The RNAStructuromeDB consists of "windows" which give a glimpse into the underlying genome (hg38/GRCh38) sequence's ability to generate a structured RNA molecule. Since the genome is so large, over 3 billion nucleotides (nt), we first fragmented it into 154 million overlapping windows (120 nt long, stepping 40 nt). Each window was computationally transcribed into RNA and analyzed for its ability to then become structured. This "ability to form a structure" is described by five metrics which were calculated via ViennaRNA and a Perl script:

1. Minimum Free Energy

The 120 nt sequence of RNA described in each window is first analyzed with an RNA folding algorithm to predict the most stable structure it could theoretically adopt. A good description of how this works by Sean R. Eddy here.

The **Minimum Free Energy (MFE)** is the free energy value of this most stable structure in kcal/mol. The more negative the value, the more stable the structure.

2. Z-Scores

It was found that, in general, structured RNA molecules adopted more stable (lower MFE) structures than shuffled versions of the same sequence (as described by Clote et. Al., in **Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency**). This was a phenomenon that would then be used to **detect regions of RNA which *could potentially* be structured**: as inferred by the thermodynamic **z-score**:

$$\boldsymbol{z-score} = \frac{MFE_{genome\ sequence} - \overline{MFE_{30\ scrambled\ versions}}}{\sigma}$$

This was calculated for each 120 nt window of the genome. We took the sequence and scrambled it 30 times. We then have two sets of sequences: native and scrambled. For each set, MFE values were calculated. If the native sequence always has a much lower MFE than the average of scrambled versions this will lead to a negative z-score (if the native sequence MFE is always more positivie, i.e., less stable, then the z-score will be positive). The equation normalizes the value by dividing by the standard deviation between all MFEs. The magnitude of the z-score then, states the number of standard deviations the native (window) MFE is from the random MFEs.

> ***Negative z-score*** **indicates a window generates a *more stable* structure than random.**
> ***Positive z-score*** **indicates a window generates a *less stable* structure than random.**

3. P-Values

This value is directly related to the z-score. It is simply the fraction of random sequences which were more stable (more negative, or less than) the original sequence:

$$\frac{\#\ of\ MFE_{30\ scrambled\ versions} < MFE_{genome\ sequence}}{30}$$

Therefore, a value of "1" indicates that all of the random sequences were actually more stable than the native and a value of "0" indicates the native was more stable than all random sequences.

4. Ensemble Diversity

When calculating an MFE structure/value for a particular sequence, we are always finding a *single* result: the *minimum* free energy structure/value. However, theoretically, there are many possible structures/values. The ensemble diversity is a metric which attempts to describe the *variety* of possible structures. How is this determined? All possible structures are first calculated. The probability of the RNA adopting one of these structures is calculated using the partition function (described by McCaskill). This partition function can then be used to measure the "diversity" of possible structures. **If the structures are very similar (different by only a few base pairs) the ensemble diversity will be low, however, if there are a wide variety of structures possible, the ensemble diversity will be high.**

A helpful program to visualize the *character* and *quantity* of these alternative structures is EnsembleRNA. Here you can generate a helpful visualization to conceptualize the conformational landscape of an RNA molecule, and determine if a particular sequence has closely related or decidedly unique alternative folds.

Alternatively, if you have a structure which you would like to "lock" into a specific conformation, we have a tool, RNA2DMut, which can be used to identify single point mutations which can minimize ensemble diversity (or maximize it).

5. Frequency of MFE

This value is related to the ensemble diversity in that it utilizes the same partition function calculations. Specifically, it measures the *frequency of the MFE structure* which was found within the calculated ensemble.