AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 15, Number 3, 1999, pp. 247–253
Mary Ann Liebert, Inc.

Associations between Amino Acids in the Evolution of HIV Type 1 Protease Sequences under Indinavir Therapy

ANDREW J. LEIGH BROWN,[1] BETTE T. KORBER,[2] and JON H. CONDRA[3]

ABSTRACT

Significant diversity exists in amino acid sequences encoding HIV-1 protease in individuals naive for protease inhibitors, which could influence the rate of evolution of resistance. High-level resistance to indinavir requires multiple substitutions among at least 11 amino acid sites, and no single substitution was observed in all of 29 resistant isolates obtained from patients on long-term indinavir monotherapy. We have analyzed the evolution of PR in these sequences. The divergence from the baseline amino acid sequence by week 24 was 4%, increasing more than 7% by week 60. The mean difference between sequences from different patients at baseline was 6% (3–9%), rising to 10% after 40 weeks (3–16%), although at all time points nonsynonymous substitutions were less frequent than synonymous nucleotide changes. Analysis of associations between variants at different amino acid sites using a mutual information statistic revealed four pairs of sites to be significantly associated. In three cases these associations included residue 82. Clusters of baseline and week 24 amino acid sequences identified by maximum parsimony did not correlate significantly with the IC_{95} to indinavir, although a weak correlation of baseline clusters with phenotype at the week 24 time point was suggested.

INTRODUCTION

While inhibitors of HIV-1 protease have been shown in clinical trials to be highly effective antiretrovirals, the evolution of resistance has been described for all inhibitors tested. In an analysis of the development of resistance to indinavir (also known as Crixivan, MK-639, and L-735,524), isolates were obtained from 21 patients on long-term monotherapy.[1] In 17 patients, isolates that showed a significant reduction in sensitivity to the drug were obtained. Substitutions at up to 11 residues in the protease sequence were observed in several isolates. Regression analysis indicated eight of these were associated with a statistically significant effect.

The requirement that several amino acid substitutions occur before high-level resistance is attained is common to several protease inhibitors.[1–10] In this they differ from many inhibitors of reverse transcriptase, either nucleoside analogs such as lamivudine[11,12] or nonnucleoside inhibitors such as nevirapine.[13] High-level resistance to these antiretrovirals is achieved by the acquisition of single substitutions, M184V[14] and Y181C[15] in reverse transcriptase (RT), respectively. This is not exclusively the case for RT inhibitors, however, as zidovudine resistance is known to require substitutions at up to five sites to achieve maximum levels of resistance.[16] Possibly in consequence, maximum levels of resistance to zidovudine are not seen in all patients even after 1 year of monotherapy.[17]

The exact combination of mutations that gives rise to resistance to indinavir can differ markedly between patients.[1] In addition, the HIV-1 protease (PR) coding sequence is highly polymorphic in untreated patients.[18,19] *In vitro* studies have suggested that some amino acid substitutions are more likely to spread in the viral population if certain others are already present in the sequence.[20] The polymorphic nature of the HIV-1 PR means that there may be particular genotypes preexisting in some patients that differ in the extent to which they predispose to the evolution of drug resistance by certain routes. This would be particularly significant if some combinations of amino acids were more frequent in resistant strains than others.

Prediction of genotypic resistance from sequence data requires the recognition of associations between amino acids at different sites in the sequence, either preexisting or that have arisen during the course of therapy. The associations of inter-

[1]Centre for HIV Research Institute of Cell, Animal, and Population Biology, University of Edinburgh, Edinburgh EH9 3JN, Scotland.
[2]Theoretical Biology and Biophysics, Group T-10, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.
[3]Department of Antiviral Research, Merck Research Laboratories, West Point, Pennsylvania 19486.

est are those that have been selected by virtue of conferring a higher fitness on the virus. Associations can also arise for other reasons; these include chance, so any analysis must incorporate a rigorous statistical element, as well as a recently shared common ancestry among some of the sequences being analyzed. To assess the importance of this last source and as a check of the integrity of the data set, a phylogenetic analysis was performed on all sequences analyzed.

METHODS

Data on patients studied and phenotypic analysis of isolates obtained have been published previously.[1,10] Sequences were grouped for analysis according to time on treatment. Baseline (week 0) isolates from 13 patients were analyzed. Intermediate time points, varying between week 12 (three patients) and week 24 (eight patients), were available for nine of these and one other (collectively referred to as "week 24" isolates). Late time point isolates from 10 patients were studied. These were referred to as "week 60 isolates," but varied between week 32 (1 patient) and week 60 (5 patients). The subtype A sequence HIV-1_{U455} was included to root the trees (the protease sequence from the subtype D strain HIV-1_{ELI} was not sufficiently distinct from B subtype sequences to act as an outgroup).

Amino acid distances were calculated using a simple metric due to Kimura (program PROTDIST in the PHYLIP package, version 3.5c[21]). Maximum parsimony analyses were performed on amino acid sequences using the program PROTPARS, and the phylogenetic analysis of the entire data set was performed using the neighbor-joining algorithm (program NEIGHBOR). A more complete discussion of the analysis that assured the integrity of this data set and the special considerations that obtain in a conserved region of HIV can be found at: *http://hiv-web.lanl.gov/HTML/Contam/contam_conserved.html*. Frequencies of synonymous and nonsynonymous nucleotide substitutions were calculated using MEGA.[22]

For analysis of association between residues at variable sites, a method developed by A. Lapedes and B. Korber (Los Alamos National Laboratory, Los Alamos, NM) was used that was earlier applied to HIV *env* V3 sequences.[23] This uses a maximum likelihood approach to estimate the association and calculates a "mutual information statistic" between all pairs of sites, analogous to the linkage disequilibrium parameter. The power of tests for association depends on both sample size and the level of variation at the sites considered. For this purpose, a variable site was defined as one where the commonest amino acid had a frequency of not more than 85%. Monte Carlo randomization of the dataset was used to estimate the probability of obtaining the observed scores by chance. One thousand such randomizations were performed.

RESULTS

Analysis of complete sequences

To summarize the evolutionary change that had occurred in these sequences, a majority-rule consensus was obtained for each isolate. The average divergence between consensus sequences from different patients in baseline samples was 6% (range, 3–9%), increasing to 10% by week 60. This variation accrued at a large number of sites in the 99-amino acid PR sequence: 19 sites varied among 13 patients at week 0, but in the week 60 isolates 28 amino acid sites were variable. Remarkably, despite the variation in protein sequence, there was more divergence at synonymous nucleotide sites than at nonsynonymous sites at all time points (Table 1).

The distance between the consensus sequence of the week 0 isolate and those obtained from the week 24 and week 60 isolates increased in almost all patients (Fig. 1), with the mean amino acid distance from baseline increasing from 4.0% (week 0–week 24) to 7.3% (week 0–week 60). However, there was substantial variability between patients in the rate at which sequence change accumulated, despite the appearance of the same amino acid changes in several patients.[1]

A phylogenetic analysis of all sequences in the complete data set was carried out using nucleotide and synonymous substitution distances to check for any possible misidentification. Of 422 sequences from different patients in the original data set, 3 sequences that clustered more clearly with those from another patient than with other sequences from the same patient were found. These sequences were removed from the analysis (the presence of these sequences would not have influenced the conclusions of the original study[1]). A neighbor-joining tree obtained for all nonidentical sequences confirmed the conclusions of the analysis of consensus sequences given above: samples from the same patient clustered together, although with relatively long branches to the later sequences and bootstrap support for groups containing all sequences from a patient was often only 40% or less (not shown). Sequences from baseline samples from different patients are more similar than are sequences of later isolates and, as expected, the baseline sequences are usually located near the root of the latter samples from the same patient. Again, despite the acquisition of a limited group of resistance-associated substitutions, later sequences from different patients do not usually become more similar to each other overall.

Association of mutations at different sites

In the analysis of associations between sites, it is important to avoid identifying associations between amino acid sites that arise solely from common ancestry. In this analysis a single

TABLE 1. DIVERSITY IN PROTEASE SEQUENCES ACCUMULATED UNDER INDINAVIR THERAPY

Time point (week)	N	*Amino acid (%)*[a]	d_n *(%)*[b]	d_s *(%)*[c]	d_s/d_n
0	13	6.0	2.8	9.1	3.2
24	13	9.1	4.3	8.3	1.9
60	11	10.3	4.9	10.0	2.1

[a]Mean proportion of amino acid differences between consensus sequences from *N* different patients at each time point.

[b]Mean proportion of nonsynonymous nucleotide differences per nonsynonymous site.

[c]Mean proportion of synonymous nucleotide differences per synonymous site.

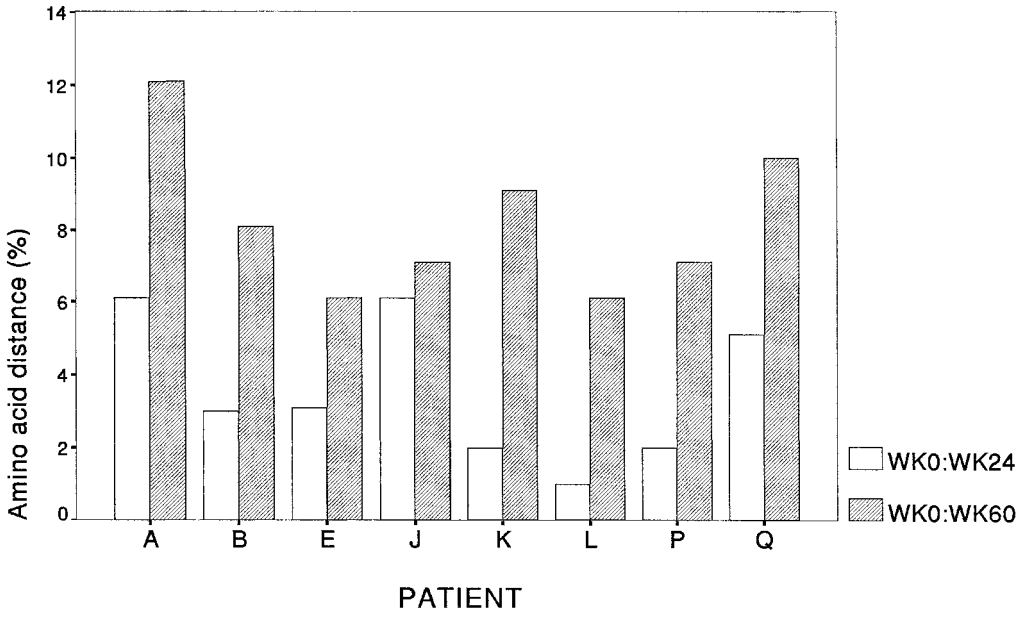

FIG. 1. Increase in amino acid distance from baseline sequence in relation to time on indinavir therapy. The amino acid distance between the consensus baseline sequence for each isolate and the consensus sequences of the week 24 and week 60 isolates is shown for eight patients for whom data from all three time points were available.

consensus sequence from each week 0 and week 60 patient isolate was used. The evolutionary distance between these two (Fig. 1) is so great for most patients that it was considered unlikely that amino acid pairs present in one would therefore be present in both. However, the main associations found were also detected in a reduced data set of week 60 sequences alone.

Associations were identified by the score of the "mutual information statistic" (M).[23] Their statistical significance was tested by a Monte Carlo procedure by comparison with the values observed by chance in multiple randomizations of the initial data set. Four significant associations were found (Table 2). The 3 strongest all involve residue 82, with residues 71, 54, and 10, respectively. Residues 10 and 54 were also significantly associated ($p < 0.05$), but this is explained by their mutual association with residue 82. No other associations between pairs of sites were significant, although for amino acids 63 and 64 the probability of observing a similar value by chance was about 20%. The relative location of these four covarying sites on the protease dimer is shown in Fig. 2.

The amino acid pairs observed at these sites are shown in Table 3, together with the associated value of the derived information statistic M' (which has a range from 0 to 1). In Table 4 the complete combinations observed at all four significant sites are shown. From these it can be seen that while substitutions at site 82 are strongly associated with high-level resistance, the appearance of six different genotypes at these four sites among only nine strains showing highest levels of resistance, confirms that viral populations can take multiple evolutionary paths to reduce sensitivity to this drug.

Predictive power of genotypic analysis

As we had detected statistically significant associations between pairs of amino acids in the PR sequence, we wished to investigate the possibility that similarities between genotypes might predict later response. Maximum parsimony clustering of amino acid sequences from each time point taken separately was performed and the phenotypic response for each group was analyzed. The lack of variation in phenotypic sensitivity observed at week 0 (all isolates fully sensitive) and week 60 (only 2 of 10 patients with an IC_{95} less than 1000 nM)[1] meant that only week 24 isolate values were suitable for analyzing associations between sequence clusters and resistance phenotype. At this time point values from 25 to 400 nM were observed. There was no clear association between genotypic clustering of week 24 samples and phenotype (data not shown).

We also examined whether the clustering of the baseline isolates is predictive of later phenotype. Among baseline sequences, 2 clusters were observed that contained 10 of the 13 patients (Fig. 3). The first of these contained patients L, E, I, and R, who at week 24 had a mean IC_{95} of 144 nM. The second contained patients J, H, N, K, P, and A, who at week 24 had a mean IC_{95} of 300 nM. Despite the small sample size, this difference approached significance ($t = 1.86$; $p = 0.1$). No sites were diag-

TABLE 2. ASSOCIATIONS BETWEEN AMINO ACID SITES DETECTED BY MUTUAL INFORMATION

Amino acid site		*Mutual information*	*Probability*
a	*b*	(M)	(P)
71	82	0.812	<0.01
54	82	0.529	<0.01
10	82	0.475	<0.05
10	54	0.472	<0.05
63	64	0.387	0.2

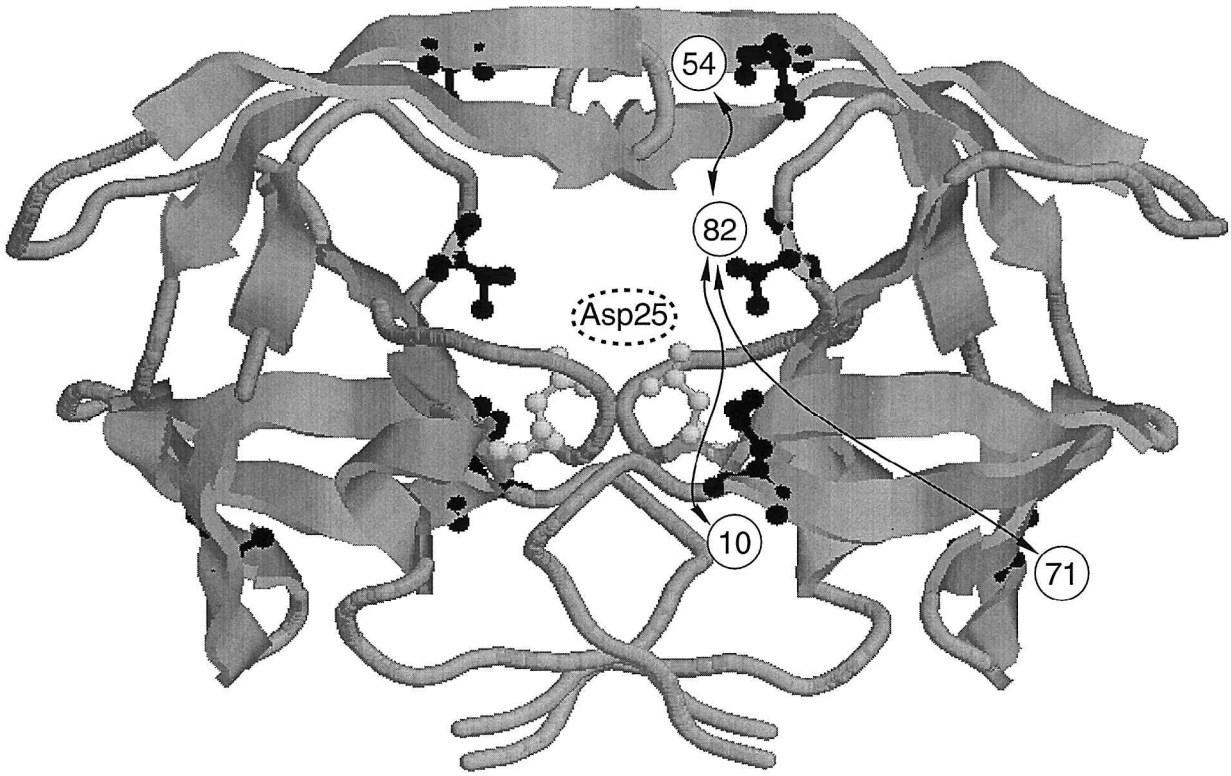

FIG. 2. Location of covarying amino acids on the protease dimer. The four covarying sites are indicated by representing each of their side chains as a black ball-and-stick structure. Note that the sites are not in particularly close proximity. The active site aspartic acid (residue 25) is indicated by a gray ball-and-stick structure. This diagram is based on the HIV_{LAI} protease structure with accession number 1hvk from the Protein Data Bank[23a] and was created using RasMol.

nostic for the two clusters but there were differences at amino acid 37, which was N (×3) or D (×1) in cluster 1 and S (×2), C (×2), or N (×2) in cluster 2. The second site that differed between the groups was site 63, which was P (×3) or H (×1) in cluster 1 and L (×3), S (×1), or P (×2) in cluster 2.

DISCUSSION

Extensive studies of variation in protease sequences and its association with phenotypic variation in sensitivity to many antivirals have clearly shown that the development of resistance

TABLE 3. PAIRWISE AMINO ACID ASSOCIATIONS AT ASSOCIATED SITES

Residue:	*10*	*54*	*63*	*64*	*71*	*82*		
Consensus:	*L*	*I*	*P*	*I*	*A*	*V*	*M′*	*No. pairs (total sequences = 34)*
Variants:					V (8)[a]	T (2)	0.01	1
					V (8)	A (7)	0.29	7
					T (1)	F (4)	0.06	1
		A (1)				A (7)	0.05	1
		V (8)				F (4)	0.23[b]	3
		V (8)				A (7)		4
	I (13)					F (4)	0.12	4
	I (13)					A (7)		5
	R (1)					T (2)	0.08	1
	I (13)	V (8)					0.1	7
	I (13)	A (1)						1
			L (6)	V (9)			0.13	5
			S (3)	M (1)			0.07	1

[a]The number of times each amino acid was observed is given in parentheses, and the number of times the pair was observed is given in the end column.

[b]The value of M' is the joint value for both pairs bracketed.

TABLE 4. OBSERVED AMINO ACID COMBINATIONS AT FOUR SIGNIFICANTLY ASSOCIATED SITES

Amino acid:	*10*	*54*	*71*	*82*	*Number of isolates*	*Mean IC_{95}*
Baseline consensus:	*L*	*I*	*A*	*V*	*15*	*118*
Variants:	V				1	100
			V	T	1	200
	I			A	1	800
			V	A	2	950
	I				4	1083
	I	V	T	A	1	1500
	I	V	V	A	4	2450
	R	V	V	T	1	3000
	I			F	1	3000
	I	V	T	F	1	3000
	I	V		F	2	3000

in this protein is multifactorial.[2–10] The previous study showed that even among the patient isolates most resistant to indinavir, many different genotypes were represented.[1]

To investigate further the evolution of indinavir resistance a number of analyses have been performed. These revealed that amino acid distance between baseline and week 60 samples was in many cases substantially higher than between baseline and week 24, indicating continued evolution of the sequence over a prolonged period and that sequence divergence between patients in PR increased over time (Table 1), although in one patient there was no increase between week 24 and week 60 (Fig. 1).

We have quantified the association between amino acid variants in the protease sequence during the evolution of resistance to indinavir *in vivo*. The strength of association has been tested by resampling the data set and four pairs of amino acid sites have been found to be significantly associated. All have previously been implicated as being associated with high-level resistance. In addition, specific pairs of amino acids at these sites that are strongly associated have been identified; these included V82A with A71V, I54V, and L10I. These results show that the contributions of the individual mutations to indinavir resistance cannot be quantified directly, because of their nonindependence. Nevertheless despite these strong and significant associations, six different genotypes were found at these four sites among the nine most resistant isolates (Table 4). The reasons why these specific amino acids covary are unclear, as they are located in different regions of the molecule, although subtle conformational effects may be involved, which might include alterations in interdomain flexibility.[24] Some may not be related to drug binding per se, but may rather be involved in the recovery of enzymatic activity lost as a consequence of the substitutions at position 82.

The identification of significant associations among amino acid sites has implications both for the evolution of resistance and its analysis. Previous studies have shown that to achieve levels of resistance to indinavir that can be detected in laboratory assays requires changes from wild type at (minimally) three amino acid sites. This raises the question of how these associations are brought about initially, at a stage when they would appear to impart little if any fitness difference. The proposal that HIV populations may have a restricted effective population site[25,26] would offer one mechanism—under these conditions chance effects could generate locally elevated frequencies of such combinations in viral subpopulations (demes) in solid tissue.[27] Once established in a deme, these variants have both an increased probability of exhibiting any fitness advantages they already have, and a higher probability of incorporating an additional mutation that would result in a significant fitness elevation.

One of the aspects of drug resistance of greatest interest from

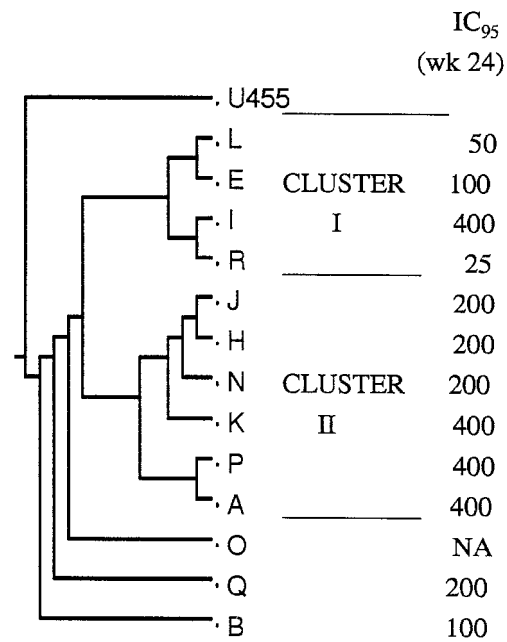

FIG. 3. Parsimony tree of sequences from baseline isolates of 13 patients compared with their week 24 IC_{95} value. A simple majority consensus-rule maximum parsimony tree was obtained using the PROTPARS program implemented in the PHYLIP package version 3.5c and the consensus sequence of each viral isolate.

the clinical perspective lies in the possibility of its prediction. The identification of significant nonrandom associations between amino acids involved in resistance is a basic requirement for such predictability and we have taken the analysis further by clustering the consensus amino acid sequences by maximum parsimony. Although there was no association between week 24 genotype clusters and resistance phenotype, analysis of consensus sequences of week 0 isolates revealed two major clades within the group that differed by a factor of 2 in their week 24 IC_{95} values, a result that approached statistical significance despite the small numbers of individuals involved.

A major paradox in the evolution of resistance to indinavir and ritonavir has been how the accumulation of multiple changes that are required for detectable levels of resistance comes about.[1,2,10] However, standard population genetics theory indicates that within populations under selection, fitness differences of as little as 10% can result in replacement of the population by the favored variant within five generations. Thus laboratory assays for drug susceptibility are probably not sufficiently sensitive to detect biologically significant differences. Such differences might, however, be detectable with appropriately designed *in vitro* competition assays.[28]

ACKNOWLEDGMENTS

We thank Carla Kuiken for help with the phylogenetic analysis, Charles Calef for assistance in the production of Fig. 2, and Daniel Holder for helpful comments on the manuscript. B.T.K. is supported by DOE-NIH interagency agreements Y01-70001 and YI-AI4058-03.

REFERENCES

1. Condra JH, Holder DJ, Schleif WA, *et al.*: Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. J Virol 1996;70:8270–8276.
2. Molla A, Korneyeva M, Gao Q, *et al.*: Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. Nature Med 1996;2:760–766.
3. Partaledis JA, Yamaguchi K, Tisdale M, *et al.*: In vitro selection and characterization of human immunodeficiency virus type 1 (HIV-1) isolates with reduced sensitivity to hydroxyethylamino sulfonamide inhibitors of HIV-1 aspartyl protease. J Virol 1995;69:5228–5235.
4. Patick AK, Rose R, Greytok J, *et al.*: Characterization of human immunodeficiency virus type 1 variant with reduced sensitivity to an aminodiol protease inhibitor. J Virol 1995;69:2148–2152.
5. King RW, Garber S, Winslow DL, *et al.*: Multiple mutations in the human immunodeficiency virus protease gene are responsible for decreased susceptibility to protease inhibitors. Antiviral Chem Chemother 1995;6:80–88.
6. Tisdale M, Myers RE, Maschera B, Parry NR, Oliver NM, and Blair ED: Cross-resistance analysis of human immunodeficiency virus type 1 variants individually selected for resistance to five different protease inhibitors. Antimicrob Agents Chemother 1995;39:1704–1710.
7. Kaplan AH, Michael SF, Wehbie RS, *et al.*: Selection of multiple human immunodeficiency virus type 1 variants that encode viral proteases with decreased sensitivity to an inhibitor of the viral protease. Proc Natl Acad Sci USA 1994;91:5597–5601.
8. Otto MJ, Garber S, Winslow DL, *et al.*: In vitro isolation and identification of human immunodeficiency virus (HIV) variants with reduced sensitivity to C-2 symmetrical inhibitors of HIV type 1 protease. Proc Natl Acad Sci USA 1993;90:7543–7547.
9. Markowitz M, Mo H, Kempf DJ, *et al.*: Selection and analysis of human immunodeficiency virus type 1 variants with increased resistance to ABT-538, a novel protease inhibitor. J Virol 1995;69:701–706.
10. Condra JH, Schleif WA, Blahy OM, *et al.*: In vivo emergence of HIV-1 variants resistant to multiple protease inhibitors. Nature (London) 1995;374:569–571.
11. Eron JJ, Benoit SL, Jemsek J, *et al.*: Treatment with lamivudine, zidovudine, or both in HIV-positive patients with 200 to 500 CD4+ cells per cubic millimeter. North American HIV Working Party. N Engl J Med 1995;333:1662–1669.
12. Ingrand D, Weber J, Boucher CAB, *et al.*: Phase I/II study of 3TC (lamivudine) in HIV-positive, asymptomatic or mild AIDS-related complex patients: Sustained reduction in viral markers. AIDS 1995;9:1323–1329.
13. Balzarini J, Karlsson A, Perez-Perez MJ, Camarasa MJ, Tarpley WG, and De Clercq E: Treatment of human immunodeficiency virus type 1 (HIV-1)-infected cells with combinations of HIV-1-specific inhibitors results in a different resistance pattern than does treatment with single drug therapy. J Virol 1993;67:5353–5359.
14. Schuurman R, Nijhuis M, van Leeuwen R, *et al.*: Rapid changes in human immunodeficiency virus type 1 RNA load and appearance of drug-resistant virus populations in persons treated with lamivudine (3TC). J Infect Dis 1995;171:1411–1419.
15. Richman DD, Havlir D, Corbeil J, *et al.*: Nevirapine resistance mutations of human immunodeficiency virus type 1 selected during therapy. J Virol 1994;68:1660–1666.
16. Kellam P, Boucher CA, Tijnagel JM, and Larder BA: Zidovudine treatment results in the selection of human immunodeficiency virus type 1 variants whose genotypes confer increasing levels of drug resistance. J Gen Virol 1994;75:341–351.
17. Richman DD, Grimes JM, and Lagakos SW: Effect of stage of disease and drug dose on zidovudine susceptibilities of isolates of human immunodeficiency virus. J Acquir Immune Defic Syndr 1990;3:743–746.
18. Lech W, Wang G, Yang YL, *et al.*: In vivo sequence diversity of the protease of human immunodeficiency virus type 1: Presence of protease inhibitor-resistant variants in untreated subjects. J Virol 1996;70:2038–2043.
19. Kozal MJ, Shah N, Shen N, *et al.*: Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nature Med 1996;2:753–759.
20. Rose RE, Gong YF, Greytok JA, *et al.*: Human immunodeficiency virus type 1 viral background plays a major role in development of resistance to protease inhibitors. Proc Natl Acad Sci USA 1996;93:1648–1653.
21. Felsenstein J: PHYLIP—phylogeny inference package (version 3.2). Cladistics 1989;5:164–166.
22. Kumar S, Tamura K, and Nei M: *MEGA: Molecular Evolutionary Genetics Analysis*, version 1.01. Pennsylvania State University, University Park, Philadelphia, 1993.
23. Korber BT, Farber RM, Wolpert DH, and Lapedes AS: Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. Proc Natl Acad Sci USA 1993;90:7176–7180.

23a. Hosur MV, Bhat TN, Kempf DJ, *et al.*: Influence of stereochemistry on activity and binding modes for C2 symmetry-based diol inhibitors of HIV protease. J Am Chem Soc 1994;116:847–855.

24. Rose RB, Craik CS, and Stroud RM: Domain flexibility in retroviral proteases: Structural implications for drug resistant mutations. Biochemistry 1998;37:2607–2621.

25. Leigh Brown AJ: Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective population number in the virus population. Proc Natl Acad Sci USA 1997;94:1862–1865.
26. Leigh Brown AJ and Richman DD: HIV-1: Gambling on the evolution of drug resistance? Nature Med 1997;3:268–271.
27. Frost SDW and Leigh Brown AJ: The genetic structure of HIV-1 populations within patients and the frequency of drug-resistant mutants. In preparation.
28. Maeda Y, Venzon DJ, and Mitsuya H: Altered drug sensitivity, fitness and evolution of human immunodeficiency virus type 1 with pol gene mutations conferring multi-dideoxynucleoside resistance. J Infect Dis 1998;177:1207–1213.

Address reprint requests to:
Andrew J. Leigh Brown
Centre for HIV Research
University of Edinburgh
Waddington Building, West Mains Road
Edinburgh EH9 3JN, Scotland