Infection, Genetics and Evolution 78 (2020) 104134

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research Paper

Evidence of genomic information and structural restrictions of HIV-1 PR and RT gene regions from individuals experiencing antiretroviral virologic failure

Elidamar Nunes de Carvalho Lima[a,b,*], Rodrigo Sucupira Andrade Lima[a], José Roberto Castilho Piqueira[a], Maria Cecília Sucupira[b], Michelle Camargo[b], Juliana Galinskas[b], Ricardo Sobhie Diaz[b]

[a] Telecommunication and Control Engineering Department, Engineering School, University of São Paulo, São Paulo, Brazil
[b] Infectious Diseases Division, Department of Medicine, Federal University of São Paulo, São Paulo, Brazil

ARTICLE INFO

Keywords:
Genomic information
Entropy
Evolutionary patterns
PR and RT HIV-1
Antiretroviral virologic failure
Resistance mutation

ABSTRACT

Objectives: This study analyzed Protease-PR and Reverse Transcriptase-RT HIV-1 genomic information entropy metrics among patients under antiretroviral virologic failure, according to the numbers of virologic failures or resistance mutations.

Methods: For this purpose, we used genomic sequences from PR and RT of HIV-1 from a cohort of chronic patients followed up at São Paulo Hospital.

Results: Informational entropy proportionally increases with the number of antiretroviral virologic failures in PR and RT ($p < .001$). Affected regions of PR were related to catalytic and structural functions, such as Fulcrum (K20) *Flap* (M46) and *Cantilever* (A71). In RT, this occurred at *Fingers* (E44) and *Palm* (K219). Informational entropy increases according to the number of resistance mutations in PR and RT ($p < .001$). Higher PR entropy was proportional to the resistance mutation numbers in *Fulcrum* (L10), *Active site* (L24) *Flap* (M46), *Cantilever* (L63) and near *Interface* (L90). In RT, they related to regions responsible for protein stability such as *Fingers* (T39) and *Palm* (L100).

Conclusions: The antiretroviral selective pressure affects HIV genomic informational entropy at the PR and RT regions, leading to the emergence of more unstable virions. Mapping the three-dimensional structure in these HIV-1 proteins is relevant to designing new antiretroviral targeting resistant strains.

1. Background

HIV evolves according to environmentally selective pressure. As an example of this evolution, resistance-associated mutations selected by antiretrovirals ultimately lead to the loss and subsequent gain of genomic information (Smith, 2006; Shafer and Schapiro, 2008; d'Ettorre et al., 2011; Das and Berkhout, 2010; Adami, 2004; Sato and Ohya, 2011; Adami, 2012; Parkhomchuk, 2007). In this sense, the selection of pre-existing resistance mutations enables virological failures, dramatically changing the HIV-1 fitness. However, the continuous selective pressure imposed by a failing ART regimen may change the acquired genomic information again, inducing progressive fitness restoration (Sato and Ohya, 2011; Potter et al., 2001; Troyer et al., 2009; Song et al., 2012; Sunshine et al., 2015).

Common antiretroviral targets are HIV-1 PR (PR) and RT (RT) (Imamichi, 2004; Arhel and Kirchhoff, 2010; Das and Arnold, 2013). Structurally, PR is a protein comprising distinct regions named *N-Terminal*, *Fulcrum*, *Flap Elbow*, *Tip Flap*, *Active Site*, and *Cantilever (*Imamichi, 2004; Arhel and Kirchhoff, 2010; Das and Arnold, 2013;Ali et al., 2010; Dam et al., 2009). HIV-1 RT is an enzyme that converts the single-strand HIV RNA to a double-strand cDNA to be integrated into the host genome. Structurally, RT is a heterodimer comprising two subunits referred to as *p*66 (560 a.a. long) and *p*51 (440 aa long); the subunit *p*66 presents a structure that resembles the anatomy of a hand, with the regions named *Fingers*, *Palm*, *Thumb*, and *Connection (*Hu and Hughes, 2012;Singh et al., 2010;Asahchop et al., 2012).

The objective of this study is to analyze the evolutive characteristics of the functional and structural patterns of HIV-1 PR and RT regions of

* Corresponding author at: Telecommunication and Control Engineering Department, Engineering School, Automation and Control Laboratory, Escola Politécnica, Universidade de São Paulo, Avenida Prof. Luciano Gualberto – travessa 3 – 158, São Paulo, SP, Brazil.
E-mail addresses: elidamarnunes@usp.br (E.N. de Carvalho Lima), piqueira@lac.usp.br (J.R.C. Piqueira), rsdiaz@catg.com.br (R.S. Diaz).

https://doi.org/10.1016/j.meegid.2019.104134
Received 30 September 2019; Received in revised form 28 November 2019; Accepted 4 December 2019
Available online 10 December 2019

the pol gene, aiming to quantify variability patterns in a population under ARVs failure, to bring light to evolutionary process escape that may be occurring. The findings were obtained using informational entropy metrics (Durston et al., 2007; Thomas, 2010; Vinga, 2014) according to the previous numbers of ART-VF and resistance-associated mutations (RM) selected by PR and RT inhibitors. Our previous study showed that the number of received ART schemes increase the number of resistance-related mutations and there is a correlation between the ART class administered and the resistance mutations; besides that, there is a negative correlation between informational entropy and the laboratorial markers of HIV-1 disease progression (2018). The novelty of this study is that it quantifies, by the entropy metric, genomic information and KLD-gain information in experienced patients according to ARV-VF failure and resistance mutation amount, on a three-dimensional structural visualization. Although there are a lot of works on resistance mutations and virologic failure in the HIV PR and RT proteins, few studies show variability in specific sites of HIV resistant individuals to ARVs. This study brings lights to possible scape biological mechanism that may occur in these proteins. Quantifying and understanding the variability patterns, by genomic information amount, in an experienced population, can help understand the evolutionary scape of this virus, and it is useful to clarify the development of new drugs based on the structural patterns.

2. Methods

2.1. Ethics statement

The Federal University of São Paulo Ethics Committee and Human Research Ethics under number #0024/11 approved this study in February 2011, and the Brazilian Ministry of Health authorized it and, informed consent has not been obtained for this study, since HIV genomic sequences were generated for clinical purposes. All experiments, as well as all methods, were performed in accordance with relevant guidelines and regulations. All data generated or analyzed during this study are included in this manuscript.

2.2. Patients-sequences

This survey includes 651 sequences of PR and RT HIV-1 among individuals presenting antiretroviral virological failure (2018). To determine the consequence of multiple ART-VF and RM in HIV-1 entropy, the average informational entropy was calculated according to the number of previous ART-VF and RM divided into 1–3, 4–6, and 7–9. The results were compared with 73 samples of RT-naïve individuals, used as controls, also infected by clade B strains. The patients were treated with an average of three therapeutic schemes of ARVs, with average treatment time of 3.11 years, and were followed prospectively (2009–2011) in the city of Sao Paulo-Brazil (Souza et al., 2011). The frequency of the therapeutic exposition with combined classes ARVs to NRTI + NNRTI + PI were 37,9%, 5,5% and 10,7%. For the combined classes NRTI + NNRTI (26,7%, 9,7%) and NRTI + PI (6,9% e2,6%). For the NRTIs, the Lamivudine was the most used in the patients (6,0%), followed by Tenofovir (5,9%), Didanosina (2,7%), Zidovudina (1,5%), Estavudina (1,3%) and Abacavir in 0,7%. For patients treated with PI, the Lopinavir was the most used (25,5%), followed by Atazanavir (13,6%), Indinavir (11,9%), Nelfinavir (9,6%), Saquinavir (4,7%), Amprenavir (1,9%) and Darunavir in 0,7% of these patients. For NNRTIs, the Nevirapine was used in 13,9% of these patients.

The length of the PR fragment analyzed is 96 a.a. (dimer-PR: residues T4 to F99) and 207 a.a. for RT (subunits *p*51 and *p*66: residue C38 to I244) according to HXB2 genomic positions. This analysis was performed for both p52 subunits. Only one genomic sequence from each 651 individuals was selected in this study to be analyzed without considering insertions, deletions, a.a. ambiguities or stop codons. By the time these patients were included in this study, they presented an average of three virologic failures. RM related to PR Inhibitors (PI), Nucleoside RT 90 Inhibitors (NRTI) and Non-Nucleoside RT Inhibitors (NNRTI) were considered according to the IAS91 USA resistance mutation list that can be considered an updated list of drug resistance mutations audited by a large number of specialists in the field of antiretroviral resistance (Wensing et al., 2017). The PR and RT of HIV-1 genomic sequences were initially aligned (*ClustalW*) (Larkin et al., 2007), subtyped (*COMET HIV-1*) (Struck et al., 2014) and, aiming exclude possible recombinant forms in this sequence set due high frequency of BF recombinants circulating in Brazil, was made complementary analysis in jPHMM (Schultz et al., 2012) and then submitted to *Stanford* DB software for RM discrimination (Rhee et al., 2003). For PR and RT were considered 23 and 32 RM respectively (Attachment-1).

2.3. Variability patterns

To quantify the variability patterns in the PR and RT of HIV-1 according to numbers of VF or RM, the informational entropy was calculated based on the *Shannon* equation (Eq. (1)) (Sato and Ohya, 2011; Piqueira et al., 2006; Funkhouser, 2012; Gatenby and Frieden, 2007; Bohlin et al., 2012). The informational entropy is a metric commonly used to quantify the uncertainty of information per site or position in the genome, it calculates the fixation of this information or frequency along the time as well as the possible replacement numbers in each a.a. According to the informational entropy equation, $p(x)$ is the probability of the base {A,T,C,G} in the sequence, in other words, p(A) is the probability of the occurrence of base "A". The frequency of each base is represented by p, and x represents each base. The value of zero for H indicates that all sequences are identical, whereas the maximum value indicates that in this position all bases have the same probability of occurrence.

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{1}$$

Eq. 1 Informational Entropy (Piqueira et al., 2006; Funkhouser, 2012).

To determine which a.a. were involved in the escape process, we used the relative entropy or the *Kullback-Leibler Divergence* (KLD) measure (Eq. (2)), comparing the groups according to the numbers of ART-VF or RM. The relative entropy – KLD is a metric mostly used to quantify the divergence (or gain) of information from a probability of distribution to another, used to detect structural and functional proteins. According to the relative entropy/ KLD formula (Eq. (2)), $\boldsymbol{p}$ is a subsequent distribution, $\boldsymbol{q}$ is an initial distribution and $\boldsymbol{x}$ represents each frequency of base, so p(A) represents the probability of base "A" occurrence on the subsequent sequence and q(A) represents the probability of base "A" occurrence on the initial sequence. KLD calculates the distance between two discrete probabilities, and it can be calculated using the information of the current and the subsequent group. If the Kullback-Leibler divergence is 0 (zero), it indicates that we can expect similar, if not the same, behavior between two different distributions, while a Kullback-Leibler divergence of 1 indicates that the two distributions behave differently. The KLD metric for the two groups of sequences was calculated as follows:

$$D(p\|q) \stackrel{\text{def}}{=} \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \tag{2}$$

Eq. 2 Relative Entropy or *Kullback-Leibler Divergence*– KLD (Gatenby and Frieden, 2007; Bohlin et al., 2012).

All analyses of informational entropy were made in the *MatLab*/ bioinformatics toolbox (2017). For the study, the sliding window technique was used based on a fixed window of size 3 base pairs and a step size equal to one. The entire genome was divided into multiple fragments, with overlapping regions of equal size. The results obtained

in this analysis, are smoothed series, representing a trend of sampling the original data.

To further describe the entropy changes on HIV-1 clade B PR and RT, a.a. were mapped and drawn onto a three-dimensional representation of these enzymes/proteins and colored according to their informational entropy score; Δ and KLD for both situations, ART-VF and RM. For this structural drawing, *Swiss* PDB viewer v.3.7 software was used, and the X-ray coordinates of the Protein Data Bank PDB are available from https://swissmodel.expasy.org/ under the access codes 2hb2.1.A for PR and 4kv8.1.B for RT (Biasini et al., 2014). We chose this model to visualize the three-dimensional structure of PR and RT HIV-1, based on genomic patterns under selective pressure occurred due to the antiretroviral fails or mutations resistance numbers. The three-dimensional structure was made through the alignment and homology of these sequences, and then color-coded (DeLano, 2002) according to the informational entropy score (ENdC et al., 2018; Chen et al., 2006) as follows: low entropy score (0- < 0.2) in gray; intermediate entropy score in lighter colors (entropy score > 0.2 - < 0.4), and high entropy (entropy score > 0.4) in darker coloration. Δ and KLD in each residue within the analyzed patient groups.

For variability and informational metrics the *Shannon Entropy* equation was used, *Relative Entropy was* calculated as *Kullback-Leibler Divergence* (KLD) equation, and the mean difference-hypothesis test was conducted using a pairwise analysis. Statistical relevance was considered when $p < .05$. The metrics, analyses, and graphics were made using *Excel®*, *MATLAB®* and *Graph Pad Prism* (MathWorks, 2017; Motulsky, 2017*)*.

3. Results

3.1. Antiretroviral virologic failure and informational entropy at HIV-1 PR and RT

A relationship between the number of antiretroviral virologic failures - ARV-V.F. and informational entropy in the PR and RT *pol* gene is shown in Fig. 1 (Panel A and B). Lines on the top of the figure indicate statistical significance with $p < .001$. The percentage distribution of sequences was calculated according to the entropy score interval in the PR and RT regions of pol HIV-1 gene (Fig. 2, Panel A and B, i-iv). This histogram serves as a background of the range of amino acids (a.a.) among the assessed groups (i). The accumulated frequencies of the entropy *vis* a *vis* (plotting overlap to better visualize the process), in each group were plotted together with the objective of understanding the variability pattern (ii). The marked area represents conserved indices – C.I.

The frequency of bases with entropy lower than a threshold equal to 0.2 (iii), considered conserved, and the frequencies of bases with entropies above 0.4, considered non-conserved/varied, were also plotted (iv). At zero ARV-V.F. about 80% sequence is conserved (entropy score < 0.20) when ARV-V.F. increase to 7–9 the conservation rate decreases 45% (Fig. iii). As shown in Figure iv, sequences at zero ARV-V.F. showed entropy increase of 5% (score > 0.4) compared to entropy increase of 20% in sequences with higher ARV-V.F. (7–9). In PR, the percentage of sequences with zero a.a. informational entropy in the ART naïve group is ± 50%, compared with ± 2% of patients with higher numbers of ART-VF (7–9) (Fig. 2 – Panel A - ii).

In RT, this frequency/percentage of zero a.a. entropy in naïve ARV-VF sequences, reaches ± 62%, compared with ± 7% of patients with higher numbers of ART-VF (7–9) (Fig. 2 – Panel B – ii). Within groups with up to 1–3 ART-VF, the percentage of sequences with a zero entropy score decreases drastically in both, PR (10%) and RT (7%) HIV-1, and it is more concentrated when entropy interval is between 0.0 and 0.2. In groups with 7–9 ART-VF, the distribution of sequences with increased entropy was higher in the 0.2–0.4 interval for both, PR and RT HIV-1.

Informational Entropy *vis a vis* was calculated in the PR and RT regions according to the number of ART-VF (Fig. 3, Panel A and B). This analysis was performed using a sliding window with 3 base pairs. The difference (Δ) in informational entropy in each residue as well as relative entropy (or *Kullback-Leibler Divergence*-KLD) was analyzed among groups (1–3 *vs* 7–9). The differential entropy (Δ) was seen in PR in six genomic regions related to antiretroviral resistance and escape involving residues in *Fulcrum*, *Active site*, *Flap Elbow* and *Flap*, *Cantilever* and near *Interface*. In RT, it occurred in the *Fingers* region and *Palm* (Attachment-2). All these site were analyzed considering the *Bioafrica-map* genome (Bioafrica, 2019).

The analysis of relative entropy (KLD) has shown that the number of ART-VF affected different genomic regions on the PR whole extension, specifically the a.a. responsible for the dimer stability and *Flap* flexibility. The same circumstance was observed in RT in the regions responsible for the structure and catalytic functions (Attachment-3). In Fig. 3, the star symbols represent the site of escape, based on immunological/hot-spots sites (Motulsky, 2017). Likewise, this was observed at the RT in *Fingers* and *Palm* (Fig. 3, Panel A and B).

In addition, a representation of the three-dimensional structure of PR and RT was drawn and the informational entropy score was color-coded (DeLano, 2002) according to the number of ART-VF (Fig. 4, Panel A and B). We can see an increased color gradation reflecting the increased informational entropy rate, according to the number of ART-VF in both PR and RT, as depicted in Fig. 4. PR presented higher

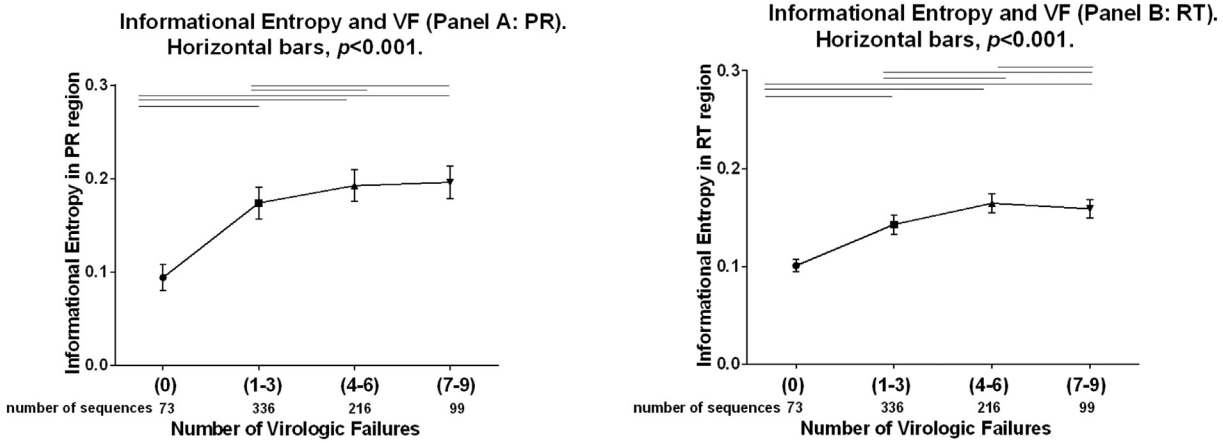

Fig. 1. Informational entropy according to the number of Virologic Failures. The analysis was performed in PR (panel A) and RT (panel B), divided into zero, 1–3, 4–6, and 7–9 VF. Horizontal bars at the top represent significant differences with $p < .001$). *n* at the button of the *x*-axis represents the number of sequences analyzed.

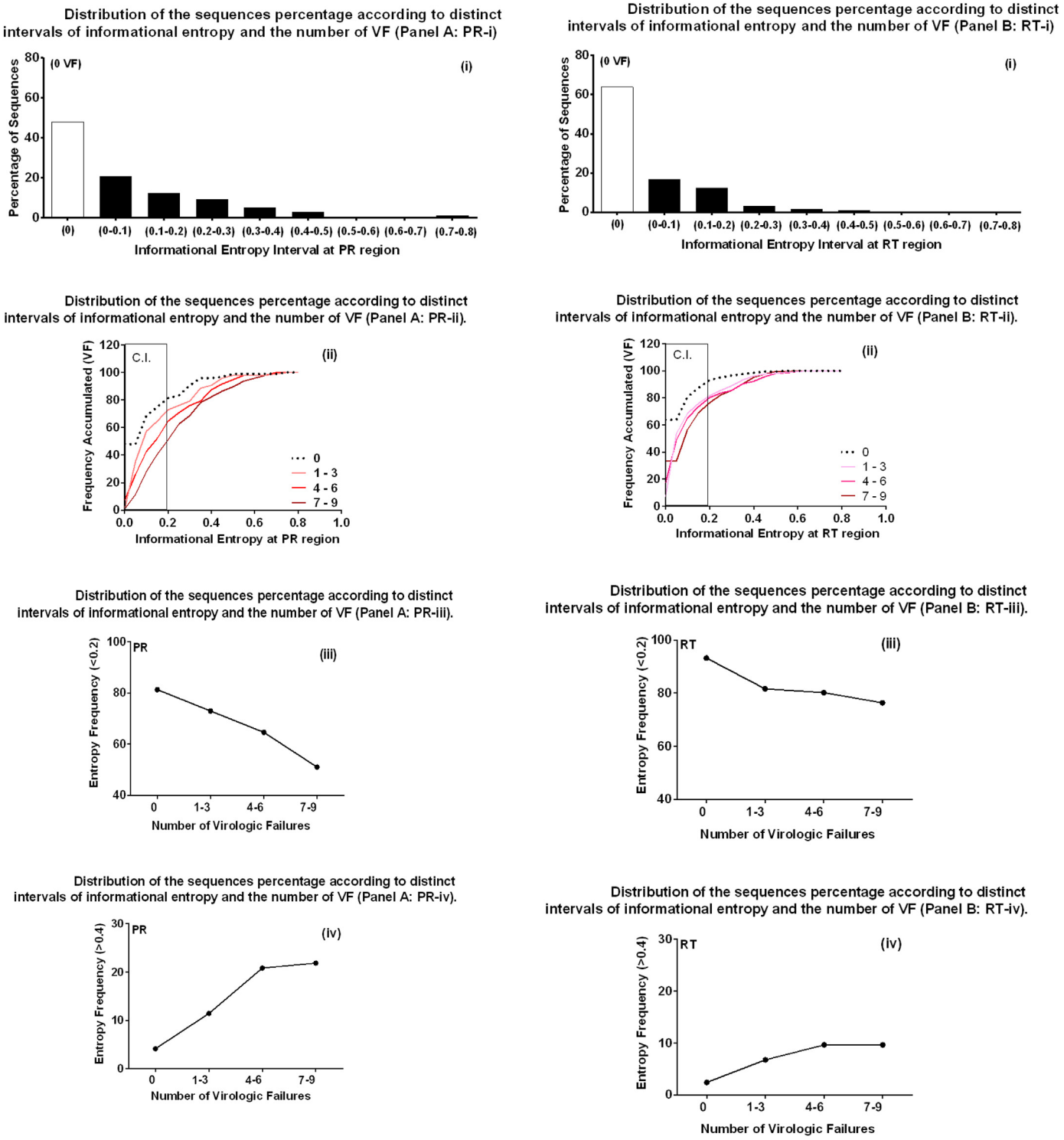

Fig. 2. Distribution of the sequence percentage according to the intervals of informational entropy. The analysis was performed *vis* a *vis* in distinct intervals of entropy according to the number of virologic failures (panels A-B). (i) Distribution in distinct intervals; (ii) Accumulated frequencies of the entropy; (iii) Frequency of bases with an entropy lower than the threshold of 0.2; (iv) Frequencies of bases with entropies above 0.4 (iv).

information in the amino acid related to the flap dynamics (FD), binding site (BS), activity and/or stability (AS) of these proteins. The majority of variable residues with high entropy were displayed in peripheral and central areas that were widely dispersed over the protein structure (Δ in red and KLD in black colors – Fig. 4 - A, B). Additionally, a conserved region (colored in gray) was present along with the structure, representing a stretch necessary for dimer formation and stabilization. KLD in this group, (colored in black) has shown an information gain to PR in *Fulcrum* (L10, L19, K20), *Flap* (M46, I54), and *Cantilever* (A71). In RT, this has occurred in *Fingers* (E44, D67) and *Palm* (L210, T215, and K219) (Fig. 4, Panel B).

3.2. Resistance Mutations (RM) and Informational Entropy at PR and RT of HIV-1

A relationship between the number of RM and informational

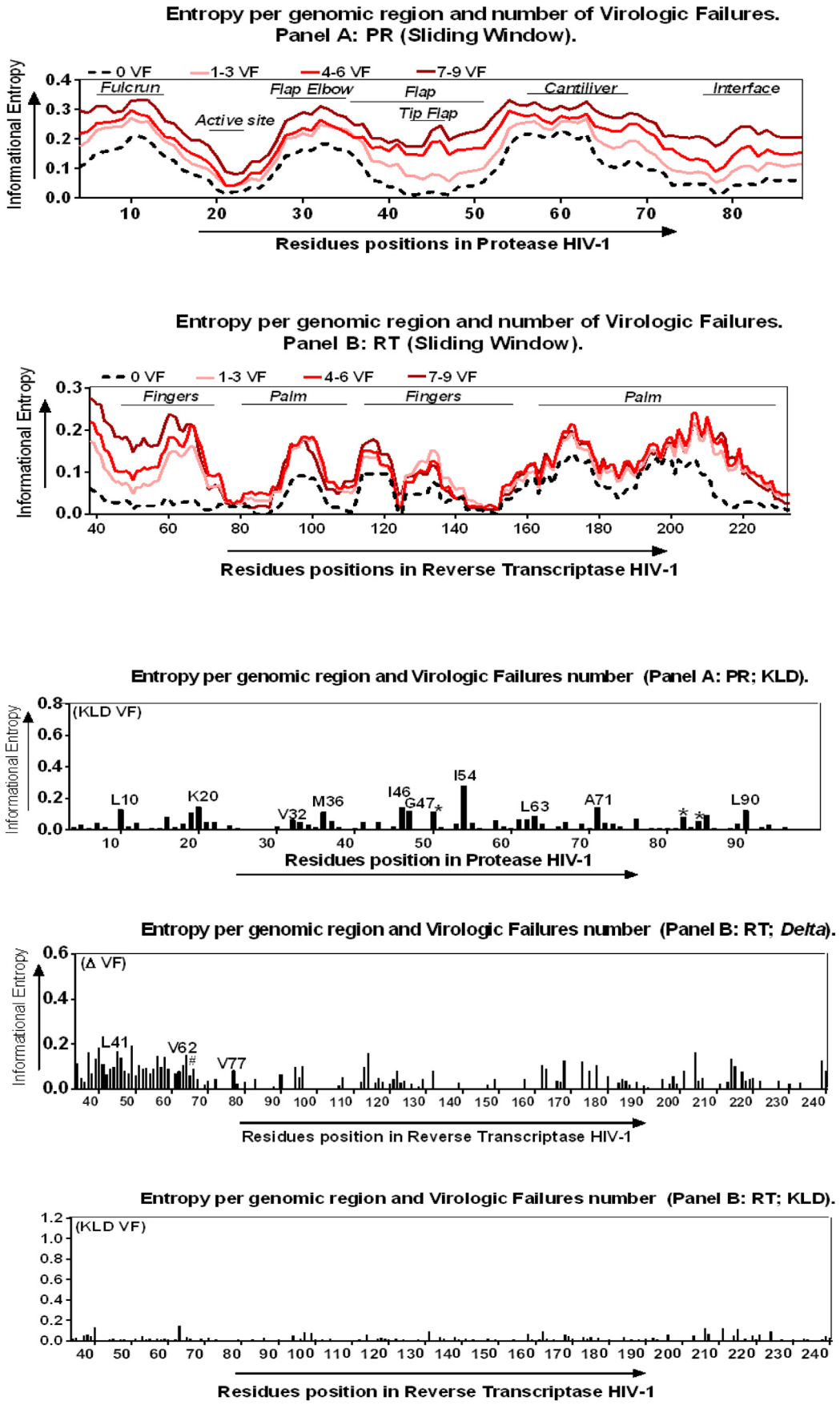

Fig. 3. Entropy per site according to the number of Virologic Failures. In PR (Panel A) and RT (Panel B). An analysis using a sliding window with 3 base pairs was performed. Differences in residues (Δ) were observed in K20, I46, F53, I54, A71, G73 and L90 and information increases (KLD) were found in K20, I46, I54, A71, and L90. Differences (Δ) in RT were observed in L41, V62, and V77, and information increases (KLD) were found in 48, 65, 210 and 220.

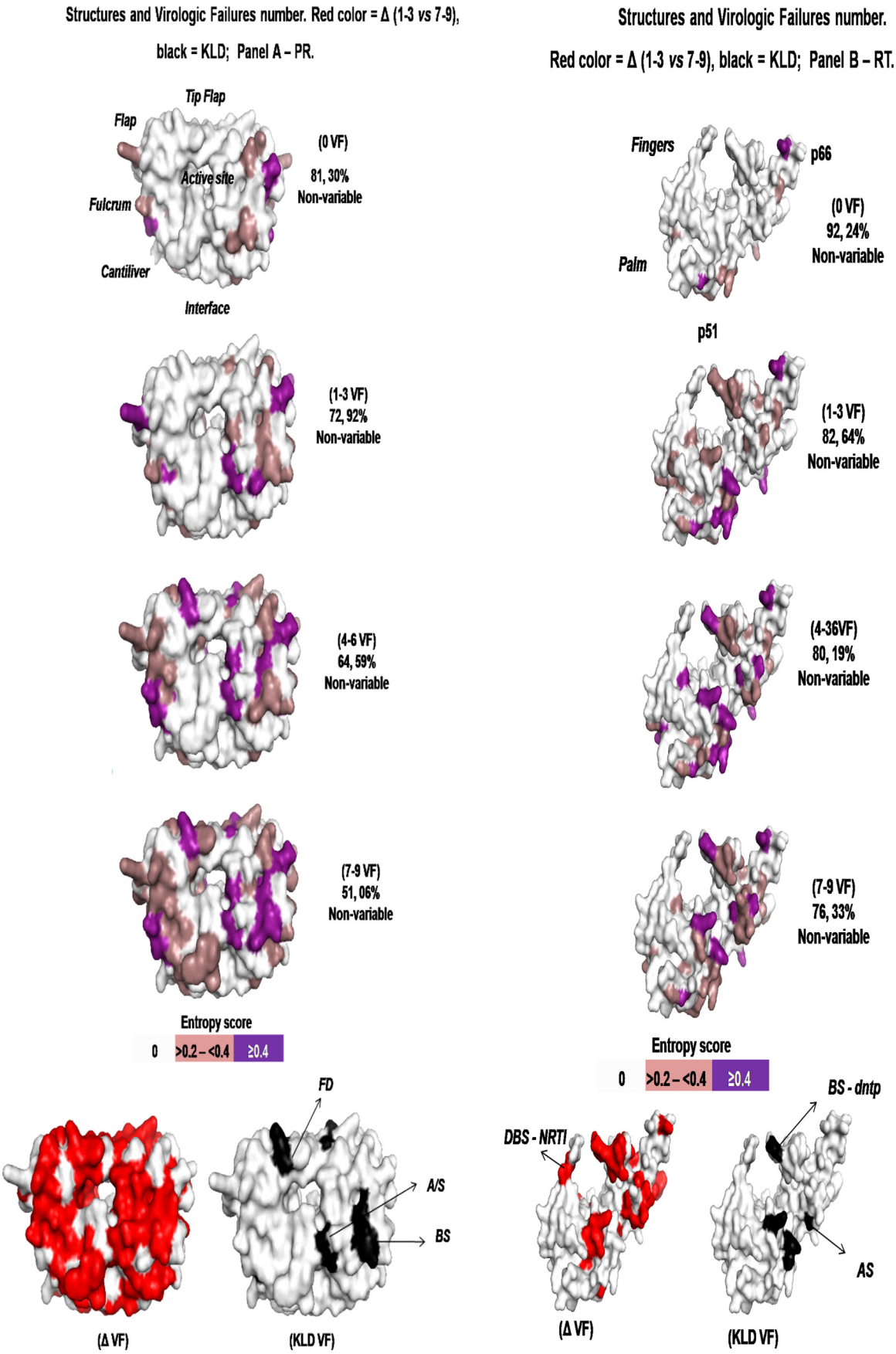

Fig. 4. Structural representation according to the number of ART-VF. The sequences were grouped and color-coded according to the entropy score (0, 1–3, 4–6, and 7–9 VF); of PR (panel A) and RT (panel B). The percentages of non-variable regions are indicated. Red represents the difference (Δ) among groups (1–3 *vs* 7–9) and black represents the KLD among these groups. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

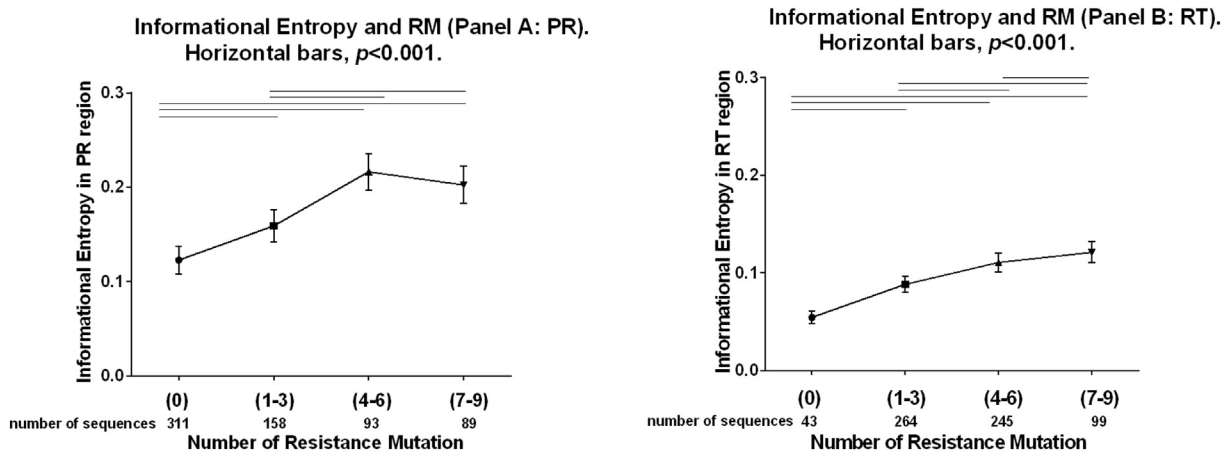

Fig. 5. Informational Entropy accords to the number of RM. The sequences were stratified as zero, 1–3, 4–6, and 7–9 in PR (Panel A) and RT (Panel B). Horizontal bars at the top represent differences with statistical significance at $p < .001$, n at the button of the x-axis represents the number of sequences analyzed.

entropy in the PR and RT *pol* gene is shown Fig. 5 (Panel A and B). Lines on the top of the figure indicate statistical significance with $p < .001$ showed a relationship between the number of RM and the informational entropy in PR and RT of HIV-1. This was evaluated in distinct intervals of informational entropy in the PR and RT regions of the pol gene among individuals with no RM (0), 1–3 RM, 4–6 RM, and 7–9 RM (Fig. 6, Panel A and B – I to iv). The percentage of PR sequences with zero a.a. informational entropy among RM naïve individuals presented less entropy close to 20% (75 zero entropy from 96 a.a.) compared with 2% in individuals with higher numbers of RM (7–9) (50 a.a. zero entropy from 96 a.a.) Fig. 6 Panel A (i).

In addition, RT sequences presented less entropy from drug-naïve individuals (193 conserved residues from 207 a.a.) compared with individuals with 7–9 VF (158 residues with zero entropy from 207 a.a.) Fig. 6 Panel B i to iv. Additionally, the RM profile at PR inhibitors (PI), Nucleoside TR Inhibitors-NRTI and Non-Nucleoside RT Inhibitors-NNRTI presented the attachment-4.

As shown in Fig. 6 Panel A and B i to iv, the groups with 1–3 RM were more concentrated at the entropy interval of 0.0–0.2, whereas the percentage of sequences with zero entropy sharply decreased both in PR and RT HIV-1. In the 7–9 RM group, the distribution of sequences with increased entropy was higher in the interval 0.2–0.4 for both the PR and RT regions (Fig. 6 - ii). The frequency of bases with entropy/considered conserved, was also plotted (iii), as were the frequencies of bases with entropies above 0.4, considered non-conserved/varied (iv).

To determine which a.a. presented higher entropy among the groups, the difference (Δ) in informational entropy in the PR and RT of HIV-1 was quantified, according to the number of RM as well as the relative entropy (KLD) between the groups with 1–3 *vs* 7–9 RM (Fig. 7 Panel A and B). In RT, the high Δ or information gain – KLD occurred in residues responsible for the binding sites and substrate functionality of these proteins/enzymes (Bioafrica, 2019; Mao, 2011) as in the *Fingers* and *Palm* domains, and in functional and structural regions related to higher resistance. In PR residues that presented high values of KLD metrics at the sites related to antigenicity *Fulcrum* near *Active site*, *Flap*, *Cantilever* and *Interface* (Fig. 7, Panel A) (Attachment-5). The star symbol in Fig. 7 represents the binding site, and (#) represents the regions related to escape.

Regarding the conserved residues in PR, five important structures related to protein stability were found and they did not undergo any alterations independent of the RM number, such as *Fulcrum*, near *Active site*, *Flap*, near *Cantilever* and *Interface*. In RT, residues presenting low entropy were found in the functional and structural regions such as *Fingers* and *Palm* (Bioafrica, 2019) (Attachment-6). For the maintenance of RT structure and function, according to previous studies, it is necessary to keep almost two-thirds of its genome to assure structural and functional stabilities, even though some a.a. have unknown functions (Ceccherini-Silberstein et al., 2009; Ceccherini-Silberstein et al., 2005).

Thus, the PR and RT HIV structures were maintained in patients with a low number of R.M. This did not occur among the sequences with an increased RM number presenting a diffuse pattern of increased entropy. Regions with zero entropy (in gray), as presented through the extension of the PR and RT, are essential to the structural stability. Thus, in the three-dimensional analysis of protein structure, we observed that the differences (Δ) in PR informational entropy were seen in the residues located in the binding and enzymatic inhibitors sites Fulcrum, Active site, Flap, Cantilever and near Interface, unlike what was observed in the patients with a lower number of RM (Fig. 8, Panel A and B).

The informational entropy in each a.a. of PR and RT according to the number of RM were structurally represented using *PyMol* software (Biasini et al., 2014), color-coded according entropy score to highlights the affected regions of the binding site and *Flap* dynamics (FD). In RT affected sites by the increase in RM number occurs specially in the binding site of deoxyribonucleotide triphosphate (BS-dNTP) in the active site, and in the drug-binding site of nucleoside RT (DBS-NRTI) as seen in Fig. 8, Panel A and B.

4. Discussion

This study describes the relationship between ART-VF or the number of RM and genomic entropy of HIV-1 in the PR and RT, among patients experiencing ART virologic failure. The number of VF was used because we recognized that as an indirect measure of resistance the determination of RM may not fully indicate the genomic correlates of HIV resistance/scape, as the knowledge of genotypic correlates of antiretroviral resistance may not be fully described or understood. In fact, some clinical trials have demonstrated that the number of virologic failures is a better predictor of the efficacy of the next salvage therapy regimen than the number of RM (Ceccherini-Silberstein et al., 2005).

In this sense, this is the first report to describe a linear and structural Map/Panel of the patterns on the informational entropy of HIV-1 PR and RT among patients with ART-VF, according to different numbers of VF or RM identifying residues under selection at the a.a. level. To obtain this informative Map/Panel, the informational entropy and relative entropy metrics were used (Sato and Ohya, 2011; Parkhomchuk, 2007; Vinga, 2014; Funkhouser, 2012; Gatenby and Frieden, 2007). According to previous studies, the number of ART-VF affects the genomic informational entropy in the PR and RT of HIV-1, in addition to being an important marker of RM levels. There is also a relationship between the numbers of ART-VF, RM, TCD4$^+$ cell counts, and viral loads (Tang and Shafer, 2012). Thus, the number of RM and the ARV therapies in

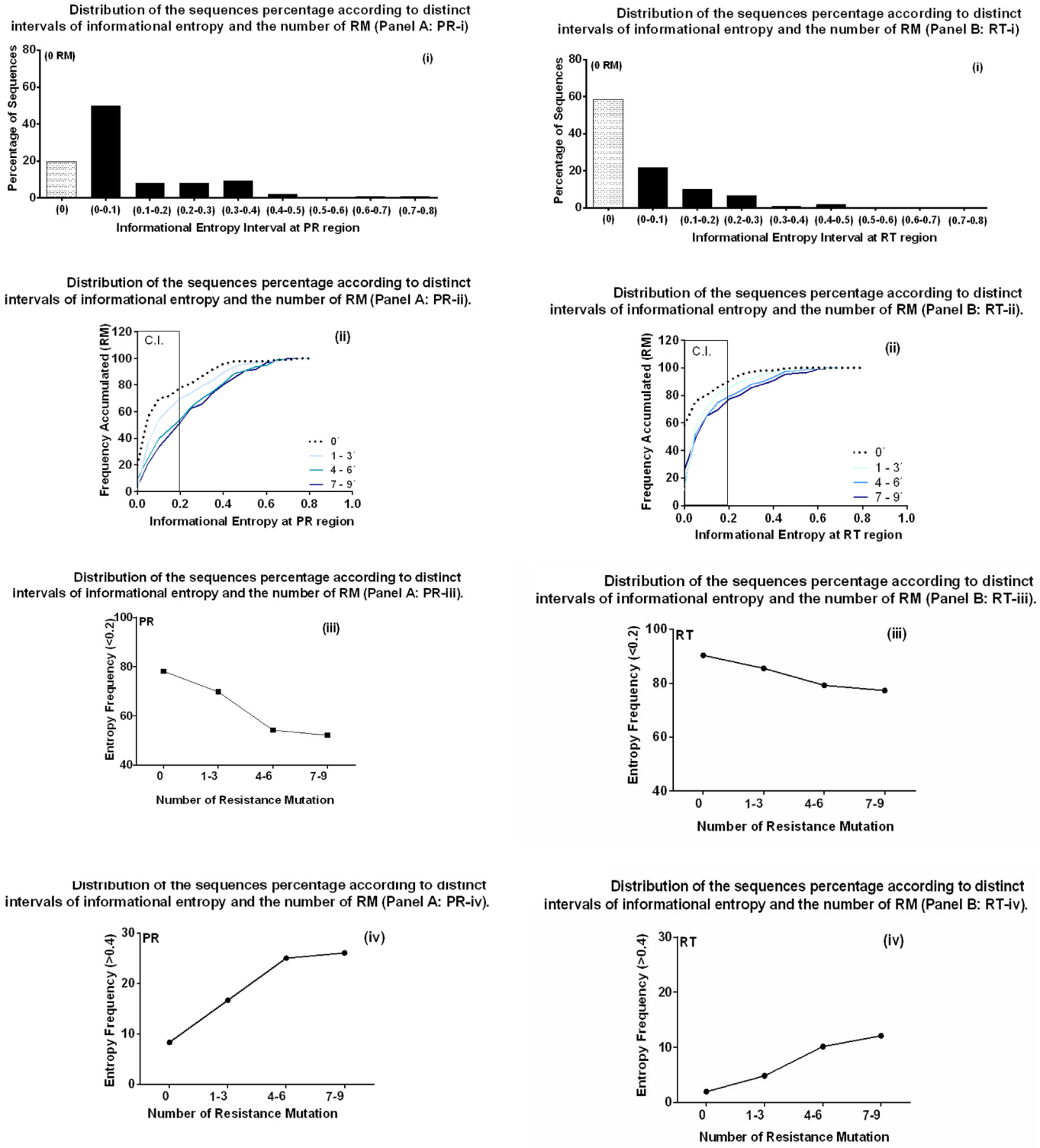

Fig. 6. Distribution of the sequence percentage according to intervals of informational entropy. The analysis was performed *vis* a *vis* in distinct intervals of entropy according to the number of resistance mutation (panels A-B). (i) Distribution in distinct intervals (ii) Accumulated frequencies of the entropy (iii) Frequency of bases with an entropy lower than the threshold of 0.2 (iv) Frequencies of bases with entropies above 0.4 (iv).

the PR and RT of HIV-1 increase the informational entropy and are frequently associated with structural changes and functional instability of these enzymes, leading to changes in HIV replicative capacity according previous study (Potter et al., 2001; Troyer et al., 2009; Tang and Shafer, 2012; Kožíšek et al., 2012; Hu and Kuritzkes, 2014).

Genomic instability (increased entropy) enables modifications/mutations that may be useful for viral escape. Thus, informational entropy is a variability validation method useful to analyze genomic patterns, which may reveal high entropy in certain regions. In addition, in an infected population, entropy pattern results may support ARV health and control systems, delimiting the choice of ARVs, which must be consistent with the monitoring of this population and strength in the adherence pattern, as well as viral load monitoring. This analysis of genomic patterns, as well as the 3D structure of the analyzed proteins,

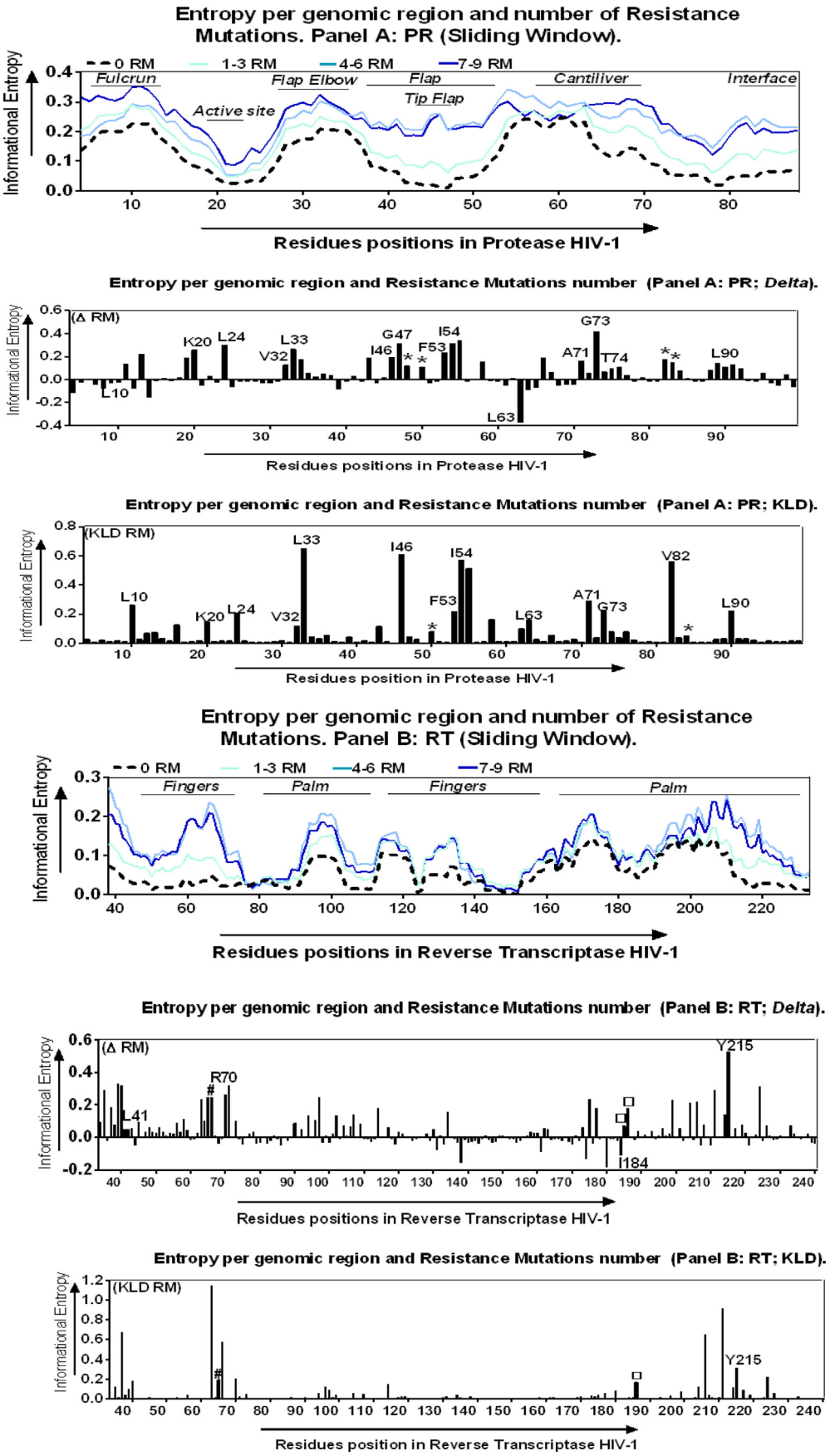

Fig. 7. Entropy per site according to the number of Resistance Mutations. PR (Panel A) and RT (Panel B) are indicated at the top left-hand side of the Y-axis. For PR, differences in residues (Δ) were observed in K20, L24, L33, G47, I53, G73 and L90, and information increase (KLD) were found in L10, L33, I46, I54, A71, V82 N, and L90. For RT, differences (Δ) were observed in L41, 68, R70, I84 and Y215, and information increases (KLD) were found in 40, 68, 69, 190, 210, 215 and 216.

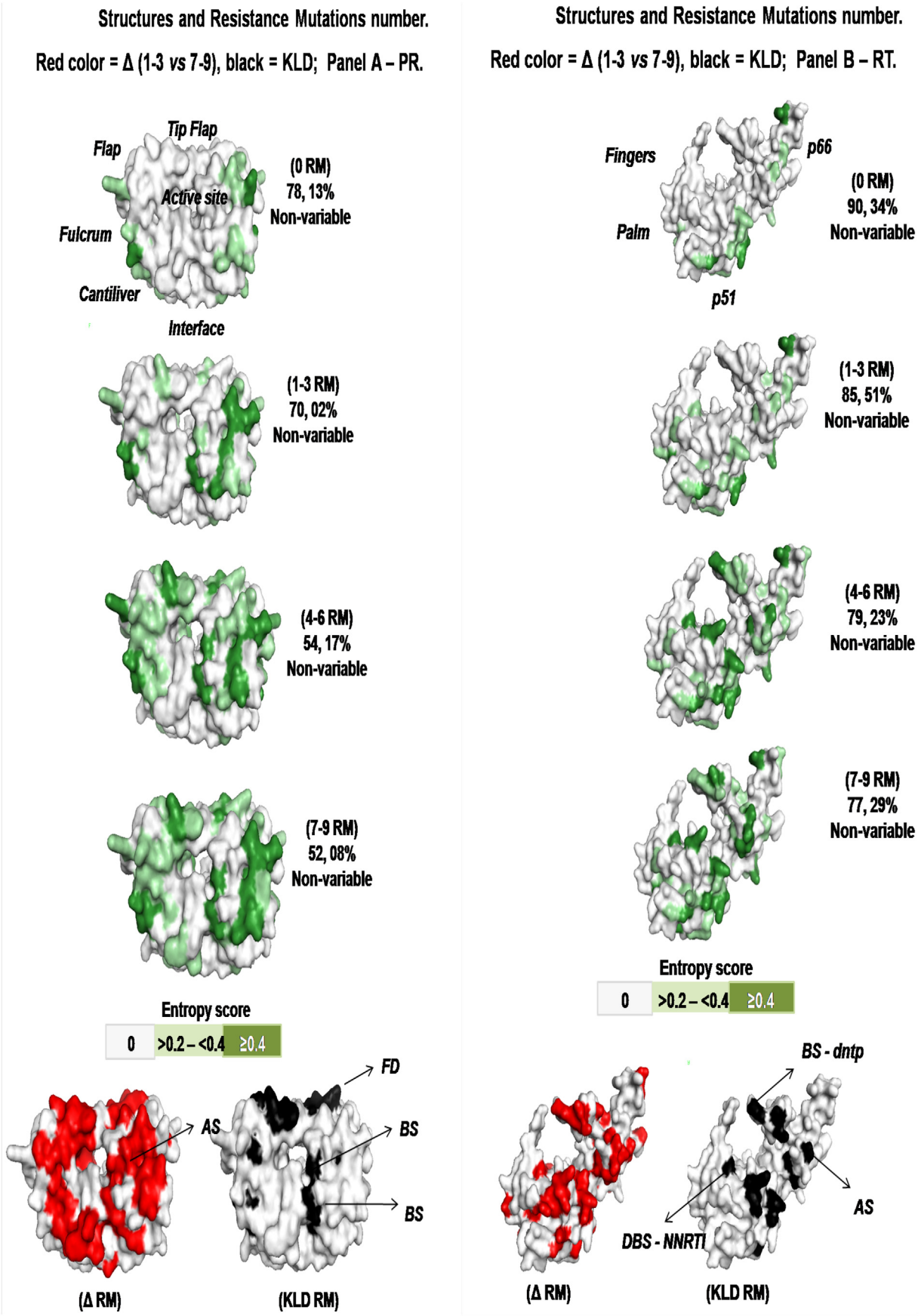

Fig. 8. Structural representation according to the number of resistance mutations. The sequences were grouped and coded according to the entropy score (0, 1–3, 4–6, and 7–9 RM). The percentages of non-variable regions are indicated. Red represents the difference (Δ) between groups (1–3 *vs* 7–9), and black represents the information gain (KLD) between groups. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

can be useful to map the variability in the conformation of these HIV proteins. Pattern quantification and definition of genomic variability may bring insight to formulate different targets and new therapeutic drugs that can evade from antiretroviral resistance.

Conserved regions were identified in the PR structure among patients with lower numbers of ART-VF in crucial regions to dimerization (*Interface*), *Active site*, *Flap* and *Cantilever*. It is clear, but not surprising, that the number of ART-VF changes PR in the region of *Active site*, and this is related to the substrate binding site, leading to an increase in the resistance to antiretrovirals (Residues L10, L19, K20, M46, I54 and A71), whereas in the absence of environmental disturbance (ART-VF or RM), the residue of the cavity binding site (D25-G27) remains totally conserved. The most conserved regions in PR were the *N* and *C*-terminals, important to the functioning and structure of proteins/enzyme (Ishima et al., 2010; Sayer et al., 2008; Ishima et al., 2001; Hornak and Simmerling, 2007; Weber and Agniswamy, 2009). In PR, the conserved residues (common among all analyzed groups independent of RM amount) were found in different regions related to structure, stability of the dimer and substrate recognition (Attachment-6).

The higher the number of ART-VF as well the number of RM, the higher the informational entropy, which interferes with the structure and function activities of PR and RT. Also, higher numbers of ART-VF presented conserved residues bounded in small blocks, suggesting that the participation of consecutive residues in structural domains is necessary to maintain the function and structure of the enzyme (Attachment-6). Thus, it is expected that structural changes in this region due to the emergence of RM will affect the viral escape (Sayer et al., 2008; Ishima et al., 2001; Hornak and Simmerling, 2007; Weber and Agniswamy, 2009).

We understand that the lack of analysis about the specific ARVs used, levels of adherence or duration of each used scheme/drug ARV may preclude many important conclusions. However, this analysis is unique, because it quantifies the patterns of genomic entropy in PR and RT according to the environmental disturbance caused by distinct levels of ART-VF or RM, and it also defines the informational entropy per site in the PR and RT residues, considering the main catalytic regions in these proteins. Although single sequences were used in an experienced ARV population for analysis of genomic variations, the main objective in this study was to quantify the variability patterns in genomic sequence variability that may be occurring in this experienced ARV population. Thus, the genomic variability can be useful to define regions of high HIV variability/escape in these individuals in chronic infection.

5. Conclusions

In conclusion, it was here determined that both the numbers of ART-VF and RM are directly related to a linear increase in informational entropy in PR and RT until a certain level of VF or resistance, demonstrating that selective pressure imposed by an environment with antiretrovirals may drive the emergence of less organized HIV offspring. This current study also depicts a linear and three-dimensional Map/Panel of HIV-1 PR and RT conservation and variability that identify the genomic regions essential to the dimer stability, substrate recognition and, catalytic activity. Therefore, we believe these results can contribute to a deeper understanding about the functional portions and structural regions of the PR and RT of evolving HIV-1 due to antiretroviral resistance.

Ethics approval and consent to participate

This study was approved by the Ethics Committee of the Federal University of São Paulo, under number #0024/11 approved in February 2011.The original obtainment and laboratory analysis of the samples have been described elsewhere (Souza et al., 2011).

Funding

This work was sponsored by Fundação de Amparo a Pesquisa do Estado de Sao Paulo (FAPESP), research grant #2011/12156-0 to R. S. D. and a PhD scholarship from Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) to E. N. d. C. L. The funders had no role in the study design, data collection, and analyses, the decision to publish or the preparation of the manuscript.

Availability of data and material

All data generated or analyzed during this study are included in this published article.

Consent for publication

Not applicable.

Author contributions

ENdCL: Conceptualization, Investigation, Methodology, Validation, Writing - Original Draft, Writing - Review & Editing; **RSAL:** Methodology, Software, Formal analysis; **JRCP**: Conceptualization, Validation, Resources, Writing - Review & Editing, Supervision, Funding acquisition; **MCS:** Investigation; **MC**: Investigation; **JG:** Investigation; **RSD**: Conceptualization, Validation, Resources, Writing - Review & Editing, Funding acquisition. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors have no declarations or conflicts of interest associated with this work.

Acknowledgements

We would like to thank the STI/AIDS and Viral Hepatitis Division of the Brazilian Ministry of Health for providing access to the genomic sequences and patient information from the Brazilian Genotyping Network (RENAGENO).

References

Adami, Christoph, 2004. Information theory in molecular biology. Physics of Life Reviews. https://doi.org/10.1016/j.plrev.2004.01.002.

Adami, C., 2012. The use of information theory in evolutionary biology. Ann N Y AcadSci. https://doi.org/10.1111/j.1749-6632.2011.06422.x.

Ali, Akbar, Bandaranayake, Rajintha M., Cai, Yufeng, King, Nancy M., Kolli, Madhavi, Mittal, Seema, et al., 2010. Molecular basis for drug resistance in HIV-1 PR. Viruses. https://doi.org/10.3390/v2112509.

Arhel, N., Kirchhoff, F., 2010. Host proteins involved in HIV infection: new therapeutic targets. Biochim. Biophys. Acta. https://doi.org/10.1016/j.bbadis.12.003.

Asahchop, E.L., Wainberg, M.A., Sloan, R.D., Tremblay, C.L., 2012. Antiviral drug resistance and the need for development of new HIV-1 RT inhibitors. Antimicrob. Agents Chemother. 56 (10), 5000–5050.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., et al., 2014. SWISS-MODEL: modeling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Research. https://doi.org/10.1093/nar/gku340.

Bioafrica Bioinformatics & Genomics For Health and Life Sciences in Africa, 2019. Available from. http://www.bioafrica.net/proteomics/POL-PRprot.html;. http://www.bioafrica.net/proteomics/POL-RTprot.html Accessed 20 May 2017b.

Bohlin, J., van Passel, M.W.J., Snipen, L., Kristoffersen, A.B., Ussery, D., Hardy, S.P., 2012. Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. BMC Genomics. https://doi.org/10.1186/1471-2164-13-66.

Ceccherini-Silberstein, F., Gago, F., Santoro, M., Gori, C., Svicher, V., Rodríguez-Barrios, F., d'Arrigo, R., et al., 2005. High sequence conservation of human immunodeficiency virus type 1 RT under drug pressure despite the continuous appearance of mutations. J Virol. https://doi.org/10.1128/JVI.79.16.10718-10729.2005.

Ceccherini-Silberstein, F., Malet, I., D'Arrigo, R., Antinori, A., Marcelin, A.G., Perno, C.F., 2009. Characterization and structural analysis of HIV-1 integrase conservation. AIDS Rev. 11 (1), 17–29.

Chen, G.-W., Chang, S.-C., Mok, C.-K., Lo, Y.-L., Kung, Y.-N., Huang, J.-H., September 2006. Et al. genomic signatures of human versus avian influenza a viruses. Emerg. Infect. Dis. 12 (9).

Dam, Elisabeth, Quercia, Romina, Glass, Bärbel, Descamps, Diane, Launay, Odile, Duval, Xavier, et al., 2009. Gag mutations strongly contribute to HIV-1 resistance to PR inhibitors in highly drug-experienced patients besides compensating for fitness loss. PLoS Pathog. https://doi.org/10.1371/journal.ppat.1000345.

Das, K., Arnold, E., 2013. HIV-1 RT and antiviral drug resistance (part 1 of 2). Curr Opin Virol. https://doi.org/10.1016/j.coviro.2013.03.012.

Das, Atze T., Berkhout, Ben, 2010. HIV-1 evolution: frustrating therapies, but disclosing molecular mechanisms. Phil. Trans. R. Soc. B. https://doi.org/10.1098/rstb.0072.

DeLano, W.L., 2002. Pymol: An open-source molecular graphics tool. CCP4 Newsletter On Protein Crystallography 40, 82–92.

d'Ettorre, G., Forcina, G., Ceccarelli, G., Andreotti, M., Andreoni, C., Rizza, C., et al., 2011. Adherence and genotypic drug resistance mutations in HIV-1-infected patients failing current antiretroviral therapy. Journal of Chemotherapy. https://doi.org/10.1179/joc.2011.23.1.24.

Durston, K.K., Chiu, D.K.Y., Abel, D.L., Trevors, J.T., 2007. Measuring the functional sequence complexity of proteins. Theoretical Biology and Medical Modelling. https://doi.org/10.1186/1742-4682-4-47.

ENdC, L., JRC, P., Camargo, M., Galinskas, J., Sucupira, M.C., Diaz, R.S., 2018. Impact of antiretroviral resistance and virologic failure on HIV-1 informational entropy. J Antimicrob Chemother. https://doi.org/10.1093/jac/dkx508.

Funkhouser, S., 2012. The entropy of a discrete real variable. Entropy 14, 1522–1538.

Gatenby, R.A., Frieden, B.R., 2007. Information theory in living systems, methods, applications, and challenges. Bull. Math. Biol. 69, 635–657.

Hornak, V., Simmerling, C., 2007. Targeting structural flexibility in HIV-1 PR inhibitor binding. Drug Discov. Today. https://doi.org/10.1016/j.drudis.12.011.

Hu, Wei-Shau, Hughes, Stephen H., 2012. HIV-1 reverse transcription. Cold Spring Harb Perspect Med. https://doi.org/10.1101/cshperspect.a006882.

Hu, Zixin, Kuritzkes, Daniel R., 2014. Altered viral fitness and drug susceptibility in HIV-1 carrying mutations that confer resistance to nonnucleoside RT and integrase strand transfer inhibitors. J Virol. https://doi.org/10.1128/JVI.00695-14.

Imamichi, T., 2004. Action of anti-HIV drugs and resistance: RT inhibitors and PR inhibitors. Curr. Pharm. Des. 10 (32), 4039–4053.

Ishima, R., Ghirlando, R., Toszer, J., Gronenborn‖, A.M., Torchia, D.A., Louis, J.M., et al., 2001. The journal of biological chemistry 52 (28), 49110–49116 276.

Ishima, R., Gong, Q., Tie, Y., Weber, I.T., Louis, J.M., 2010. Highly conserved glycine 86 and arginine 87 residues contribute differently to the structure and activity of the mature HIV-1 PR. Proteins. https://doi.org/10.1002/prot.22625.

Kožíšek, Milan, Henke, Sandra, Šašková, Klára Grantz, Jacobs, Graeme Brendon, Schuch, Anita, Buchholz, Bernd, et al., 2012. Mutations in HIV-1 *gag* and *pol* compensate for the loss of viral fitness caused by a highly mutated PR. Antimicrob Agents Chemother. https://doi.org/10.1128/AAC.00465-12.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Mao, Y., 2011. Dynamical basis for drug resistance of HIV-1 PR. BMC Struct. Biol. https://doi.org/10.1186/1472-6807-11-31.

MathWorks, Bioinformatics ToolboxTM: User's Guide, 2017. https://www.mathworks.com/help/pdf_doc/bioinfo/bioinfo_ug.pdf.

Motulsky, H.J. GraphPad Statistics Guide. http://www.graphpad.com/guides/prism/7/statistics/index.htm.

Parkhomchuk, Dmitri V., 2007. Information Theory of Genomes. arXiv:q-bio/0612038 [q-bio.GN], available from. https://arxiv.org/abs/q-bio/0612038.

Piqueira, J.R.C., Serboncini, F.A., Monteiro, L.H.A., 2006. Biological models: measuring variability with classical and quantum information. Journal of Theoretical Biology. https://doi.org/10.1016/j.jtbi.2006.02.019.

Potter, S.J., Chew, C.B., Steain, M., Dwyer, D.E., Saksena, N.K., 2001. Obstacles to successful antiretroviral treatment of HIV-1 infection: problems & perspectives. Indian J. Med. Res. 19, 217–237.

Rhee, S.Y., Gonzales, M.J., Kantor, R., Betts, B.J., Ravela, J., Shafer, R.W., et al., 2003. Nucleic Acids Research. https://doi.org/10.1093/nar/gkg100.

Sato, Keiko, Ohya, Masanori, 2011. Evolution of HIV-1 from the viewpoint of information theory. IEEE Xplore. https://doi.org/10.1109/ISABEL.2010.5702806.

Sayer, J.M., Liu, F., Rieko Ishima, I.T., Weber, J.M., 2008. Louis. Effect of the active site D25N mutation on the structure, stability, and ligand binding of the mature HIV-1 PR. J. Biol. Chem. 283 (19), 13459–13470.

Schultz, A.-K., Bulla, I., Abdou-Chekaraou, M., Gordien, E., Morgenstern, B., Zoulim, F., Dény, P., Stanke, M., 2012. jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. Nucleic Acids Res. 40, W193–W198.

Shafer, R.W., Schapiro, J.M., 2008. HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. AIDS Rev. 10 (2), 67–84.

Singh, K., Marchand, B., Kirby, K.A., Michailidis, E., Sarafianos, S.G., 2010. Structural aspects of drug resistance and inhibition of HIV-1 RT. Viruses. https://doi.org/10.3390/v2020606.

Smith, R.J., 2006. Adherence to antiretroviral HIV drugs: how many doses can you miss before resistance emerges? Proc. R. Soc. B. https://doi.org/10.1098/rspb.2005.3352.

Song, H., Pavlicek, J.W., Cai, F., Bhattacharya, T., Li, H., Iyer, S.S., et al., 2012. Impact of immune escape mutations on HIV-1 fitness in the context of the cognate transmitted/founder genome. Retrovirology 9, 89.

Souza, D.C.F., Sucupira, M.C.A., Brindeiro, R.M., Fernandez, J.C.C., Sabino, E.C., Inocencio, L.A., Diaz, R.S., 2011. The Brazilian network for HIV-1 genotyping external quality control assurance programme. J. Int. AIDS Soc. 14, 45.

Struck, D., Lawyer, G., Ternes, A.M., Schmit, J.C., Bercoff, D.P., 2014. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res. 42, e144.

Sunshine, Justine E., Larsen, Brendan B., Maust, Brandon, Casey, Ellie, Deng, Wenje, Chen, Lennie, et al., 2015. Fitness-balanced escape determines resolution of dynamic founder virus escape processes in HIV-1 infection. Journal of Virology 89 (20).

Tang, M.W., Shafer, R.W., 2012. HIV-1 antiretroviral resistance scientific principles and clinical applications. Drugs 72 (9) e1-e25 0012–6667/12/0009–0001.

Schneider, Thomas D., 2010. A brief review of molecular information theory. Nano Commun Netw. https://doi.org/10.1016/j.nancom.2010.09.002.

Troyer, R.M., McNevin, J., Liu, Y., Zhang, S.C., Krizan, R.W., Abraha, A., et al., 2009. Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. PLoS Pathog. https://doi.org/10.1371/journal.ppat.1000365.

Vinga, Susana, 2014. Information theory applications for biological sequence analysis. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbt068.

Weber, I.T., Agniswamy, J., 2009. HIV-1 PR: structural perspectives on drug resistance. Viruses. https://doi.org/10.3390/v1031110.

Wensing, A.M., Calvez, V., Günthard, H.F., Johnson, V.A., Paredes, R., Pillay, D., et al., 2017. Update of the drug resistance mutations in HIV-1. Topics Ant. Med. 24 (Suppl. 4), 132–141.