様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １７ 日現在

機関番号：８２４０１
研究種目：若手研究
研究期間：2018～2019
課題番号：１８Ｋ１８１５６

研究課題名（和文）A Deep Learning framework for cancer precision medicine

研究課題名（英文）A Deep Learning framework for cancer precision medicine

研究代表者
Ｌｙｓｅｎｋｏ　Ａｒｔｅｍ（Lysenko, Artem）

国立研究開発法人理化学研究所・生命医科学研究センター・研究員

研究者番号：８０７５３８０５

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：肝細胞癌（HCC）は世界で3番目に顕著な癌であり、この癌の不均一性により治療が困難になるため、関連データの分析に計算手法を使用することが非常に重要です。このプロジェクトの目的は、人工知能の最近の進歩を利用して、このタイプの癌をよりよく理解し、より効果的な治療法の開発を促進することです。この目標は、高次元の癌オミクスプロファイリングデータから癌の転帰を予測するための新しいタイプのディープラーニングアーキテクチャを開発することで達成されました。次に、他の計算方法と組み合わせて、患者の転帰に影響を与えるすべての要因を含む的に調査しました。

研究成果の学術的意義や社会的意義
The main contribution of this project is in making advances in computational analysis methods for large biomedical cancer datasets. These innovations will potentially lead to new discoveries necessary for better cancer diagnosis and treatment strategies.

研究成果の概要（英文）：Hepatocellular carcinoma (HCC) is a third most prominent cancer world-wide and is characterized by very high tumor heterogeneity making development of effective treatments particularly challenging. The problem of HCC treatment naturally fits into precision medicine paradigm, where different sub-types of the disease are identified by computational analysis and treatments are customized using this prior knowledge to achieve optimal outcomes.
The aim of this project is to facilitate better understanding of this type of cancer and development of more effective treatments by leveraging recent advances in Artificial Intelligence. This goal was achieved by developing a new type of deep learning architecture for predicting cancer outcome from high-dimensional cancer ‘omics profiling data, which was then applied in conjunction with other computational methods to comprehensively explore whole range of factors affecting patient outcomes.

研究分野：Computational medicine

キーワード：Deep Learning　Precision Medicine　Cancer　Multiomics

１．研究開始当初の背景

Hepatocellular carcinoma (HCC) is a third most prominent cancer world-wide and is characterized by very high tumor heterogeneity making development of effective treatments particularly challenging. The problem of HCC treatment naturally fits into precision medicine paradigm, where different sub-types of the disease are identified by computational analysis and treatments are customized using this prior knowledge to achieve optimal outcomes. The aim of this project was to facilitate better understanding of this type of cancer using both established and novel approaches, with particular focus on the development of a new method by leveraging recent advances in Artificial Intelligence and Deep Learning. This was done by developing a novel Deep Artificial Neural Networks that for the first time successfully employed a meta-learning approach in survival analysis of high-dimensional cancer transcriptomics data. From the more general technological perspective, the aspiration was to facilitate the application of Deep Learning algorithms in biomedical domain by developing an architecture that can have far greater synergy with the types and structure of such data. This was motivated by an observation that outside of medical image analysis, application of DL in biomedicine has proven particularly challenging (Miotto, et al., 2017). Given that DL models are so powerful because of their complexity, as a consequence they also typically require vast numbers of annotated samples to parametrize, whereas in typical multi-omics studies these numbers are usually at best in the low hundreds. Amount of available data is most frequently a key limitation preventing application of DL in biomedical domain.

However, recent research in DL has identified some promising strategies for relaxing this limitation, in particular a family of approaches known as meta-learning. In principle, meta-learning can be very effective in some specific cases where very small amounts of data are available (Finn, et al., 2017). Meta-models are parametrized using much wider collections of somewhat related data – or even thematically similar types of outcomes -- and then tailored to the specific case of interest. The purpose of the meta-learning model is to specifically capture the overarching finer patterns common to all of the sub-tasks on which it is trained rather than to optimally solve just one of those tasks. Though, of course, then optimal performance in a specific task can then be achieved by tuning this generalized model on some more focused set of exemplar samples. Crucially, even very small number of samples can typically be sufficient for this final tuning step.

２．研究の目的

From the technology development perspective, this project has comprehensively explored the possibility of adapting the meta-learning DNN for applications in predictive modelling of cancer survival and disease progression, with the goal of answering key scientific question of how can meta-learning deep neural network algorithms be used to develop high-quality predictive models for biomedical applications. Whereas from the biomedical perspective, the objective was to specifically focus on solving current limitations specific to the areas of cancer precision medicine, with particular aspiration of gaining new insights into anti-cancer immune response in hepatocellular carcinoma type

of cancers. The overlap of these two directions resulted in the focus of the work being eventually consolidated around the following intermediate tasks:

(1) Discovery and identification of immune-related patterns in cancer multi-omics datasets of interest. As immune response-specific data was not sufficiently profiled by the experimental methods, it had to be inferred from the transcriptomics by appropriate inference methods, in particular immune cell proportions, known sub-types of immune responses and immunologically hot/cold tumor categorizations. Additionally, other potentially informative types of analysis included pathway activity-based methods (PARADIGM, SPIA) and “immunoduct” tumor immunology pipeline.

(2) Development of a mathematical formalism suitable for realizing survival analysis on censored time-to-event data using Deep Learning networks. Disease-free and overall survival are among the most important clinical outcomes in oncology, however preliminary review done during early stages of this project indicated so far only very few such methods have been proposed and that they also did not offer sufficiently good levels of performance in case of high-dimensional multi-omics datasets.

(3) Development of a meta-learning framework that can allow this formalism to be trained on a broader collection of the data; demonstrating that the proposed approach can successfully be used on real data and can achieve superior levels of performance.

３．研究の方法

Due to very high computational requirements of Deep Learning algorithms, the first hurdle the overcome was to acquire and deploy the necessary computer hardware sufficiently powerful to complete the planned experiments. In fact, almost all of the funding awarded for this project went towards these equipment costs. To that end, in the first year of the project a specialized GPU server was purchased, which was further upgraded in the second year (made possible by a decline in price of GPU graphics cards at that time). The implementation and training of the Deep Learning models produced during this work was done on this system, using the CUDA / TensorFlow + Keras DL software stack in R and Python 3 programming environments.

In terms of the data, in line with the original plan, the work has used the extensive collection of public cancer data, chiefly from The Cancer Genome Atlas Program (TCGA). In all, after QC this dataset had necessary variables (transcriptomics and overall / disease-free survival annotations) for 32 different cancer types across 10,002 individual samples – a sufficient size to reasonably expect an additional performance boost from using Deep Learning over other, simpler methods. At the same time, an additional smaller closed-access hepatocellular carcinoma dataset from another group at RIKEN was also analyzed, with particular emphasis on discovering immunological features that can influence survival and can also be replicated in TCGA hepatocellular carcinoma (LIHC) cancer type.

After evaluating several possible meta-learning designs from the most recent meta-learning studies, it was decided that a two-branch “Siamese”-type architecture showed the greatest promise. Some advantages of this meta-learning architecture are in its ability to: (1) discover meaningful embeddings for the data, which can then be potentially used to identify clusters/subtypes for the samples; (2) summarize multiple features(genes) into distinctive components that could be helpful in identifying potential mechanisms. An additional work has then been done to develop a survival modelling strategy specifically compatible with this formalism, and finally experiments were conducted to refine and optimize the design using real data (types and numbers of layer, type of regularization, design of the connection between two branches, best training strategy and, finally, hyperparameter optimization).

4．研究成果

The analysis of the HCC dataset from the immunological perspective did identify several potentially interesting patters that were not previously reported for this dataset. Specifically, it was possible to confirm that “immunologically hot” HCC tumor type was associated with much better prognosis for overall survival and appeared to be an important factor for this caner type. Among all considered methods, the best possible immunologically-relevant characterization was achieved based on the gene signature

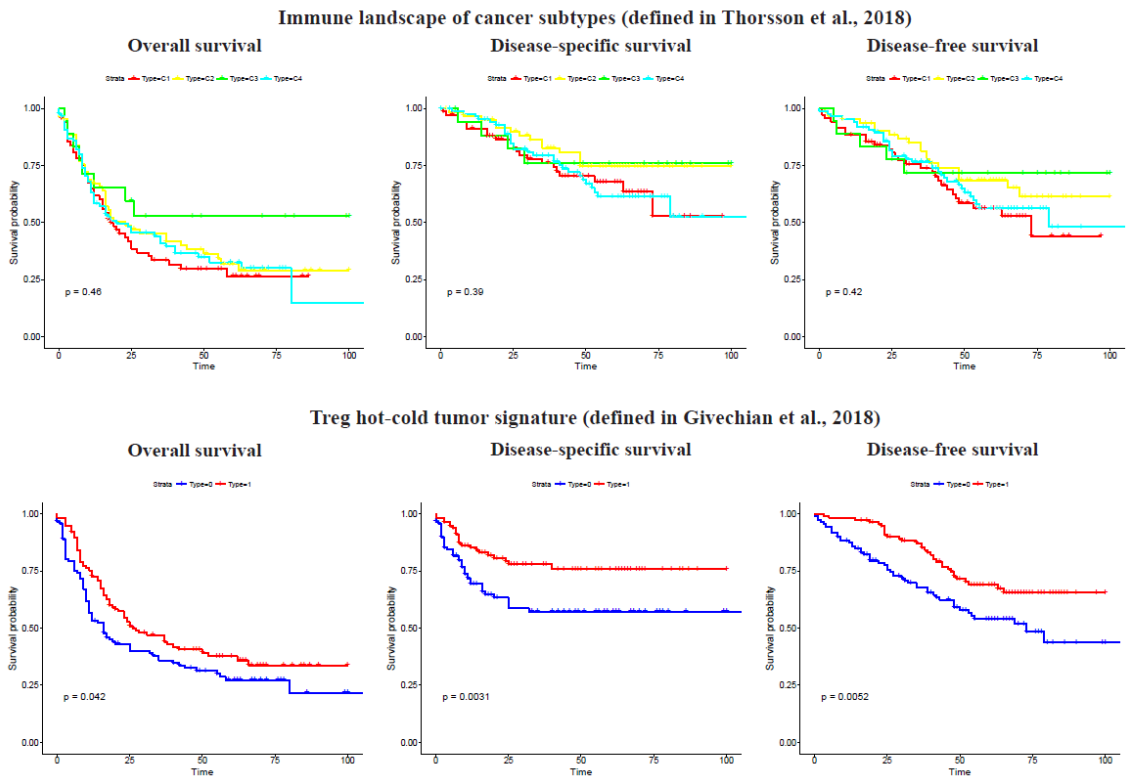

Figure 1. An example of some immunological features predictive of survival identified during the analysis of an original HCC dataset. Substantial number of better prognosis samples with the “Inflammatory” subtype (top panel, light-green) were also assigned a “hot tumor” subtype (bottom panel, red). However, the overlap was only partial, suggesting that neither categorization currently offers the best possible accuracy vs. granularity tradeoff.

from the Givechian, et al. (2018) study, and the pattern was also successfully replicated in the LIHC dataset from TCGA. Profiling with respect to known sub-types identified partial overlap with a known improved prognosis subtype from “Immune Landscape of Cancer” categorization and consistent patterns were observed with respect to inferred immune cell proportions (e.g. CIBERSORT and other similar methods), though for HCC categorization derived from this signature had much better predictive performance that any of these other approaches (Fig. 1).

The developed meta-learning DNN architecture for survival analysis was evaluated on the TCGA datasets and several additional, stand-alone cancer studies. It was possible to demonstrate that both generalized and fine-tuned models are capable of achieving good results. When evaluated for its ability to predict overall and disease-free survival for a particular cancer type using data exclusively from other cancer types, the best-performing model was able to reach Somer’s D scores of at least above 0.4 for the majority of all 32 cancer types in both cases. Generally, the performance was better than several other leading algorithms that where compared in similar setups (RandomForestSRC, XGBoost and Coxnet). Interestingly, even a generalized model (i.e. the one prior to being fine-tuned on the samples from cancer type for which it was evaluated) was capable of producing good results, suggesting that it may have identified a potentially large set of common survival determining mechanisms across all cancer type. Currently a methodology-focused paper describing the model architecture development and these outcomes is in preparation, in particular one aspect that it comprehensively explores is the role of anti-cancer immune response with respect to the outcome groups suggested by the DNN model, and thus consolidates the two sets of outcomes produced during this project.

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　7件／うち国際共著　7件／うちオープンアクセス　7件）

１．著者名 Sharma Alok*、Lysenko Artem*、Lopez Yosvany、Dehzangi Abdollah、Sharma Ronesh、Reddy Hamendra、Sattar Abdul、Tsunoda Tatsuhiko	４．巻 19
２．論文標題 HseSUMO: Sumoylation site prediction using half-sphere exposures of amino acids residues	５．発行年 2019年
３．雑誌名 BMC Genomics	６．最初と最後の頁 1-7
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1186/s12864-018-5206-8	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Lysenko Artem、Sharma Alok、Boroevich Keith A、Tsunoda Tatsuhiko	４．巻 1
２．論文標題 An integrative machine learning approach for prediction of toxicity-related drug safety	５．発行年 2018年
３．雑誌名 Life Science Alliance	６．最初と最後の頁 1-14
掲載論文のDOI（デジタルオブジェクト識別子） 10.26508/lsa.201800098	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Saqi Mansoor*、Lysenko Artem*、Guo Yi-Ke、Tsunoda Tatsuhiko、Auffray Charles	４．巻 20
２．論文標題 Navigating the disease landscape: knowledge representations for contextualizing molecular signatures	５．発行年 2018年
３．雑誌名 Briefings in Bioinformatics	６．最初と最後の頁 609～623
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1093/bib/bby025	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Choobdar Sarvenaz、The DREAM Module Identification Challenge Consortium (including Lysenko, Artem)、et al.	４．巻 16
２．論文標題 Assessment of network module identification across complex diseases	５．発行年 2019年
３．雑誌名 Nature Methods	６．最初と最後の頁 843～852
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1038/s41592-019-0509-5	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Menden Michael P. AstraZeneca-Sanger Drug Combination DREAM Consortium (including Lysenko, Artem)、et al.	4．巻 10
2．論文標題 Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen	5．発行年 2019年
3．雑誌名 Nature Communications	6．最初と最後の頁 2674
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1038/s41467-019-09799-2	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Janowska-Sejda Elzbieta I.*、Lysenko Artem*、Urban Martin*、Rawlings Chris、Tsoka Sophia、Hammond-Kosack Kim E.	4．巻 10
2．論文標題 PHI-Nets: A Network Resource for Ascomycete Fungal Pathogens to Annotate and Identify Putative Virulence Interacting Proteins and siRNA Targets	5．発行年 2019年
3．雑誌名 Frontiers in Microbiology	6．最初と最後の頁 2721
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.3389/fmicb.2019.02721	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Mason Mike J.、Multiple Myeloma DREAM Consortium (including Lysenko, Artem)、et al.	4．巻 1
2．論文標題 Multiple Myeloma DREAM Challenge reveals epigenetic regulator PHF19 as marker of aggressive disease	5．発行年 2020年
3．雑誌名 Leukemia	6．最初と最後の頁 1~9
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1038/s41375-020-0742-z	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計3件（うち招待講演　1件／うち国際学会　0件）

1．発表者名 Artem Lysenko
2．発表標題 Machine learning-driven analysis of biological networks for predictive modelling of drug toxicity
3．学会等名 IB-2018, Harpenden, UK（招待講演）
4．発表年 2018年

１．発表者名 Lysenko, Artem, Keith A. Boroevich, Tatsuhiko Tsunoda
２．発表標題 Towards computational drug screening: profiling drug toxicity in the context of a biological network (poster)
３．学会等名 ISMB /ECCB Basel, Switzerland
４．発表年 2019年

１．発表者名 Lysenko, Artem, Keith A. Boroevich, Tatsuhiko Tsunoda
２．発表標題 Towards discovery of human disease mechanisms by graph-based contextual integration of ‘omics signatures (poster)
３．学会等名 ISCB, OIST, Okinawa
４．発表年 2019年

〔図書〕　計1件

１．著者名 Lysenko, Artem, Keith A. Boroevich, Tatsuhiko Tsunoda	４．発行年 2019年
２．出版社 Springer Nature	５．総ページ数 16
３．書名 "Genotyping and Statistical Analysis" in Genome-Wide Association Studies (book chapter)	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考