Perspective

Modern Deep Learning in Bioinformatics

Haoyang Li[1,2,#], Shuye Tian[3,#], Yu Li[4,#], Qiming Fang[5], Renbo Tan[1], Yijie Pan[6], Chao Huang[6], Ying Xu[1,2,7,*], and Xin Gao[4,*]

[1] Cancer Systems Biology Center, the China-Japan Union Hospital, Jilin University, Changchun 130033, China

[2] MOE Key Laboratory of Symbolic Computation and Knowledge Engineering, College of Computer Science and Technology, Jilin University, Changchun 130012, China

[3] Department of Biology, Southern University of Science and Technology, Shenzhen 518055, China

[4] Computational Bioscience Research Center (CBRC), Computer Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia

[5] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China

[6] Ningbo Institute of Information Technology Application, Chinese Academy of Sciences, Ningbo, China

[7] Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

[#] These authors contributed equally to this work.

[*] Correspondence to: Xin Gao, E-mail: xin.gao@kaust.edu.sa; Ying Xu, E-mail: xyn@uga.edu

Deep learning (DL) has shown explosive growth in its application to bioinformatics and has demonstrated thrillingly promising power to mine the complex relationship hidden in large-scale biological and biomedical data. A number of comprehensive reviews have been published on such applications, ranging from high-level reviews with future perspectives to those mainly serving as tutorials. These reviews have provided an excellent introduction to and guideline for applications of DL in bioinformatics, covering multiple types of machine learning (ML) problems, different DL architectures, and ranges of biological/biomedical problems. However, most of these reviews have focused on previous research, whereas current trends in the principled DL field and perspectives on their future developments and potential new applications to biology and biomedicine are still scarce. We will focus on

modern DL, the ongoing trends and future directions of the principled DL field, and postulate new and major applications in bioinformatics.

Introduction

ML has been the main contributor to the recent resurgence of artificial intelligence (AI). The most essential piece in modern machine learning technology is DL. DL is founded on artificial neural networks (ANNs), which have been theoretically proven to be capable of approximating any nonlinear function within any specified accuracy (Hornik 1991) and have been widely used to solve various computational tasks (Li et al. 2019). However, they have been criticized for being black boxes. This lack of interpretability has limited their applications, particularly when their performance did not stand out among other more interpretable machine learning methods, such as linear regression, logistic regression, support vector machines (SVMs), and decision trees.

During the past decade, three important advances in science and technology have led to the rejuvenation of ANNs, particularly via DL. First, unprecedented quantities of data have been generated in modern life, mostly imaging and natural language data. The complex nature of information derivation from such data has posed great challenges to other machine learning methods but has been handled well by ANNs. Similarly, high-throughput biological data such as next-generation sequencing, metabolomic data, proteome data, and electron microscopic structural data, has raised equally challenging computational problems. Second, computational power has been increasing rapidly with affordable costs, including the development of new computing devices, such as GPUs and FPGAs. Such devices provide ideal hardware platforms for highly parallel models. Third, a range of proposed optimization algorithms have made deep ANNs stand out as an ideal technique for large and complex data analyses and information discovery compared to competing techniques in the big data era. Here are also some problems in the bioinformatics field as follows which need to be tackled. First, the interpretability of model is essential to biologists to understand how model helps solve the biological problem, e.g. predicting DNA-protein binding (Luo et al. 2019). Second, the clinical expect accuracy of computational model related to the healthcare or disease diagnosis is ~98%–99% and it is tough to reach that high accuracy. Moreover, two fundamental breakthroughs have tremendously increased the applicability of ANN techniques: convolutional neural networks (CNNs) for imaging data and recurrent neural networks (RNNs) for natural language data, which will be introduced in the supplementary materials with other well-known architectures. We surveyed the literature and tabulated the number of publications in log-scale for 14 commonly studied biological topics appearing together with ‘RNN’, ‘CNN’, or ‘deep learning’ according to PubMed, which are detailed in Figure 1. As expected, ‘image’ is the most commonly approached topic by DL, and ‘disease’ and ‘imaging’ follow closely. CNNs are much more frequently used in bioinformatics than

RNNs because CNNs can easily capture local features, solving fundamental issues, such as identifying and applying conserved sequence motifs.

Here, we focus on the ongoing trends and future directions of modern DL, perspective on future developments and potential new applications to biology and biomedicine.

Current Trend in Principled DL

Attention mechanism

Attention mechanisms, which were first proposed to conduct machine-based translation tasks (Vaswani et al. 2017), can alleviate the problems faced by RNNs when applied to bioinformatics problems, thus expanding their domain of applications in bioinformatics. The self-attention layer can translate the original representation of an input sequence (e.g. one-hot encoding for RNA, DNA, or protein sequences) into another representation of the sequence. For each position in the sequence, the other positions in the input sequence try to better characterize that position for capturing the semantic meaning of the sequence and interactions between different sequential positions.

Attention mechanisms can potentially be used in a wide range of biosequence analysis problems, such as RNA sequence analysis and prediction (Park et al. 2017), protein structure and function prediction from amino acid sequences (Zou et al. 2018), and identification of enhancer–promoter interactions (EPIs) (Hong et al. 2020). For example, EPIs show great significance to human development because they are critical to the regulation of gene expression and are closely related to the occurrence of human diseases. However, experimental methods to identify EPIs require too much time, manpower, and money. EPIVAN (Hong et al. 2020) was designed to predict long-range EPIs using only genomic sequences via DL methods and attention mechanisms. This method has been tested on six cell lines, and the AUROC (area under the ROC) and AUPR (area under the PR curve) values of EPIVAN are higher than those without the attention mechanism, which indicates that the attention mechanism is more concerned with cell line-specific features and can better capture the hidden information from the perspective of sequences.

Reinforcement learning

Reinforcement learning (Mnih et al. 2015) considers what actions to take, given the current state of the partial solution to maximize the cumulative reward. After each action, the state can change. Observations about the set of change-of-state become guiding information for future actions. This type of reinforcement learning has recently been incorporated into the DL paradigm, referred to as deep reinforcement learning (DRL). Note that a key distinguishing feature is that users do not have to predefine all the states, and a model can

be trained in an end-to-end manner, which has become an increasingly active research field with numerous algorithms being developed.

Reinforcement learning can be applied in collective cell migration (Hou et al. 2019), DNA fragment assembly (Bocicor et al. 2012), and characterizing cell movement (Wang et al. 2018). DNA fragment assembly is a technique that aims to reconstruct the original DNA sequence from a large number of fragments by determining the order in which the fragments have to be assembled back into the original DNA molecule, and it is also an NP-hard optimization problem. Bocicor et al. (2012) proposed a new reinforcement learning-based model for solving this problem. Reinforcement learning in this problem was formulated as training the agent to find a path during assembling fragments from the initial to a final alignment state, maximizing the performance measure, one of the fitness functions, which sums the overlap scores over all adjacent fragments. This reinforcement learning model shows less computational complexity and unnecessary external supervision in the learning process compared with the genetic algorithm and supervised approach, respectively.

Few-shot learning

Although there is a large amount of data in the bioinformatics field (Li et al. 2019), data scarcity still occurs in biology and biomedicine. For example, under the enzyme commission (EC) classification (Li et al. 2017a), only one enzyme belongs to the class of phosphonate dehydrogenase (EC 1.20.1.1). In this case, standard DL algorithms cannot work because one needs numerous data for each class to train a generalizable DL model (Li et al. 2018). Few-shot learning, as its name indicates, is designed to handle these cases. In principle, few-shot learning trains a machine learning model with a very small quantity of data. In extreme cases, there is only one training sample for one class, referred to as one-shot learning (Fei-Fei et al. 2006). Similarly defined is zero-shot learning (Socher et al. 2013) when a class has no training sample. Using few-shot learning algorithms, a model can be trained with reasonable performance on some difficult problems by utilizing only the existing limited data.

Few-shot learning is suitable for many problems in bioinformatics that have limited data, such as protein function prediction (Li et al. 2017a) and drug discovery (Joslin et al. 2018). For instance, the drug discovery problem is to optimize the candidate molecule that can modulate essential pathways to achieve therapeutic activity by finding analog molecules with increased pharmaceutical activity. Due to the limitation of small biological data, it is challenging to form accurate predictions for novel compounds. As we searched, one-shot learning has been used to significantly lower the quantity of data required and achieves precise predictions in drug discovery (Altae-Tran et al. 2017). The method proposed in this work combines iterative refinement long short-term memory (LSTM) and graph convolutional neural networks and can improve the learning of meaningful distance metrics

over small molecules. Iterative refinement LSTMs can generalize to new experimental assays related but not identical to assays in the training collection, and graph convolutional networks are useful for transforming small molecules into continuous vectorial representations. The results of applying one-shot models to a number of assay collections show strong performance compared to other methods, such as random forest and graph convolutional neural networks. Consequently, this one-shot method is capable of transferring information between related but distinct learning tasks.

Deep generative models

In biology, high-throughput omic data tend to have high dimensionality and be intrinsically noisy, such as single-cell transcriptomic data (Lopez et al. 2018). The widely used dimensionality reduction methods, such as principal component analysis (PCA), may not work well with such data because of those properties. Deep generative models, such as variational autoencoders (VAEs)(Doersch 2016), are powerful networks for information derivation using unsupervised learning, which has achieved remarkable success in recent years. Generally, it is almost impossible to model the exact distributions of any property of such datasets; those methods are designed to model an approximate distribution that is as similar to the true distribution as possible, implicitly or explicitly. When training a VAE, a low-dimensional latent representation of the raw data with latent variables can be learned, which were assumed to generate the real data. Those generated samples, which do not exist in the real world, can be useful for various biological data modeling problems, such as drug design and protein design.

Deep generative models can be applied to problems related to protein structure design (Anand and Huang 2018)(Ingraham et al. 2019), 3D compound design (Imrie et al. 2019), protein loop modeling (Li et al. 2017b), and DNA design (Killoran et al. 2017). The structure and function of proteins is a key feature of understanding biology at the molecular and cellular levels. However, there might be missing regions that need to be reconstructed, and the prediction of those missing regions is also called the loop modeling problem. A generative adversarial network (GAN) is applied for this problem, which can capture the context of the loop region and predict the missing area (Li et al. 2017b). The 3D protein structure is represented by the 2D distance map in which each value is a real Euclidean distance of Cα atoms of two amino acids. The root-mean-square deviation score of their GAN method has 44% improvement compared to other tools, and their GAN method obtains the smallest standard deviation compared to other tools, which show the stability of their prediction.

Meta learning

Meta learning (Finn et al. 2017), also known as 'learn-to-learn', attempts to produce such models, which can quickly learn a new task with a few training samples based on models trained for related tasks. A good meta learning model should generalize to a new task even if the task has never been encountered during the training time. The key idea is that when training a model is finished, the model needs to be exposed to a new task during the testing phase, several steps of fine-tuning are performed, and then the model's performance on the new task is checked. In brief, meta learning outputs a machine learning model that can learn quickly.

For instance, the ability of an antibody to respond to an antigen depends on the antibody's specific recognition of an epitope (Hu et al. 2014). Thus, meta learning can be used in B-cell conformational epitope prediction in continuously evolving viruses, which is useful for vaccine design. The proposed meta learning approach is based on stacked and cascade generalizations. In the hierarchical architecture, the meta learner of each level will input the meta features outputted from a low level and output the meta features to successive levels until the top level which will output the final classification result. Low correlation among these meta learners indicates that these learners truly have complementary predictive capabilities, and the ablation analysis indicates that these learners differentially interacted and contributed to the final meta model. Consequently, the meta learner can analyze the complementary predictive strengths in different prediction tools and integrate these tools to outperform the single best-performing model through meta learning.

Symbolic reasoning empowered DL

It is noteworthy that until recently, DL has yet to include symbolic reasoning or logic as part of its toolkit, hence having omitted the essential information provided by logic reason and the associated explainability (Hu et al. 2016). In recent years, ML researchers have developed a number of methods to incorporate symbolic reasoning with DL. For example, SATNet (Wang et al. 2019) uses a differentiable satisfiability solver to bridge DL and logic reasoning; NLM (Hamilton et al. 2018) exploits the power of both DL and logic programming, utilizing it to perform inductive learning and logic reasoning efficiently.

In the bioinformatics field, symbolic reasoning is applied and evaluated on structured biological knowledge, which can be used for data integration, retrieval, and federated queries in the knowledge graph (Alshahrani et al. 2017). This method combines symbolic methods, in particular, knowledge representation using symbolic logic and automated reasoning, with neural networks that encode for related information within knowledge graphs, and these embeddings can be applied to predict the edges in the knowledge graph, such as drug target relations. The performance combining symbolic methods outperforms traditional approaches.

Conclusion

DL is a relatively new field compared to traditional machine learning, and the application of DL in bioinformatics is an even newer field. However, the last decade has witnessed the rapid development of DL with thrillingly promising power to mine complex relationships hidden in large-scale biological and biomedical data. In this article, we reviewed some selected modern and principled DL methodologies, some of which have recently been applied to bioinformatics, while others have not yet been applied. This perspective may shed new light on the foreseeable future applications of modern DL methods in bioinformatics.

[Supplementary material is available at Journal of Molecular Cell Biology online. The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST), under award numbers FCC/1/1976-18-01, FCC/1/1976-23-01, FCC/1/1976-25-01, FCC/1/1976-26-01, URF/1/3450-01-01, URF/1/3454-01-01, URF/1/4098-01-01, URF/1/4077-01-01, and REI/1/0018-01-01. Y.X. and X.G. conceived the study; H.L., S.T., and Y.L. wrote the paper together; Q.F., R.T., Y.P., and C.H. contributed materials and critical revisions to the paper.]

Reference

Alshahrani M, Khan MA, Maddouri O, et al (2017) Neuro-symbolic representation learning on biological knowledge graphs. Bioinformatics 33, 2723–2730.

Altae-Tran H, Ramsundar B, Pappu AS, et al. (2017) Low Data Drug Discovery with One-Shot Learning. ACS Cent Sci 3, 283–293.

Anand N, and Huang P (2018) 'Generative modeling for protein structures'. In: Advances in Neural Information Processing Systems 31: 32nd Conference (NeurIPS 2018), Montreal, Canada. pp 7494–7505. La Jolla, USA: Neural Information Processing Systems (NIPS).

Bocicor M-I, Czibula G, and Czibula I (2012) 'A Reinforcement Learning Approach for Solving the Fragment Assembly Problem'. In: 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Timisoara, Romania, 2011. pp 191–198. Los Alamitos, USA: IEEE Computer Society.

Doersch C (2016) Tutorial on variational autoencoders. arXiv, https://arxiv.org/abs/1606.05908.

Fei-Fei L, Fergus R, and Perona P (2006) One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 28, 594–611.

Finn C, Abbeel P, and Levine S (2017) 'Model-agnostic meta-learning for fast adaptation of deep networks'. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017. PMLR 70, 1126–1135. http://www.jmlr.org/.

Hamilton W, Bajaj P, Zitnik M, et al (2018) 'Embedding logical queries on knowledge graphs'. In: Advances in Neural Information Processing Systems 31: 32nd Conference (NeurIPS

2018), Montreal, Canada. pp 2026–2037. La Jolla, USA: Neural Information Processing Systems (NIPS).

Hong Z, Zeng X, Wei L, et al. (2020) Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. Bioinformatics 36, 1037–1043.

Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural networks 4, 251–257.

Hou H, Gan T, Yang Y, et al (2019) Using deep reinforcement learning to speed up collective cell migration. BMC Bioinformatics 20, 571.

Hu Y-J, Lin S-C, Lin Y-L, et al (2014) A meta-learning approach for B-cell conformational epitope prediction. BMC Bioinformatics 15, 378.

Hu Z, Ma X, Liu Z, et al (2016) Harnessing deep neural networks with logic rules. arXiv, https://arxiv.org/abs/1603.06318.

Imrie F, Bradley AR, van der Schaar M, et al. (2019) Deep Generative Models for 3D Compound Design. BioRxiv, doi: https://doi.org/10.1101/830497.

Ingraham J, Garg VK, Barzilay R, et al. (2019) 'Generative Models for Graph-Based Protein Design'. In: Advances in Neural Information Processing Systems 32: 33rd Conference (NeurIPS 2019), Vancouver, Canada. pp 15741–15752. La Jolla, USA: Neural Information Processing Systems (NIPS).

Joslin J, Gilligan J, Anderson P, et al (2018) A fully automated high-throughput flow cytometry screening system enabling phenotypic drug discovery. SLAS Discov Adv Life Sci R&D 23, 697–707.

Killoran N, Lee LJ, Delong A, et al (2017) Generating and designing DNA with deep generative models. arXiv, https://arxiv.org/abs/1712.06148.

Li Y, Huang C, Ding L, et al (2019) Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods 166, 4–21.

Li Y, Li Z, Ding L, et al (2018) Supportnet: solving catastrophic forgetting in class incremental learning with support data. arXiv, https://arxiv.org/abs/1806.02942.

Li Y, Wang S, Umarov R, et al (2017a) DEEPre: sequence-based enzyme EC number prediction by deep learning. Bioinformatics 34, 760–769

Li Z, Nguyen SP, Xu D, et al. (2017b) 'Protein Loop Modeling Using Deep Generative Adversarial Network'. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI 2017), Boston, USA. pp 1085–1091. New York, USA: IEEE.

Lopez R, Regier J, Cole MB, et al (2018) Deep generative modeling for single-cell transcriptomics. Nat Methods 15, 1053.

Luo X, Tu X, Ding Y, et al (2019) Expectation pooling: An effective and interpretable pooling method for predicting DNA-protein binding. BioRxiv, doi: https://doi.org/10.1101/658427.

Mnih V, Kavukcuoglu K, Silver D, et al (2015) Human-level control through deep reinforcement learning. Nature 518, 529

Park S, Min S, Choi H-S, et al. (2017) 'Deep recurrent neural network-based identification of precursor micrornas'. In: Advances in Neural Information Processing Systems 30: 31st Annual Conference (NIPS 2017), Long Beach, USA. pp 2891–2900. La Jolla, USA: Neural Information Processing Systems (NIPS).

Socher R, Ganjoo M, Manning CD, et al. (2013) 'Zero-shot learning through cross-modal transfer'. In: Advances in neural information processing systems 26: 27th Annual Conference 2013, Lake Tahoe, USA. pp 935–943. La Jolla, USA: Neural Information Processing Systems (NIPS).

Vaswani A, Shazeer N, Parmar N, et al (2017) 'Attention is all you need'. In: Advances in neural information processing systems 30: 31st Annual Conference (NIPS 2017), Long Beach, USA. pp 5998–6008. La Jolla, USA: Neural Information Processing Systems (NIPS).

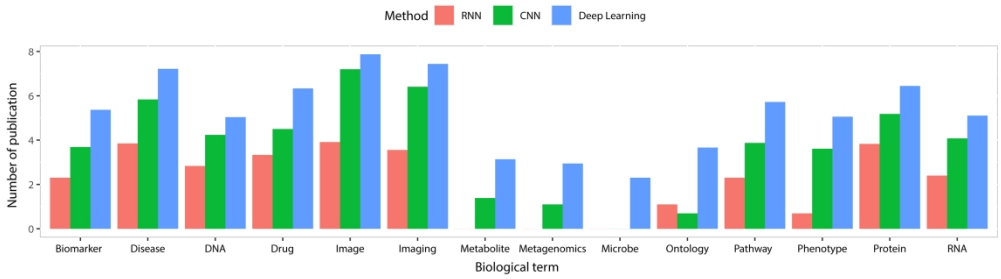
Wang P-W, Donti PL, Wilder B, et al. (2019) SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. arXiv, https://arxiv.org/abs/1905.12149.

Wang Z, Wang D, Li C, et al (2018) Deep reinforcement learning of cell movement in the early stage of C.elegans embryogenesis. Bioinformatics 34, 3169–3177.

Zou Z, Tian S, Gao X, et al. (2018) mldeepre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning. Front Genet 9, 714

Figure legend

Figure 1 Number of publications (log-scale) for 14 biological topics. For each topic, the three bars show the number of publications mentioning the terms ‘RNN’, ‘CNN’, and ‘deep learning’, respectively.



291x83mm (300 x 300 DPI)