Data analytics and visualization for inspecting cancers and genes

Victor Chang[1]

Received: 21 July 2017 / Revised: 9 August 2017 / Accepted: 30 August 2017 /
Published online: 9 September 2017

Abstract This paper describes our latest research in data analytics and visualization for bioinformatics and healthcare. Each year many patients have suffered cancers. Analytics and visualization can help to simulate the development of malignant tumors and help identify weak spots of tumor for treatment, inspect malignant tumors in general and inspect whether genes have cancerous cells. Related literature, technologies, simulation results with explanation, performance evaluation and comparisons with other work have been discussed in details. We can process training data with a low completion time to achieve simulations of malignant tumors and genes to inspect their status, as well as the querying the output data within seconds. Our malignant tumor and gene simulation can achieve 360 degrees for an inspection of cancerous presence. We conclude that data analytics and visualization can provide effective and efficient healthcare research and also other type of interdisciplinary research.

Keywords Analytics and visualization for bioinformatics and healthcare · Simulation for malignant tumor inspection · Simulation for gene inspection · Big data processing an analytics for bio-science

1 Introduction

This paper presents the latest data analytics and visualization for bioinformatics and healthcare. In human bodies, there are many complex but self-regulated mechanism that every cell, organ, gene, protein and body defense system are on function every second. Although it is happening now and then, some of these facts cannot be easily comprehend, such as how the cell divisions can take place by itself with orders; cell growth can be maintained; how cells can detect foreign invaders and prompts self-defense mechanism

✉ Victor Chang
Victor.Chang@xjtlu.edu.cn

[1] Xi'an Jiaotong-Liverpool University, Suzhou, China

and how each body organ can always operate smoothly every second. In this paper, the emphasis is to study how cells and genes can become malignant and harmful to human bodies. While these concepts can be complicated and require a significant period to fully understand, the use of analytics and visualization can be extremely useful, since all complex concepts and sequence of events (taking place every second) can be presented by analytics [11]. This allows the following benefits. First, complex biological simulations can represent details of what can happen or what has happened every second [28]; Second, it can make explanations easier. If there is anything that cannot be understood, learners can repeat analytics and simulations again. Educators can also explain any areas that learners have in doubt. This can provide a two-way process to improve the quality and experience of learning [12, 21]. But most importantly, if patients are in health hazards such as battling with cancer, any advancement to know the causes of malignant tumors, identifying weak spots of malignant tumors, inspecting any cancerous genes and tumors involved and finally making more accurate judgement on the treatment based on all the information received earlier [2, 10, 19]. However, all these work can be challenging and complex. There is a need to identify the well-beings of the patients, make progress for health research and present complex work that can be viewed, checked and understood by the stakeholders involved, such as analytics and visualization for bioinformatics and healthcare. There are reports of effectiveness and positive impacts to medical researchers, students and healthcare communities after adopting analytics and visualization [11, 12, 21, 28]. In this paper, we demonstrate proofs-of-concept on simulating malignant tumor, inspecting tumours and genes with cancerous presence, as well as malignant tumor's weak spots via analytics and visualization. The Improved MapReduce Framework has been used to process training data, present the outputs, discuss the implications and undertake performance evaluation. The breakdown of this paper is as follows. Section 2 presents the related literature and Section 3 discusses the MapReduce Framework. Section 4 and 5 presents simulations for cancerous tumor and their inspection. Similarly, Section 6 follows the same principle to attempt detect any cancerous cells on two genes prone to live cancer and breast/ovarian cancer. Section 7 demonstrates results of our performance evaluation and Section 8 presents topics of discussion. Finally, Section 9 sums up Conclusion and Future Work.

2 Literature

This section describes literature with two sub-sections. The first sub-section describes formulas for cancer growth models and the second sub-section presents the technologies behind to make biological analytics happen.

2.1 Growth models for malignant tumors

While there are many biological and physiological science subjects to study, the focus of this paper is to investigate any forms of malignant and harmful impacts, which include the growth of malignant tumors, or cancer cells. There are three commonly known mathematical models. The first model is an "Exponential Growth Model", meaning the growth can be expanded in the exponential rate [16, 22]. This takes account into the general growth rate, but does not measure the size of tumour in details.

Exponential Growth Model:

$$n(t) = N_0 \exp(\lambda t) \tag{1}$$

The second model is a "Gompertz Model" that describes tumor dynamics based on a time-series model [16, 22]. Growth is at the fastest at the beginning and eventually slows down and becomes the slowest when the tumor has reached its size limit. In formulas (2), N∞ is the plateau cell number at large values of r. The initial tumor growth rate is represented by b. The limitation of this model is that a single set of growth parameters is insufficient to model tumor, and either multiple sets will be used, or blend this model with another model to represent the real life situations experienced by patients.

Gompertz Model:

$$n(t) = N_0 \exp\left[\ln\left(\frac{N\infty}{N_0}[1-\exp(-bt)]\right)\right] \tag{2}$$

The third model is called "Model based on metabolic considerations" [16, 22]. This is closer to real life situations in patients, since the tumor growth is also dependent on patients' health conditions and metabolism. The cells follow a pattern based on the patients' metabolism for cell division. To facilitate this, energy conservation will be required and updated in the formula. Formula 3 show that regardless of the masses, all malignant cells follow the same growth pattern. Additionally, variable B is the amount of energy required for the cells at rest. Variables Bc and Nc refer to metabolic rate for individual cell and number of cells from the patients, and NcBc is the energy required to maintain existing malignant tissues. Ec is referred as the energy to create new tissue within a cell. The total mass of a patient is m, which is calculated by the mass of a cell and the total number of cells, m = mcNc. With a number of variable rearrangement, formula 4 is the final output calculating the malignant tumor over a period of time. When it reaches a certain threshold, computer simulations will trigger alarm to patients to take actions as soon as possible.

Model based on metabolic considerations:

$$B = \underset{c}{\Sigma}\left[N_c B_c + E_c \frac{dN_c}{dt}\right] \tag{3}$$

$$\frac{dm}{dt} = am^{\frac{3}{4}} - bm \tag{4}$$

2.2 Supporting technologies

This section describes elate technologies to support the biological analytics service a follows.

- Artificial intelligence and machine learning: Advanced algorithms will be used behind the scenes. Selected methods such as MapReduce and some extents of deep learning can be used jointly to ensure that the optimization of data processing with analytics can be achieved [4, 15].
- Databases: A large quantity of training data can be stored in the databases before the simulations and can be all queried and be ready for data processing [23].

- Mathematical modelling and Application Program Interfaces (APIs): All mathematical models with different variables can be blended effectively in the form of APIs for calculations and computation [5, 26].
- Data privacy: Training data similar to the real tumors will be used so that there are no any data privacy and anonymization issues [8].

For this paper, we have developed a unique MapReduce Framework to accelerate the data processing and ensure the completion of our outputs can be presented by analytics, as described in Section 3.

3 System architecture and MapReduce framework

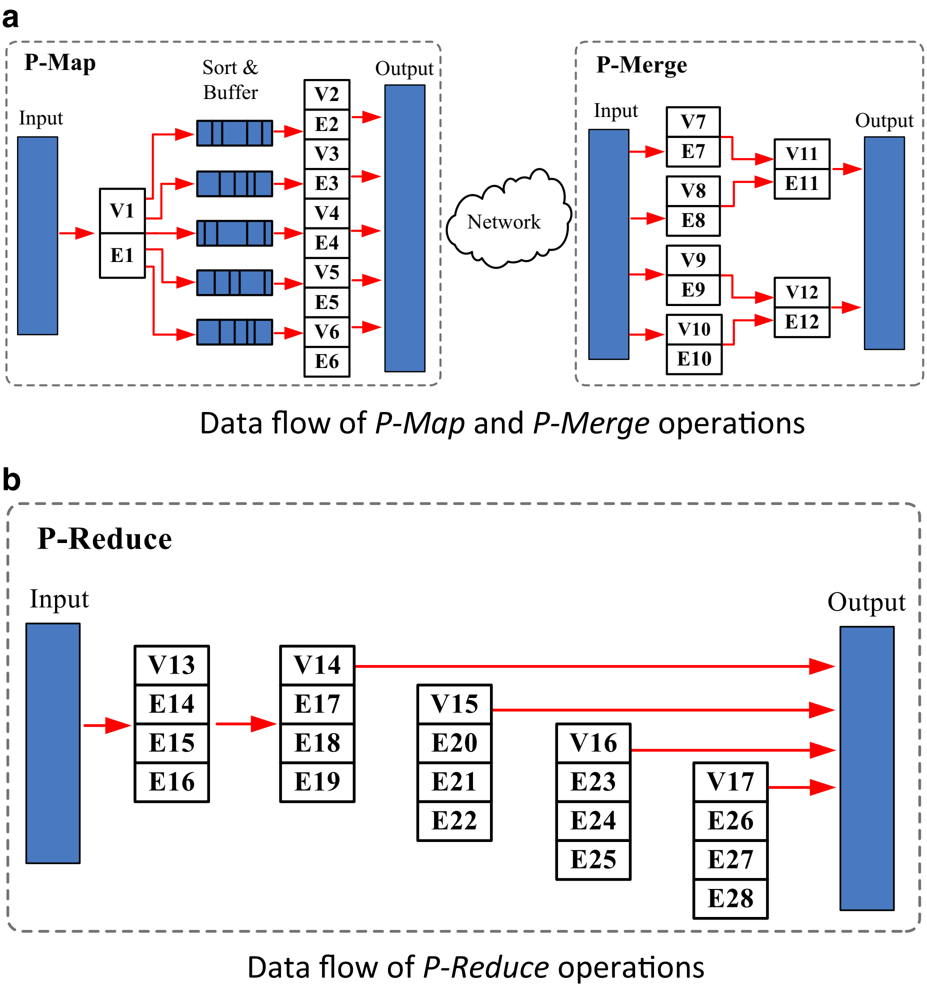
Understanding how malignant and harmful body functions have been triggered is an important aspect. While the emphasis for this paper is to present the reality in a form of analytics and visualization, so that doctors can understand their patients' extent of health conditions and make better judgement in treatment prescription. To facilitate this, modern computing infrastructure in [6] has been built and used to simulate complex biological functions and growth. The resources include the high-end Cloud Computing, with 300 GHz multi-core CPUs, 10 Bps network speed, full virtualization Cloud Virtual Machine (VM) Center, up to 100 nodes per VM to be used for testing and simulation, previously it was used for weather simulations and forecasting [6].

For this medical research, all the cancer cells have been simulated thousands of times based on patients' different types of tumors. It has been simulated repetitively many times, so that the final outcome can be close to a better accuracy. As a result, all these thousands of simulations can become very valuable training data as described in this section for simulating high-performance malignant tumors.

To facilitate a faster and more smooth big data processing, improved MapReduce functions have been used. To further divide the work, it has split MapReduce into three steps: Partitioning-Map (P-Map), Partitioning-Merge (P-Merge) and Partitioning-Reduce (R-Reduce). An additional command, P-Query, can be used after executing P-Map, P-Merge and P-Reduce once to retrieve all the processed data directly for the next experiments or next services. The main goal of the P-Map, P-Merge and P-Reduce is to enhance the data processing and reducing the processing time. MapReduce framework can present inputs as **<key, value>,** which can be summed up as <Ky, Vy>, in which y is the iteration number of MapReduce service. The proposed P-MapReduce can get <Ky, Vy > in each parallel pipeline. Each pipeline can determine the iteration number of inputs and outputs. By using partitioning approach, it can perform optimization of data processing and each task is divided as follows.

- P-Map: which maps all data into the grouped clusters ready processing
- P-Merge: All the processed outputs are collected;
- P-Reduce: All the outputs can be condensed into one or fewer clusters.
- P-Query: It can be executed once all the three functions that have been executed once. The command can directly retrieve data

To demonstrate the split of tasks and data processing, Fig. 1a and b shows the data flow of our proposed optimization algorithm. *P-Map* will split all the inputs from *V* and *E* (from our



Data flow of *P-Map* and *P-Merge* operations

Data flow of *P-Reduce* operations

Fig. 1 Data flow of the optimization algorithm

Procedure 1, Partitioning algorithm) into both sorting and splitting process. The outputs for V and E in *P-Map* reshuffled the outcomes of $V$and E sequentially into pairs (e.g., V_2 and E_2 and V_3 and E_3) until the end of the processing. The maximum number of pairs in each processing operation can reach up to 256 pairs. Subsequently, the output of *P-Map* will become the input of *P-Merge*, and same or identified groups of outputs are classified together. For illustration, the inputs of *P-Merge* become $\{V_7, E_7\}$ all the ways to$\{V_{10}, E_{10}\}$, all of which are further processed and summed as $\{V_{11}, E_{11}\}$ and $\{V_{12}, E_{12}\}$.

The uniqueness of this approach can allow two commands to be simultaneously executed: P-Map and P-Merge which then trigger the P-Reduce function to take further actions. To demonstrate how to process a large quantity of data, Table 1 shows a list of commands. After reading all the data, the three-step process of P-Map, P-Merge and P-Reduce can be executed. Then the status of all data processing can be updated in real-time. After getting the outputs of the data processing, the final output can be returned to the big data platform to confirm the completion of data processing tasks. After successfully completing the entire data processing,

Table 1 Command steps to accelerate data processing and update results

read(*data*);
p-map(*data*);
p-merge(*data*);
p-reduce(*data*);
update(*data*);
return(*data*);

the outputs have been stored in the database. Thus, the P-Query can be used to query the results directly, as shown in Table 2. The advantage is that it allows researchers to query for the outputs without the need to re-run the entire experiment. Performance evaluation for using P-Map, P-Merge and P-Reduce, as well as P-Query will be presented in Section 7.

4 Malignant cancer cell simulations and Anaytics

This section presents different outputs of malignant cancer cell simulation, after following the steps in Section 2 for formulation and Section 3 for simulation and data processing. Since the emphasis of this paper is on data analytics and visualization, the outputs of malignant tumor can be presented followed by discussion. Figure 2a shows the rapid growth of a tumor, which has experienced all three phases presented in three mathematical models in Section 2. While it is still growing, it can absorb the energy and sugar from patients' body. Figure 2b shows the malignant tumor has reached its equilibrium for the maximum growth due to the limitation of its size and ability to obtain more energy. At this level, it has reached a stable status and may not join other malignant tumor before metastasis (movement of malignant of cancer cells from one part of the body to another) happen. For the treatment, the stable status can be treated with radiotherapy, chemotherapy and/or surgeries for removal. For strong radiotherapy and chemotherapy before surgeries, a malignant tumor may react differently [19]. If it is responding positively, meaning it is on the way to be destroyed, it will look like Fig. 2c. It means either the malignant tumor is about to be split or become unstable. However, the body mechanism and defense can be weaken as a result of strong radiotherapy and chemotherapy that kill all cells at the area of cancer, regardless of good and bad cells in the patient.

It is often that malignant tumors in "full strength" can be resistant to chemotherapy and radiotherapy. Repeated treatments are commonly practiced, but it can make patients' health very weak and vulnerable [20]. Figure 3 shows a malignant tumor in full strength by analytics and visualization. The entire brightness shows the extent the strength of the malignant tumor, the brighter the more active and energy hungry as it has become. The red spots represent regions of active energy production and the blue spot indicate the place to store its energy and reproduction. The red spots have higher metabolism rates and active malignant activities, and thus it is harder to break. With analytics and visualization, medical researchers can magnify those regions and identify where they can "attack" the malignant tumor without harming patient's body.

Table 2 Command steps to query outputs directly after the successful completion of the data processing

read(*data*);
p-query(*data*);
update(*data*);
return(*data*);

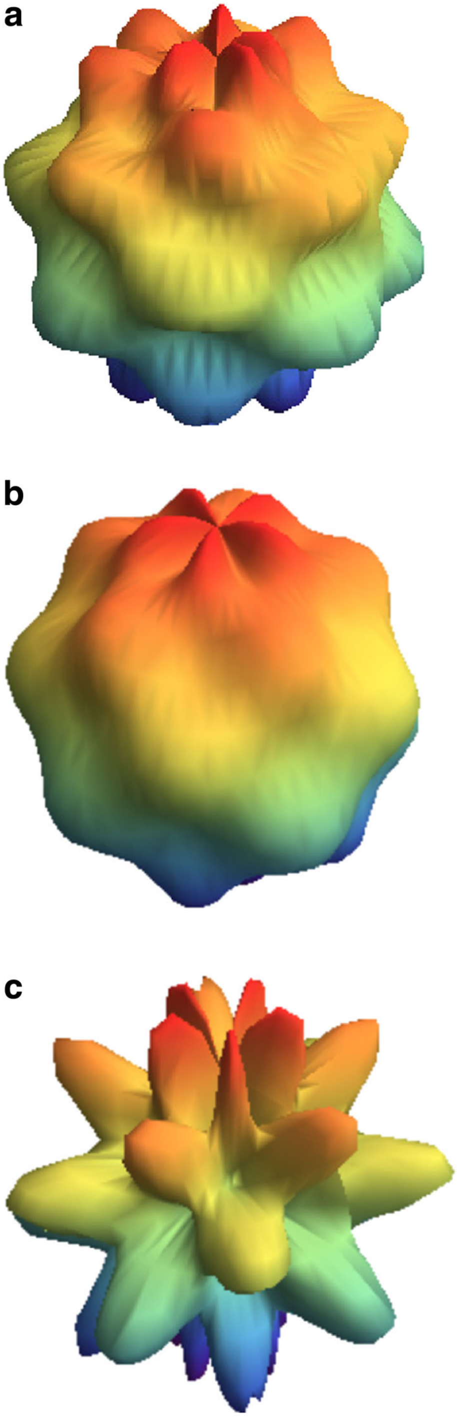
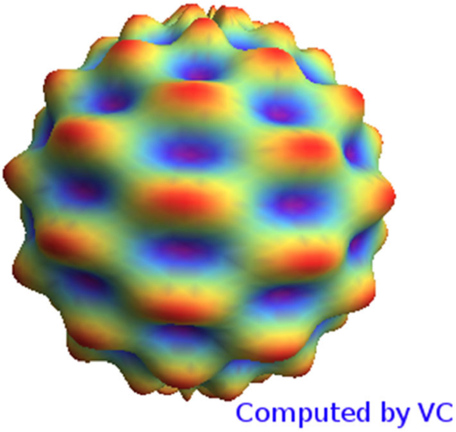


Fig. 2 **a** Malignant tumor in rapid growth. **b** Malignant tumor in stable status after growth. **c** Malignant tumor to be destroyed and become unstable

Fig. 3 Tumor cell with active and inactive regions



5 Malignant cancer cell inspection

The concept shown in Fig. 3 is in an ideal situation that there are plenty of weak regions equally distributed to be attacked. In real life situations, inspecting each tumor cell can be a different experience. There is an important step before identifying where to attack a malignant tumor. The darker the color, the less active section that the malignant tumor is, and thus it is better for doctors to attack the malignant tumor. Figure 4a shows an inspection for a malignant tumor, in which the lower region is more inactive. It is also a common practice to double check several times and rotate that tumor computationally. Figure 4b confirms the original lower region of the tumor is inactive, derived from a different angle of views. Figure 4c shows the focus of the weak areas on the tumor cell to be attack. Those darker areas on the concave region can be the spot. Attacking the malignant cell in its weaker spot can ensure a higher rate of success, since it will be easier for anti-bodies or medical injection to be more active to attack. Destroying the malignant cells can then obtain a better outcome by not harming and destroying all good and bad cells in the patient's body.

6 Gene inspection for genes on the way to become malignant for liver and breast/ovarian cancers

Each gene is important for the human bodies. There are particularly a few genes with a very crucial role to our health. Hence, analytics and visualization should also study their impacts to human health. A paper has identified that there are six genes that can cause liver cancer if malfunctioned [13]. One such a gene is called S100P and the output data has been processed and stored in our databases, after the first completion of all data processing. This can be achieved by executing commands in Table 2, including P-Query. Figure 5a shows the S100P gene in inspection to see any cancerous cells identified as black spots. The gene is in an active and healthy status Figure 5b then shows that after 360 degrees of rotation inspection of S100P gene, there is no black spot and patient's sample can be clear of liver cancer cells. As a result,

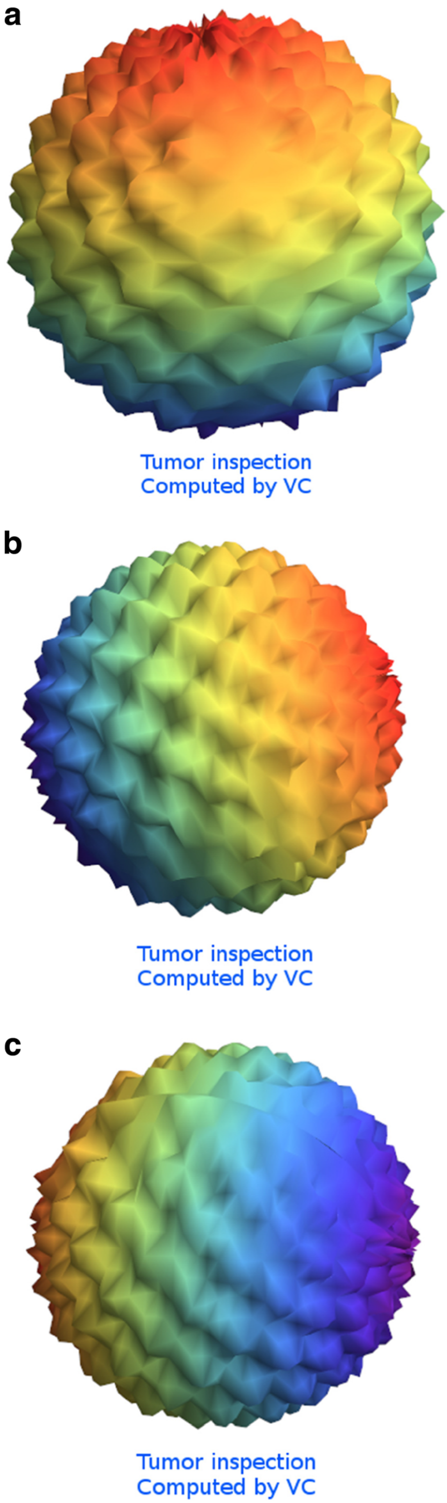


Fig. 4 **a** A tumor in inspection. **b** A tumor in rotation for a further inspection. **c** Identifying possible spot to attack on the tumor cell

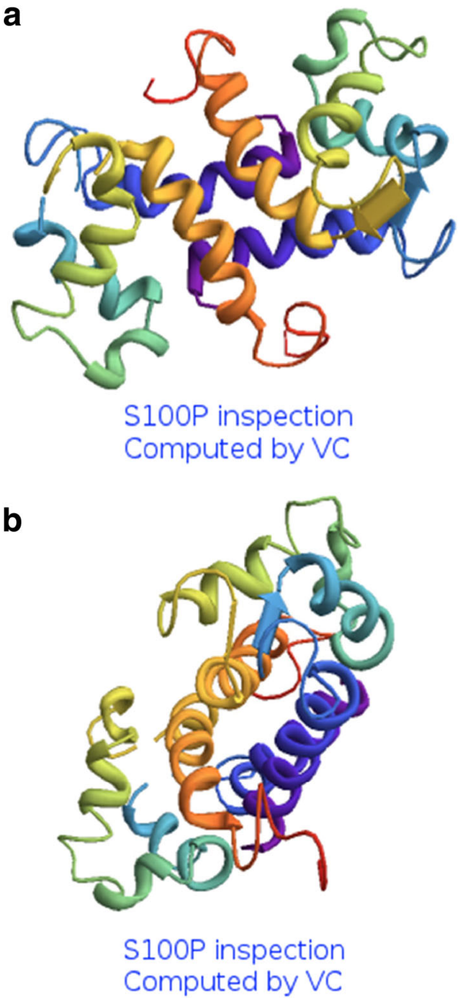
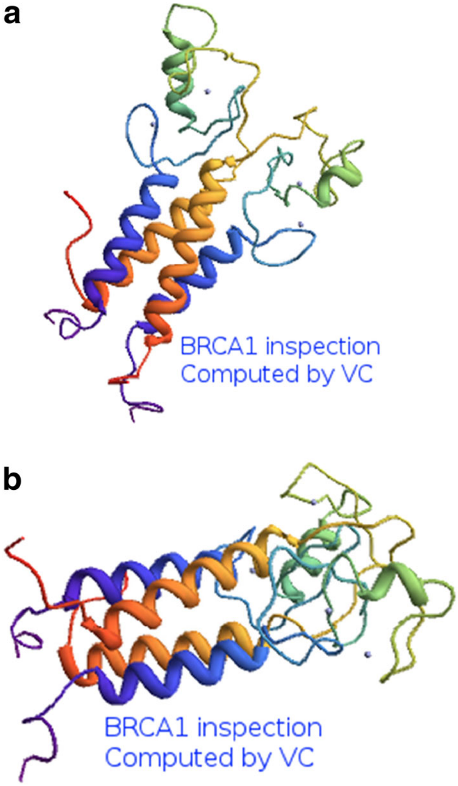


Fig. 5 **a** S100P inspection. **b** S100P inspection confrming no cancerous (black) spots

the benefits of adopting analytics can be positive and immediate since results can be known in a short period of time.

Breast or ovarian cancers are common among women. Reasons may vary such as genetic reason, disorder in food and life style and so on, but out of all these possibilities, genetic disorder is a main factor. Thus, it is important to inspect breast and ovarian causing genes such as BRCA1, which can cause either or both cancers [27]. The approach is similar to other data analysis, where training data would be obtained after going through sample taking, simulations and anonymization. After running all data processing, the outputs can be queried directly. The aim is to check any cancerous black spots. Figure 6a shows that BRCA1 gene is in inspection. There are a few tiny spots. After rotating 360 degrees, Fig. 6b shows that those tiny spots are clear of cancer activities to confirm medical diagnosis made to this patient earlier. In this example, analytics and visualization can be very effective in diagnosis and inspection.

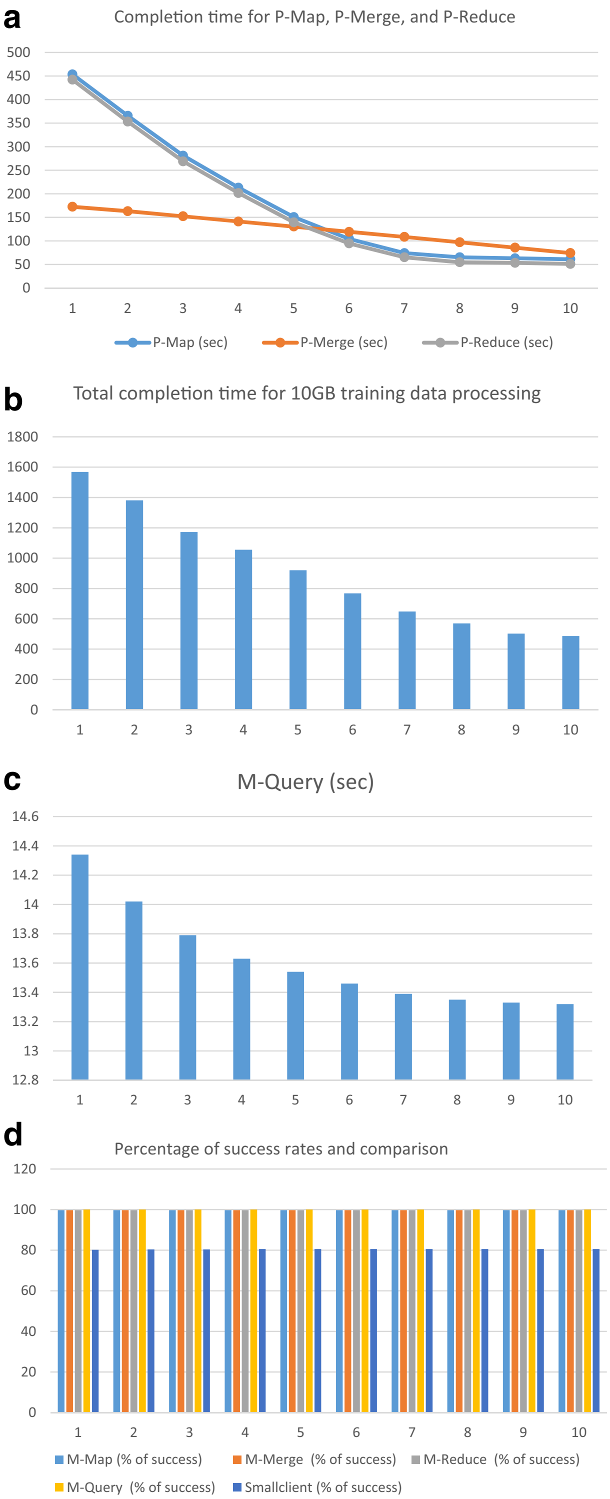
Fig. 6 **a** BRCA1 inspection. **b** S100P inspection confrming no cancerous (black) spots



7 Performance evaluation

It is crucial to keep an excellent performance while using thousands of training data at the same time to ensure a stable and reliable high performance can be guaranteed throughout using simulations and computational analysis. Section 3 has described the techniques behind processing a large quantity of training data. Each training data of 10GB has been processed by P-Map, P-Merge and P-Reduce functions respectively. Each experiment was then undertaken five times and the mean completion time was recorded for each successful experiment. The focus is to identify the duration of completion time, success rate of data processing for each of P-Map, P-Merge, P-Reduce and P-Query functions.

Figure 7a shows the completion time for executing P-Map, P-Merge and P-Reduce between 10 and 100 nodes. Completion time for P-Map and P-Reduce can be further reduced when the number of nodes has increased up to 60 nodes. Further this point, the impact on the completion time is seen less. However, P-Merge stays consistently reduced in the completion time for the increasing number of nodes. All the completion time for successful P-Map, P-Merge and P-Reduce functions have been added up. Figure 7b shows the total completion time, which has been consistently reduced when the number of nodes has increased up to 100. In summary of all these experiments, P-Map, P-Merge and P-Reduce have been useful to process thousands of tumor training data and can provide reliable resources and outputs to simulate malignant cells and proteins. In comparison with similar work in [7, 27], completion time is short and competitive, after processing a large quantity of data. But more importantly, it is not only


a
Completion time for P-Map, P-Merge, and P-Reduce
500
450
400
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
P-Map (sec)
P-Merge (sec)
P-Reduce (sec)
b
Total completion time for 10GB training data processing
1800
1600
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
c
M-Query (sec)
14.6
14.4
14.2
14
13.8
13.6
13.4
13.2
13
12.8
1
2
3
4
5
6
7
8
9
10
d
Percentage of success rates and comparison
120
100
80
60
40
20
0
1
2
3
4
5
6
7
8
9
10
M-Map (% of success)
M-Merge (% of success)
M-Reduce (% of success)
M-Query (% of success)
Smallclient (% of success)

◀ **Fig. 7** **a** Performance evaluation between P-Map, P-Merge and P-Reduce (1 unit = 10 nodes). **b** Total completion time for 10GB (P-Map, P-Merge, P-Reduce) training data processing (1 unit = 10 nodes). **c** Total completion time for 10GB training (P-Query, after data processing first time) data processing (1 unit = 10 nodes). **d** Percentage of success rates and comparison between P-Map, P-Merge, P-Reduce and SmallClient (1 unit = 10 nodes)

the numerical computational outputs but analytics and visualization that can help doctors and scientists to understand the outputs easily and quickly, since the weak spots to attack malignant tumor can be identified and decisions for the better actions can be made effectively and almost instantaneously after showing up the final output in analytics. Figure 7c shows the time to perform P-Query after the first completion of all data processing. The completion time is all under 15 s for all nodes since all output data can be directly queried out from the system. The impacts to the completion time with the increased nodes are not significant.

We also have our previous wok of processing big data by SmallClient [25]. Comparison between P-Map, P-Merge, P-Reduce and SmallClient have been undertaken to measure the success rate. Experiments were performed more than 20 times to ensure a fair comparison of success rate for data processing. Each P-Map, P-Merge and P-Reduce can get as high as 99.7% regardless of the number of nodes and the SmallClient can only get between 80.2% and 80.6% success rate for data processing. P-Query always stays 100% to retrieve processed outputs successfully.

8 Discussion

Analytics and visualization are useful for healthcare and bioinformatics research with the following reasons. First, it can reduce costs of running very expensive experiments and medical equipment to support. Simulations can be done many times and outputs can be computed or queried at any time [1, 11]. Second, it can show to the patients with different cancer status of their treatment, similarly it can help medical doctors and researchers more precisely know the weak spots of malignant tumor to begin their treatment [9]. Third, the general public can understand the complex biological science more with ease. It can also better educate to the public about awareness of health [14]. Fourth, analytics and visualization can also be used for education and training medical students. It can also save lecturers time to repetitively explain the development of malignant tumors [1, 9, 12, 14]. In general, implementing analytics and visualization for healthcare and bioinformatics research can help more people understand biological science and complex health problems such as cancer, AIDS and other major diseases.

Comparison with other work with similar approaches is as follows. Schatz [24] uses CloudBurst to investigate DNA sequencing. MapReduce is the technique behind data processing. There are experimental results with performance evaluation but there is no any presentation of visualized results. Matsunaga et al. [17] use CloudBlast to integrate MapReduce and Visualization. They show a parallelism but the emphasis is on the DNA sequencing with experimental results, without visualization presented in their outputs. McKenna [18] develop a Genome Analysis Toolkit that combines MapReduce framework to analyse DNA sequencing data. They have shown a few diagrams that require high computational resources. Approaches illustrated by this paper do not consume very high throughput and consumption of energy and can demonstrate results as visualization within minutes as demonstrated by results in Section 7.

Our research contributions are as follows. First, we provide unique visual analytics with data processing techniques. Second, mathematical models can be blended successful with our improved MapReduce framework to process training data originally derived from the real data. Third, this research is involved with real-world challenge of processing, interpreting and visualizing a large scale data, particularly inspecting malignant tumors and genes in details and 360 degrees of rotation. Last, the completion time for all data processing is considered low with 99.7% success rate in data processing in this research. Compared to other work [3, 7, 9, 14, 27], which report that data processing, interpretation and visualization can be difficult and prolong tasks. Our work can demonstrate effective data processing, analytics and interpretation as a handy process.

9 Conclusion and future work

This paper demonstrates proofs-of-concept of simulating complex biological science in the form of analytics and visualization. The emphasis is to investigate malignant tumors, understand its development, inspect it and identify weak spots, where that treatments can be focused to get a greater chance of success. Training data of 10GB each was generated as a result of previous work involved with thousands of sample taking, data reading and updating as well as simulation. To make this happen, improved MapReduce framework, dividing the big data processing into P-Map, P-Merge and P-Reduce. P-Query can be used to query the final outputs directly after the first full completion of data processing. Outputs for malignant tumor inspection, growth and weak spot checks, together with outputs of gene inspection checking cancer presence, have been demonstrated. Performance evaluation was undertaken. The individual and total completion time for P-Map, P-Merge and P-Reduce between 10 and 100 nodes have been recorded with explanations. Similarly, P-Query time was taken all under 15 s to query output processed data.

Our wok has demonstrated a cost-effective, useful and effective way of developing analytics and visualization to influence bioinformatics and healthcare research. Analytics and visualization can be used in other disciplines to create a greater positive impact and influence, particularly if the “apps” or analytics can demonstrate updates in real-time, which will be the focus of our future work.

References

1. Agostinelli S, Allison J, Amako KA, Apostolakis J, Araujo H, Arce P et al (2003) GEANT4—a simulation toolkit. Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment 506(3):250–303
2. Boik J (2001) Natural compounds in cancer therapy. Oregon Medical Press, Princeton
3. Bostock M, Ogievetsky V, Heer J (2011) D^3 data-driven documents. IEEE Trans Vis Comput Graph 17(12): 2301–2309
4. Cao J, Cui H, Shi H, Jiao L (2016) Big data: a parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce. PLoS One 11(6):e0157551
5. Chang V (2014) The business intelligence as a service in the cloud. Futur Gener Comput Syst 37:512–534
6. Chang V (2017) Towards data analysis for weather cloud computing. Knowl-Based Syst 127:29–45
7. Christopher R, Dhiman A, Fox J, Gendelman R, Haberitcher T, Kagle D et al (2004) Data-driven computer simulation of human cancer cell. Ann N Y Acad Sci 1020(1):132–153
8. Cios KJ, Moore GW (2002) Uniqueness of medical data mining. Artif Intell Med 26(1):1–24
9. Collett D (2015) Modelling survival data in medical research. CRC Press, Boca Raton
10. Cuomo MI (2012) A world without cancer: the making of a new cure and the real promise of prevention. Rodale, Emmaus
11. Green TM, Ribarsky W, Fisher B (2008) Visual analytics for complex concepts using a human cognition model. In: Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on. IEEE, p 91–98
12. Green TM, Ribarsky W, Fisher B (2009) Building and applying a human cognition model for visual analytics. Inf Vis 8(1):1–13
13. Hu J, Sharma S, Gao Z, Chang V (2017) Gene-based collaborative filtering using recommender system. Comput Electr Eng. https://doi.org/10.1016/j.compeleceng.2017.04.010

14. Huang D, Tory M, Aseniero BA, Bartram L, Bateman S, Carpendale S et al (2015) Personal visualization and personal visual analytics. IEEE Trans Vis Comput Graph 21(3):420–433
15. Lin J, Dyer C (2010) Data-intensive text processing with MapReduce. Synth Lect Hum Lang Technol 3(1):1–177
16. Marušić M (1996) Mathematical models of tumor growth. *Mathematical*. Communications 1(2):175–188
17. Matsunaga A, Tsugawa M, Fortes J (2008) Cloudblast: combining mapreduce and virtualization on distributed resources for bioinformatics applications. In: eScience, 2008. eScience'08. IEEE Fourth International Conference on. IEEE, p 222–229
18. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303
19. Pienta KJ (ed) (2012) *Diagnosis and Treatment of Genitourinary Malignancies* (Vol. 88). Springer Science & Business Media, New York
20. Priestman T (2012) Cancer chemotherapy in clinical practice. Springer Science & Business Media, New York
21. Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press, Cambridge
22. Roose T, Chapman SJ, Maini PK (2007) Mathematical models of avascular tumor growth. SIAM Rev 49(2):179–208
23. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11(9):647
24. Schatz MC (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25(11):1363–1369
25. Siddiqa A, Karim A, Chang V (2017) SmallClient for big data: an indexing framework towards fast data retrieval. Clust Comput 20(2):1193–1208
26. Suresh P, Hsu SH, Reklaitis GV, Venkatasubramanian V (2010) OntoMODEL: ontological mathematical modeling knowledge management in pharmaceutical product development, 2: applications. Ind Eng Chem Res 49(17):7768–7781
27. Venkitaraman AR (2002) Cancer susceptibility and the functions of BRCA1 and BRCA2. Cell 108(2):171–182
28. Wilkinson DJ (2011) Stochastic modelling for systems biology. CRC Press, Boca Raton

Prof. Victor Chang is an Associate Professor (Reader) and Director of PhD at IBSS, Xi'an Jiaotong-Liverpool University, Suzhou, China, after working as a Senior Lecturer at Leeds Beckett University, UK, for 3.5 years. Within 4 years, he completed Ph.D. (CS, Southampton) and PGCert (Higher Education, Fellow, Greenwich) while working for several projects at the same time. Before becoming an academic, he has achieved 97% on average in 27 IT certifications. He won a European Award on Cloud Migration in 2011, IEEE Outstanding Service Award in 2015, best papers in 2012 and 2015, the 2016 European award: Best Project in Research, Outstanding Young Scientist award in 2017 and numerous awards since 2012. He is widely regarded as a leading expert on Big Data/Cloud/IoT/ security. He is a visiting scholar/PhD examiner at several universities, an Editor-in-Chief of IJOCI & OJBD journals, Editor of FGCS, Associate Editor of TII, founding chair of two international workshops since Year 2014 and founding Conference Chair of IoTBDS http://www.iotbd.org and COMPLEXIS http://www.complexis.org since Year 2016. He was involved in different projects worth more than £12.5 million in Europe and Asia. He has published 3 books as sole authors and the editor of 2 books on Cloud Computing and related technologies. He has given and will give 14 keynotes at international conferences.