31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Use of Average Mutual Information for studying changes in HIV populations

Khalid Sayood, Federico Hoffman, and Charles Wood

Abstract— **Average mutual information (AMI) has been used in a number of applications in bioinformatics. In this paper we present its use to study genetic changes in populations; in particular populations of HIV viruses. Disease progression of HIV-1 infection in infants can be rapid resulting in death within the the first year, or slow, allowing the infant to survive beyond the first year. We study the development of rapid and slow progressing HIV population using *AMI charts* based on average mutual information among amino acids in the *env* gene from a population of 1142 clones derived from seven infants with slow progressing HIV-1 infection and four infants with rapidly progressing HIV-1 infection. The AMI charts indicate the relative homogeneity of the rapid progressor populations and the much greater heterogeneity of the slow progressor population, especially in later samples. The charts also show the distinct regions of covariation between residues without the need for aligning the sequences. By examining the changes in AMI between populations we can distinguish between clones obtained from rapid progressor and slow progressor. A measure of this change can be used to enhance prediction of disease progression.**

I. INTRODUCTION

Average mutual information (AMI) $I(X;Y)$ defined as

$$I(X;Y) = \sum_{X,Y \in \mathscr{A}} P(X,Y) \log \left(\frac{P(X,Y)}{P(X)P(Y)} \right) \qquad (1)$$

is a measure of information contained in the random variable Y about the random variable X, where $\mathscr{A}$ is the alphabet from which X and Y take their values. From the definition it is clear that the average mutual information is a symmetric measure, that is, $I(Y;X) = I(Y;X)$. Developed by Shannon [1] for the analysis of communication systems, it has been used in a variety of applications in the biological fields. It has been used to examine covariation of different sites in the V3 loop of the HIV genome [2], to investigate correlations between sites in protein sequences[3], [4], [5], and to differentiate between coding and noncoding regions[6], to investigate long range correlations [7], to develop species signatures [8], for fragment assembly [9] to study coevolving sites in polypeptide sequences [2], [10], for secondary structure prediction [11], [12], and to study relationships between genes and their phenotypes [13].

The application of AMI to DNA and protein sequences have generally used used one of two formulations. In the first the random variables X and Y are taken to be nucleotides which are at some distance, or lag, k apart. The AMI thus becomes a function of the distance between the nucleotides. In other words

$$I(X;Y) = I(k)$$

This approach has been used for investigating long range relationships in DNA sequences and to develop species signatures or otherwise characterize DNA sequences [3], [4], [5], [6], [8], [9]. In this approach $\mathscr{A} = \{A,G,T,C\}$ and the probabilities needed for computing the AMI are estimated using a wide sense stationarity assumption from a single sequence. Hence the approach is generally employed for long sequences.

The second approach is generally employed where the length of the sequence is small but there are multiple sequences available as in the case of studying the co-variance of residues in a protein sequence. In such applications sequences from different clones are considered to be different realizations of a discrete valued random process. The k^{th} and m^{th} residues of the proteins can be thought of as samples of the random process at "times" k and m and can be viewed as random variables X and Y, and

$$I(X;Y) = I(k,m)$$

This means a sequence of length N can be characterized by N^2 values. In the current work we have organized these N^2 values in the form of an $N \times N$ matrix whose $(i,j)^{th}$ element is I(i,j) as shown in Table I.

$I(1,1)$	$I(1,2)$	$I(1,3)$	$\dots$	$I(1,N)$
$I(2,1)$	$I(2,2)$	$I(2,3)$	$\dots$	$I(2,N)$
$I(3,1)$	$I(3,2)$	$I(3,3)$	$\dots$	$I(3,N)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$I(N,1)$	$I(N,2)$	$I(N,3)$	$\dots$	$I(N,N)$

TABLE I

DISPLAYING THE AVERAGE MUTUAL INFORMATION

We refer to these $N \times N$ matrices as *AMI charts* and display them as grayscale images with lower values in black and higher values in white as shown in Figure 1. This figure depicts the AMI chart for a collection of *env* proteins obtained from infant 1690 (details below). The number of clones used to generate the chart was 128.

This work was supported in part by a grant from the National Institutes of Health K25AI068151

K. Sayood is with the Department of Electrical Engineering, University of Nebraska, Lincoln NE 68588-0511 ksayood1@unl.edu

F. Hoffman is with the Instituto Carlos Chagas, ICC-Fiocruz, Curitiba, Parana, Brazil federico.g.hoffmann@gmail.com

C. Wood is with Nebraska Center for Virology, School of Biological Sciences, University of Nebraska, Lincoln, NE 68583-0900 cwood@unlnotes.unl.edu

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

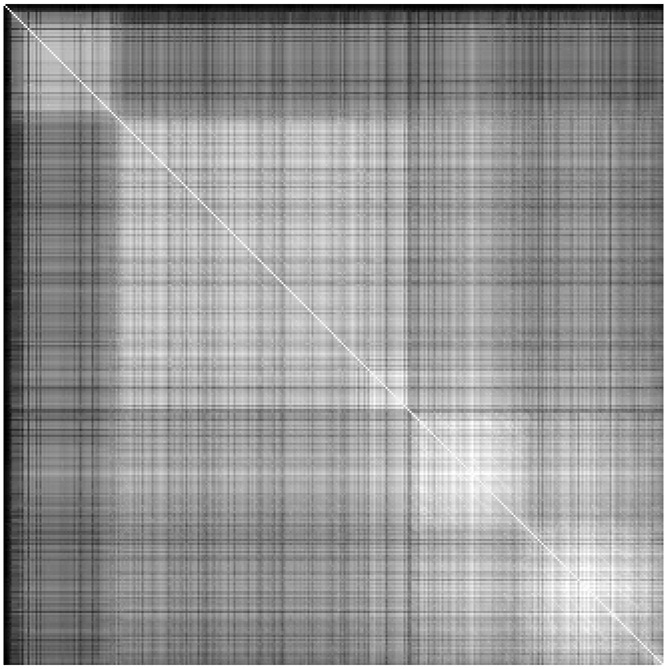

Fig. 1. AMI values for 1690 sequence

Larger values of $I(X;Y)$ can be viewed as indicating greater dependence between the amino acids in the positions k and m. However, lower values of $I(X;Y)$ do not necessarily mean lower dependence. To see why this is so we write the average mutual information in terms of entropy and conditional entropy

$$I(X;Y) = H(X) - H(X|Y)$$

Here $H(X)$ is a measure of information contained in X, or equivalently, the amount of uncertainty about X. $H(X|Y)$ can be seen as the uncertainty remaining about X after Y is known. Thus the difference is the amount of information contained in Y about X. If Y is unrelated to X then the uncertainty remaining about X after Y is known will be the same as the uncertainty about X prior to Y being known. In other words $H(X) = H(X|Y)$ and $I(X;Y)$ is zero. However, if there was no uncertainty about X to start with, even if there was a heavy dependence of X on Y, $I(X;Y)$ would still be zero. To take into account this possibility studies of co-evolving residues generally use a normalized version of the average mutual information [10]. In this work we have avoided any normalization as it tends to obscure developments in time which was the main focus of our study.

One of the difficulties with the use of functions of probability estimates is the lack of sufficient data for obtaining reliable estimates of the probabilities. In our particular application the results seem to indicate that for our purposes the amount of data available for probability estimates was sufficient.

II. Data

The data used in this study were obtained from HIV-1 populations isolated from the venous blood of eleven infants born to mothers infected with the HIV-1 Type C virus. The infants were all breast fed. Of the eleven, four infants (1449, 2669, 2873, and 2617) died within the first year due to HIV related complications. Seven infants (1984, 1084, 1690, 1157, 2660, 2953, and 834) remained asymptomatic four years after birth. We will call the first group rapid progressors to indicate the rapid progression of the disease, and the second group slow progressors. Blood was drawn from the infants at irregular time intervals ranging from two months in the first year, to one year for slow progressors in their fourth year of life. More details about these samples can be found in [14], [15], [16].

All the data for the rapid growers was combined to generate the AMI charts shown in Figure 2. Each chart is labeled with the identifying number of the infant from whom the data was collected. The AMI charts for four of the seven slow growers is shown in Figure 3. While there is variation between the charts, the AMI charts for the rapid progressors are generally darker indicating low AMI values. The reason for this was a lack of variability between the clones indicating perhaps that the population had achieved an optimum configuration for infecting the patient and was therefore not changing very rapidly. The AMI charts for the slow progressors were generally lighter indicating higher AMI values. Furthermore there was a checkerboard pattern which indicated that certain regions of the *env* protein were changing more rapidly than other regions of the protein.

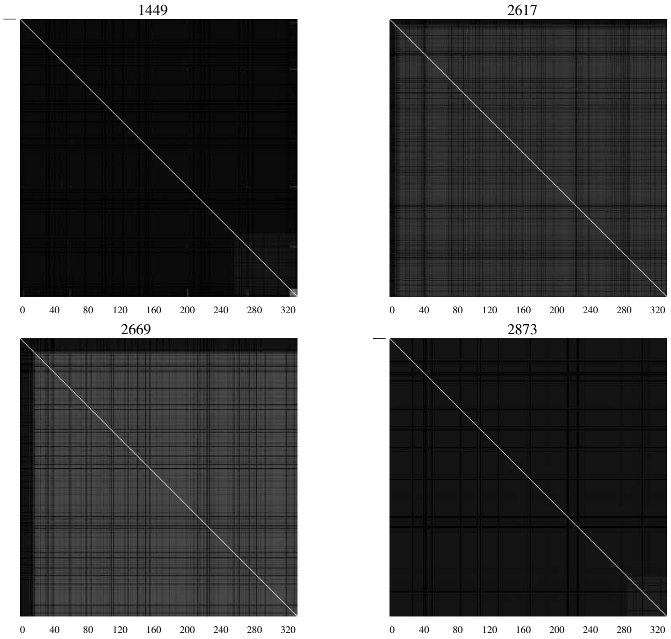

Fig. 2. AMI charts for rapid growers 1449, 2669, 2873, and 2617

III. Results

If we look at the histogram of the AMI values for the various populations of clones as shown in Figures 4 and 5 we find that there is a much wider range of values obtained from the slow grower population than the range of values in the rapid grower population. One of the reasons for this

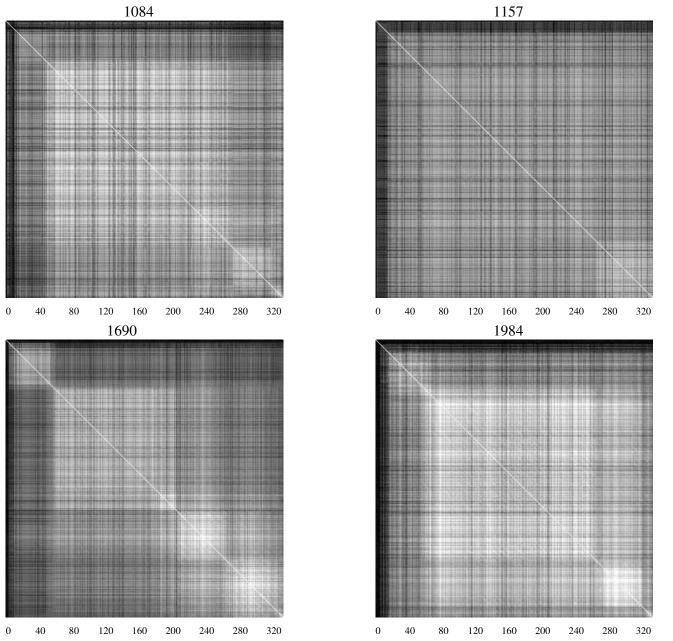

Fig. 3. AMI charts for slow growers. Sequences from the mother infant pairs 1084, 1157, 1690, 1984 were used to generate the chart. Note the ”checkerboard” pattern. The white squares correspond to regions of higher variability

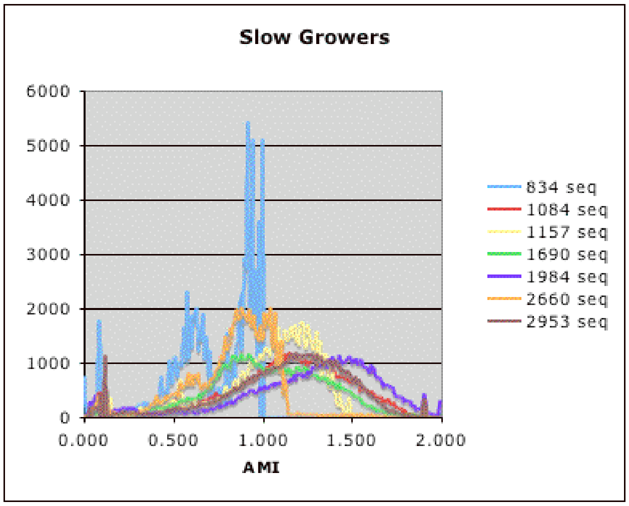

Fig. 4. Histogram of the AMI values of the slow growers. The units for the AMI values is nats

difference is that the there is not much change in the AMI charts at different sampling points for the rapid grower population as shown for one case in Figure 6. The slow grower population on the on the other hand shows a much greater variation between time points as can be seen in Figure 7.

In order to get a single number which would reflect the degree of change between time points we computed the range of differences in the AMI values between AMI charts at different time points. Tabulating the difference in the range of AMI values between the AMI charts obtained for the HIV-1 population from the first sample for an infant and the second sample we see a definite correlation between the range of values and whether the HIV population is a rapid or slow grower population. As shown in Table II six of the seven slow grower population have a range greater than 1.5 while the differences in AMI values between the first and second time points for three of the four rapid growers is about a third that of the slow growers.

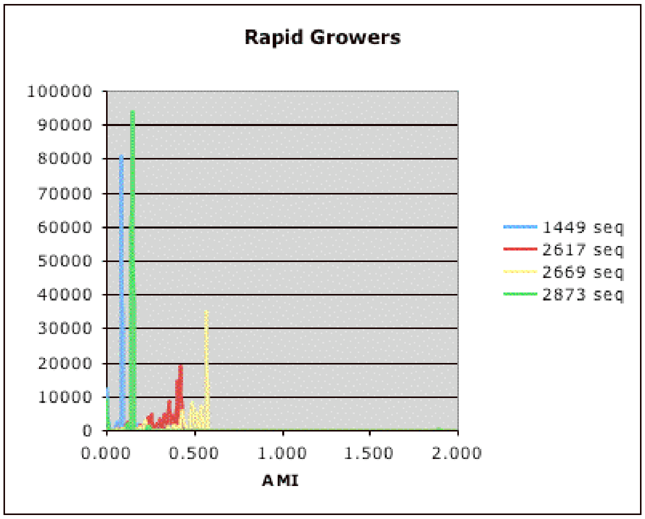

Fig. 5. Histogram of the AMI values of the fast growers. Notice the difference in the range of the y-axis. The AMI values are computed in nats.

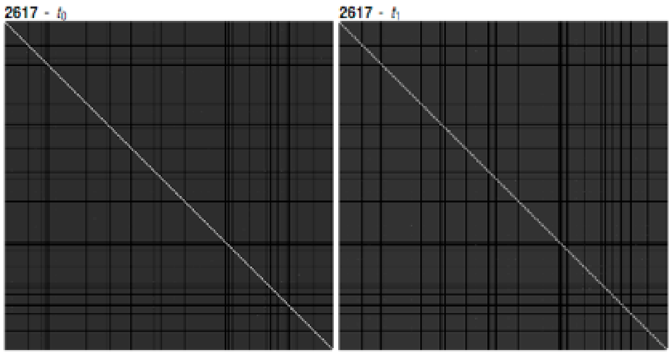

Fig. 6. AMI charts for clones from infant 2617 at the initial time point and two months later.

The results in Table II seems to indicate a clear difference between the rate of change of slow progressors and rapid progressors. However, there are several factors that might be influencing these results. The timepoints t_0 and t_1 are different for the different sets of clones. Therefore, the interval between t_0 and t_1 are also different. Therefore, we need to look at how much of the difference between rapid and slow progressor is because of intrinsic differences between rapid and slow progressors, and how much is due to the different intervals.

In the data available to us we have three cases of rapid progressors for which there are three samples that were taken two months apart and two cases of slow progressors for

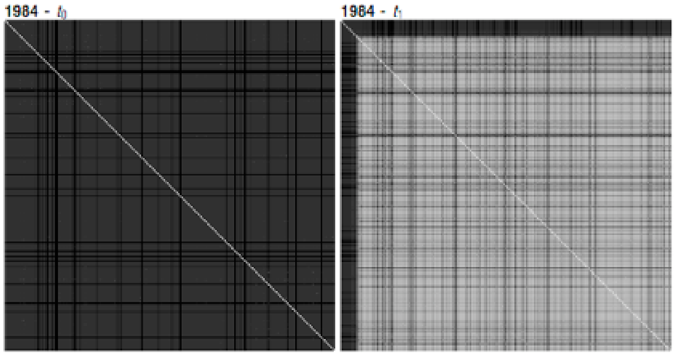

Fig. 7. AMI charts for clones from infant 1984 at the initial time point and two months later.

Sequence	Type	Range
1690	*Slow*	2.34
1157	*Slow*	2.20
2953	*Slow*	2.09
1084	*Slow*	1.70
2660	*Slow*	1.61
1984	*Slow*	1.58
2669	**Rapid**	1.00
834	*Slow*	0.86
2617	**Rapid**	0.51
1449	**Rapid**	0.45
2873	**Rapid**	0.38

TABLE II

RANGE OF DIFFERENCES BETWEEN t_0 AND t_1 FOR DIFFERENT POPULATIONS. NOTICE THE MUCH LARGER RANGE OF DIFFERENCES FOR THE MOST OF THE SLOW PROGRESSOR POPULATIONS WHEN COMPARED WITH THE RAPID PROGRESSOR POPULATION.

which we have samples taken two months apart. The results are shown in Table III

In this case the rapid growers and slow growers can be clearly differentiated.

IV. CONCLUSIONS AND FUTURE WORK

The AMI charts described in this work provide a useful tool for monitoring the behavior of populations. By looking at the AMI charts at different time points we can monitor changes in the genetic makeup of populations of clones. This in turn may be a useful way of monitoring disease progression. Furthermore, the range of difference between AMI values is seen to be an effective statistic for predicting the disease outcome. We are currently testing the approach with more cases. The results presented in this paper are a promising beginning and indicate the usefulness of this approach.

Sequence	Type	Range
1984	Slow	1.58
834	Slow	0.85
2617	**Rapid**	0.51
1449	**Rapid**	0.45
2873	**Rapid**	0.38

TABLE III

RANGE OF DIFFERENCES OF AMI FOR $|t_0 - t_1| = 2$MONTHS FOR DIFFERENT POPULATIONS. NOTICE THE MUCH LARGER RANGE OF DIFFERENCES FOR THE MOST OF THE SLOW PROGRESSOR POPULATIONS WHEN COMPARED WITH THE RAPID PROGRESSOR POPULATION.

REFERENCES

[1] C.E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[2] B.T.M. Korber, R.M. Farber, D.H. Wolpert, and A.S. Lapedes. Covariation of Mutations in the V3 Loop of Human Immunodeficiency Virus Type I Envelope Protein: An Information Theoretic Analysis. *Proceedings of the National Academy of Sciences*, 90:7176–7180, August 1993.

[3] B.G. Giraud, A. Lapedes, and L.C. Liu. Analysis of Correlations Between Sites in Models of Protein Sequences. *Physical Review E*, 58(5):6312–6322, 1998.

[4] H. Herzel and I. Grosse. Correlations in DNA Sequences: The Role of Protein Coding Segments. *Physical Review E*, 55(1):800–810, 1997.

[5] R. Roman-Roldan, P. Bernaolo-Galvan, and J.L. Oliver. Application of Information Theory to DNA Sequence Analysis: A Review. *Pattern Recognition*, 29(7):1187–1194, 1996.

[6] I. Grosse, H. Herzel, S.V. Buldyrev, and H.E. Stanley. Species Independence of Mutual Information in Coding and Noncoding Regions. *Physical Review E*, 61(5):5624–5629, 2000.

[7] M.J. Berryman, A. Allison, and D. Abbot. Mutual Information for Examining Correlations in DNA. *Fluctuation and Noise Letters*, 4(2):237–246, June 2004.

[8] M. Bauer. *A Distance Measure for DNA Sequences*. PhD thesis, University of Nebraska - Lincoln, 2001.

[9] H.H. Otu and K. Sayood. A Divide and Conquer Approach to Sequence Assembly. *Bioinformatics*, 19(1):22–29, January 2003.

[10] L.C. Martin, G.B. Gloor, S.D. Dunn, and L.M. Wahl. Using Information Theory to Search for Co-evolving Residues in Proteins. *Bioinformatics*, 21:4116–4124, 2005.

[11] Hofacker I, Fekete M, Stadler P. Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*,319:1059–1066, 2002.

[12] Lindgreen S, Gardner P, Krogh A. Meauring covariation in RNA alignments: physical realism improves information measure. *Bioinformatics* 22:2988–2995, 2006.

[13] Slonim N, Elemento O, Tavazole S. *Ab initio* genotype-phenotype association reveals intrinsic modularity in genetic networks. *Molecular Systems Biology* 2006.

[14] Zhang H, Orti G, Du Q, He J, Kankasa C, Bhat G, Wood C. Phylogenetic and Phenotypic Analysis of HIV Type 1 Env gp120 in Cases of Subtype C Mother-to-Child Transmission. *AIDS Research and Human Retrovirus* 18:1415-1423, 2002.

[15] Hoffman FG, He X, West JT, Lemey P, Kankasa C, Wood C. Genetic Variation in Mother-Child Acute Seroconverter Pairs from Zambia *AIDS* 22(7)817-824, 2008.

[16] Zhang H, Hoffman FG, He J, He X, Kankasa C, West JT, Mitchell CD, Ruprecht RM, Orti G, Wood C. Characterization of HIV-1 subtype C envelope glycoproteins from perinatally infected children with different courses of disease *Retrovirology*, 3:73, 2006.