Using Information Theory to Find Codon and Codon Pair Bias Networks

Clement, N.[1], Katagiri, N.[1], Bar, H.[2], DiCuccio, M.[3], Komar, A.A.[4] & Kimchi-Sarfaty, C.[1]

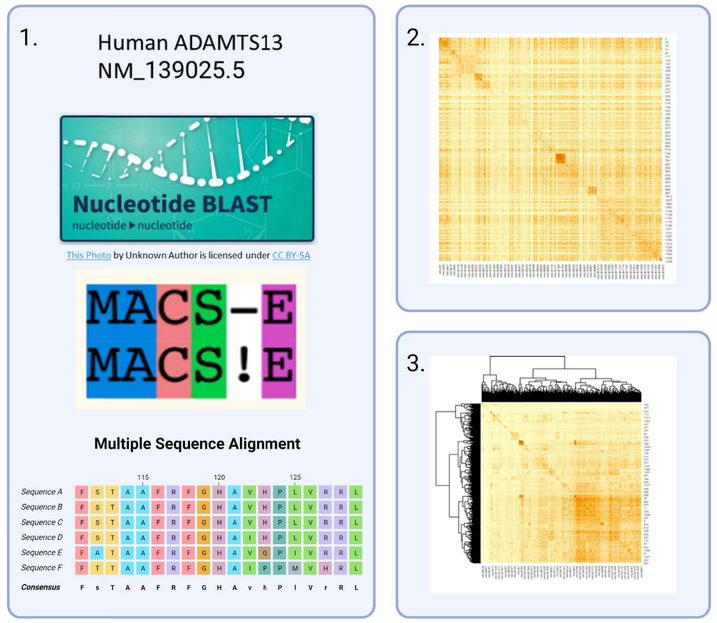
[1] OTAT/DPPT/HB in the CBER, US FDA, Silver Spring, MD • [2] Department of Statistics, University of Connecticut, Storrs, CT • [3] NCBI, Bethesda, MD • [4] Center for Gene Regulation in Health and Disease, Cleveland State University, Cleveland, OH

Introduction

Bias is observed in codon usage (CU) and codon pair usage (CPU) in genetic coding sequences, but the statistical and evolutionary relationship between them remains unclear. It is possible that different regions within a protein exhibit unique CU - CPU relationships which can be defined statistically. ADAMTS13, an anti-clotting factor, was used as the model for developing the CU/CPU relationship analytical method. This study aims to explore the relationship between CU and CPU by identifying different regions in ADAMTS13 with unique codon and/or codon pair information networks.

Symmetric uncertainty (SU) was quantified for each pairwise combination of codons in a multiple sequence alignment of ADAMTS13 sequences. Symmetric uncertainty quantifies the statistical dependence between two variables with a score of one indicating perfect dependence and zero indicating independence. Unsupervised machine learning classification is used to identify clusters of codons with similar information content. The ADAMTS13 sequences used in this project were obtained from the NCBI standard nucleotide database and are comprised of all available high quality, complete mRNA homolog and ortholog mammalian sequences. Positions clustered by SU were observed to exhibit different codon usage from one another which will be used to enhance codon and codon pair optimization schemes.

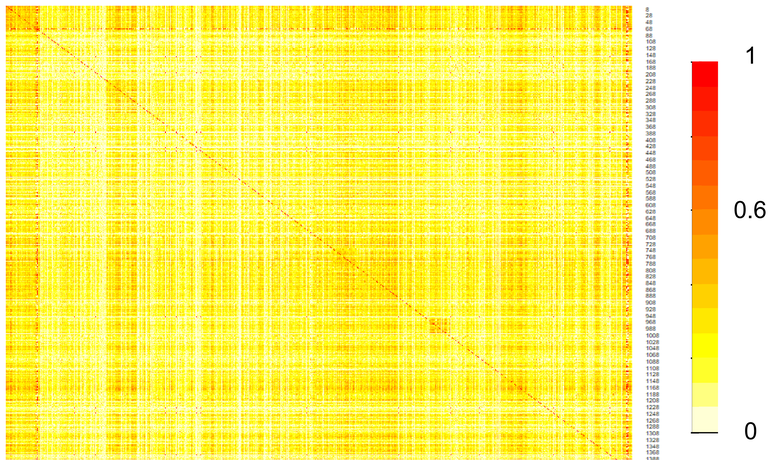
Methods



1. Megablast retrieved sequences that are similar to human ADAMTS13 (NM_139025.5). MACSE was used to conserve codons during alignment.
2. Symmetric Uncertainty (SU) is calculated for each pairwise combination of codon positions, and a symmetric SU heatmap is generated. SU ranges from 0 (statistically independent) to 1 (perfectly statistically dependent).
3. CLARANS clustering was used to search the SU heatmap for groups of positions whose codon bias are related. Next, biological factors on each codon were incorporated into the computation of statistical relationships of the clusters.

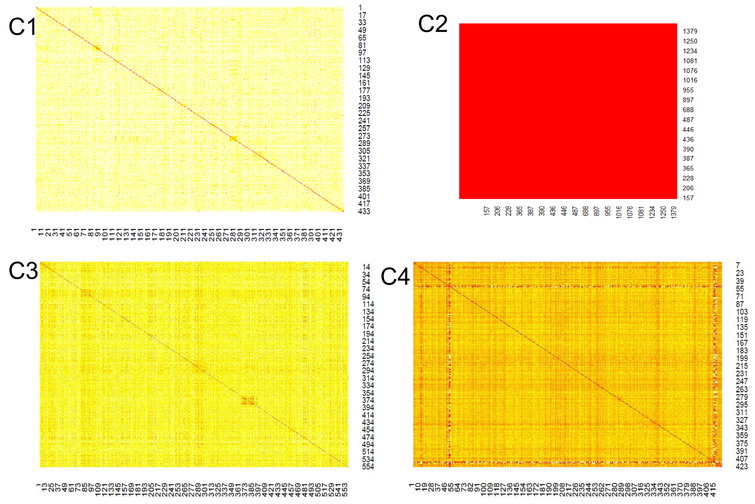
Results

Fig. 1 Short and long-range Symmetric Uncertainty interactions in ADAMTS13



- The symmetric uncertainty (SU) heatmap of pairwise positional codon usage in *ADAMTS13*. The heatmap is derived from the ADAMTS13 sequence for 157 species and aligned based on the human sequence (NM_139025.5). Both axes in the SU heatmap are a codon position in *ADAMTS13*. The color scale of the heatmap runs from white (low SU) to red (high SU).

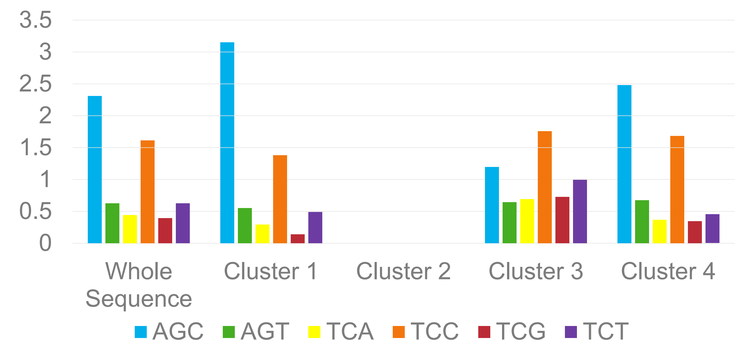
Fig. 2 High, medium, and low information clusters have been identified in ADAMTS13



- Low (C1), medium (C3), high (C4), and perfectly dependent (C2) clustered heatmaps were identified during analysis. No details can be observed in C2 because all 18 positions were perfectly dependent upon one another and the heatmap is completely red.

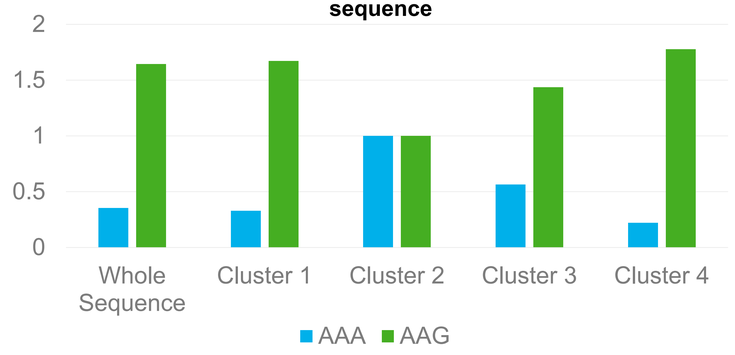
Results

Fig. 3 Serine's Relative Synonymous Codon Usage in the clusters and whole sequence



- As a synonymous codon's Relative Synonymous Codon Usage (RSCU) increases it is used more preferentially than those with lower RSCU. While a synonymous codon whose RSCU is below 1 is used less preferentially than other synonymous codons. A RSCU of 1 indicates that a synonymous codon is used with equal preference to others. The clusters exhibit different codon usage bias from each other and the entire ADAMTS13 sequence. An example of the differing usage can be seen in differences in Serine's RSCU plot.

Fig. 4 Lysine's Relative Synonymous Codon Usage in clusters and whole sequence



- The only codons observed in Cluster 2 coded for Lysine and Asparagine. Both amino acids are coded by two codons, and these codons were equally represented in Cluster 2.

Conclusions

- Clusters of codons with similar information relationships were identified in ADAMTS13. These groups display unique codon usage which may be used to predict codon and codon pair usage more accurately than previous methods.
- The next steps of this project will be to evaluate codon pair predictability in the clusters and develop a positional observed to expected ratio quantification method.
- Enhanced understanding of the relationship between codon and codon pair usage will be used to enhance codon and codon pair optimization schemes for applications in therapeutics.