Reinforcement Learning Based Control of Tumor Growth with Chemotherapy

Amin Hassani, M.B Naghibi.S
Department of Control Engineering
Ferdowsi university of Mashhad
Mashhad, Iran
amin_hassani@ieee.org, mb-naghibi@um.ac.ir

***Abstract*— In this paper, optimal drug schedule for patients in progressive cancer phase who take the drug through infusion pump is obtained. An objective of control is reducing tumor cell numbers effectively while minimizing total amount of drug regimen. This is done because of the known serious side effects and major damages resulting from chemotherapy. Chemotherapy brings about weakness of the patient's immune system which is one of the most dangerous side effects. The optimal control problem is to design an effective drug-schedule to reduce the size of the tumors 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 year, especially in medical applications. Performance evaluation of the proposed algorithm has been performed by simulating on the mathematical model of tumor cells interacting with immune system. Simulation results show that a burst of treatment at the beginning is the best way to battle the tumor and constant decreasing the dosage of drug let the immune system to be reconstructed.**

***Keywords*—drug regimen, immun system, optimal control, reinforcement learning.**

I. INTRODUCTION

During recent years, diagnosis and treatment of cancer has become one of the favorite topics among the academic associations. Cancer is the uncontrolled growth of cells which occurs when certain cells keep dividing and forming more cells without the ability to stop this process [1]. Cancers are capable of spreading throughout the body by two mechanisms: invasion and metastasis. Invasion refers to the direct migration and penetration by cancer cells into neighboring tissues. Metastasis refers to the ability of cancer cells to penetrate into lymphatic and blood vessels, circulate through the bloodstream, and then invade normal tissues elsewhere in the body [2]. Cancer is thought to be caused by the interaction between genetic susceptibility and environmental toxins [1]. Even though there are a number of treatment options for cancer patients such as surgery, chemotherapy, immunotherapy, and radiotherapy, the life expectancy for the cancer patient will be diminished due to the disease and quite possibly the treatments as well. These treatment rules cannot in general provide a cure for cancer but may bring about remission that can later relapse. The effects of these treatments can vary from cancer to cancer and individual to individual, which further complicates the situation for effective eradication of cancer in any given patient [3].

Chemotherapy, in its most general sense, refers to treatment of disease by chemicals that kill cells, specifically those of micro-organisms or cancer. In popular usage, it will usually refer to antineoplastic drugs used to treat cancer or the combination of these drugs into a standardized treatment regimen. Chemotherapy types are categorized by the time applying which can be before, after, or instead of surgery or radiotherapy. Chemotherapy can be used before an operation (this is known as neo-adjuvant or primary chemotherapy) to shrink a cancer that is too large – or too attached to surrounding healthy tissue – to be removed easily during an operation. This can make removing the cancer easier during a later operation. Chemotherapy can be given after an operation (this is known as adjuvant chemotherapy) when all the visible cancer has been removed but there is a risk that some cancer cells, which are too small to be seen, may have been left behind. The aim is to destroy these cancer cells. Chemotherapy may also be given if a cancer cannot be completely removed during an operation. In this situation chemotherapy may not be able to cure the cancer but may shrink it and so reduce symptoms [4].

An important problem in chemotherapy is to design drug dosage regimens such that at the end of a treatment, the tumor burden is minimized. The importance is because virtually all chemotherapeutic regimens can cause depression of the immune system, often by paralyzing the bone marrow and leading to a decrease of white blood cells, red blood cells and platelets. The other side effects that may occur are Pain, Nausea and vomiting, Diarrhea or constipation, Anemia, Malnutrition ,Hair loss ,Memory loss . A proper dosage regimen has to balance the benefits of the treatment against the, often serious, toxic side effects. At present, treatments are developed and evaluated through empirical clinical trials. This process has led to a large number of patients being treated in sub-optimal ways.

This optimal control problem has solved using classical mathematical model-based methods. During the recent years, application of artificial intelligence approaches has been increased. Martin [5] used non-linear programming techniques. The results were improved by Bojkov et al. [6], who used an intuitive approach coupled with direct search procedure proposed in [7]. Direct search procedure combined with random numbers and contraction search region techniques was

used in [8]. Artificial intelligence methods were used by Tan et al. [9] who applied distributed evolutionary computing methods. In [10], Authors applied adaptive Neural Networks to solve the problem. While real modeling of such a complex human mechanism is very difficult, the model-free control approaches outperform other schemes. 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 chemotherapy 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

In this section, we are going to derive the mathematical model for simulation of environment and the. The general form for the evolution of a cell population, N(t), in absence of treatment with general growth rate $f(N(t),t)$ per cell is given by:

$$\frac{dN(t)}{dt} = N(t).f(N(t),t) \tag{1}$$

The per cell growth rate function, f, represents the net growth. The following three forms for the growth rate in (1) are typically used:

$$f(\mathrm{p},\mathrm{t}) = \begin{cases} \mathrm{C}_1 & , \text{ Exponential growth} \\ \mathrm{C}_2\left(1-\frac{\mathrm{P}}{\mathrm{K}}\right) & , \text{ Logistic growth} \\ \mathrm{C}_3 \log\left(\frac{\mathrm{K}}{\mathrm{P}}\right) & , \text{ Gompertz growth} \end{cases} \tag{2}$$

Where in (2), K and C_i , are positive constants that represent the leading order exponential growth and the carrying capacity, respectively.

In developing treatments models the role of pharmacokinetics, the way in which the drugs interact with the human body, is paramount. The present investigation assumes that the drug administrations and the effects of chemotherapy are instantaneous. These assumptions, although not realistic, are somewhat justified because the model time scale used (one day) is relatively large enough for the majority of effects of treatment to be affected in one time unit. Gyllenberg [11] proposed a mathematical model for the growth of solid tumors which employs quiescence as a mechanism to explain characteristic Gompertz-type growth curves. The model distinguishes between two types of cells within the tumor, proliferating and quiescent.

dePillis develop and analyze a mathematical model, in the form of a system of ordinary differential equations (ODEs), governing cancer growth on a cell population level with combination immune, vaccine, and chemotherapy treatments [12].

The following nonlinear model which is similar and a bit different to one used by Pillis is implemented in this paper:

$$\frac{dT}{dt} = aT(1-bT) - c_1NT - K_TMT \tag{3}$$

$$\frac{dN}{dt} = \alpha_1 - fN + g\frac{T}{h+T}N - pNT - K_NNM \tag{4}$$

$$\frac{dC}{dt} = \alpha_2 - \beta C - K_cMC \tag{5}$$

$$\frac{dM}{dt} = -\gamma M + V_M(t) \tag{6}$$

T(t) is tumor cell population, N(t) is total Natural Killers cell population, C(t) is number of circulating lymphocytes (or white blood cells), and M(t) is chemotherapy drug concentration in the bloodstream.

There are three differences between this model and the DePillis's one [12]. First, instead of normal cell population which was a criteria for patient's health, we choose natural killers cell population, N(t), whithout loss of generality. Second, for simplifying analyze in RL algorithm, we approximate exponential term of $-K_T(1-e^{-M})T$ with $-K_TMT$. third, we consider the effects of natural killers on equation (3), in the term $-c_1NT$, and also the natural activation of immune system in presence of tumor growth which was shown by $g\frac{T}{h+T}N$ in equation (4).

Tumor growth is assumed to be logistic, based on data gathered from immunodeficient mice [13]. We assume that circulating lymphocytes are generated at a constant rate, and that each cell has a natural lifespan. This gives us the term $\alpha_2 - \beta C$ in equation (5). Chemotherapy drug, after injection, will be eliminated from the body over time at a rate proportional to its concentration, giving an exponential decay $-\gamma M$. $V_M(t)$ is the external source of drug which is injected by infusion pump and indeed it is the only term that can be under our control. Constant system parameters which are listed in Table 1 have been extracted from several references within the [14].

TABLE I. CONSTANT SYSTEM PARAMETERS

Parameter	*Value*	*Unit*
a	4.31×10^{-2}	day^{-1}
b	1.02×10^{-14}	$cells^{-1}$
c_1	3.41×10^{-10}	$day^{-1}.cells^{-1}$
f	4.12×10^{-2}	day^{-1}
g	1.5×10^{-2}	day^{-1}
h	2.02×10^{1}	$cells^{2}$
K_c, K_N	6.00×10^{-1}	day^{-1}
K_T	8.00×10^{-1}	day^{-1}
p	2.00×10^{-11}	$day^{-1}.cells^{-1}$
α_1	1.2×10^{4}	$cell.day^{-1}$
α_2	7.50×10^{8}	$cell.day^{-1}$
β	1.20×10^{-2}	day^{-1}
γ	9.00×10^{-1}	day^{-1}

III. REINFORCEMENT LEARNING

Reinforcement Learning is the problem faced by an agent that must learn behavior through trial-and-error interactions with a dynamic environment which may has dynamic model or may not. In the standard reinforcement learning model, an agent is connected to its environment via perception and action, as depicted in Fig 1.

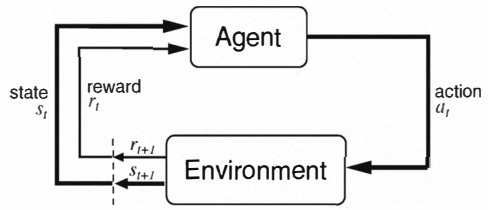

Figure 1. standard reinforcement learning structure.

Reinforcement learning techniques rely on feedback from the environment in order to learn. Feedback takes the form of a numerical reward signal, and guides the agent in developing its policy. The reinforcement learning agent's sole objective is to maximize the total reward it receives in the long run. The environment is usually modeled as an MDP, which is defined by a set of states, actions, transition probabilities, and expected values. Each action has a probability of being the selected action, defined by policy table which is extracted from a value function. The value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from the state, so the value function indicates what is good in long run. A greedy action is an action that has the greatest value. In order to learn, the agent must balance exploration and exploitation of the environment. During exploration, the agent tries non-greedy actions in hopes of improving its estimates of their values.

Value functions allow agent to compute how "good" it is to be in a given state. $V_\pi(S)$ is called the state-value function, and allows the agent to compute the expected reward of being in state s, and following policy π. $Q_\pi(S,a)$ is called the action-value function, and allows the agent to compute the expected reward of being in state s, taking action a, and thereafter following policy π. An optimal policy consists of the actions that lead to the greatest reward over time. The optimal state-value function is denoted by $V_\pi^*(S)$, and the optimal action-value function is denoted by $Q_\pi^*(S,a)$ [15].

In action-value methods the true (actual) value of action a is denoted by $Q^*(a)$, and the estimated value at the tth play by $Q_t(a)$. Also the true value of an action is the mean reward received when that action is selected. In control problems, the main idea of one popular approach is simply to learn action values, $Q_t(s,a)$, rather than state values, $V_t(s)$.

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 this 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 (ε-greedy)
Take action a, observe r , and S(new)
$$Q_{K+1} = Q_K + \alpha[r_{K+1} - Q_K]$$
Set S(new) to S
Until S is terminal

Complete control procedure flowchart is illustrated in Fig.2.

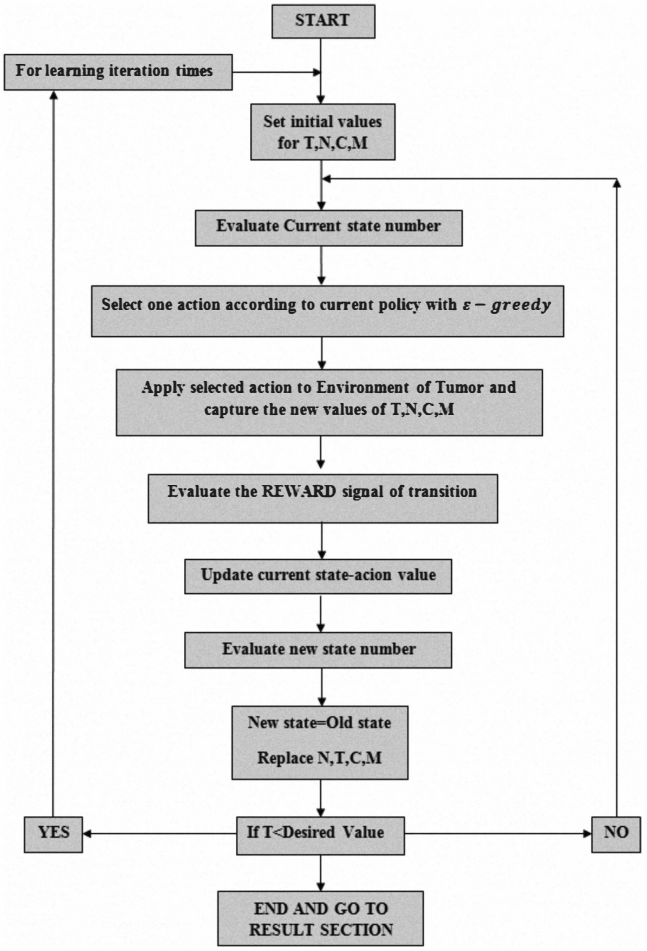

Figure 2. Complete control procedure flowchart.

In this work, two dimensional states are considered with respect to T, M. All the state variables are normalized to their initial values, and the whole interval of T, M-[0, 1]- is divided into 11 equal intervals in which each has a 0.1 length. So the total number of the states is more than 120. Decoding of state number is performed by structure in Fig.3.

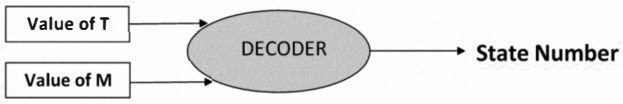

Figure 3. Decoding Structure

While action-value is implemented in this work, Policy table should be updated during the learning steps in each episode. The following formula is proposed for decoding sub-system:

$$State\ Number = 10 \times [T \times 10] + [M \times 10] + 1 \qquad (7)$$

The numerical reward signal is defined as $log\ (T_{old}/T_{new})$ which guide agent through learning the optimal drug dosage schedule. The main feature of this logarithmic signal is that it will be positive if the tumor cells decrease after drug implementation and will be negative if they increase. Actions are divided into discrete normalized drug dosage {0.1,…, 1} and also a constant rate decremental function, which both are the input of dynamic system, V_M.

V. SIMULATION RESULTS

Dynamic model equations (3) – (6) has been discretized with sampling time of T=1 [16,17]. In simulation we assume $\varepsilon = 0.01$, $\alpha = 0.6$ in algorithm.

The Terminal state is defined on when the number of Tumor cell population fall under the specific threshold. The learning loop has been iterated 200 times. The drug schedule and tumor cell population during the treatment was obtained as Fig. 4 and Fig.5 . Numerical results of natural killer cells and circulating lymphocytes are shown in Fig.6 and Fig.7.

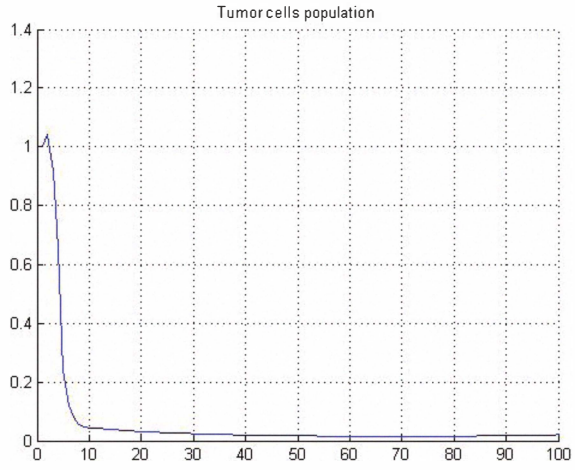

Figure 4. tumor cell population during the treatment.

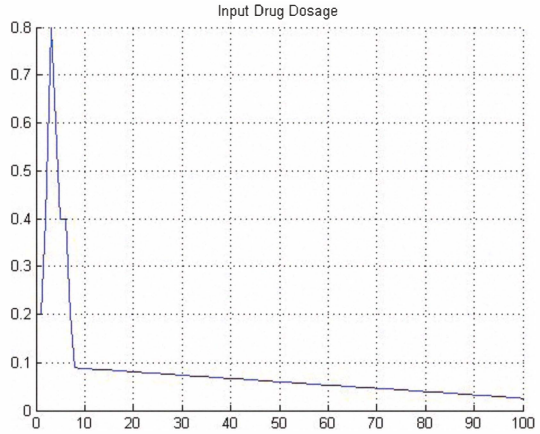

Figure 5. drug schedule during the treatment.

The control demonstrates that a burst of treatment at the beginning is the best way to fight the tumor. Drug concentration in bloodstream during the treatment is shown in Fig.8.

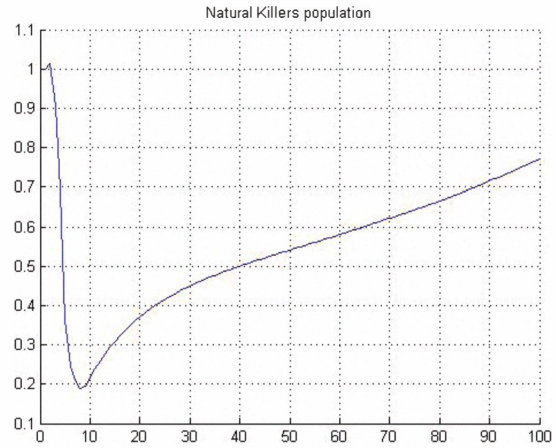

Figure 6. Natural killer cells during the treatment.

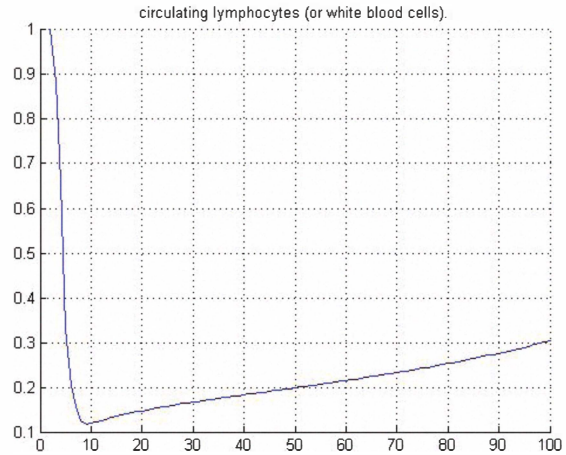

Figure 7. Circulating lymphocytes during the treatment

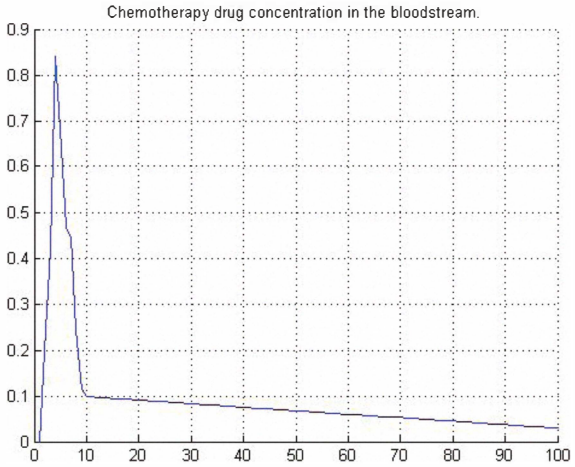

Figure 8. drug concentration in bloodstream during the treatment.

Since the numbers of immune system cells are never gone under 10% of their initial stages, we don't need to consider them in states. The constant decreasing scheme of drug dosage helps natural killer cells and circulating lymphocytes to reconstruct themselves. Although the population of tumor cell is much less than the amount which can be detectable clinically ($8\times10^2 cells$), the number of cells has not been reached zero yet. So if the treatment is cut in specific terminal times, the tumor cells begin to grow again. Clinically, this means that the specific and low dosage of drug should be prescribed to the whole life of patient.

VI. CONCLUSION

In this paper, optimal control of chemotherapy drug dosage for patients with progressive cancer has been studied. Although the dynamic model of ordinary differential equations was implemented for the simulation of 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. Generally, The RL algorithm need not the dynamic model of the system and indeed, this powerful feature can be applied to model-free environments. 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.

Another possible direction of work can be an extraction of optimal strategies directly from clinical data without relying on the identification of any accurate mathematical models, unlike approaches based on adaptive design. Tremendous potential of reinforcement learning can better be realized when it is applied to the problems in which even the relationship between actions and outcomes is not fully known.

REFERENCES

[1] "Healthbase Medical Tourism Resources Site." Available at http://www.healthbase.com/

[2] "National Cancer Institute." Available at http://www.cancer.gov/

[3] J. J. W. Westman Fabijonas Kern , B. R. Fabijonas , D. L. Kern "Compartmental Model for Cancer Evolution: Chemotherapy and Drug Resistance," *Association of College*, vol. 61, pp. 1023--29, 2001.

[4] " cancer e- library cancer databases " Available at http://www.cancermkn.scot.nhs.uk/

[5] R.B. Martin,"Optimal control drug scheduling of cancer chemotherapy ", *Automatica* , Vol .28,pp. 1113-1123,1992

[6] B. Bojkov, R. Hansel, and R. Luus, "Application of direct search optimization to optimal control problems", *Hung J. Ind. Chem.*, vol. 21, pp. 177-185, 1993.

[7] R. Luus, "Comments on dynamic optimization of batch reactors using adaptive stochastic algorithms", *Ind. Eng. Chem. Res.*, vol. 37, pp. 305-311.

[8] R. Luus, F . Harting, F. J. Keil, "Optimal drug scheduling of cancer chemotherapy by direct search optimization", *Hung J. Ind. Chem.*, vol. 23, pp. 55-58, 1995.

[9] K. C. Tan,E. F. Khor, J.Cai,C. M. Heng and T. H. Lee,"Automating the drug scheduling of cancer chemotherapy via evolutionary computation", *Art Intel Med,*Vol25,pp. 169-185 ,2002

[10] A. F. Floares, C. Cucu, M. Lazar, L., "Adaptive neural networks control of drug dosage regimens in cancer chemotherapy," *2003 IEEE Neural Networks Proceedings*, vol. 1, pp. 154- 159, 2003.

[11] M. Gyllenberg, "Quescence as an explanations of Gompertzian tumor growth", *Helsinki university*, 1989.

[12] L.G. dePillis, W.GU and A.E. Radunskaya, "Mixed immunotherapy and chemotherapy of tumors: Modeling, applications and biological interpretations", *Journal of Theoretical Biology*, 238(4):841–862, February 2006.

[13] A. Diefenbach, E.R. Jensen, A.M. Jamieson, and D. Raulet. Rae1 "H60 ligands of the NKG2D receptor stimulate tumor immunity", *Nature*, 413:165 171, September 2001.

[14] B. Hauser. Blood tests. Technical report, *International Waldenstrom s Macroglobulinemia Foundation*, January 2001. Available at `http://www.iwmf.com/`

[15] S. a. Barto, *Reinforcement Learning: An Introduction.* Cambridge, MA: The MIT Press, , 1998.

[16] Robert F. Stengel, Raffaele Ghigliazza," Optimal control of innate immune response ", Optim. Control Appl. Meth., 2002; 23: 91–104

[17] C.-T. Chen, Linear system theory and design, Third ed: Oxford university press, 1999.