be populated with drug susceptibility results from published articles.

Table 2. Published protease and RT sequences obtained from patients receiving anti-HIV therapy

		Received drug ± other drugs		Received no additional drugs*	
Drug	Gene	Amino Acid Sequence	Nucleic Acid Sequence	Amino Acid Sequence	Nucleic Acid Sequence
zidovudine	RT	168	136	124	105
didanosine	RT	83	71	17	15
zalcitibine	RT	20	16	2	2
stavudine	RT	5	5	0	0
lamivudine	RT	38	34	4	0
nevirapine	RT	0	0	0	0
delavirdine	RT	30	0	16	0
indinavir	Protease	39	39	35	35
saquinavir	Protease	38	38	34	34
ritonavir	Protease	59	11	55	7
nelfinavir	Protease	4	0	4	0
adefovir	RT	4	0	1	0
abacavir	RT	0	0	0	0
efavirenz	RT	0	0	0	0
amprenavir	Protease	0	0	0	0
zidovudine + didanosine	RT	47	47	42	42
zidovudine + lamivudine	RT	29	24	5	0

One of the isolates from an individual receiving zidovudine + lamivudine belonged to group O. The remaining isolates belong to group M and are presumed to belong to subtype B based on country of origin and phylogenetic analysis.

*The 5th and 6th columns include sequences from individuals who received no drugs other than the drug indicated in the first column.

CITING THE DATABASE

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

REFERENCES

1 Tantillo,C., Ding,J., Jacobo-Molina,A., Nanni,R.G., Boyer,P.L., Hughes,S.H., Pauwels,R., Andries,K., Janssen,P.A. and Arnold,E. (1994) *J. Mol. Biol.*, **243**, 369–387.
2 Erickson,J.W. and Burt,S.K. (1996) *Annu. Rev. Pharmacol. Toxicol.*, **36**, 545–571.
3 Hirsch,M.S., Conway,B., D’Aquila,R.T., Johnson,V.A., Brun-Vezinet,F., Clotet,B., Demeter,L.M., Hammer,S.M., Jacobsen,D.M., Kuritzkes,D.R., Loveday,C., Mellors,J.W., Vella,S. and Richman,D.D. (1998) *J. Am. Med. Assoc.*, **279**, 1984–1991.
4 Coffin,J.M. (1995) *Science*, **267**, 483–489.
5 Korber,B., Hahn,B., Foley,B., Mellors,J.W., Leitner,T., Myers,G., McCutchan,F. and Kuiken,C.L. (eds) (1997) *Human retroviruses and AIDS 1997: a compilation and analysis of nucleic and amino acid sequences.* Theoretical Biology and Biophysics Group. Los Alamos National Laboratory, Los Alamos, NM, pp. II-A-20–II-A-31.
6 Hu,D.J., Dondero,T.J., Rayfield,M.A., George,J.R., Schochetman,G., Jaffe,H.W., Luo,C.C., Kalish,M.L., Weniger,B.G., Pau,C.P., Schable,C.A. and Curran,J.W. (1996) *J. Am. Med. Assoc.*, **275**, 210–216.
7 Shafer,R.W., Kozal,M.J., Winters,M.A., Iversen,A.K., Katzenstein,D.A., Ragni,M.V., Meyer,W.A., Gupta,P., Rasheed,S., Coombs,R. and Merigan,T.C. (1994) *J. Infect. Dis.*, **169**, 722–729.
8 Shafer,R.W., Winters,M.A., Palmer,S. and Merigan,T.C. (1998) *Ann. Intern. Med.*, **128**, 906–911.
9 Learn,G.H.J., Korber,B.T., Foley,B., Hahn,B.H., Wolinsky,S.M. and Mullins,J.I. (1996) *J. Virol.*, **70**, 5720–5730.
10 Korber,B., Theiler,J. and Wolinsky,S. (1998) *Science*, **280**, 1868–1871.

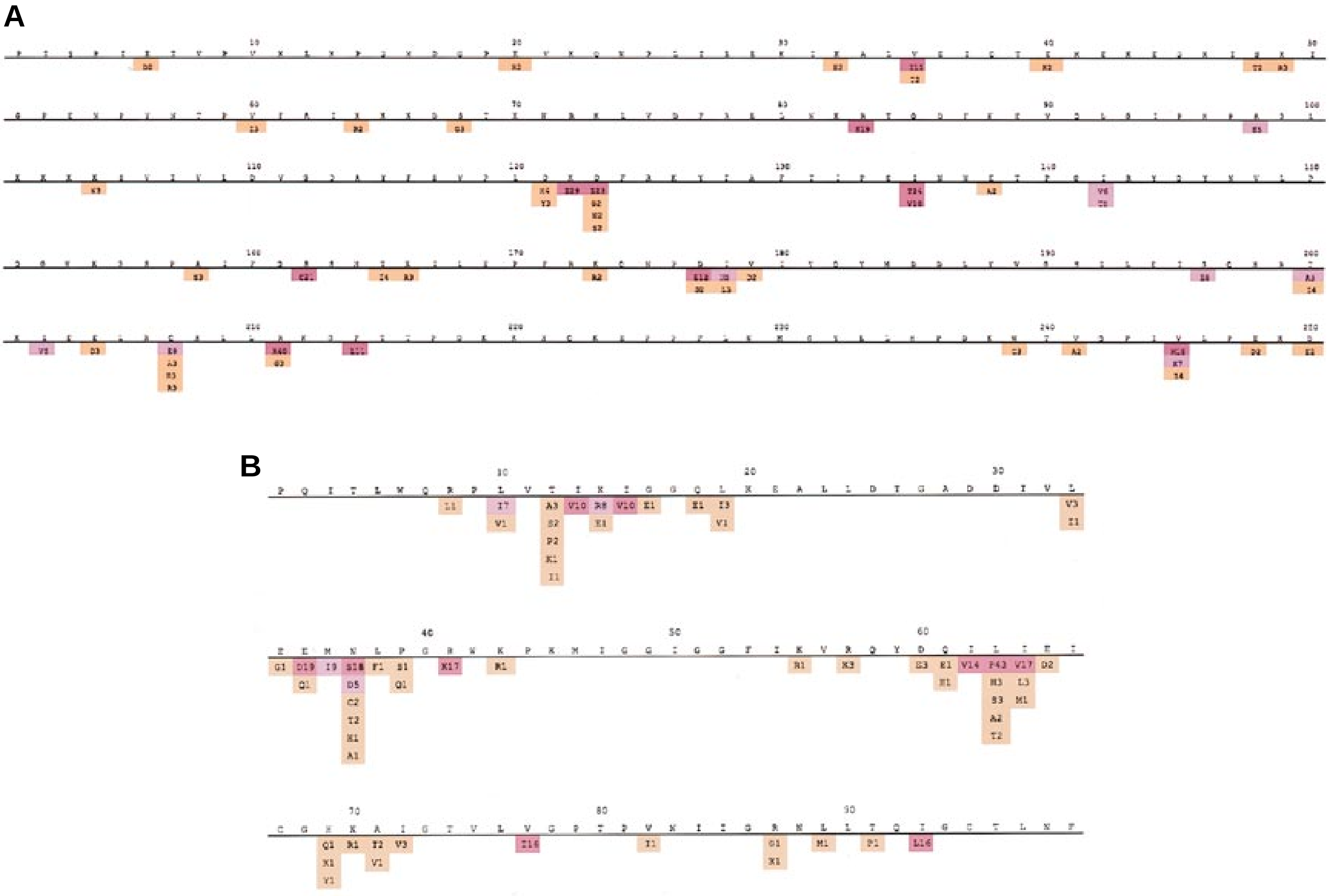
A
B
10
20
30
P Q I T L W Q R P L V T I K I G G Q L K E A L L D T G A D D T V L
40
50
60
E E M N L P G R W K P K M I G G I G G F I K V R Q Y D Q I L I E I
70
80
90
C G H K A I G T V L V G P T P V N I I G R N L L T Q I G C T L N F

Figure 4. HIV-1 subtype B reverse transcriptase (RT) (**A**) and protease (**B**) consensus sequence and polymorphisms of isolates from untreated individuals. Panel (A) is based on an alignment of 110 RT sequences from different individuals. To be included in this analysis, sequences had to encompass codons 40–220. Panel (B) is based on an alignment of 297 protease sequences from different individuals. Polymorphisms are shown beneath the consensus sequence. The number following the polymorphism is the percentage of isolates with that polymorphism. Only those polymorphisms that were present in at least two isolates are shown.