Mutation Analysis of HIV-1 Primary Protein Sequences

R.S. Sergeev* [1], A.V. Tuzikov [2], V.F. Eremin [3]
1) Faculty of Applied Mathematics and Computer Science, Belarus State University, Minsk, Belarus, e-mail: roma.sergeev@gmail.com
2) United Institute of Informatics Problems, Minsk, Belarus

3) Research Institute for Epidemiology and Microbiology, Minsk, Belarus

Abstract: *Many human immunodeficiency virus type-1 (HIV-1) infected persons are treated with multiple protease and reverse transcriptase (RT) inhibitors. Due to high variability of the virus strains these drug classes often lead to the development of drug resistance. This process is accompanied by intensive mutations in the (HIV pol gene) virus protease and reverse transcriptase. But only a subset of treatment-associated mutations is responsible for establishing drug-resistance. The proposed method is designed to detect protease and RT mutations responsible for drug-resistance surveillance and support decisions on further treatment regimen.*

***Keywords*:** Phylogeny, HIV-1, antiretroviral therapy, contingency tables, drug resistance.

1. INTRODUCTION

Drug resistance is a major obstacle to the effective treatment of human immunodeficiency virus type 1 (HIV-1) infection. Although some of the antiretroviral drugs have been approved for the treatment of HIV-1, cross-resistance within each of the three antiretroviral drug classes—nucleoside reverse transcriptase (RT) inhibitors, nonnucleoside RT inhibitors, and protease inhibitors—often leads to the development of multidrug resistance ([1]). This process is accompanied by intensive mutations in the virus protease and reverse transcriptase. But possible HIV-1 protein variations tend to be limited by patterns of amino acid covariations ([2]).

Most of the published data on drug resistance to protease and RT inhibitors cover mutations in HIV-1 Subtype B isolates. On the other hand majority of infection cases in Belarus is dealt with HIV-1 Subtype A (svetlogorsk variant). Inconsistency of analytical data on drug-resistance mutations in HIV-1 Subtype A strains does not allow developing effective treatment regimen.

This work is a part of a project that aims to develop methods and propose some tools to analyze HIV-1 Subtype A strains. The goals of the project are: 1) analyze HIV-1 Subtype A protease and reverse transcriptase from treated and untreated persons; 2) recognize protease and RT- inhibitor associated mutations and clusters of correlated mutations and evaluate their contribution to the development of drug-resistance; 3) establish mutations both in Subtype B and Subtype A HIV-1 strains that lead to similar effects; 4) develop software to recognize and analyze HIV-1 Subtype A mutations.

In this paper we describe the developed statistical method that can be used to detect and analyze HIV-1 protein correlated mutations.

2. METHODS AND ALGORITHMS

Formalization of the subject domain is a very important stage for successful realization of the task. We carried out decomposition of the initial problem into subtasks as the first step to get mathematical problem definition.

The following subtasks should be solved sequentially to get solution of the initial task:

1. Detect mutations that were caused by antiretroviral therapy;
2. Determine correlated mutations;
3. Detect mutations that resulted in development of drug-resistance surveillance.

We used different methods to solve each subtask.

2.1 Gene sequencing

Aligned gene pol (HIV-1, subtype B) sequences from NCBI and Los-Alamos databanks were used for the purpose of the experiment Detailed patient's histories of antiretroviral treatment were taken into consideration to acquire essential information for the analysis.

Mutations are defined as changes to the nucleotide sequence of the considered consensus sequence compared to the reference sequence of the analyzed HIV-1 subtype.

In this paper we will speak only about a gene pol segment which is responsible for synthesis of viral protease. Data on every person are obtained in the form of aligned HIV-1 primary protein sequences and reflect the dynamic of mutation development in connection with inhibitor treatment. The appropriate protein sequence length is 99 amino acids. In case of more than two sequences we deal with multiple sequence alignment (MSA).

2.2 Association with antiretroviral therapy

To analyze association of protease mutations with antiretroviral therapy we used contingency tables.

The table of occurrences (Fig. 1) is used to determine, whether a mutation in position i of the sequence was caused by treatment. The table is constructed separately for each drug and each tested site in the sequence. Thus we can clarify which drug exactly caused the mutation.

X\Y	**Yes**	**No**	Σ
Yes	n_{11}	n_{12}	n_{1o}
No	n_{21}	n_{22}	n_{1o}

Σ	n_{o1}	n_{o2}	n

Fig.1 - Contingency table of characteristics *X* and *Y*.

In this notation X indicates if any treatment was carried out {Yes/No}, Y indicates if mutation is detected {Yes/No}, n_{ij} - number of sequences, for which characteristic X has gradation i, and characteristic Y has gradation j.

The situation above can be describe by a probability model, where the i-th line $(p_{i1}, p_{i2}) = \left(\frac{n_{i1}}{n}, \frac{n_{i2}}{n}\right)$ is a random sample from polynomial distribution with probabilities (q_{i1}, q_{i2}) and fixed number of observations n_{io}.

Introduce null hypothesis of independence of characteristics X and Y:

$$H^0 : q_{ij} = \frac{q_{0j}}{r}, \ \text{где} \ q_{0j} = \sum_{i=1}^{r} q_{ij} \qquad (1)$$

Based on the data from contingency table we build an assessment:

$$\chi^2 = n \cdot \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - n_{i0} n_{0j})^2}{n_{i0} n_{0j}} \sim \chi_1^2 \ npu \ n \to \infty \qquad (2)$$

Statistical hypotheses testing looks like:

$$\begin{cases} H^0 : P(\chi^2) \geq \alpha \\ \overline{H}^0 : P(\chi^2) < \alpha \end{cases}, \qquad (3)$$

where $\alpha \in (0,1)$ - assigned significance level.

P-value is calculated as follows to get numerical criteria:

$$P(\chi^2) = 1 - F_1(\chi^2), \qquad (4)$$

where $F_1(\chi^2)$ is a cumulative distribution function of Chi-square with one degree of freedom.

It is possible to apply procedure of a multiple hypothesis testing ([3]) to test *k* sites at once where mutations accrued. In this case *k* p-values should be ordered from the least to the greatest $p_{(1)} \leq \ldots \leq p_{(k)}$. Let $H^0_{(1)}, \ldots, H^0_{(k)}$ are the corresponding hypotheses. Assume that i^* is the least integer from 1 to k, such that $p_{(i^*)} > \frac{\alpha}{k - i^* + 1}$. Then hypotheses $H^0_{(1)}, \ldots, H^0_{(i^*-1)}$ are rejected, and hypotheses $H^0_{(i^*)}, \ldots, H^0_{(k)}$ are accepted.

2.3. Detecting the correlated mutations

To identify correlated mutations we divided sequences from patients into groups according to the treatment they accepted. This justifies assumption that similar environmental factors affected sequences from the same group.

To investigate correlations between site i and site j (positions in a sequence), we used evolutionary information obtained from phylogenetic tree restored for each group of sequences.

Consider a group of sequences obtained from patients with similar treatment histories. Let N is a number of sequences in the group. Assume that all sequences are aligned. So we deal with multiple sequence alignment (MSA). Phylogenetic trees are built by use of maximum likelihood method ([5]) with γ-distribution of site evolution rates. Bootstrapping procedure is applied to estimate confidence value.

For each site i vector $V_i = (V_{i1}, V_{i2}, \ldots, V_{iN})$ is introduced, where N is a number of sequences in MSA, V_{ik} is amount of mutations accrued on the route from the tree root to leaf k. In case of positions i and j which should be analyzed, we have vector $V_i = (V_{i1}, V_{i2}, \ldots, V_{iN})$ that corresponds to site i, and vector $V_j = (V_{j1}, V_{j2}, \ldots, V_{jN})$ that corresponds to site j.

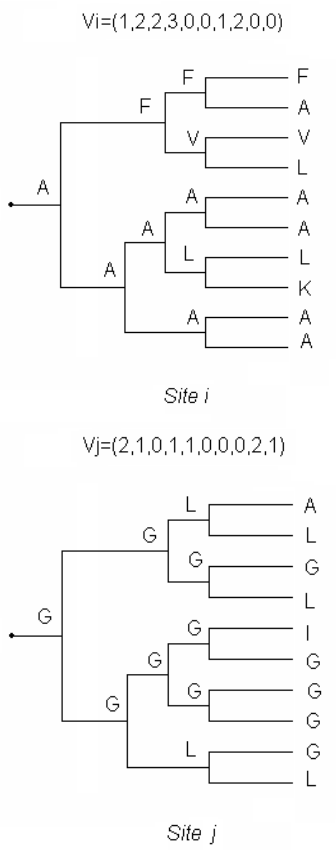

Fig.2 – Example of tree topology and calculated vectors V_i, V_j to find correlation between sites i and j.

Connectivity between sites can be measured by use of Pearson correlation coefficient:

$$\rho_{ij} = \frac{\mathrm{cov}(V_i, V_j)}{\sigma_{V_i} \cdot \sigma_{V_j}}, \quad (5)$$

So, there is a sample of size m consisting of vectors of N components. Assume that the observed characteristics obey Gaussian distribution.

Define hypothesis H^0 of independence of vectors V_i, V_j, i.e. $H^0: \rho_{ij} = 0$ and alternative hypothesis $H^1 = \overline{H}^0$. Statistical hypothesis test for H^0, H^1 is:

$$\begin{cases} H^0 : P(\rho_{ij}) \geq \alpha \\ \overline{H}^0 : P(\rho_{ij}) < \alpha \end{cases}, \quad (6)$$

where $\alpha \in (0,1)$ is a given significance level, p-value can be calculated under the following formula ([4]):

$$P(\rho_{ij}) = 1 - \int_{-|\rho_{ij}|}^{+|\rho_{ij}|} p_0(u)du, \quad (7)$$

where

$$p_0(u) = \frac{\Gamma\left(\frac{N-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{N-2}{2}\right)}\left(1-u^2\right)^{\frac{N-4}{2}}, \; u \in [-1,+1]$$

is density of ρ_{ij} if null-hypothesis is correct.

If site i correlates with site j, and site j correlates with site k, we have a cluster of correlated mutations in sites i, j, k. The strength of correlation can be estimated as follows:

$$\rho = \min_{i,j \in S}\{\rho_{ij}\} \quad (8)$$

2.4. Detecting the resistant mutations

Based on patient charts and well-characterized histories of antiretroviral treatment, this stage consists in detecting whether mutation in position i leads to development of drug resistance. Again, we propose to use contingency tables for this purpose.

For each site i of the sequence a table of occurrences is built (Fig. 3). In this case X denotes virus load, Y indicates whether mutation in the considered position occurred.

X\Y	$< \delta$	$\geq \delta$	Σ
Yes	n_{11}	n_{12}	n_{1o}
No	n_{21}	n_{22}	n_{1o}
Σ	n_{o1}	n_{o2}	n

Fig.3 – Contingency table for drug resistance analysis

Further we apply statistical hypothesis test as described above in 2.2.

3. RESULTS

Here we report results of numerical experiments that were carried out under a number of HIV-1 Subtype B protease sequences.

Data from NCBI and Los-Alamos databanks were used to test the correctness of the method. Patient's histories of antiretroviral treatment and analytical conclusions on the provided data were taken into consideration.

Firstly, we detected mutations associated with antiretroviral therapy in given samples of HIV-1 protease sequences. For instance, mutations in positions Q7, L10, L33, I84 were analyzed in one of the experiments. Labels AiB should be read as follows: substitution of amino acid A with amino acid B in position i of the sequence. To make final decision multiple hypothesis test was carried out with significance level $\alpha = 0.05$. Results are summarized in Table 1.

Table 1. Detecting treatment associated mutations

Mutation	Drug name	Amount of sequences	p-value	Accepted hypothesis
Q7H	APV	47	0.19	H^0
	LPV	49	0.21	H^0
	NFV	995	0.0	H^1
L10I	APV	47	0.0	H^1
	IDV	865	0.0	H^1
	NFV	997	0.0	H^1
L10V	APV	47	0.78	H^0
	IDV	865	0.002	H^1
	NFV	990	0.62	H^0
L33F	APV	47	0.001	H^1
	LPV	50	0.0	H^1
	NFV	1010	0.0	H^1
L33V	IDV	873	0.26	H^0
	SQV	276	0.51	H^0
	NFV	1007	0.89	H^0
I84V	APV	49	0.0	H^1
	LPV	51	0.0	H^1
	NFV	1006	0.0	H^1

Thus, mutations L10I, L33F, I84V have been detected as treatment associated. Moreover, these mutations occurred in sequences from patients treated by any protease inhibitor type. This agrees with information obtained from other sources ([6], [7]).

Phylogenetic tree was constructed by use of maximum likelihood method to recognize correlated mutation in

HIV-1 Subtype B protease. The segment of gene pol that is responsible for coding of protease was taken for this purpose. Only aligned sequences were used.

Because of extensive procedure we are listing result only for a few sites: Q7, L10, T12, K20, V32, L33, E35, M36, N37, P39, I47, G48, I54, V82, I84. This test was carried out using MSA consisting of 67 sequences obtained from patients treated by protease inhibitors. Results are listed in Table 2.

Table 2 - Analysis of sites for co-evolution

Site i		Site j		Coeff. ρ
Amino acid	**PI-associated mutations**	**Amino acid**	**PI-associated mutations**	
Q7	-	L10	10F,10I,10R,10V	0.06
Q7	-	I84	84V,84C	0.07
L10	10F,10I,10R,10V	L33	33I,33F	0.12
L10	10F,10I,10R,10V	I84	84V,84C	0.3
T12	-	K20	20R,20T,20V,20I, 20M	0.14
T12	-	P39	-	0.01
K20	20R,20T,20V,20I, 20M	M36	36V,36I	0.39
V32	32I	I47	47V,47A	0.56
L33	33I,33F	I84	84V,84C	0.11
L33	33I,33F	Q7	-	0.02
E35	35G,35D	N37	37D,37E	0.24
G48	48V	I54	54V,54A, 54M,54L, 54T	0.23
I54	54V,54A,54M, 54L,54T	L10	10F,10I,10R,10V	0.41
V82	82A,82T,82F	L10	10F,10I,10R,10V	0.3

Sites with correlation value $|\rho| \geq 0.2$ were considered to be related. In this case positions L10, I82; L10, I84; K20, M36; M32, I47; E35, N37; G48, I54; L10, I54 appeared to be correlated. This agrees with the results in other researches ([1], [2]).

3. CONCLUSION

Treatment-associated mutations have different effect in drug-resistance development. Some of them may serve as specific markers of HIV-1 drug resistance or make a synergetic effect being clustered as correlated mutations.

Different statistical methods including chi-square and regression analyses, hypothesis testing and some exhaustive search techniques are used to analyze protein and DNA sequences for different purposes. Here we proposed an adequate method to analyze RT and protease sequences from patients with known treatment histories.

Nowadays, there are some software packages used in Belarus to analyze HIV-1 sequences and mutations. Analytical facilities of these packages allow comparing of the input data with already known results published in different internet databanks ([8]). But poor or no offline decision making facilities are available. In this work we proposed a method that can be implemented and become a supplementary source of information for decision making on treatment regimen. This makes such offline analytical facility extremely valuable in highly variable environment and lack of analytical data on HIV-1 Subtype A strains which is mostly spread in Belarus.

The drawback of this approach is high computational complexity to get evolutionary information about the sequence sites. Because of statistical methods used for data analysis, availability of sufficient amount of sequences provided with detailed patient charts is critical to get correct results.

4. AVAILABILITY

At present the project is in progress. Published HIV-1 Subtype B protein (pol gene) sequences from Los Alamos HIV Database http://www.hiv.lanl.gov/content/ and Stanford HIV Drug Resistance Database http://hivdb.stanford.edu/ are available and are used in tests.

On the other hand we started data collection and HIV-sequencing form infected patients in Belarus. As soon as HIV-1 sequencing is finished, this method will be applied to analyze the data from patients in Belarus.

Finally, we expect that results of this work will be used in developing of antiretroviral therapy and treatment regimen to prevent drug-resistance establishment. Software implementation is planned to be available.

5. REFERENCES

[1] S.-Y. Rhee, W.J. Fessel et al. *HIV-1 Protease and Reverse Transcriptase Mutations: Correlations with Antiretroviral Therapy in Subtype B Isolates and Implications for Drug-Resistance Surveillance*, The Journal of Infectious Diseases, 2005;192:456-65.

[2] S.-Y. Rhee, T.F. Liu, S.P. Holmes, R.W. Shafer *HIV-1 Subtype B Proteaseand Reverse Transcriptase Amino Acid Covariation*, PLoS Comput Biol 3(5): e87. doi:10.1371/journal.pcbi.0030087

[3] Benjamini, Y., and Y. Hochberg. 1995. *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J. R. Stat. Soc. Ser. B 57:289–300.

[4] Y.S. Kharin, M.S. Abramovich, V.I .Maliugin *Statistics: electronic textbook*, Minsk, BSU, 2000

[5] Abbas A., Holmes S. *Bioinformatics and Management Science: Some Common Tools and Techniques*. Operation Research, Vol.52, No.2, March-April 2004, pp.165-190

[6] *National Center for Biotechnology Information* [Electronic resource] – 1988-2008. - Mode of access: http://www.ncbi.nlm.nih.gov/

[7] *Los Alamos HIV Database* [Electronic resource] – 2005-2008. - Mode of access: http://www.hiv.lanl.gov/content/