Stochastic Modeling of Temporal Variability of HIV-1 Population

Ilia Kiryukhin[1], Kirill Saskov[1], Alexander Boukhanovsky[1], Wilco Keulen[2], Charles Boucher[3], and Peter M.A. Sloot[4]

[1] Institute for High Performance Computing and Information Systems,
191186 St.Petersburg, Russia
{ilia, kirills, avb}@fn.csa.ru
http://www.csa.ru

[2] Virology Network,
69042 Utrecht, The Netherlands
wilco.keulen@vironet.com

[3] University Medical Center,
3508 GA Utrecht, The Netherlands
boucher@lab.azu.nl
http://www.azu.nl

[4] University of Amsterdam,
1098 SJ Amsterdam, The Netherlands
sloot@science.uva.nl
http://www.uva.nl

Abstract. A multivariate stochastic model for describing the dynamics of complex non-numerical ensembles, such as observed in Human Immunodeficiency Virus (HIV) genome, is developed. This model is based on principle component analyses for numberized variables. The model coefficients are presented in the terms of deterministic trends with correlated lags. The results indicate that we may use this model in short-term forecast of HIV evolution, for evaluation of HIV drug resistance and for testing and validation of diagnostic expert rules. The model also reproduces the specific shape of the bi-modal distribution for the mutations number.

1 Introduction

Recently HIV genome analysis has become a routine medical procedure. The obtained relevant drug resistance mutations allow finding drugs suitable for antiretroviral treatment based on expert experience. Recent experience indicate that virological response is significantly larger when highly active antiretroviral therapy (HAART) is applied on the basis of resistance testing, although this approach does not always predict virological success [1–3]. In any case, if a patient takes antiretroviral drugs, the individual HIV population evolutes during disease history. Most studies on individual

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2657, pp. 125–135, 2003.

HIV population dynamics are based on numerical simulations such as population dynamics based models [4] or cellular automata based models [5].

These approaches consider the individual (for each patient) HIV population only. The practice of HIV treatment however, has shown that patients may be infected by mutated viruses from other patients [6]. This implies that the evolution of total world populations of HIV and the associated changing of the related drug resistance levels, should be taken into account. Since the characteristics of world HIV population dynamics are determined by a huge amount of detailed, specific factors, one of the most promising approaches for the study of these phenomena is probabilistic modelling, based on recent HIV statistics. For the analysis we describe in this paper, the large databases of HIV-infected patients, collected over several years in USA, is used [7]. These databases contain genotypes of 43620 patients examined from August 9, 1998 to May 5, 2001. We observed 59 different mutations in the RT genome, including 17 mixed mutations, and 77 different mutations in the protease genome, including 34 mixed mutations. The developed probabilistic model described in this paper takes into account the peculiarities of initial data and specifics of the underlying dynamics. Recent probabilistic models for genome ensembles are mainly directed towards the evaluation of specific parts in the genome, or to choose the closest related pattern [8]. In the development of a stochastic model for the temporal variability of the global HIV-population we have to address the follow problems:

- The drug resistance depends on combinations of mutations. So, the probabilistic model must take into account the total variability of genome. The dimensionality of data is high (all amino acids and positions of mutations in viral genome),
- The initial data is non-numerical; therefore the well-developed standard procedures of multivariate statistics are inconvenient.

2 Multivariate Stochastic Model of HIV Genome Ensemble

Let us consider the following model of a data representation: all genome samples ($k = \overline{1, M}$, M is number of patients in a considered time interval, e.g. month) consist the literal corteges given by $X_k = \{x_{kj}\}_{j=1}^{n}$, where n is number of relevant positions in the genome. Each $x_{kj} \in V$, where $V = \{v_i\}_{i=1}^{m}$ are literal marks for the amino acids (A, C, D, E, F, …). When mutations are absent, the corresponding cortege X_k (so called "wild-type" virus), may be associated with some initial value (centroid) for example, $\overline{X}$. Such consideration allows us to compare several terms in the sample, taking into account its proximity to the "wild-type" virus.

Note, that the analysis of the marginal mutations is not enough for general description of all genome ensemble variability, because some positions of genome may be statistically dependent [8], especially in accordance with viral fitness. For the reduction of the dimensionality and further modeling of such data, powerful procedures (principal component of factor techniques) of multivariate statistical analysis have been developed [9,10]. In reference [11] the generalization of these methods for analy-

sis of temporal tendencies is described. However, the general problem we face is that all these procedures are developed only for *numerical* values.

Therefore we propose a three-stage statistical procedure for the HIV genome model: discretization and enumeration, reduction of dimensionality and temporal analysis. The general scheme of the proposed procedure is shown in Fig. 1.

Discretization and enumeration of literal corteges. One of the possible ways to apply the classical MSA procedures is the numberization (generation of number marks for non-numerical values) of non-numerical data. An adequate numberization procedure is based on the estimation of contingence (probability) tables $F_{(m)}$ for groups of m mutations. If $m = 1$ then $F_{(1)} = \{p_{ij}\}$, where p_{ij} is probability of the amino acid occurrence with a literal mark v_i in relevant position j. If $m = 2$, then table $F_{(2)} = \{p_{pj}^{si}\}$, where p_{pj}^{si} is probability of simultaneous occurrence of i-th and j-th mutations in positions p and s. Table $F_{(2)}$ consists of $n(n-1)/2$ independent blocks F_{ps}.

In accordance with [12], for reduction of dimensionality it is better to use number marks obtained by a procedure of matrix $F_{(2)}$ "reflection" on the n-dimensional Gaussian distribution $N_n(\mu, K)$, where p_{pj}^{si} corresponds to the "wild-type" virus $\overline{X}$, and correlation matrix $K = \{\rho_{ij}\}$, obtained from the optimization problem

$$Q = \sum_{i<j} \rho_{ij}^2 \to \max, \quad \rho_{ij} = C' F_{ij} C . \tag{1}$$

where *C* is a vector of numbered marks for the amino acids indices. Because vector C defines the numeric scale only, the realizations of marks in scale gradations are obtained by means of a Monte-Carlo simulation (such as the uniform distributed value in each gradation).

The result for the first stage of this procedure (see Fig. 1) is (a) the transformation of initial literal corteges to sample Gaussian random vectors $U = \{U_k\}_{k=1}^n$, where n is number of relevant positions in genome and (b) a simplification of the model.

Principal component analysis and the factor model. For reducing of the data dimensionality a principal component (PC) approach is widely used [13]. It allows us to represent each centered vector $U^{(0)} = U - \mu$ as orthogonal expansion on the basis $\varphi_m = \{\varphi_{mk}\}_{k=1}^n$

$$U^{(0)} = \sum_{k=1}^{n} a_k \varphi_k . \tag{2}$$

Here a_k is the coefficient of expansion. The eigen-basis of expansion (2) (so called empirical orthogonal functions, EOF) are given by the principal axes of multivariate genome ensemble. The eigenvectors of the correlation matrix is given by

$$K_U \varphi_m = \lambda_m \varphi_m . \tag{3}$$

Here the eigenvalues λ_m are the variances of the principal components.

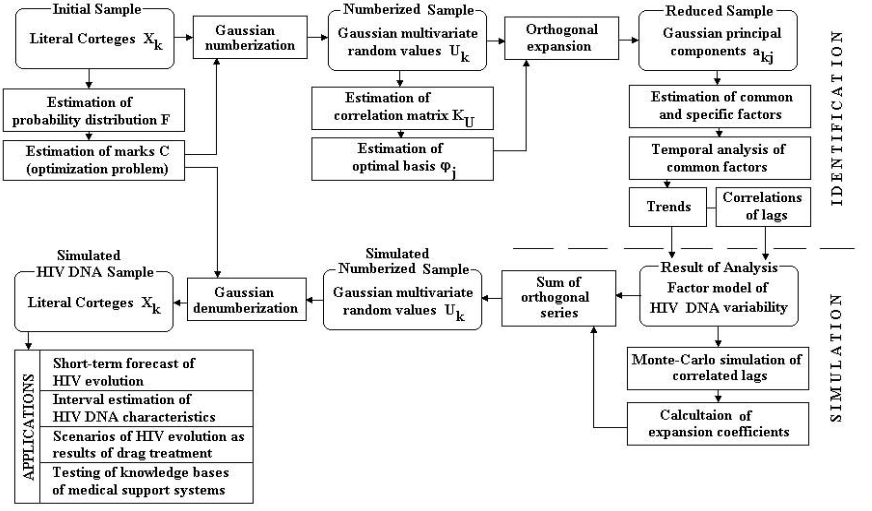

Fig. 1. General scheme for stochastic modeling of HIV population variability

The convergence of expansion (2) is associated with the index:

$$D_m = \frac{\sum_{k=1}^{m} \lambda_k}{\sum_{k=1}^{n} \lambda_k} \cdot 100\% . \tag{4}$$

Using the PC expansion (2) with M first terms only (where M is obtained by the PC significance tests, see [13]), allows us to consider the factor model of temporal variability of the genome ensemble

$$U(t) = \mu + \sum_{k=1}^{M} a_k(t)\varphi_k + \varepsilon_t . \tag{5}$$

Here $a_k(t)$ – are the time series of the expansion coefficients (that will be obtained by inverse transformation of (2) based on the orthogonal properties of the basis), ε_t is the Gaussian white noise. In terms of factor analysis the values $a_k(t)$ may be considered as common factors, driving the temporal variability of the genome ensemble, and ε_t – as the specific factor.

Thus, the results of the second stage is that we only need to consider a set of independent factors $a_l, l = \overline{1,M}$, instead of all the high-dimensional samples of $U = \{U_k\}_{k=1}^{n}$, where $M < n$.

Temporal variability of common factors. The main advantage of the proposed representation (5) is that all the common factors are independent. So, it allows us to

reduce the description of temporal variability of the HIV genome ensemble to analysis of time series of each factor independently.

Note that the individual HIV population dynamics has rather slow temporal changes, so, for temporal analysis monthly time intervals are required, however, all the data is distributed non-uniformly (per month). Thus, for probabilistic modeling the combined distribution approach can be used:

$$F_a(x) = \int_{-\infty}^{\infty} G_a(x,\xi) f_a(\xi) d\xi . \tag{6}$$

where $F_a(x)$ is the total distribution of each common factor, $G_a(x,\xi)$ is the short-term (intra-month) distribution of factor a, and $f_a(\xi)$ is the long-term (month-to-month) distribution for parameter ξ of the short-term distribution. Thus, once we know the type of $G_a(x,\xi)$, all the analyses of temporal variability can be done in terms of parameters ξ (mean value, variance, characteristic quantiles) only.

The model for the temporal variability of ξ is presented in the form of [11]:

$$\xi(t) = \xi^*(t) + \delta(t) . \tag{7}$$

where

$$\xi^*(t) = \sum_k \alpha_k \phi_k(t) . \tag{8}$$

is the deterministic part (trend) with fixed coefficients α_k, defined on some basis functions $\phi_k(t)$, e.g. $\phi_k(t) = t^k$,

$$\delta(t) = \sum_j \beta_j \delta(t-k) + \gamma(t) . \tag{9}$$

is the stochastic part, presented as a autoregressive model [14] with coefficients β_j and white noise $\gamma(t)$.

Thus, the third (final) stage results in a parametric model (5,7–9) of common factors with parameters α_k, β_j (see Fig. 1).

3 Identification and Interpretation

The stochastic model (5,7–9) is applied to the above-mentioned database. Let us note, that protease relevant mutations are independent from RT relevant mutations, because they are caused by different groups of antiretroviral drugs – protease inhibitors (PIs) and RT inhibitors (RTIs.

Table 1. Input of PCs (%) in total variability of protease and RT parts of genome ensemble

#PC	1	2	3	4	5	6	7	8	9	10
Protease	16.5	15.3	11.1	8.8	8.7	7.5	7.2	6.8	4.7	3.8
RT	18.5	15.8	10.9	9.8	6.7	5.8	5.2	4.9	4.5	3.5

In table 1 the input (4) of each PC in total variability are shown for protease and RT separately. It is seen, that in both cases convergence of the expansions are satisfactory, e.g. for RT the five PCs explain only 61.7% of total variability, and the ten PCs – 85.6%. For protease these values are 60.4% and 90.4%. Following the criterion $D_M < (1/n)\%$, in (5) for RT is enough M=9 (82.1%), and for protease M=8 (81.9%). The deviation from 100% for these values is explained by the specific factor ε_t only.

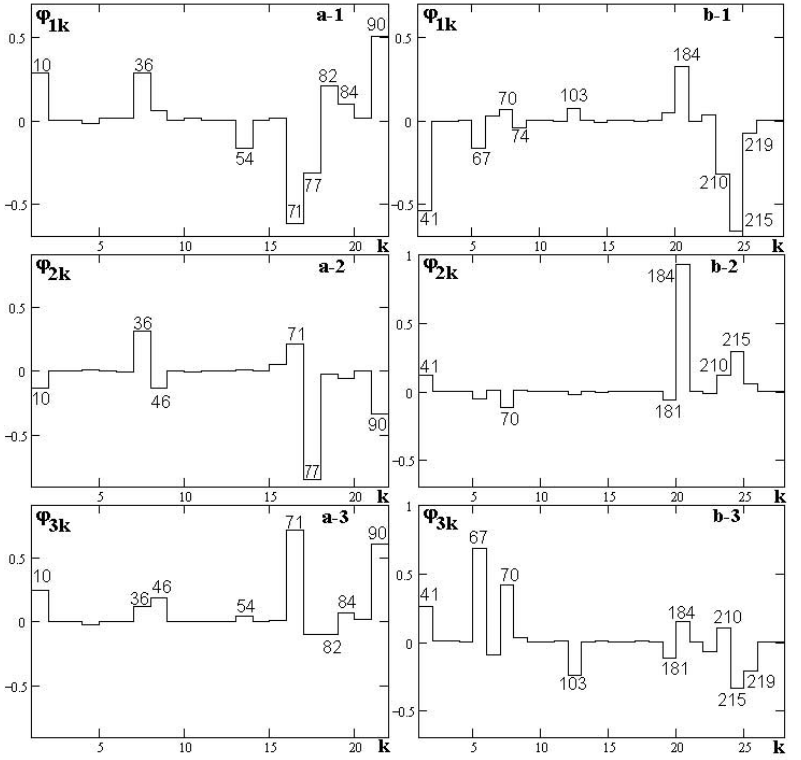

Fig. 2. First 3 empirical orthogonal functions for the protease (a) and RT (b) relevant mutations.

In the Fig. 2 the first 3 empirical orthogonal functions are shown for protease and RT. The influence of different codons is readily observed from these figures, for instance the1st EOF for protease shows a balance between the main groups of mutations in positions (10,36,82,90) and (54,71,77) of protease. However, for defining the concrete types of mutations it is necessary to do a 'denumberization', (see Fig. 1).

The temporal analysis of the common factors in terms equations (6–9) show that for protease only the first coefficient in the expansion (5) has a statistically significant trend. For RT the behavior of the coefficient is more complicated because, although the hypothesis of trends for the 2nd and 3rd components are approved by Fisher's criterion, the expression of these trends (by determination coefficient R^2) are very weak. For example, in table A.1 results are shown of the linear trend (7) analysis for some of the quantiles (e.g. 25%, 50%, 75%) of (5) for the first 3 coefficients for protease and RT. In both cases the coefficient $a_1(t)$ has a significant deterministic part (7) (trend); the behavior of other coefficients may be considered in terms of stationary time series (9) only.

Therefore, the results of this identification are given by the set of model parameters: mean vector μ, M empirical orthogonal functions φ_k, vector of variances D_ε of specific factor, coefficients β_{kj} for expression (9) of each common factor $a_k, k = \overline{1,M}$, and two coefficients α_{11}, α_{12} of linear trend (8) for a_1. These two coefficients are enough for the description of all of the evolutionary part of the HIV genome temporal variability.

4 Simulations and Verification

After the definition of the model given in (5) with basis function from (3) and coefficients with distributions (6-9), we can now perform stochastic simulations of the model ensembles of HIV genome. As seen from Fig. 1, the initial step of the simulation is the calculation of time series for the stochastic part (9) by means of a autoregressive approach, see [14]. After that the sum (7) of trend (8) and lag (9) are calculated for all *t*. The next step is a Monte-Carlo generation of all the statistically independent coefficients $a_k(t)$ with the distributions $G(x,\xi_k)$ in (6), where parameters ξ_k were estimated in the previous step. Finally, the sum of the orthogonal series (5) with coefficients $a_k(t)$ is computed.

The result of this procedure is a Gaussian random vector; for obtaining the literal representations of genome the inverse procedure (denumberization) is used. This procedure associates the numerical value in a fixed position with a concrete scale vector C, obtained from (1). Thus, the result of the simulation is an ensemble of corteges, which consist of relevant HIV genome mutations.

The model (5,7–9) can be verified on probabilistic characteristics of the ensemble, that were not used in the identification procedure. Here for illustration we consider the integral characteristic of the genome variability – distribution P(k) for a number k of all the mutations. Obviously, P(k) is a result of the joint occurrence of mutations, and its use to model the verification is valid.

In Fig. 3 these distributions for the whole ensemble (1998-2001) are shown. It is clearly seen, that all the curves are bi-modal (first maximum is 2–3 mutations, and second one – 5–6 mutations). Nevertheless, this bi-modal shape is conservative for

monthly distributions $P_t(k)$ (see boundaries of tolerant (min, max) intervals in fig. 3). Therefore we expect that there are two independent groups of genomes, corresponding to the low and high number of mutations.

The discovered bi-model distribution is approximated by a mixture of Bernoulli distributions [15]:

$$P(k) = pC_{m_1}^k q_1^k (1-q_1)^{m_1-k} + (1-p)C_{m_2}^k q_2^k (1-q_2)^{m_2-k}. \quad (10)$$

where p is an entry of the first group of mutations (and (1–p) is an entry of the second group, m_1, m_2 – are maximal numbers of mutations in groups and q_1, q_2 – are probabilities of a single mutation in the groups. The results of the approximation given by (10) are shown in Fig.3. It is seen, that the approximated and sample data are close to each other. Also shown in fig.3 are the tolerant intervals, obtained as **(min, max)** of the monthly distribution. These values reflect the boundaries of variability of the distribution shape in different months.

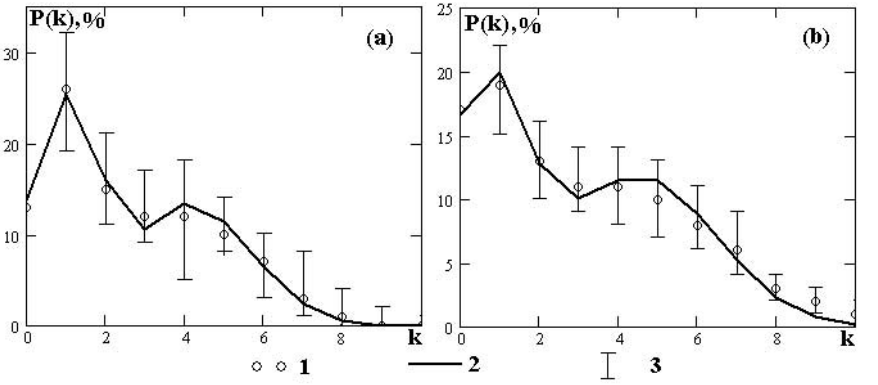

Fig. 3. Distributions of total number of mutations in protease (a) and RT (b). (1 – sample estimates for 1998-2001; 2 – approximation by (3); 3 – tolerant intervals (min, max of monthly data)

The parameters $(p, q_1, q_2, m_1, m_2)_t$ of the P(k) approximation for the total ensemble (1998-2001), and characteristics of its temporal trends are shown in table A.2. It is seen that in both cases only the weight values p have a clear significant trend. For protease the weight of the left part (group of m_1 mutations) increased from 39% in Summer, 1998 to 62% in Summer 2001 (with average increment a=0.74% per month). One interpretation is that we have two groups of patients. One group is the "new" patients, that had one or two treatments, thus their genotype contains relative small numbers of mutations. The second group is the "old" patients, which have a long treatment history. In the same table the parameters of the simulated ensembles are shown.

Let us note, that the stochastic model (5,7–9) is very sensitive to the reproduction of the value P(k), because the joint occurrence of k>>1 mutations is a rare event. For example, in table 2 the results of verification on the most simple value – mean number of mutations (mathematical expectation of P(k)) are shown.

Table 2. Model verification: prediction for mean number of mutations (point estimates)

Genome fragment	Order of model					Sample
	5	10	15	20	25	
Protease	1.57	2.43	2.61	2.70	–	2.64
RT	1.95	2.68	2.95	3.05	3.08	3.14

From table 2 we observe that for low order M the mean number of mutations is less then sample estimated one. But for increasing M the simulated and sample estimates became comparable, thus indicating the validation of the applied method.

5 Conclusions

A multivariate stochastic model, based on principle component analyses for numberized variables, is proposed to describe the variability of HIV genome populations. The temporal analysis of the model coefficients in terms of (6-9) show that only the first coefficients have significant trends. It allows to use this fact in short-term forecast of HIV evolution.

Verification of the proposed model indicated that this model may be used for simulations in future studies of HIV drug resistance, and for testing and validation of diagnostic expert rules (see Fig.1).

References

1. Durant J, Clevenbergh P, Halfon P, et al. Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomised controlled trial. Lancet 1999; 353:21959.
2. Baxter JD, Mayers DL, Wentworth DN, et al. A randomized study of antiretroviral management based on plasma genotypic antiretroviral resistance testing in patients failing therapy. CPCRA 046 study team for the Terry Beirn Community Programs for Clinical Research on AIDS. AIDS 2000; 14:F8393.
3. Zollner B, Feucht HH, Weitner L, Adam A, Laufs R. Drug-resistant genotyping in HIV-1 therapy. Lancet 1999; 354:112021.
4. Maree AF, Keulen W, Boucher CA, De Boer RJ. Estimating relative fitness in viral competition experiments. J Virol 2000 Dec;74(23):11067–02.

5. P.M.A. Sloot, F. Chen and C.A. Boucher: Cellular Automata Model of Drug Therapy for HIV Infection, in S. Bandini; B. Chopard and M. Tomassini, editors, 5th International Conference on Cellular Automata for Research and Industry, ACRI 2002, Geneva, Switzerland, October 9–11, 2002. Proceedings, in series Lecture Notes in Computer Science, vol. 2493, pp. 282–293. October 2002.
6. Little SJ, Holte S, Routy JP, Daar ES, Markowitz M, Collier AC, Koup RA, Mellors JW, Connick E, Conway B, Kilby M, Wang L, Whitcomb JM, Hellmann NS, Richman DD. Antiretroviral-drug resistance among patients recently infected with HIV. N Engl J Med 2002 Aug 8;347(6):385–94
7. Genotype database is obtained from a large service testing laboratory from the US. It contains the resistance profiles of the Protease and Reverse Transcriptase genes of the HIV-1 virus obtained from plasma samples of HIV-1 infected patients. No clinical background information on medication or drug history is available.
8. Mathematical Methods for DNA Sequences // Ed. M.S. Waterman. CRC Press Inc., Boca Raton, Florida, 1999.
9. Anderson T.W. An introduction to multivariate statistical analysis. John Wiley, NY, 1948.
10. Bartlett M.S. Multivariate analysis. J. Roy. Stat. Soc. Suppl. 9(B), 1947, 176–197.
11. Brilinger D. Time series. Data analysis and theory. Holt, Renehart and Winston, Inc., New York, 1975.
12. Aivazyan S.A., Buchstaber V.M., Yenyukov I.S., Meshalkin L.D. Applied statistics. Classification and reduction of dimensionality. Finansy i statistica, Moscow, 1989, 608 p. (in Russian)
13. Johnson R.A., Wichern D.W. Applied multivariate statistical analysis. Prentice-Hall Internalional, Inc., London, 1992, 642 pp.
14. Ogorodnikov V.A., Prigarin S.M. Numerical modelling of random processes and fields: algorithms and applications. VSP, Utrecht, the Netherlands, 1996, 240 p.
15. Wolfe J.H. Pattern clustering for multivariate mixture analysis. Miltiv. Behav. Res., 1969, 22, pp. 165–170.

Appendix

Table A.1. Trend analysis of first 3 PCs for protease and RT

	1-st PC			2-nd PC			3-rd PC		
	25%	50%	75%	25%	50%	75%	25%	50%	75%
				Protease					
Trend **a**	-0.024	-0.030	-0.031	0.015	0.003	0.001	0	0.001	0
F_{sample}	74.2	64.1	36.4	1.30	0.79	0.05	0.16	3.87	0.34
R^2	0.70	0.67	0.54	0.04	0.03	0.01	0.01	0.11	0.01
				Reverse transcriptase					
Trend **a**	0.086	0.038	0.020	0.001	-0.001	-0.013	-0.001	-0.001	-0.02
F_{sample}	175.7	88.8	44.35	3.56	6.09	17.6	5.37	11.39	22.33
R^2	0.85	0.74	0.59	0.10	0.16	0.36	0.15	0.26	0.42

Table A.2. Results of verification of stochastic model for protease and RT: sample and simulated trends of mutations distribution

Parameter	Total sample (1998-2001)	Monthly (min–max)	Trend **a** (%/month)	95% CI for **a**	F_{sample}	R^2
			Protease			
p ,%	48	28–67	0.74	0.57–0.91	64.0	0.67
q_1 ,%	47	29–65	-0.20	-0.48–0.09	1.97	0.06
q_2 ,%	46	39–65	0.18	-0.01–0.46	3.29	0.09
m_1	2	1–4	–	–	–	–
m_2	9	6–9	–	–	–	–
			Reverse transcriptase			
p ,%	47	37–59	0.49	0.34–0.63	94.3	0.75
q_1 ,%	16	10–20	-0.03	-0.32–0.25	0.57	0.02
q_2 ,%	34	31–38	0.07	-0.20–0.34	4.94	0.13
m_1	6	5–9	–	–	–	–
m_2	14	13–14	–	–	–	–

Both for protease and RT, F_{sample} is compared with Fisher's test $F(1,31,95\%) = 4.14$