Thesis Porject

Eftychia Thomaidou
4182219

June 19, 2015

1 Motivation

The DNA is transcribed to RNA and that translated into a protein. Protein is the actual product of the DNA. How often the RNA is translated or how much protein is produced, is highly related with the Ribosome. The ribosome is the actual 'molecular machine' that commits the action of the translation.
The part of the RNA, that becomes a protein, is called coding sequence. The ribosome binds to the RNA at the RBS (Ribosomal Binding Site). The RBS is located at the 5UTR, which is the untranslated region in front of the coding sequence. That is according to the traditional way of the translation and doesn't apply in all cases.
Other elements that are located in the 5' Untranslated Region affect the binding of the ribosome and thus the translation. It is known that this is the case, but those elements are unknown or not confirmed, so it is of hight interest to be discovered.
Aim of this project is to build a regression model, that has as input different features, that can be found in the 5' UTR sequences, and predicts the translation initiation rates. The reason why it is important to do that is dual. To begin with, it is very useful to be able to predict the initiation rates for new sequences and consequently be able to synthesize new sequences with high initiation rate. Additionally, it is of high importance to understand which of the elements, located in the 5' UTR, influence the translation initiation rates and so the translation.
The data of this project are: an aggregation of names of yeast genes and their translation initiation rates. In this model are used the log values of the latter. The algorithm is applied on the data provided by the following papers: [3] and [2].

2 Methodology

In order to retrieve more specific information, like the starting and ending position of the 5 UTR of each gene, are used the data provided by [4] and [5]. A parsing algorithm is used, that reads the two GFF3 files and combines them in an array. From this array are selected only those genes that their initiation rate is known and provided in the first data. With the use of 'start' and 'end' columns of the array is calculated the length of the sequences. The actual sequences are also known and provided by the SGD database [1]. As it it known there are 16 chromosomes saved in fasta files and one extra named 'chrmt.fsa'. For personal ease is used the numeric way of naming the chromosome files and not the roman. Similarly the

last file is named as the 17. Now all the important data are provided and are ready to be used.
In need for features for the regression model is used the length of the sequences. Are used only the sequence that have length ¡ 11bp.
Features: sequenceLength, Afrequency = baseCounter['A']/float(sequenceLength), Tfrequency = baseCounter['T']/float(sequenceLength), Gfrequency = baseCounter['G']/float(sequenceLength), Cfrequency = baseCounter['C']/float(sequenceLength), Target: Initiation Rates

References

[1] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic acids research*, page gkr1029, 2011.

[2] Luca Ciandrini, Ian Stansfield, and M Carmen Romano. Ribosome traffic on mrnas maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation. *PLoS computational biology*, 9(1):e1002866, 2013.

[3] Alexey A Gritsenko, Marc Hulsman, Marcel J T Reinders, and De Ridder. An unbiased quantitative model of S. cerevisiae protein translation derived from ribosome profiling data. pages 1–7, 2014.

[4] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349, 2008.

[5] Moran Yassour, Tommy Kaplan, Hunter B Fraser, Joshua Z Levin, Jenna Pfiffner, Xian Adiconis, Gary Schroth, Shujun Luo, Irina Khrebtukova, Andreas Gnirke, et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mrna sequencing. *Proceedings of the National Academy of Sciences*, 106(9):32643269, 2009.