Integrating Codon Usage Bias into Protein Folding Stability Analysis

Understanding protein folding stability is crucial in molecular biology, as it impacts protein function and cellular processes. One significant factor influencing protein folding is codon usage bias—the non-random usage of synonymous codons in the coding sequences of genes. Codon usage bias arises due to evolutionary pressures such as translational efficiency, accuracy, and gene expression levels.

Research, such as the comprehensive review by Parvathy et al. (2022), highlights how codon usage bias affects not just the speed of translation but also the co-translational folding of proteins. Rare codons, which are less frequently used in the genome, can slow down translation at specific sites, allowing the nascent polypeptide more time to fold properly. This mechanism suggests that the proportion of synonymous rare codons in an mRNA sequence could be a primary determinant of protein folding stability.

Hypothesis

Based on these insights, our hypothesis is:

The proportion of synonymous rare codons (codon usage bias) in the spliced mRNA sequence is a primary factor in determining protein folding stability.

Testing the Hypothesis

To test this hypothesis, we propose a systematic approach involving computational analysis and classification of proteins based on their folding characteristics and intrinsic disorder tendencies.

Classification of Proteins

We will classify proteins into four categories based on two criteria:

Co-translational Folding: Whether the protein begins folding during translation.

Intrinsic Disorder: Determined by the median disorder tendency score (with a threshold of > 0.5 indicating intrinsic disorder).

This results in a 2x2 matrix:

Intrinsically Disordered (Yes), Intrinsically Disordered (No)

Co-translationally Folded (Yes), Class 1, Class 2

Co-translationally Folded (No), Class 3, Class 4

Calculating Disorder Tendency Scores

We use the IUPred2A software to calculate the disorder tendency scores of proteins. For example, for the tumor suppressor protein p53, we obtain the following JSON data:

{
"aiupred": [0.31906798, 0.3666469, 0.422062, 0.48549908, 0.54222655, 0.57280475, 0.52941346, 0.51476824, 0.49363083, 0.5071435, 0.52531326, 0.5269308, 0.53040254, 0.5677328, 0.58607996, 0.5794885, 0.4932471, 0.42208833, 0.36165988, 0.40498608, 0.4689802, 0.50728387, 0.49116153, 0.45801705, 0.42430568, 0.44740328, 0.48831582, 0.53219223, 0.5042823, 0.4971833, 0.48980412, 0.47431025, 0.49798086, 0.57478184, 0.6027384, 0.6129357, 0.5939865, 0.56798977, 0.5125191, 0.5032883, 0.56562865, 0.6721967, 0.7139982],
"seq": "YSPALNKMFCQLAKTCPVQLWVDSTPPPAPASAPWPSTSSHST",
"smoothing": "Default smoothing"
}

Calculating the median of the aiupred scores gives us the median disorder tendency score for p53. If this value is greater than 0.5, p53 is classified as intrinsically disordered.

Analyzing Codon Usage Bias

Using codon usage data from the Kazusa Codon Usage Database, we can determine the frequency of each codon in a given organism.

Synonymous codons that are used less frequently are considered rare codons.

For each protein's mRNA sequence, we:

Count the Total Codons: The total number of codons in the mRNA sequence.

Identify Rare Codons: Based on codon usage tables, we identify which codons are rare.

Calculate Proportion of Rare Codons: The number of rare codons divided by the total number of codons.

Developing the Predictive Model

Our model predicts protein folding stability based on the proportion of rare codons:

High Proportion of Rare Codons: May lead to slower translation, allowing proper co-translational folding and potentially resulting in stable protein structures.

Low Proportion of Rare Codons: Faster translation may not provide sufficient time for proper folding, possibly leading to unstable proteins.

Validation and Confusion Matrix

We validate the model by comparing its predictions with known data on protein folding stability. This involves creating a confusion matrix for each of the four classes:

Predicted Stable, Predicted Unstable

Actual Stable, True Positives, False Negatives

Actual Unstable, False Positives, True Negatives

We calculate the accuracy for each class and hypothesize that the model will be most accurate for proteins that are co-translationally folded and intrinsically disordered (Class 1).

Executing the Model: A Step-by-Step Example

Select a Protein

Let's consider the p53 protein as an example.

Calculate Median Disorder Score

From the aiupred array, calculate the median value.

If the median score > 0.5, classify p53 as intrinsically disordered.

Determine Co-translational Folding

Based on literature, determine if p53 is known to fold co-translationally.

Assume p53 folds co-translationally (for this example).

Analyze Codon Usage in p53 mRNA

Obtain the mRNA sequence for p53.

Translate the mRNA sequence into codons.

Use codon usage data to identify rare codons.

Calculate the proportion of rare codons.

Predict Folding Stability

If the proportion of rare codons is above a certain threshold (determined from statistical analysis), predict that p53 will have stable folding.

Compare with Actual Data

Use experimental data on p53 folding stability to validate the prediction.

Update the confusion matrix accordingly.

Expected Outcome

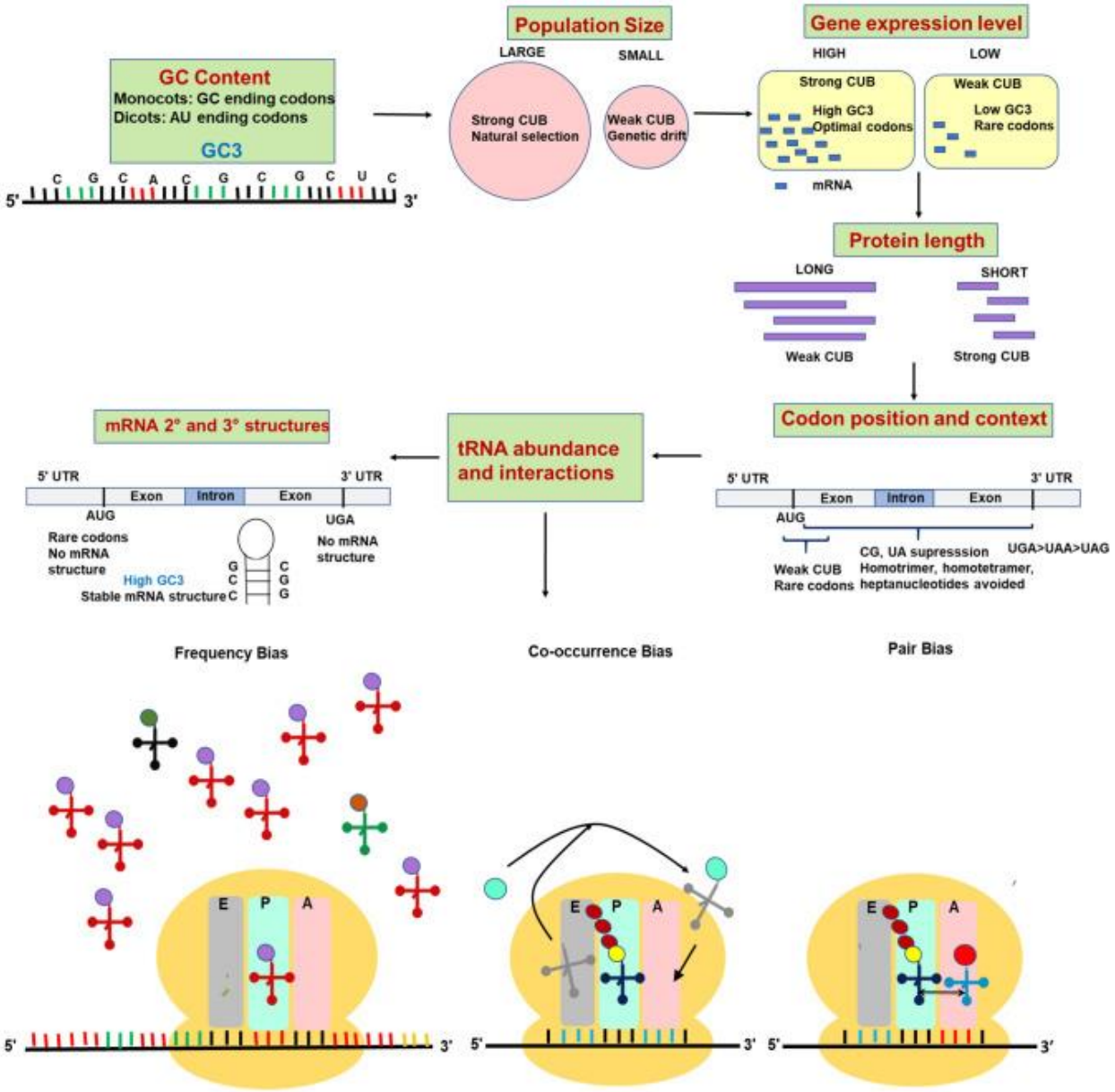
Given that intrinsically disordered proteins (IDPs) like p53 often rely on specific folding pathways and may be sensitive to translation kinetics, we expect our model to perform well in predicting folding stability for Class 1 proteins (Yes-Yes).

Conclusion

Our approach integrates the concept of codon usage bias into the analysis of protein folding stability. By examining the proportion of synonymous rare codons in mRNA sequences and correlating it with protein folding characteristics, we aim to uncover patterns that contribute to protein stability.

This method not only tests our hypothesis but also provides a framework for predicting protein folding outcomes based on genetic sequence data. The logical flow from codon usage to protein folding stability underscores the intricate relationship between genetics and protein biochemistry.

Limitations



Next Steps

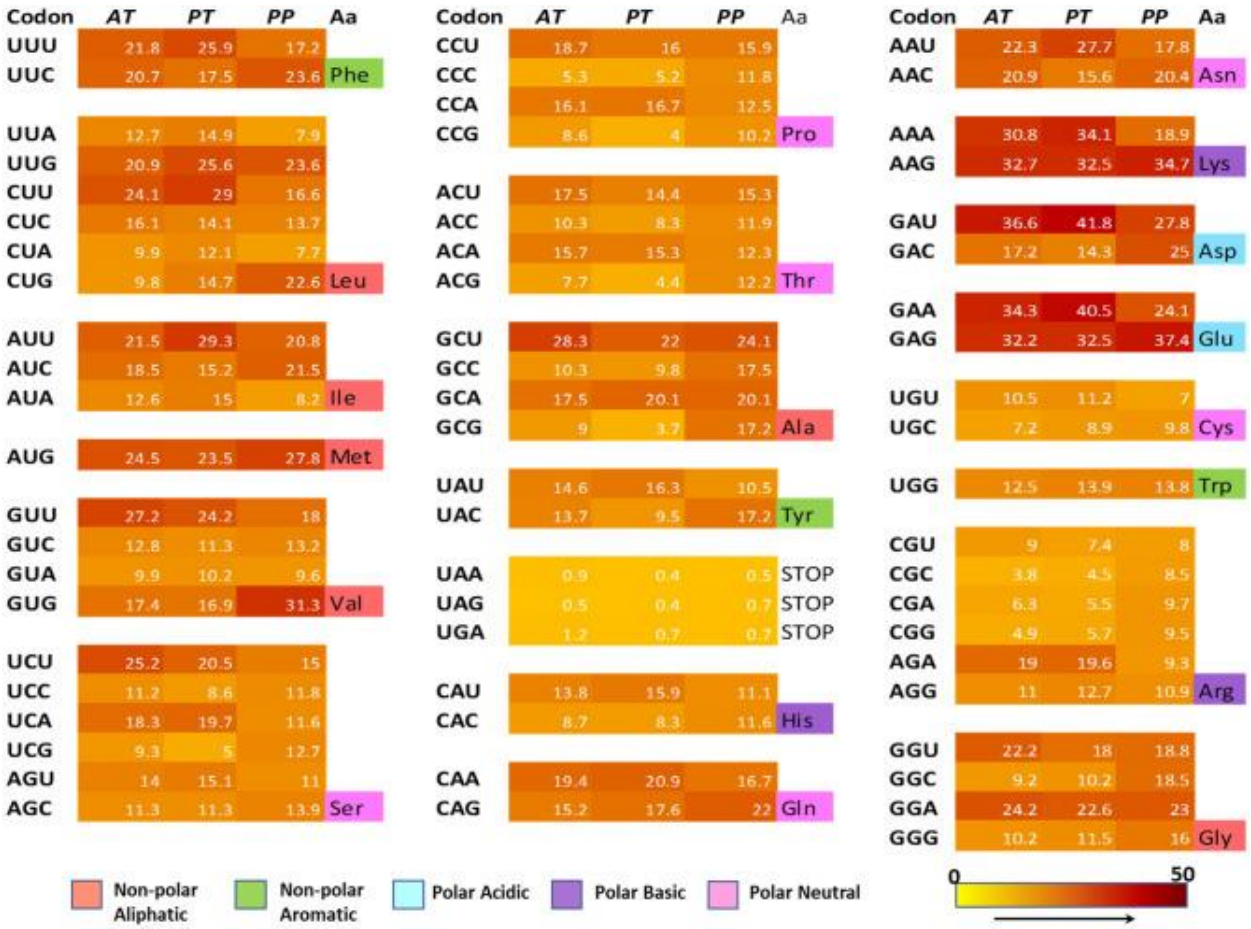
Data Collection: Compile a dataset of proteins with known folding characteristics and mRNA sequences.

Model Refinement: Adjust thresholds and parameters in the model based on initial results.

Statistical Analysis: Perform statistical tests to assess the significance of the findings.

Further Validation: Expand the analysis to more proteins across different organisms to generalize the model.

Metrics used:



Relative synonymous codon usage (RSCU) is used to analyse codon usage variation between genes. RSCU value for a codon is the observed frequency of the codon divided by the expected frequency of the same codon within a synonymous codon group in the entire coding sequence of the gene, under the assumption of equal usage of the synonymous codons for an amino acid or in the absence of codon usage bias. Expected number of occurrences of a codon is ratio of the number of times the encoded amino acid is present in the protein sequence to the number of synonymous codons for the amino acid encoded by codon. An RSCU value of 1 shows no codon bias and greater than 1 means that a codon is used more often than expected, while values less than 1 indicate its relative rarity [9]. RSCU values can be 2, 3, 4 and 6 when a single

codon is used to encode aminoacids having 2,3,4 and 6 synonymous codons respectively.

Data

Identified synonymous rare codons in Homosapiens based on calculated RSCU values

Class 1 Proteins

Class 2 Proteins

Class 3 Proteins

Class 4 Proteins