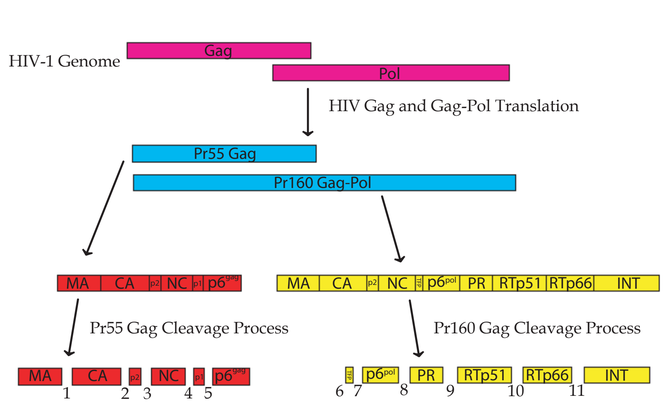
Covariation analysis of the amino acid sequence of HIV-1 subtype B protease and its Gag-Pol cleavage sites

Júlia Kornai[1], István Bartha[1], Rita Hírmondó[1], Jochen Bodem[2] and Viktor Müller[1]

[1]Institute of Biology, Eötvös Loránd University, Budapest, Hungary

[2]Institute of Virology and Immunobiology, University of Würzburg, Germany

The cleavage of the Gag-Pol polyprotein by the viral protease (PR) is essential for the infectivity of HIV virions. Protease inhibitor (PI) therapy can give rise to resistance mutations in the protease, which is often associated with a decreased activity of the enzyme. Impaired function can be partially restored by compensatory mutations in the cleavage sites (CS), probably by providing a better substrate for the mutated proteases. We performed a statistical analysis on publicly available HIV sequences to detect associations between specific protease and cleavage site mutations, which might identify variations at cleavage site as potential compensatory mutations.



SEQUENCES

HIV-1 subtype B nucleotide sequences containing the protease region were downloaded from the Los Alamos HIV Sequence Database (www.hiv.lanl.gov). Alignment was generated using MUSCLE (www.drive5.com/muscle) and then refined by manual inspection. Our final alignment contained 30305 sequences and spanned the entire Gag-Pol region. Translation of nucleotide sequences and further analyses were performed using PERL programs and PAUP* (http://paup.csit.fsu.edu).

COVARIATION ANALYSIS:

We used chi-square tests of independence to detect associations between the occurrence of mutations at each individual position in the cleavage sites and in the protease. The strength and direction of the association is characterized by the phi-correlation coefficient; a positive coefficient indicates that mutations at the two positions occur together preferentially. The basic scheme of the method is shown below.

	Mutation at the cleavage site position	***Consensus AA at the cleavage site position***
Mutation at the PR position	***a:*** number of sequences containing mutation at both positions	***b:*** number of sequences containing mutation at the PR position and consensus AA at the cleavage site position
Consensus AA at the PR position	***c:*** number of sequences containing consensus AA at the PR position and mutation at the cleavage site position	***d:*** number of sequences containing consensus AA at both positions

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$\chi^2 = 0$ - total independence

$\chi^2 \geq 3{,}84$ has the p - value $p \leq 0{,}05$.

$$\varphi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

$-1 \leq \varphi \leq 1$

$\varphi < 0$ - negative correlation

$\varphi = 0$ - no correlation

$\varphi > 0$ - positive correlation

$0{,}1 \leq |\varphi| < 0{,}3$ - weak correlation

$0{,}3 \leq |\varphi| < 0{,}5$ - moderate correlation

$0{,}5 \leq |\varphi|$ - strong correlation

IDENTIFICATION OF PI RESISTANT SEQUENCES:

To pinpoint correlated mutations that may be induced by drug treatment, we repeated the analysis on the subset of sequences that contained PI resistant protease. We classified a protease as PI resistant, if it contained mutations at two or more of the following resistance-associated positions: 23, 24, 30, 32, 33, 46, 47, 48, 50, 53, 54, 73, 76, 82, 84, 88, 90.

Table #1: Polymorphisms in the cleavage sites (parentheses contain the frequency of non-consensus amino acids and the most prevalent substitutions).
Note: we used an extended definition of a cleavage site (+/- 10 AA positions) to allow the identification of positions that are important in PR processing but fall outside the classical definition (+/- 5 positions).

Cleavage site	P10	P9	P8	P7	P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	P5'	P6'	P7'	P8'	P9'	P10'
MA/CA	G (6.48%, E)	N (16.70%, K,H,S)	S (30.35%, N)	S (20.49%, N,G)	Q (14.43%, K,P)	V (4.45%, A)	S	Q (5.12%, H)	N	Y (2.96%, F)	P	I	V	Q	N	L (24.61%, M,I)	Q	G	Q	M
CA/p2	G	G	P	G (35.94%, S)	H	K	A	R	V (21.58%, I)	L	A	E	A	M	S	Q (2.60%)	V (23.46%, M,A,I)	T (3.39%)	N (22.19%, Q,S,G)	S (28.48%, Q,A,P,T)
p2/NC	S	Q (2.60%)	V (23.46%, M,A,I)	T (3.39%)	N (22.19%, Q,S,G)	S (28.48%, Q,A,P,T)	A (29.38%, V,P,N,T)	T (38.97%, S,N,A)	I (19.19%, V)	M (2.54%, L)	M (3.14%, I)	Q	R (29.10%, K)	G (3.72%, S)	N (2.10%)	F (6.86%, Y)	R (15.35%, K)	N (14.01%, S,G)	Q (2.51%, P)	R (6.93%, K)
NC/TFP	M	K	D (5.80%, E)	C	T (13.67%, A,N,I,S)	E (2.25%)	R	Q	A	N	F	L	R	E	N (49.22%, D)	L	A	F (2.56%)	P (17.50%, Q,L)	Q
TFP/p6 pol	A	N	F	L	R	E	N (49.22%, D)	L	A	F (2.56%)	P (17.50%, Q,L)	Q	G (4.56%, R)	K (31.29%, E)	A	R (3.74%, G)	E (7.64%, K)	F (10.89%, L)	S (22.56%, P)	S (12.24%, T,P)
p6 pol/PR	D (28.17%, E,N,G)	R (7.43%, G)	Q (5.15%, P)	G	T (39.36%, I,P,A,S,D,N)	V (13.53%, I)	S	F (20.45%, L)	S (29.74%, D,N,G)	F (19.82%, L)	P	Q	I	T	L	W	Q	R	P	L (25.19%, V,F,I)
PR/p51	L (14.61%, M)	T	Q (2.29%)	I (33.93%, L)	G	C	T	L	N	F	P	I	S	P	I	E (4.19%, D)	T	V	P	V
p51/p66	K (3.99%)	E (2.38%)	P	I	V (37.46%, T,L,A,E,I)	G	A (2.27%, V)	E	T	F	Y	V	D	G	A	A (3.46%, S)	N (4.10%, S)	R	E (5.52%, D)	T
p66/INT	L	V	S	A (44.84%, N,T,S)	G	I (4.55%, V)	R	K (9.71%, R)	V (8.99%, I)	L	F	L	D	G	I	D (3.53%)	K (3.54%)	A	Q	E (10.44%, D)
NC/p1	M	K	D (5.80%, E)	C	T (13.67%, A,N,I,S)	E (2.25%)	R	Q	A	N	F	L	G	K (3.72%, R)	I (5.97%, L)	W	P	S	H (22.36%, L,S,N,Y)	K (4.20%, R)
p1/p6gag	P	S	H (22.36%, L,S,N,Y)	K (4.20%, R)	G (2.69%, E)	R	P	G	N	F	L (8.59%, P)	Q	S (19.48%, N)	R (3.17%)	P (12.69%, T,L)	E (3.66%, A)	P	T (21.85%, P,S)	A	P

Results #1: associations between PR and CS mutations (all sequences)

Cleavage site		PR position	All sequences			PI-resistant sequences		
name	position		phi	p-value	n	phi	p-value	n
MA/CA	**P6**	**10**	0.2668	<0.0001	492	NA	NA	<100
CA/p2	**P7**	**77**	0.3243	<0.0001	648	NA	NA	<100
		93	0.3097	<0.0001	646	NA	NA	<100
	P2	**72**	0.4506	<0.0001	648	NA	NA	<100
		93	0.3156	<0.0001	653	NA	NA	<100
p2/NC	**P3**	**72**	0.4087	<0.0001	652	NA	NA	<100
RTp51/RTp66	**P9`**	**35**	0.2621	<0.0001	630	NA	NA	<100
		37	0.2225	<0.0001	632	NA	NA	<100

Results #2: associations between PR and CS mutations (PI resistant sequences)

Cleavage site		PR position	All sequences			PI-resistant sequences		
name	position		phi	p-value	n	phi	p-value	n
NC/TFP (or TFP/p6pol)	**P5' (P4)**	**19**	0.0964	0.0004	1326	0.5983	<0.0001	120
		23	0.1708	<0.0001	1340	0.5839	<0.0001	120
		32	0.1605	<0.0001	1344	0.5709	<0.0001	120
		47	0.1686	<0.0001	1344	0.5725	<0.0001	120
		72	0.0796	0.0038	1320	0.5308	<0.0001	119
TFP/p6pol	**P9'**	**23**	0.2734	<0.0001	1666	0.5879	<0.0001	122
		32	0.2526	<0.0001	1670	0.5047	<0.0001	122
		47	0.2638	<0.0001	1670	0.5404	<0.0001	122
		93	0.0452	0.0653	1662	0.5158	<0.0001	122
	P10'	**12**	0.1212	<0.0001	1769	0.6704	<0.0001	123
		19	0.1430	<0.0001	1768	0.5659	<0.0001	123
		20	0.2781	<0.0001	1771	0.6473	<0.0001	122
		23	0.3530	<0.0001*	1784	0.6822	<0.0001	123
		32	0.3415	<0.0001*	1788	0.6419	<0.0001	123
		36	0.2182	<0.0001	1771	0.5137	<0.0001	121
		46	0.3302	<0.0001	1786	0.5251	<0.0001	123
		47	0.3530	<0.0001*	1788	0.7032	<0.0001	123
		53	0.3645	<0.0001*	1786	0.6927	<0.0001	123
		72	0.1741	<0.0001	1763	0.5624	<0.0001	122
		93	0.0474	0.0455	1779	0.5029	<0.0001	123

Results #3: negative covariation between PR and CS mutations (PI resistant sequences)

Cleavage site		PR position	All sequences			PI-resistant sequences		
name	position		phi	p-value	n	phi	p-value	n
NC/TFP (or TFP/p6pol)	**P5' (P4)**	**41**	-0.0195	0.4778	1325	-0.4099	<0.0001	120
		82	-0.1699	<0.0001	1337	-0.4633	<0.0001	119
TFP/p6pol	**P1'**	**10**	-0.0831	0.0022	1357	-0.4082	<0.0001	120
	P4'	**12**	0.0175	0.4773	1648	-0.4116	<0.0001	121
		47	-0.1030	<0.0001	1667	-0.4573	<0.0001	121
	P9'	**82**	-0.0554	0.0240	1663	-0.4443	<0.0001	121
	P10'	**37**	-0.0159	0.5058	1762	-0.4123	<0.0001	122
		82	-0.0382	0.1072	1781	-0.4098	<0.0001	122

Shading indicates resistance-associated PR positions. Asterisks(*) indicate p-values calculated with Fisher's exact test.

Contact address: viktor.mueller@env.ethz.ch

PHYLOGENETIC JACKKNIFING

We implemented a phylogenetic test to exclude non-functional covariations that arose by "common descent" only. Such apparent covariations may arise if a pair of mutations (without functional association) is present in an intensively sampled patient or patient group. For each position pair that had $|\varphi| \geq 0.2$ and $p \leq 0.05$ in the main analysis, we selected the sequences that contained the two positions of the pair. We removed the two positions of the correlating pair and all protease positions that are strongly associated with resistance, and calculated the distance matrix for the sequences from the remaining alignment (the best-fit substitution model was determined with the MODELTEST block of PAUP). Our "phylogenetic jackknifing" test looped over all sequences and selected the 10% of sequences that were closest to the center of selection in terms of the estimated evolutionary distance. We removed this 10% from the alignment, and repeated the chi-square test for the given mutation pair on the remaining 90% of sequences. Further selection of significant results was based on the smallest phi value in any of the jackknived samples.This way, we excluded covariations that were generated by any closely related 10% of the available sequences. This test excluded the majority of the covariations obtained in the main analyses, which indicates that sampling is indeed biased and the phylogenetic jackknifing test was needed and justified.

Of 59 pairs that had $|\varphi| \geq 0.2$ in the complete set, only 8 passed the jackknifing test. Of 140 pairs that had $|\varphi| \geq 0.2$ in the PI resistant set, 75 passed the jackknifing test. This indicates that most associations are indeed generated by drug pressure. Because associations were generally stronger in the PI resistant set, we used the more stringent thresholds of $|\varphi| \geq 0.5$ or $|\varphi| \leq -0.4$ for listing associations in the PI resistant set.

DISCUSSION

We found significant association between several pairs of PR and cleavage site mutations. Most associations were positive, which indicates that the CS mutations may compensate the effect of the PR mutations. Most such mutation pairs involved a known drug resistance position in the PR, and associations in the PI resistant set were stronger, which suggests that most associations were indeed generated by drug pressure. Interestingly, some mutation pairs demonstrated a negative correlation, indicating that changes in these cleavage sites are less tolerated by the mutant than by the wild-type protease. Remarkably, a large fraction of the correlated pairs involved a CS position that falls outside the classical definition of a cleavage site, which suggests that positions that lie farther from the cleaved bond may also influence the efficiency of cleavage.

This work has been made possible through the support of the European Commission grants HIV Protease (QLK2-CT-2001-02360) and Virolab (027446): www.virolab.org.