AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 11, Number 11, 1995
Mary Ann Liebert, Inc.

Sequence Note

A Computer Program Designed to Screen Rapidly for HIV Type 1 Intersubtype Recombinant Sequences

ADAM C. SIEPEL, AARON L. HALPERN, CATHERINE MACKEN, and BETTE T.M. KORBER

Detection of recombinant genomes involving the genetically distinct HIV-1 subtypes is a current problem of immense importance[1–3]; if such intersubtype recombination occurs with notable frequency, the already formidable challenge of developing vaccines against multiple subtypes is further complicated. Here we describe a simple and rapid heuristic method for detection of recombinant sequences (or more broadly, mosaic sequences), and its application to the *env* and *gag* coding sequences in the HIV Sequence Database.[4]

The computer program developed, called the Recombinant Identification Program (RIP), quickly provides summary information describing large sets of sequences (e.g., Table 1) and more detailed output describing particular sequences (e.g., Fig. 1). Two sequence sets with compatible alignments are required as input: a "background" alignment with subtype designations (currently HIV-1 subtypes A–H have been defined[4]), and an alignment of query sequences to be screened one at a time. Alignments may be either of nucleotide or amino acid sequences. The RIP first generates a consensus sequence to represent each subtype as delineated in the background alignment. A consensus threshold option allows the user to select a minimal frequency for the most common character in a given position for inclusion in the consensus sequence. As different thresholds occasionally suggested an additional recombinant, it is important to explore a range of thresholds. Second, the RIP slides a window of user-specified size, one position at a time, along the length of the query sequence and the aligned subtyped consensus sequences, and evaluates which subtype consensus is most similar to the query within each window. Similarity is defined as the fraction of matching characters. The center position of each window is marked with the appropriate subtype (see Fig. 1). A query sequence is identified as a potential intersubtype recombinant in the summary output if it bears a significant resemblance to one subtype in one region and another subtype in another region (Table 1b).

Two program options are important for how the RIP defines window boundaries. The first option is an "informative mode," which can decrease the number of ambiguous results in regions that are conserved across subtypes. When the informative mode is active, the RIP counts only informative positions, defined as positions at which at least one subtype consensus sequence differs from the others, when it determines window boundaries in relation to a central or reference position. The second option relates to the handling of gaps inserted to maintain alignment. The RIP can either "squeeze" gaps, meaning that gaps are considered except in positions where the query and all subtype consensus sequences are represented by a gap, or it can "strip" gaps, meaning that it ignores positions at which the query or any one of the subtype consensus sequences is represented by a gap.

The best-matching subtype within each window is qualified by a measure of confidence that is obtained by comparing the distance between the query and the subtype consensus it matches best, to the distance between the query and the subtype consensus it matches second-best. Confidence is calculated using a z test, assuming that (1) each site evolves independently according to the same process, and (2) the binomial distribution that theoretically results from the use of Hamming distances assuming independence of sites can be approximated by a normal distribution. Note that the calculations with respect to overlapping windows are not independent. For this reason and others, possible recombinants must be subjected to further phylogenetic analysis for confirmation. The detailed output of the RIP, as shown in Fig. 1, can help identify the region of potential cross-over sites to facilitate such subsequent analysis.

The RIP was used to scan 211 HIV-1 *env* sequences and 85 HIV-1 *gag* nucleotide sequences, using the most current alignments in the HIV Sequence Database,[4] and results were compared with the recent findings of Robertson and co-workers based on phylogenetic analyses.[2,3] (Table 1 provides a summary of the *env* analysis; *gag* intragenic and intergenic analyses are not shown due to space limitations, but will be reported in the 1995 edition of *Human Retroviruses and AIDS.*) Most sequences showed significant matches to only one subtype (Table 1a). Those that matched more than one subtype included 9 of the 12 intragenic recombinants noted by Robertson *et al.*[2,3]

Theoretical Biology and Biophysics (T10), Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

Table 1. Program Output Summarizing a Scan of 211 *env* Sequences in Los Alamos Database[a]

a. Representatives of the 202 sequences that showed no evidence of recombination

	Subtype							
Locus	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
KENYA	1540(996)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
SF1703	1461(990)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
GUN	0(0)	1500(1329)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
TB132	0(0)	1740(1740)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
ZAM18A	0(0)	0(0)	2095(2095)	0(0)	0(0)	0(0)	0(0)	0(0)
92BR025	0(0)	0(0)	462(462)	0(0)	0(0)	0(0)	0(0)	0(0)
UG269A	0(0)	0(0)	0(0)	1698(1698)	0(0)	0(0)	0(0)	0(0)
92UG046	0(0)	0(0)	0(0)	1051(1051)	0(0)	0(0)	0(0)	0(0)
CAR4071	0(0)	0(0)	0(0)	0(0)	1150(1150)	0(0)	0(0)	0(0)
CARELO	0(0)	0(0)	0(0)	0(0)	983(983)	0(0)	0(0)	0(0)
BZ126A	0(0)	0(0)	0(0)	0(0)	0(0)	1432(679)	0(0)	0(0)
CAR4067	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	290(166)	0(0)
CA13	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	631(631)
Z3	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
b. Nine possible recombinants								
CAR286A	418(418)	0(0)	0(0)	0(0)	0(0)	0(0)	53(53)	0(0)
CAR423A	345(345)	0(0)	0(0)	0(0)	0(0)	0(0)	13(13)	0(0)
DJ263A	1619(1560)	0(0)	0(0)	0(0)	9(9)	0(0)	0(0)	0(0)
UG266A*	94(94)	0(0)	0(0)	1677(1677)	0(0)	0(0)	0(0)	0(0)
CAR4039	116(116)	0(0)	0(0)	0(0)	722(387)	0(0)	0(0)	0(0)
VI525A*	0(0)	0(0)	0(0)	0(0)	9(9)	0(0)	744(742)	0(0)
CAR4081	445(406)	0(0)	0(0)	0(0)	0(0)	0(0)	32(32)	0(0)
K124A*	583(336)	0(0)	0(0)	1143(1143)	0(0)	0(0)	0(0)	0(0)
ZAM184*	380(264)	0(0)	205(205)	0(0)	0(0)	0(0)	0(0)	0(0)

[a]Those previously identified by Robertson *et al.*[2,3] are marked by asterisks. Values outside parentheses indicate the number of windows in which a query sequence is significantly similar to the consensus of a given subtype; values inside parentheses indicate the maximum number of those that are consecutive. Note that the unclassified sequence among these, Z3,[4] had no window that gave a significant association with any one subtype. Subsequent phylogenetic analysis could not confirm that four of the sequences identified as possible recombinants (CAR286A, CAR423A, DJ263A, and CAR4081) actually were hybrids. For this run, the window size was 200 nucleotides, informative mode was off, and matches were considered significant if they occurred with at least 90% certainty, as calculated by the method described in this article.

(5 of 7 in *gag* and 4 of 5 in *env*; Table 1b). The two disagreements with Robertson *et al.* in *gag* may have been due to a lack of divergence among subtypes F, G, and H near the 3′ terminus of that region, which led the RIP to evaluate best matches as uncertain. The third disagreement, an A/D recombination site in HIVMAL envelope, presumably occurred because the seg-

```
CAR4039         ATTAGATCTGAAAATATCACAAACAATGCCAAAACCATAATAGTACAGCTGGTTACGCCTGTACAAATTAATTGTACCAGACCC...TCCAACAATATAAGA...ACAAGT.........GTAAGGATA......
A               ..................................................T.??.A.......A?..................T...AA........C........A................C.T.........
B               ...............T....GG.......T.....................AA?GAAT.....G.............A.........AA........C........A.............A..CAT.........
C               ...............C.G..........T......A...........T..TAA.?AAT.....G.....GTG.....A.........AA...T....C........A.............A..............
D               ...............C.............T....................TAA.GA.T.....AC............A..G.......A...?............CA...G.........AC.CCT.........
E               ..C............C............................G..C..TAA..AAT.....G....C............................C...........................T.........
F               ..................T..G.T...A.A....................TAA.GAAT.....G.............A.........AA........C........A.............A..CAT.........
G               ..............CT.....G...........GT.........G.....TAA..ATAG.A..G.......................AAT.......C........A.............A...CA..C......
H               ..C..................G.....A......A...............TAAA.AT......GT......C...............AA...T..C.CG.......G....ATGCGT...A..G.T.........
BEST_MATCH        A           AA AA AA A   AAA   AAA        A  A  AA    AAA A  a    aaaa     a  a  a   eeE     E EE      EE   E         EE EEE  E
.                                ^^ ^^ ^   ^++   +++        ^                                      e              ^      ^^   +         ++ +++  ^
```

FIG. 1. A representative segment from the program output for CAR4039, including the putative 3′ cross-over site. The top line is the query sequence, the next eight lines represent signature patterns[9] of subtype consensus sequences with respect to the query sequence, and the bottom two lines indicate the subtypes that the query most resembles in various windows. In the signature patterns, periods (".") represent identity between the consensus and query sequences, letters represent consensus bases not found in the query sequence, and question marks ("?") indicate positions lacking a clearly defined consensus, with a consensus threshold level set at 50%. Best-matching subtype letters are positioned at the center of corresponding windows. They are provided only for columns that (1) lack gaps and question marks, and (2) are "informative," meaning that all subtype signature bases at that position are not identical. Capital letters indicate an absolute similarity greater than or equal to 90%, and lower-case letters a similarity less than 90%. Here the cross-over site can be approximated by the point at which "a"s change to "e"s. Carets ("∧") beneath the subtype names indicate a statistical confidence of 90% and plus signs ("+") indicate a confidence of 95%. The window size is 30 informative sites.

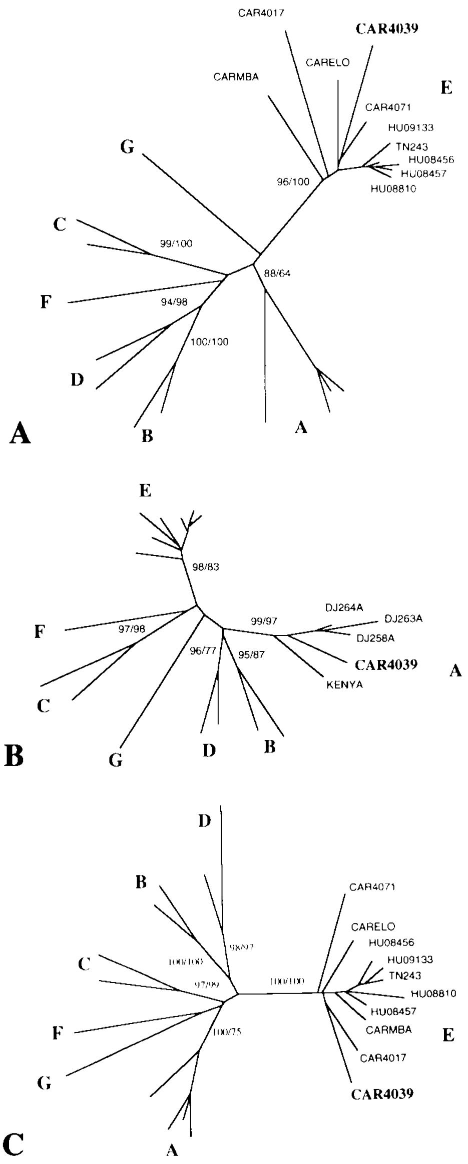

FIG. 2. Neighbor-joining trees, constructed using PHYLIP[10] with intersequence distances computed by the Kimura two-parameter method, for three *env* segments of CAR4039, an A/E hybrid: (a) the 5′ segment, positions 1–612; (b) the interior segment, positions 613–878; and (c) the 3′ segment, positions 879–1483. The positions are numbered according to the CAR4039[8] sequence. Likely cross-over sites were estimated using a chi-square optimization procedure,[2] with a four-sequence alignment including CAR4039, DJ264A, CARMBA, and SIVCPZGAB (the last three being arbitrary representatives of an A, an E, and an outgroup). Subtypes represented by more than one sequence are labeled with bootstrap values of the form neighbor-joining/parsimony (e.g., 94/98). Neighbor-joining bootstrap values were calculated with PHYLIP, and parsimony bootstrap values with PAUP[11]; 100 replicates were used in both cases. Note the high bootstrap values grouping the interior segment of CAR4039 with the A-clade. Sequence names of the form HU0XXXX contain the accession numbers of DAIDS- and WHO-sequenced samples from Thailand.

ment putatively belonging to subtype A is short (98 bp)[2,3] compared to the window sizes that we found practical. The RIP did correctly identify HIVMAL as an intergenic recombinant, as the subtype designations of A in *gag* and D in *env* were clear. The hybrid character of HIVMAL has been known since 1988.[5,6] In addition to the nine corroborations with Robertson *et al.*, five novel mosaic sequences were identified in *env* (Table 1b). In one of these, the *env* portion of CAR4039,[7] the case for intra-*env* recombination between subtypes A and E was strongly supported by phylogenetic analysis (Fig. 2). CAR4039 is from the Central African Republic, where E and A subtypes are cocirculating. The other four were only weakly supported by tree analyses, perhaps because the regions in question are not phylogenetically informative enough for confirmation, or because spurious results were obtained with RIP due to the lack of clear local subtype definition. Over small regions, the question of whether a mosaic sequence is the result of recombination or lack of divergence can be difficult to resolve. Using our method, the percentage of sequences with intragenic, intersubtype recombination events observed in *env* is about 4% (9 of 211). We are currently exploring whether a more detailed representation of subtypes than a simple consensus sequence will permit a more accurate detection of recombinants.

The automatic, sliding-window method for detecting intersubtype recombination as implemented by the RIP allows extremely rapid scanning of large sets of sequences with reasonable accuracy. It is likely that the window-sliding increment of 1 bp will allow detection of some recombinants that would be missed using tree analyses and "bootscanning" methods,[8] which due to their more labor-intensive nature use a larger window-sliding increment (e.g., 150 bp). A version of the RIP written in C++ for UNIX is available at the Human Retroviruses and AIDS ftp site: atlas.lanl.gov, in the directory pub/aids-db/PROGS/RIP.

ACKNOWLEDGMENTS

This work was funded by an NIH/DAIDS and DOE interagency agreement (3-Y01-AI-70001-13). We thank Gerald Myers for his thoughtful suggestions concerning both the manuscript and the program, and Beatrice Hahn and Jonathan Bingham for helpful conversations.

REFERENCES

1. Sabino EC, Shpaer EG, Morgado MG, Korber BTM, Diaz RS, Bongertz V, Cavalcante S, Galvao-Castro B, Mullins JI, and Mayer A: Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. J Virol 1994;68:6340–6346.
2. Robertson DL, Hahn BH, and Sharp PM: Recombination in AIDS viruses. J Mol Evol 1995;40:249–259.
3. Robertson DL, Sharp PM, McCutchan FE, and Hahn BH: Recombination in HIV-1. Nature (London) 1995;374:124–126.
4. Myers G, Korber B, Jeang K-T, Henderson L, Wain-Hobson S, and Pavlakis G (eds.): *Human Retroviruses and AIDS 1994.* Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico, 1994.

5. Myers G, Rabson AB, Josephs SJ, Smith TF, and Wong-Staal F (eds.): *Human Retroviruses and AIDS 1988.* Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico, 1988.
6. Li W-H, Tanimura M, and Sharp PM: Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 1988;5:313–330.
7. Schmitt D, Mathiot C, Levy J-F, Girard M, You B, Barre-Sinoussi F, Deslandres A, and Kieny MP: This sequence was provided to the HIV database (Ref. 4) prior to its publication in a journal.
8. Salminen MO, Carr JK, Burke DS, and McCutchan FE: Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses 1995; 11:1423–1425.
9. Korber B and Myers G: Signature pattern analysis: A method for assessing viral sequence relatedness. AIDS Res Hum Retroviruses 1992;8:1549–1559.
10. Felsenstein J: PHYLIP—phylogeny inference package. Cladistics 1989;5:164–166.
11. Swofford D: PAUP: Phylogenetic analysis using parsimony, version 3.1. Center for Biodiversity, Illinois Natural History Survey, Champaign, Illinois, 1993.

Address reprint requests to:
Adam Siepel or Bette Korber
T10 MS K710
Theoretical Biology and Biophysics
Los Alamos National Laboratory
Los Alamos, New Mexico 87545