Summer Review 5

FBGAN

Anvita Gupta, James Zou

Stanford University

Reviewed by : Arshdeep Sekhon

[1]Department of Computer Science, University of Virginia

https://qdata.github.io/deep2Read/

The Task: GANs for synthetic biology

generate genes (protein sequences) that can encode proteins with specific properties

- generate antimicrobial peptides: lower molecular weight peptides with less than 50 amino acids
- optimize secondary structure for peptides– alpha helical peptides

De nitions

secondary structure: the three dimensional form of local segments of proteins.

important for protein functions

two types: alpha[50 amino acids] and beta

The Dataset generation

Uniprot database: Select 3655 proteins with length $\leq$ 50 residues[1]

Uniprot database: protein sequence records with functional information

cluster by sequence similarity

Select one from each cluster as a representative protein

convert into cDNA sequences

get a codon for each amino acid, start codon, stop codon[2]

[1] limit the length to 50 to avoid long-term dependencies + observation of secondary structure etc.

[2] The start codon marks the site at which translation into protein sequence begins, and the stop codon marks the site at which translation ends.

GAN

Loss function:

$$\min_G \max_D V(D;G) = E_{x \sim P_{data}(x)}[\log(D(x))] + E_{z \sim P(z)}[\log(1 \quad D(G(z)))] \tag{1}$$

WGAN more stable during training.

5 residual layers.

2 1 D convolutions of 5 1

Use Gumbel Softmax instead of Softmax

GAN

Figure: the GAN model

Train GAN to produce valid sequences for a few epochs

The Analyzer

Figure: Function Analyzer

Analyzer to select sequences that are desirable properties

Pretrain Analyzer

The model

Figure: The model

Use feedback mechanism to select sequences that are desirable proper

Analyzer training

2600 experimentally veri ed antimicrobial properties from APD3 Database

negative set from Uniprot

Translate to cDNA and train Analyzer

can be potentially non di erentiable

Analyzer for Antimicrobial Peptide Coding genes

Classi er

Input: Gene Sequences Output: 1/0 codes for AMP or not

Positive Set of 2600 AMPs from APD3 Database

Negative set of 2600 random peptides from Uniprot translated to cDNA

Analyzer: Architecture

RNN architecture: 2 GRU layers of 128 size

Last time step to dense layer

sigmoid activation function: whether gene belongs to positive class not

Analyzer: Check secondary structure

Wrapper around PSIPRED secondary structure predictor

PSIPRED: predict secondary structure of each aminoacid

Wrapper: gene sequence to protein sequence to PSIPRED

predicts structure of amino acids inside the protein: total number of alpha helix tagged residues: choose above a certain cuto

If gene to protein not possible: output 0

Results: Generate protein coding genes

Train GAN to produce 156 nucleotides

Correct gene if start codon, some codons, stop codins

Before training, 31.25% sequences follow the correct gene structure

After training, 77.08% sequences follow correct structure

Results: Generate protein coding genes

Figure: A set of 500 valid genes were sampled from the trained WGAN, and 10
physiochemical features were calculated for the proteins encoded by the synthe
genes. The same 10 features were also calculated for the cDNA sequences fro
Uniprot proteins. PCA was performed on the features of the natural cDNA

Feedback analyzer: Results

training accuracy = 0.9447 and validation accuracy = 0.8613

test accuracy = 0.842

Results: Feedback-Loop to Optimize Antimicrobial Properties

After GAN and AMP analyzer are trained, link with feedback loop

analyzer selects sequences with $P(AMP) > 0.8$ and feed into discriminator as real sequences

Replace n oldest with selected n newest

Criteria to evaluate selected genes

does the Analyzer predict more sequences antimicrobial over time?

are the generated genes similar to real AMP wrt properties and sequences of proteins?

After 60 epochs, nearly all predicted as anitmicrobial

93:3% of the generated sequences after closed loop training have correct gene structure

Figure:

Results: Similarity of generated vs exprimental sequenc

Figure: Edit distance results

a larger proportion of sequences with a lower edit distance from the AMP sequences

the sequences after feedback have a higher edit distance within themselves than the antimicrobial sequences do with each other

Results: physiochemical properties

the proteins encoded by the closed-loop sequences shift to be close
to the positive antimicrobial peptides in ve out of ten
physiochemical properties such as Length, Hydrophobicity, and
Aromaticity, and remains as similar as the sequences without feedba
for properties such as Charge and Aliphatic index.
This is true even though the analyzer operated directly on the gene
sequence rather than these physiochemical properties
the feedback mechanism did not directly optimize the physiochemic
properties that show a shift.

Results: Secondary Structure with Black-Box PSIPRED Analyzer

Use Secondary Structure Anayzer

secondary structure more attractive to optimize for since it arises in short peptides of length less than 50

gene sequences with more than 5 alpha-helical residues were input back into the discriminator as real data.

After 43 epochs of feedback, the helix length in the generated sequences was signi cantly higher than the helix length without feedback and the helix length of the original Uniprot proteins,

Results: Secondary Structure with Black-Box PSIPRED Analyzer

helix length was greater with feedback than without feedback

Figure: Before

Results: Secondary Structure with Black-Box PSIPRED Analyzer

helix length was greater with feedback than without feedback

Figure: After FBGAN