Reinforcement Learning for Multifocal Tumour Targeting

Yi Hao[1], Zhijing Wang[2], Minghao Liu[1], Yifan Chen[1,3(✉)], and Yue Sun[1,4(✉)]

[1] School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China
yifan.chen@uestc.edu.cn,sunyuestc90@126.com
[2] Glasgow College, University of Electronic Science and Technology of China, Chengdu, China
[3] Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Chengdu, China
[4] School of Mechanical and Electrical Engineering, Chengdu University of Technology, Chengdu, China

Abstract. This paper implements a reinforcement learning (RL) targeting strategy for multifocal tumour lesions in the framework of computational nanobiosensing (CONA). Multi-tumours are promoted by the metastatic interaction between the surrounding tissues and the tumour suppressor. Nanorobots, regarded as computing agents, aim to search the multi-tumour lesions within the complicated vessel network. The Biological information gradient fields (BGFs) indicate the formation of the tumour microenvironment regulated by the nearby vessel network. By using reinforced learning and applying the knowledge of BGFs, this work achieves a higher tumour targeting efficiency than the previous work. The Markov and BGFs rewards are included in the total RL reward, in which the Markov reward is utilized for training nanorobots to find the path and avoid colliding with vessel walls, allowing them to learn the vascular network's topology, whereas the knowledge of BGFs incentive benefits faster convergence of the searching process. Therefore, this method enables the discovery of the path planning for the multi-tumour in a heterogeneous vessel network by combining viable vessel path planning with BGFs information.

Keywords: Reinforcement Learning · Biological Gradient Field · Markov Rewards · Multifocal Tumours

1 Introduction

Tumour is the primary disease that threatens humans health worldwide [1]. According to [1], more than fifty thousand people may have died because of cancer in 2020. Detecting tumours early and precisely delivering medication treatment could significantly reduce fatalities. However, this is a great challenge

Y. Chen et al. (Eds.): BICT 2023, LNICST 512, pp. 22–30, 2023.
https://doi.org/10.1007/978-3-031-43135-7_3

for traditional medical imaging techniques such as MRI and CT for early-stage tumours detection due to the limited facility resolution and acquired information.

In emerging nanotechnology, nanoparticles are used to enhance image effects based on physiochemical properties of *in vivo* environment, such as temperature, optical characteristic, tissue elasticity [2]. However, injected nanoparticles can only rely on human system circulation without external control, which results in low efficiency with 0.7% achieving tumour targeting [3]. This outcome is unsatisfactory in a practical situation, so overcoming the previous hurdle is crucial to moving forward nanomedicines into practice. Computational nanoparticle-mediated drug delivery has great potential in cancer diagnoses [4]. Therefore, we propose a computational nanobiosensing (CONA) framework as a "smart" strategy for targeting tumours, in which external manipulable nanorobots replace nanoparticles [5,6]. According to [7], pH and oxygen tension in the human body, nutrition distribution, enzymatic activity may become heterogeneous. The changes in biological gradient fields (BGFs) resulting from these factors can therefore be used to define the location of tumours, and the visualized tumour-triggered BGFs can be considered "objective functions". In these objective functions, the high-risk tissue is the domain, the targeted tumour sites are the optimal solutions, and the nanoparticles are the computing agents [6]. By this means, tumour targeting can be performed under the guidance of an external steering field to manipulate the internal nanorobots. Besides, a previous study showed improved efficiency and feasibility [5]. Conclusively, the knowledge of BGFs is essential in the CONA framework for tumour detection.

Indeed, the knowledge of BGFs is helpful to *in vivo* tumour detection; nonetheless, the vessel walls are failed to be considered [8]. As a result, the agents that adopt CONA strategies may adhere to the vessel walls during the searching process. Combining the BGFs with human vasculature during the tumour targeting process to avoid the obstacle caused by blood vessel walls can overcome the challenge mentioned above. The previous work mainly contributes to the detection of the single cancer [8], while we focus on the multimodal biodetection scenario in this paper. Different from the single tumour, multifocal tumours originated from a specific cellular cone and migrated to other lesions in the metastatic interaction process [9]. Based on our previous work in [8], reinforcement learning (RL) achieved considerable results with sphere BGFs function, which has only one optimal solution in the global domain. The current work focuses on the BGFs having optimal solutions using the RL algorithm with Q-learning, which means that nanorobots can search multifocal tumours in more complex BGFs with higher accuracy.

This paper is organized as follows. Section 2 describes the model of vascular network and BGFs. In Sect. 3, methodology including the Markov decision process, RL algorithm, and multifocal tumour search strategies are introduced. Following that, simulations with different scenarios are presented to illustrate the performance of the smart strategy in a more complex environment, and the results are shown in Sect. 4. Finally, Sect. 5 shows the main outcomes and the conclusion is drawn.

2 Models of Vascular Network and BGFs

2.1 Tumour Vascular Network

Due to the high demand for oxygen and nutrition during tumour growth, the vascular network around the tumour, high-density interconnected, is more complicated than normal vascular networks [10].

Additionally, tumour vascular networks have many unique properties, such as the tortuous vessels and wide range of avascular spaces in tumours, which are modelled by the fractal dimension [10,11]. Moreover, the percolation structure indicates the similarity with the local growth process [11]. Therefore, the invasion percolation algorithm is used to represent the growth process of the tumour vascular network.

When using the invasion percolation algorithm, first randomly assign uniformly distributed intensity values to each lattice point. Next, from any location as the starting point, the network gradually occupies the minimum lattice point adjacent to the current location and iteratively grows until the desired lattice occupancy is reached. The blood vessels are interconnected to all the occupied lattice points, while the blood flows in from the initial entry point and flows out from the specified outlet point. Then the network is pruned, and only the blood vessels of the non-zero blood flow part are retained to obtain the required tumour vascular network.

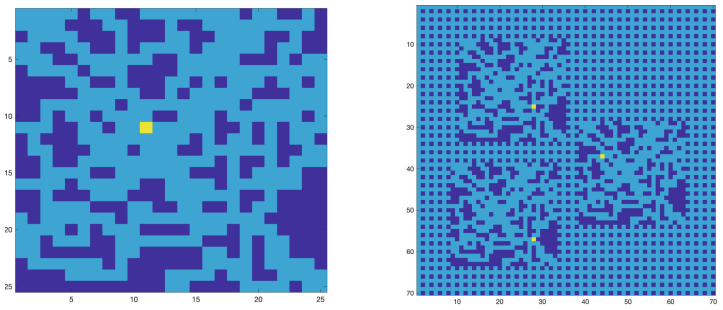
The occupancy of the grid indicates the fractal dimension of tumour vessels. According to [11], the lattice occupancy corresponding to fractal dimensions 1.6, 1.8, 1.9 and 2.0 are 40%, 60%, 80% and 10%, respectively. An example of a tumour vascular network with 77.00% occupancy is shown in Fig. 1(a), and the distance between two adjacent vessels is 50 μm.

For the multi-tumour vascular network, as shown in Fig. 1(b), the length per grid side is 100 μm, which represents the distance between blood vessels of healthy tissue, and the adjacent area with neovascularization is represented as three small squares.

2.2 Biological Gradient Fields

As mentioned above, oxygen tension and pH in the tumour microenvironment are heterogeneous. Besides, the distribution of glucose and other nutrients such as growth factors may be uneven or deficient [7]. These passive physical properties of tumour tissue like blood flow velocity and vascular network structure can be utilized to generate BGFs.

In the RL computation, BGFs in the high-risk tissues are transferred into representative objective functions used to evaluate the tumour detection performance, such as convergence, accuracy, and robustness of touch computing. Previously published studies have primarily focused on the experimental observation of the changes in the tumour microenvironment; however, it lacks the proper quantitative BGFs models. Therefore, this paper focuses on representative objective functions shown in Eq. (1), which could evocatively represent the



(a) (b)

Fig. 1. Vessel Framework Model (a) Simulated tumour vascular network using invasion percolation algorithm. The level of occupancy is 0.77 and the distance of tumour vascular is set to 50 μm. The light blue area represents the vascular spaces around the tumour, and the dark blue area indicates the vessels. (b) Represents the three tumour vascular network regions generated by invasive infiltration techniques, and the scope of the search space is 70×70. (Color figure online)

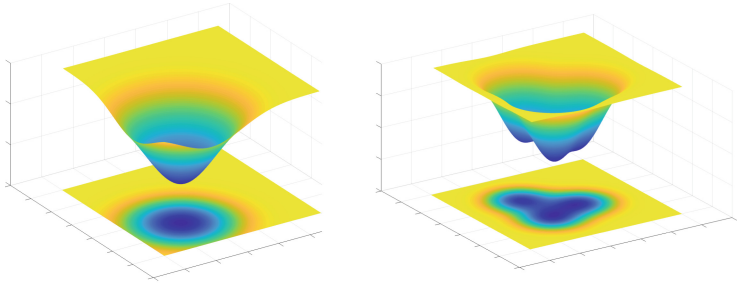
presence and absence of random fluctuations in the BGFs around tumours [9]. Besides, they are consistent with the qualitative observation results in the existing literature and can represent the BGFs in different change modes, which is used to verify the effectiveness of *in vivo* computing strategy in different BGFs. The expression of BGFs is shown in Eq. (2), and the corresponding landscape is shown in Fig. 2. It is worth noting that this BGF landscape is different from the traditional optimization problems in the concept test function [9]. Since traditional optimization problems require comprehensive features in the functions, while this model only focuses on the biological characteristics of high-risk tumour areas, thus, this paper proposed in vivo computing concept has a significant difference from the traditional optimization problem. The detail of this model is illustrated and discussed in Sect. 4.

$$U_T(x,y) = U_{T1}(x,y) + U_{T2}(x,y) + U_{T3}(x,y) \tag{1}$$

$$U_{T1}(x,y) = 1 - exp((-(x-56)^2 - (y-27)^2)/300) \tag{1-1}$$

$$U_{T2}(x,y) = 1 - exp((-(x-36)^2 - (y-43)^2)/300) \tag{1-2}$$

$$U_{T3}(x,y) = 1 - exp((-(x-24)^2 - (y-27)^2)/300) \tag{1-3}$$



(a) (b)

Fig. 2. The landscape of tumour BGFs, and the minimal values 0 and maximum values 1 are the range of the functions, and the tumour is located in the optimal minimal values. (a) The landscape of singular tumour with coordinates (70,70). (b) The landscape of three tumours, with location at coordinates $(56, 27), (36, 43)$, and $(24, 27)$.

3 Multifocal Tumour Searching Model

The searching scenario is a 70×70 grid-like vascular network with three tumours, where the nanorobots find the shortest path to the tumour locations. This optimisation problem can be mapped into a reward function as part of a Markov decision process (MDP), which the RL algorithm can solve.

In conventional mathematical optimization, continuous niche technology can effectively solve multi-solution computing problems. As an overall training process, the tumours are examined one by one [12]. Firstly, we use an optimization algorithm to find the optimal global solution. Secondly, the objective function is modified in the optimal solution location to update the knowledge of BGFs. Thirdly, re-run the optimization algorithm with updated BGFs to find the next tumour. Repeat the above process until all tumours are found. It is worth noting that Q-learning is adopted for each step of the optimization algorithm.

3.1 Markov Decision Process

The optimization problem, as mentioned before, can be regarded as an MDP, which is defined through the tuple (S, A, R) with state-space S, action space A, and reward function R [13].

The state at mission time t in the vascular network is given by $s_t = (p_t, p_t) \in \mathbb{R}^2$, which is the location of the nanorobot. There are five actions that are allowed nanorobots to take:

$$A = \{up, down, left, right, idle\} \tag{2}$$

If the nanorobot collides with the wall of a vessel, then it goes to idle mode; otherwise, it takes a step forward in one of the four directions specified in (2).

The reward function maps the state-actions to a real-valued reward, i.e., $S \times A \rightarrow R$. The mission goals, R, consist of the following components:

– Markov reward, which is utilized to avoid colliding with the vessel wall during the tumour searching process.

$$r_{Markov} = \begin{cases} 2, & \text{one tumour found} \\ 3, & \text{two tumours found} \\ 4, & \text{three tumours found} \\ 0.01, & \text{tumour not found \&} \\ & \text{vessel wall not reached} \\ -1, & \text{vessel wall reached} \end{cases} \tag{3}$$

– BGFs reward, which is used to speed up the convergence and is defined by (1).

To sum up, the total reward, divided into two parts, the Markov reward and the BGFs reward of the multifocal tumours, is defined as:

$$r = r_{Markov} + \beta \times r_{BGFs} \tag{4}$$

where $\beta (0 \leqslant \beta \leqslant 100)$ regulates the weight of the Markov reward and BGFs reward in the model.

3.2 Q-Learning

Q-learning is one of the basic algorithms of reinforcement learning [14]. A model-free learning method allows agents to select optimal actions using experienced action sequences in a Markov environment. The interaction between the agent and the multi-tumour environment can be regarded as an MDP, a critical assumption in Q-learning. The agent could perform the tasks in iterations, firstly observe the state $s_t \in S$ and then perform an action $a_t \in A$ at time t and subsequently receive a reward $r(s_t, a_t) \in R$ to the agent from the environment, and finally restart in a new state $s_t + 1$. The behavioural policy of the agent is to obtain the highest reward, which means finding the optimal paths from the starting point to the destination. A probabilistic policy $\pi(a|s)$ is a distribution over actions based on the state such that $\pi : S \times A \rightarrow R$. It reduces to $\pi(a|s)$ in the predictable situation, resulting in $\pi : S \rightarrow A$.

To learn the policy π, Q-Learning updates the state-action value function by iterating, given as:

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}[R_t | s_t = s, s_t = a] \tag{5}$$

which denotes an expectation of the discounted cumulative return R_t from the current state s_t up to a terminal state at time T given by

$$R_t = \sum_{k-t}^{T} \gamma^{k-t} r(s_k, a_k) \tag{6}$$

with $\gamma \in [0,1]$ being the discount factor, balancing the priority of current and future rewards. s_t and a_t are simplified to s and a, s_{t+1} and a_{t+1} are simplified

to s' and a' for the sake of clarity. We use the Bellman equation to train our model [15]:

$$Q_{k+1}(s,a) = r_s^a + \gamma Q_k(s', \pi(s')) \tag{7}$$

where k is the number of steps and r_s^a is the same as r in (3). In practice, a learning rate of $0 \leq \alpha \leq 1$ is used to keep the agent from becoming stuck in a locally optional solution:

$$Q_{k+1}(s,a) = Q_k(s,a) + \alpha(r_s^a + \gamma Q_k(s', \pi(s) - Q_k(s,a)) \tag{8}$$

4 Simulation and Results

4.1 Simulation Setup

Each new iteration is conducted in a 70×70 discretized vascular network with $10\,\mu\text{m} \times 10\,\mu\text{m}$ in each grid cell size. Considering the limited lifespan of the nanorobot, we set the maximum exploration steps to 3000. During the simulation process, training steps are used to determine the performance of nanorobots in different scenarios.

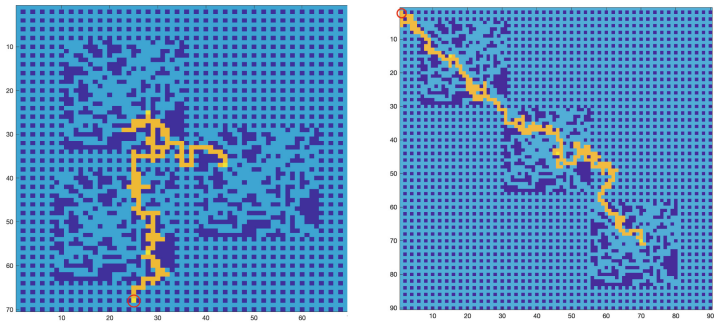
4.2 Simulation Results

Figure 3(a) shows the trajectories of nanorobots, which are represented by the colour orange. As shown in the figure, the movement of nanorobots is coordinated towards the maximum-gradient direction estimated in the Markov reward and BGFs. The method of Q-learning can detect all three tumour centres with high accuracy and efficiency. It is shown that the simulation result is correlated with the tumour locations and the weight β, these factors will be introduced in the following paragraphs.

Q-learning is assessed regarding its tumour targeting efficiency with different multi-tumour locations considering the practical situation. Figure 3(b) presents different tumour locations and different nanorobot injection locations, which verifies the robustness of this tumour research strategy.

4.3 Parameter Optimization in Different Scenarios

According to Eq. (3), the value of the Markov reward is between -1 and 1. The maximum value of the function is set to 0, and the minimum value is set to -1. As a result, when the weight β is set to 1, the Markov and BGFs rewards will be of the same size. The number of training steps for the convergent model decreases as the weight β goes from -3 to -1.5. As shown in Fig. 4, the choice of β value is critical, and both much large and much small cannot achieve the goal. Taking [15,15], [44,46], and [70,70] tumour locations, for example, the training steps reached the low bound when β was equal to -1.5 because the reward was applied to avoid attaching to the vessel wall and the gradient reward was used to speed up the convergence. Furthermore, given the restricted resource of computing capacity, setting the incentive weight $\log \beta$ to -1.5 is the best option.



(a) (b)

Fig. 3. The simulation demonstrated multifocal tumour targeting. (a) and (b) are simulations of the search process in the environment with different tumour locations. The yellow dots and red circles represent the tumour locations and injection point of the agent, and the trajectory of the agent is marked by orange. (Color figure online)

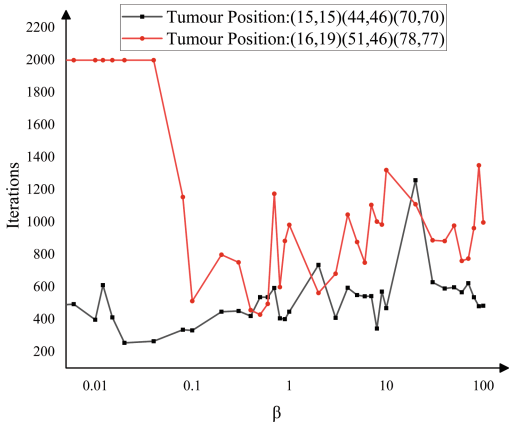


Fig. 4. Training steps (iteration) related to different β values, two different sets of three-tumour locations are tested.

5 Conclusion

This paper investigated the novel CONA strategy using a reinforcement learning algorithm to target multi-tumours. The main contributions of this paper are as follows:

- This paper introduces an RL algorithm for multi-tumour detection, which shows a better performance with fewer iterations than the brute-force searching strategy.
- The model uses the reward of BGFs in multifocal tumour targeting processes.
- For different scenarios, the nanorobots are injected in different locations to search multi-tumours simultaneously.

However, the algorithm is unstable, and its simulation results depend in part on the location of the tumour. Therefore, further research could use other reinforcement learning algorithms to select optimal actions, or focus on nanorobot swarms with RL algorithms. In addition, considering the complex microenvironment of the human body, multi-tumour BGFs may not be the result of direct superposition of BGFs, so more accurate vascular models and physiological environment models are needed.

References

1. Siegel, R.L., et al.: Colorectal cancer statistics, 2020. CA Cancer J. Clin. **70**(3), 145–164 (2020). https://doi.org/10.3322/caac.21601
2. Chen, H., Zhang, W., Zhu, G., Xie, J., Chen, X.: Rethinking cancer nanotheranostics. Nat. Rev. Mater. **2**(7), Art. no. 17024 (2017)
3. Wilhelm, S., Tavares, A.J., Dai, Q., Ohta, S., Chan, W.C.W.: Analysis of nanoparticle delivery to tumours. Nat. Rev. Mater. **1**(5), 16014 (2016)
4. Seidi, K., Neubauer, H.A., Moriggl, R., Jahanban-Esfahlan, R., Javaheri, T.: Tumour target amplification: implications for nano drug delivery systems. J. Control. Release **275**, 142–161 (2018)
5. Shi, S., Sharifi, N., Cheang, U.K., Chen, Y.: Perspective: computational nanobiosensing. IEEE Trans. Nanobiosci. **19**(2), 267–269 (2020). https://doi.org/10.1109/TNB.2019.2956470
6. Shi, S., Chen, Y., Yao, X.: In vivo computing strategies for tumour sensitization and targeting. IEEE Trans. Cybern. **52**(6), 4970–4980 (2020). https://doi.org/10.1109/TCYB.2020.3025859
7. Kwon, E.J., Lo, J.H., Bhatia, S.N.: Smart nanosystems: bio-inspired technologies that interact with the host environment. Proc. Natl. Acad. Sci. **112**(47), 201508522 (2015)
8. Liu, L., Sun, Y., Shi, S., Chen, Y.: Smart tumour targeting by reinforcement learning. In: 2021 IEEE International Conference on Nano/Molecular Medicine & Engineering (NANOMED), Virtual, 15–18 November 2021 (2021)
9. Shi, S., Chen, Y., Yao, X.: NGA-inspired nanorobots-assisted detection of multifocal cancer. IEEE Trans. Cybern. (2020). https://doi.org/10.1109/TCYB.2020.3024868
10. Gazit, Y., et al.: Fractal characteristics of tumour vascular architecture during tumour growth and regression. Microcirculation **4**(4), 395–402 (1997). https://doi.org/10.3109/10739689709146803
11. Baish, J.W., et al.: Role of tumour vascular architecture in nutrient and drug delivery: an invasion percolation-based network model. Microvasc. Res. **51**(3), 327–46 (1996). https://doi.org/10.1006/mvre.1996.0031
12. Beasley, D., Bull, D.R., Martin, R.R.: A sequential niche technique for multimodal function optimization. Evol. Comput. **1**, 101–125 (1993)
13. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. Artif. Intell. **101**(1–2), 99–134 (1998)
14. Watkins, C.: Technical note: Q-learning. Mach. Learn. **8** (1992)
15. Ford, R., Delbert, F.: A simple algorithm for finding maximal network flows and an application to the Hitchcock problem. Rand Corporation (1955)