Genome Dreaming

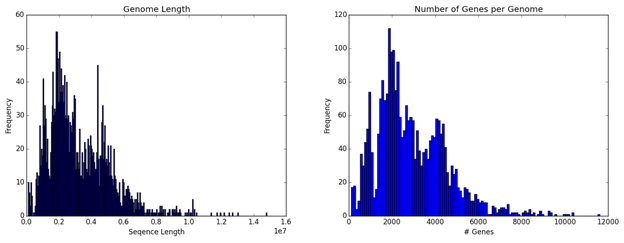
Akshay Maheshwari, Bohan Wu, Oğuz H. Elibol, Drew Endy

Introduction and Motivation

- The ability to generate novel genomes and sequences with predictable characteristics would revolutionize synthetic biology
- We build an end-to-end pipeline to:
  1. Learn the properties of unlabeled prokaryotic genomes in an unbiased way
  2. Generate new sequences with predictable properties
  3. Visualize and evaluate learned representations of generated sequences

Data

- 4131 unlabeled prokaryotic genomes and their gene annotations from the KEGG database



Acknowledgements

We would like to thank Dr. Drew Endy, Dr. Anshul Kundaje, Dr. James Zou, Namrata Anand, Bo Wang, Jesse Zhang, Volodymyr Kuleshov, and all members of the Endy lab

Model & Visualization

Pipeline:

- Learn structure using Recurrent Neural Networks (LSTM)
- Optimize model
- Generate sequence

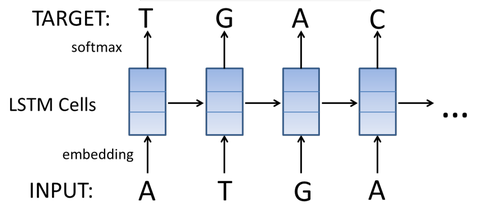
- Extract hidden layers for all *E. coli* genes
- Reduce hidden layers into 2D using t-SNE
- Cluster using Mixture of Gaussians

- Extract proportion of points in each cluster
- Create vector embedding for each gene (100-D)

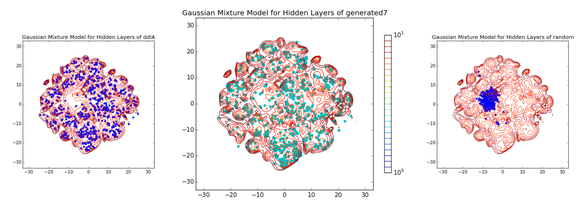
- Reduce gene embeddings to 2D
- Evaluate similarities between genes and generated sequence

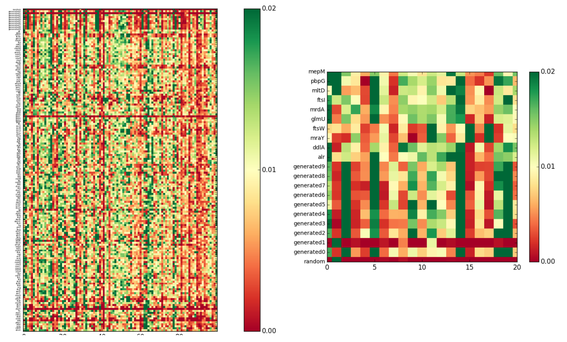
Learning and Sequence Generation

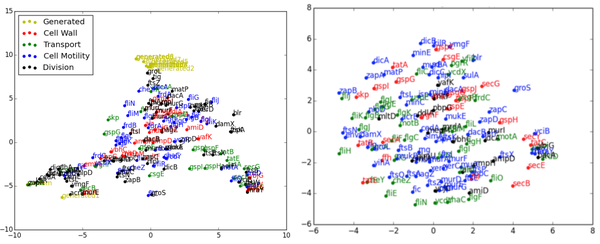
$$L_{CE}(y, \hat{y}) = -\sum_i y_i \log \hat{y}_i$$



Sequence Visualization and Evaluation

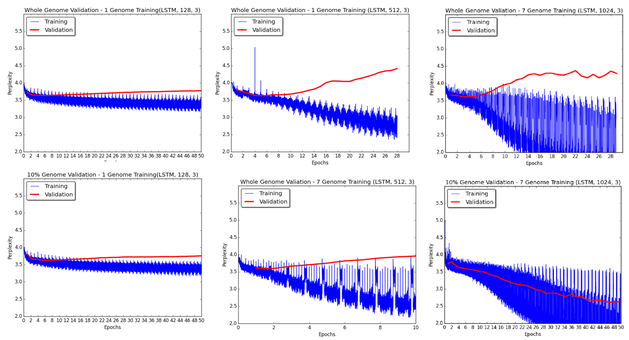






Results

Model tuning and optimization

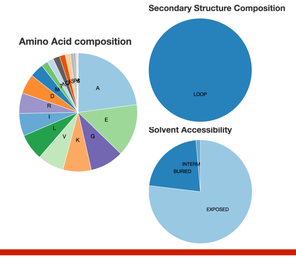


Model Validation

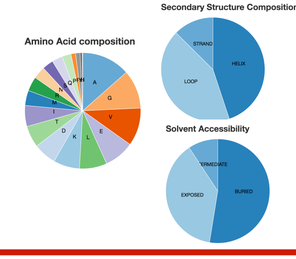
	Escherichia Model	Streptococcus Model
ECL (E)	3.59998	4.17365
ESO (E)	3.62222	4.1728
EBL (E)	3.56904	4.17352
EAL (E)	3.63504	4.15488
SAH (S)	3.67812	3.42402
SXY (S)	3.68177	3.47255
SEQO (S)	3.68418	3.47396
SAUR (S)	3.67919	3.43899

Sequence Generation and Property Prediction

Generated Protein



groL (Closest Neighbor)



Next Steps

- Build a stronger model: explore alternatives such as WaveNet, Neural Turing Machines, and Attention Models
- Identify the properties and motifs that the neural network is learning