Mathematical Meta-Model for HIV - From Cell to Society

ABOUBEKEUR HAMDI-CHERIF (1)
Computer College
Computer Science Department
Qassim University
PO Box 6688 – 51452 Buraydah
SAUDI ARABIA

(1) Permanent Address : Université Ferhat Abbas Setif (UFAS)
Faculty of Science
Computer Science Department
19000 Setif
ALGERIA

(1) email: shrief@qu.edu.sa, elhamdi62@gmail.com

Abstract: - The Human Immunodeficiency Virus (HIV) is responsible for one of the most crucial pending public health issues of our age. We describe a set of existing computational techniques that might improve our understanding of the dynamics within individual cells along with its epidemiological profiling. While HIV-contaminated cell mathematical modeling leads to constantly improved models, the study of the social spread of the disease will hopefully lead to ad-hoc regional and global policies for its control. Some mathematical models and their corresponding proven computer implementation, with their mutual interactions, are discussed and integrated within a coherent framework allowing their cross-comparison through an information-theoretic meta-model.

Key-Words: - Mathematical modeling, Information-theoretic modeling, HIV dynamics, Cell modeling, Cell regulation, HIV epidemiology

1. Introduction

In this paper, we propose to present and critically analyse the major Human Immunodeficiency Virus (HIV) mathematical models and their corresponding efficient computer implementation. A meta-model is presented that takes into account the relative merit of models with respect to each other. By meta-model, we mean a mathematical framework that is capable of ranking models using some sort of metric. Two complementary lines of research can be considered. The first one concerns the study of cell behavior via the integration of various mathematical models within a unified framework. The complementary perspective investigates the social spread of the disease and some pertinent models that describe HIV dynamics. Although concentration is now given to the first perspective, the second one can easily be incorporated within the framework. This is so because, at the methodological level, the meta-model does not distinguish between the two perspectives. Indeed, the meta-model considers real data, modelled data and models as abstracted entities. In any case, models are to be checked against real data. The tangible benefits of the proposed approach are:

- A better understanding of cell contamination / cure, with the possibility of choosing / adapting / developing improved ad hoc mathematical model(s) resulting in a better control of the disease at the cellular level leading to its dynamic monitoring.
- The possibility of choosing / adapting / developing a good epidemiological model entailing a better social control of the spread of the disease, through regional and global policies.

The far-reaching purpose of this multidisciplinary research is to contribute to the eradication of HIV through better understanding of the virus within the cell and its spread within society using advanced mathematical models and efficient computational techniques. What are the mathematical and computational means to control the virus’ spread as a first step towards its eradication? The main goal of the research is to contribute to elucidating this fundamental question. One of the earliest mathematical models in HIV dynamics modeling is presented in [14]. This seminal work catalyzed a large number of studies over the next several years, with the more theoretical work consisting primarily of variations. It is this body of research that we wish to concentrate on. Later, three mathematical models with their comparison have been compared in [25]. An introduction to information theory-based model
selection and application of the technique to mathematical models of HIV infection dynamics are reported in [3], whereby six previously published models are used and compared with respect to their ability to represent patient data from a study employing reverse transcriptase mono-therapy. A mathematical ranking of each of the model’s ability to represent patient data using Akaike Information Criterion (AIC) is obtained showing what features of the model are important. Despite the existence of these mathematical models and many others, an efficient meta-model for comparing models is still lacking. For developing a useful meta-model, we propose to use a set of parameterized metrics for comparing the available mathematical models. The present article is structured as follows. In the next section, we describe an overview of relevant research work stressing the importance of bioinformatics, optimization and epidemiological modeling. Section 3 summarizes the biological and mathematical prerequisites. In Section 4, the proposed meta-model is presented. Section 5 reports the main known mathematical models to be initially used as basic entities of the meta-model.

2. Relevant recent research
2.1 Bioinformatics settings
The present study stems from recent research that mainly concerns the integration of novel mathematical tools within a bioinformatics framework, now under process:
- Bridging the gap between machine learning and bioinformatics [9], [10],
- Integration of control methods within bioinformatics with special reference to biological control [11], [12].
- Adaptation of one evolutionary multiobjective optimization (EMO) method in HIV settings based on the applications done for the classification of two types of cancer [13].

2.2 Optimal control applications
Many HIV optimal control applications have been devised e.g. [16]. A nonlinear third-order mathematical model of HIV dynamics has recently been proposed whose goal is to generate a control strategy to keep the number of HIV virions under a pre-specified level and to reduce the total amount of medications that patients receive [22].

2.3 Computational intelligence application
Computational intelligence is a machine learning sub-field that uses past data to infer, statistically, or otherwise, future behavior of the data. In this respect, an evolutionary method has been proposed for designing multidrug therapies for HIV [20].

2.4 Epidemiological example of models
Various models are used to predict spread of the disease at the social level. In [15], the authors use a model to predict its spread in Australia by 2015, on the basis of the data available in 2006. At the abstract level the meta-model does not make any difference between models describing cellular functioning and those relative to its social impact.

3. HIV: from virology to modeling
Before describing the relevant HIV-related mathematical models, we give a brief account of the biological principles behind HIV. The main biological reference is [26].

3.1 HIV as a retrovirus
3.1.1 Healthy cell functions
In healthy living cells, the genomic sequence information ordinarily flows as a result of the following biological operations:
- Replication: i.e., from deoxyribonucleic acid (DNA) to DNA, knowing that DNA is an informational molecule encoding the genetic instructions and representing the basic building block in the development and functioning of all known living organisms and many viruses.
- Transcription: i.e., from DNA to messenger ribonucleic acid (mRNA); this latter carries information about a protein sequence to the ribosomes, the protein synthesis factories in the cell. mRNA is coded so that every three nucleotides, or a codon, correspond to one amino acid.
- Translation: i.e., from RNA to protein. When the mRNA is exported from the nucleus to the cytoplasm, it is bound to ribosomes and translated into its corresponding protein form with the help of transfer RNA (tRNA) [6].

3.1.2 HIV infected cell
HIV is a member of the genus lentivirus, a family of retroviruses. At the opposite of ordinary flow of genomic sequence information, exposed above in the case of healthy cell, retroviruses transfer their information via the process of reverse transcription i.e., from RNA to DNA. The reverse transcription process of viral RNA changes the functions and genomic structure of CD4+T-cells. As a retrovirus, HIV primarily infects varieties of immune cells such as macrophages and lymphocyte T-cells which make up a quarter of the white blood cell count. T-cells
can be further divided into CD8 and CD4+T-cells. CD4+T-cells are the main target of HIV infection (by T-trophic strains). Basically, HIV infection reduces levels of CD4+T-cells using three main steps, namely:

- **Direct killing:** producing viral killing of infected cells;
- **Apoptosis:** increased rates of apoptosis, or programmed cell death (PCD) in infected cells;
- **CD4-CD8 interaction:** killing of infected CD4+T-cells by CD8 cytotoxic lymphocytes that recognize infected cells. These changes damage the immune system and lead to low CD4+T-cell counts. When CD4+T-cell numbers decreases below a certain critical level, less than 200 cells per mm³, normal cell-supported immunity is lost, and the body becomes more susceptible to many kinds of infections [26].

### 3.1.3 Retrovirus life cycle

The virus has no independent existence and cannot replicate outside living cells. Its replication can be broken down into three successive stages:

- **Entry to the cell:** The virus binds itself to the target monocyte/macrophage and CD4+T-cells by adsorbing its surface proteins (the envelop gp120 protein) to two host-cell receptors (proteins). After HIV has bound to the target cell, the HIV RNA and various enzymes, including reverse transcriptase, integrase, ribonuclease, and protease, are injected into the cell. During the microtubule-based transport to the nucleus, the viral single-strand RNA genome is transcribed into double-strand DNA, which is then integrated into a host chromosome [5].

- **Replication and transcription:** after the virus enters the cell, an enzyme called reverse transcriptase (RT) liberates the single-stranded RNA genome from the attached viral proteins and copies it into a complementary DNA (cDNA) molecule. The process of reverse transcription produces many errors resulting in mutations that may cause drug resistance or allow the virus to evade the body’s immune system. This integrated viral DNA may then lie dormant, in the latent stage of HIV infection [27].

- **Assembly and release:** The new mRNA codes for the new viral proteins that will contribute to the reconstruction of the HIV-RNA. The viral proteins help the mRNA and the reconstruction proteins to transport into the cell membrane side. The structural components of the virus accumulate at the membrane of the infected cell to construct the HIV virion. Leftover proteins associated with the inner surface of the host-cell membrane, along with the HIV-RNA, are released to form a bud from the host cell, and can proceed to infect other healthy cells.

### 3.2 Mathematical modeling prerequisites

When studying a complex phenomenon, such as the one addressed here, the corresponding mathematical model(s) need(s) to be rich enough to describe relevant biological aspects but not richer so as to allow for mathematical tractability and simulations applicability. As a result, some sort of philosophical razor is essential before embarking on modeling, and especially when making the initial assumptions. For instance, when building a model, Occam’s razor makes an ordering of competing hypotheses and chooses the one that makes the fewest assumptions. In Hitchens’ razor, the so-called burden of proof lies with the model. If the latter is not supported by evidence, it will remain a claim – certainly vain until proved. In any scientific debate, the opponent does not need to argue against an unfounded claim. In Popper’s falsifiability principle, a theory can only be scientific if it is falsifiable; this latter term coarsely understood as ‘testable’. Falsifiability of a hypothesis (here hypothesis is understood as mathematical model) means that if it false then its falsehood will be discovered through observation or by experimentation. Far from being a negative characteristic, falsifiability is considered a positive and perhaps essential quality of a model. As a result of falsifiability, the model is testable by empirical experiment and thus conforms to the standards of the scientific method.

Furthermore, any mathematical model, like the ones describing a complex biological phenomenon, needs the following fundamental characteristics:

- **Approximation:** simplifying assumptions are necessary for the description of the real world; this is done using an appropriate philosophical razor.
- **Specificity:** the model has to address a specific target, in order to answer a precise question about the phenomenon.

For instance, some HIV models are used to estimate specific immune and virological parameters, an optimized therapy or the expected number of newly infected cells at the individual scale, or newly infected individuals at the social scale. Other models concentrate on the understanding of cause-effect relation between different biological processes of the virus, once again either at the individual level or social level. These models allow us to study various combinations of dynamics of cell behavior or disease spread, such as linear versus non-linear,
proliferation versus no proliferation or time delays versus no time delays. In this way, we are able to find what mathematical features best describe the data that are modeled and provide a useful line of research for others to consider when modeling infectious diseases. Obviously, other questions about HIV give rise to totally different models with ad-hoc approximations regarding the overall phenomenon [24].

4 Meta-model construction
A meta-model is necessary to explore the validity of a given model or when comparing it with other similar models. We need to measure the estimates of the expected error between a given model, being tested, and the actual data provided in real life. This error is naturally produced during the process of approximation / identification. As a result, a metric is necessary. Metrics encompass properties that are the prerequisites for several important convergence theorems for iterative algorithms, i.e., Banach’s fixed point theorem, which is the basis of several pattern-matching algorithms. Another important property is boundedness when numerical applications are considered. The methodological steps of the proposed meta-model can described as follows:

// Meta-Model Construction //
Methodological Steps

Given
- Real data from the environment

Do
1. Initialization
- Identify metadata (the way data are structured)
- Construct a set of simplifying hypotheses describing data
- Identify goals of model
- Construct an initial model
- Test and obtain modeled data

2. Comparison
- Choose a metric
- Use chosen metric to compare real and modeled data

3. Iteration
- Iterate on hypotheses by adapting/adopting a philosophical razor
- Iterate on metric choice

4. Improvement/Termination
- Continuously improve model
- Terminate if satisfied by model results
- Otherwise restart by re-initialization
- Iterate on goals sought by model

4.1 Information-theoretic approach
One of the popular methods for conducting the present study is the so-called Akaike Information Criterion (AIC) which is an information-theoretic measure [1], [2], [17]. AIC is measured using the so-called Kullback-Leibler distance.

4.1.1 Akaike information criterion (AIC)
Given a set of candidate models describing the data, the preferred model is the one with the minimum AIC value. The AIC value [1] is given by:

$$AIC = 2K - \ln(L(\hat{\theta}/y))$$

(1)

where $\hat{\theta}$ is the vector of $K$ best fit maximum likelihood estimator (MLE) parameters in the model and $L$ is the log-likelihood function conditioned on $y$, the given data. The AIC estimates the expected relative “discrepancy” between a fitted model and the unknown true mechanism. In this context, “discrepancy” is the quantification of the information lost when a model is used to approximate data and is denoted as the Kullback-Leibler (K-L) distance, as defined below. Hence, AIC rewards goodness of fit. It also includes a penalty represented by an increasing function of the number of estimated parameters.

4.1.2 Kullback-Leibler divergence / distance
Kullback and Leiber proposed a measure of information from the statistical aspect viewpoint. This measure involves two probability distributions associated with the same experiment. The Kullback-Leibler divergence (KL-Div) measures the difference between two probability distributions taken over the same event space. The KL-Div of the probability distributions $P$ and $Q$, on a finite set $X$ is defined by Equation (2) below:

$$D_{KL-Div}(P // Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

(2)

Clearly from the equation above, KL-Div is a non-symmetric information theoretical measure of $P$ from $Q$. since the KL-Div between $P$ and $Q$ depends on the order in which $P$ and $Q$ are given. Therefore, KL-Div is not a distance metric because it is not symmetric. Various measures have been introduced in the literature generalizing this measure. The Kullback-Leibler distance simply symmetrizes the equation above to obtain a symmetric non-negative quantity characterizing any distance. This distance is
given by:

\[ D_{KL}(P//Q) = D_{KL-Div}(P//Q) + D_{KL-Div}(Q//P) \] (3)

As a result, the KL-divergence of a model from reality may be estimated, to within a constant additive term, by a function (e.g. sum of absolute values, sum of squares) of the deviations observed between the model’s results and data. Estimates of such divergence for models that share the same additive term can in turn be used to choose between models. When trying to fit parametrized models to data there are various estimators which attempt to minimize Kullback–Leibler divergence, such as maximum likelihood and maximum spacing estimators [4].

4.1.3 AIC and Fisher’s function

AIC writes out a simple relationship between the Kullback–Leibler distance and Fisher’s maximized log-likelihood function \( L \). Minimizing the AIC also minimizes the information loss between the model and the data. However, the criteria only estimates the K-L distance up to a phenomena-dependent, additive constant. Thus, while we cannot calculate the K-L distance directly, AIC approach allows that a highly sophisticated theory makes direct comparison and ranking of proposed models via a simple calculation [3].

4.2 Information-theoretic model extensions

One of the major lines of research in building the meta-model is the information-extension as proposed in [4]. An attempt to identify the most appropriate mathematical model that describes HIV dynamics using the AIC has been presented and five mathematical models of HIV infection dynamics investigated in [3], based on the use of Nowak-May-like differential equations [18], [19]. As Kullback–Leibler divergence is not the only divergence addressed in the proposed meta-model, we consider various other metrics.

4.2.1 \( \lambda \) divergence

If \( P \) or \( Q \) have probabilities \( \lambda \) and \( (1-\lambda) \) respectively, \( \lambda \) divergence represents the expected information gain about \( X \) from discovering which probability distribution \( X \) is drawn from either of them. \( \lambda \) divergence is given by:

\[ D_{\lambda}(P//Q) = \lambda D_{KL-Div}(P//M) + (1-\lambda)D_{KL-Div}(Q//M) \]

\[ M = \lambda P + (1-\lambda)Q \] (4)

4.2.2 Jensen–Shannon divergence

For the value \( \lambda = 0.5 \), we obtain the Jensen–Shannon divergence, defined by Equation (5) below:

\[ D_{JS}(P//Q) = \frac{1}{2}D_{KL-Div}(P//N) + \frac{1}{2}D_{KL-Div}(Q//N) \]

\[ N = \frac{1}{2}(P + Q) \] (5)

4.2.3 Endres-Schindelin Metric

Closely-related to the capacitory discrimination and Jensen–Shannon divergence, the Endres-Schindelin metric is a metric for probability distributions, which is bounded, information-theoretically motivated, and has a natural Bayesian interpretation [7]. The square root of the well-known \( \chi^2 \) distance is an asymptotic approximation to it. Let \( X \) be a discrete random variable which can take on \( N \) different values. We now draw an independent and identically distributed sample \( \hat{X} \), where each observation is drawn from one of two known distributions, \( P \) and \( Q \). Each of those is used with equal probability. However, we do not know which one is used when. The coding strategy that gives the shortest average code length for the representation of the data is given by the Endres-Schindelin metric defined by Equation (6):

\[ D_{pq}^2 = \sum_{i=1}^{N} (p_i \log \frac{2p_i}{p_i + q_i} + q_i \log \frac{2q_i}{p_i + q_i}) \] (6)

It is a squared quantity because \( D_{pq} \) does not fulfill the triangular inequality.

4.2.4 Relations between various divergences

As a summary, Kullback divergence \( D_{KL-Div}(P//Q) \) can be interpreted as the inefficiency of assuming that the true distribution is \( Q \) when it really is \( P \), while the Endres-Schindelin metric \( D_{pq} \) could be seen as a minimum inefficiency distance. In fact, \( D_{pq} \) can be seen as a special case of Jensen–Shannon divergence, itself a particular case of \( \lambda \) divergence. The summarizing result is:

\[ \frac{1}{2} D_{pq}^2 = D_{\lambda=0.5}(P, Q) \] (7)

4.2.5 Hellinger distance

The Hellinger distance between two discrete probability distributions \( P(p_1...p_k) \) and \( Q(q_1...q_k) \), is defined as:

\[ D_h^2 = \sum_{i=1}^{k} (p_i \log \sqrt{\frac{2p_i}{p_i + q_i}} + q_i \log \sqrt{\frac{2q_i}{p_i + q_i}}) \]
\[ H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - p_i)^2} \]  

and can be expressed using the Euclidean norm of the difference of the square root vectors, i.e.

\[ H(P, Q) = \frac{1}{\sqrt{2}} \left\| \sqrt{p_i} - \sqrt{q_i} \right\|_2 \]

5 HIV dynamics
The previous Section considered the mathematical tools available for constructing the proposed meta-model. Now, what are the models that are to be tested or compared? In this Section, we report HIV dynamics [23], [21] and discuss limitations of the so-called Nowak-May HIV mathematical model [18] and its extensions [24].

5.1 Early HIV mathematical models
As stressed above, one of the earliest models is due to [14]. This work was the first to show that the infection pathogenesis was a rapidly varying dynamical process during which about twelve billion viral particles per day were being produced in infected individuals. The data from twenty patients were modeled by the following simple linear first-order differential equation:

\[ \frac{dV}{dt} = P - cV \]  

where \( V \) represents the time-dependent viral concentration, \( P \) the daily production rate, and \( c \) the viral clearance rate. By evaluating each of the patient’s viral loads before receiving anti-viral therapy a steady state level for the virus was calculated, thus yielding an initial condition for \( V(t) \).

5.2 Nowak-May HIV model
The standard Nowak-May model takes into account the interaction between the dynamics of three variables, namely the non-infected lymphocytes (\( T \)), the infected lymphocytes (\( I \)) and the viral population (\( V \)) [18]. The dynamics of these variables are given by the set of equations:

\[ \frac{dT}{dt} = \beta (1 - T) - \gamma_{NM} VT \]

\[ \frac{dI}{dt} = \gamma_{NM} VT - \alpha I \]  

\[ \frac{dV}{dt} = \alpha I - \xi_{NM} V \]  

(11)

Three detailed and more involved models are further expanded in [25].

6. Conclusion
Mathematical models are used to explore possible reasons for HIV dynamics. We have described a meta-model for HIV for the purpose of integrating and choosing mathematical models according to the goals under considerations. A set of various parameterized metrics is chosen from available literature in order to give more flexibility in the cross-comparison between candidate models. Thus an information-theoretic-like framework is made available that is able to test models validity at the cellular and epidemiological levels, on the basis of clinical, behavioral and biological data available. The main benefit is HIV dynamics understanding for the purpose of monitoring and control. Testing the meta-model with real data, especially in vivo for cellular dynamics and in real-time for epidemiology, remains a challenging task.

References: