

Identification of Correlating Conserved Regions on HIV Proteins by Mutation Analysis

Yaoyu E. Wang¹ Charles DeLisi^{1,2}
yew@bu.edu delisi@bu.edu

Departments of Bioinformatics¹ and Biomedical Engineering²,
Boston University, Boston Massachusetts, 02215

Keywords: Correlated Mutation, Protein Conservation, HIV

1. Introduction

The RNA genomes of retroviruses such as Human Immunodeficiency Virus undergo rapid evolutionary change upon human infection. The enormous diversity and evolutionary progression of these viruses make the development of reliable detection tests, effective vaccines, and pharmaceutical agents difficult. Understanding the interacting relationship between viral proteins becomes important to identify protein regions that may be susceptible to drug intervention. The conserved regions of viral proteins are particularly of interest because of their roles in producing viable virus. We are interested to examine the interacting relationship of highly conserved region pairs between proteins, and attempt to identify inter-protein regions with high level of mutation correlation.

2. Method and Results

The distribution of interest is the probability of observing mutation at two conserved regions from different proteins by random chance with no selective pressure. Consider the situation of proteins X and Y from the same virus has been sequenced in N_X and N_Y strains respectively. Supposed a conserved region of X is mutated at x number of strains in N_X ; similar notations are used for Y . If there are z pairs of conserved regions from each of X and Y mutated simultaneously on the same strain, the probability of observing z , assuming no mutation correlation, can be expressed as $P(z | x, y, N_X, N_Y)$, where z is constrained by

$$\eta = N_X \cap N_Y \quad (1)$$

Then, the number of possible distribution of z *indistinguishable* coincidences over η is:

$$w_z = \binom{\eta}{z} \quad (2)$$

Supposed x_1 and y_1 of mutations occur within η , where $x_1 \subseteq x$ and $y_1 \subseteq y$, the number of possible distribution of remaining mutation within η but not in z , denoted as $\overline{w_z}$, is

$$\overline{w_z} = \frac{(\eta - z)!}{(\eta + z - x_1 - y_1)!(x_1 - z)!(y_1 - z)!} \quad (3)$$

The number of ways $\overline{w_z}$ can be distributed in N_X and N_Y is dictated by number of mutation outside of η . Define w_x to be the number of ways $x - x_1$ can occur in $N_X - \eta$ strains and similarly for y :

$$w_x = \frac{(N_X - \eta)!}{(x - x_1)![N_X - \eta - (x - x_1)]!} \quad (4)$$

$$w_y = \frac{(N_Y - \eta)!}{(y - y_1)![N_Y - \eta - (y - y_1)]!} \quad (5)$$

Then the total number of ways of distributing a set of x, x_1, y, y_1 given z number of strains overlapping is $w_z w_x w_y$. This must be summed over all possible values of x_1 and y_1 . Thus, the total number of ways in which z overlaps can occur given N_X, N_Y, x, y is:

$$W_z = w_z \sum_{x_1=z}^i \sum_{y_1=z}^j \overline{w_z} w_x w_y \quad i = \begin{cases} x & \eta \geq x \\ \eta & \eta \leq x \end{cases}, \quad j = \begin{cases} y & \eta \geq y \\ \eta & \eta \leq y \end{cases} \quad (6)$$

Finally to obtain $P(z | x, y, N_X, N_Y)$, we must divide equation (6) by the number of ways of distributing x mutations over N_X and y mutations over N_Y without restriction:

$$W = \binom{N_X}{x} \binom{N_Y}{y} \quad (9)$$

We therefore obtain:
$$P(z | x, y, N_1, N_2) = \frac{W_z}{W} = \frac{w_z \sum_{x_1=z}^i \sum_{y_1=z}^j \overline{w_x w_y}}{\binom{N_x}{x} \binom{N_y}{y}} \quad (10)$$

2.2 Tables

Table 1. Conserved decemer is defined to be peptide with over 90% conservation across all HIV strains. Conserved Segment is the aggregate of overlapping conserved decemers.

	Capsid	Integrase	RT	RNase H	gp120
Conserved Decemer	11	31	19	15	2
Conserved Segment	4	5	7	2	1

Table 2. Correlating Conserved HIV Protein regions. The table represents the number of conserved decemer pairs, where $P(z|x,y,N_1,N_2) \leq 0.01$, observed within pairs of conserved segments. Only inter-protein pairs are considered.

	EVNIVTDSQYALGHIIQAQPD (52-71)	WVOAHKGGIGNEQ (95-108)	SPRTLNAWVKV (16-27)	VGGHQAAMQMLK (59-71)	PRGSDIAGTT (99-109)	PRGSDIAGTTVDRF(155-169)	WVTVYYGVPPVW (35-46)	PGIWQLDCTHLEGG (58-72)	EAEVIPAETGQ (85-96)	GIPYNPQSQGV (140-151)	QAEHLKTAVQMAVFIHNFKRKGGIG (168-192)	WKGPAKLLWKGEGAVVIQD (235-250)	KQWPLTEEKIKAL (22-35)	IGPENPYNTP (50-60)	RKLVDFRELNK (72-82)	SVTVLDVGDGA (105-115)	FRKYTAFTIPS (124-135)	EPPFLWMGYEL (224-235)	QKLVGKLNWASQIY (258-272)
RNase	EVNIVTDSQYALGHIIQAQPD (52-71)		20	33	11	36	22	55	0	17	75	66	44	3	0	0	0	0	40
	WVOAHKGGIGNEQ (95-108)		8	12	1	11	8	8	0	8	17	6	4	1	3	2	8	6	20
Capsid	SPRTLNAWVKV (16-27)						4	10	0	4	12	17	8	1	2	2	4	2	9
	VGGHQAAMQMLK (59-71)						6	15	0	6	33	33	12	1	11	3	1	6	15
	PRGSDIAGTT (99-109)						2	5	0	2	6	6	4	0	4	1	0	2	5
	PRGSDIAGTTVDRF(155-169)						10	25	10	10	0	25	0	8	5	5	10	10	25
gp120	WVTVYYGVPPVW (35-46)							10	0	0	14	22	8	2	8	2	4	4	10
INT	PGIWQLDCTHLEGG (58-72)												25	1	4	1	0	8	25
	EAEVIPAETGQ (85-96)												0	0	0	0	0	0	2
	GIPYNPQSQGV (140-151)												4	2	3	0	0	4	10
	QAEHLKTAVQMAVFIHNFKRKGGIG (168-192)												28	0	6	2	0	20	4
	WKGPAKLLWKGEGAVVIQD (235-250)												33	4	28	8	16	8	47
RT	KQWPLTEEKIKAL (22-35)																		
	IGPENPYNTP (50-60)																		
	RKLVDFRELNK (72-82)																		
	SVTVLDVGDGA (105-115)																		
	FRKYTAFTIPS (124-135)																		
	EPPFLWMGYEL (224-235)																		
	QKLVGKLNWASQIY (258-272)																		

Discussion

All five proteins show some number of correlating conserved decemers. Furthermore, all of the three catalytic proteins contain conserved segment that display high number of correlating decemer pairs with other proteins. For example, QKLVGKLNWASQIY of RT has more than 10 correlating decemer pairs with 8 out of 12 conserved segments on other proteins. By analyzing the correlating mutation between inter-protein regions, we identify correlating conserved segments that their interaction can be further studied.

References:

1. Kelleher AD, Long C, Holmes EC, Allen RL, Wilson J, Conlon C, Workman C, Shaunak S, Olson K, Goulder P, Brander C, Ogg G, Sullivan JS, Dyer W, Jones I, McMichael AJ, Rowland-Jones S, and Phillips RE. Clustered Mutations in HIV-1 gag Are Consistently Required for Escape from HLA-B27-restricted Cytotoxic T Lymphocyte Responses. *J. Exp. Med.*, **193**, 375-385 (2001)
2. Human Retroviruses and AIDS 2000: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. Kuiken CL, Foley B, Hahn B, Korber B, McCutchan F, Marx PA, Mellors JW, Mullins JI, Sodroski J, and Wolinsky S, Eds. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.