CS 287r Project Abstract

Sam Goldman
sgoldman@college.harvard.edu

David Yang
yangd@college.harvard.edu

Project Advisor: Eli Weisntein (Marks Lab)

February 21, 2019

1 Area

Title: A Seq2Seq Framework for Codon Optimization

In genetic engineering, a common goal is to take a protein-coding gene from one organism and insert it into another organism for "heterologous" expression. While genes are made up of base pairs, *A*, *T*, *G*, or *C*, biology naturally segments genetic sequences into triplets of base pairs, referred to as codons (e.g. *GCG*), which map directly to proteins. The redundancy in the genetic code allows for multiple codons to encode for the same protein segment, which means that the same protein can be expressed from a variety of DNA sequences (e.g. the protein produced from DNA sequence *GCG* is equivalent to *GCA*). Though the resulting protein may be the same, the specific DNA sequence directly affects the amount of functioning protein produced. In other words, organisms use codons that are best fit for their cellular environment.

The codons that underlie the protein can be optimized when re-purposing a protein from one species to another in order to maximize protein production, a task known as codon optimization. Traditional approaches often use a naive replacement strategy, which chooses the most frequently used codon for each position of the amino acid. In practice, this crude method is often ineffective.

We propose a Seq2Seq translation framework for the task of codon optimization. Here, each species is considered a language (e.g. Human, E. coli), and DNA sequence for a protein as a sentence. Sequences between two organisms that code for the same protein, which are known as homologues, can be considered as parallel sentences. Our task is to "translate" the DNA sequence from a species to another, so that the "translated" sequence would be optimized for the target species for expression.

2 Papers

1. **MAIN:** Zhao et al. (2019) use a transformer architecture with copy attention mechanisms to achieve state of the art on grammar correction. We seek to reimplement this approach in a biological domain, as codon optimization likely involves similar constraints around copying various codons from source to target.

2. Schmaltz et al. (2016) demonstrate that a seq2seq model can be used for grammar correction. It uses a combination of character and word level models, which can correspond to DNA and Amino acid level models in our case.

3. Zoph et al. (2016) show how transfer learning in language with a parent and child language pairs can be an effective method of transfer. We hope to use similar methods here. We may need to leverage similar transfer approaches due to the size of our datasets (e.g. *E.coli* only has 4000 genes).

4. Tian et al. (2018) provide the most recent computational approach to codon optimization we could find, using prepackaged random forest methods. While the paper is somewhat inscrutable, they offer a helpful strategy for generating a dataset.

3 Time

We are scheduled to present in class on April 17. We propose a rough timeline of deadlines for certain tasks:

March 5: Collect dataset of homologous protein sequences[1]
March 12: Train baseline, NMT translation models on our corpus of protein families
March 19: Implement more sophisticated copy attention models
March 26: Repurpose models for different downstream, baseline tasks and interpret

4 Baseline

To our knowledge, codon optimization has not previously been considered as a seq2seq, translation problem. Recent efforts have focused on experimental approaches that score metrics such as expression level for each constructed sequence. For example, a recent work experimentally tested 244,000 biological sequences that created similar proteins and measured the total protein production for each sequence (Cambray et al. (2018)), but our model will be focused on generation of translations and not regression tasks.

After constructing a list of homolog sequences between species, we can test our output sequences on the held out data (with a metric similar to BLEU).

In order to show the power of our model, we plan to transfer our model to perform one of two downstream tasks with a well defined baseline: horizontal gene transfer detection or mutation prediction.

Horizontal gene transfer: Bacteria are confusingly able to freely swap groups of genes with other bacteria in physical proximity. Detecting when these transfer events occur has important applications in clinical medicine and microbiology. Given our model $p(y|x, \theta_z)$, that translates sequence x to a sequence y optimized for organism z, we can score the likelihood that a given sequence belongs to organism z. This can be used on the baseline task of horizontal gene transfer (Becq et al. (2010)).

[1] Either precurated or generated with Jackhmmer search tool

Mutation prediction: There is a similarly strong mutation prediction literature, and we can use our model on this downstream task as well for additional baseline of comparison [2].

References

Becq, J., Churlaud, C., and Deschavanne, P. (2010). A benchmark of parametric methods for horizontal transfers detection. *PLoS One*, 5(4):e9989.

Cambray, G., Guimaraes, J. C., and Arkin, A. P. (2018). Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in escherichia coli. *Nature biotechnology*, 36(10):1005.

Schmaltz, A., Kim, Y., Rush, A. M., and Shieber, S. M. (2016). Sentence-level grammatical error identification as sequence-to-sequence correction.

Tian, J., Li, Q., Chu, X., and Wu, N. (2018). Presyncodon, a web server for gene design with the evolutionary information of the expression hosts. *International journal of molecular sciences*, 19(12):3872.

Zhao, W., Wang, L., Shen, K., Jia, R., and Liu, J. (2019). Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

[2] https://marks.hms.harvard.edu/evmutation/downloads.html