ISSN 2644-3228 (online)

Artificial Intelligence in *Cancer*

Artif Intell Cancer 2021 October 28; 2(5): 51-78

Published by Baishideng Publishing Group Inc

A I C

Artificial Intelligence in Cancer

Contents

Bimonthly Volume 2 Number 5 October 28, 2021

OPINION REVIEW

MINIREVIEWS

Artificial Intelligence in Cancer

Bimonthly Volume 2 Number 5 October 28, 2021

Contents

ABOUT COVER

Editorial Board Member of *Artificial Intelligence in Cancer*, Anca Maria Cimpean, MD, PhD, Associate Professor, Department of Histology, Victor Babes University of Medicine and Pharmacy, Timisoara 300041, Romania. ancacimpean1972@yahoo.com

AIMS AND SCOPE

The primary aim of *Artificial Intelligence in Cancer* (*AIC*, *Artif Intell Cancer*) is to provide scholars and readers from various fields of artificial intelligence in cancer with a platform to publish high-quality basic and clinical research articles and communicate their research findings online.

AIC mainly publishes articles reporting research results obtained in the field of artificial intelligence in cancer and covering a wide range of topics, including artificial intelligence in bone oncology, breast cancer, gastrointestinal cancer, genitourinary cancer, gynecological cancer, head and neck cancer, hematologic malignancy, lung cancer, lymphoma and myeloma, pediatric oncology, and urologic oncology.

INDEXING/ABSTRACTING

There is currently no indexing.

RESPONSIBLE EDITORS FOR THIS ISSUE

NAME OF JOURNAL
Artificial Intelligence in Cancer

ISSN
ISSN 2644-3228 (online)

LAUNCH DATE
June 28, 2020

FREQUENCY
Bimonthly

EDITORS-IN-CHIEF
Mujib Ullah, Cedric Coulouarn, Massoud Mirshahi

EDITORIAL BOARD MEMBERS
https://www.wjgnet.com/2644-3228/editorialboard.htm

PUBLICATION DATE
October 28, 2021

INSTRUCTIONS TO AUTHORS
https://www.wjgnet.com/bpg/gerinfo/204

GUIDELINES FOR ETHICS DOCUMENTS
https://www.wjgnet.com/bpg/GerInfo/287

GUIDELINES FOR NON-NATIVE SPEAKERS OF ENGLISH
https://www.wjgnet.com/bpg/gerinfo/240

PUBLICATION ETHICS
https://www.wjgnet.com/bpg/GerInfo/288

PUBLICATION MISCONDUCT
https://www.wjgnet.com/bpg/gerinfo/208

ARTICLE PROCESSING CHARGE
https://www.wjgnet.com/bpg/gerinfo/242

STEPS FOR SUBMITTING MANUSCRIPTS
https://www.wjgnet.com/bpg/GerInfo/239

ONLINE SUBMISSION
https://www.f6publishing.com

Artificial Intelligence in Cancer

Submit a Manuscript: https://www.f6publishing.com

Artif Intell Cancer 2021 October 28; 2(5): 51-59

DOI: 10.35713/aic.v2.i5.51

ISSN 2644-3228 (online)

OPINION REVIEW

Artificial neural network for prediction of acute kidney injury after liver transplantation for cirrhosis and hepatocellular carcinoma

Luis Cesar Bredt, Luis Alberto Batista Peres

ORCID number: Luis Cesar Bredt 0000-0002-8487-1790; Luis Alberto Batista Peres 0000-0001-5863-6720.

Author contributions: Bredt LC and Peres LAB contributed equally to this review article; all authors equally contributed to this paper with conception and design of the study, literature review and analysis, drafting and critical revision and editing, and final approval of the final version.

Conflict-of-interest statement: No potential conflicts of interest. No financial support.

Manuscript source: Invited manuscript

Specialty type: Transplantation

Luis Cesar Bredt, Department of Surgical Oncology and General Surgery, University Hospital of Western Paraná, State University of Western Paraná, Cascavel 85819-110, Paraná, Brazil

Luis Alberto Batista Peres, Department of Nephrology, University Hospital of Western Paraná, State University of Western Paraná, Cascavel 85819-110, Paraná, Brazil

Corresponding author: Luis Cesar Bredt, FRCS (Gen Surg), MD, PhD, Full Professor, Surgeon, Department of Surgical Oncology and General Surgery, University Hospital of Western Paraná, State University of Western Paraná, Tancredo Neves Avenue, Cascavel 85819-110, Paraná, Brazil. lcbredt@gmail.com

Abstract

Acute kidney injury (AKI) has serious consequences on the prognosis of patients undergoing liver transplantation (LT) for liver cancer and cirrhosis. Artificial neural network (ANN) has recently been proposed as a useful tool in many fields in the setting of solid organ transplantation and surgical oncology, where patient prognosis depends on a multidimensional and nonlinear relationship between variables pertaining to the surgical procedure, the donor (graft characteristics), and the recipient comorbidities. In the specific case of LT, ANN models have been developed mainly to predict survival in patients with cirrhosis, to assess the best donor-to-recipient match during allocation processes, and to foresee postoperative complications and outcomes. This is a specific opinion review on the role of ANN in the prediction of AKI after LT for liver cancer and cirrhosis, highlighting potential strengths of the method to forecast this serious postoperative complication.

Key Words: Liver transplantation; Acute kidney injury; Artificial neural network; Prediction; Hepatocellular carcinoma; Postoperative

Core Tip: This opinion review aims to explore the potential benefits of artificial neural network models in predicting the occurrence of acute kidney injury in the postoperative period of liver transplantation for cirrhosis and hepatocellular carcinoma.

Country/Territory of origin: Brazil

Peer-review report's scientific quality classification
Grade A (Excellent): 0
Grade B (Very good): B
Grade C (Good): 0
Grade D (Fair): 0
Grade E (Poor): 0

Received: October 12, 2021
Peer-review started: October 12, 2021
First decision: October 20, 2021
Revised: October 22, 2021
Accepted: October 27, 2021
Article in press: October 27, 2021
Published online: October 28, 2021

P-Reviewer: Song B
S-Editor: Wang JL
L-Editor: A
P-Editor: Wang JL

Citation: Bredt LC, Peres LAB. Artificial neural network for prediction of acute kidney injury after liver transplantation for cirrhosis and hepatocellular carcinoma. *Artif Intell Cancer* 2021; 2(5): 51-59
URL: https://www.wjgnet.com/2644-3228/full/v2/i5/51.htm
DOI: https://dx.doi.org/10.35713/aic.v2.i5.51

INTRODUCTION

Liver transplantation (LT) is the best treatment option for patients with early stages of hepatocellular carcinoma (HCC) and cirrhosis[1-4]. Mainly, the use of LT depends on maintaining a balance between patient-specific survival benefit, the availability of alternative treatment modalities[5,6], and the equitable distribution of donor organs[5,7-12]. Current selection criteria aim to avoid transplant futility by excluding patients at a high risk of tumor recurrence[10,11]. Selecting patients with HCC within Milan criteria has been shown to provide excellent patient outcomes[13-15].

Among the possible complications related to LT for cirrhosis and HCC, acute kidney injury (AKI) is a common complication, with extremely variable reported incidence rates (4% to 94%)[16-22], and is associated with several immediate complications, including volume overload, metabolic acidosis and electrolyte disturbances. Although most patients eventually recover after an episode of AKI, many patients may not return to baseline renal function, and the occurrence of AKI has been shown to be an independent risk factor for the development of chronic kidney disease and death, as well as for the reduction of survival rates of liver receptors[23]. In addition, transplant patients who require temporary renal replacement therapy (RRT) have a prolonged hospital stay, with subsequent need for more resources and higher costs related to LT[24].

Artificial neural network (ANN) is commonly used to solve complex problems, where the behavior of variables is not rigorously known. One of its main characteristics is the ability to learn through examples and generalize the information learned, generating a non-linear model, making its application in spatial analysis very efficient [25]. ANN can be an alternative with high performance to the logistic regression (LR) model, where the relative risk term is parameterized by an ANN instead of regression, enabling the application of deep learning. ANN models have been developed mainly to predict survival in patients with cirrhosis, to assess the best donor-to-recipient match during allocation processes, and to foresee postoperative complications and outcomes[26-32], but studies evaluating such a promising tool, as ANN, for predicting AKI following LT for cirrhosis and HCC, are scarce.

The multifactorial origin of AKI after LT makes it complex to predict which candidate for the procedure has an increased risk of this complication[33,34]. In the face of this complexity, ANN would be a very reliable prognostic tool for AKI risk assessment, enabling, therefore, early or even prophylactic therapies for AKI, improving patients outcomes[35]. This is a specific opinion review on the role of ANN in the prediction of AKI after LT for liver cancer and cirrhosis, highlighting potential strengths of the method to forecast this serious postoperative complication.

OVERVIEW OF RISK FACTORS FOR AKI AFTER LT

The etiology of AKI after LT is multifactorial and not fully understood, with several risk factors related to the organ receptor[20,22,24,35], graft-related characteristics[36], and finally some perioperative have been identified over the past few years[20,33,34]. Similarly, the use of postoperative nephrotoxic immunosuppression can further provoke or aggravate kidney damage[20].

Based on these risk factors, various models have been developed using LR for predicting AKI after LT. However, because several of these models address postoperative parameters, their utility in predictive modeling appears to be of questionable relevance. Regardless of the variability of the triggering factors, it is of fundamental importance to identify patients at risk ideally by the set of preoperative clinical assessment and complementary information of the intraoperative period, thus enabling the adoption of preventive measures or early therapies for AKI, such as reduced doses and postponing postoperative patients immunosuppression, and also early RRT, thus reducing mortality and accelerated recovery of renal function[20].

Among the potential AKI predictors that can be evaluated at the time of transplant indication, the severity of the recipient's liver disease stands out[20-37], expressed by the Model for End-Stage Liver Disease (MELD) score. The MELD score determines the allocation of the organ prioritizing the "sickest first" patient, with high values of the score conferring a greater risk for the occurrence of ARF after TH, thus reflecting an interrelationship between liver and renal functions in cirrhotic patients[38]. Similarly, another predictors related to the recipient have been identified, such as high levels of pre-transplant serum creatinine, high body mass index (BMI) of the recipient (BMI values above 30 kg/m^2), and the presence of pre-existing diabetes mellitus[33,35,37].

In addition to the clinical characteristics of the recipient, there are predictive factors of AKI that are related to the functional quality of the graft. The first situation refers to the modality of TH performed, as living-donor LT, in general, offers a graft that is functionally superior to deceased-donor LT, where the critical clinical conditions of the donor confer a greater potential risk to the occurrence of postoperative AKI[20]. Moreover, "marginal grafts" from "extended criteria donors" have increasingly been used, including steatotic grafts, grafts from clinically critical donors, grafts with high ischemia time, both "warm ischemia time" and "cold ischemia time"[20,37,39].

There are some intraoperative events that can be crucial for the occurrence of AKI. The main factor concerns the occurrence of intraoperative arterial hypotension (IOAH) with consequent renal hypoperfusion during LT[22]. Patients undergoing LT often experience IOAH as a result of several factors, including the duration of surgery, the severity of bleeding, the severity of post-reperfusion syndrome of the graft, and the severity of liver disease[33,35,39]. On some occasions, this renal hypoperfusion occurs in patients with previous renal dysfunction[34], and can often be aggravated by the deleterious renal effects of blood transfusion[22,34,37] and the use of vasoactive drugs in the intraoperative period[40].

BASICS OF ANN

An ANN lies under the umbrella of reinforcement machine learning, and comprises 'units' arranged in a series of layers, each of which connects to layers on either side. ANNs are inspired by biological systems, such as the brain, and how they process information. The original concept of ANNs is derived from neurobiological models. ANNs are massively parallel, computer-intensive and data-driven algorithmic system that is composed of multitude of highly interconnected nodes (neurons). Each elementary node of a neural network is able to receive an input from external sources, according to the relative importance and different weight, which transforms into an output signal to other nodes by different activation function[25].

In terms of topology, to implement an ANN, different variables must be defined, among which: (1) the number of nodes in the input layer (such variable corresponds to the number of variables that will be used to feed the neural network, being normally the variables of greater importance for the problem under study); (2) the number of hidden layers and the number of neurons to be placed in these layers; and (3) the number of neurons in the output layer[41].

The process of learning of an ANN is a process where free parameters are adapted through a process of stimulation by the environment in which the network is inserted. With this, the type of learning is determined based on the way in which the modification of the parameters takes place. In summary, there is the following sequence of events: (1) the neural network is stimulated by an environment; (2) the neural network undergoes modifications in its free parameters as a result of this stimulation; and (3) the neural network responds in a new way to the environment, due to changes in its internal structure[25].

Considering the interactions of linked nodes, an output obtained from one node can serve as an input for other nodes, and the conversion of inputs into outputs is activated by virtue of certain transforming function that is typically monotone. The specified working function depends on parameters determined for the training set of inputs and outputs. The network architecture is the organization of nodes and the types of connections permitted. The nodes are arranged in a series of layers with connections between nodes in different layers, but not between nodes in the same layer[42].

ANNs can be classified into feedforward and feedback networks categories, and back-propagation updating algorithm with adjustment of connection weights between the neurons during the training process, is a widely used feedforward networks. Feedforward networks is included within the supervised learning network, essentially

using a gradient descent-training algorithm[43,44].

Multilayer perceptron

The perceptron, introduced by Rosenblatt in 1958, is a simple form of RNA whose main application is in pattern classification problems. The single-layer perceptron is only capable of classifying linearly separable patterns. In practice, the problem to be worked on does not admit an exact linear separation, making it necessary to use a multilayer perceptron. Multilayer perceptron (MLP)-type architectures are the most used and known artificial neural models. An MLP network is subdivided into layers: input layer, intermediate or hidden layer(s) and output layer. In the multilayer ANN architecture, inputs are extended from the input layer to the output layer, passing through one or more hidden layers. In this same sense, a multilayer neural network is typically composed of aligned layers of neurons. The input layer distributes the input information to the hidden layer(s) of the network. At the output layer, the solution to the problem is obtained. Hidden layers are intermediate layers, whose function is to separate the input and output layers. Neurons in one layer are connected only to neurons in the immediately posterior layer, with no feedback or connections between neurons in the same layer. Also, characteristically, the layers are fully connected[45].

In Figure 1 it is possible to observe an MLP-type architecture with two intermediate layers. The presented network has all connections, which means that a neuron in any layer of the network is connected to all other neurons in the previous layer. Signals flow through the network positively, from left to right, layer by layer.

The learning process of MLP networks by back-propagation consists of two steps: propagation and back-propagation. In the propagation step, an activation pattern is applied to the nodes of the network's input layer and its effect propagates through the network, layer by layer. In the last layer, a set of outputs is produced, configured as the real network response. In the and back-propagation step, all synaptic weights are adjusted according to an error correction rule. The error signal is propagated backwards through the network, against the direction of the synaptic connections, the synaptic weights being adjusted to make the actual response of the network approach the desired response, in a statistical sense[25]. An important characteristic of MLP networks is the non-linearity of neuron outputs. This nonlinearity is obtained using a sigmoid-type function as an activation function, usually the logistic function[25].

ANNS FOR AKI PREDICTON AFTER LT FOR CIRRHOSIS AND HCC

Over the past two decades, machine learning algorithms have been increasingly applied for cancer diagnosis, prognostication, and treatment outcome prediction[46-49]. For example, recently, an MLA approach based on a random forest workflow has been developed by a group in Germany to predict disease-free survival after liver resection for HCC[50].

Studies regarding ANNs in the field of LT for cirrhosis and HCC, researchers[26-31] have already conducted studies with LR models and ANN for the prediction of survival of these patients (Table 1). In 1992, Doyle *et al*[26] introduced a 10 feed forward back-propagation ANN model to predict LT survival. Marsh *et al*[27] presented a three layer feed forward fully connected ANN model to predict the survival analysis and time to recurrence of HCC after LT. Parmanto *et al*[28] conducted a study with time series sequence of medical data of patients that undergone LT with ANNs using back-propagation through time algorithm, and their results were compared with 6-fold cross validation. Cucchetti *et al*[29] proposed an ANN survival prognosis model for patients with cirrhosis at a LT unit, and proved that ANN is better than MELD for this proposal. Zhang *et al*[30] proposed a MLP model of patients with cirrhosis and compared the performance of the model with MELD and Sequential Organ Failure Assessment score. In 2013, Cruz *et al*[31] conducted a study with radial basis function ANNs using multi-objective evolutionary algorithm in order to match the donor-recipient pairs.

The results of the researchers above demonstrate that the ANNs predictive models can be capable of using live data of cirrhotic patients with or without HCC, and perform both diagnostic and predictive tasks[32]. Because of the simplicity in structure, ability to do parallel processing tasks, having long term memory, having fault tolerant ability and getting collective output, ANN models can do better than LR models[51].

In the specific scenario of AKI after LT for cirrhosis and HCC, in 2018, Lee *et al*[52] compared the performance of machine learning approaches with that of LR analysis to

Table 1 Studies with artificial neural networks and logistic regression models for the prediction of survival of patients in the field of cirrhosis and liver transplantation

Ref.	Year	Model and endpoint
Doyle *et al* [26]	1992	10 feed forward back-propagation ANN model to predict LT survival
Marsh *et al* [27]	1997	ANN for survival analysis and time to recurrence of HCC after LT
Parmanto *et al*[28]	2001	Back-propagation through time ANN algorithm to predict outcomes after LT
Cucchetti *et al* [29]	2007	ANN for survival prognosis of patients with cirrhosis
Zhang *et al* [30]	2012	MLP model for predicting outcomes of patients with cirrhosis and compared the performance with MELD and SOFA scores
Cruz *et al*[31]	2013	Radial basis function ANNs using multi-objective evolutionary algorithm to match the donor-recipient pairs
Lee *et al*[52]	2018	Compared the performance of ML approaches (decision tree, random forest, gradient boosting machine, support vector machine, naïve Bayes, MLP, and deep belief networks) with that of LR analysis to predict AKI after LT for cirrhosis and HCC (49%)
He *et al*[53]	2021	LR analysis as a conventional model, and random forest, support vector machine, classical decision tree, and conditional inference tree algorithms to predict AKI after LT for cirrhosis and HCC (40.7%)

ANN: Artificial neural network; LR: Logistic regression; LT: Liver transplantation; HCC: Hepatocellular carcinoma; MLP: Multilayer perceptron; MELD: Model for end-stage liver disease; SOFA: Sequential Organ Failure Assessment; AKI: Acute kidney injury.

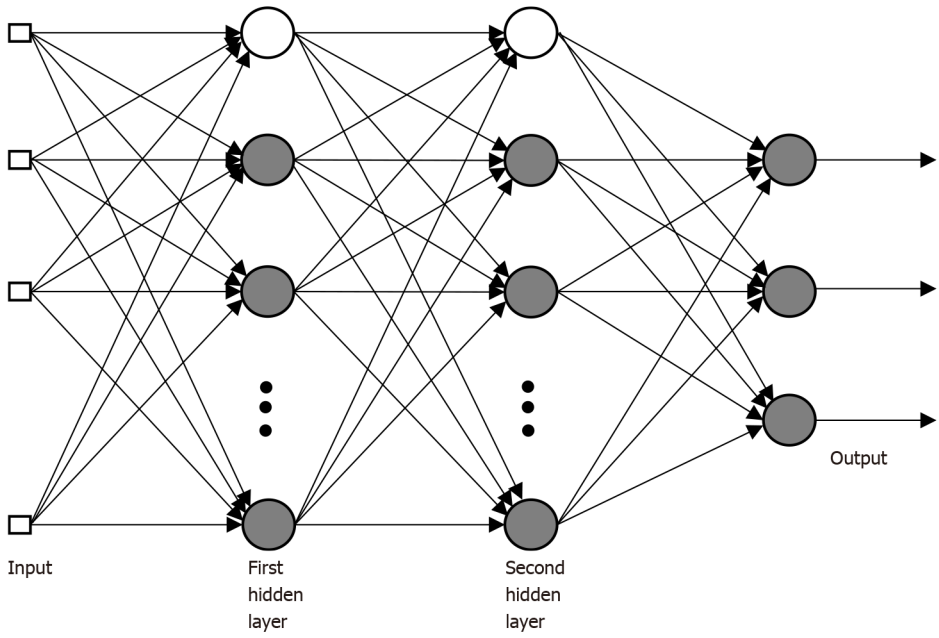

Figure 1 Multilayer perceptron-type architecture with two intermediate layers.

predict AKI after LT for cirrhosis and up to 49% of total patients with HCC. This huge analysis of 1211 patients adopted preoperative and intraoperative input variables. The primary outcome was postoperative AKI defined by Acute Kidney Injury Network criteria. The following machine learning techniques were used: decision tree, random forest, gradient boosting machine, support vector machine, naïve Bayes, MLP, and deep belief networks. These techniques were compared with LR analysis regarding the area under the receiver operating characteristic (AUROC). AKI incidence was 30.1%. The performance in terms of AUROC was best in gradient boosting machine among all analyses to predict AKI of all stages (0.90, 95%CI: 0.86–0.93), and decision tree and random forest techniques showed moderate performance (AUROC 0.86 and 0.85, respectively). The AUROC of the MLP was 0.64 (0.59–0.69), vector machine was 0.62 (0.57–0.67), naïve Bayes was 0.60 (0.54–0.65), and deep belief network was 0.59 (0.53–0.64). The AUROC of LR analysis was 0.61 (95%CI: 0.56–0.66), concluding that

MLP model showed best performance than LR analysis, with a slight higher, but significant, AUROC.

He *et al*[53] evaluated a total of 493 patients (40.7% of patients with HCC) with donation after cardiac death LT. In this study, AKI was defined according to the clinical practice guidelines of Kidney Disease Improving Global Outcomes, and the clinical data of patients with AKI and without AKI were compared through LR analysis as a conventional model, and four predictive machine learning models were developed using random forest, support vector machine, classical decision tree, and conditional inference tree algorithms. The predictive power of these models was then evaluated using the AUROC. The reported incidence of AKI was 35.7% (176/493) during the follow-up period. Compared with the non-AKI group, the AKI group showed a remarkably lower survival rate ($P < 0.001$). The random forest model demonstrated the highest prediction accuracy of 0.79 with AUROC of 0.850 (95%CI: 0.794–0.905), which was significantly higher than the AUCs of the other machine learning algorithms and LR models ($P < 0.001$).

As the standard ANN workflow involves model performance monitoring and re-training to account for model drift, a multidisciplinary partnership between clinicians and data scientists is required, with a commitment to the curation and iterative maintenance of datasets to allow for the development of meaningful decision-support tools[54]. This process should involve, first and foremost, a robust, consistent, and objective means of collecting data. The data in the case of postoperative AKI, are mainly laboratorial and clinicopathologic characteristics from electronic medical records, and clinicians and surgeons must to establish interdisciplinary partnerships that strive towards a common goal and synergism. For instance, clinicians and surgeons help provide a clinically relevant outcome, and data scientists can identify the optimal methodology to make predictions for the outcome based on the available data.

CONCLUSION

The reported high incidence of AKI after LT for cirrhosis and HCC in numerous studies highlights the importance of this issue. The prediction of this complication may provide a focus for further research, mainly in the development of ANNs predictive models that may be applied immediately after LT.

ANNs are essentially a large number of interconnected processing elements, working in unison to solve specific problems, and its use for this specific purpose is directly related to the efficiency with which it provides responses close to real output data. ANN methods may provide feasible tools for forecasting AKI after LT in this population, and perhaps provide a high-performance predictive model that may ultimately improve perioperative management of these patients at risk for this serious complication.

REFERENCES

1 **Bruix J**, Sherman M; American Association for the Study of Liver Diseases. Management of hepatocellular carcinoma: an update. *Hepatology* 2011; **53**: 1020-1022 [PMID: 21374666 DOI: 10.1002/hep.24199]

2 **Forner A**, Reig M, Bruix J. Hepatocellular carcinoma. *Lancet* 2018; **391**: 1301-1314 [PMID: 29307467 DOI: 10.1016/S0140-6736(18)30010-2]

3 **Bruix J**, Reig M, Sherman M. Evidence-Based Diagnosis, Staging, and Treatment of Patients With Hepatocellular Carcinoma. *Gastroenterology* 2016; **150**: 835-853 [PMID: 26795574 DOI: 10.1053/j.gastro.2015.12.041]

4 **European Association For The Study Of The Liver**; European Organisation For Research And Treatment Of Cancer. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. *J Hepatol* 2012; **56**: 908-943 [PMID: 22424438 DOI: 10.1016/j.jhep.2011.12.001]

5 **Johnson RJ**, Bradbury LL, Martin K, Neuberger J; UK Transplant Registry. Organ donation and transplantation in the UK-the last decade: a report from the UK national transplant registry. *Transplantation* 2014; **97** Suppl 1: S1-S27 [PMID: 24356460 DOI: 10.1097/01.TP.0000438215.16737.68]

6 **Neuberger J**, James O. Guidelines for selection of patients for liver transplantation in the era of donor-organ shortage. *Lancet* 1999; **354**: 1636-1639 [PMID: 10560692 DOI: 10.1016/S0140-6736(99)90002-8]

7 **Kwong A**, Kim WR, Lake JR, Smith JM, Schladt DP, Skeans MA, Noreen SM, Foutz J, Miller E, Snyder JJ, Israni AK, Kasiske BL. OPTN/SRTR 2018 Annual Data Report: Liver. *Am J Transplant*

2020; **20** Suppl s1: 193-299 [PMID: 31898413 DOI: 10.1111/ajt.15674]
8 **Freeman RB**, Edwards EB, Harper AM. Waiting list removal rates among patients with chronic and malignant liver diseases. *Am J Transplant* 2006; **6**: 1416-1421 [PMID: 16686765 DOI: 10.1111/j.1600-6143.2006.01321.x]
9 **Northup PG**, Intagliata NM, Shah NL, Pelletier SJ, Berg CL, Argo CK. Excess mortality on the liver transplant waiting list: unintended policy consequences and Model for End-Stage Liver Disease (MELD) inflation. *Hepatology* 2015; **61**: 285-291 [PMID: 24995689 DOI: 10.1002/hep.27283]
10 **Mehta N**, Dodge JL, Goel A, Roberts JP, Hirose R, Yao FY. Identification of liver transplant candidates with hepatocellular carcinoma and a very low dropout risk: implications for the current organ allocation policy. *Liver Transpl* 2013; **19**: 1343-1353 [PMID: 24285611 DOI: 10.1002/lt.23753]
11 **Mazzaferro V**, Sposito C, Coppa J, Miceli R, Bhoori S, Bongini M, Camerini T, Milione M, Regalia E, Spreafico C, Gangeri L, Buzzoni R, de Braud FG, De Feo T, Mariani L. The Long-Term Benefit of Liver Transplantation for Hepatic Metastases From Neuroendocrine Tumors. *Am J Transplant* 2016; **16**: 2892-2902 [PMID: 27134017 DOI: 10.1111/ajt.13831]
12 **Yeh H**, Smoot E, Schoenfeld DA, Markmann JF. Geographic inequity in access to livers for transplantation. *Transplantation* 2011; **91**: 479-486 [PMID: 21200366 DOI: 10.1097/TP.0b013e3182066275]
13 **Mazzaferro V**. Squaring the circle of selection and allocation in liver transplantation for HCC: An adaptive approach. *Hepatology* 2016; **63**: 1707-1717 [PMID: 26703761 DOI: 10.1002/hep.28420]
14 **Mazzaferro V**, Llovet JM, Miceli R, Bhoori S, Schiavo M, Mariani L, Camerini T, Roayaie S, Schwartz ME, Grazi GL, Adam R, Neuhaus P, Salizzoni M, Bruix J, Forner A, De Carlis L, Cillo U, Burroughs AK, Troisi R, Rossi M, Gerunda GE, Lerut J, Belghiti J, Boin I, Gugenheim J, Rochling F, Van Hoek B, Majno P; Metroticket Investigator Study Group. Predicting survival after liver transplantation in patients with hepatocellular carcinoma beyond the Milan criteria: a retrospective, exploratory analysis. *Lancet Oncol* 2009; **10**: 35-43 [PMID: 19058754 DOI: 10.1016/S1470-2045(08)70284-5]
15 **Mazzaferro V**, Sposito C, Zhou J, Pinna AD, De Carlis L, Fan J, Cescon M, Di Sandro S, Yi-Feng H, Lauterio A, Bongini M, Cucchetti A. Metroticket 2.0 Model for Analysis of Competing Risks of Death After Liver Transplantation for Hepatocellular Carcinoma. *Gastroenterology* 2018; **154**: 128-139 [PMID: 28989060 DOI: 10.1053/j.gastro.2017.09.025]
16 **Hamada M**, Matsukawa S, Shimizu S, Kai S, Mizota T. Acute kidney injury after pediatric liver transplantation: incidence, risk factors, and association with outcome. *J Anesth* 2017; **31**: 758-763 [PMID: 28766021 DOI: 10.1007/s00540-017-2395-2]
17 **Chae MS**, Lee N, Park DH, Lee J, Jung HS, Park CS, Choi JH, Hong SH. Influence of oxygen content immediately after graft reperfusion on occurrence of postoperative acute kidney injury in living donor liver transplantation. *Medicine (Baltimore)* 2017; **96**: e7626 [PMID: 28767577 DOI: 10.1097/MD.0000000000007626]
18 **Mizota T**, Hamada M, Matsukawa S, Seo H, Tanaka T, Segawa H. Relationship Between Intraoperative Hypotension and Acute Kidney Injury After Living Donor Liver Transplantation: A Retrospective Analysis. *J Cardiothorac Vasc Anesth* 2017; **31**: 582-589 [PMID: 28216198 DOI: 10.1053/j.jvca.2016.12.002]
19 **Kim WH**, Lee HC, Lim L, Ryu HG, Jung CW. Intraoperative Oliguria with Decreased SvO_2 Predicts Acute Kidney Injury after Living Donor Liver Transplantation. *J Clin Med* 2018; **8** [PMID: 30597881 DOI: 10.3390/jcm8010029]
20 **Kalisvaart M**, Schlegel A, Umbro I, de Haan JE, Polak WG, IJzermans JN, Mirza DF, Perera MTP, Isaac JR, Ferguson J, Mitterhofer AP, de Jonge J, Muiesan P. The AKI Prediction Score: a new prediction model for acute kidney injury after liver transplantation. *HPB (Oxford)* 2019; **21**: 1707-1717 [PMID: 31153834 DOI: 10.1016/j.hpb.2019.04.008]
21 **Zhou J**, Zhang X, Lyu L, Ma X, Miao G, Chu H. Modifiable risk factors of acute kidney injury after liver transplantation: a systematic review and meta-analysis. *BMC Nephrol* 2021; **22**: 149 [PMID: 33888081 DOI: 10.1186/s12882-021-02360-8]
22 **Joosten A**, Lucidi V, Ickx B, Van Obbergh L, Germanova D, Berna A, Alexander B, Desebbe O, Carrier FM, Cherqui D, Adam R, Duranteau J, Saugel B, Vincent JL, Rinehart J, Van der Linden P. Intraoperative hypotension during liver transplant surgery is associated with postoperative acute kidney injury: a historical cohort study. *BMC Anesthesiol* 2021; **21**: 12 [PMID: 33430770 DOI: 10.1186/s12871-020-01228-y]
23 **Chawla LS**, Bellomo R, Bihorac A, Goldstein SL, Siew ED, Bagshaw SM, Bittleman D, Cruz D, Endre Z, Fitzgerald RL, Forni L, Kane-Gill SL, Hoste E, Koyner J, Liu KD, Macedo E, Mehta R, Murray P, Nadim M, Ostermann M, Palevsky PM, Pannu N, Rosner M, Wald R, Zarbock A, Ronco C, Kellum JA; Acute Disease Quality Initiative Workgroup 16. Acute kidney disease and renal recovery: consensus report of the Acute Disease Quality Initiative (ADQI) 16 Workgroup. *Nat Rev Nephrol* 2017; **13**: 241-257 [PMID: 28239173 DOI: 10.1038/nrneph.2017.2]
24 **Hobson C**, Ozrazgat-Baslanti T, Kuxhausen A, Thottakkara P, Efron PA, Moore FA, Moldawer LL, Segal MS, Bihorac A. Cost and Mortality Associated With Postoperative Acute Kidney Injury. *Ann Surg* 2015; **261**: 1207-1214 [PMID: 24887982 DOI: 10.1097/SLA.0000000000000732]
25 **Haykin S**. Neural networks: Principles and Practice. New York: Bookman, 2001
26 **Doyle HR**, Dvorchik I, Mitchell S, Marino IR, Ebert FH, McMichael J, Fung JJ. Predicting outcomes after liver transplantation. A connectionist approach. *Ann Surg* 1994; **219**: 408-415 [PMID: 8161267

DOI: 10.1097/00000658-199404000-00012]
27 **Marsh JW**, Dvorchik I, Subotin M, Balan V, Rakela J, Popechitelev EP, Subbotin V, Casavilla A, Carr BI, Fung JJ, Iwatsuki S. The prediction of risk of recurrence and time to recurrence of hepatocellular carcinoma after orthotopic liver transplantation: a pilot study. *Hepatology* 1997; **26**: 444-450 [PMID: 9252157 DOI: 10.1002/hep.510260227]
28 **Parmanto B**, Doyle HR. Recurrent neural networks for predicting outcomes after liver transplantation: representing temporal sequence of clinical observations. *Methods Inf Med* 2001; **40**: 386-391 [PMID: 11776736]
29 **Cucchetti A**, Vivarelli M, Heaton ND, Phillips S, Piscaglia F, Bolondi L, La Barba G, Foxton MR, Rela M, O'Grady J, Pinna AD. Artificial neural network is superior to MELD in predicting mortality of patients with end-stage liver disease. *Gut* 2007; **56**: 253-258 [PMID: 16809421 DOI: 10.1136/gut.2005.084434]
30 **Zhang M**, Yin F, Chen B, Li YP, Yan LN, Wen TF, Li B. Pretransplant prediction of posttransplant survival for liver recipients with benign end-stage liver diseases: a nonlinear model. *PLoS One* 2012; **7**: e31256 [PMID: 22396731 DOI: 10.1371/journal.pone.0031256]
31 **Cruz-Ramírez M**, Hervás-Martínez C, Fernández JC, Briceño J, de la Mata M. Predicting patient survival after liver transplantation using evolutionary multi-objective artificial neural networks. *Artif Intell Med* 2013; **58**: 37-49 [PMID: 23489761 DOI: 10.1016/j.artmed.2013.02.004]
32 **Dreiseitl S**, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 2002; **35**: 352-359 [PMID: 12968784 DOI: 10.1016/s1532-0464(03)00034-0]
33 **Park MH**, Shim HS, Kim WH, Kim HJ, Kim DJ, Lee SH, Kim CS, Gwak MS, Kim GS. Clinical Risk Scoring Models for Prediction of Acute Kidney Injury after Living Donor Liver Transplantation: A Retrospective Observational Study. *PLoS One* 2015; **10**: e0136230 [PMID: 26302370 DOI: 10.1371/journal.pone.0136230]
34 **Zongyi Y**, Baifeng L, Funian Z, Hao L, Xin W. Risk factors of acute kidney injury after orthotopic liver transplantation in China. *Sci Rep* 2017; **7**: 41555 [PMID: 28134286 DOI: 10.1038/srep41555]
35 **Utsumi M**, Umeda Y, Sadamori H, Nagasaka T, Takaki A, Matsuda H, Shinoura S, Yoshida R, Nobuoka D, Satoh D, Fuji T, Yagi T, Fujiwara T. Risk factors for acute renal injury in living donor liver transplantation: evaluation of the RIFLE criteria. *Transpl Int* 2013; **26**: 842-852 [PMID: 23855657 DOI: 10.1111/tri.12138]
36 **Jochmans I**, Meurisse N, Neyrinck A, Verhaegen M, Monbaliu D, Pirenne J. Hepatic ischemia/reperfusion injury associates with acute kidney injury in liver transplantation: Prospective cohort study. *Liver Transpl* 2017; **23**: 634-644 [PMID: 28124458 DOI: 10.1002/lt.24728]
37 **Thongprayoon C**, Kaewput W, Thamcharoen N, Bathini T, Watthanasuntorn K, Lertjitbanjong P, Sharma K, Salim SA, Ungprasert P, Wijarnpreecha K, Kröner PT, Aeddula NR, Mao MA, Cheungpasitporn W. Incidence and Impact of Acute Kidney Injury after Liver Transplantation: A Meta-Analysis. *J Clin Med* 2019; **8** [PMID: 30884912 DOI: 10.3390/jcm8030372]
38 **Tinti F**, Umbro I, Meçule A, Rossi M, Merli M, Nofroni I, Corradini SG, Poli L, Pugliese F, Ruberto F, Berloco PB, Mitterhofer AP. RIFLE criteria and hepatic function in the assessment of acute renal failure in liver transplantation. *Transplant Proc* 2010; **42**: 1233-1236 [PMID: 20534269 DOI: 10.1016/j.transproceed.2010.03.128]
39 **Romano TG**, Schmidtbauer I, Silva FM, Pompilio CE, D'Albuquerque LA, Macedo E. Role of MELD score and serum creatinine as prognostic tools for the development of acute kidney injury after liver transplantation. *PLoS One* 2013; **8**: e64089 [PMID: 23717537 DOI: 10.1371/journal.pone.0064089]
40 **Wyssusek KH**, Keys AL, Yung J, Moloney ET, Sivalingam P, Paul SK. Evaluation of perioperative predictors of acute kidney injury post orthotopic liver transplantation. *Anaesth Intensive Care* 2015; **43**: 757-763 [PMID: 26603801 DOI: 10.1177/0310057X1504300614]
41 **Santos AM**, Seixas JM, Pereira BB, Medronho RA. Usando Redes Neurais Artificiais e Regressão Logística na predição da Hepatite A. *Revista Brasileira de Epidemiologia* 2005; **8**: 117-126
42 **Eyng E**, Fileti AMF. Control of absorption columns in the bioethanol process: Influence of measurement uncertainties. *Eng Appl Artif Intell* 2010; **23**: 271-282 [DOI: 10.1016/j.engappai.2009.11.002]
43 **Eyng E**, Silva FV, Palú F, Fileti AMF. Neural Network Based Control of an Absorption Column in the Process of Bioethanol Production. *Braz Arch Biol Techn* 2009; **52**: 961-972 [DOI: 10.1590/S1516-89132009000400020]
44 **Liew PL**, Lee YC, Lin YC, Lee TS, Lee WJ, Wang W, Chien CW. Comparison of artificial neural networks with logistic regression in prediction of gallbladder disease among obese patients. *Dig Liver Dis* 2007; **39**: 356-362 [PMID: 17317348 DOI: 10.1016/j.dld.2007.01.003]
45 **Minsky ML**, Papert SA. Perceptrons. Cambridge: MIT Press, 1969
46 **Ivanics T**, Patel MS, Erdman L, Sapisochin G. Artificial intelligence in transplantation (machine-learning classifiers and transplant oncology). *Curr Opin Organ Transplant* 2020; **25**: 426-434 [PMID: 32487887 DOI: 10.1097/MOT.0000000000000773]
47 **Kourou K**, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015; **13**: 8-17 [PMID: 25750696 DOI: 10.1016/j.csbj.2014.11.005]
48 **Singal AG**, Mukherjee A, Elmunzer BJ, Higgins PD, Lok AS, Zhu J, Marrero JA, Waljee AK. Machine learning algorithms outperform conventional regression models in predicting development

of hepatocellular carcinoma. *Am J Gastroenterol* 2013; **108**: 1723-1730 [PMID: 24169273 DOI: 10.1038/ajg.2013.332]

49 **Rajkomar A**, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019; **380**: 1347-1358 [PMID: 30943338 DOI: 10.1056/NEJMra1814259]

50 **Schoenberg MB**, Bucher JN, Koch D, Börner N, Hesse S, De Toni EN, Seidensticker M, Angele MK, Klein C, Bazhin AV, Werner J, Guba MO. A novel machine learning algorithm to predict disease free survival after resection of hepatocellular carcinoma. *Ann Transl Med* 2020; **8**: 434 [PMID: 32395478 DOI: 10.21037/atm.2020.04.16]

51 **Chandra V**, Girijadevi R, Nair AS, Pillai SS, Pillai RM. MTar: a computational microRNA target prediction architecture for human transcriptome. *BMC Bioinformatics* 2010; **11** Suppl 1: S2 [PMID: 20122191 DOI: 10.1186/1471-2105-11-S1-S2]

52 **Lee HC**, Yoon SB, Yang SM, Kim WH, Ryu HG, Jung CW, Suh KS, Lee KH. Prediction of Acute Kidney Injury after Liver Transplantation: Machine Learning Approaches vs. Logistic Regression Model. *J Clin Med* 2018; **7** [PMID: 30413107 DOI: 10.3390/jcm7110428]

53 **He ZL**, Zhou JB, Liu ZK, Dong SY, Zhang YT, Shen T, Zheng SS, Xu X. Application of machine learning models for predicting acute kidney injury following donation after cardiac death liver transplantation. *Hepatobiliary Pancreat Dis Int* 2021; **20**: 222-231 [PMID: 33726966 DOI: 10.1016/j.hbpd.2021.02.001]

54 **Sendak M**, Gao M, Nichols M, Lin A, Balu S. Machine Learning in Health Care: A Critical Appraisal of Challenges and Opportunities. *EGEMS (Wash DC)* 2019; **7**: 1 [PMID: 30705919 DOI: 10.5334/egems.287]

Artificial Intelligence in Cancer

Submit a Manuscript: https://www.f6publishing.com

Artif Intell Cancer 2021 October 28; 2(5): 60-68

DOI: 10.35713/aic.v2.i5.60

ISSN 2644-3228 (online)

MINIREVIEWS

Repairing the human with artificial intelligence in oncology

Ian Morilla

ORCID number: Ian Morilla 0000-0002-5100-5990.

Author contributions: Morilla I designed and wrote the full manuscript.

Conflict-of-interest statement: Dr. Morilla has nothing to disclose.

Manuscript source: Invited manuscript

Specialty type: Mathematical and Computational Biology

Country/Territory of origin: France

Peer-review report's scientific quality classification

Grade A (Excellent): A, A
Grade B (Very good): 0
Grade C (Good): 0
Grade D (Fair): 0

Ian Morilla, Laboratoire Analyse, Géométrie et Applications - Institut Galilée, Sorbonne Paris Nord University, Paris 75006, France

Corresponding author: Ian Morilla, PhD, Assistant Professor, Senior Research Fellow, Laboratoire Analyse, Géométrie et Applications - Institut Galilée, Sorbonne Paris Nord University, 13 Sorbonne-Paris-Cité, Villetaneuse, Paris 75006, France. morilla@math.univ-paris13.fr

Abstract

Artificial intelligence is a groundbreaking tool to learn and analyse higher features extracted from any dataset at large scale. This ability makes it ideal to facing any complex problem that may generally arise in the biomedical domain or oncology in particular. In this work, we envisage to provide a global vision of this mathematical discipline outgrowth by linking some other related subdomains such as transfer, reinforcement or federated learning. Complementary, we also introduce the recently popular method of topological data analysis that improves the performance of learning models.

Key Words: Cancer research; Data analysis; Feature classification; Artificial intelligence; Machine learning; Healthcare systems

Core Tip: In this review, we explore powerful artificial intelligence based models enabling the comprehensive analysis of related problems on oncology. To this end, we described an asserted set of machine learning architectures that goes from the most classical multiple perceptron or neural networks to the novel federated and reinforcement learning designs. Overall, we point out the outgrowth of this mathematical discipline in cancer research and how computational biology and topological features can boost the general performances of these learning models.

Citation: Morilla I. Repairing the human with artificial intelligence in oncology. *Artif Intell Cancer* 2021; 2(5): 60-68

URL: https://www.wjgnet.com/2644-3228/full/v2/i5/60.htm

DOI: https://dx.doi.org/10.35713/aic.v2.i5.60

Grade E (Poor): 0

Received: October 15, 2021
Peer-review started: October 15, 2021
First decision: October 24, 2021
Revised: October 26, 2021
Accepted: October 27, 2021
Article in press: October 27, 2021
Published online: October 28, 2021

P-Reviewer: Guo XY, Saraiva MM
S-Editor: Wang JL
L-Editor: A
P-Editor: Wang JL

INTRODUCTION

The flourishing proliferation of artificial intelligence (AI) worldwide over the last decade has disrupted the way oncologists face cancer. More and more every day, the contribution of AI-based models to different axes of cancer research is not only improving their ability to stratify patients early on or discover new drugs but also influences its fundamentals. By integrating novel structures of data organisation, exploitation, and sharing of clinical data among health institutions, AI is achieving in the short-term to successfully accelerate cancer research. Medical practitioners are becoming familiar with some few mathematical concepts, such as machine learning (ML) or (un/semi) supervised learning. The former is a collection of data-driven techniques with the goal of building predictive models from high-dimensional datasets[1,2], while the latter refers to the grade of human intervention that these models require to make predictions.

These methods are being successfully used in cancer at many levels by simply analysing clinical data, biological indicators, or whole slide images[3-5]. Their application has revealed themselves as an effective way to tackle multiple clinical questions, from diagnosis to prediction of treatment outcomes. For instance, in Morilla *et al*[3], a minimal signature composed of seven miRNAs and two biological indicators was identified using general linear models trained at the base of a deep learning model to predict treatment outcomes in gastrointestinal cancer. In Schmauch *et al*[4], 2020, the authors predicted the RNA-Seq expression of tumours from whole slide images using a deep learning model as well.

Indeed, in this particular discipline, ML algorithms have evolved faster. Several approaches have succeeded in the classification of cancer subtypes using medical imaging[6-8]. Mammography and digital breast tomosynthesis have enabled a robust breast cancer detection by means of annotation-efficient deep learning approaches[9]. Epigenetic patterns of chromatin opening across the stem and differentiated cells across the immune system have also been predicted by deep neural networks in ATAC-seq analysis. In Maslova *et al*[10], solely from the DNA sequence of regulatory regions, the authors discovered ab initio binding motifs for known and unknown master regulators, along with their combinatorial operation.

Another domain where the application of AI-based models has largely been used is single-cell RNA sequencing (sc-RNAseq) analysis. In Lotfollahi *et al*[11] (2020), a new method based on transfer learning (TL) and parameter optimisation is introduced to enable efficient, decentralised, iterative reference building, and the contextualization of new datasets with existing single-cell references without sharing raw data. In addition, few methods have emerged around genetic perturbations of outcomes at the single-cell level in cancer treatments[12,13].

Finally, some computational topology techniques grouped under the heading of "topological data analysis" (TDA) have also been successfully proven as efficient tools in some cancer subtype classifications[14].

Thus, AI has turned the oncologists and co-workers' lives around providing them with a new perspective, which was once developed by only a bulk of specialists and is rapidly becoming a reference in the domain. This work revisits, then, most of those techniques and provides a quick overview of their applications in cancer research.

AI OR ML

ML or AI models, sometimes a philosophical matter, is a branch of mathematics concerned by numerically mimicking the human brain reasoning as it resolves a given problem. There are many examples of this practice; from those most classic techniques of regression or classification of dataset[15] to the current ground-breaking algorithms as "Deep-Mind, Alpha Fold" for protein-folding prediction[16]. In any case, all of these methods share a common objective: the ML problem. This problem can be mathematically expressed as: $$\hat{C}=\underset{C\in\mathcal{M}}{argmin}\mathbb{E}_{x,y\in\mathcal{X}\times\mathcal{Y}} [\mathcal{B}_{l} (C(x),y)]$$.

For example, if we select the particular loss function binary cross entropy, $-B_l-$, this equation describes the parameter misapplication of the neural network C by diminishing the expected value of the loss function between the output of this network $C(x)$ and the true label y.

INTERPRETABLE AI MODELS

Frequently, the intricate design of models based on any ML technique (*i.e.*, neural networks) makes them more difficult to interpret than simpler traditional models. Hence, if we want to fully exploit the potential of these models, a deeper understanding of their predictions would be advisable in practice. Thus, the predicted efficacy of a personal therapy on a cancer must be well explained, since its decisions directly influence human health. From a methodological point of view, we need to ensure model development with proper interpretations of their partial outputs in order to prevent undesirable effects of the models[17,18]. The two main streams of this discipline are the so-called "feature attribution" and "feature interaction" methods. The former[19-22] individually rewards input features depending on its local causal effect in the model output, whereas the latter examines those features with large second-order derivatives at the input or weight matrices of feed-forward and convolutional architectures[23,24]. However, the robustness of all these approaches may be compromised by the presence of specific types of architecture.

DEEP LEARNING

One class of ML models broadly used in current computational cancer research is deep neural networks. Overall, they have succeeded over other non-linear models[25] in the analysis of pathologic image recognition and later patient stratification based on the learned models[26,27]. In brief, deep neural models work in a large number of layers of information that is progressively passing by from one layer to another (*i.e.*, the backpropagation algorithms) to extract relevant features from the original data according to a non-linear model, which is associated with the selected optimisation problem. Their designs can encompass a wide range of algorithms from the classic multiple perceptron networks[28-30] and convolutional neural networks[31-36] to the most recently established long short-term memory (LSTM) recurrent neural networks (RNNs) that are put into the spotlight in the next section[37,38].

RNNs: A different and convenient design other than the more classical neural networks in which the information flows forward are the RNNs. These are computationally more complex models with the skill of capturing hidden behaviours other methods in cancer studies cannot do[39-41]. Recurrent models exhibit an intrinsic representation of the data that allows the exploitation of context information. Specifically, a recurrent network is designed to maintain information about earlier iterations for a period that depends only on the weights and input data at the model's entrance[42]. In particular, the network's activation layers take advantage of inputs that come from chains of information provided by previous iterations. This influences the current prediction and enables the gathering of network flops that can retain contextual information on a long-term scale. Thus, by following this reasoning, RNNs can dynamically exploit a contextual interval over the input training history[43].

LSTM: An improvement in of RNNs is the construction of LSTM networks. LSTMs can learn to sort the interexchange between dependencies in the predictive problems addressed by batches. These models have had a major impact on the biomedical domain, particularly in cancer research[44-48]. LSTMs have been successfully proven in analysis where the intrinsic technical drawbacks associated with RNNs have prevented a fair performance of the model[49]. There are two main optimisation problems that must be avoided during the training stage when applying LSTM to solve a problem, namely: (1) vanishing gradients; and (2) exploding gradients[50]. In this sense, LSTM specifically provides an inner structural amelioration concerning the units leveraged in the learning model[51]. However, there is an improvement in the LSTM network calibration that is increasingly used in biomedical research: LSTM bidirectional networks. In these architectures, a bidirectional recurrent neural lattice is applied in order to be able to separately pass by two forward and backward recurrent nets sharing the same output layer during the training task[51].

TL

Recycling is always a significant issue! In ML, we can also reuse a model that was originally envisaged for solving a different task other than the problem that we might

be currently facing, but both share a similar structural behaviour. This practice is called TL in ML. Its usage has been progressively increasing in problems whose architecture can consume huge amounts of time and computational resources. In these cases, pre-trained networks are applied as a starting learning point, which largely boosts the performance of new models to approach related problems. Then, TL should ameliorate the current model in another setting if such a model is available for learning features from the first problem in a general way[52,53]. Regarding its benefits in oncology, we can outstand its usage in large datasets of piled images to be recognised for patient stratification, as previously described in the following works[54-61].

REINFORCEMENT LEARNING

Reinforcement learning (RL) is one of the latest ML extensions that ameliorates the global performance of learning models when making decisions. In RL, a model learns a given objective in an a priori fixed uncertainty by means of trial and error computations until a solution is obtained. Then, to guide the model, the AI algorithm associates rewards or penalties with the local performance of the model. The final goal was to maximise the amount of rewards obtained. Remarkably, the ML architecture provides no clues on how to find the final solution, even if it rules the reward conditions. Thus, the model must smooth the optimisation problem from a totally random scenario to a complex universe of possibilities. However, if the learning algorithm is launched into a sufficiently powerful computational environment, the ML model will be able to store thousands of trials to effectively achieve the given goal. Nevertheless, a major inconvenience is that the simulation environment is highly dependent on the problem to be computed.

To sum it up, although RL should not be taken as the definitive algorithm, it promises to blow up the current concept of deep learning in oncology[62-64]. An example with no precedents is the DeepMind algorithm very famous nowadays by performing alpha protein folding[16] predictions at a scale ever done before.

FEDERATED LEARNING

A simple description of federated learning (FL) could be a decentralised approach to ML. Thus, FL boosts and accelerates medical discoveries on partnerships with many contributors while protecting patient privacy. In FL, we only improve and calibrate the results and not the data. Thus, what FL really promises it is a new era in secured AI in oncology: Training, testing, or ensuring privacy that way of learning is an efficient method of using data from a comprehensive network of resources belonging each time to a node of many interconnected hospital institutions[65-68].

TOPOLOGICAL ML

Topological ML (TML) is an interaction that has been recently established between TDA and ML. Owing to new advances in computational algorithms, the extraction of complex topological features, such as persistence homology or Betti curves, has become progressively feasible in large datasets. In particular, TDA is commonly referred to as capturing the shape of the data. This method fixes their topological invariants as hotspot to look up relevant structural and categorical information. Indeed, TDA provides ideal completeness in terms of multi-scalability and globalisation missed from the rigidness of their geometric characteristics. In that sense, the use of this tool has been growing in cancer research until it is considered as contextually informative in the analysis of massive biomedical data[69-74]. Multiple studies have exploited the complementary information that emerged from different prisms to gain new insights into the datasets. Its association with ML has enhanced both classical ML methods and deep learning models[75,76].

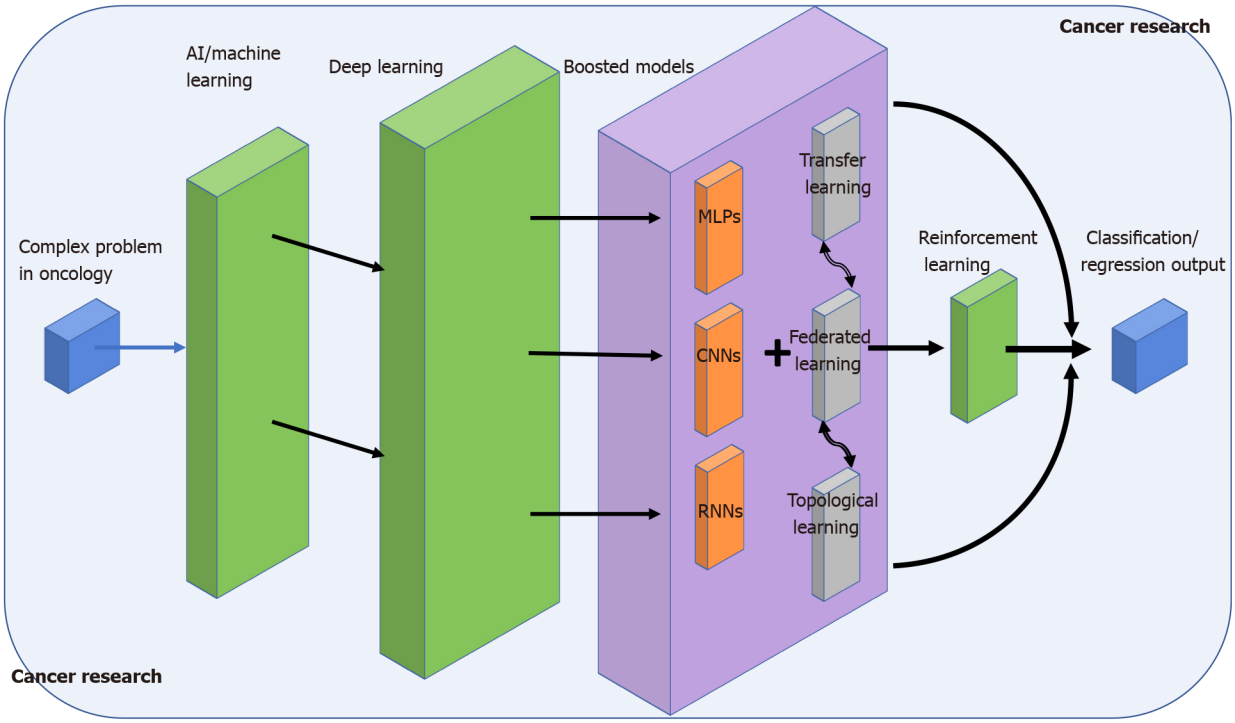

Figure 1 Relational overview of the artificial intelligence-based models introduced in this work. To solve any given complex problem in cancer research by means of machine learning models we can use many deep layers. Then, depending on the particular structures of data, we can empower the performance of the selected architecture, *i.e.*, multilayer perceptron, convolutional or recurrent networks by adding learning strategies such as transfer, federated or topological learning. These strategies are interchangeable (double banded black arrows). As well, we can directly go directly from the selected architecture to the problem's solution using reinforcement learning. AI: Artificial intelligence; MLPs: Multi-layer perceptrons; CNNs: Convolutional neural networks; RNNs: Recurrent neural networks.

CONCLUSION

In this work, we summarise the conclusions of some major references of AI in cancer research (Figure 1). Overall, we wanted to point out the rapid AI outgrowth in the biomedical domain and how AI has systematically become familiar to anyone in the domain, expert, or not. This is possibly due to recent advances in learning-oriented algorithms, which have enabled the transformation of data analysis to any scale and complexity provided a suitable environment is available. We have provided many examples of a varied set of learning models (Multi-layer perceptron, convolutional neural networks, RNNs, *etc.*) that have been successfully proven for related cancer problems such as patient stratification, image-based classification, or recording-device optimisation[77,78]. We have compared different approaches to solve similar questions, and we have introduced novel concepts such as TL, FL, or RL that prevent some of the most classical constraints regarding network architectures or information privacy on high dimensional datasets. Finally, the combination of TDA and ML has also been shown to be a promising discipline where to exploit extra topological features extracted at a higher level. Such tandem promises to contribute to the improvement of the AI algorithm's performance from a totally different perspective. Although data-driven based AI models have the potential to change the world of unsupervised learning, some limitations could endanger a promising future. The three major issues that hamper a better optimisation and general performance in AI models are related to: (1) the high dependency of the model on the data scale; (2) choice of a proper computational environment, and (3) practical problems of time or computational cost should be assumed. Thus, the future challenges in this discipline begin by smoothing such obstacles as much as possible, which will ultimately end up with AI as the tool of reference in healthcare institutions for a much broader analysis in oncology.

ACKNOWLEDGEMENTS

We acknowledge the financial support from the Institut National de la Sant e et de la Recherche Medicale (INSERM), and the Investissements d'Avenir programme ANR-

10-LABX-0017, Sorbonne Paris Nord, Laboratoire d'excellence INFLAMEX.

REFERENCES

1 **Camacho DM**, Collins KM, Powers RK, Costello JC, Collins JJ. Next-Generation Machine Learning for Biological Networks. *Cell* 2018; **173**: 1581-1592 [PMID: 29887378 DOI: 10.1016/j.cell.2018.05.015]

2 **Morilla I**, Léger T, Marah A, Pic I, Zaag H, Ogier-Denis E. Singular manifolds of proteomic drivers to model the evolution of inflammatory bowel disease status. *Sci Rep* 2020; **10**: 19066 [PMID: 33149233 DOI: 10.1038/s41598-020-76011-7]

3 **Morilla I**, Uzzan M, Laharie D, Cazals-Hatem D, Denost Q, Daniel F, Belleannee G, Bouhnik Y, Wainrib G, Panis Y, Ogier-Denis E, Treton X. Colonic MicroRNA Profiles, Identified by a Deep Learning Algorithm, That Predict Responses to Therapy of Patients With Acute Severe Ulcerative Colitis. *Clin Gastroenterol Hepatol* 2019; **17**: 905-913 [PMID: 30223112 DOI: 10.1016/j.cgh.2018.08.068]

4 **Schmauch B**, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, Kamoun A, Sefta M, Toldo S, Zaslavskiy M, Clozel T, Moarii M, Courtiol P, Wainrib G. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat Commun* 2020; **11**: 3877 [PMID: 32747659 DOI: 10.1038/s41467-020-17678-4]

5 **Saillard C**, Schmauch B, Laifa O, Moarii M, Toldo S, Zaslavskiy M, Pronier E, Laurent A, Amaddeo G, Regnault H, Sommacale D, Ziol M, Pawlotsky JM, Mulé S, Luciani A, Wainrib G, Clozel T, Courtiol P, Calderaro J. Predicting Survival After Hepatocellular Carcinoma Resection Using Deep Learning on Histological Slides. *Hepatology* 2020; **72**: 2000-2013 [PMID: 32108950 DOI: 10.1002/hep.31207]

6 **Saillard C**, Delecourt F, Schmauch B, Moindrot O, Svrcek M, Bardier-Dupas A, Emile JF, Ayadi M, De Mestier L, Hammel P, Neuzillet C, Bachet JB, Iovanna J, Nelson DJ, Paradis V, Zaslavskiy M, Kamoun A, Courtiol P, Nicolle R, Cros J. Identification of pancreatic adenocarcinoma molecular subtypes on histology slides using deep learning models. *J Clin Oncol* 2021; **39** suppl 15: 4141 [DOI: 10.1200/jco.2021.39.15_suppl.4141]

7 **Rhee S**, Seo S, Kim S. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2018: 3527–3534 [DOI: 10.24963/ijcai.2018/490]

8 **Schutte K**, Moindrot O, Hérent P, Schiratti JB, Jégou S. Using stylegan for visual interpretability of deep learning models on medical images. 2021 Preprint. Available from: ArXiv:2101.07563

9 **Lotter W**, Diab AR, Haslam B, Kim JG, Grisot G, Wu E, Wu K, Onieva JO, Boyer Y, Boxerman JL, Wang M, Bandler M, Vijayaraghavan GR, Gregory Sorensen A. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med* 2021; **27**: 244-249 [PMID: 33432172 DOI: 10.1038/s41591-020-01174-9]

10 **Maslova A**, Ramirez RN, Ma K, Schmutz H, Wang C, Fox C, Ng B, Benoist C, Mostafavi S; Immunological Genome Project. Deep learning of immune cell differentiation. *Proc Natl Acad Sci U S A* 2020; **117**: 25655-25666 [PMID: 32978299 DOI: 10.1073/pnas.2011795117]

11 **Lotfollahi M**, Naghipourfar M, Luecken MD, Khajavi M, Uttner MB, Avsec Z, Misharin AV, Theis FJ. Query to reference singlecell integration with transfer learning. 2020 Preprint. Available from: bioRxiv:2020.07.16.205997 [DOI: 10.1101/2020.07.16.205997]

12 **Hou W**, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol* 2020; **21**: 218 [PMID: 32854757 DOI: 10.1186/s13059-020-02132-x]

13 **Zhang Y**, Wang D, Peng M, Tang L, Ouyang J, Xiong F, Guo C, Tang Y, Zhou Y, Liao Q, Wu X, Wang H, Yu J, Li Y, Li X, Li G, Zeng Z, Tan Y, Xiong W. Single-cell RNA sequencing in cancer research. *J Exp Clin Cancer Res* 2021; **40**: 81 [PMID: 33648534 DOI: 10.1186/s13046-021-01874-1]

14 **Dindin M**, Umeda Y, Chazal F. Topological data analysis for arrhythmia detection through modular neural networks. In: Goutte C, Zhu X, editors. Advances in Artificial Intelligence. Cham: Springer International Publishing, 2020: 177-188 [DOI: 10.1007/978-3-030-47358-7_17]

15 **Azzaoui T**, Santos D, Sheikh H, Lim S. Solving classification and regression problems using machine and deep learning. Technical report, University of Massachusetts Lowell, 2018 [DOI: 10.13140/RG.2.2.21723.21288]

16 **Deepmind**. Deepmind alphafolding. [Accessed: 2021-10-12] Available from: https://deepmind.com/research

17 **Sundararajan M**, Taly A, Yan Q. Axiomatic attribution for deep networks. In: ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70. JMLR.org, 2017: 3319–3328

18 **Janizek JD**, Sturmfels P, Lee SI. Explaining explanations: Axiomatic feature interactions for deep networks. *J Mach Learn Res* 2021; **22**: 1-54

19 **Binder A**, Montavon G, Lapuschkin S, Muller KR, Same W. Layer-wise relevance propagation for neural networks with local renormalization layers. In: Villa A, Masulli P, Pons Rivero A, editors. Artificial Neural Networks and Machine Learning – ICANN 2016. ICANN 2016. Lecture Notes in Computer Science, vol 9887. Cham: Springer, 2016: 63-71 [DOI: 10.1007/978-3-319-44781-0_8]

20 **Ribeiro MT**, Singh S, Guestrin C. "why should i trust you?" explaining the predictions of any classifier. In: KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2016: 1135–1144 [DOI: 10.1145/2939672.2939778]

21 **Shirkumar A**, Greenside P, Kundaje A. Learning important features through propagating activation diferences. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. JMLR org, 2017: 3145–3153

22 **Lundeberg SM**, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Red Hook: Curran Associates Inc., 2017: 4768-4777

23 **Cui T**, Marttinen P, Kaski S. Recovering pairwise interactions using neural networks. 2019 Preprint. Available from: arXiv:1901.08361

24 **Greenside P**, Shimko T, Fordyce P, Kundaje A. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics* 2018; **34**: i629-i637 [PMID: 30423062 DOI: 10.1093/bioinformatics/bty575]

25 **Shavitt I**, Segal E. Regularization learning networks: Deep learning for tabular datasets. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems, volume 31. Red Hook: Curran Associates Inc., 2018 [DOI: 10.7551/mitpress/7503.003.0107]

26 **Devlin J**, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019: 4171-4186 [DOI: 10.18653/v1/n19-1423]

27 **He K**, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2015 Preprint. Available from: arXiv:1512.03385

28 **Kumar G**, Pawar U, O'Reilly R. Arrhythmia detection in ECG signals using a multilayer perceptron network. In: Curry E, Keane MT, Ojo A, Salwala D, editors. AICS, volume 2563 of CEUR Workshop Proceedings. CEUR-WS.org, 2019: 353-364

29 **Alsmadi MK**, Omar KB, Noah SA, Almarashdah I. Performance comparison of multi-layer perceptron (back propagation, delta rule and perceptron) algorithms in neural networks. In: 2009 IEEE International Advance Computing Conference. IEEE, 2009: 296–299 [DOI: 10.1109/iadcc.2009.4809024]

30 **Freund Y**, Schapire RE. Large margin classification using the perceptron algorithm. *Mach Learn* 1999; **37**: 277-296 [DOI: 10.1023/A:1007662407062]

31 **Hadush S**, Girmay Y, Sinamo A, Hagos G. Breast cancer detection using convolutional neural networks. 2020 Preprint. Available from: arXiv:2003.07911

32 **Chaturvedi SS**, Tembhurne JV, Diwan T. A multi-class skin cancer classification using deep convolutional neural networks. *Multim Tools Appl* 2020; 28477-27498 [DOI: 10.1007/s11042-020-09388-2]

33 **Santos C**, Afonso L, Pereira C, Papa J. BreastNet: Breast cancer categorization using convolutional neural networks. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2020: 463-468 [DOI: 10.1109/CBMS49503.2020.00094]

34 **Yoo S**, Gujrathi I, Haider MA, Khalvati F. Prostate Cancer Detection using Deep Convolutional Neural Networks. *Sci Rep* 2019; **9**: 19518 [PMID: 31863034 DOI: 10.1038/s41598-019-55972-4]

35 **Kumar N**, Verma R, Arora A, Kumar A, Gupta S, Sethi S, Gann PH. Convolutional neural networks for prostate cancer recurrence prediction. In: Gurcan MN, Tomaszewski JE, editors. Proceedings Volume 10140, Medical Imaging 2017: Digital Pathology. SPIE, 2017: 101400H [DOI: 10.1117/12.2255774]

36 **Dumoulin V**, Francesco V. A guide to convolution arithmetic for deep learning. 2016 Preprint. Available from: arXiv:1603.07285

37 **Graves A**. Long Short-Term Memory. Berlin: Springer Berlin Heidelberg 2012; 37-45 [DOI: 10.1007/978-3-642-24797-2_4]

38 **Hochreiter S**, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; **9**: 1735-1780 [PMID: 9377276 DOI: 10.1162/neco.1997.9.8.1735]

39 **Lane N**, Kahanda I. DeepACPpred: A Novel Hybrid CNN-RNN Architecture for Predicting Anti-Cancer Peptides. In: Panuccio G, Rocha M, Fdez-Riverola F, Mohamad MS, Casado-Vara R, editors. PACBB, volume 1240 of Advances in Intelligent Systems and Computing. Springer, 2020: 60-69 [DOI: 10.1007/978-3-030-54568-0_7]

40 **Moitra D**, Mandal RK. Automated AJCC (7th edition) staging of non-small cell lung cancer (NSCLC) using deep convolutional neural network (CNN) and recurrent neural network (RNN). *Health Inf Sci Syst* 2019; **7**: 14 [PMID: 31406570 DOI: 10.1007/s13755-019-0077-1]

41 **Chiang JH**, Chao SY. Modeling human cancer-related regulatory modules by GA-RNN hybrid algorithms. *BMC Bioinformatics* 2007; **8**: 91 [PMID: 17359522 DOI: 10.1186/1471-2105-8-91]

42 **Bengio Y**, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw* 1994; **5**: 157-166 [PMID: 18267787 DOI: 10.1109/72.279181]

43 **Sak H**, Senior AW, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Li H, Meng HM, Ma B, Chng E, Xie L, editors. Proc. Interspeech 2014. ISCA, 2014: 338–342 [DOI: 10.21437/Interspeech.2014-80]

44 **Agrawal P**, Bhagat D, Mahalwal M, Sharma N, Raghava GPS. AntiCP 2.0: an updated model for predicting anticancer peptides. *Brief Bioinform* 2021; **22** [PMID: 32770192 DOI: 10.1093/bib/bbaa153]

45 **Jiang L**, Sun X, Mercaldo F, Antonella Santone A. Decablstm: Deep contextualized attentional bidirectional lstm for cancer hallmark classification. *Knowl Based Syst* 2020; **210**: 106486 [DOI: 10.1016/j.knosys.2020.106486]

46 **Asyhar AH**, Foeady AZ, Thohir M, Arifin AZ, Haq DZ, Novitasari DCR. Implementation LSTM Algorithm for Cervical Cancer using Colposcopy Data. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2020: 485-489 [DOI: 10.1109/icaiic48513.2020.9065068]

47 **Bichindaritz I**, Liu G, Bartlett CL. Survival prediction of breast cancer patient from gene methylation data with deep LSTM network and ordinal cox model. In: Bartk R, Bell E, editors. FLAIRS Conference. AAAI Press, 2020: 353–356 [DOI: 10.32473/flairs.v34i1.128570]

48 **Gao R**, Huo Y, Bao S, Tang Y, Antic S, Epstein ES, Balar A, Deppen S, Paulson AB, Sandler KL, Massion PP, Landman BA. Distanced LSTM: Time-distanced gates in long short-term memory models for lung cancer detection. In: Suk HI, Liu M, Yan P, Lian C, editors. Machine Learning in Medical Imaging. MLMI 2019. Lecture Notes in Computer Science, vol 11861. Cham: Springer, 2019: 310-318 [DOI: 10.1007/978-3-030-32692-0_36]

49 **Gers FA**, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. *Neural Comput* 2000; **12**: 2451-2471 [PMID: 11032042 DOI: 10.1162/089976600300015015]

50 **Graves A**, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans Pattern Anal Mach Intell* 2009; **31**: 855-868 [PMID: 19299860 DOI: 10.1109/TPAMI.2008.137]

51 **Graves A**, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 2005; **18**: 602-610 [PMID: 16112549 DOI: 10.1016/j.neunet.2005.06.042]

52 **Olivas ES**, Guerrero JDM, Sober MM, Benedito JRM, Lopez AJS. Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques - 2 Volumes. Hershey: IGI Publishing, 2009 [DOI: 10.4018/978-1-60566-766-9]

53 **Yosinski J**, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. Cambridge: MIT Press, 2014: 3320-3328

54 **Khamparia A**, Bharati S, Podder P, Gupta D, Khanna A, Phung TK, Thanh DNH. Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. *Multidimens Syst Signal Process* 2021; 1-19 [PMID: 33456204 DOI: 10.1007/s11045-020-00756-7]

55 **Jayachandran S**, Ghosh A. Deep transfer learning for texture classification in colorectal cancer histology. In: Schilling FP, Stadelmann T, editors. Artificial Neural Networks in Pattern Recognition. ANNPR 2020. Lecture Notes in Computer Science, vol 12294. Cham: Springer, 2020: 173-186 [DOI: 10.1007/978-3-030-58309-5_14]

56 **Shaikh TA**, Ali R, Sufyan Beg MM. Transfer learning privileged information fuels CAD diagnosis of breast cancer. *Mach Vis Appl* 2020; **31**: 9 [DOI: 10.1007/s00138-020-01058-5]

57 **de Matos J**, de Souza Britto A, Oliveira LES, Koerich AL. Double transfer learning for breast cancer histopathologic image classification. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019: 1-8 [DOI: 10.1109/IJCNN.2019.8852092]

58 **Obonyo S**, Ruiru D. Multitask learning or transfer learning? application to cancer detection. In: Merelo JJ, Garibaldi JM, Linares-Barranco A, Madani K, Warwick K, editors. Computational Intelligence. ScitePress, 2019: 548–555

59 **Kassani SH**, Kassani PH, Wesolowski MJ, Schneider KA, Deters R. Breast cancer diagnosis with transfer learning and global pooling. In: 2019 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2019: 519-524 [DOI: 10.1109/ICTC46691.2019.8939878]

60 **Dhruba SR**, Rahman R, Matlock K, Ghosh S, Pal R. Application of transfer learning for cancer drug sensitivity prediction. *BMC Bioinformatics* 2018; **19**: 497 [PMID: 30591023 DOI: 10.1186/s12859-018-2465-y]

61 **Vesal S**, Ravikumar N, Davari A, Ellmann S, Maier AK. Classification of breast cancer histology images using transfer learning. In: Campilho A, Karray F, ter Haar Romeny B, editors. Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science, vol 10882. Cham: Springer, 2018: 812–819 [DOI: 10.1007/978-3-319-93000-8_92]

62 **Kalantari J**, Nelson H, Chia N. The Unreasonable Effectiveness of Inverse Reinforcement Learning in Advancing Cancer Research. *Proc Conf AAAI Artif Intell* 2020; **34**: 437-445 [PMID: 34055465 DOI: 10.1609/aaai.v34i01.5380]

63 **Daoud S**, Mdhaffar A, Jmaiel M, Freisleben B. Q-Rank: Reinforcement Learning for Recommending Algorithms to Predict Drug Sensitivity to Cancer Therapy. *IEEE J Biomed Health Inform* 2020; **24**: 3154-3161 [PMID: 32750950 DOI: 10.1109/JBHI.2020.3004663]

64 **Balaprakash P**, Egele R, Salim M, Wild S, Vishwanath V, Xia F, Brettin T, Stevens R. Scalable reinforcement-learning-based neural architecture search for cancer deep learning research. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. New York: Association for Computing Machinery, 2019: 1-33 [DOI: 10.1145/3295500.3356202]

65 **Beguier C**, Andreux M, Tramel EW. Efficient Sparse Secure Aggregation for Federated Learning. 2020 Preprint. Available from: arXiv:2007.14861

66 **Andreux M**, du Terrail JO, Beguier C, Tramel EW. Siloed federated learning for multi-centric histopathology datasets. 2020 Preprint. Available from: arXiv:2008.07424

67 **Andreux M**, Manoel A, Menuet R, Saillard C, Simpson C. Federated Survival Analysis with Discrete-Time Cox Models. 2020 Preprint. Available from: arXiv:2006.08997

68 **Rieke N**, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, Ourselin S, Sheller M, Summers RM, Trask A, Xu D, Baust M, Cardoso MJ. The future of digital health with federated learning. *NPJ Digit Med* 2020; **3**: 119 [PMID: 33015372 DOI: 10.1038/s41746-020-00323-1]

69 **Hensel F**, Moor M, Rieck B. A Survey of Topological Machine Learning Methods. *Front Artif Intell* 2021; **4**: 681108 [PMID: 34124648 DOI: 10.3389/frai.2021.681108]

70 **Groha S**, Weis C, Gusev A, Rieck B. Topological data analysis of copy number alterations in cancer. 2020 Preprint. Available from: arXiv:2011.11070

71 **Loughrey C**, Fitzpatrick P, Orr N, Jurek-Loughrey A. The topology of data: Opportunities for cancer research. *Bioinformatics* 2021 [PMID: 34320632 DOI: 10.1093/bioinformatics/btab553]

72 **Gonzalez G**, Ushakova A, Sazdanovic R, Arsuaga J. Prediction in Cancer Genomics Using Topological Signatures and Machine Learning. In: Baas N, Carlsson G, Quick G, Szymik M, Thaule M, editors. Topological Data Analysis. Abel Symposia. Cham: Springer, 2020: 247-276 [DOI: 10.1007/978-3-030-43408-3_10]

73 **Yu YT**, Lin GH, Jiang IHR, Chiang CC. Machine learning-based hotspot detection using topological classification and critical feature extraction. *IEEE Trans Comput Aided Des Integr Circuits Syst* 2015; **34**: 460-470 [DOI: 10.1109/TCAD.2014.2387858]

74 **Matsumoto T**, Kitazawa M, Kohno Y. Classifying topological charge in SU(3) YangMills theory with machine learning. *Prog Theor Exp Phys* 2021; **2**: 023D01 [DOI: 10.1093/ptep/ptaa138]

75 **Bukkuri A**, Andor N, Darcy IK. Applications of Topological Data Analysis in Oncology. *Front Artif Intell* 2021; **4**: 659037 [PMID: 33928240 DOI: 10.3389/frai.2021.659037]

76 **Huang CH**, Chang PM, Hsu CW, Huang CY, Ng KL. Drug repositioning for non-small cell lung cancer by using machine learning algorithms and topological graph theory. *BMC Bioinformatics* 2016; **17** Suppl 1: 2 [PMID: 26817825 DOI: 10.1186/s12859-015-0845-0]

77 **Morilla I**. A deep learning approach to evaluate intestinal fibrosis in magnetic resonance imaging models. *Neural Comput Appl* 2020; **32**: 14865-14874 [DOI: 10.1007/s00521-020-04838-2]

78 **Morilla I**, Uzzan M, Cazals-Hatem D, Colnot N, Panis Y, Nancey S, Boschetti G, Amiot A, Tréton X, Ogier-Denis E, Daniel F. Computational Learning of microRNA-Based Prediction of Pouchitis Outcome After Restorative Proctocolectomy in Patients With Ulcerative Colitis. *Inflamm Bowel Dis* 2021; **27**: 1653-1660 [PMID: 33609036 DOI: 10.1093/ibd/izab030]

Artificial Intelligence in Cancer

Submit a Manuscript: https://www.f6publishing.com

Artif Intell Cancer 2021 October 28; 2(5): 69-78

DOI: 10.35713/aic.v2.i5.69

ISSN 2644-3228 (online)

MINIREVIEWS

Artificial intelligence reveals roles of gut microbiota in driving human colorectal cancer evolution

Xue-Hua Wan

ORCID number: Xue-Hua Wan 0000-0002-6367-848X.

Author contributions: Wan XH wrote and revised the manuscript; Wan XH has read and approve the final manuscript.

Conflict-of-interest statement: Author declares no conflict of interest.

Manuscript source: Invited manuscript

Specialty type: Microbiology

Country/Territory of origin: China

Peer-review report's scientific quality classification

Grade A (Excellent): A

Xue-Hua Wan, TEDA Institute of Biological Sciences and Biotechnology, Nankai University, Tianjin 300457, China

Corresponding author: Xue-Hua Wan, PhD, Assistant Professor, Senior Researcher, TEDA Institute of Biological Sciences and Biotechnology, Nankai University, No. 23 Honda Street, Tianjin 300457, China. xuehua.wan@hotmail.com

Abstract

With the rapid development of high-throughput sequencing and artificial intelligence (AI) techniques, gut mucosal microbiota begins to be recognized as critical drivers of human colorectal cancer (CRC). Various AI approaches have been designed to obtain effective information from enormous numbers of microbial cells residing in gut mucosal as well as cancer cells. These mainly include detection of microbial markers for early clinical diagnosis of stage-specific CRC, characterization of pathogenic bacterial activities *via* genomic and transcriptomic analyses, and prediction of interplay between bacterial drivers and host immune systems. Here I review the current progresses of AI applications in profiling gut microbiomes linked to CRC initiation and development. I further look forward to future AI research for improving our understanding of the roles of gut microbiota in CRC evolution.

Key Words: Artificial intelligence; Colorectal cancer; Gut microbiome; High-throughput sequencing

Core Tip: In this review, the author reviews the current progresses of artificial intelligence (AI) applications in profiling gut microbiomes linked to colorectal cancer (CRC) initiation and development. The author further looks forward to future AI research for improving our understanding of the roles of gut microbiota in CRC evolution.

Citation: Wan XH. Artificial intelligence reveals roles of gut microbiota in driving human colorectal cancer evolution. *Artif Intell Cancer* 2021; 2(5): 69-78

URL: https://www.wjgnet.com/2644-3228/full/v2/i5/69.htm

Grade B (Very good): 0
Grade C (Good): 0
Grade D (Fair): 0
Grade E (Poor): 0

Received: October 19, 2021
Peer-review started: October 19, 2021
First decision: October 24, 2021
Revised: October 24, 2021
Accepted: October 27, 2021
Article in press: October 27, 2021
Published online: October 28, 2021

P-Reviewer: Herold Z
S-Editor: Wang JL
L-Editor: A
P-Editor: Wang JL

DOI: https://dx.doi.org/10.35713/aic.v2.i5.69

INTRODUCTION

Colorectal cancer (CRC) continuously receives public and academic attentions due to its high prevalence and mortality rate[1]. Understanding the genetic mechanisms behind CRC initiation and progression is important to the development of early diagnosis and new therapy for CRC and its recurrence. The concept of the adenoma-carcinoma sequence, which refers to a sequential activation of oncogenes and inactivation of tumor suppressor genes, is well recognized for CRC progression[2,3]. The adenoma-carcinoma sequence involves genetic mutations and epigenetic modification of human genome in vivo, which have been believed to be caused by exogenous and endogenous mutagens for decades[4-6]. However, it is still not fully understood which exogenous mutagens induce cancers and the induction mechanisms behind them remain largely unknown, especially when the questions go deep to a defined type of cancer.

Growing evidences indicate that gut mucosal microbiota is strongly linked to CRC development and may serve as a primary driver to induce inflammation in the human colon[7-13]. High-throughput sequencing (HTS) of 16S ribosomal RNA (rRNA) gene fragments is widely applied to profile microbial communities and used to study the composition structures of gut mucosal microbiota associated with human CRC (Figure 1)[14-17]. Moreover, metagenome sequencing of gut mucosal microbiomes coupled with binning strategies and other downstream analysis are able to reveal metabolism pathways in potential pathogenic bacteria at lineage levels, which are critical to screening microbial biomarkers (*e.g.*, taxa and gene) for CRC and understanding the microbe-host interactions (Figure 1)[18-20]. Emerging meta transcriptomic sequencing, which examines large-scale gene expressions in microbial communities, is able to provide comprehensive insights into microbial population activities in host. Based on these in silico analyses and following wet-lab validations, species such as *Fusobacterium nucleatum*, *Peptostreptococcus anaerobius*, pks$^+$ *Escherichia coli* and *Eubacterium rectale* have been identified as pathogenic drivers responsible for CRC progression[9,10,12,21]. However, due to the expensive and time-consuming wet-lab experiments, a list of CRC-associated species is on the way to be examined for the physiological roles in CRC progression. Instead, AI approaches can serve as efficient methods to detect potential roles of these microbes in microbe-host interactions and provide clues for wet-lab validation.

With its increasingly wide applications in our everyday life, *e.g.* self-driving cars, facial recognition, and medical diagnosis, AI becomes one of the most popular fields that are heavily invested and supported in a number of countries. AI is capable of mimicking and going beyond human capabilities. In some biological fields such as genomics and transcriptomics, AI is able to complete the complex tasks that are impossible for human to finish[22]. AI technique encompasses machine learning (ML) as a major branch that includes deep learning as a subset of ML[23,24]. In essence, ML are computing algorithms that are either supervised by training datasets or designed as unsupervised algorithms. They are widely applied in gut microbiome field. Here I review the current progresses of AI applications in detection of pathogenic drivers for CRC and prediction of their driving roles in CRC evolution.

TAXONOMIC PROFILING OF GUT MICROBIOMES BASED ON 16S RRNA GENE SEQUENCING

Classification algorithms to categorize operational taxonomic unit

To understand the roles of pathogenic bacterial species in initiating and driving CRC progression, the first and most important step is to identify the spectrum of indigenous bacterial taxonomy in human gut. Current HTS technology has developed sufficiently mature methods and is able to extensively characterize bacterial taxonomy in samples collected from diverse environments and various hosts, including human gut mucosal[14-20,25,26]. As a key step for taxonomic assignment, classification of operational taxonomic units (OTUs) from large datasets of HTS 16S rRNA sequencing reads employs various AI algorithms. Classical algorithms for OTU classification include long-sequence-fist list removal algorithm[27,28], uclust algorithm[29], random

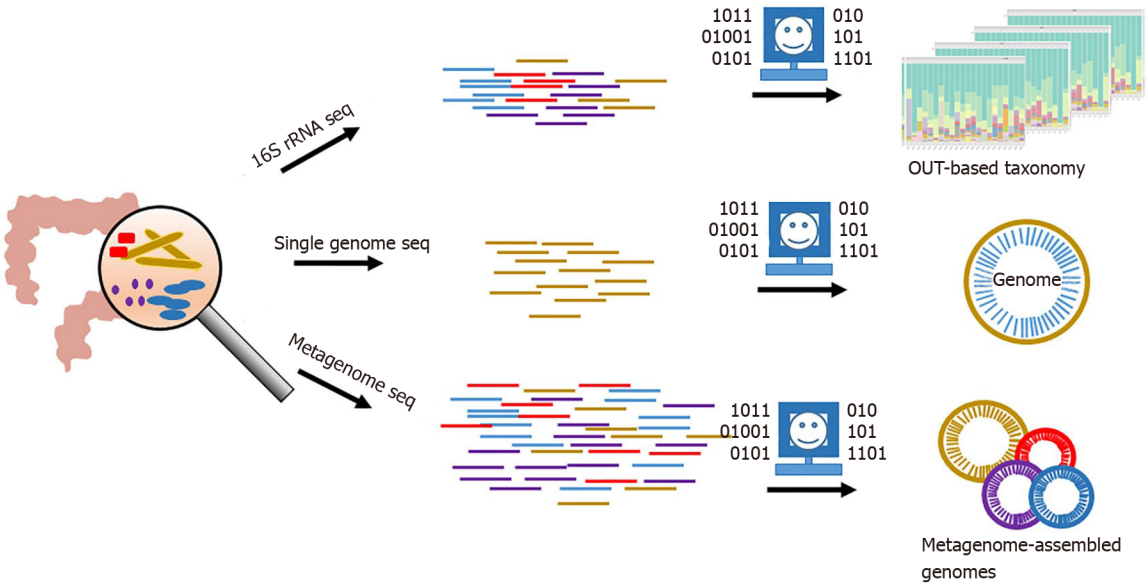

Figure 1 Schematic of artificial intelligence applications in characterizing the traits of gut microbiota associated with colorectal cancer.
OTU: Operational taxonomic unit.

forest algorithm[30], and RDP naïve Bayesian classifier algorithm[31]. Because the datasets are usually generated in large scales, both accuracy and computation speed must be considered for trade off. Long-sequence-fist list removal algorithm implements a super-fast heuristic to identity DNA segments with high identity between sequences, to avoid costly computational alignments of full sequences[27,28]. Uclust algorithm sorts k-mer of sequencing reads to rapidly identify sequences in common[29]. Random forest algorithm builds an ensemble of decision trees that are trained with a combination of learning models[30]. RDP naïve Bayesian classifier algorithm classifies based on the multinomial model in both training and testing for computing classification probabilities[31]. However, challenges still remain to accurately determine the species using 16S rRNA sequences. Errors introduced due to experimental limitations such as polymerase chain reaction amplification and HTS sequencing need to be considered. In addition, although hypervariable regions in 16S rRNA sequences were used for taxonomic assignment, some sequences from bacterial species within the same genus are highly homologous or identical, leading to problems for taxonomic assignment. To solve these issues, new algorithms are also developed. For example, Bayesian-like operational taxonomic unit examiner algorithm employs a grammar-based assignment strategy to deal with sequencing reads errors, in which unsupervised Bayesian models are built based on k-mers split from sequencing reads[32]. To solve homology issues of hypervariable regions in 16S rRNA, Gwak and Rho used a k-nearest neighbor algorithm and the species consensus sequence models to determine species-level taxonomy[33]. Further development of AI methods for OTU classification will help improve the accuracy for taxonomic assignment and speed for dealing with large-scale dataset.

Neighbor-joining and maximum-likelihood based phylogenetic trees

Since gut microbiome OTUs may represent novel species/strains, placing them on a phylogenetic tree can shed light on their taxonomic positions. The computation of phylogenetic likelihood for reconstruction of evolutionary tress from sequence data is both memory and computing consuming. Both Neighbor-Joining (NJ) and maximum-likelihood algorithms are the most popular methods in resolving topology of OTU sequences[34-38]. The NJ tree inference method belongs to distance-based method and takes a matrix of pairwise distance between the sequences to build evolutionary tree. The maximum-likelihood algorithm calculates all the possible tree topologies based on the probability.

Principal component analysis based dimension reduction of big data

The composition structure of gut microbiome is highly complex, containing high-dimensional information for hundreds of bacterial species and their abundances[39]. To apply data mining strategies on looking for critical factors that distinguish gut

microbiomes, large numbers of samples were usually collected from patients in different CRC conditions, such as various intestinal locations and CRC stages. To examine the differences among samples that belong to specific conditions, the high-dimensional information from each sample need to be reduced and presented on a two-dimensional space. As an unsupervised algorithm, principal component analysis is a dimensionality reduction algorithm that transforms and compresses matrix consisting of high-dimensional interrelated variables to a new set of two-dimensional variables[40,41]. By plotting the compressed two-dimensional variables, the microbiome patterns of gut mucosal samples collected from different conditions can be evaluated.

CLINICAL MICROBIAL GENOMIC ASSEMBLY ALGORITHM

To understand gut microbiome functions, bacteria residing in gut mucosal ecosystem need to be isolated and cultivated in laboratory for experimental validation[42]. Sequencing the genomes of these bacteria can reveal their metabolism traits and guide downstream functional analyses. For whole genome shotgun sequencing, bacterial genomic DNA is fragmented into small pieces for 2 × 100 or 2 × 150 bp paired-end sequencing. Various de novo assemblers, including Velvet, SPAdes and SoapDeNovo, have been designed to assemble a large number of short sequence reads to form a set of contiguous sequences representing the genome[43-45]. Because the reads are short, they are usually generated in large quantities with a high coverage depth. To deal with such a large dataset, the assemblers are not designed to assemble the short reads directly. Instead, the reads are splitted to form a set of k-mers and then mapped through de Bruijn graph. Although de Bruijn graph is suggested for short read assembly (100-200 bp), it is not recommended to assemble very short reads (25-50 bp). Velvet was designed to manipulate de Bruijn graph algorithm efficiently for very short reads assembly[43]. Elimination of errors and resolving repeats regions were considered in Velvet[43]. Reconstruction of consensus sequences from k-mers based on de Bruijn algorithm may lead to fragmented assembly. To deal with the issues, paired de Bruijn graphs using read-pairs (bireads) was designed. Inspired by paired de Bruijn graphs, SPAdes uses paired assembly graph algorithm by introducing k-bimer adjustment that reveals exact distances for the adjusted k-bimers[44]. SOAPdevo2, as the version 2 of SOAPdenovo, also utilizes de Bruijn graph algorithm but is designed to reduce memory consumption in de Bruijn graph constructions[45]. The algorithm supports error correction for long k-mers to improve accuracy and sensitivity during the assembly process. Moreover, the program benefits the assembly of repeat regions with high coverage depth and regions with low coverage depth *via* application of a k-mer size selection strategy. Therefore, these assembly algorithms have their specific advantages and are widely utilized in practical applications.

METAGENOMICS ASSEMBLY AND BINNING

Gut mucosal microbiomes comprise hundreds of bacterial species, of which some are uncultivable in laboratory conditions[46,47]. Sequencing these mixed bacterial populations facilitates discovery of the genomic traits of these uncultivable bacteria. Although assembling the reads and reconstructing genes from these complex mixtures are challenging, metagenomic assembly algorithms and downstream binning strategies are under developing progresses to solve the technique problems.

Metagenomic assembly algorithms

Genome assembly for sequencing reads from a single species assumes that all the reads are sequenced from the same genomic DNA and contaminations can be screened out during quality control process[48]. The genome size of single species can be estimated based on the sizes of close phylogenetic neighbors and k-mer counting, and the required sequencing depth can be calculated according to the genome size. During assembly process, de Bruijn algorithm is designed to simply consider nodes or edges with low coverage depth as contamination and remove them[48,49]. In the same way, nodes with high coverage depth are considered by the algorithm as repetitive regions in the genome sequence. In contrast, metagenomic assembly cannot make such a simple assumption to decide nodes with low and high coverage depths to be from contamination sequences or repetitive regions. This is because metagenomic

sequencing reads are generated from mixed bacterial populations, in which certain species grow better than the rest and show high abundances in the mixed communities, whereas rare species show low abundances. Therefore, the coverage depths of heterogeneous reads cannot facilitate the assumption of their origins.

Currently, the most popular assemblers for metagenomics assembly include MEGAHIT and metaSPAdes[50,51]. MEGAHIT utilizes a fast parallel algorithm for succinct de Bruijn graphs to assemble k-mers from metagenomics reads[50]. To avoid k-mer singletons caused by sequencing error, MEGAHIT sorts and counts all (k + 1)-mers splitted from the sequencing reads and only counts (k + 1)-mers with > 2 occurrences[50]. In addition, MEGAHIT utilizes a mercy-kmers strategy to recover low-depth edges for the assembly of rare species[50]. MetaSPAdes uses de Bruijn graph of all reads using SPAdes, transforms it into the assembly graph using various simplification procedures[51]. The algorithm works across a wide range of coverage depths.

Binning strategy

Since assembled metagenomic scaffolds/contigs are derived from each species and show sequence composition characteristics such as GC content and coverage depth, various binning strategies are designed for the reconstruction of metagenome-assembled genome (MAG). MAGs represent genomes from monophyletic lineages and can be used to analyze taxonomic and metabolic potentials. A number of programs have been designed for MAG binning, including MetaBat2, Maxbin2, CONCOCT, MyCC, and BinSanity[52-56]. MetaBat2 is a user-friendly program that does not need to tune the parameters for its sensitivity and specificity[52]. It utilizes a new adaptive binning algorithm to tune these parameters automatically, and uses a graph based structure for contig clustering. MetaBat2 is optimized for extensive low-level computation and works very efficiently for very large datasets. MaxBin 2.0 employs an Expectation-Maximization algorithm to recover draft genomes from metagenomes [53]. It measures the tetranucleotide frequencies of the contigs and their coverages and then classifies the contigs into each bins. CONCOCT uses Gaussian mixture models to cluster contigs into bins[54]. Sequence composition and coverage are considered for assigning contigs to bins. A variational Bayesian approach is used to determine the number of clusters. MyCC works in a way using metagenomics signatures, contig/scaffold coverage depths, and Barnes-Hut-SNE-based dimension reduction [55]. MyCC predicts genes in metagenomic contigs using Prodigal and then identifies single-copy marker genes using Hidden Markov Model trained FetchMG along with UCLUST. The reduced genomic signatures *via* Barnes-Hut-SNE algorithm are then clustered using affinity propagation for binning. Similarly, BinSanity utilizes affinity propagation algorithm to generate bins based on coverage depth, tetranucleotide frequency, and GC content[56]. Although these bin extraction algorithms are designed based on their own specific principles, the resulted bins from the same dataset can be combined, evaluated, modified, and improved to generate high-quality final set of bins using metaWRAP[57].

Quality checking and taxonomic inference for MAGs

Quality evaluation of the assembled MAGs determines the reliability of downstream annotation analyses. Because the concept of metagenome sequencing is quite new, not many programs have been developed with matured principles to determine MAG qualities. Currently, the most popular program is CheckM, which uses a set of lineage-specific marker genes within a reference genome tree[58]. By this way, CheckM estimates the completeness and contamination of the assembled MAGs and determines which MAGs are useful for downstream analyses. To determine the set of marker genes, CheckM reconstructed a genome tree based on 5656 reference genomes and then inferred the marker gene set using HMMER based on hidden Markov models and FastTree based on WAG and GAMMA models. To evaluate a MAG, the marker gene set is identified in the MAG using hidden Markov models. The identified homologous genes of the marker genes are further aligned, concatenated, and then placed into the reference genome tree using pplacer for taxonomic inference and quality checking[59]. Another evaluation method for the assembled MAG is MetaQUAST, which aligns contig sequences of MAG to a close reference genome[60]. This program is able to detect potential taxonomic position of MAG by BLASTN searches against 16S rRNA sequences from the SILVA database[61,62]. Then it automatically downloads close reference genomes from the on-line NCBI database and aligns them against MAG for evaluation.

Different from the taxonomic assignment based on 16S rRNA sequencing, metagenome sequencing and assembly contain much more information than 16S rRNA sequences. Data mining strategies to obtain taxonomic information from large-scale metagenome assembly need to be considered and designed. As discussed above, both CheckM and MetaQUAST provide lineage hints for taxonomic assignment of MAGs[58,60]. Additionally, PhyloFlash maps sequencing reads to small-subunit rRNA (SSU rRNA) database for taxonomic assignment and can be performed before the metagenomes are assembled[63]. FOCUS uses non-negative least squares algorithm to compare k-mers between references genomes and MAGs, and determine taxonomic position for contigs binned in MAGs[64].

PREDICTION OF MICROBE-HOST INTERACTIONS

Gut microbes living in intestine mucosal, including commensals and pathogens, regulate homeostasis of host immunity[65]. Their activities are able to alter host signaling and immunity by interacting with the host proteins. Deciphering how microbe and host interact *via* protein-protein interactions and through which microbial and host proteins they work are important to development of novel strategies for prevention of CRC. Since wet-lab experiments are time-consuming and laborious, experimentally determining the microbe-host interactions is still challenging. On the other hand, genome-wide computational methods can efficiently provide hints to enhance our understanding of this challenging task[66-71]. One category of these computational methods are AI based methods for determining protein-protein interactions (PPI) between microbes and host[69,70]. Currently, AI based methods for PPI predictions are still new and only a few of them have been developed. Most of them are supervised methods, which utilizes well-recognized datasets as standards to train AI models and determine parameters. These training datasets are either collected from high-throughput experiments or obtained from literatures by text mining. Supervised PPI methods utilize various AI models such as logistic regression, random forests, support vector machine, artificial neural networks, and K-nearest neighbors [72-76]. However, these AI-based PPI methods are designed for the PPI relationship between specific pathogen and human such as human-*Bacillus anthracis*, human-*Yersinia pestis* and human-*Fusobacterium nucleatum*[67,77-79]. Because high abundances of *F. nucleatum* are associated with CRC patients and especially associated with specific CRC stages, *F. nucleatum* is proposed for its causal role in CRC development. Computational scanning of *F. nucleatum* genome and human proteins identified FusoSecretome proteins and their targets in the host network[67]. PPI-coupled network analysis identified that *F. nucleatum* perturbed host cellular pathways including immune and infection response, homeostasis, cytoskeleton organization, and gene expression regulation[67]. However, AI-based PPI studies for human-microbiome interactions still need more efforts due to the complex mixed-population of species within gut microbiome.

CONCLUSION

Rapid development of high-throughput sequencing and high-throughput screening experiments generate large-scale datasets and largely improve our understanding of functional roles of gut microbiomes in CRC evolution. Using AI-based analyses, potential pathogenic species from gut microbiome have been identified to play critical roles in driving CRC. However, there are still limitations in current methods and challenges remain for them to be improved. These include but not limited to the questions as follows. How to accurately identify bacterial species/strains that reside in gut mucosal? How to use metagenomics sequencing data to assemble complete or nearly complete MAGs for bacterial single species? How to build AI models to interpret human-microbiome interactions under different environmental conditions? And many more challenges remain to be solved. I believe that continuous improvement of AI technology in CRC diagnosis as well as many more diseases will facilitate answering the above questions and help develop clinical treatment and prevention of CRC in advance.

REFERENCES

1 **Bray F**, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018; **68**: 394-424 [PMID: 30207593 DOI: 10.3322/caac.21492]

2 **Fearon ER**, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990; **61**: 759-767 [PMID: 2188735 DOI: 10.1016/0092-8674(90)90186-i]

3 **Smit WL**, Spaan CN, Johannes de Boer R, Ramesh P, Martins Garcia T, Meijer BJ, Vermeulen JLM, Lezzerini M, MacInnes AW, Koster J, Medema JP, van den Brink GR, Muncan V, Heijmans J. Driver mutations of the adenoma-carcinoma sequence govern the intestinal epithelial global translational capacity. *Proc Natl Acad Sci U S A* 2020; **117**: 25560-25570 [PMID: 32989144 DOI: 10.1073/pnas.1912772117]

4 **Morley AA**, Turner DR. The contribution of exogenous and endogenous mutagens to *in vivo* mutations. *Mutat Res* 1999; **428**: 11-15 [PMID: 10517973 DOI: 10.1016/s1383-5742(99)00026-5]

5 **Stratton MR**, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458**: 719-724 [PMID: 19360079 DOI: 10.1038/nature07943]

6 **Esteller M**. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* 2007; **8**: 286-298 [PMID: 17339880 DOI: 10.1038/nrg2005]

7 **Tjalsma H**, Boleij A, Marchesi JR, Dutilh BE. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. *Nat Rev Microbiol* 2012; **10**: 575-582 [PMID: 22728587 DOI: 10.1038/nrmicro2819]

8 **Song M**, Chan AT. Environmental Factors, Gut Microbiota, and Colorectal Cancer Prevention. *Clin Gastroenterol Hepatol* 2019; **17**: 275-289 [PMID: 30031175 DOI: 10.1016/j.cgh.2018.07.012]

9 **Zhang S**, Cai S, Ma Y. Association between *Fusobacterium nucleatum* and colorectal cancer: Progress and future directions. *J Cancer* 2018; **9**: 1652-1659 [PMID: 29760804 DOI: 10.7150/jca.24048]

10 **Long X**, Wong CC, Tong L, Chu ESH, Ho Szeto C, Go MYY, Coker OO, Chan AWH, Chan FKL, Sung JJY, Yu J. Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. *Nat Microbiol* 2019; **4**: 2319-2330 [PMID: 31501538 DOI: 10.1038/s41564-019-0541-3]

11 **Rhee KJ**, Wu S, Wu X, Huso DL, Karim B, Franco AA, Rabizadeh S, Golub JE, Mathews LE, Shin J, Sartor RB, Golenbock D, Hamad AR, Gan CM, Housseau F, Sears CL. Induction of persistent colitis by a human commensal, enterotoxigenic Bacteroides fragilis, in wild-type C57BL/6 mice. *Infect Immun* 2009; **77**: 1708-1718 [PMID: 19188353 DOI: 10.1128/IAI.00814-08]

12 **Wang Y**, Wan X, Wu X, Zhang C, Liu J, Hou S. Eubacterium rectale contributes to colorectal cancer initiation *via* promoting colitis. *Gut Pathog* 2021; **13**: 2 [PMID: 33436075 DOI: 10.1186/s13099-020-00396-z]

13 **Wang Y**, Zhang C, Hou S, Wu X, Liu J, Wan X. Analyses of Potential Driver and Passenger Bacteria in Human Colorectal Cancer. *Cancer Manag Res* 2020; **12**: 11553-11561 [PMID: 33209059 DOI: 10.2147/CMAR.S275316]

14 **Nakatsu G**, Li X, Zhou H, Sheng J, Wong SH, Wu WK, Ng SC, Tsoi H, Dong Y, Zhang N, He Y, Kang Q, Cao L, Wang K, Zhang J, Liang Q, Yu J, Sung JJ. Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat Commun* 2015; **6**: 8727 [PMID: 26515465 DOI: 10.1038/ncomms9727]

15 **Dadkhah E**, Sikaroodi M, Korman L, Hardi R, Baybick J, Hanzel D, Kuehn G, Kuehn T, Gillevet PM. Gut microbiome identifies risk for colorectal polyps. *BMJ Open Gastroenterol* 2019; **6**: e000297 [PMID: 31275588 DOI: 10.1136/bmjgast-2019-000297]

16 **Saito K**, Koido S, Odamaki T, Kajihara M, Kato K, Horiuchi S, Adachi S, Arakawa H, Yoshida S, Akasu T, Ito Z, Uchiyama K, Saruta M, Xiao JZ, Sato N, Ohkusa T. Metagenomic analyses of the gut microbiota associated with colorectal adenoma. *PLoS One* 2019; **14**: e0212406 [PMID: 30794590 DOI: 10.1371/journal.pone.0212406]

17 **Zhang M**, Lv Y, Hou S, Liu Y, Wang Y, Wan X. Differential Mucosal Microbiome Profiles across Stages of Human Colorectal Cancer. *Life (Basel)* 2021; **11** [PMID: 34440574 DOI: 10.3390/Life11080831]

18 **Dai Z**, Coker OO, Nakatsu G, Wu WKK, Zhao L, Chen Z, Chan FKL, Kristiansen K, Sung JJY, Wong SH, Yu J. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* 2018; **6**: 70 [PMID: 29642940 DOI: 10.1186/s40168-018-0451-2]

19 **Chen F**, Dai X, Zhou CC, Li KX, Zhang YJ, Lou XY, Zhu YM, Sun YL, Peng BX, Cui W. Integrated analysis of the faecal metagenome and serum metabolome reveals the role of gut microbiome-associated metabolites in the detection of colorectal cancer and adenoma. *Gut* 2021 [PMID: 34462336 DOI: 10.1136/gutjnl-2020-323476]

20 **Mizutani S**, Yamada T, Yachida S. Significance of the gut microbiome in multistep colorectal carcinogenesis. *Cancer Sci* 2020; **111**: 766-773 [PMID: 31910311 DOI: 10.1111/cas.14298]

21 **Pleguezuelos-Manzano C**, Puschhof J, Rosendahl Huber A, van Hoeck A, Wood HM, Nomburg J, Gurjao C, Manders F, Dalmasso G, Stege PB, Paganelli FL, Geurts MH, Beumer J, Mizutani T, Miao Y, van der Linden R, van der Elst S; Genomics England Research Consortium, Garcia KC, Top J, Willems RJL, Giannakis M, Bonnet R, Quirke P, Meyerson M, Cuppen E, van Boxtel R, Clevers H. Mutational signature in colorectal cancer caused by genotoxic pks$^+$ E. coli. *Nature* 2020; **580**: 269-

273 [PMID: 32106218 DOI: 10.1038/s41586-020-2080-8]
22 **Sharma A**, Rani R. A systematic review of applications of machine learning in cancer prediction and diagnosis. *Arch Comput Methods Eng* 2021 [DOI: 10.1007/s11831-021-09556-z]
23 **Kourou K**, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015; **13**: 8-17 [PMID: 25750696 DOI: 10.1016/j.csbj.2014.11.005]
24 **Choi RY**, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. *Transl Vis Sci Technol* 2020; **9**: 14 [PMID: 32704420]
25 **Pallen MJ**, Loman NJ, Penn CW. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr Opin Microbiol* 2010; **13**: 625-631 [PMID: 20843733 DOI: 10.1016/j.mib.2010.08.003]
26 **Lightbody G**, Haberland V, Browne F, Taggart L, Zheng H, Parkes E, Blayney JK. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Brief Bioinform* 2019; **20**: 1795-1811 [PMID: 30084865 DOI: 10.1093/bib/bby051]
27 **Li W**, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 2001; **17**: 282-283 [PMID: 11294794 DOI: 10.1093/bioinformatics/17.3.282]
28 **Li W**, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006; **22**: 1658-1659 [PMID: 16731699 DOI: 10.1093/bioinformatics/btl158]
29 **Edgar RC**. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010; **26**: 2460-2461 [PMID: 20709691 DOI: 10.1093/bioinformatics/btq461]
30 **Schloss PD**, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; **75**: 7537-7541 [PMID: 19801464 DOI: 10.1128/AEM.01541-09]
31 **Wang Q**, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007; **73**: 5261-5267 [PMID: 17586664 DOI: 10.1128/AEM.00062-07]
32 **Koparde VN**, Adkins RS, Fettweis JM, Serrano MG, Buck GA, Reimers MA, Sheth NU. BOTUX: bayesian-like operational taxonomic unit examiner. *Int J Comput Biol Drug Des* 2014; **7**: 130-145 [PMID: 24878725 DOI: 10.1504/IJCBDD.2014.061652]
33 **Gwak HJ**, Rho M. Data-Driven Modeling for Species-Level Taxonomic Assignment From 16S rRNA: Application to Human Microbiomes. *Front Microbiol* 2020; **11**: 570825 [PMID: 33262743 DOI: 10.3389/fmicb.2020.570825]
34 **Saitou N**, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; **4**: 406-425 [PMID: 3447015 DOI: 10.1093/oxfordjournals.molbev.a040454]
35 **Mailund T**, Brodal GS, Fagerberg R, Pedersen CN, Phillips D. Recrafting the neighbor-joining method. *BMC Bioinformatics* 2006; **7**: 29 [PMID: 16423304 DOI: 10.1186/1471-2105-7-29]
36 **Dhar A**, Minin VN. Maximum likelihood phylogenetic inference. *Ency Evol Bio* 2016; **2**: 499-506 [DOI: 10.1016/B978-0-12-800049-6.00207-9]
37 **Price MN**, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010; **5**: e9490 [PMID: 20224823 DOI: 10.1371/journal.pone.0009490]
38 **Sagulenko P**, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* 2018; **4**: vex042 [PMID: 29340210 DOI: 10.1093/ve/vex042]
39 **Thursby E**, Juge N. Introduction to the human gut microbiota. *Biochem J* 2017; **474**: 1823-1836 [PMID: 28512250 DOI: 10.1042/BCJ20160510]
40 **Swets DL**, Weng J. Using discriminant eigenfeatures for image retrieval. *IEEE Trans Pattern Anal Mach Intell* 1996; **18**: 831-836 [DOI: 10.1109/34.531802]
41 **Turk M**, Pentland A. Eigenfaces for recognition. *J Cogn Neurosci* 1991; **3**: 71-86 [PMID: 23964806 DOI: 10.1162/jocn.1991.3.1.71]
42 **Forster SC**, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, Dunn M, Mkandawire TT, Zhu A, Shao Y, Pike LJ, Louie T, Browne HP, Mitchell AL, Neville BA, Finn RD, Lawley TD. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol* 2019; **37**: 186-192 [PMID: 30718869 DOI: 10.1038/s41587-018-0009-7]
43 **Zerbino DR**, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008; **18**: 821-829 [PMID: 18349386 DOI: 10.1101/gr.074492.107]
44 **Bankevich A**, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; **19**: 455-477 [PMID: 22506599 DOI: 10.1089/cmb.2012.0021]
45 **Luo R**, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 2012; **1**: 18 [PMID: 23587118 DOI: 10.1186/2047-217X-1-18]
46 **Kenny DJ**, Plichta DR, Shungin D, Koppel N, Hall AB, Fu B, Vasan RS, Shaw SY, Vlamakis H,

Balskus EP, Xavier RJ. Cholesterol Metabolism by Uncultured Human Gut Bacteria Influences Host Cholesterol Level. *Cell Host Microbe* 2020; **28**: 245-257.e6 [PMID: 32544460 DOI: 10.1016/j.chom.2020.05.013]

47 **Almeida A**, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. A new genomic blueprint of the human gut microbiota. *Nature* 2019; **568**: 499-504 [PMID: 30745586 DOI: 10.1038/s41586-019-0965-1]

48 **Ayling M**, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. *Brief Bioinform* 2020; **21**: 584-594 [PMID: 30815668 DOI: 10.1093/bib/bbz020]

49 **Sedlar K**, Kupkova K, Provaznik I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput Struct Biotechnol J* 2017; **15**: 48-55 [PMID: 27980708 DOI: 10.1016/j.csbj.2016.11.005]

50 **Li D**, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly *via* succinct de Bruijn graph. *Bioinformatics* 2015; **31**: 1674-1676 [PMID: 25609793 DOI: 10.1093/bioinformatics/btv033]

51 **Nurk S**, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017; **27**: 824-834 [PMID: 28298430 DOI: 10.1101/gr.213959.116]

52 **Kang DD**, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019; **7**: e7359 [PMID: 31388474 DOI: 10.7717/peerj.7359]

53 **Wu YW**, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016; **32**: 605-607 [PMID: 26515820 DOI: 10.1093/bioinformatics/btv638]

54 **Alneberg J**, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014; **11**: 1144-1146 [PMID: 25218180 DOI: 10.1038/nmeth.3103]

55 **Lin HH**, Liao YC. Accurate binning of metagenomic contigs *via* automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* 2016; **6**: 24175 [PMID: 27067514 DOI: 10.1038/srep24175]

56 **Graham ED**, Heidelberg JF, Tully BJ. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* 2017; **5**: e3035 [PMID: 28289564 DOI: 10.7717/peerj.3035]

57 **Uritskiy GV**, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 2018; **6**: 158 [PMID: 30219103 DOI: 10.1186/s40168-018-0541-1]

58 **Parks DH**, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015; **25**: 1043-1055 [PMID: 25977477 DOI: 10.1101/gr.186072.114]

59 **Matsen FA**, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 2010; **11**: 538 [PMID: 21034504 DOI: 10.1186/1471-2105-11-538]

60 **Mikheenko A**, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 2016; **32**: 1088-1090 [PMID: 26614127 DOI: 10.1093/bioinformatics/btv697]

61 **Quast C**, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013; **41**: D590-D596 [PMID: 23193283 DOI: 10.1093/nar/gks1219]

62 **Yilmaz P**, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res* 2014; **42**: D643-D648 [PMID: 24293649 DOI: 10.1093/nar/gkt1209]

63 **Gruber-Vodicka HR**, Seah BKB, Pruesse E. phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. *mSystems* 2020; **5** [PMID: 33109753 DOI: 10.1128/mSystems.00920-20]

64 **Silva GG**, Cuevas DA, Dutilh BE, Edwards RA. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* 2014; **2**: e425 [PMID: 24949242 DOI: 10.7717/peerj.425]

65 **Pickard JM**, Zeng MY, Caruso R, Núñez G. Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease. *Immunol Rev* 2017; **279**: 70-89 [PMID: 28856738 DOI: 10.1111/imr.12567]

66 **Zuñiga C**, Zaramela L, Zengler K. Elucidation of complexity and prediction of interactions in microbial communities. *Microb Biotechnol* 2017; **10**: 1500-1522 [PMID: 28925555 DOI: 10.1111/1751-7915.12855]

67 **Zanzoni A**, Spinelli L, Braham S, Brun C. Perturbed human sub-networks by Fusobacterium nucleatum candidate virulence proteins. *Microbiome* 2017; **5**: 89 [PMID: 28793925 DOI: 10.1186/s40168-017-0307-1]

68 **Magnúsdóttir S**, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jäger C, Baginska J, Wilmes P, Fleming RM, Thiele I. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol* 2017; **35**: 81-89 [PMID: 27893703 DOI: 10.1038/nbt.3703]

69 **Lian X**, Yang S, Li H, Fu C, Zhang Z. Machine-Learning-Based Predictor of Human-Bacteria Protein-Protein Interactions by Incorporating Comprehensive Host-Network Properties. *J Proteome*

Res 2019; **18**: 2195-2205 [PMID: 30983371 DOI: 10.1021/acs.jproteome.9b00074]

70 **Guven-Maiorov E**, Hakouz A, Valjevac S, Keskin O, Tsai CJ, Gursoy A, Nussinov R. HMI-PRED: A Web Server for Structural Prediction of Host-Microbe Interactions Based on Interface Mimicry. *J Mol Biol* 2020; **432**: 3395-3403 [PMID: 32061934 DOI: 10.1016/j.jmb.2020.01.025]

71 **Jansma J**, El Aidy S. Understanding the host-microbe interactions using metabolic modeling. *Microbiome* 2021; **9**: 16 [PMID: 33472685 DOI: 10.1186/s40168-020-00955-1]

72 **Prasasty VD**, Hutagalung RA, Gunadi R, Sofia DY, Rosmalena R, Yazid F, Sinaga E. Prediction of human-Streptococcus pneumoniae protein-protein interactions using logistic regression. *Comput Biol Chem* 2021; **92**: 107492 [PMID: 33964803 DOI: 10.1016/j.compbiolchem.2021.107492]

73 **Wei ZS**, Yang JY, Shen HB, Yu DJ. A Cascade Random Forests Algorithm for Predicting Protein-Protein Interaction Sites. *IEEE Trans Nanobioscience* 2015; **14**: 746-760 [PMID: 26441427 DOI: 10.1109/TNB.2015.2475359]

74 **Chakraborty A**, Mitra S, De D, Pal AJ, Ghaemi F, Ahmadian A, Ferrara M. Determining Protein–Protein Interaction Using Support Vector Machine: A Review. *IEEE Access* 2021; **9**: 12473-12490 [DOI: 10.1109/ACCESS.2021.3051006]

75 **Tsuchiya Y**, Tomii K. Neural networks for protein structure and function prediction and dynamic analysis. *Biophys Rev* 2020; **12**: 569-573 [PMID: 32166610 DOI: 10.1007/s12551-020-00685-6]

76 **Suratanee A**, Plaimas K. Reverse Nearest Neighbor Search on a Protein-Protein Interaction Network to Infer Protein-Disease Associations. *Bioinform Biol Insights* 2017; **11**: 1177932217720405 [PMID: 28757797 DOI: 10.1177/1177932217720405]

77 **Ahmed I**, Witbooi P, Christoffels A. Prediction of human-Bacillus anthracis protein-protein interactions using multi-layer neural network. *Bioinformatics* 2018; **34**: 4159-4164 [PMID: 29945178 DOI: 10.1093/bioinformatics/bty504]

78 **Dyer MD**, Neff C, Dufford M, Rivera CG, Shattuck D, Bassaganya-Riera J, Murali TM, Sobral BW. The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. *PLoS One* 2010; **5**: e12089 [PMID: 20711500 DOI: 10.1371/journal.pone.0012089]

79 **Yang H**, Ke Y, Wang J, Tan Y, Myeni SK, Li D, Shi Q, Yan Y, Chen H, Guo Z, Yuan Y, Yang X, Yang R, Du Z. Insight into bacterial virulence mechanisms against host immune response *via* the Yersinia pestis-human protein-protein interaction network. *Infect Immun* 2011; **79**: 4413-4424 [PMID: 21911467 DOI: 10.1128/IAI.05622-11]

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-3991568
E-mail: bpgoffice@wjgnet.com
Help Desk: https://www.f6publishing.com/helpdesk
https://www.wjgnet.com