VastBiome™

REVERSE ENGINEERING THE MICROBIOME

A Long Short-Term Memory Network to Improve the Selection of Immunotherapy in Cancer

Behnam Barabadi MS*, Ayin Vala MS*, Lanying Ma PhD, Kovi Bessoff MD PhD, Maria Gomez MPH, Kareem Barghouti MBA, Peter McCaffrey MD

*Co-first Authors

BACKGROUND

Abstract

Immune checkpoint inhibition (ICI) therapy has shown considerable promise in treating patients with advanced malignancies . The early clinical success of these drugs, combined with the potential for their application across a wide range of oncologic indications has resulted in an explosive increase in the number of clinical studies investigating the use of anti-PD1/PDL1 therapies both as monotherapy and in combination with other agents . Despite their promise, the utility of ICI drugs is threatened by variable responses. The often dire prognosis of eligible patients coupled with the extreme cost and serious side-effect profile of these therapies warrants the development of improved tools for patient selection. While efforts to identify predictive biomarkers remain an active area of research, the human intestinal microbiome has emerged as a potential source for such predictive signatures.

Efforts to mine the microbiome for generalizable signatures have been impeded by a population and regional-level variability which confounds predictions. Much of this can be explained by the fact that approaches to microbiome analysis are historically based upon 16S marker gene sequencing which suffers from low resolution. Even in cases where whole genome shotgun sequencing is employed, approaches use taxonomic assignment as the primary feature space. This fails to capture functional aspects of gut microbial inhabitants and is fragile in the setting of horizontal gene transfer and evolution that diverges from known reference sequences.

In this work, we present an LSTM network trained using a feature space that encompasses both microbial genomic and proteomic features and we demonstrate the superior performance of such a model by deploying it against previously published immuno-oncology cohort data.

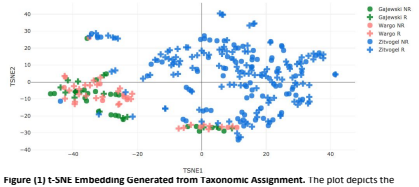

Figure (1) t-SNE Embedding Generated from Taxonomic Assignment. The plot depicts the overlay of three previously published cohorts and illustrates the difficulty of identifying a meaningful discriminator between response phenotypes. Responders (R) to therapy are shown with positive sign and Non-responders (NR) are shown with dots.

METHODS

Data

We aggregated data from three publicly-available, previously published studies examining the gut microbiome and immune checkpoint blockade therapy in patients with advanced epithelial malignancy (melanoma, non-small-cell lung cancer, renal cell carcinoma).
These data comprised 318 total patient samples with whole genome shotgun sequencing performed on patient stool at the start of immunotherapy and, for a subset of patients, intercurrently at the second month of therapy. For each patient, response to checkpoint blockade therapy was assessed radiographically using the RECIST criteria which composites number and size of appreciable tumors.

We applied a customized bioinformatics pipeline to convert patient whole genome shotgun sequences into collections of related synthetic genes and protein motifs encoded by those genes. This pipeline relies upon de novo assembly of biosynthetic gene clusters and extraction of protein motif families resulting in a collection of gene and protein vectors for each sample.

Approach

Because the space of protein motifs is very large and because gene clusters may differ in whole but share functionally relevant subsets , we reduced the dimensionality of the protein motif space by embeddings, and ordering of those contents. Using this approach, a new, we trained a LSTM Network model to predict binary response status (with responders defined as stable disease, partial or complete response) based upon embeddings generated from these transformed vectors. The rationale for embedding synthetic genes and gene clusters prior to training is the likelihood of assembling two identical gene clusters is very low and thus, in addition to a large gene vocabulary, very few samples would contain the same words. Moreover, the evolution of biosynthetic genes preserves groups of especially fit synthesizers even if those groups comprise only a subset of an entire observed gene cluster. Thus, re-labeling samples according to cluster membership based upon gene similarity instroductes some helpful constraints to the breadth and sparsity of this feature space.

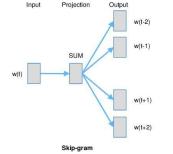

Figure (2) Pfam2Vec.
Our hypothesis is that the pfam domains are positional. Therefore, we trained a Pfam2Vec model to assign similar vectors to pfams co-occurring together more often. Skipgram approach (predict context given target word) was employed since it performs better with rare pfams, and limited amount of observations.
Pfam2Vec was achieved by training window size of 4 and early stopping with minimum percentage of of 0.5%, inputs of context, target and 8 negative samples. The output embeddings were vectors of size 200. Model was trained with CrossEntropyLoss and Adam optimizer with learning rate of 0.001.

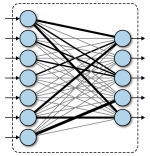

Figure (3A) Linear Model for text classification
The Linear Model's architecture consisted of 3 fully connected layers with 1456, 256, 64 neurons using the Relu activation Function. Dropout layer was implemented between layer 2 and 3 and Sigmoid activation function at the last layer. Binary Cross Entropy Loss function and Adam as Optimizer with learning rate of 0.001 was used.

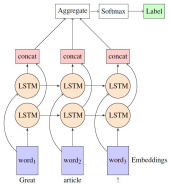

Figure (3B) LSTM Network Model
We designed a bidirectional RNN with 2 layers of GRU with size of 250 each. Embedding layer is the pretrained Pfam2Vec from the previous model of size 200. The training parameters were CrossEntropyLoss and Adam optimizer with learning rate of 0.001

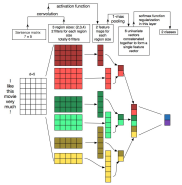

Figure (3C) CNN for Text Classification Model
CNN was employed using the same Pfam2Vec embedding weights. All BGCs converted into vectors of the size 121(equal to the highest number of pfams found in a BGC). CNN model has 4 filters of sizes 1, 2, 3, 5 followed by Maxpool1D after each convolutional layer. Trained with CrossEntropyLoss and Adam optimizer with learning rate of 0.001.

RESULTS

Models were trained on NVIDIA GeForce GTX 1080. LSTM Network was constructed using PyTorch. Each sample-level vector of gene cluster memberships was fed into an embedding layer based upon pfam2vec after which embeddings were fed into the LSTM Network.

The CNN model using the Pfam2Vec embeddings produced the highest AUC of 0.81 based on model running on 100 randomly selected samples.

The top ranked pfams were used to construct corresponding Genetic Blueprints. These blueprints include top responder pfams that are correlated to positive outcome, and exclude the non-responder pfams. The Genetic Blueprints can be used to identify and even produce metabolites synthetically in host organisms such as E coli, a starting point for sourcing new drug candidates for ICI therapy.

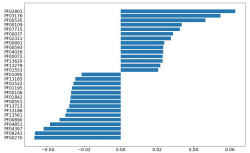

Figure (4) Shows feature importance of protein families ranked in relevance to response. Permutation importance and Integrated Gradient were incorporated in the Captum package in order to obtain feature importance in the CNN model.

CONCLUSIONS

Our prediction models demonstrate the feasibility of making meaningful predictions about immunotherapy response using signatures from the gut microbiome as well as the merit of using functional signatures to do so. Importantly, such functional genomic and proteomic features reach across taxonomic boundaries and allow for generalizable predictions even when no clear discriminative function is possible with taxonomic information. Equally important is the fact that this model performance was possible by processing and representing existing data in a new way and does not require generation of additional datasets beyond what is typically generated for microbiome studies. Finally, the ability for a network to learn a generalizable hypothesis about response in this setting illustrates the additional possibility of using such a model for hypothesis generation as a means to guide further mechanistic research about microbiome host immune control and may potentially unlock both biomarkers and novel therapies.

REFERENCES

1. Routy, B. et al. Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. Science eaan3706 (2017).
2. Gopalakrishnan, V. et al. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. Science 359, 97–103 (2018).
3. Matson, V. et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. Science 359, 104–108 (2018).