Identifying Transcription Unit Structure from Rend Sequencing Data

Category: Life Sciences

Travis Horst (thorst)

In bacteria, many genes are expressed together on the same RNA transcript called a transcription unit. Identifying these transcription units can play a role in better understanding biology as related genes are typically expressed together and coexpression can have important physiological implications. Although many transcription units have been identified, there is not an efficient way to determine all the transcription units. Sequencing techniques can provide high throughput data of the transcripts in populations of bacteria. Recent work (Fig. 1),[1] has shown promise in identifying transcription units however the analysis (peak z-score) was not generally applicable to entire genomes. Rend seq data can provide benefits over other sequencing data as it does not suffer from 3' end bias and shows enrichment at the start and end of transcription units, which should capable of capturing with a model.

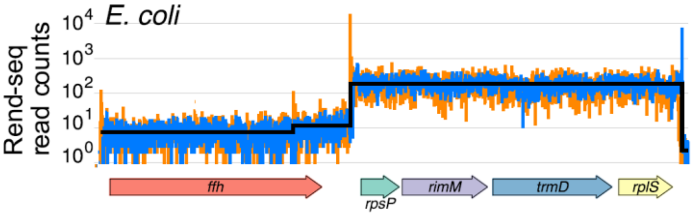

Fig 1: Example data - orange and blue are reads. Black indicates transcription units with genes below.

Using the sequencing reads dataset from the above paper, my plan would be to explore different machine learning techniques to identify the transcription units. This would largely consist of unsupervised learning methods so that models do not need to be trained on annotated datasets that might be incorrect or incomplete. The goal would be to assign each position in the genome to a set of transcription units (clusters) through the counts at each position. A good candidate would be a mixture model with a Poisson distribution to capture the counts. Another option would be creating a Hidden Markov Model to identify transitions between different read intensity levels. Although typically associated with time series, the position in the genome could behave in a similar manner. Another possible method would be using DBSCAN clustering which could identify outliers at the start or end of transcription units. Clusters could be formed using sequence position, a moving average of nearby locations and the read intensities at a given position.

Some data processing will be required. The sequencing data exists as unnormalized counts for two different strands. Normalizing the data and combining the reads from the different strands could provide different performance on the models. Identifying gene locations will also require some processing of the data to map the output of the model to a relevant biological interpretation. Identification of transcription units can be compared to curated transcription units from Ecocyc for a subset of the genome as well as the output from the paper (both regions with good and bad performance). This will provide a metric for how the algorithms are performing and their accuracy against previously identified transcription units. For Poisson mixture, the number of populations could be varied depending on the number of genes in a region of interest. For the HMM, transition probabilities and the number of states will be varied. The parameters for DBSCAN (maximum distance between points, number of points in a group) can be varied to assess performance.

[1] Lalanne JB, *et al.* (2018). Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. *Cell 173*(3) 749-761.