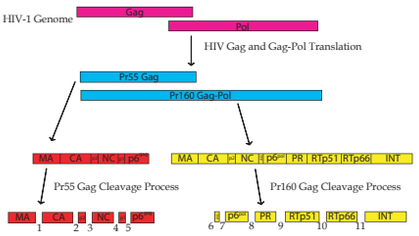
Covariation analysis of the amino acid sequence of HIV-1 subtype B protease and its Gag-Pol cleavage sites

Júlia Kornai[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). Accurate alignment of sequences was achieved by manual refinement of automatic alignment generated by ClustalW. Translation of nucleotide sequences and further analysis were performed using PERL programs. 14172 sequences remained after the discarding of corrupted or ambiguous sequences. Analyses were restricted to positions covered by at least 100 sequences, which yielded Gag-pol positions 407-1434, including 8 cleavage sites: NC/p1, p1/p6gag, NC/TFP, TFP/p6pol, p6pol/PR, PR/RTp51, RTp51/RTp66 and RTp66/INT. We defined 'wild type' as the consensus sequence in our dataset.

Polymorphisms in the cleavage sites: parentheses contain the frequency of non-consensus amino acids and the most prevalent substitutions.

Cleavage site	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	P5'
NC/TFP	E	R	Q	A (4.52%; V)	N	F	L	R	E (2.67%; K)	N/D
TFP/p6pol	E (2.67%; K)	N/D	L (5.63%; V)	A	F	P (20.09%; L)	Q	G (6.41%; R)	K (20.09%; E)	A
p6pol/PR	V (20.02%; LS,L)	S	F (17.6%; L)	S (23.78%; N,G)	F (17.47%; L)	P	Q	I	T	L
PR/p51	C	T	L	N	F	P	I	S	P	I
p51/p66	G	A	E	T	F	Y	V	D	G	A
p66/INT	I (4.74%; V)	R	K (8.49%; R)	V (3.77%; I)	L	F	L	D	G	I
NC/p1	E	R	Q	A (4.52%; V)	N	F	L	G	K (2.67%; R)	I (6.14%; M)
p1/p6gag	R	P	G	N	F	L (14.46%; F,P)	Q	S (10.66%; N)	R (2.03%)	P(8.40%; L,T)

COVARIATION ANALYSIS #1

We used chi-square tests of independence to detect associations between PI treatment and cleavage site mutations. The strength and direction of the association is characterized by the phi-correlation coefficient; a positive coefficient indicates an association between PI treatment and mutations at the given CS. The basic scheme of the method is shown below.

	PI resistant PR	***Wild type PR***
Mutation in the CS	***a***: number of sequences containing mutation in the CS and at least two resistance mutations in the PR	***b***: number of sequences containing mutation in the CS, but no resistance mutation in the PR
No mutation in the CS	***c***: number of sequences containing no mutation in the CS, but at least two resistance mutations in the PR	***d***: number of sequences containing no mutation in the CS and no resistance mutation in the PR

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

$\chi^2 = 0$ - 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

Results #1: associations between PI resistance and CS mutations

	Full data set		PI-resistant seqs[a]		Wild type PR[b]		phi	p-value
	# of wt sequences	# of mutated sesuqences	# of wt sequences	# of mutated sesuqences	# of wt sequences	# of mutated sesuqences		
NC/TFP	98	113	9	9	85	98	-0.02	0.7773
TFP/p6pol	76	143[c]	14	5	59	131	-0.26	0.0002
p6pol/PR	406	476[c]	63	71	316	375	-0.01	0.7913
PR/p51	9788	305	1773	103	7136	170	0.08	<0.0001
p51/p66	209	7	0	0	206	7	NA	NA
p66/INT	49	3	0	0	48	3	NA	NA
NC/p1	180	31	5	13	168	15	0.53	<0.0001[d]
p1/p6gag	173	54	18	1	147	51	-0.14	0.0496[d]

[a]Sequences containing at least 2 resistance-associated mutations in PR.
[b]Sequences containing no resistance-associated mutations in PR.
[c]These two cleavage sites contained 2 or 3 mutations in a considerable number of sequences.
[d]p-values were calculated by Fisher's exact test due to the small number of sequences.

COVARIATION ANALYSIS #2

Chi-square tests of independence were employed for all PR-CS position pairs to test for associations between the occurrence of mutations at the two positions. A positive coefficient indicates that mutations at the two positions occur together preferentially.

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

Results #2: associations between PR and CS mutations

Cleavage site		PR position	All sequences			PI-resistant sequences			Cleavage site associated with treatment (Data in the literature)
name	position		phi	p-value	n	phi	p-value	n	
NC/TFP (or TFP/p6pol)	P5' (P4)	13	0.2845	<0.0001	224	NA	NA	<100	+
		16	-0.2798	<0,0001	224	NA	NA	<100	
		37	-0.235	0.0004	222	NA	NA	<100	
		41	-0.2813	<0.0001	224	NA	NA	<100	
		54	-0.2445	0.0002	223	NA	NA	<100	
		62	0.25	0.0002	223	NA	NA	<100	
		64	0.355	<0.0001	225	NA	NA	<100	
		77	-0.2286	0.0006	223	NA	NA	<100	
		82	-0.2567	0.0001	223	NA	NA	<100	
TFP/p6pol	P1'	12	0.2763	<0.0001	229	NA	NA	<100	
		13	0.422	<0.0001	231	NA	NA	<100	
		16	-0.2313	0.0004	231	NA	NA	<100	
		41	-0.2868	<0.0001	231	NA	NA	<100	
		62	0.3073	<0.0001	229	NA	NA	<100	
		63	0.2934	<0.0001	230	NA	NA	<100	
		64	0.4797	<0.0001	232	NA	NA	<100	
		93	-0.2141	0.0011	231	NA	NA	<100	
TFP/p6pol	P4'	16	-0.2313	0.0004	231	NA	NA	<100	+
		41	-0.2012	0.0021	232	NA	NA	<100	
		63	-0.329	<0.0001	230	NA	NA	<100	
		93	0.2989	<0.0001	231	NA	NA	<100	
p1/p6gag	P1'	12	0.2877	<0.0001	236	NA	NA	<100	++
		13	0.251	<0.0001	239	NA	NA	<100	
		41	-0.2119	0.001	239	NA	NA	<100	
		62	0.3522	<0.0001	237	NA	NA	<100	
		64	0.2447	0.0001	240	NA	NA	<100	
p6pol/PR	P5	10	-0.1363	<0.0001	909	-0.239	0.0047	138	
		16	0.3182	<0.0001	921	0.628	<0.0001*	139	
		37	0.0984	0.0031	904	0.2814	0.001	134	
		41	0.1718	<0.0001	912	0.3784	<0.0001	139	
		54	0.0586	0.0767	914	0.282	0.0009	136	
		77	0.0286	0.395	886	0.4442	<0.0001	134	
		82	0.0249	0.4521	916	0.2071	0.0152	137	
		90	-0.0803	0.0152	914	-0.276	0.0011	137	
		93	-0.0583	0.0776	917	-0.2057	0.0155	138	
	P3	39	0.2911	<0.0001	1787	0.239	0.0067*	224	-
		46	0.0556	0.0192	1777	0.2029	0.0027	217	
		57	0.2979	<0.0001	1783	0.1061	0.123*	224	
		64	-0.0568	0.0168	1768	0.2196	0.001	220	
		71	0.171	<0.0001	1774	-0.2214	0.0011	213	
	P2	57	-0.2012	<0.0001	1792	-0.0329	0.6251	223	

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

Associations with $|\varphi| \geq 0.2$ were included in the table.

DISCUSSION

We have found significant association between resistance to PIs and the presence of mutations in four cleavage sites. The positive association of NC/p1 mutations with treatment is well-documented in the literature; the weak association demonstrated for the PR/p51 CS remains to be validated experimentally. Interestingly, two cleavage sites (TFP/p6pol and p1/p6gag) demonstrated a negative correlation with PI resistance, indicating that changes in these CSs are less tolerated by the mutant PR enzymes. Both positive and negative correlations were also detected in the analysis of individual AA-positions. Positive covariation might indicate compensatory mutations, while negative covariation indicates that the given mutations in PR actually narrow the specificity of the enzyme for the wt amino acids in the CS positions.

We note that the dataset used for the analysis might not constitute a balanced representation of treated and untreated Gag-Pol sequences, and we had no direct information on the treatment history of individual samples.

Acknowledgement

This work was initiated within the frame of the EU FP5 grant QLK2-CT-2001-02360, and is now being developed further under the FP6 grant IST-027446. VM was also supported by the Hungarian National Science Fund (grant No. D45948).

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