3rd Iranian Conference on Electrical and Electronics Engineering
(ICEEE2011)
Islamic Azad University, Gonabad-Branch
14-15 July 2011

Hepatitis B Virus Infection Control Using Reinforcement Learning

Amin Noori, Mohammad Bagher Naghibi-sistani, Naser Pariz
Department of Control Engineering
Ferdowsi university of Mashhad
Mashhad, Iran
Amin.noori@ieee.org, mb-naghibi@um.ac.ir, n-pariz@um.ac.ir

***Abstract*— In this paper, optimal drug schedule for patients infected by hepatitis B virus (HBV) is obtained. An objective of the control is reducing infected cells and free virions. The optimal control problem is to design an effective drug-schedule to reduce the number of infected cells and free virions in a time-optimal fashion. To achieve this goal, a reinforcement learning (RL), which is one of the best unsupervised machine learning algorithms, is proposed for control. Because RL has no need of environment model, i.e. it is model-free; it has absorbed interests during the recent years, especially in medical applications. Performance evaluation of the proposed algorithm has been performed by simulating on the mathematical model of drug dosage of hepatitis therapy.**

Keywords- drug therapy; hepatitis B virus infection; optimal control; reinforcement learning

I. Introduction

Hepatitis B, caused by the hepatitis B virus (HBV), represents an enormous challenge to global public health, which can lead to cirrhosis and primary hepatocellular carcinoma (HCC). More than 2 billion people alive today have been infected by HBV [1], with 5 million new cases each year [2]. The disease's clinical course varies, in adults usually resulting in self-limiting acute hepatitis or, rarely, fatal fulminant disease [2-4]. Acute hepatitis B infection does not usually require treatment because most adults clear the infection spontaneously [5]. Early antiviral treatment may only be required in less than 1% of patients, whose infection takes a very aggressive course. On the other hand, treatment of chronic infection may be necessary to reduce the risk of cirrhosis and liver cancer. Chronically infected individuals with persistently elevated serum alanine aminotransferase, a marker of liver damage, and HBV deoxyribonucleic acid (DNA) levels are candidates for therapy [6].

Although none of the available drugs can clear the infection, they can stop the virus from replicating, and prevent liver damage such as cirrhosis and liver cancer. Treatments include antiviral drugs such as lamivudine, adefovir, tenofovir, telbivudine and entecavir, and immune system modulators such as interferon alpha. However, some individuals are much more likely to respond than others and this might be because of the genotype of the infecting virus or the patient's heredity. The treatment works by reducing the viral load, which in turn reduces viral replication in the liver [7].

The optimal control problem is to design an effective drug-schedule to reduce the number of infected cells and free virions for patients infected by HBV. The intention of this paper is to introduce a model-free based Reinforcement Learning (RL) control approach, based on agent-environment interaction feedbacks, to optimizing drug dosage.

The rest of the paper is organized as follows. Section II discusses the mathematical issues of the compartment model. In Section III basic concepts of reinforcement learning will be outlined. Section IV discusses the control strategy. Section V illustrates the results of the work. Section VI concludes the paper by admiring the model-free feature of the algorithm.

II. Mathematical Model of HBV

The study of anti-HBV infection treatment may benefit from the use of mathematical modeling. Several models have been introduced for understanding HBV dynamics [8-11]. Among those models, the basic virus infection model (BVIM) introduced by Zeuzem et al. [10] and Nowak et al. [8] is widely used in the studies of virus infection dynamics. The BVIM with three variables takes the form of

$$\begin{cases} \frac{dT}{dt} = \lambda - dT - \beta VT \\ \frac{dI}{dt} = \beta VT - \delta I \\ \frac{dv}{dt} = pI - cV \end{cases}, \qquad (1)$$

Where T, I and V are numbers of uninfected cells, infected cells and free virions, respectively. Uninfected cells are assumed to be produced at the constant rate λ. Uninfected cells are assumed to die at the rate of dT and become infected at the rate of βVT, where β is a constant rate that describing the infection process. Infected cells are thus produced at the rate of βVT and are assumed to die at the rate of δI. Free virions are assumed to be produced from infected cells at the rate of pI and are removed at the rate of cV. This model can describe some aspects of the viral dynamics in HBV infection.

In this paper we use an mathematical model of HBV that introduced by Hattaf et al. [7]. The HBV model is given by the following nonlinear system of differential equations

$$\begin{cases} \frac{dT}{dt} = \lambda - dT - (1 - u_1(t))\beta VT \\ \frac{dI}{dt} = (1 - u_1(t))\beta VT - \delta I \\ \frac{dV}{dt} = (1 - u_2(t))pI - cV \end{cases} , \qquad (2)$$

Where $u_1(t)$, represents the efficiency of drug therapy in blocking new infection, so that infection rate in the presence of drug is $(1 - u_1(t))\beta$. The control $u_2(t)$, represents the efficiency of drug therapy in inhibiting viral production, such that the virion production rate under therapy is $(1 - u_2(t))p$.

III. Reinforcement Learning

The reinforcement learning problem is the learning problem of agent interacts with its environment so as to achieve its goal. In fact, RL is learning policy which means what to do or, in other words, how to map each situation to action to maximize the received long run reward. The trial and error and delayed reward are the most important characteristics of RL. In order to find the best policy, at first, the agent must obtain new experiences based on trial and error. In each state of environment, it evaluates how good the chosen action is considering the immediate reward and value of new state. The value of a state is the long term reward expected to be acquired over the future starting from that state [12],[13].

Assume that an agent interacts with its environment at a sequence of discrete time steps, t=0, 1, 2, . . ., as shown in Fig. 1. Also, assume that $S = \{s_1, s_2 . . . s_n\}$ is the finite set of possible states of the environment and $A = \{a_1, a_2, . . ., a_m\}$ is the finite set of admissible actions, which the agent can take. At each time step t, the agent senses the current state $s_t = s \in S$ of its environment and accordingly selects an action $a_t = a \in A$. As a result of its action, the state of environment changes to the new state $s_{t+1} = s' \in S$, with a transition probability $P_{ss'}(a)$, and the agent receives an immediate reward r_{t+1}.

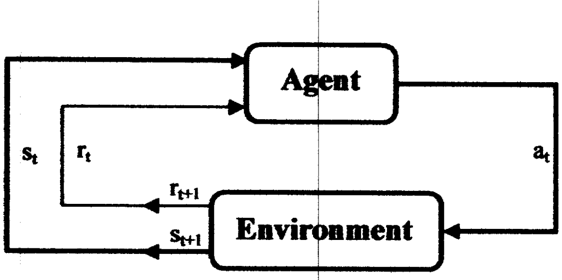

Figure 1. The agent-environment interaction in RL

However, the reinforcement learning is established on an important assumption that the interaction of agent and dynamic environment satisfies the Markov property and the reinforcement learning task is a Markov Decision Process (MDP). In this condition, the value of a state s under policy π, denoted by $V^{\pi}(s)$, as given in (3), is the expected return when starting in state s and afterward following policy π

$$V^{\pi}(s) = E\left\{\sum_{k=0}^{\infty} \gamma^{k} r_{t+k+1} \mid s_t = s\right\}. \qquad (3)$$

Where, γ is discounted factor. Considering the Bellman equation in dynamic programming, the above equation can be written as:

$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{s'} P_{ss'}^{a}\left[R_{ss'}^{a} + \gamma V^{\pi}(s')\right]. \qquad (4)$$

In which $\pi(s,a)$ is the probability of taking action a in state s and $R^{a}_{ss'}$ is the expected value of reward r_{t+1}. Now, the optimal value and the best action in state s are calculated by taking the maximum value of (4) over action space $A(s)$ and building Bellman optimality equation as follows

$$V^{*}(s) = \max_{a \in A(s)} \sum_{s'} P_{ss'}^{a}\left[R_{ss'}^{a} + \gamma V^{*}(s')\right]. \qquad (5)$$

In the next section, it is explained that how the RL algorithm is used to.

IV. Control Strategy

The Dynamic Programming (DP) is one of the most applicable reinforcement learning methods with great computational expense in which the environment model is assumed to be perfect. Because of these limitations it is not very suitable for biomedical applications. Temporal Difference (TD) methods have an advantage over DP methods in that they do not require a model of the environment and also they are naturally implemented in an online, fully incremental fashion. As a consequence, TD approaches have a merit of being used on medical cases. We proposed The following Temporal Difference RL-based approach for solving the optimal control of chemotherapy drug dosage regimen.

```
Initialize Q(s,a) arbitrarily
  Repeat (for each episode):
    Initialize S
    Repeat (for each step of episode):
      Choose a from s using policy derived from Q (e-greedy)
      Take action a, observe r , and S(new)
      $Q_{K+1}= Q_K + a[r_{K+1}- Q_K]$
    Set S(new) to S
  Until S is terminal
```

Complete control procedure flowchart is illustrated in Fig. 2.

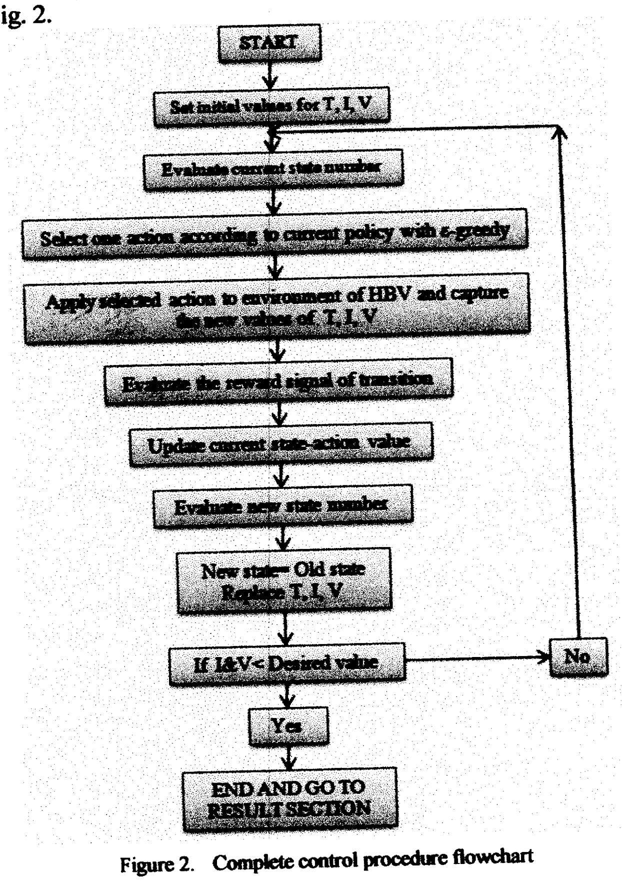
START
Set initial values for T, I, V
Evaluate current state number
Select one action according to current policy with ε-greedy
Apply selected action to environment of HBV and capture the new values of T, I, V
Evaluate the reward signal of transition
Update current state-action value
Evaluate new state number
New state= Old state
Replace T, I, V
If I&V< Desired value
No
Yes
END AND GO TO RESULT SECTION

Figure 2. Complete control procedure flowchart

In this work, three dimensional states are considered with respect to T, I, V. All the state variables are normalized to their initial values, and the whole interval of T, I, V-[0, 1]- is divided into 5 equal intervals. So the total number of the states is 125.

While action-value is implemented in this work, Policy table should be updated during the learning steps in each episode. The numerical reward signal is defined as $log(I_{old}/I_{new}) + log(V_{old}/V_{new})$ which guide agent through learning the optimal drug dosage schedule [14]. The main feature of this logarithmic signal is that it will be positive if the infected cells and free virions decrease after drug implementation and will be negative if they increase. Actions are divided into discrete normalized drug dosage {0.1,... 1}

V. Simulation Results

The following parameters and initial values are used for the simulation which we have taken from [2]: T_0 =5.5×10^7 cells, I_0 =1.1×10^7 cells, V_0 =6.3×10^9 copies/ml, $c = 0.67$, $h = 1$, $d = 3.78 \times 10^{-3}$, $\delta = 3.259d$, $\lambda = (2/3) \times 10^8\, d$, $R_0 = 1.33$, $p = \frac{cV_0\delta R_0}{\lambda(R_0-1)}$, $\beta = \frac{d\delta c R_0}{\lambda p}$.

Dynamic model equations (2) have been discretized with sampling time of T=1 [15]. In simulation we assume ε = 0.09, α = 0.6 in algorithm. The Terminal state is defined on when the number of infected cells (I) and free virions (V) population fall under the specific threshold. The learning loop has been iterated 500 times. The infected cells and free virions population during the treatment was obtained as Fig.3 and Fig. 4.

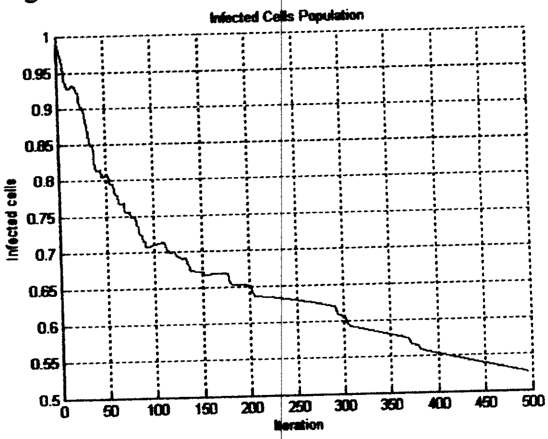

Figure 3. Infected cells population during treatment

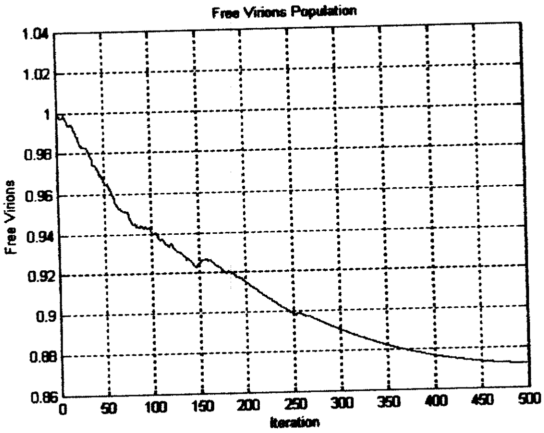

Figure 4. Free virions population during treatment

The control demonstrates that infected cells population and free virions are decreased during treatment. Uninfected target cells population is shown in Fig. 5.

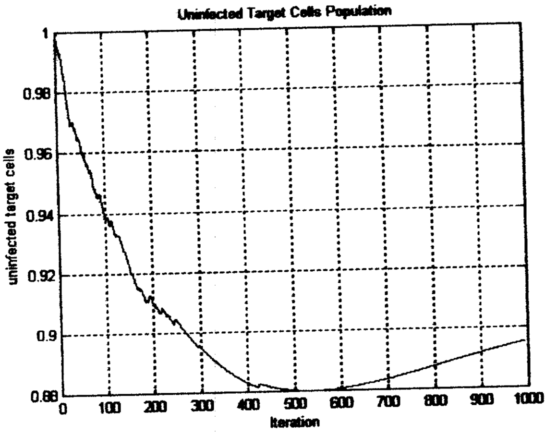

Figure 5. Uninfected target cells population during treatment

Fig. 5 shows that uninfected target cells population during treatment, at first decreases and then increases.

VI. Conclusion

In this work, we discuss an efficient numerical method based on RL to identify the best treatment strategy of hepatitis B viral (HBV) in order to block new infection and prevent viral production by using drug therapy with minimum side effects. Although the dynamic model of ordinary differential equations was implemented for the simulation of the dynamic environment and reward signal, showing the ability of RL algorithms in solving optimal control problems was the main purpose. Furthermore, as it was cleared, there isn't any relation between the usage of model and the learning algorithm. The flexibility and relative simplicity of this technique can lead to improved therapy for individuals, whose unique characteristics can be taken into account when establishing treatment protocols.

References

[1] S. Eikenberry, S. Hews, J. D. Nagy, and Y. Kuang, "The Dynamics of a delay model of hepatitis B virus infection with logistic hepatocyte growth," Mathematical Biosciences and engineering, Vol. 6, No. 2, April 2009.

[2] L. Min, Y. Su, and Y. Kuang, "Mathematical analysis of a Basic virus infection model with application to HBV Infection," Rocky Mountain J. Math, vol. 38, no. 5, 2008, pp. 1573-1585, doi:10.1216/RMJ-2008-38-5-1.

[3] E.H.C.J. Buster, and H.L.A. Janssen, "Antiviral treatment for chronic hepatitis B virus infection – immune modulation or viral suppression?," The Netherlands Journal of Medicine, vol. 64 , no. 6, pp. 175-185, june 2006.

[4] C. Castillo-Chavez, K. Cooke, W. Huang, and S. A. Levin, "On the role of long incubation periods in the dynamics of acquired immunodeficiency syndrome (AIDS)," Journal of Mathematical Biology, vol. 27, no. 4, pp. 373-398, 1989.

[5] F. Hollinger, and D.Lau, "Hepatitis B: the pathway to recovery through treatment," Gastroenterol. Clin. North Am. 35 (4): 895931, 2006.

[6] C. Lai, and M. Yuen, "The natural history and treatment of chronic hepatitis B: a critical evaluation of standard treatment criteria and end points," Ann. Intern. Med. 2007.

[7] K. Hattaf, M. Rachik, S.Saadi, and N. Yousfi, "Optimal control of treatment in a basic virus infection model," Applied Mathematical Sciences, vol. 3, no. 20, pp. 949-958, 2009.

[8] M. A. Nowak and R. M. May, "Viral dynamics," Oxford University Press, Oxford, 2000.

[9] S. Zuezem, J. M. Schmidt, and J.H. Lee, "Effect of inferformalt on the dynamics of Hepatitis C virus turnover in vivo," J. Hepatology, vol. 23, pp 366-371, 1996.

[10] S. Zuezem, R. A. de Man, and P. Honkoop, "Dynamics of Hepatitis B virus infection in vivo," J. Hepatology, vol. 27, pp 431-436, 1997.

[11] Y. Zheng, L. Min, Y. Ji, Y. Su, and Y. Kuang, "Global Stability of endemic equilibrium point of basic virus infection model with application to HBV infection," Springer-Verlag Berlin Heidelberg, vol. 23, January 2009, pp. 1221-1230, doi: 10.1007/s11424-010-8467-0

[12] R. S. Sutton, and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press, 1998.

[13] L. P. Kaelbling, M.L. Littman, and A.W. Moore, "Reinforcement learning: a survey," Journal of Artificial Intelligence, pp. 237–285, 1996.

[14] Sh. Yasini, M. B. Naghibi-Sistani, and A. Karimpour, "agent-based simulation for blood glucose control in diabetic patients," International Journal of Applied Science, Engineering and Technology 5:1, pp. 40-47, 2009.

[15] C. T. Chen, Linear system theory and design, 3rd ed., Oxford university press, 1999.