Insights into the Dynamical Evolution of the HIV genome

Santa Fe Institute Complex Systems Summer School 2008 - Bariloche - Argentina

Carlo Altamirano, Guillermo Espinoza, Peter Klimek, Tomas Perez-Acle, Miguel Ponce de Leon, Alejandro Rozenfeld

INTRODUCTION AND MOTIVATION

Concepts and methods originating from the study of dynamical systems have exhibited an ever-growing scope, crossing boundaries from physics to biology, economy and social behavior. In a pioneering attempt Manfred Eigen (Eigen and Schuster, 1979) showed how to impose evolutionary dynamics onto a space of sequences. Pursuing his ideas, we attempt to identify dynamical patterns within mutations of genomic sequences with the ultimate goal to understand how selective constraints on the phenotype are reflected in the space of potential genotypes and thus favoring 'trajectories of evolution'.

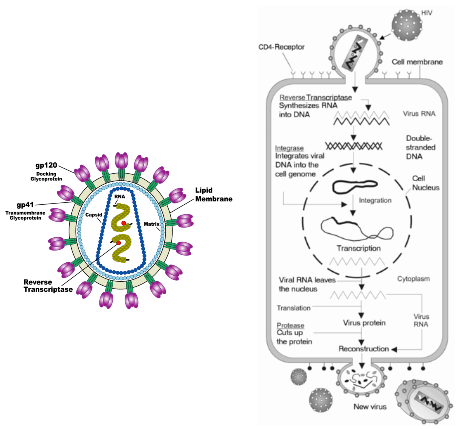

Figure 1. Molecular structure of the human inmunodeficiency virus type 1 (left) and replicative cycle into the host cells (right). Mutant HIV variants appears by the selective pressure produced by antiviral drugs mainly targeted to the protease and the reverse transcriptase.

MATERIALS

- 539 HIV genome sequences, corresponding to virus isolates from around the world, were downloaded from the HIV database at LANL.
- A phylogenetic tree developed with the Maximum Likelihood method, including 163 variants of the hyper mutant HIV-1 reverse transcriptase, was obtained from the HIV database at LANL (see Figure 2).

INFORMATION THEORY METHODS

- In order to test the predictive power of the Information Theory methods, the complete dataset was analyzed calculating the block entropy (Eq. 1). The Log was taken to a base of 4.
- The source entropy was calculated in accordance to (Eq. 2). By taking the second derivative of *H(L)* we obtained the predictability gain in accordance to (Eq. 3), in units or *quarts* per symbol. The excess of entropy was calculated in accordance to (Eq. 4). The mutual information between two sequences was calculated in accordance to (Eq. 5).

$$H(L) = -\sum_{s^L \in \mathbf{A}^L} \Pr(s^L) \log_4 \Pr(s^L) \quad , \tag{1}$$

$$h_\mu = \lim_{L\to\infty} \frac{H(L)}{L} = \lim_{L\to\infty} \Delta H(L) \quad , \tag{2}$$

$$\Delta^2 H(L) = h_\mu(L) - h_\mu(L-1) \tag{3}$$

$$E = \sum_{L=2}^{\infty} (L-1)\Delta^2 H(L) \quad . \tag{4}$$

$$E(i,j) = \lim_{L\to\infty} I(s_i, s_j) \quad . \tag{5}$$

where $I[X;Y] \equiv H[X] - H[X|Y]$

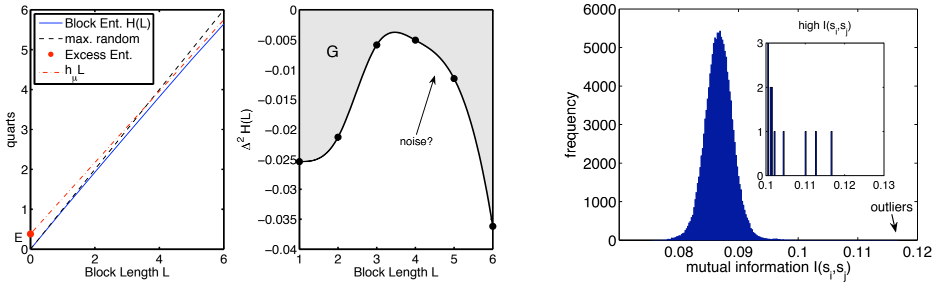

Figure 4. Left. H(L) and its first over L. The slashed line shows the upper bound given by a maximal random configuration of letters. We further show the excess entropy. Middle, the second derivative, the predictability gain, over L. We found a peak at L = 3; 4. Beyond this size we expect our data to suffer from a too sparse dataset. The shaded area denotes the total predictability. Right, histogram of the excess entropies or mutual informations between each pair of sequences. The positive outliers deserve a closer inspection.

DELAY-COORDINATE EMBEDDING

- Given a time series from a sensor or a single state variable $x_i(t)$ in an *n*-dimensional dynamical system. delay-coordinate embedding lets one reconstruct a useful version of the internal dynamics of that system. The Taken’s theorem guarantee that the reconstructed dynamics is topologically identical to the true dynamics of the system.
- To apply the delay coordinate-embedding method, the following equation was applied taking into account the temporal order of the HIV sequences provided by the phylogenetic reconstruction shown in figure 2.

$$\vec{r}(t) = [x_i(t),\ x_i(t-\tau),\ x_i(t-2\tau),\ \ldots\ ,\ x_i(t-(m-1)\tau)]$$

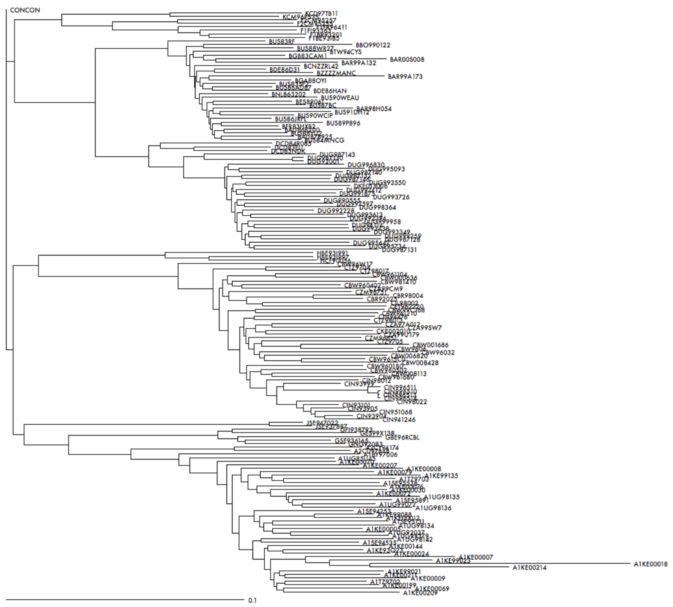

Figure 2. Phylogenetic reconstruction of HIV type I isolates. The phylogenetic tree was obtained from the HIV Database at Los Alamos National Laboratory (USA). (http://hiv.lanl.gov) (parameters; order: TCAG / freq: 0.230787 0.197985 0.378874 0.192354 / ratepar: 0.865745 0.123350 0.186650 0.274382 0.138878)

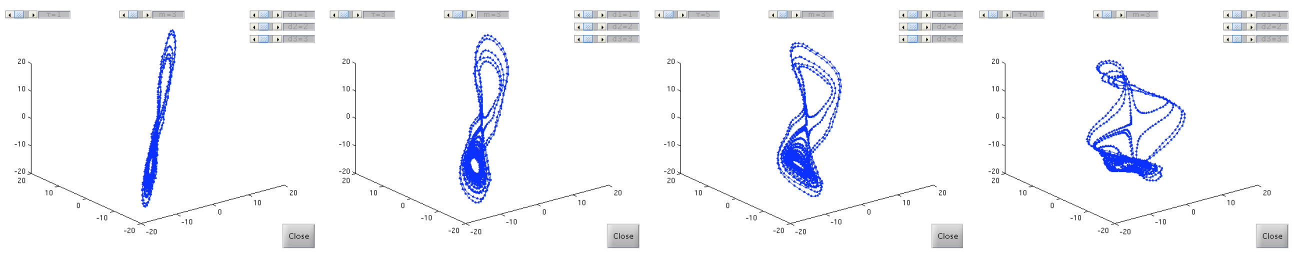

Figure 3. Test of the applied method. Results of the time-embedded methodology applied to the reconstruction of the Lorenz attractor for different time delays (from left to right).

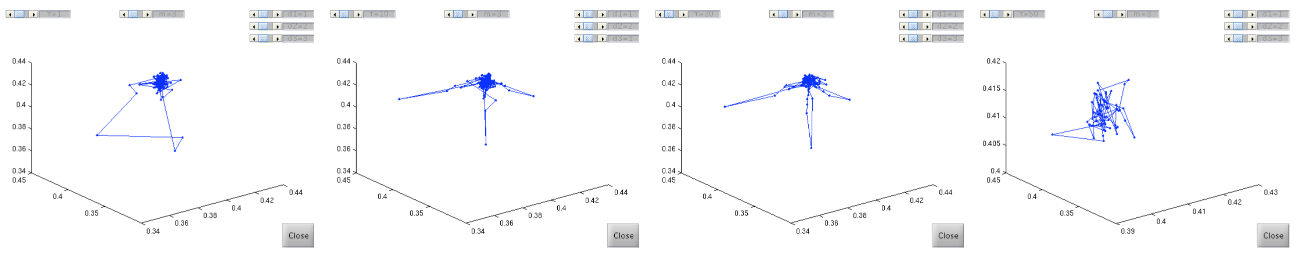

Figure 6. Time-embedded analyses applied to the HIV RT sequences arranged in figure 2.

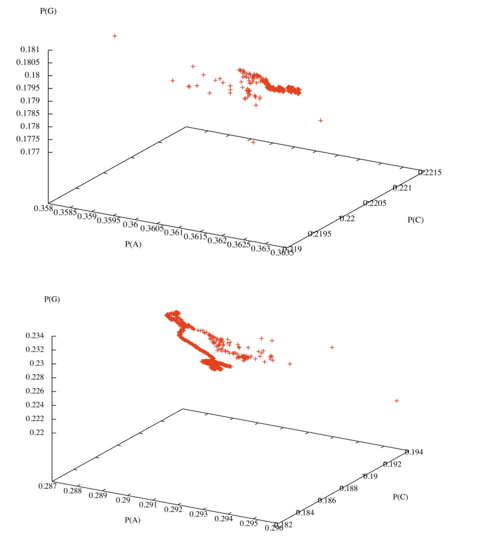

Figure 5. HIV genome sequences plotted in the probability space. Top, 539 HIV genome obtained from the HIV genome database at LANL. Bottom, hypermutant HIV RT sequences.

Table 1. (on the right) Fractal dimension for both datasets used in this work, calculated in accordance to the box-counting method. HIV* represents the 539 HIV complete genomes. HIV** represents the 163 sequences obtained from the phylogenetic analysis depicted in figure 2. HIV-RT gene represents the dataset of the hypermutant reverse transcriptase dataset.

DNA Sequence	Fractal Dimension
HIV*	0.864
HIV**	1.038
HIV-RT gene	0.967