Identification of Correlating Conserved Regions on HIV Proteins by Mutation Analysis

Yaoyu E. Wang1 Charles DeLisi1,2
yw@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

\[
P(X = x, Y = y | N_X, N_Y) = \frac{w_x w_y}{w_z},
\]

Then, the number of possible distribution of \( z \) indistinguishable coincidences over \( \eta \) is:

\[
w_z = \binom{\eta}{z}
\]

Supposed \( x_i \) and \( y_j \) of mutations occur within \( \eta \), where \( x_i \subseteq x \) and \( y_j \subseteq y \), the number of possible distribution of remaining mutation within \( \eta \) but not in \( z \) denoted as \( w_z \), is

\[
w_z = \frac{(\eta - z)!}{(\eta + z - x_{i} - y_{j})(x_{i} - z)!(y_{j} - z)!}
\]

The number of ways \( w_x \) can be distributed in \( N_X \) and \( N_Y \) is dictated by number of mutation outside of \( \eta \). Define \( w_x \) to be the number of ways \( x-x_i \) can occur in \( N_X-\eta \) strains and similarly for y:

\[
w_x = \frac{(N_X - \eta)!}{(x - x_i)!(N_X - \eta - (x - x_i))!}
\]

\[
w_y = \frac{(N_Y - \eta)!}{(y - y_j)!(N_Y - \eta - (y - y_j))!}
\]

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

\[
W_z = \sum_{x = \eta}^{\eta} \sum_{y = \eta}^{\eta} w_x w_y w_z
\]

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:

\[
W = \binom{N_X}{x} \binom{N_Y}{y}
\]
We therefore obtain:

\[ P(z | x, y, N_1, N_2) = \frac{W_z}{W} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} w_x w_y \sum_{\zeta=x}^{N_x} \sum_{\eta=y}^{N_y} \mu_{x,\eta} \]

(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.

<table>
<thead>
<tr>
<th>Capsid</th>
<th>Integrase</th>
<th>RT</th>
<th>RNase H</th>
<th>gp120</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conserved Decemer</td>
<td>11</td>
<td>31</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>Conserved Segment</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>RNase</th>
<th>FYNTFESQALGVRQXQY (52-71)</th>
<th>VVFQAMYKLGQ (54-95)</th>
<th>PGRSH (99-109)</th>
<th>PGRSH (115-169)</th>
<th>WVTVVYQYEV (135-203)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capsid</td>
<td>Sprilnawvkv (16-27)</td>
<td>VQGHRQAMXQMLK</td>
<td>PGRSH (99-109)</td>
<td>PGRSH (115-169)</td>
<td>WVTVVYQYEV (135-203)</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>11</td>
<td>30</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>gp120</td>
<td>WVTVVYQYEV (25-40)</td>
<td>WVTVVYQYEV (25-40)</td>
<td>WVTVVYQYEV (25-40)</td>
<td>WVTVVYQYEV (25-40)</td>
<td>WVTVVYQYEV (25-40)</td>
</tr>
<tr>
<td>INT</td>
<td>PGWQLCTHLE (58-72)</td>
<td>EAFVQAETGQ (85-96)</td>
<td>PGRSH (99-109)</td>
<td>PGRSH (115-169)</td>
<td>WVTVVYQYEV (25-40)</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>11</td>
<td>30</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>RT</td>
<td>KQWLITEEK (22-35)</td>
<td>GIPYQFST (50-60)</td>
<td>RKLVT (72-82)</td>
<td>SYTVL (95-115)</td>
<td>FKYAT (143-155)</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>11</td>
<td>30</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>4</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

### 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, QKLVGKLNWSA/QY 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:
