The link between gene expression and machine learning

openaccessgovernment.org

24 March 2023

Professor Y-h. Taguchi uses tensor decomposition to identify genes associated with altered gene expression caused by drug treatment

Machine learning has attracted much interest lately. Last year, generative AIs such as stable diffusion, which can create beautiful images from a given prompt and ChatGPT, which can mimic human conversations very well, were developed. All these popular AIs are based on a machine learning technique called deep learning (DL), inspired by the human brain's structure, also known as neural networks. However, these DL-based generative AIs have a hidden problem: they require massive amounts of data to learn.

ChatGPT is based on 300 billion tokens

For instance, stable diffusion was trained on 2.3 billion images with English captions, while GPT-3.5, which ChatGPT is based on, used 300 billion tokens (word fragments) to learn. Of course, they also needed a long training time and a huge memory size that could only be stored on remote servers, which users could access via the internet. It was reported that GPT-3, the previous version of GPT-3.5, cost four point five billion dollars to train.

At this point, no one is sure if these new generative AIs can produce enough value compared to their incurred cost.

Human genes and machine learning

Unlike these popular topics, some cases are not well-known but important, where we do not have enough samples to learn from. For example, medicine is a typical case that belongs to this category. As we know, humans have tens of thousands of genes. As we also know well, we need more samples than variables to apply machine learning techniques. This is because if the number of samples is less than the number of variables, it is always possible to predict the outcome. For example, suppose that we are required to derive a function that can "predict" the past seven days' weather using ten variables, such as air temperature at morning, noon and evening; air pressure at the same time points; sunshine duration; and so on. If we have more than seven variables, we can always relate weather to any seven variables out of ten variables. This is completely useless since there is no reason that we can use this function to predict future weather based on the currently available ten variables. In this sense, if we want to relate genes to disease outcomes correctly, we need samples from as many as ten thousand people, which is definitely hopeless. Recently, we proposed a mathematical method based on some mathematical

trick [1-2], tensor decomposition, to address this problem. I also reported the application of this method to COVID-19 problems [3-8]. This year I will report other applications of my method in the current and upcoming manuscripts.

Applying our method to in silico drug discovery in cancer

The first example of applying our method to in silico drug discovery, although the publication was delayed significantly, is for cancer [9]. In this study, we applied our method to various publicly available gene expression profiles of cancer cell lines treated with different drug candidate compounds. Our method, tensor decomposition [1,2,8], enables us to identify genes associated with altered expression caused by drug treatment. The identified genes are not necessarily genes of proteins that the small compounds are supposed to bind to. Usually, although compounds can function by binding to proteins, gene expression measured is not that of proteins but that of RNA. RNA is a molecule that can mediate from DNA which consists of the genome, including genetic information, to proteins. Thus, measuring the amount of RNA is not directly related to proteins. We also consider an external database that records gene expression when various genes are suppressed (knock out, KO). Then we try to find which KO gene alters the expression of genes whose expression is changed by drug treatment. This procedure enables us to identify proteins that drugs are supposed to bind to. The identified proteins significantly overlap with those targeted by identified drugs statistically. Thus, our method was effective in identifying effective drugs from gene expression profiles.

Identifying drug-target proteins

Our method has more additional contributions. For example, our data set includes expressions of as few as less than 1,000 genes among more than 20,000 human genes. Despite that, we could identify drug-target proteins outside these 1,000 genes since we identify drug-target proteins based on whether binding to these proteins can affect the expression of these 1,000 genes. In addition, drugs identified are common among various cancer cell lines. Since gene expression profiles of individual cancer cell lines are independent, the fact that our method can identify common drugs as promising ones supports the effectiveness of our methodology. The drugs we identified in our in-silico approach have not been verified experimentally. Please support our efforts to test the efficiency of our identified drugs for cancers *in vitro* or *in vivo*.

References

1. Y-h. Taguchi, Unsupervised feature extraction applied to bioinformatics, Research Outreach, No. 115, pp. 154-157, (2020), [https://doi.org/10.32907/ro-115-154157]
2. Y-h. Taguchi, Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach, Springer International, (2020).
3. Y-h. Taguchi, In Silico Drug Discovery for COVID-19 Using an Unsupervised Feature Extraction Method, Scientia, (2021) [https://doi.org/10.33548/SCIENTIA727]

4. Y-h. Taguchi, How to compete with COVID-19 with a computer? Open Access Government, issue 33. pp 210- 211, (2021)
5. Y-h. Taguchi, Can mice be an effective model animal for Covid-19? Open Access Government, issue 34, April (2022) pp.112-113.
6. Y-h. Taguchi, Is human blood better than cell lines as a COVID-19 infection model? Open Access Government, issue 35, July (2022) pp.182-183.
7. Y-h. Taguchi, Slight changes can improve much for algorithms looking at gene expressions. Open Access Government, issue 36, October (2022) pp.130-131. [https://doi.org/10.56367/OAG-036-10026]
8. Y-h. Taguchi, Kernel Tensor Decomposition can improve the drug discovery process, Open Access Government, isseu 37, Jan (2023) pp.202-203 [https://doi.org/10.56367/oag-037-10026]
9. Y-h. Taguchi, Drug candidate identification based on gene expression of treated cells using tensor decomposition- based unsupervised feature extraction for large-scale data. BMC Bioinformatics 19 (Suppl 13), 388 (2019). [https://doi.org/10.1186/s12859-018-2395-8]

Please Note: This is a Commercial Profile

More About Stakeholder

In scilico drug repositioning targeting SARS-CoV-2

Development of effective drugs toward COVID-19 is urgently required, and so research is being implemented with in scilico drug repositioning.