BIOINFORMATICS

Vol. 19 no. 8 2003, pages 981–986
DOI: 10.1093/bioinformatics/btg123

Mining HIV dynamics using independent component analysis

Sorin Draghici[1], Frank Graziano[2], Samira Kettoola[3], Ishwar Sethi[4] and George Towfic[5,*]

[1]Department of Computer Science, Wayne State University, [2]Univ. of Wisconsin Hospital Madison, [3]Department of Computer Science & Software Engineering, UW-Platt, Wisconsin, [4]Department of Computer Science and Engineering, Oakland University, Rochester and [5]Department of Computer Science, Clarke College, Dubuque, IA, USA

Received on September 26, 2002; revised on December 16, 2002; accepted on January 13, 2003

ABSTRACT

Motivation: We implement a data mining technique based on the method of Independent Component Analysis (ICA) to generate reliable independent data sets for different HIV therapies. We show that this technique takes advantage of the ICA power to eliminate the noise generated by artificial interaction of HIV system dynamics. Moreover, the incorporation of the actual laboratory data sets into the analysis phase offers a powerful advantage when compared with other mathematical procedures that consider the general behavior of HIV dynamics.

Results: The ICA algorithm has been used to generate different patterns of the HIV dynamics under different therapy conditions. The Kohonen Map has been used to eliminate redundant noise in each pattern to produce a reliable data set for the simulation phase. We show that under potent antiretroviral drugs, the value of the CD4+ cells in infected persons decreases gradually by about 11% every 100 days and the levels of the CD8+ cells increase gradually by about 2% every 100 days.

Availability: Executable code and data libraries are available by contacting the corresponding author.

Implementation: Mathematica 4 has been used to simulate the suggested model. A Pentium III or higher platform is recommended.

Contact: gtowfic@clarke.edu.

INTRODUCTION

Many approaches have been used to model and analyze the enormous amount of HIV data available in different libraries. In general, current HIV research focuses on three main objectives: (1) providing a clear and easy access to different HIV databases (Huba , 1998); (2) accelerating the process of designing efficient drugs that target the HIV virus (Noever and Baskaran, 1992); and (3) predicting the future behavior of different HIV parameters (Hraba and Dolezal, 1996; Kirschner *et al.*, 2000; Perelson and Nelso, 1999). Since our work aims at achieving the third objective, we will discuss briefly the mathematical approaches currently used in this area.

Mathematical modeling used for HIV simulations, under different therapies, involves a set of simultaneous ordinary differential equations (ODE) that take into account the dynamic of disease processes both at population and cellular levels. Many attempts have been made to establish a computer paradigm based on the derivation of a governing ODE, together with its initial conditions, using a considered HIV data set. The resulting ODE is then used to provide a mean to understand and predict the dynamics of human immunodeficiency virus by simulating the behavior of the CD4+, CD8+ T-cells and the viral load. In general, mathematical modeling has the following two drawbacks: (1) under the same treatment and patient conditions (at a particular point in time), different mathematical models produce different outcomes depending on the parameters and the type of the differential equations used; and (2) relying on one data set sample to setup a set of ODE will produce a specific rather than a generic model that can cover a wide spectrum of cases.

In this work we consider a coupled mathematical model that incorporates three algorithms:

(1) Independent Component Analysis Model (Amari *et al.*, 1996; Wolfram, 2002; Aapo *et al.*, 1998) is used for data refinement and normalization. The model accepts a mixture of CD4+, CD8+, and viral load data for a particular patient and normalizes it with broader data sets obtained from other HIV libraries. The objective of the ICA algorithm is to isolate groups of independent patterns embedded in the considered libraries.

*To whom correspondence should be addressed.

(2) Kohonen Map recurrent networks (Maillet and Rousset, 2001; Pollock *et al.*, 2002) are used to select those patterns chosen by ICA algorithm that are close (within an acceptable precision) to the considered set of input data. Kohonen Map identifies two mechanisms for a network to self-organize spatially:
   (a) locate the unit that best responds to the given input (the winning unit);
   (b) modify the strength of the connections to the winning unit and its connection neighborhood.

   These two mechanisms help not only in selecting similar data sets but also to iteratively improve the system by throwing away unrelated data sets.

(3) Finally, a non-linear regression model is used to predict future mutations in the CD4+, CD8+ and viral loads. This is currently done for a period of 6 months. The data set CD8+ selected in step 2 above is used to predict the required regression parameters that can be used for the prediction. The Dynafit medical package (Kuzmic, 1996) has been used in order to accomplish this goal.

THE ALGORITHM

Figure 1 represents a block diagram that describes the overall activities involved in the considered model. As shown in Figure 1, initial input data sets are processed first using ICA to produce independent sets that embed different behaviors of the system dynamics under different conditions. These sets are then processed by the Kohonen Map to eliminate noise and further refine each of the independent sets. The considered input data representing individual patients' record is then compared with each group of the resulting independent sets to select those members that are close within some degree of precision to the considered patient record. Finally, non-linear regression is applied on this set of data to further advance the dynamic system in time. Mathematically, this can be expressed as follows. Let

$$\Omega(\zeta) = \{\Omega_1(\zeta), \Omega_2(\zeta), \ldots, \Omega_n(\zeta)\}$$

where $\Omega(\zeta)$ represents a set of data that contains all considered data sets that are related to the parameter ζ in the considered libraries (Pennisi and Gohen, 1996; Kavacs *et al.*, 1996; Juriaans *et al.*, 1994; http://www.huba.com; and the UW–Madison's data set). The parameter ζ represents one of the considered HIV components to be analyzed. Thus ζ could be CD4+, CD8+, or VL.

Each of the subsets $\Omega_1(\zeta), \Omega_2(\zeta), \ldots, \Omega_n(\zeta)$ contains a set of closely identical patterns of data (for example nearly identical CD4+ profiles) that belongs to $\Omega(\zeta)$. i.e.:

$$\Omega_i(\zeta) = \{\delta_{i1}(\zeta), \delta_{i2}(\zeta), \ldots, \delta_i p(\zeta)\}$$

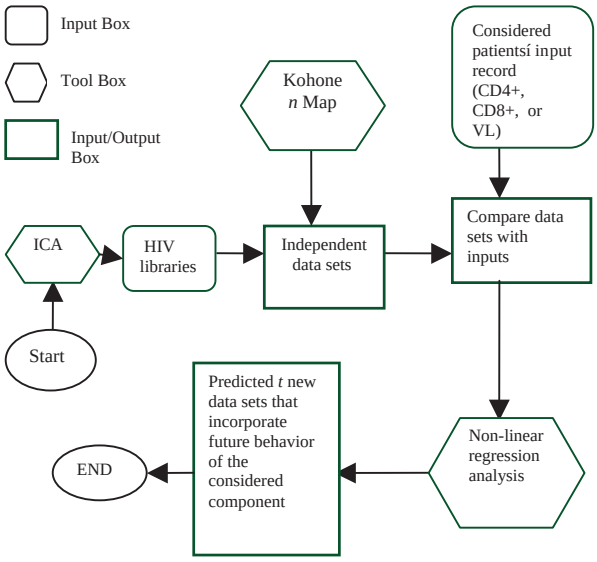

Fig. 1. Block diagram for the overall activities involved in the HIV prediction.

$$(i = 1, 2, \ldots, n)$$

p is the number of nearly identical sets in $\Omega_i(\zeta)$), Here,

$$\Omega_1(\zeta) \cap \Omega_2(\zeta) \cap, \ldots \cap \Omega_n(\zeta) \approx \phi$$

where ϕ is the empty set, and

$$\delta_{i1}(\zeta), \cong \delta_{i2}(\zeta), \ldots \cong \delta_{ip}(\zeta)(i = 1, 2, \ldots n)$$

The ICA model applied in this work is a modification of the blind source separation approach discussed in Amari *et al.* (1996). In this model, an observed vector of mixed components, correspond to a realization of m-dimensional discrete-time dependent variables, is considered. A combination of neural networks approach and eigenvector analysis is used to predict the set of components that constructed the considered mixed vector. We apply the suggested ICA model on a set of data obtained from the considered HIV libraries.

Figure 2 below shows an abstract representation to the Independent component analysis algorithm used in this work. In Figure 2, $Y[m, n]$ is a mixing matrix such that 'm' represents the total number of samples considered and 'n' represents total number of data sources. In our case 'm' represents the total number of CD4+, CD8+, or viral load vectors for a considered time interval and 'n' represents the total number of patients.

RESULTS AND DISCUSSION

Here we analyze the impact of permanent treatment at different time periods on the dynamic behavior of the

begin
{
'Normalize the mixing matrix $Y[m, n]$ with respect to its overall mean value'
'Ignore non-significant data sets in $Y[m, n]$ by throwing away all vectors associated with minimum Eigenvalues'.
'Initialize a random weighting matrix $W[m, n]$ such that it has an orthogonal, unit norm columns'
for $t > 0$ until $w(t+1) \simeq w(t)$ **do**
{
'calculate $w(t+1) = w(t) \pm \mu(t)[y(t)\Psi(w(t)^{\mathrm{T}} y(t) - w(t))]$'
// **where** $0 < \mu(t) < 1$; $\Psi(x) = x \exp(-x^2/2)$; T is the transpose operator,
}
'Estimate the independent components s that compose the mixing matrix Y using the formula: $s = w^{\mathrm{T}} \bullet y$'
// where $\bullet$ is the dot product operation
}

Fig. 2. ICA algorithm for HIV data separation

CD4+, CD8+, and the viral load. In each study case we consider the following:

(1) the use of a combination of three-medication treatment of nucleoside analogs, protease inhibitors and non-nucleoside reverse transcriptease inhibitors.

(2) HIV dynamics are studies for the period of 10 years (using the non-linear regression analysis provided by Kuzmic (1996). This is considered as a suitable period as indicated by many researchers (Blower *et al.*, 1999)).

(3) CD4+ lymphocyte dynamics is considered because the depletion of this T-cell subpopulation and the parallel decrease in the helper activities of T lymphocytes seemed to be the major immune system effect caused by HIV infection. Cytokines produced by CD8+ lymphocytes is considered since it inhibits HIV proliferation (Baier *et al.*, 1995).

(4) It is assumed, as indicated by the preliminary analysis on the considered data sets that the T-helper cell activity does not decrease linearly with the decline of the CD4+ lymphocytes but faster.

(5) For better test of system dynamics, we test different parameters when the CD4+ count is under the 200 measures. This provides a better test for the system under critical conditions.

(6) To get an unbiased model, we randomly choose different subsets from the considered laboratory data for modeling purposes. The remaining data sets are used to compare the obtained simulation result with that of the remaining laboratory data sets.

Effect of permanent treatment on CD4+ counts

Figure 3 shows a set of laboratory data at different time intervals. When combination treatments are applied on the ICA vector, which represents a set of selected

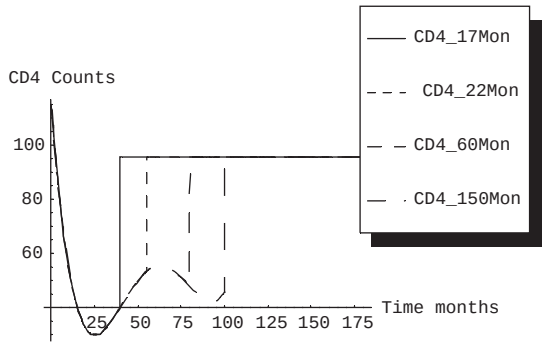

Fig. 3. Laboratory CD4+ data for different time intervals.

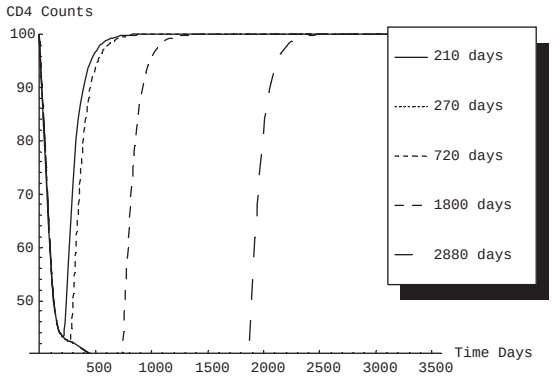

Fig. 4. Simulation of CD4+ obtained at different time intervals.

CD4+ count, we obtain the set of data that we show in Figure 4 for different time internals. An analysis of the data obtained from different simulations considered in Figure 4, shows that although we started treatments at different times, the final steady state of the CD4+ cells is the same $\cong$ 99.96. The final steady state value seems to depend only on the effectiveness of the therapy,

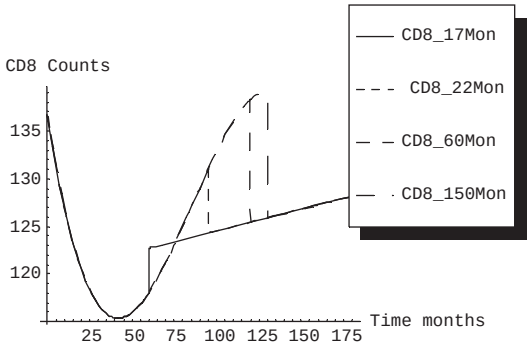

Fig. 5. Laboratory CD8+ data sets for different time intervals.

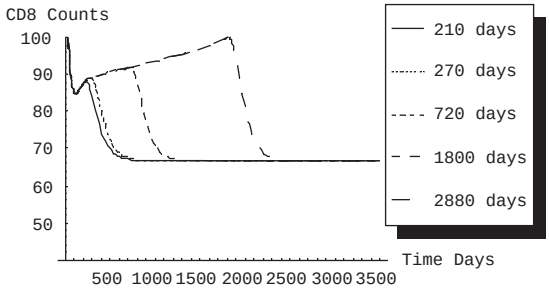

Fig. 6. Simulation of the CD8+ T-cells obtained for different time intervals.

regardless of the onset of treatment. This is consistent with the laboratory data. When CD4+ lymphocyte numbers are too low (<40) therapy can no longer reverse the CD4+ cell depletion at the required time (<10 years). The symptomatic phase begins when the concentration and diversity of helper T-lymphocytes becomes too low, which causes a collapse in cellular immunity.

The maximum value of the CD4+ count can reach the baseline value even when treatment starts years after the initial HIV acquisition. The only limit to this is when CD4+ T-helper cells reach a very low value (<40) to the point where the T-helper activity deceases non-linearly with the number of CD4+ count (as indicated in Fig. 4). Here, we consider a permanent therapy lasts 7 months, 9 months, 2, 5 and 8 years.

Effect of permanent treatment on CD8+ counts

Figure 5 shows a set of laboratory CD8+ data at different time intervals. The data in Figure 6 shows that at this critical stage of HIV mutation, where the CD4+ count reaches a low level, the CD8+ T-cells start to increase at a rate proportional to that of the CD4+ count. This is an indication that the T-cell activities can no longer prevent the accumulation of the CD8+ T-cells. A comparison between the corresponding cases in Figures 4 and 6 show that the value of the CD8+ count is higher than that of the CD4+ count when the virus infection has a dominant effect. The value of the CD4+ cells decreases gradually by about 11% every 100 days and the levels of the CD8+ cells increase gradually by about 2% every 100 days. Otherwise, the normal situations where the CD4+ count are greater than the CD8+ count is satisfied. This is consistent with the considered laboratory data sets shown in Figures 3 and 5.

In all considered cases shown in Figure 6, the final stable values of CD8+ count are less than that of the initial values of these cells before treatment started. An exception for this is in the last case where treatment could not suppress the CD8+ T-cells and hence could not improve on the CD4+ T-cells.

Effect of treatment on viral loads

We consider here the viral load data sets classified by the ICA algorithm and refined by the Kohonen procedure. The non-linear regression algorithm has been used on the resulting data sets for the prediction phase. Figure 7 shows the dynamics of the viral loads for the same time period considered for the CD4+, and CD8+ T-cells. Figure 7 shows that the viral load increases in the first stage of the viral infection at a rate of about 400%. It then keeps increasing at a low percentage of about 2% in a 1-month interval. This is due to the behavior of immune systems under HIV infections where the immune system reacts to the sudden increase in the viral loads and thus restricts its propagation. When treatment is delayed, the viral load count starts to increase exponentially again (last experiment in Fig. 7). This exponential increase in the viral load will reduce the effect of the immune system and thus prevents the increase of the CD4+ count (as can be verified in Fig. 4).

It is clear from Figures 4, 6 and 7 that the dominant factor in the HIV virus control is the number of the CD4+ count rather than the viral load count. Although the rate of increase in the viral loads is higher than that of the CD4+, it is still possible to control the viral load count when treatment is started at an early stage. The rate of change in the CD4+, under treatment, is much less than that of both the CD8+ count and the viral load.

CONCLUSIONS

Independent component analysis (ICA), Kohonen networks and non-linear regression analysis have been used to study the dynamic behavior of different components that have a major impact on the immune system in HIV-infected persons. The main advantage of the ICA method (as compared with the traditional mathematical modeling using a set of simultaneous differential equations) is that

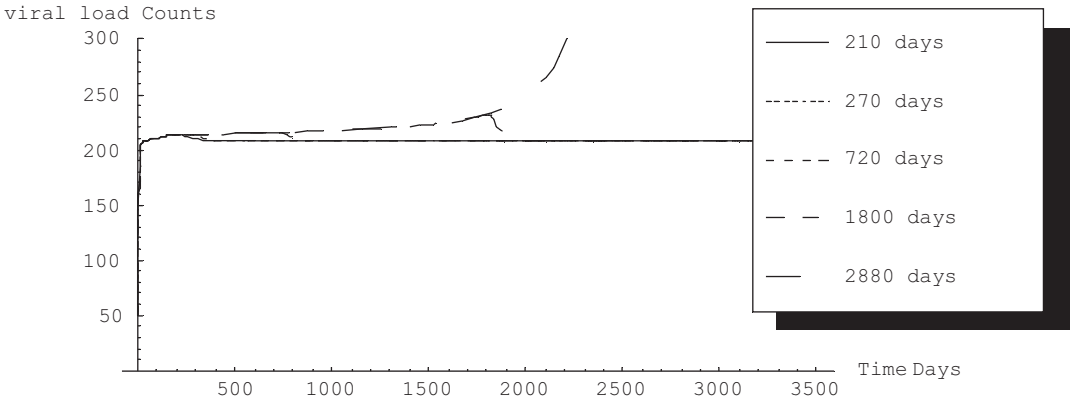

Fig. 7. Simulation of the viral load for different time durations.

while mathematical modeling requires an expert to incorporate the dynamic behavior in the differential equations, ICA analyzes the system by mining into the real data and automatically selects the appropriate components to be incorporated. Another advantage of the ICA as compared to the mathematical model is that modifications to the data sets are automatically incorporated into the model. This is not the case when simultaneous equations are considered where the model has to be modified, by a mathematical expert, each time new data or analysis concepts are considered.

The use of the Kohonen Map helped in selecting a set of data that is close (up to a required degree of precision) to the considered input data. While the ICA model is efficient in isolating different sets of data that have an independent behavior, the Kohonen model provided an efficient selection of a particular set of data that has a high degree of similarity with the considered patients' input record. The main advantage of the Kohonen model, for our application, is that it offers a 'best selection' algorithm that produces a common behavior of different sets of data rather than a behavior that simulates a particular data set. Another advantage of the Kohonen model is that it offered more refined and accurate data for the regression phase.

The non-linear regression model has helped in advancing the historical data by predicting future behavior of the system dynamics (the behavior of CD4+, CD8+, and VL in the foreseeable future, which is considered in our case to be within a 10 years limit).

Combination therapy (using a combination of nucleoside analogs; protease inhibitors and non-nucleoside reverse transcriptase inhibitors) has been implemented at different stages of the virus infection. It is shown in Figures 4, 6 and 7 that when substantial decline of CD4+ lymphocytes starts, it progresses rapidly to its total depletion, and then a permanent steady state of the CD4+ cell level is established. A steady state of the CD4+ lymphocyte level is accompanied by a steady state of the CD8+ level. On the other hand, when T-helper activity decreases non-linearly with the CD4+ and CD8+ lymphocyte dynamics, the value of the CD4+ cells decreases gradually by about 11% every 100 days and the levels of the CD8+ cells increase gradually by about 2% every 100 days (a ratio of about 5.5/1). In the case of combination therapy, the final steady state value seems to depend only on the effectiveness of the therapy, regardless of the onset of treatment.

ACKNOWLEDGEMENTS

We acknowledge Dr Bob Schatz, UW-Platteville Coordinator of Corporate Relations and Jennifer Bellehumeur, R.N.,M.S. Research Coordinator, Immunology Clinic, UW-Madison Hospital and Clinics, for their valuable coordination between the UW-Platteville University and the medical school at the University of Wisconsin and in providing the HIV data and coordinating the research requirements in different stages. The third author would like to express her gratitude to UW-Platteville for their SAIF grant support.

REFERENCES

Aapo,H., Rinen, and Erkki,O. (1998) Independent component analysis by general. *Signal Processing*, Vol. 61.

Amari,S., Cichocki,A. and Yang,H. (1996) A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, Vol. 8, MIT Press, Cambridge.

Back,A. and Cichocki,A. (1997) Blind source separation and deconvolution of fast sampled signals. In Kasabov,N. (ed.), *Proceedings of the International Conference on Neural Information Processing, ICONIP-97, New Zealand*, Vol. I, Springer, New York, pp. 637–641.

Baier,M., Werner,A., Bannert,N., Metzner,K. and Kurt,R. (1995) HIV suppression by interleukin-16. *Nature*, 378–563.

Blower,S., Koelle,K., Kirschner,D. and Mills,J. (1999) Live attenuated HIV vaccines predicting the tradeoff between efficacy and safety. *PNAS*, **98**, 3618–3623. 2017, HIVAIDS Surveillance Database (Population Division), US Bureau of the Census.

Hraba,T. and Dolezal,J. (1996) A mathematical model and CD4+ lymphcyte dynamics in HIV infection. Vol. 2.

Huba,G. (1998) *AIDS Capitation*. Cherin,D. and Huba,G. (eds), Haworth Press, New York.

Juriaans,S., Van Gemen,B., Weverling,G., Van Strijp,D., Nara,P. Coutin Coutinho,R. *et al.* (1994) The natural history of HIV-1 infection: virus load and virus phenotype independent determinants of clinical course? *Virology*, **204**, 223–233.

Kavacs,J., Vogel,S. Albert,J. *et al.* (1996) Controlled trial of IL-2 infusions in patients infected with HIV. *New Eng. J. Med.*, **335**, 1350–1356.

Kirschner,D., Webb,G. and Cloyd,M. (2000) A model of HIV-1 disease progression based on virus-induced and homing-induced apoptosis of CD4+ T lymphocytes. *J. AIDS Human Retrov.*, **24**, 352–362.

Kuzmic,P. (1996) Program DYNAFIT for the analysis of enzyme kinetic data application to HIV proteinase. *Anal. Biochem.*, **237**, 260–273.

Maillet,B. and Rousset,P. (2001) Classifying hedge funds with Kohonen maps: a first attempt, Working Paper, University of Paris I Pantheon-Sorbonne.

Noever,D. and Baskaran,S. (1992) *Steady-state vs. generational genetic algorithms*, a comparison of time complexity and convergence properties, Santa Fe Institute, paper # 92-07-032.

Pennisi,E. and Gohen,J. (1996) Eradication of HIV from a patient: not just a dream? *Science*, **272**, 1884.

Perelson,A. and Nelso,P. (1999) Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Rev.*, **41**, 3–44.

Pollock,R., Lane,T. and Watts,M. (2002) A Kohonen self-organizing map for the functional classification of proteins based on one-dimensional sequence information. In *Proceedings of IJCNN*. pp. 189–192.

Wolfram,L. (2002) Linear modes of gene expressions determined by ICA. *Bioinformatics*, **18**, 51–60.