Zac Webel

Data Scientist - Applied Mathematician
zow2@georgetown.edu - 561-225-4251
GitHub: zac-webel

CURRENT WORK

Food & Drug Administration (FDA)

Washington D.C

- Building a nucleotide transformer language model predicting mutations in virus DNA coding for spike proteins.
- Sequence to sequence model with an end goal of simulating the evolution of the virus genome coding for the spike proteins.
- Goal for the generative model to guide antibody engineering towards future versions of the virus before the virus has evolved to that stage.

EDUCATION

Georgetown University

Washington D.C

Data Science & Health Informatics - GPA: 4.0 – **Master of Science**

- AI for Health Applications
- Precision Health Informatics (Genomic Data Science)
- Imaging Informatics (Imaging AI)
- Massive Health Data Fundamentals (Machine Learning)
- Evidence Based Data Analysis in Population Health (Statistics)
- Utilizing Data in EMRs
- Digital Health Applications
- Advanced Health Informatics

Florida State University – Honors College

Tallahassee FL – London England

Applied & Computational Mathematics - GPA: 3.894 – **Bachelor of Science**

- **Graduated with honors at 20 years old - 3 Year Completion with 0 credits entering.**
- Honors Multivariable Calculus: A
- Applied Linear Algebra: A
- Programming in Python: A
- Programming in C++: A
- Partial Differential Equations: A
- Ordinary Differential Equations: A

Shanghai Jiao Tong University

Shanghai China

- Machine Learning Course
- One of 11 Students Chosen to be an Oxbridge AI Scholar

SKILLS

Python – Tensorflow – Keras – Numpy – Pandas - SQL – R
Unique Problem Solving Skills – Creative Feature Design
Ability to Comprehend and Apply Scientific Literature
Ability to Learn Complex Topics – Strong Work Ethic – Team Leader
No Ego – Desire to Learn – Desire to Improve – Desire to Innovate

EXPERIENCE

All projects below are done entirely by me and the code is available on my GitHub

SEQUENTIAL LANGUAGE MODEL GENERATING NOVEL VIRUS TAIL FIBER SEQUENCES

- Created a two branched neural network which has shown the ability to generate a novel protein sequence with a third party (ProteInfer) predicted function of virus tail fiber and cell binding when given an entirely random seed sequence and user desired molecular properties.
- User inputs 30 amino acid seed and desired molecular properties for the sequence, the model, at each time steps generates a 21-class probability distribution representing 20 amino acids and a stop flag.
- Created script to pull and clean protein sequences of interest from the NCBI Entrez protein database.
- Tokenized all sequences and performed word2vec to generate custom amino acid embeddings.
- Wrote a sub-sequence data collection script, translating each amino acid into its embedding vector with a window size of 30 amino acids and a step size of 1. The ending position + 1 amino acid was one hot encoded for output. At each time step three molecular properties vectors are concatenated: target properties, window properties and difference properties.
- Created, trained, and evaluated own model architecture containing only 4 million parameters.
- Wrote a generation script.

SELF ATTENTION – LSTM HYBRID GENERATES PROTEIN SEQUENCES WITH DESIRED SECONDARY STRUCTURE PROPERTIES

- Created a neural network which takes input a seed sequence and 3 desired secondary structure percentages and is shown to construct novel polypeptide sequences with an average absolute percentage difference of under 3% for alpha helix, turn and beta sheet desired properties calculated by bio seqUtils protein analysis.
- The model constructs a transformed representation of the input sequence through three branches. A self-attention layer is applied to the sequence of embeddings without positional encodings, stacked LSTM layers analyze the embeddings sequentially and the third branch expands the desire secondary structure properties into nonlinear higher dimensions. These branches are concatenated and used to generate a 21-class probability distribution (20 amino acids and a stop flag).
- Created script to pull and clean protein sequences of interest from the NCBI Entrez protein database.
- Tokenized all sequences and performed word2vec to generate custom amino acid embeddings.
- Wrote a sub-sequence data collection script, translating each amino acid into its embedding vector, then concatenating in sequence with a window size of 30 amino acids and a step size of 1. The ending position + 1 amino acid was one hot encoded for output. At each time step three molecular properties vectors are concatenated: Target properties, window properties and difference properties all of which only contains 3 percentages corresponding to helix, turn and sheet.
- Created, trained, and evaluated own model architecture containing less than 6 million parameters.
- Wrote a generation script.

GENETIC BIOMARKER DISCOVERY – STAGE 2 INVASIVE BLADDER CANCER

- Ran statistical analysis on gene expression data from stage 2 invasive bladder cancer vs clinically verified normal samples in R.
- Isolated 4 significantly differentially expressed genes: CDC20, TOP2A, SRPX, IQGAP3
- Performed Principal Component Analysis on the gene expression data, reducing to 2 dimensions then running the K-means algorithm which separated invasive cancer samples from normal samples. Signals a possible genetic transcript for predicting the occurrence of stage 2 invasive bladder cancer.
- Ran a systems biology analysis in R, Finding the G Beta: Gamma Signaling Through BTK pathway R-HSA-8964315 relevant to stage 2 invasive bladder cancer.