Conservation of Codon Frequencies Across Domains of Life

John Poncini, Roger Volden, Jairo Navarro, David Bernick
Department of Biomolecular Engineering at University of California, Santa Cruz

Abstract

Here, we describe FOCUS, a set of visualization tools that display protein sequences and structures with amino acids colored depending on the respective codon frequency. We approximate the rate of translation by constructing codon usage tables from the open reading frames of genomes deposited in the NCBI GenBank, then applying the global and relative codon frequencies to a gene of interest. We hypothesize that the constructed codon usage tables, when normalized by relative codon bias, can serve as a proxy for the abundance of each amino-acyl tRNA in the cell. Using these tools, we show that the selection of high and low frequency codons appear to be structurally conserved in orthologs across all domains of life. This evidence suggests that codon frequency plays a vital role in protein folding, and thus has broad implications in drug discovery, amyloid diseases, protein structure prediction, and synthetic biology.

Introduction

The conjoining factor between the codons and their respective amino acids is called transfer RNA (tRNA). tRNA molecules appear in different concentrations within cells and overall frequencies of tRNAs can differ between organisms. Each organism has a codon bias, which is just another way of referring to the frequency of each of its codons. We used Python to calculate ORFs (open reading frames), split them into codons, count how many times each codon occurs, and divide each value of our codon dictionary by the total number of codons to get a global codon frequency. We believe that the frequency of each codon relates to its concentration in the cytoplasm, and in turn affects the rate of translation. This stems from the idea that it will take longer for a low frequency tRNA to come by the ribosome during translation than a high frequency codon because of their relative concentrations in the cell. We aim to show that the positioning of low frequency codons is strategic, and that there is evolutionary pressure for rare codons to appear in certain positions in proteins. It has been hinted at that codon frequency adds another layer of depth to the protein folding code. If this is true, then we expect to see that low frequency codons appear in conserved positions in proteins. In this study, we looked at arginyl-tRNA synthetase because they are well conserved across domains of life. We expect to see that low frequency codons are structurally conserved across domains of life.

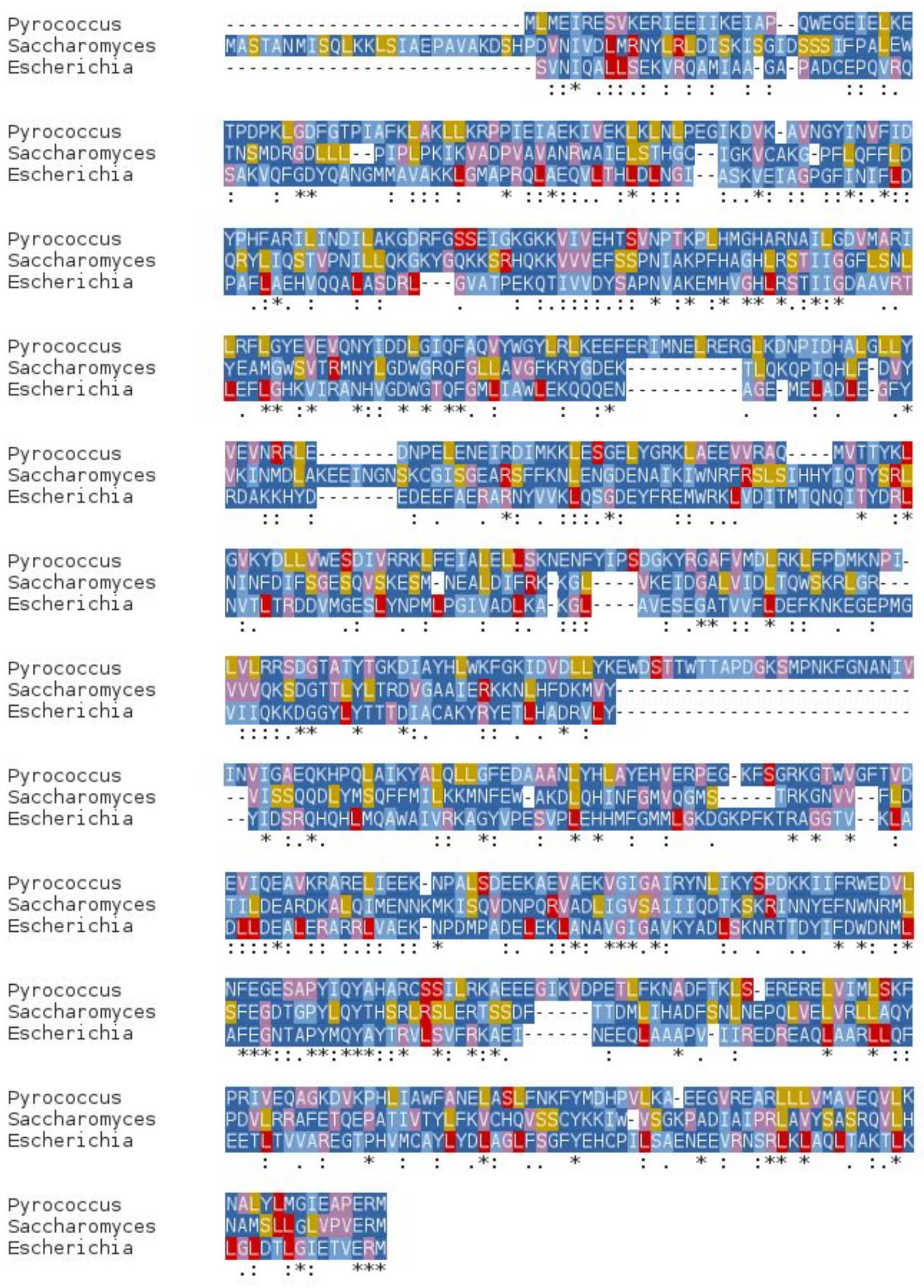


Figure 1. Multiple sequence alignment showing the frequency of the codons between *Pyrococcus horikoshii, Saccharomyces cerevisiae*, and *Escherichia coli*. Blue amino acids represent those that have a global frequency of >40%. Light blue is between 40 and 30%. Pink is between 30 and 20%.Orange is between 20 and 10%. Red represents those with a frequency <10%.

Results

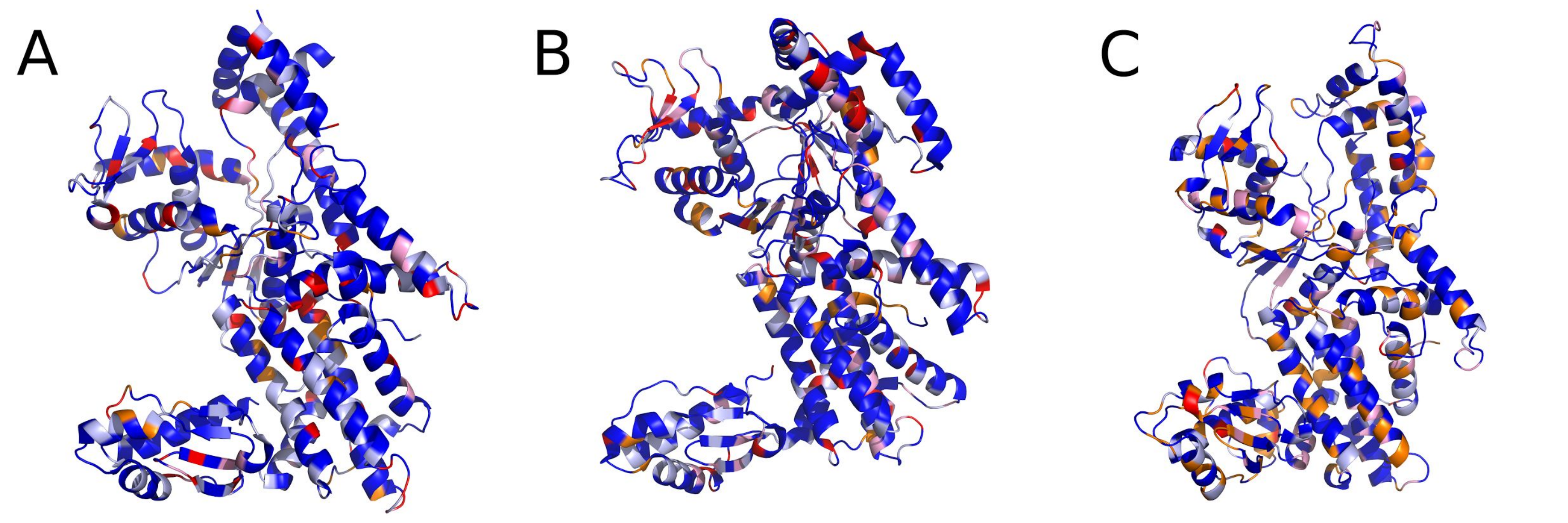
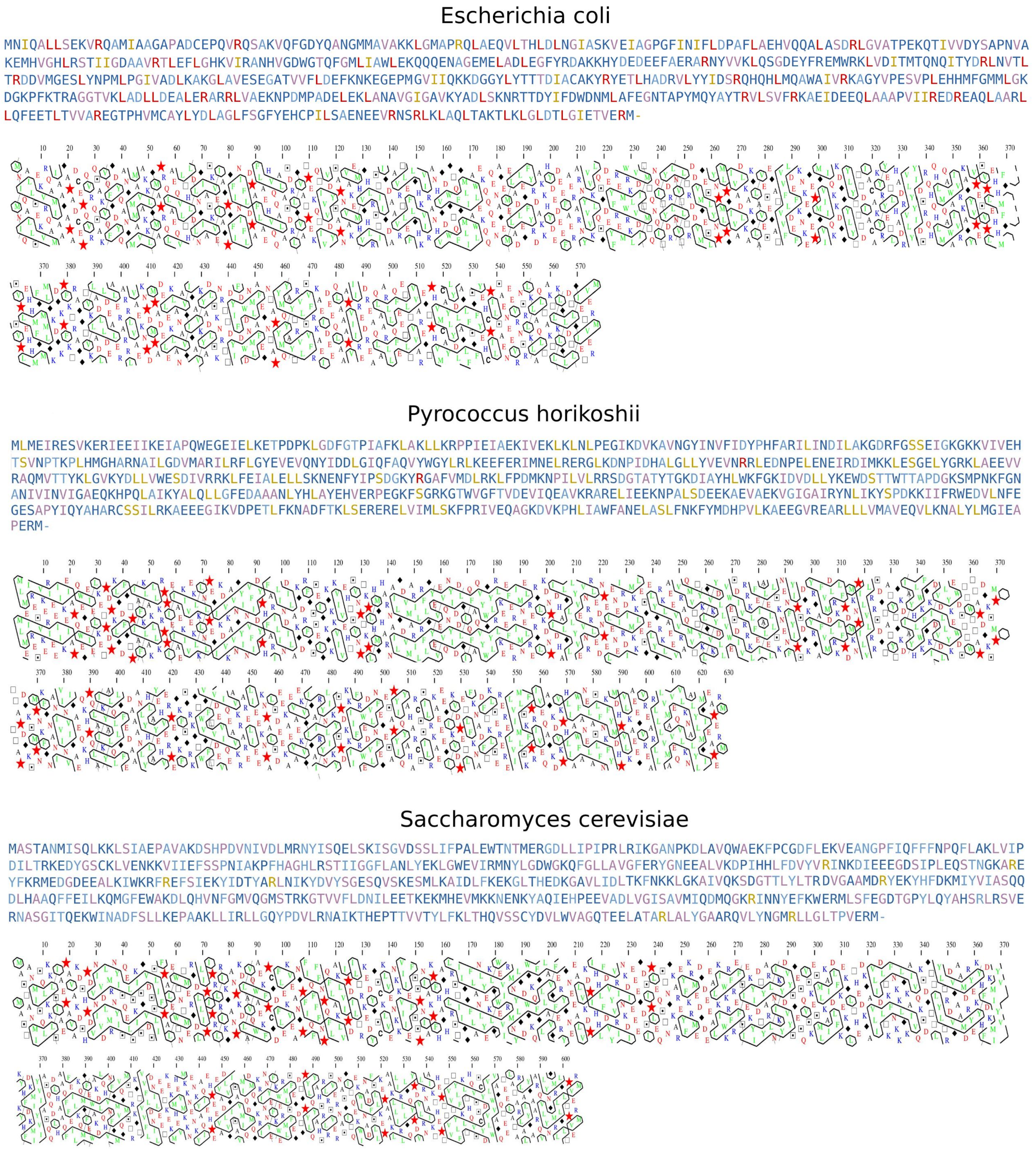


Figure 2. Output of the colorFreq.py program when applied to each arginyl-tRNA synthetase. **(A)** *Escherichia coli*, representing bacteria (PDB ID 4OBY). **(B)** *Pyrococcus horikoshii*, representing archaea (PDB ID 2ZUE). **(C)** *Saccharomyces cerevisiae* (brewer's yeast), representing eukarya (PDB ID 1BS2). These structures are colored according to frequency of codons. The same color for each codon appears in the same areas across these three organisms. Many rare codons appear around loop regions and around β-strands.



Escherichia coli

MNIQALLSEKVRQAMIAAGAPADCEPQVRQSAKVQFGDYQANGMMAVAKKLGMAPRQLAEQVLTHLDLNGIASKVEIAGPGFINIFLDPAFLAEHVQQALASDRLGVATPEKQTIVVDYSAPNVA
KEMHVGHLRSTIIGDAAVRTLEFLGHKVIRANHVGDWGTQFGMLIAWLEKQQQENAGEMELADLEGFYRDAKKHYDEDEEFAERARNYVVKLQSGDEYFREMWRKLVDITMTQNQITYDRLNVTL
TRDDVMGESLYNPMLPGIVADLKAKGLAVESEGATVVFLDEFKNKEGEPMGVIIQKKDGGYLYTTTDIACAKYRYETLHADRVLYYIDSRQHQHLMQAWAIVRKAGYVPESVPLEHHMFGMMLGK
DGKPFKTRAGGTVKLADLLDEALERARRLVAEKNPDMPADELEKLANAVGIGAVKYADLSKNRTTDYIFDWDNMLAFEGNTAPYMQYAYTRVLSVFRKAEIDEEQLAAAPVIIREDREAQLAARL
LQFEETLTVVAREGTPHVMCAYLYDLAGLFSGFYEHCPILSAENEEVRNSRLKLAQLTAKTLKLGLDTLGIETVERM-

Pyrococcus horikoshii

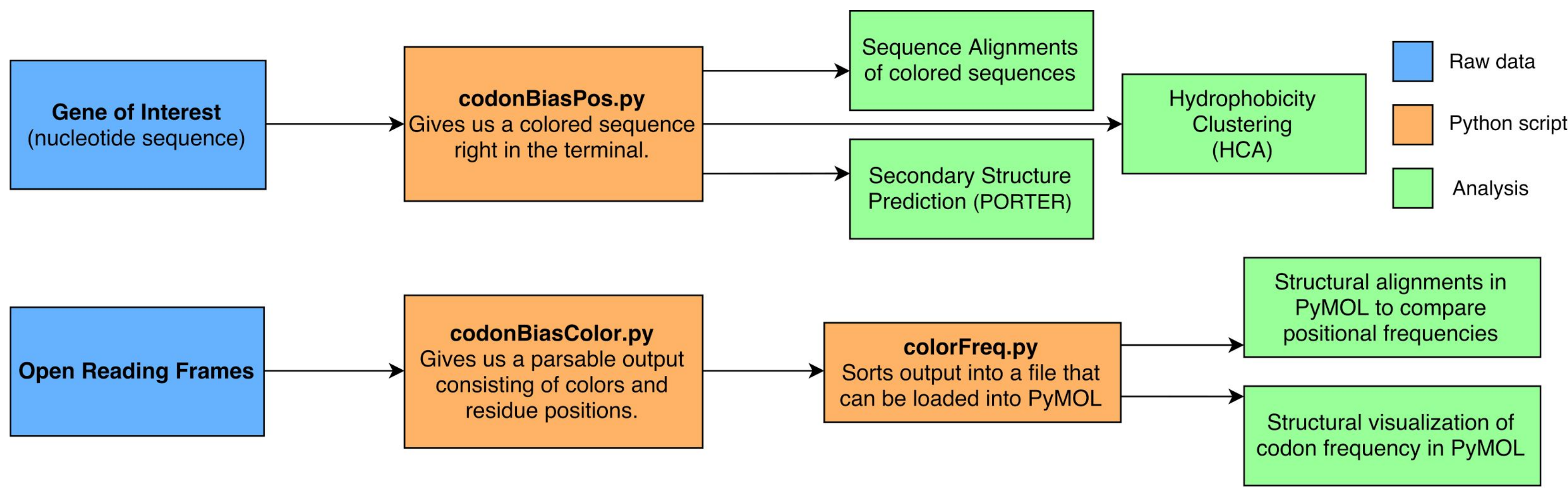
MLMEIRESVKERIEEIIKEIAPQWEGEIELKETPDPKLGDFGTPIAFKLAKLLKRPPIEIAEKIVEKLKLNLPEGIKDVKAVNGYINVFIDYPHFARILINDILAKGDRFGSSEIGKGKKVIVEH
TSVNPTKPLHMGHARNAILGDVMARILRFLGYEVEVQNYIDDLGIQFAQVYWGYLRLKEEFERIMNELRERGLKDNPIDHALGLLYVEVNRRLEDNPELENEIRDIMKKLESGELYGRKLAEEVV
RAQMVTTYKLGVKYDLLVWESDIVRRKLFEIALELLSKNENFYIPSDGKYRGAFVMDLRKLFPDMKNPILVLRRSDGTATYTGKDIAYHLWKFGKIDVDLLYKEWDSTTWTTAPDGKSMPNKFGN
ANIVINVIGAEQKHPQLAIKYALQLLGFEDAAANLYHLAYEHVERPEGKFSGRKGTWVGFTVDEVIQEAVKRARELIEEKNPALSDEEKAEVAEKVGIGAIRYNLIKYSPDKKIIFRWEDVLNFE
GESAPYIQYAHARCSSILRKAEEEGIKVDPETLFKNADFTKLSERERELVIMLSKFPRIVEQAGKDVKPHLIAWFANELASLFNKFYMDHPVLKAEEGVREARLLLVMAVEQVLKNALYLMGIEA
PERM-

Saccharomyces cerevisiae

MASTANMISQLKKLSIAEPAVAKDSHPDVNIVDLMRNYISQELSKISGVDSSLIFPALEWTNTMERGDLLIPIPRLRIKGANPKDLAVQWAEKFPCGDFLEKVEANGPFIQFFFNPQFLAKLVIP
DILTRKEDYGSCKLVENKKVIIEFSSPNIAKPFHAGHLRSTIIGGFLANLYEKLGWEVIRMNYLGDWGKQFGLLAVGFERYGNEEALVKDPIHHLFDVYVRINKDIEEEGDSIPLEQSTNGKARE
YFKRMEDGDEEALKIWKRFREFSIEKYIDTYARLNIKYDVYSGESQVSKESMLKAIDLFKEKGLTHEDKGAVLIDLTKFNKKLGKAIVQKSDGTTLYLTRDVGAAMDRYEKYHFDKMIYVIASQQ
DLHAAQFFEILKQMGFEWAKDLQHVNFGMVQGMSTRKGTVVFLDNILEETKEKMHEVMKKNENKYAQIEHPEEVADLVGISAVMIQDMQGKRINNYEFKWERMLSFEGDTGPYLQYAHSRLRSVE
RNASGITQEKWINADFSLLKEPAAKLLIRLLGQYPDVLRNAIKTHEPTTVVTYLFKLTHQVSSCYDVLWVAGQTEELATARLALYGAARQVLYNGMRLLGLTPVERM-

Figure 3. Comparisons of codon frequency with hydrophobic clusters analyses (HCA) of highly conserved arginyl-tRNA synthetases in *Escherichia coli, Pyrococcus horikoshii, Saccharomyces cerevisiae* show that rare codons are often found between hydrophobic clusters. The data provide evidence that protein folding occurs progressively, such that the number of exposed hydrophobic residues is minimized and structurally significant polar contacts are made. Notice that residues colored in gold and red (codon frequencies <10% and <5%) appear to occur at the beginning and end of hydrophobic clusters. The HCA plot is a bidimensional plot that is associated with the alpha helicoidal pitch. This allows for ease in visualization of protein secondary structure. The plot depicts hydrophobic residues as green, and groups them if they form a cluster. Negatively charged residues are colored red, and positively charged residues are colored blue. Prolines are depicted as red stars. Serine and threonine are depicted as a black box within a white box, and alanine is depicted as a white box. Glycine is depicted as a black rhombus.

Methods



Gene of Interest (nucleotide sequence)

codonBiasPos.py Gives us a colored sequence right in the terminal.

Sequence Alignments of colored sequences

Hydrophobicity Clustering (HCA)

Secondary Structure Prediction (PORTER)

Open Reading Frames

codonBiasColor.py Gives us a parsable output consisting of colors and residue positions.

colorFreq.py Sorts output into a file that can be loaded into PyMOL

Structural alignments in PyMOL to compare positional frequencies

Structural visualization of codon frequency in PyMOL

Figure 4. Block diagram that outlines our underlying bioinformatics pipeline.

Discussion

Throughout our work on this project, the correlation between codon frequency and protein structure has become increasingly clear. While there appears to be some conservation of rare codons in the MSA (Fig. 1), there is a lot of noise, and the link between codon frequency and structure is not apparent. This is due to the methods the sequence alignments use, not fully taking into consideration the 2D and 3D structures of the proteins being aligned. With this in mind, we look towards structural models to shine light upon the underlying mechanism that governs the conservation of codon frequencies.

In each crystal structure (Fig. 2), the rare codons appear to outline secondary structures, but there are many exceptions to this observation. Many alpha helices have rare codons embedded within the helix, and therefore we concluded that there must be something more to it that we are missing.

We came across the HCA program in search for a better way to analyze our data and found that rare codons appear to occur directly after hydrophobic residues, and in some cases, after structurally important features, such as kinks. This evidence suggests that the occurrence of rare codons plays a vital role in ensuring correct protein folding; more specifically, their occurrence marks a ribosomal pausing event in which the nascent polypeptide is able to find its lowest energy conformation.

This work has widespread implications in the life sciences, most notably in protein engineering and synthetic biology as we plan to incorporate our findings into a codon optimization program. This work also applies to the study of prion diseases such as Alzheimer's and Parkinson's disease, as it shines light upon a potential underlying mechanism for protein aggregation. Furthermore, we plan to incorporate this work into a molecular dynamics simulation of protein folding to gain a more fundamental understanding of the mechanism central to our understanding of life.

Acknowledgements

We would like to thank David Bernick for providing us with the opportunity to work on this project through UCSC iGEM. We would also like to thank the rest of the 2015 UCSC iGEM team for contributing to the project.

References

1. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141,** 344–354 (2010).
2. López, D. & Pazos, F. Protein functional features are reflected in the patterns of mRNA translation speed. *BMC Genomics* **16,** 513 (2015).
3. Chartier, M., Gaudreault, F. & Najmanovich, R. Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics* **28,** 1438–1445 (2012).
4. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of co-translational folding. *Nat. Struct. Mol. Biol.* **20,** 237–243 (2013).
5. Chan, C. T. Y. *et al.* Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nat. Commun.* (2013). doi:10.1038/ncomms1938.Reprogramming
6. Chan, C. T. Y. *et al.* A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. *PLoS Genet.* **6,** 1–9 (2010).
7. Chevance, F. F. V, Le Guyon, S. & Hughes, K. T. The Effects of Codon Context on In Vivo Translation Speed. *PLoS Genet.* **10,** (2014).
8. Englander, S. W. & Mayne, L. The nature of protein folding pathways. *Proc. Natl. Acad. Sci.* **111,** 15873–15880 (2014).
9. Gaboriaud C, Bissery V, Benchetrit T, Mornon JP. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett (1987) 224 (1): 149-155