Discovering B/T Cell-Cancer Antigen Affinity for Targeted Cancer Drug Therapy and Diagnosis Through Deep Neural Network Models

Aryansh Shrivastava[1]
[1]Washington High School, Fremont, CA

Abstract

Cancer, one of the leading causes of death worldwide, derives from an uncontrolled division of abnormal cells in a given part of the body. Targeted immunotherapy is a promising avenue of cancer treatments, galvanizing the body's own immune system to marshal B and T cells against abnormal cell growth, pathologically inhibiting antigen function and proliferation. In the current landscape of cancer immunotherapy, however, medicine industries are belabored with the need to use painstaking trial and error and numerous wet-lab investigations to test amino acid sequences' affinities to cancer antigens. Furthermore, most cancer antigens and the structures of their epitopes are unknown, and although most malignancies can be cured when diagnosed early, organ-specific assessments cannot be used for early-stage cancers. To bridge this gap, I innovate a deep convolutional neural network (CNN) model pipeline, which analyzes the complex amino acid makeup of tumor infiltrating B/T cell receptors based on the relationships of biochemical properties among adjacent amino acids predictive of how these receptor polypeptides fold in three-dimensional space, computing high affinity amino acid sequences to revolutionize both targeted drug discovery and early diagnosis.

Introduction

As in Figure 1, the human adaptive immune system consists of B and T cell lymphocytes that utilize their receptors to identify, attach, and eliminate pathogenic invaders from our bodies (Sompayrac, 2019). The key hypervariable region of their receptors that these lymphocytes rearrange to target and attach with higher and higher affinities to these invaders is the CDR3 region. As such, in cancer patients, the CDR3 regions of cancer-infiltrating lymphocytes undergo numerous genomic differentiation processes prior to protein synthesis, including class switch recombination and somatic hypermutations, to increase their affinity toward cancer antigens (Dong et. al, 2018). Additionally, lymphocytes divide mitotically prior to differentiation, ensuring a wide variety of CDR3 sequences. For instance, B cells possess transmembrane immunoglobulin protein receptors that synaptically bind to antigens by interfacing their CDR3 region with the specific epitopes of antigens, exemplifying the importance of lymphocytes in cancer prevention.

The sum of all B cell receptors (BCRs) or all T cell receptors (TCRs) in the individual cell sample of a patient is called the BCR/TCR repertoire. This temporal snapshot of repertoire analysis of immune receptors can provide valuable insights of the current state of the cancer patient's immune system. The recombination process that rearranges the gene segments for the construction of the receptors is key to the development of the immune response, and the correct formation of the rearranged receptors is critical to their future binding affinity to antigens.

Therefore, there is a functional correlation between the CDR3 region data in BCR/TCR repertoire of cancer patients and the type of cancer (Li et. al, 2016; Hu et. al, 2019; Sompayrac, 2019; Dong et. al, 2018; Yuen et. al, 2016). Deep convolutional neural networks (CNNs) are powerful tools to study functional genomics and protein structures and identify hidden patterns associated with difficult classification problems, achieving better performance than traditional methods (Alipanahi et. al, 2015; Krizhevsky et. al, 2012). Moreover, the Cancer Genome Atlas (TCGA) has compiled big patient BCR and TCR repertoire data across 33 cancer types (1,2) as in Figure 1, which could be partitioned into training and validation datasets for a CNN model once obtained.

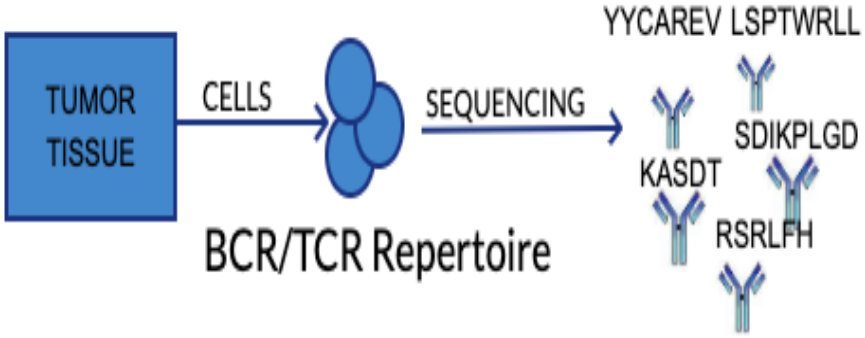

Figure 1. TCGA-compiled BCR and TCR repertoires of CDR3 sequences are sourced from sequencing reads of solid patient tumor tissue.

Materials and Equipment

After designing the architecture of the CNN model pipeline, its implementation requires Python libraries for machine learning, such as NumPy to process and aggregate the data and PyTorch to write a modular implementation easily extensible to more layers. Python libraries such as Pandas and Matplotlib can be used for data visualization and analysis. Based on data size, a system for GPU-accelerated computation such as AWS may be desirable for efficiency in training and validation.

Methods

Data Processing Layer: Analysis and Filtration

The first layer of architecture I articulate is the data processing layer, which should conduct the necessary exploratory data analysis to determine the distribution and trends of CDR3 data. Grouping the data by cancer and patient and counting the number of unique occurrences for each CDR3 sequence as well as its length for a given cancer removing null and duplicate values, allows for the filtration of data. Then, I intend to calculate a probability metric based on the frequency of each unique CDR3 sequence given by the ratio of the number of its unique occurrences within a cancer to the total number of patients in that cancer, and plot histograms for the distributions of length and probability for each given cancer type to analyze the data. The top 10 cancer types should be selected based on the number of patients and unique CDR3 regions per cancer for an ample magnitude of data. Thereafter, I write an API to retrieve data by cancer type,

sequence length, probability, etc., with functions to provide data objects to the CNN model for training and testing. To compare CDR3 sequences of different lengths, I also write padding functions to increase the lengths of shorter CDR3 sequences.

Feature Engineering: Amino Acid Properties as Predictive Markers

It is essential to capture the relationship between the biochemical properties of adjacent amino acids because their interactions will determine how the overall CDR folds in three-dimensional space, into a complementary shape that interfaces with high affinity to antigens of a particular cancer. The Amino Acid (AA) Index [1] catalogs quantitative measures of the biochemical properties of amino acids measured by scientists in vitro, such as electrostatic charge, hydrophobicity, polarity, and frequency into secondary formation of alpha helical coils or beta pleated sheets.

To statistically validate, up to a p value significance of 0.05, that these amino acid biochemical properties could be used as predictive markers to differentiate whether or not CDR sequences are associated with a given cancer type, I apply a z-transform to normalize all biochemical indices, followed by a two-tailed Wilcoxon rank sum test to find the top 50 most significant biochemical indices and a heatmap visualization to ensure a clear distinction between CDRs from a given cancer and CDRs not from that cancer.

After taking the top 50 biochemical properties by statistical significance, I perform a principal component analysis to further reduce the number of features down to 20 independent biochemical features.

Peptide Encoding Layer: Two-Dimensional Biochemical Matrices

The 20 computed biochemical features must now be incorporated into the CNN architecture. I innovate a very flexible encoding mechanism to transform each CDR3, which is an amino acid sequence of length L, into a 2-dimensional biochemical feature matrix with dimensions L x 20. For each of the L constituent amino acids corresponding to columns in the matrix, there are 20 rows of features derived from their biochemical properties. The two-dimensional biochemical encoding matrix can be viewed as an image of pixels, an ideal input architecture for the CNN

Input and Output Layers

The input to the model is the L x 20 encoded biochemical matrix corresponding to a given CDR, while the output of the model is a continuous value between 0 and 1 corresponding to the performance of the input CDR for the given cancer. The design decision of using regression to produce a continuous output rather than classification to produce a binary one stems from the goal of obtaining how likely each patient is to obtain a particular cancer. By the same token, the corresponding probabilities to use as outputs in the training set come from the probability metric in the data processing layer described earlier.

Intermediate Hidden Layers

The intermediate layers in the model are two sets of convolution and pooling layers, which are then fed into dense linear ANN layers with random dropout, ultimately connecting to one output neuron upon application of the softmax function for a regression calculation of probabilities. Internally, with each layer, CNN filters will convolve over receptive fields to generate increasingly specific feature maps, and backpropagation and dropout will finetune the filter parameters to maximize accuracy and performance while minimizing over or underfitting.

Application to Drug Discovery via B Cell Data

Since B cells, not T cells, are the lymphocytes that create antibodies that identify and attach to cancer antigens to kill them, and the end goal of a medicine is exactly to create antibodies, the model must be trained on these BCR CDR3 antibody sequences. Normally, in a machine learning problem, one would go forward and calculate the affinity of a CDR3 region for a particular cancer type, but a key challenge in drug discovery is to go backwards and reverse engineer an array of high affinity amino acid sequence motifs that can be used by industries to streamline drug development.

I implement a random k-mer generator that can generate numerous BCR CDR3 sequences of random amino acids. A large spectrum of k-mers can be generated, input into the model for cancer affinity evaluation, and collected in a hashtable data structure to avoid duplicates.

Then, agglomerative hierarchical clustering by Levenshtein edit distance, a metric that closely resembles the number of somatic hypermutations, can be used to analyze the general overlapping patterns associated with high affinity CDRs based on taking the position weight matrix for each cluster. The position weight matrix tallies, for each amino acid, its frequency of occurrence for a particular position in the CDR sequences in a cluster.

Overall, after motifs are calculated, they can be fed back into the model for further affinity evaluation or even recycled in the pipeline in a selective affinity approach. The overall pipeline is shown in flowchart format in Figure 2.

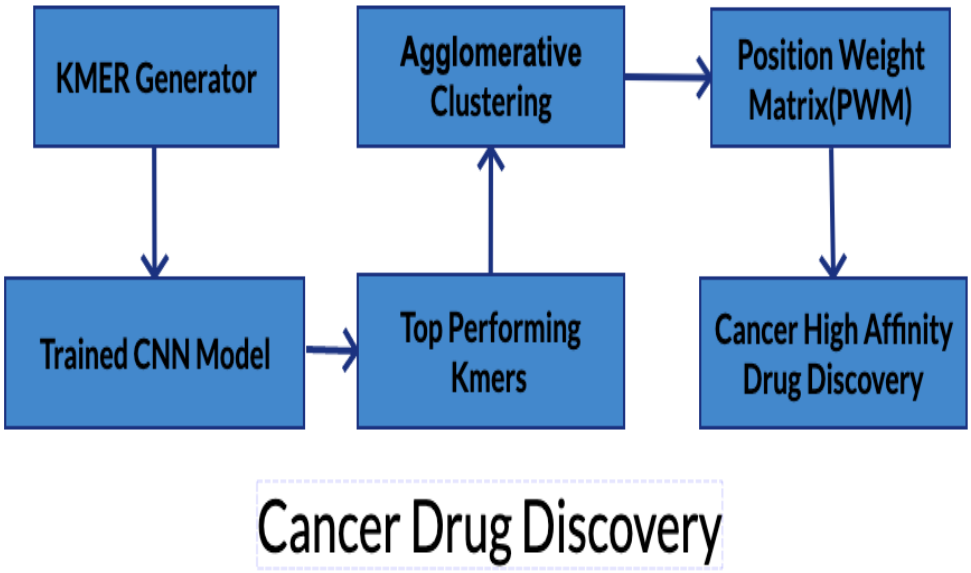

Figure 2. The cancer drug discovery pipeline leverages the trained CNN model to reverse engineer high affinity amino acid sequence motifs

Application to Early Diagnosis via T Cell Data

For very early stage cancers, our T cells already begin to exhibit predictive markers, even though B cells take longer to capture the general motifs to create high affinity antibodies for predictive diagnostic markers (Coulie et. al, 2014). This inspires me to extend my model from B cell receptors to T cell receptors using the same architecture but in a novel diagnostic pipeline. Because my model learns the high affinity TCR sequences with a given cancer type, it can also be used to detect high affinity CDR3 regions from the patient's TCR repertoire, yielding the probability of early cancer diagnosis.

I design a test in which the patient's TCR repertoire is extracted from a peripheral blood test and screened for public CDR3 regions. Then, each CDR is passed through my model and the top N% score is calculated for the top performing CDR3 regions for varying values of N, averaging them together to reduce the sensitivity of false positives. Today, with the advent of sequencing technologies, a peripheral blood test for a TCR repertoire costs only around $200, so this method is very reliable for early cancer diagnosis. The overall pipeline is shown in flowchart format in Figure 3.

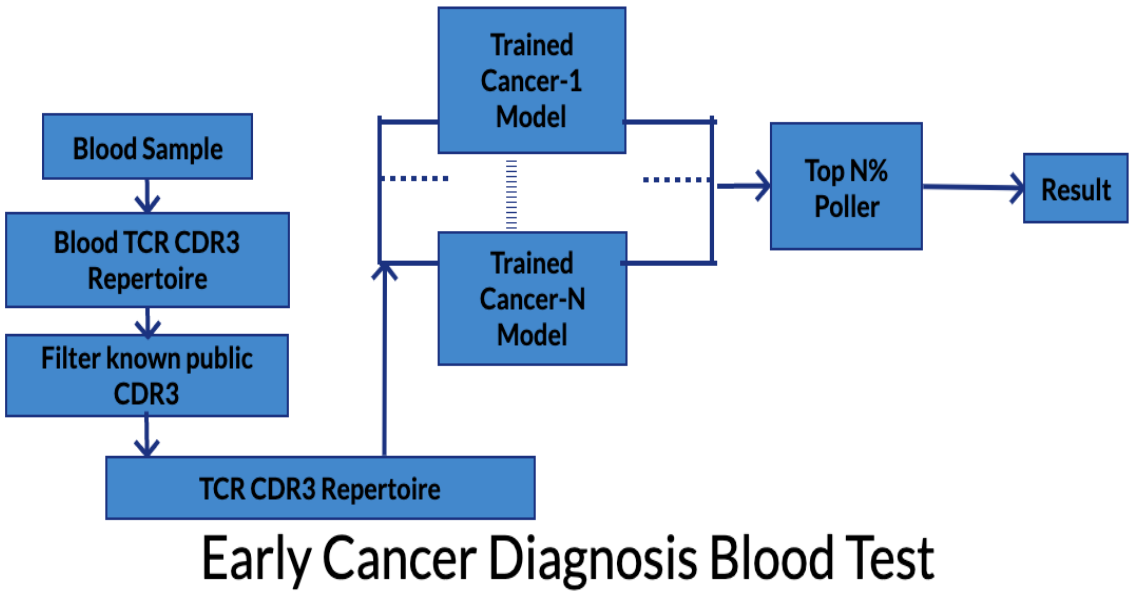

Figure 3. The early cancer diagnosis pipeline leverages a peripheral blood test to extract the TCR CDR3 repertoire of a patient, parallel processing among numerous trained CNN cancer-specific models, and ultimately cumulative polled results of diagnostic probabilities for all cancer types.

Results

As in Figure 4, the overall CNN pipeline is designed with a deep paradigm to include many layers for feature extraction. There is an extensible design because the parameters will be finetuned based on sensitivity and specificity results on the training and validation datasets. In particular, to determine the ideal number of filters for each of the two convolutional layers, the Loss vs. Epoch and receiver operating characteristic (ROC) graphs for the training and validation sets can be plotted to prevent over or underfitting. It can be inferred that convolutional layers toward the end of the model will have more filters than those toward the beginning due to the increasing number and specificity of features. Similarly, the error function to penalize loss and the inclusion of a random dropout for the linear ANN layers can be evaluated as well.

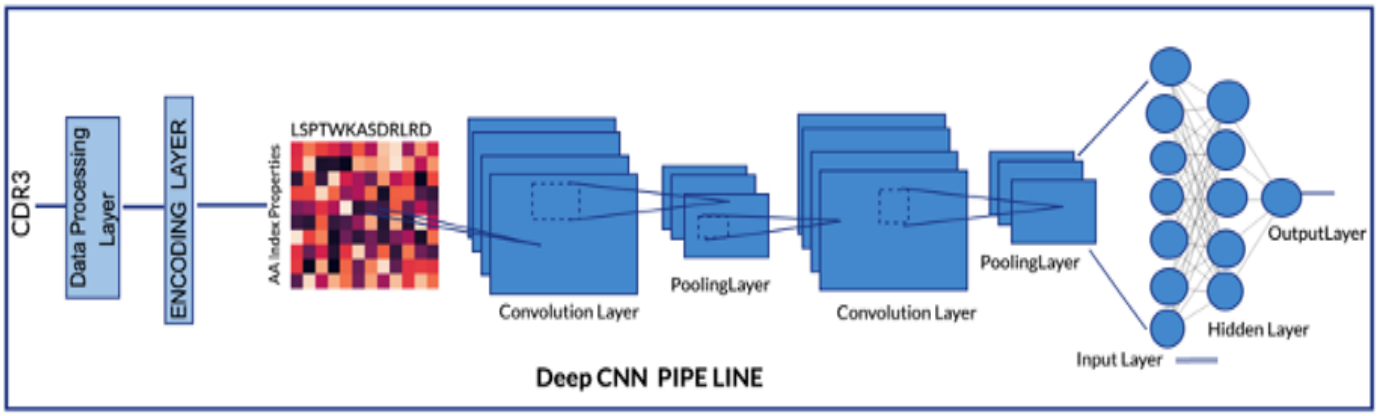

Figure 4. The complete deep CNN pipeline takes in a CDR3 as input, converts it into a two-dimensional biochemical matrix, processes it to extract features, and outputs a vector of probabilities across the 33 different cancer types.

Discussion

In conclusion, a novel CNN pipeline for both cancer drug discovery and early diagnosis is architecturally developed, along with conceptual analyses for training and validation. When trained and validated using existing immune repertoire datasets, the pipeline revolutionizes cancer treatment with two facets. First, high affinity CDR3 motifs can be reverse-engineered to guide the process of industrial drug discovery and development. Despite the fact that industries today have the tools to validate whether a given drug has the proper physical properties for immunotherapy, they lack a method beyond trial and error to identify promising inhibitory sequences. In contrast, this model is very effective in generating a broad spectrum of drug sequences based on the overlapping patterns of existing drugs, a strong alternative to expedite the process. Second, a quick, accurate, and noninvasive method for the early detection of cancer through the immune report of a $150-200 peripheral blood test can be developed by extending the same CNN model architecture to TCR data. Access to such an early diagnosis is an integral factor for some cancer treatments, such as pancreatectomy for pancreatic ductal adenocarcinoma.

In the future, ADMET analyses such as lipophilicity and solubility can be added in a separate layer at the end of the drug design process to ensure bioavailability and safety of human administration. In addition, other predictive markers such as bond length and bond angle that inform secondary through quaternary protein structures, or personalized markers such as age profile, can also be integrated with the pipeline to enhance specificity. Finally, to verify that the model is accurate for real drug sequences, it can be further validated by testing on the sequences of known drugs, such as PD-1/PD-L1 checkpoint inhibitors, based on proportionality of their computed affinities through my model to their experimentally determined kD values.

Data Availability Statement

The original contributions presented in the study are included in the article/supplementary files; further inquiries can be directed to the author.

Author Contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

No funding was required.

Conflict of Interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Footnotes

1. Available online at: www.genome.jp/aaindex/

References

Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. 2015. "Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning." Nature Biotechnology 33 (8): 831–38.

Coulie, Pierre G. 2014. "Tumour Antigens Recognized by T Lymphocytes: At the Core of Cancer Immunotherapy." " Nature Reviews Cancer 14 (2): 135–146.

Dong, Haidong, and Svetomir N. Markovic. 2018. The Basics of Cancer Immunotherapy. Springer.

Hu, Xihao, Jian Zhang, Jin Wang, Jingxin Fu, Taiwen Li, Xiaoqi Zheng, Binbin Wang, et al. 2019. "Landscape of B Cell Immunity and Related Immune Evasion in Human Cancers." Nature Genetics 51 (3): 560–67.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2017. "ImageNet Classification with Deep Convolutional Neural Networks." Communications of the ACM 60 (6): 84–90.

Li, Bo, Taiwen Li, Jean-Christophe Pignon, Binbin Wang, Jinzeng Wang, Sachet A. Shukla, Ruoxu Dou, et al. 2016. "Landscape of Tumor-Infiltrating T Cell Repertoire of Human Cancers." Nature Genetics 48 (7): 725–32.

Sompayrac, Lauren M. 2019. How the Immune System Works. 6th ed. Hoboken, NJ: Wiley-Blackwell.

Yuen, Grace J. 2016. "B Lymphocytes and Cancer: A Love–Hate Relationship." " Trends in Cancer 2 (12): 747–757.