ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

VIROLOGY

Virology 314 (2003) 536–548

www.elsevier.com/locate/yviro

Covariation of amino acid positions in HIV-1 protease

Noah G. Hoffman,[a] Celia A. Schiffer,[b] and Ronald Swanstrom[a,*]

[a] *UNC Center for AIDS Research, University of North Carolina, Chapel Hill, NC 27599-7295, USA*

[b] *Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01655, USA*

Received 28 January 2003; returned to author for revision 8 April 2003; accepted 10 June 2003

Abstract

We have examined patterns of sequence variability for evidence of linked sequence changes in HIV-1 subtype B protease using translated sequences from protease inhibitor (PI) treated and untreated subjects downloaded from the Stanford HIV RT and Protease Sequence Database (http://hivdb.stanford.edu). The final data set size was 648 sequences from untreated subjects (notx) and 531 for PI-treated subjects (tx). Each subject was uniquely represented by a single sequence. Mutual information was calculated for all pairwise comparisons of positions with nonconsensus amino acids in at least 5% of sequences; significance of pairwise association was assessed using permutation tests. In addition pairs of positions were assessed for linkage by comparing the observed occurrences of amino acid combinations to expected values. The mutual information statistic indicated linkage between nine pairs of sites in the untreated data set (10:93, 12:19, 35:38, 37:41, 62:71, 63:64, 71:77, 71:93, 77:93). Strong statistical support for linkage in the treated data set was seen for 32 pairs, eight involving position 10:7 involving position 71, with the rest being 12:19, 15:77, 20:36, 30:88, 35:36, 35:37, 36:62, 36:77, 46:82, 46:84, 48:54, 48:82, 54:82, 63:64, 63:90, 73:90, 77:93, and 84:90. Most associations were positive, although negative associations were seen for five pairs of interactions. Structural proximity suggests that numerous pairs may interact within a local environment. These interactions include two distinct clusters around 36/77 and 71/93. While some of these interactions may reflect fortuitous linkage in heavily treated subjects with many resistance mutations, others will likely represent important cooperative interactions that are amenable to experimental validation.

Keywords: Protease; HIV-1; Resistance; Covariation; Mutual information; Protease inhibitor; Variation

Introduction

Activity of the human immunodeficiency virus type I (HIV-1) protease is required for the maturation of budding virions into infectious particles, making this protein an important target of antiretroviral drugs. Although the use of protease inhibitors (PIs) does not select for resistance as part of a fully suppressive regimen, viruses containing protease mutations conferring drug resistance are at a replicative advantage when complete suppression is not achieved.

Mutations in protease associated with reduced PI sensitivity are often classified as either primary or secondary. Primary mutations appear early in the evolution of inhibitor resistance: the major primary resistance mutations occur at positions 30, 48, 50, 82, 84, and 90. Although these positions have varying degrees of specificity for different inhibitors, substantial cross-resistance has been reported for most of them, particularly for positions 84 and 90 (inhibitor specificity of each of these mutations is reviewed in Hirsch et al., 2000; Kantor et al., 2001; Resch and Swanstrom, 2000; Shafer et al., 2000). Highly resistant viruses may accumulate more than one of these mutations, and the evolution to resistance for any one inhibitor can proceed along several different paths. All of these positions occur near the substrate-binding cleft of the protease except 90 (Wlodawer and Vondrasek, 1998) and usually directly reduce inhibitor binding to protease. In general, mutations conferring significant resistance are absent in untreated populations; this is consistent with the idea that these substitutions are themselves deleterious to virus replication and are maintained only in the presence of the inhibitor.

* Corresponding author. UNC Center for AIDS Research, Lineberger Bldg. Rm 22-006, CB 7295, University of North Carolina, Chapel Hill, NC 27599-7295. Fax: +1-919-966-8212.

E-mail address: risunc@med.unc.edu (R. Swanstrom).

0042-6822/03/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0042-6822(03)00484-7

No single mutation provides high-level resistance, which is generally attained only following the acquisition of one or more primary mutations in combination with one or more "secondary" mutations. Secondary mutations by themselves can be associated with low-level resistance and are common polymorphisms in untreated populations (Barrie et al., 1996; Birk and Sonnerborg, 1998; Kozal et al., 1996; Lech et al., 1996; Wegner et al., 2000). These latter mutations may compensate for a reduction in fitness (and enhance resistance) or may further reduce drug sensitivity but with an additional fitness loss (Mammano et al., 2000; Martinez-Picado et al., 1999; Resch et al., 2002; Zennou et al., 1998).

Both resistance mutations and other mutations occurring as part of the natural variability of protease often occur together at frequencies that would not be predicted according to chance; i.e., they covary. Covariation between amino acid positions in protease can be caused by (i) the specificity of the compensatory effect of one mutation for another; (ii) and additive reduction in sensitivity to a particular PI by two or more mutations in the presence of drug selection; or (iii) the preexistence of a mutation creating a favorable context for the emergence of a specific resistance mutation at another position (examples of these phenomena can be found in Condra et al., 1996; Mammano et al., 2000; Molla et al., 1996; Rose et al., 1996; Winters et al., 1998).

A number of studies have noted the existence of covariation between amino acid positions in protease (Jacobsen et al., 1996; Lech et al., 1996; Leigh Brown et al., 1999; Wu et al., 2003; Yahi et al., 1999). Data sets containing a relatively small number of sequences have limited the sensitivity of most of these previous studies for detecting covariation. Thanks to the availability of many carefully curated protease sequences at the Stanford HIV RT and Protease Sequence Database (Kantor et al., 2001), we have been able to search for evidence of covariation within a large data set of protease sequences isolated from both PI-naive subjects and those who have received PI therapy. Because covariation can also appear by chance or due to the recent common ancestry of groups of sequences, we have incorporated both a rigorous statistical element and phylogenetic analyses into our study.

Results and discussion

Initial measurements of amino acid variability among protease sequences from treated and untreated subjects

Amino acid positions in protease range in variability from completely conserved to having nonconsensus substitutions in nearly 50% of sequences. We restricted our analysis to positions containing nonconsensus residues in at least 5% of sequences. Thirty such "variable" positions were identified between the two data sets; position 88, a known resistance-associated position with 4.7% variability in the tx data set, was also included. Each of these 31 positions was assigned to one of three classes based on the relative frequency of amino acid substitution in the notx and tx sets (Fig. 1A). Class I positions were of approximately the same variability among notx and tx sequences (≤3% difference between the two data sets). Class II positions had different frequencies of substitution between notx and tx sequences (>3% difference between the two data sets) with substantial variability among notx sequences (>5%). In this class, positions 10 and 71 had the biggest increase in variability between the two data sets (three to fourfold), while the other positions differed by less than twofold, with variability at positions 72 and 77 changing the least (approaching Class I). Class III positions had very little variability in notx sequences (≤3%), with greater variability among tx sequences (Fig. 1A). In this last class, position 20 showed the greatest variability in the notx data set (approaching Class II). The results of the covariation analysis will be discussed below in the context of these three classes. These 31 positions are dispersed throughout the entire protease structure (Fig. 1B).

Analysis of covariation between positions in protease using mutual information

There are 465 possible pairwise combinations of the 31 variable positions; to permit a direct comparison of patterns of covariation within the notx and tx sequence sets, we calculated mutual information (*M*) values and uncertainty coefficients (*U*) for all 465 combinations of these positions in both data sets. As noted by Korber et al. (1993), the magnitude of *M* alone for any given pair of positions is a poor indicator of covariability without some measure of significance, which is the probability that the value of *M* as great as that observed for that pair of positions could have been achieved by chance. Accordingly, we used a permutation test to assess the significance of *M* for each pair of positions. An additional statistic, the uncertainty coefficient (*U*), was calculated for each pair of positions (Theil, 1972). *U* is a measure of the proportional reduction in error (PRE) in predicting one position based on another position and ranges from 0 (no correlation) to 1 (perfect correspondence). Tables 1 and 2 list $M(x, y)$, $U(x|y)$, $U(y|x)$, and *P* values of these measures for all pairs of positions reaching a significance of $<10^{-4}$ in the notx and tx sequence sets, respectively.

To describe the associations between specific amino acid substitutions, we constructed contingency tables indicating the number of sequences containing each combination of amino acids at pairs of positions with significant mutual information scores (Table 3). The magnitude of the positive or negative association (i.e., more or fewer sequences than expected containing the combination, respectively) between amino acids is indicated by ratios of observed to expected numbers of sequences. The significance of departures from expected values was calculated using another permutation test. For example, the bottom right cell of the first contingency table (describing 10:91) indicates that 41 sequences in the notx set contain both L101 and 193L; this combination was observed in 2.3 times more sequences than expected and has a significance of $P < 10^{-5}$ (as indicated by the asterisk).

There are two important limitations to consider in this analysis resulting from potential sources of bias in the data set. First, no attempt has been made to sort the sequences based on subject treatment. Ideally (for this type of analysis) a progression of mutations that accumulated during treatment with a single protease inhibitor could be analyzed. This has been done with smaller numbers of sequences for a few inhibitors (Condra et al., 1996; Jacobsen et al., 1996; Leigh Brown et al., 1999; Molla et al., 1996; Patick et al., 1998), but such data are not available in larger numbers. Thus the data set as we used it is biased in an uncontrolled way for the inhibitor used or for the use of multiple inhibitors either sequentially or together. The second concern is that there has been no attempt to look at the addition of sequential mutations. Our understanding of the resistance phenomenon incorporates an important temporal aspect to the accumulation of mutations. In some cases we have tried to infer a temporal relationship based on asymmetry in the appearance of single versus double mutations, but this interpretation must be made with caution.

Interactions involving positions with preexisting variability

Pairs varying together in both sets

The notx set of sequences represents the “natural” variability in protease; in this background, covariation between positions can be observed in the absence of the positive selection imposed on specific residues by protease inhibitor therapy. As expected, the majority of position-pairs covarying in the notx set showed a similar interaction in the tx set. The occurrence of similar patterns of covariation in both sets not only provides independent verification of specific interactions, but also suggests that certain positions play a role in general compensatory mechanisms. Pairs of statistically associated positions appearing in both sets included 35:37, 12:19, 71:93, 15:77, 36:77, 63:64, 77:93, 37:41, and 10:93 (some of these pairs did not achieve the strict significance cutoffs for both tests, but exhibited similar trends, as discussed below).

Plausible structural explanations for covariation demonstrated by a subset of these pairs (35:37, 12:19, 71:93, 15:77, and 36:77) are possible. For example, the E35D N37D combination occurred around twice as often as expected in both the notx and the tx sequence sets (Table 3). N37D appears less often than E35D, suggesting that N37D is stabilized in the presence of E35D to a greater extent than the reverse. This is also reflected in the values of $U(35|37)$ and $U(37|35)$ for these two positions which are 0.087 and 0.047, respectively, in the tx set (i.e., knowing the composition of position 37 reduces the uncertainty of position 35 more than vice versa). Positions 35 and 37 are near one another in the protease structure in the hinge region of the flap, but the side chains are 5.5 Å apart. E35 forms an ionic bond with R57. In the context of an E35D mutation, R57 might extend to maintain an interaction with the shorter aspartic acid side chain; this extension could also bring R57 within bonding distance of the side chain of position 37. Thus N37D would be stabilized by interacting with R57.

The significance of the interaction between positions 12 and 19 seems to be a need to maintain van der Waals packing between the most common single substitutions at these two sites, as the single mutations TI2S:L19 and TI2: L191 were relatively infrequent, occurring 1.9 and 1.5 fold less often than expected. In the wild-type structure (Vondrasek and Wlodawer, 1996) the side chains of 12 and 19, which project into the solvent from parallel strands of a β-sheet, make van der Waals contacts. The loss of the methyl group in T12S could be compensated by the repositioning of the methyl group in L19I in the double mutant.

Although $M(15,77)$ narrowly missed the significance cutoff in the notx set ($P = 0.0002$), the trend was similar in both sets, with a strong negative interaction between I15V and V771 (this combination of mutations occurred 2.6 and 3.4 times less often than expected in the notx and tx sets, respectively); thus I15V appears in the V77 background but is suppressed in the V77I background. The closest approach between these positions is 6.1 Å, so they probably do not interact directly (Fig. 2A). However, residues 15 and 77 both flank M36, and it is possible that the V77I mutation pushes M36 into an unfavorable interaction with I15V. Similarly, the presence of the V771 mutation makes M361 less frequent, likely due to an interaction between the 36 and 77 side chains. The side chain of V77 makes van der Waals contact with the side chain and backbone of M36. The M361:V771 double mutation is strongly suppressed, with the double mutant occurring 3.4- and 3.9-fold less often than expected in the notx and tx sets, respectively, probably due to a van der Waals clash ($M(36,77)$ was not significant in the notx set, but showed a trend with $P = 0.003$). This set of interactions is illustrated in Fig. 2A.

Two additional pairs of positions (63:64 and 77:93) covaried significantly in both sets; another pair, 37:41, covaried significantly in the notx set but narrowly missed our strict significance cutoff in the tx set (Table 3). There was no clear physical basis for the covariation displayed by any of these pairs of positions.

Pairs varying together in the untreated set only

Although pairs covarying in the absence of therapy might be expected to maintain similar interactions in the presence of protease inhibitors, we found that some detectable associations in the notx set were lost or diminished in the tx set, including 10:93, 71:93, 71:77, and 62:71. This could have been due to selection bias or a loss of power in the smaller tx data set (531 sequences versus 648 in the notx set). Alternatively, competing interactions or functions

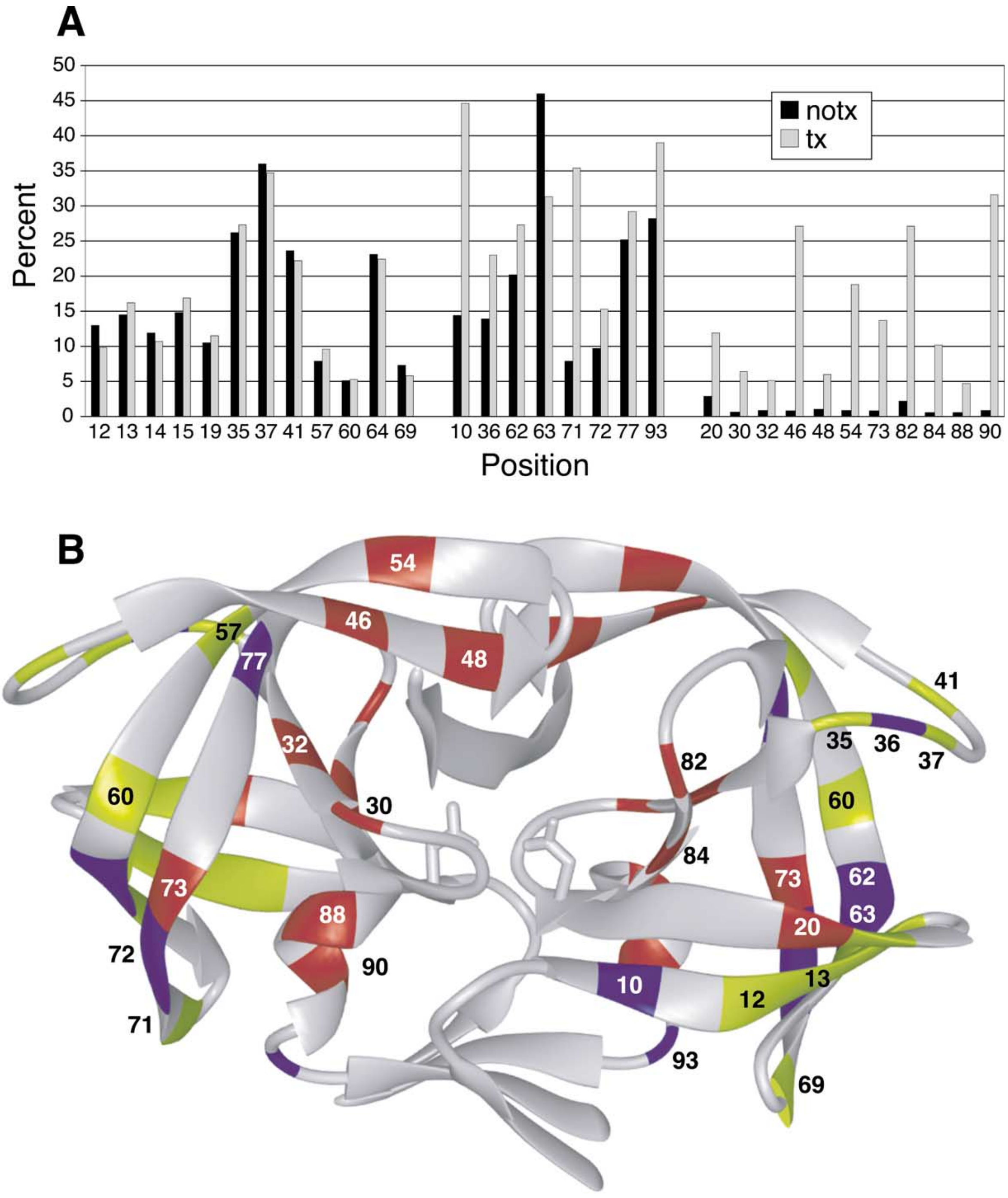

Fig. 1. (A) Frequency of nonconsensus amino acid substitution in the notx and tx sequence sets. Positions shown contained at least 5% nonconsensus amino acids (with the exception of position 88 at 4.7%) and are grouped into three classes from left to right: Class I positions were of approximately the same variability among notx and tx sequences; Class II positions had different frequencies of substitution between notx and tx sequences with substantial variability among notx sequences; Class III positions had very little variability in notx sequences, with higher levels of variability among tx sequences. (B) A ribbon diagram, made with MIDAS (Ferrin et al., 1988), of a substrate complex of the HIV-1 protease dimer (Prabu-Jeyabalan et al., 2000), showing the locations of the three different classes. Yellow are Class I; blue are Class II, and red are Class III.

emerging after the onset of therapy might diminish the importance of these interactions under selective pressure from protease inhibitors. This possibility is especially noteworthy for positions 10 and 71.

The most striking example of this phenomenon occurred for 71:93. Although the mutual information score describing this interaction reached significance in both sets, the value of $U(71|93)$, the largest among all pairs in the notx set

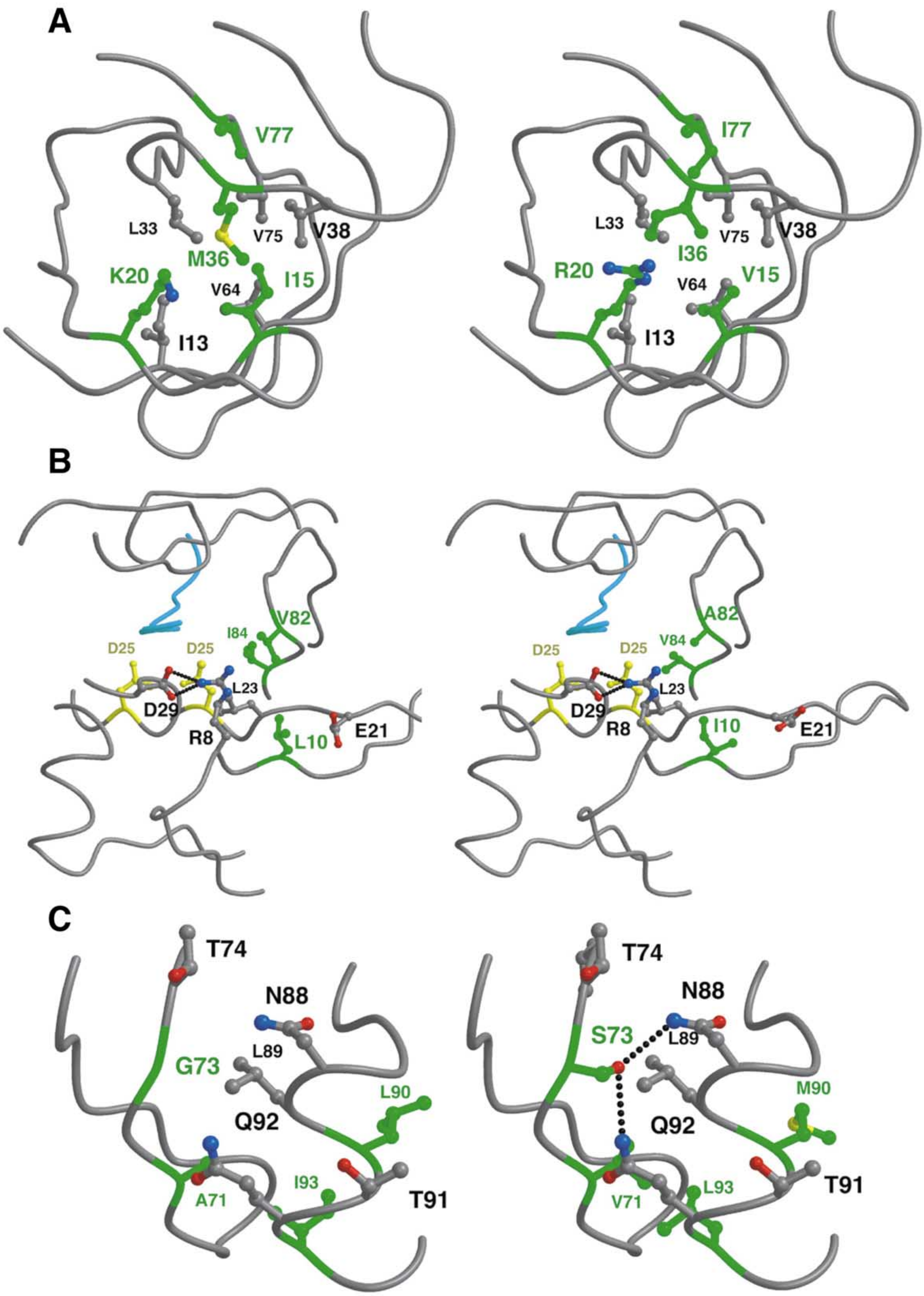

Fig. 2. Examination of sidechains of residues whose physical proximity may explain the associations of their respective mutational patterns. The highlighted residues are shown in green. Hydrogen or ionic bonds between side chains are shown with black dotted lines. The figures were made with MIDAS (Ferrin et al., 1988). In each panel the left figure shows the region within the crystal structure of a substrate complex (Prabu-Jeyabalan et al., 2000) and the right figure is a model with the most common mutations replaced at their respective sites using MIDAS. When necessary, side chains were rotated to alleviate a strong van der Waals clash. (A) Interactions among hydrophobic residues within the core of each monomer, including I15V, K20R, M36I, and V77I. The close proximity of these side chains could account for the interdependency of the mutational patterns. (B) L10I may act to prevent R8 from interacting with E21, especially in the context of either V82A and I84V. The catalytic aspartic acid residues are shown in yellow. (C) A71V, G73S, L90M, and I93L, all pack around the alpha helix, and the interdependence of the mutations may indicate a need to preserve the conformation of this region.

Table 1
Results of tests for covariation among amino acid positions in protease exhibiting variability in the notx data set

p2	dist	c1	c2	Drug-naïve subjects (notx)					Treated subjects (tx)				
				M	U(1|2)	U(2|1)	p(M)	p(E)	M	U(1|2)	U(2|1)	p(M)	p(E)
19	3.7	1	1	0.066	0.107	0.132	*	*	0.09	0.175	0.186	*	*
77	6.1	1	2	0.02	0.046	0.029	0.0002	0.00009	0.028	0.058	0.04	*	*
37	4.8	1	1	0.054	0.081	0.043	*	*	0.059	0.087	0.047	*	*
77	3.4	2	2	0.027	0.054	0.04	0.00304	0.00004	0.061	0.085	0.088	*	*
41	9.9	1	1	0.05	0.039	0.085	*	0.00002	0.044	0.035	0.078	0.00032	*
64	1.3	2	1	0.104	0.076	0.149	*	*	0.087	0.074	0.127	*	*
93	17.9	2	2	0.046	0.068	0.072	*	*	0.029	0.042	0.041	0.00004	0.00001
93	10.5	2	2	0.046	0.081	0.071	*	*	0.028	0.026	0.041	0.00208	0.00001
71	4.0	2	2	0.021	0.04	0.062	0.00008	0.00003	0.017	0.027	0.018	0.1564	0.00099
77	14.0	2	2	0.032	0.097	0.048	*	*	0.014	0.014	0.02	0.46496	0.012
93	4.4	2	2	0.055	0.163	0.086	*	*	0.032	0.033	0.046	0.00004	0.00018
71	15.4	2	2	0.011	0.019	0.032	0.2706	0.00966	0.081	0.074	0.083	*	*
36	1.3	1	2	0.025	0.037	0.049	0.00348	0.00219	0.066	0.098	0.093	*	*
62	7.5	2	2	0.016	0.033	0.031	0.01292	0.00574	0.053	0.074	0.082	*	*

Note. Pairs of positions shown received scores for *M* that were larger than expected with a significance of $P < 10^{-4}$ in at least one set of sequences. Columns are defined from left to right as follows: p1, p2, the positions compared, dist, atomic distance in angstroms of closest approach between positions *p*1 and *p*2 calculated from PDB 1A30 (pairs with dist ≤ 5.1 Å shaded black); c1, c2, the designated classes of *p*1 and *p*2 based on Fig. 2A; M, mutual information score; U(1|2), U(2|1), the uncertainty coefficient for position *p*1 given position *p*2, and *p*2 given *p*1, respectively; p(M), the probability that the observed value of *M* was due to chance; p(E), the lowest probability that the difference between observed and expected counts of a combination of amino acids represented at a pair of positions was due to chance (see Table 3). An asterisk (*) indicates $P < 10^{-5}$. Cells containing values corresponding to *M*, *U*, and *p*(*M*) are shaded if $p(M) < 10^{-4}$, likewise, cells corresponding to *p*(*E*) are shaded if $p(E) < 10^{-4}$.

at 0.163, was reduced by almost a factor of 5 in the tx set. In addition, the interaction between specific amino acids was significant only in the notx set, in which A71V was overrepresented in the 193L background by more than threefold (A71T193L was also overrepresented, but not significantly). It is possible that the lengthening of the side chain of 71 by A71T or A71V is better accommodated by a leucine than an isoleucine at position 93 (Fig. 2C).

Table 2
Results of tests for covariation between pairs of amino acid positions in protease in which at least one position varied in the tx set only (i.e., Class II/Class III and Class III/Class III interactions)

p2	dist	c1	c2	M	U(1|2)	U(2|1)
46	21.8	2	3	0.079	0.072	0.096
48	18.0	2	3	0.039	0.035	0.155
54	16.2	2	3	0.116	0.106	0.177
73	18.9	2	3	0.071	0.065	0.124
82	5.0	2	3	0.121	0.11	0.146
84	7.7	2	3	0.045	0.041	0.121
90	10.0	2	3	0.054	0.05	0.081
36	3.5	3	2	0.104	0.191	0.145
71	21.1	3	2	0.091	0.139	0.093
90	11.5	2	3	0.071	0.06	0.105
73	4.5	2	3	0.093	0.095	0.162
82	17.3	2	3	0.105	0.107	0.126
84	11.7	2	3	0.046	0.047	0.124
90	6.7	2	3	0.112	0.114	0.167
88	3.5	3	3	0.027	0.103	0.114
82	14.2	3	3	0.056	0.068	0.068
84	11.4	3	3	0.035	0.043	0.094
54	5.1	3	3	0.091	0.362	0.138
82	12.0	3	3	0.046	0.184	0.056
82	8.0	3	3	0.23	0.35	0.277
90	8.9	3	3	0.118	0.207	0.176
90	7.1	3	3	0.051	0.137	0.076

Note. Column labels are as defined in the legend for Table 1. *P* values are not shown; the significance of both *M* and *E* for all pairs shown achieved $P < 10^{-5}$ except for *M*(10,84) and *M*(30,88), for which *P* values were 0.00004 and 0.00008, respectively.

Similarly, no significant covariation was observed for 10:93, 71:77, or 62:71 in the tx set, despite significant interactions in the untreated set. None of these pairs contain positions that are near each other in the protease structure (Fig. 1B).

Pairs varying together in the treated set only

A final set of pairs involving positions that exhibit substantial variability in the notx set (i.e., Class I and Class II) covary only in the treated set, suggesting that their interaction is dependent on selective pressure from inhibitor therapy. These pairs include 10:71, 35:36, and 36:62. The E35D:M361 combination appeared 1.9 times more often than expected among tx sequences. These side chains lie on opposite sides of the loop at the hinge of the flap, but do not directly interact. In addition, L101:A71V and M361:162V each appeared 1.8-fold over expected values in the tx set. Neither of these pairs contains positions that are close in the structure and both positions are fairly variable, especially in the tx data set, making it difficult to assess the possible significance of these interactions. An alternative explanation for their statistical association is that these positions are independently selected through interactions with other positions and thus appear linked.

Table 3
Contingency tables for selected pairs of positions

amino acid and position [frequency]
count of sequences (*obs/exp*) p-value

notx (N = 648)

10,93	I93 [0.72]	I93L [0.27]
L10 [0.86]	423(1.1) *	128(-1.2) *
L10I [0.10]	25(-1.9) *	41(2.3) *

12,19	L19 [0.90]	L19I [0.06]
T12 [0.87]	522(1.0) *	21(-1.5) *
T12S [0.03]	10(-1.9) *	10(10.0) *

15,77	V77 [0.75]	V77I [0.23]
I15 [0.85]	399(-1.0) p=9e-05	138(1.1) p=0.0001
I15V [0.15]	85(1.2) p=9e-05	8(-2.6) p=0.0001

35,37	N37 [0.64]	N37D [0.10]
E35 [0.74]	334(1.1) *	31(-1.6) *
E35D [0.24]	76(-1.3) *	33(2.1) p=1e-05

37,41	R41 [0.76]	R41K [0.23]
N37D [0.10]	64(1.2) p=6e-05	3(-5.3) p=2e-05
N37H [0.02]	2(-4.0) p=6e-05	9(3.0) p=5e-05

62,71	A71 [0.92]	A71V [0.04]
I62 [0.80]	488(1.0) p=8e-05	11(-1.8) p=3e-05
I62V [0.20]	108(-1.1) p=0.0005	13(2.6) p=0.0002

63,64	I64 [0.77]	I64V [0.19]
P63 [0.54]	306(1.1) *	29(-2.2) *
P63L [0.27]	106(-1.3) *	68(2.1) *
P63Q [0.02]	3(-2.7) p=0.0007	6(3.0)

71,77	V77 [0.75]	V77I [0.23]
A71 [0.92]	466(1.0) *	116(-1.2) *
A71T [0.04]	9(-2.1) p=3e-05	15(2.5) p=4e-05
A71V [0.04]	10(-1.9) p=0.0002	14(2.3) p=0.0003

71,93	I93 [0.72]	I93L [0.27]
A71 [0.92]	453(1.1) *	140(-1.2) *
A71T [0.04]	10(-1.8) p=0.0007	14(2.0)
A71V [0.04]	2(-9.0) *	23(3.3) *

77,93	I93 [0.72]	I93L [0.27]
V77 [0.75]	384(1.1) *	96(-1.4) *
V77I [0.23]	71(-1.5) *	75(1.9) *

tx (N = 531)

10,46	M46 [0.73]	M46I [0.17]	M46L [0.08]
L10 [0.55]	251(1.2) *	25(-2.0) *	13(-1.8) p=0.0007
L10I [0.34]	107(-1.2) *	44(1.4)	24(1.7)

10,48	G48 [0.94]	G48V [0.05]
L10 [0.55]	290(1.1) *	3(-5.3) *
L10I [0.34]	153(-1.1) *	24(2.4) *

10,54	I54 [0.81]	I54V [0.14]	I54T [0.02]
L10 [0.55]	277(1.2) *	11(-3.8) *	0(0/6) p=0.0003
L10I [0.34]	108(-1.3) *	56(2.2) *	9(3.0) p=0.0003

10,71	A71 [0.65]	A71V [0.23]
L10 [0.55]	227(1.2) *	36(-1.9) *
L10I [0.34]	79(-1.5) *	74(1.8) *

10,71	A71 [0.65]	A71V [0.23]
L10 [0.55]	227(1.2) *	36(-1.9) *
L10I [0.34]	79(-1.5) *	74(1.8) *

10,73	G73 [0.86]	G73S [0.08]	G73T [0.02]
L10 [0.55]	276(1.1) *	12(-2.1) p=3e-05	0(0/6) p=0.0001
L10I [0.34]	133(-1.2) *	30(2.0) *	10(2.5) p=0.0001

10,82	V82 [0.73]	V82A [0.20]	V82T [0.03]
L10 [0.55]	256(1.2) *	27(-2.2) *	4(-2.5)
L10I [0.34]	92(-1.4) *	71(2.0) *	6(-1.0)
L10R [0.02]	4(-1.8)	0(0/2)	5(5/0) *

10,84	I84 [0.90]	I84V [0.09]
L10 [0.55]	282(1.1) *	10(-2.7) *
L10I [0.34]	144(-1.1) *	33(1.9) *

10,90	L90 [0.68]	L90M [0.31]
L10 [0.55]	238(1.2) *	53(-1.7) *
L10I [0.34]	93(-1.3) *	84(1.5) *

10,93	I93 [0.61]	I93L [0.39]
L10 [0.55]	205(1.1) p=1e-05	87(-1.3) p=1e-05
L10I [0.34]	87(-1.3) p=2e-05	92(1.3) p=2e-05

12,19	L19 [0.89]	L19I [0.07]	L19V [0.01]
T12 [0.90]	441(1.0) *	26(-1.3) p=1e-05	2(-3.0) p=0.0001
T12S [0.03]	11(-1.4)	2(2.0)	3(3/0) p=0.0009

15,77	V77 [0.71]	V77I [0.27]
I15 [0.83]	294(-1.1) *	137(1.1) p=1e-05
I15V [0.17]	81(1.3) *	7(-3.4) *

20,36	M36 [0.77]	M36I [0.19]
K20 [0.88]	391(1.1) *	60(-1.5) *
K20R [0.05]	5(-4.4) *	22(4.4) *

30,88	N88 [0.95]	N88D [0.03]
D30 [0.94]	480(1.0) p=0.0001	7(-2.0) *
D30N [0.06]	23(-1.3) p=6e-05	8(8.0) *

35,36	M36 [0.77]	M36I [0.19]
E35 [0.73]	328(1.1) *	48(-1.5) *
E35D [0.25]	78(-1.3) *	47(1.9) *

35,37	N37 [0.65]	N37D [0.14]
E35 [0.73]	281(1.1) *	36(-1.4) p=3e-05
E35D [0.25]	60(-1.5) *	33(1.8) p=6e-05

36,62	I62 [0.73]	I62V [0.26]
M36 [0.77]	322(1.1) *	83(-1.3) *
M36I [0.19]	53(-1.4) *	46(1.8) *

36,77	V77 [0.71]	V77I [0.27]
M36 [0.77]	265(-1.1) *	136(1.2) *
M36I [0.19]	92(1.3) *	7(-3.9) *

37,41	R41 [0.78]	R41K [0.21]
N37D [0.14]	66(1.2) p=0.0008	5(-3.0) p=0.0003
N37H [0.02]	1(-8.0) *	9(4.5) *

46,82	V82 [0.73]	V82A [0.20]
M46 [0.73]	308(1.1) *	59(-1.3) p=1e-05
M46L [0.08]	16(-1.9) *	24(3.0) *

46,84	I84 [0.90]	I84V [0.09]
M46 [0.73]	363(1.0) *	23(-1.6) p=4e-05
M46I [0.17]	66(-1.2) *	21(2.6) p=2e-05

48,54	I54 [0.81]	I54V [0.14]	I54T [0.02]
G48 [0.94]	423(1.0) *	61(-1.1) p=1e-05	0(0/9) *
G48V [0.05]	7(-3.4) *	12(3.0) p=0.0002	10(10.0) *

48,82	V82 [0.73]	V82A [0.20]
G48 [0.94]	379(1.0) *	84(-1.2) *
G48V [0.05]	7(-3.0) *	21(3.5) *

54,71	A71 [0.65]	A71V [0.23]
I54 [0.81]	307(1.1) *	74(-1.4) *
I54V [0.14]	27(-1.8) *	36(2.0) *
I54T [0.02]	0(0/6) p=3e-05	9(4.5) p=3e-05

54,82	V82 [0.73]	V82A [0.20]
I54 [0.81]	373(1.2) *	34(-2.6) *
I54V [0.14]	9(-6.1) *	57(3.8) *
I54T [0.02]	0(0/7) p=1e-05	10(5.0) *

63,64	I64 [0.78]	I64V [0.18]
P63 [0.69]	310(1.1) *	38(-1.8) *
P63L [0.15]	41(-1.5) *	36(2.6) *

63,90	L90 [0.68]	L90M [0.31]
P63 [0.69]	216(-1.2) *	147(1.3) *
P63L [0.15]	70(1.3) *	6(-4.0) *

71,73	G73 [0.86]	G73S [0.08]	G73T [0.02]
A71 [0.65]	313(1.1) *	7(-4.1) *	7(-1.0)
A71V [0.23]	89(-1.2) *	34(3.1) *	1(-3.0)
A71I [0.01]	1(-3.0)	0(0/0)	3(3/0) *

71,82	V82 [0.73]	V82A [0.20]
A71 [0.65]	285(1.1) *	35(-2.0) *
A71V [0.23]	55(-1.6) *	59(2.4) *

71,84	I84 [0.90]	I84V [0.09]
A71 [0.65]	323(1.0) p=2e-05	17(-1.9) *
A71V [0.23]	98(-1.1) p=2e-05	26(2.4) *

71,90	L90 [0.68]	L90M [0.31]
A71 [0.65]	286(1.2) *	53(-2.0) *
A71V [0.23]	48(-1.8) *	75(2.0) *

71,93	I93 [0.61]	I93L [0.39]
A71 [0.65]	229(1.1) p=0.0002	114(-1.2) p=0.0004

73,90	L90 [0.68]	L90M [0.31]
G73 [0.86]	352(1.1) *	102(-1.4) *
G73S [0.08]	6(-5.2) *	39(2.8) *
G73T [0.02]	1(-8.0) p=0.0001	10(3.3) p=6e-05

77,93	I93 [0.61]	I93L [0.39]
V77 [0.71]	254(1.1) p=1e-05	122(-1.2) p=1e-05
V77I [0.27]	67(-1.3) p=2e-05	76(1.4) p=4e-05

84,90	L90 [0.68]	L90M [0.31]
I84 [0.90]	348(1.1) *	124(-1.2) *
I84V [0.09]	12(-2.8) *	37(2.5) *

Note. The identity of the pair of positions is indicated in bold in a box at the top left of each table. Row and column labels in each contingency table indicate a specific amino acid followed by its overall frequency at that position in square brackets. Rows and columns contain amino acids at each of the two positions being compared in descending order of abundance; thus the first row and column of each table corresponds to the consensus amino acids. Each cell of the table contains the three following values: (i) the observed count of sequences containing the indicated combination of amino acids; (ii) in parentheses, the ratio of observed (*obs*) to expected (*exp*) counts of sequences calculated as *obs/exp* if *obs* > *exp* and $-1 \times$ *exp/obs* if *exp* > *obs* (hence negative values indicate fewer sequences containing this combination of amino acids than expected. If either value is 0, the raw count values are shown instead); (iii) the uncorrected *p* value describing the significance of the difference between the observed and expected values. An asterisk (*) indicates $P < 10^{-5}$. The contingency table describing the amino acid composition of positions 10 and 93 in the notx set (top left of the page) is annotated. For example, the bottom right cell of the first contingency table (describing 10:91) indicates that 41 sequences in the notx set contain both L10I and 193L; this combination was observed in 2.3 times more sequences than expected and has a significance of $P < 10^{-5}$ (as indicated by the asterisk). Rows or columns not containing at least one amino acid combination with $P < 0.001$ are not shown, nor are individual P values > 0.001.

Pairs involving at least one Class III position

The majority of position linkages identified, and those with the greatest overall magnitude and significance, involved Class III positions, i.e., positions which were variable only in the treated set (Tables 2 and 3).

Interactions involving Class II positions: 10, 36, 71, and 63

Positions 10 and 71 are statistically linked with a number positions, most of which are not nearby in the structure. Mutations at both positions 10 and 71 frequently appear in clinical samples from subjects who have failed protease inhibitor therapy, and these mutations can be selected by passage of HIV-1 in the presence of a protease inhibitor in vitro (Gong et al., 2000; Patick et al., 1996; Smidt et al., 1997; Vaillancourt et al., 1999; Watkins et al., 2003). Mutations at either one of positions 10 and 71 have previously been shown to rescue or improve the infectivity of viruses carrying other mutations associated with protease inhibitor resistance (Mammano et al., 2000; Nijhuis et al., 1999; Rose et al., 1996). The association of these positions with a broad array of other substitutions in the tx data set (10 with 46, 48, 54, 71, 73, 82, 84, 90, and 93; and 71 with 10, 54, 73, 82, 84, 90, and 93) suggests that both may function as compensatory mutations through an indirect mechanism. Similar to other Class II positions, substantial variability is present at 10 and 71 in the notx data set; however, the especially large increase in variability at these positions in the tx set distinguishes these two positions from the others in this class (Fig. 1A).

L10I is associated with a number of different residues. However, these interactions are complicated by the accumulation of multiple mutations during the selection for higher levels of resistance. Virtually all of the examples of L10I in the Treated data set are in the background of active site mutations at position 82 and/or 84. Both the L10I:V82A and the L10I:I84V combinations occur about twice as often as expected (Table 3). One structural explanation for the statistical association between 10 and 82/84 is as follows: R8 forms a conserved ionic bond with D29 at the base of the active site. Flanking R8, L10, together with L23, V82 and 184, form a hydrophobic barrier. On the other side of this barrier is the electronegative E21. The loss of hydrophobic surface when either V82A or 184V mutations occur could cause an opening in this barrier which may allow increased interaction between the side chains of R8 and E21, potentially disrupting the active site (Fig. 2B). The L10I mutation may provide steric interference to E21, preventing it from adopting a conformation that would allow it to attract the positively charged R8 and therefore compensating for the loss of hydrophobic surface resulting from a mutation at 82 or 84.

The apparent association of position 10 with mutations at positions 46, 48, 54, 73, and 90 may reflect either a more specific interaction with mutations at 82 or 84, with which many positions covary, or may reflect a general effect of fitness enhancement by changes at position 10 in the presence of resistance mutations. The role of L10I as a general enhancer could also explain its appearance in the untreated data set in which mutations at positions 82 and 84 are rare.

Like L10I, A71V is associated with specific changes at a number of different positions. It is difficult, therefore, to distinguish specific associations from a more general effect on protease activity. There is one example of specificity in the pattern of A71V associations that is suggested by structural proximity involving positions 73, 90, and 93. A71V shows an association with both G73S and L90M. In addition, both G73S and G73T were overrepresented in an L90M background (by 2.8- and 3.3-fold, respectively); the statistical correlation between these two Class III positions was among the strongest seen. Two-thirds of all occurrences of G73S in the tx set were in an A71V/L90M background, 10-fold above the expected number. Thus, while A71V may have a more global effect on enzyme activity, the effect on L90M is likely to be enhanced by the presence of the G73S mutation. Covariation between positions 71 and 90 (Jacobsen et al., 1996; Yahi et al., 1999) and 71 and 82 (Jacobsen et al., 1996; Leigh Brown et al., 1999) has been described.

Neither A71V nor G73S makes direct contact with the L90M side chain. However, residues 71 and 73 are located on a strand adjacent to the helix containing 90 and 93. An L90M mutation results in an increase in the size of the residue, and an expansion of this region by this substitution, might be stabilized from the other side of this helix by A71V and G73S substitutions. The replacement of G73 with either Ser or Thr introduces a hydroxyl group that is in a position to interact with the side-chain amines of positions N88 and Q92. Most mutations at 73 occur in the background of L90M mutations, suggesting that they are compensatory for L90M. The new 73/88/92 interactions may increase the stability of the alpha helix to allow it to accommodate the larger L90M residue. A71V could provide stability to the helix by packing against the base of the helix; a resulting clash with I93 might select for the I93L mutation (Fig. 2C).

Like certain mutations at 10 and 71, P63 has been shown to have an effect on protease activity and also to act as a compensatory mutation for some distal protease inhibitor resistance mutations: P63 is known to have a positive impact on the fitness of a virus carrying D30N and L90M (Martinez-Picado et al., 1999) and confers modest resistance to indinavir and nelfinavir in combination with other substitutions (Ziermann et al., 2000). Consistent with these previous observations, we detected a negative interaction between P63L and L90M, indicating the importance of P63 in the presence of resistance mutations. Although it has been reported that a crystal structure of protease containing M46I and P63 exhibited conformational differences in the flap region compared to the “wild-type” protein (Chen et al., 1995), the specific mechanism by which P63 influences protease activity is unknown. Note that although the wild-type amino acid at position 63 has historically been considered Leu, the consensus was clearly Pro in both data sets: the notx and tx sequence sets contained P63 in 54 and 69% of sequences, respectively, compared to 27 and 15% for P63L; other studies have also noted that Pro is the most abundant amino acid at position 63 among sequences from PI-naive subjects (Kozal et al., 1996; Shafer et al., 1999). It will be important to determine the mechanism by which mutations at positions 10, 71, and 63 modify protease activity in the presence of a wide array of other mutations.

Another instance of covariation in this group appears to have a structural basis: the side chains of K20 and M36 appear to interact (over 115). In the K20R:M36I double mutant, the longer Arg side chain would compensate for the shorter Ile side chain, allowing these two residues to maintain an interaction (Fig. 2A). The K20R substitution rarely occurs in the absence of M361, while the reverse is not true, suggesting that the latter change precedes the former.

Class III/Class III interactions

Class III mutations are those most strongly associated with resistance. While the detected interactions are statistically supported, other important interactions are likely not included due to the bias of the data set for isolates obtained following therapy failure with only a subset of the available protease inhibitors. It is also possible that some of the interactions detected in the current data set will not be selected as paired mutations with newer inhibitors even though one of the mutations is selected. Thus it will be important to determine the relevance of these and other interactions as data become available for the newer inhibitors. The current data set is dominated by sequences from subjects who failed therapy with saquinavir, ritonavir, and indinavir with a smaller component of nelfinavir failures.

Covariation between positions 30 and 88 in the tx data set is likely an extreme example of an inhibitor-specific interaction. Among subjects treated with currently available protease inhibitors, the D30N mutation is found only in those receiving nelfinavir (Kantor et al., 2001). In the tx set, half of all N88D mutations were in the D30N background, consistent with a previous report that most isolates from subjects failing nelfinavir therapy containing N88D usually contained D30N as well (Patick et al., 1998). Thus the association of D30N and N88D likely represents an enhancement of the D30N effect. The D30N mutation reduces a hydrogen bond with bound nelfinavir (Kaldor et al., 1997). N88D may contribute to nelfinavir resistance by further drawing the position 30 side chain away from the vicinity of the active site and the inhibitor, or by compensating for the loss of the negative charge on the side from the D30N mutation, which could affect the long-range electrostatics of the enzyme. N88S can also serve as a primary resistance mutation for nelfinavir and other protease inhibitors (Gong et al., 2000; Smidt et al., 1997; Ziermann et al., 2000). The observation that no examples of N88S were found with D30N (data not shown) suggests that N88D and N88S contribute to resistance by different mechanisms.

Interactions with position 46 represent another example of distinct patterns of covariation between different mutations at the same position. M46L is overrepresented threefold in a V82A background and is somewhat suppressed (although not significantly) in an I84V background. In contrast, M46I appears 2.6 times more often than expected with I84V ($P = 2 \times 10^{-5}$) and appears with V82A at near-expected frequencies. Conversely, nearly half of the I84V mutations in the data set appear with M46I. Position 46 is not near the active site residues 82 and 84, and thus the physical basis for this interaction is unclear. Indeed, it is likely that M46I and V82A/I84V owe their association to a general effect of the former mutation: an M46I mutant exhibited enhanced catalytic activity over the wild-type enzyme and improved the activity of a protease mutant containing both V82T and I84V (Schock et al., 1996). A different combination of mutations at 46 and 82, M46F and V82I, has also been reported to occur together more often than expected (Lech et al., 1996).

The pattern of codon usage at position 46 provides evidence that M46I and M46L lie along divergent evolutionary pathways, and that neither mutation typically gives rise to the other. Isoleucine, present in 18.5% of tx sequences, is always encoded by ATA, presumably the result of a third-position G-to-A transition from the wild-type ATG (methionine). Leucine, present in 8.2% of sequences, is encoded by TTG (6.8% of sequences), CTG (1.1%), or TTA (0.3%). Only the rare TTA codon is likely to arise from isoleucine; TTG and CTG probably result from first-position transversions from the consensus ATG. Thus neither M46I nor M46L typically serve as intermediates for each other in a stepwise trajectory to high-level resistance.

Positions 48, 54, and 82 represent a hierarchy of interactions. V82A is a common initial mutation for indinavir and ritonavir, and the subsequent addition of 154V is frequently seen, particularly in the evolution of resistance to ritonavir (Condra et al., 1996; Molla et al., 1996). Covariation between positions 54 and 82 has been reported among sequences from subjects treated with indinavir (Leigh Brown et al., 1999) and unspecified combination therapy (Yahi et al., 1999). The G48V mutation was first described as being associated with resistance to SQV (Jacobsen et al., 1996). In this data set, G48V mutations are overrepresented in the background of V82A and both 154V (narrowly missing our strict significance cutoff at $P = 0.0002$) and 154T. The 48:54 and 54:82 interactions were characterized by the highest values of U seen in this study. Because most 154V mutations occur in the absence of a G48V change, the latter is likely a late mutation added after changes at both 54 and 82 have occurred.

In contrast to 154V, all 10 examples of 154T mutations occur with G48V (and V82A), suggesting that the interaction of I54T with G48V is different than that of 154V. The addition of a hydroxyl group with the 154T mutation has the potential to provide backbone interactions at G49 and I47 and likely confers added bridging stability to the antiparallel

β-strands of the flap. The structure of the flap is likely to be somewhat perturbed by G48V as it slightly displaces F53, which normally packs against, and likely stabilizes, the outer surface of the flap.

Finally, I84V and L90M are associated. This association was noted previously among sequences from subjects treated with saquinavir (Jacobsen et al., 1996). L90M is an initial mutation in the evolution of resistance to SQV (Jacobsen et al., 1996), but its structural contribution is poorly understood. Changes at position 90 also occur in conjunction with therapy with most currently available protease inhibitors (reviewed in Shafer et al., 2000). I84V appears predominately in the L90M background, suggesting that L90M typically precedes I84V.

A recent study by Wu et al. (2003) has described covariation between positions in protease using different statistical methods; the results of that study are mostly in agreement with those reported here, particularly for the most strongly linked positions. A consistent feature of both studies is the extensive covariation between positions that are not close enough for a direct physical interaction. Although these authors tested the hypothesis that the covariation between physically distant positions could be explained by a chain of covarying residues, they reported that all potential chains that were identified could have been explained by chance. Thus the mechanism of the selective pressures responsible for covariation between distant positions in protease remains an open question.

Summary

In summary, we have used statistical tools to find significant associations between pairs of variable positions in HIV-1 protease sequences. Some associations are not affected by the presence of resistance mutations, some exist only as resistance mutations, and other associations appear to be obscured or lost when resistance mutations are present. Interactions around residues 36 and 77 and also around 71 and 93 represent distinct but linked clusters that may be important for regulating protease activity. We have provided a structural interpretation for many of these interactions, although not all of the interactions lend themselves to such an interpretation. Overall, this analysis provides a framework for a systematic characterization of the role of individual and linked amino acid changes in the evolution of resistance to protease inhibitors.

Materials and methods

All subtype B protease amino acid sequences from the Stanford HIV RT and Protease Resistance Database (Kantor et al., 2001) that had corresponding nucleotide sequences were downloaded on February 2, 2002. Most (approximately 75%) of downloaded sequences were identified as molecular clones in the database; to insure that other sequences were not derived from viral populations containing complex mixtures of *pro* genotypes, we excluded sequences containing more than two ambiguous codons. In the final data sets, no position contained an ambiguous codon in more than 2.5% of sequences. Sequences were classified as originating from subjects who had never received protease inhibitors (notx) or from those who had been treated with at least one protease inhibitor (tx). All but one representative of sets of identical amino acid sequences were discarded, and one sequence from each subject represented in the database was chosen at random. The resulting data sets contained 648 and 531 notx and tx sequences, respectively, and are available from the authors.

A total of 31 positions within protease was analyzed, based on the presence of a nonconsensus substitution in at least 5% of sequences in at least one of the sequence sets (position 88 was also included). Mutual information (*M*) was calculated for all possible 465 pairwise combinations of these 31 variable positions. This statistic has been used as a measure of covariation in the V3 region of envelope (Bickel et al., 1996; Korber et al., 1993) and protease sequences from patients treated with indinavir (Leigh Brown et al., 1999). *M* is the sum of the Shannon entropies at each of two positions minus the joint entropy of the two positions:

$$M(x, y) = H(x) + H(y) - H(x, y)$$

where

$$H(x) = -\sum_{i=1}^{m} P(x_i)\log P(x_i)$$

and

$$H(x, y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} P(x_i, y_j)\log P(x_i, y_j)$$

Here *m* and *n* are the numbers of different amino acids represented at positions *x* and *y*, respectively. $P(x_i)$ is the frequency of amino acid *i* at position *x*, and $P(x_i, y_j)$ is the frequency of each combination of amino acids x_i and y_j. If the amino acids at the two positions vary independently they will form many combinations and *H(x, y)* will be large, reducing the value of *M*(*x, y*), if the positions covary, there will be fewer combinations, and *M*(*x, y*) will remain relatively large (since *H*(*x, y*) is small). The significance of *M* was determined using a permutation test, in which the columns of positions are shuffled in place (breaking any association between the positions) and *M*(*x, y*) is recalculated for each shuffle to generate a reference distribution. Uncorrected *P* values describing the significance of *M*(*x, y*) were calculated as the number of shuffles in which the permuted value of *M* was greater than the original value, divided by the total number of shuffles performed. We used 10^5 permutations for the test, allowing us to assess significance to $P \geq 0.00001$. If no permuted values of *M* exceeded the

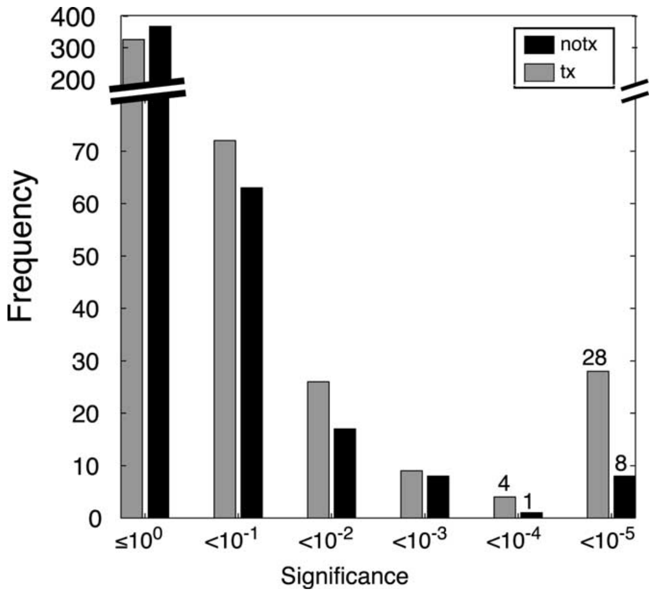

Fig. 3. Histogram of uncorrected *P* values describing the significance of *M* for all pairwise comparisons of positions in protease. *M* was calculated for all 465 pairs of 31 variable positions in the notx (black bars) and tx (gray bars) sequence sets, and the significance of each value of *M* was determined with a permutation test. Each bar indicates the number of pairs achieving a significance less than the value indicated below it on the *x*-axis, and ≥ the next label to the right (i.e., *P* values represented by the bar marked $< 10^{-2}$ fall in the range $10^{-3} \leq P < 10^{-2}$).

original value, then $P < 10^{-5}$. Note that the reference distribution against which any *M*(*x*, *y*) was compared was generated from random rearrangements of amino acids in position *x* with respect to *y*. We agree with Bickel et al., (1996) that this strategy for calculating significance is reasonable and less likely to exclude weak but real associations than the approach used in an earlier study (Korber et al., 1993).

We considered pairs of positions achieving $P < 10^{-4}$ as significantly covariant. At this cutoff, the significance level of the entire analysis (or the probability that at least one of the 465 possible combinations of 31 positions achieved $P < 10^{-4}$ by chance) is $<10^{-4} \times 465$, or less than 0.05 (Bickel et al., 1996). Fig. 3 shows the distribution of *P* values of the 465 pairs of positions in the notx and tx sequence sets. Most comparisons resulted in a *P* value between 0.1 and 1.0. A total of 9 (1.9%) and 32 (6.9%) of the 465 possible pairs of positions reached significance in the notx and tx sequence sets, respectively. Of these, five pairs were significant in both sets.

Although this permutation test describes the *significance* of an interaction between positions, it does not provide information about the *magnitude* of the observed covariation. Indeed, the correlation between the magnitude of *M* for any pair of positions and the significance of that value is poor ($R^2 = 0.26$ and 0.33 for the notx and tx sets, respectively). In addition, *M* is expressed in arbitrary units and fails to account for differences in variability between positions and therefore does not provide an easily interpretable measure of the strength of the interaction between two positions. Consequently, we also calculated the uncertainty coefficients *U(x|y)* and *U(y|x)* for each pair of positions (Theil, 1972)

$$U(x|y) = M(x, y)/H(x); U(y|x) = M(x, y)/H(y)$$

U(x|y) can be interpreted as the proportional reduction in error in predicting the composition of position *x* given the composition of position *y*. *U* is an asymmetrical measure and for positions *x* and *y*, *U(x|y)* is often not equal to *U(y|x)*. *U* is scaled from 0 to 1 and provides a more easily interpretable measure of the strength of the association between the positions than *M*. In addition, *U* also has a somewhat better correlation with significance than *M* ($R^2 = 0.39$ and 0.47 for the notx and tx sets, respectively). The significance of *M(x, y)*, *U(x|y)*, and *U(y|x)* for any *x* and *y* are identical, since the values of *H(x)* and *H(y)* are constant in each permutation.

Neither *M* nor *U* describe the underlying linkage between specific amino acids at a pair of positions. Therefore, for all pairs of positions achieving a value of *M* with significance of $P \leq 10^{-4}$ in either the notx or the tx sets (36 pairs), the specific associations among amino acids were examined in greater detail. For each pair of positions *x* and *y*, the occurrence of each combination of *m* amino acids at position *x* and *n* amino acids at position *y*, $N(x_i, y_j)$ ($1 \leq i \leq m$, $1 \leq j \leq n$), was tallied in $m \times n$ contingency tables. New contingency tables were then constructed from 10^5 randomizations of the data in which all columns of the alignment were shuffled in place, again with the intention of breaking any association between each pair of positions. For each permutation, a new $m \times n$ contingency table was calculated to generate a reference distribution of expected tallies of each possible combination of amino acids. Combinations of amino acids at *x* and *y* were present either more or less often than expected, and a significance level, $P(x_i, y_j)$, was assigned to each by counting the number of permuted values that were either ≤ or $\geq N(x_i, y_j)$, respectively, and by dividing by the number of permutations. As was the case for *M*, some combinations of amino acids were seen either more rarely or more frequently in the original sequences than in any of the permutations, resulting in a significance level of $P(x_i, y_j) < 10^{-5}$. For comparison, expected values (*E*) for each combination of amino acids at positions *x* and *y* were also calculated as $E(x_i, y_j) = F(x_i)F(y_j)/N$ (where *F* is a count of sequences containing amino acid x_i or y_j, and *N* is the total number of sequences); *E* was rounded to the nearest integer. As expected, $E(x_i, y_j)$ was approximately equal to the mean of the reference distribution of permuted values of $N(x_i, y_j)$.

Acknowledgments

All protease sequences used in this study were obtained from the Stanford HIV RT and Protease Sequence Database (http://hivdb.stanford.edu); we thank the curators of this site and its creator, Robert Shafer. This work was supported in part through NIH Grant R01-AI50485 to R.S, NIH Training Grants T32-A107419 and T32-A107001 to N.H., and NIH/NIGMS Grant P01 GM66524-01 to Celia Schiffer, and was performed under the auspices of the UNC Center for AIDS Research, supported by NIH Grant P30 A150410.

References

Barrie, K.A., Perez, E.E., Lamers, S.L., Farmerie, W.G., Dunn, B.M., Sleasman, J.W., Goodenow, M.M., 1996. Natural variation in HIV-1 protease, Gag p7 and p6 and protease cleavage sites within gag/pol polyproteins: amino acid substitutions in the absence of protease inhibitors in mothers and children infected by human immunodeficiency virus type 1. Virology 219, 407–416.

Bickel, P.J., Cosman, P.C., Olshen, R.A., Spector, P.C., Rodrigo, A.G., Mullins, J.I., 1996. Covariability of V3 loop amino acids. AIDS Res. Hum. Retroviruses 12, 1401–1411.

Birk, M., Sonnerborg, A., 1998. Variations in HIV-1 pol gene associated with reduced sensitivity to antiretroviral drugs in treatment-naive patients. AIDS 12, 2369–2375.

Chen, Z., Li, Y., Schock, H.B., Hall, D., Chen, E., Kuo, L.C., 1995. Three-dimensional structure of a mutant HIV-1 protease displaying cross-resistance to all protease inhibitors in clinical trials. J. Biol. Chem. 270, 21433–21436.

Condra, J.H., Holder, D.J., Schleif, W.A., Blahy, O.M., Danovich, R.M., Gabryelski, L.J., Graham, D.J., Laird, D., Quintero, J.C., Rhodes, A., Robbins, H.L., Roth, E., Shivaprakash, M., Yang, T., Chodakewitz, J.A., Deutsch, P.J., Leavitt, R.Y., Massari, F.E., Mellors, J.W., Squires, K.E., Steigbigel, R.T., Teppler, H., Emini, E.A., 1996. Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. J. Virol. 70, 8270–8276.

Ferrin, T.E., Huang, C.C., Jarvis, L.E., Langridge, R., 1988. The MIDAS display system. J. Mol. Graph. 6, 13–27.

Gong, Y.F., Robinson, B.S., Rose, R.E., Deminie, C., Spicer, T.P., Stock, D., Colonno, R.J., Lin, P.F., 2000. In vitro resistance profile of the human immunodeficiency virus type 1 protease inhibitor BMS-232632. Antimicrob. Agents Chemother. 44, 2319–2326.

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., Demeter, L.M., Vella, S., Jacobsen, D.M., Richman, D.D., 2000. Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society-USA Panel. J. Am. Med. Assoc. 283, 2417–2426.

Jacobsen, H., Hanggi, M., Ott, M., Duncan, I.B., Owen, S., Andreoni, M., Vella, S., Mous, J., 1996. In vivo resistance to a human immunodeficiency virus type 1 proteinase inhibitor: mutations, kinetics, and frequencies. J. Infect. Dis. 173, 1379–1387.

Kaldor, S.W., Kalish, V.J., Davies 2nd, J.F., Shetty, B.V., Fritz, J.E., Appelt, K., Burgess, J.A., Campanale, K.M., Chirgadze, N.Y., Clawson, D.K., Dressman, B.A., Hatch, S.D., Khalil, D.A., Kosa, M.B., Lubbehusen, P.P., Muesing, M.A., Patick, A.K., Reich, S.H., Su, K.S., Tatlock, J.H., 1997. Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease. J. Med. Chem. 40, 3979–3985.

Kantor, R., Machekano, R., Gonzales, M.J., Dupnik, K., Schapiro, J.M., Shafer, R.W., 2001. Human immunodeficiency virus reverse transcriptase and protease sequence database: an expanded data model integrating natural language text and sequence analysis programs. Nucleic Acids Res. 29, 296–299.

Korber, B.T., Farber, R.M., Wolpert, D.H., Lapedes, A.S., 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc. Natl. Acad. Sci. USA 90, 7176–7180.

Kozal, M.J., Shah, N., Shen, N., Yang, R., Fucini, R., Merigan, T.C., Richman, D.D., Morris, D., Hubbell, E., Chee, M., Gingeras, T.R., 1996. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat. Med. 2, 753–759.

Lech, W.J., Wang, G., Yang, Y.L., Chee, Y., Dorman, K., McCrae, D., Lazzeroni, L.C., Erickson, J.W., Sinsheimer, J.S., Kaplan, A.H., 1996. In vivo sequence diversity of the protease of human immunodeficiency virus type 1: presence of protease inhibitor-resistant variants in untreated subjects. J. Virol. 70, 2038–2043.

Leigh Brown, A.J., Korber, B.T., Condra, J.H., 1999. Associations between amino acids in the evolution of HIV type 1 protease sequences under indinavir therapy. AIDS Res. Hum. Retroviruses 15, 247–253.

Mammano, F., Trouplin, V., Zennou, V., Clavel, F., 2000. Retracing the evolutionary pathways of human immunodeficiency virus type 1 resistance to protease inhibitors virus fitness in the absence and in the presence of drug. J. Virol. 74, 8524–8531.

Martinez-Picado, J., Savara, A.V., Sutton, L., D'Aquila, R.T., 1999. Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1. J. Virol. 73, 3744–3752.

Molla, A., Korneyeva, M., Gao, Q., Vasavanonda, S., Schipper, P.J., Mo, H.M., Markowitz, M., Chernyavskiy, T., Niu, P., Lyons, N., Hsu, A., Granneman, G.R., Ho, D.D., Boucher, C.A., Leonard, J.M., Norbeck, D.W., Kempf, D.J., 1996. Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. Nat. Med. 2, 760–766.

Nijhuis, M., Schuurman, R., de Jong, D., Erickson, J., Gustchina, E., Albert, J., Schipper, P., Gulnik, S., Boucher, C.A., 1999. Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. AIDS 13, 2349–2359.

Patick, A.K., Duran, M., Cao, Y., Shugarts, D., Keller, M.R., Mazabel, E., Knowles, M., Chapman, S., Kuritzkes, D.R., Markowitz, M., 1998. Genotypic and phenotypic characterization of human immunodeficiency virus type 1 variants isolated from patients treated with the protease inhibitor nelfinavir. Antimicrob. Agents Chemother. 42, 2637–2644.

Patick, A.K., Mo, H., Markowitz, M., Appelt, K., Wu, B., Musick, L., Kalish, V., Kaldor, S., Reich, S., Ho, D., Webber, S., 1996. Antiviral and resistance studies of AG1343, an orally bioavailable inhibitor of human immunodeficiency virus protease. Antimicrob. Agents Chemother. 40, 292–297.

Prabu-Jeyabalan, M., Nalivaika, E., Schiffer, C.A., 2000. How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease. J. Mol. Biol. 301, 1207–1220.

Resch, W., Swanstrom, R., 2000. Protease inhibitor resistance in HIV-1 infection. Rev. Med. Microbiol. 11, 211–221.

Resch, W., Ziermann, R., Parkin, N., Gamarnik, A., Swanstrom, R., 2002. Nelfinavir-resistant, amprenavir-hypersusceptible strains of human immunodeficiency virus type 1 carrying an N88S mutation in protease have reduced infectivity, reduced replication capacity, and reduced fitness and process the Gag polyprotein precursor aberrantly. J. Virol. 76, 8659–8666.

Rose, R.E., Gong, Y.F., Greytok, J.A., Bechtold, C.M., Terry, B.J., Robinson, B.S., Alam, M., Colonno, R.J., Lin, P.F., 1996. Human immunodeficiency virus type 1 viral background plays a major role in development of resistance to protease inhibitors. Proc. Natl. Acad. Sci. USA 93, 1648–1653.

Schock, H.B., Garsky, V.M., Kuo, L.C., 1996. Mutational anatomy of an HIV-1 protease variant conferring cross-resistance to protease inhibitors in clinical trials. Compensatory modulations of binding and activity. J. Biol. Chem. 271, 31957–31963.

Shafer, R.W., Hsu, P., Patick, A.K., Craig, C., Brendel, V., 1999. Identification of biased amino acid substitution patterns in human immunodeficiency virus type 1 isolates from patients treated with protease inhibitors. J. Virol. 73, 6197–6202.

Shafer, R.W., Kantor, R., Gonzales, M.J., 2000. The genetic basis of HIV-1 resistance to reverse transcriptase and protease inhibitors. AIDS Rev. 2, 211–228.

Smidt, M.L., Potts, K.E., Tucker, S.P., Blystone, L., Stiebel Jr., T.R., Stallings, W.C., McDonald, J.J., Pillay, D., Richman, D.D., Bryant, M.L., 1997. A mutation in human immunodeficiency virus type 1 protease at position 88, located outside the active site, confers resistance to the hydroxyethylurea inhibitor SC-55389A. Antimicrob. Agents Chemother. 41, 515–522.

Theil, H., 1972. Statistical Decomposition Analysis. North-Holland Publishing Co., Amsterdam, pp. 115–120.

Vaillancourt, M., Irlbeck, D., Smith, T., Coombs, R.W., Swanstrom, R., 1999. The HIV type 1 protease inhibitor saquinavir can select for multiple mutations that confer increasing resistance. AIDS Res. Hum. Retroviruses 15, 355–363.

Vondrasek, J., Wlodawer, A., 1996. New database. Science 272, 337–338.

Watkins, T., Resch, W., Irlbeck, D., Swanstrom, R., 2003. Selection of high-level resistance to human immunodeficiency virus type 1 protease inhibitors. Antimicrob. Agents Chemother. 47, 759–769.

Wegner, S.A., Brodine, S.K., Mascola, J.R., Tasker, S.A., Shaffer, R.A., Starkey, M.J., Barile, A., Martin, G.J., Aronson, N., Emmons, W.W., Stephen, K., Bloor, S., Vingerhoets, J., Hertogs, K., Larder, B., 2000. Prevalence of genotypic and phenotypic resistance to anti-retroviral drugs in a cohort of therapy-naive HIV-1 infected US military personnel. AIDS 14, 1009–1015.

Winters, M.A., Schapiro, J.M., Lawrence, J., Merigan, T.C., 1998. Human immunodeficiency virus type 1 protease genotypes and in vitro protease inhibitor susceptibilities of isolates from individuals who were switched to other protease inhibitors after long-term saquinavir treatment. J. Virol. 72, 5303–5306.

Wlodawer, A., Vondrasek, J., 1998. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. Annu. Rev. Biophys. Biomol. Struct. 27, 249–284.

Wu, T.D., Schiffer, C.A., Gonzales, M.J., Taylor, J., Kantor, R., Chou, S., Israelski, D., Zolopa, A.R., Fessel, W.J., Shafer, R.W., 2003. Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. J. Virol. 77, 4836–4847.

Yahi, N., Tamalet, C., Tourres, C., Tivoli, N., Ariasi, F., Volot, F., Gastaut, J.A., Gallais, H., Moreau, J., Fantini, J., 1999. Mutation patterns of the reverse transcriptase and protease genes in human immunodeficiency virus type 1-infected patients undergoing combination therapy: survey of 787 sequences. J. Clin. Microbiol. 37, 4099–4106.

Zennou, V., Mammano, F., Paulous, S., Mathez, D., Clavel, F., 1998. Loss of viral fitness associated with multiple Gag and Gag-Pol processing defects in human immunodeficiency virus type 1 variants selected for resistance to protease inhibitors in vivo. J. Virol. 72, 3300–3306.

Ziermann, R., Limoli, K., Das, K., Arnold, E., Petropoulos, C.J., Parkin, N.T., 2000. A mutation in human immunodeficiency virus type 1 protease, N88S, that causes in vitro hypersensitivity to amprenavir. J. Virol. 74, 4414–4419.