# A short statistical investigation of RNA features used in classification with RNAz

Christopher

November 25, 2020

## 1 Introduction

The identification of features useful in the discrimination of ncRNA loci in whole genome screens remains a challenging task. RNAz 2.0 uses three distinct features to allow for clear classification: A measure for structural conservation (SCI), the z-score of minimum free folding energy and the shannon entropy as a metric for sequence diversity. However, to this point, no comparative statistical investigation regarding the overall distribution and feasibility of these properties of RNA has been published. For this reason, this short study should serve as an entry point to get a glimpse of the underlying picture. Using the aforementioned features, it has been shown that the creation of a classifier that satisfies normal discriminatory precision requirements is possible without further modifying the background model. Evidence suggests however (refer to the Master Thesis for further details), that even with a hypothetical ideal classifier based on these features, there are certain classes of ncRNA that are more reliably classified than others (most notably miRNA). On the other hand, C/D-box snoRNAs were shown to be exceedingly difficult to reliably identify,potential reasons were already discussed in the mentioned thesis. This leads to the assumption, that the map painted by the feature tuple of SCI, z-score and entropy is incomplete or not yet optimal for full identification of ncRNA properties. For this reason, a representative data set was investigated to obtain information regarding distribution, correlation and ultimately 'usefulness' of individual ncRNA properties.

## 2 Data

For this first study, a secondary test data set included with the original RNAz 2.0 release paper was used. It contains 4303 instances of curated structurally conserved RNA alignments, as well as 4303 negative samples, which were generated by choosing subsequences with approximately identical dinucleotide composition from random genomic locations.

# 3 Results & Discussion

For a first overview, features used in classification were plotted in a 3D coordinate system and the distribution of features for native and control set visualized with histograms. These results can be viewed in figures 1 and 2.

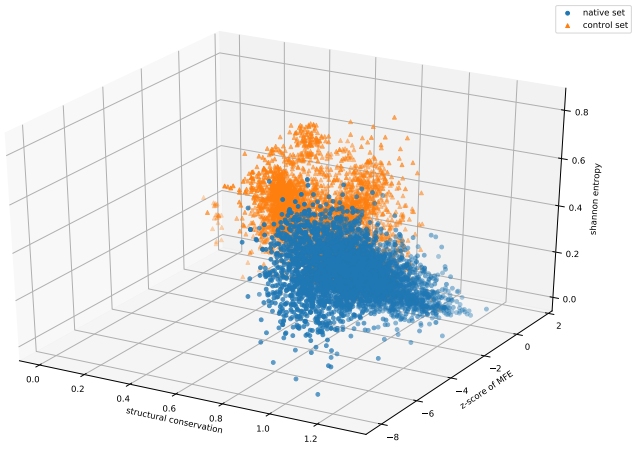



Figure 1: SCI, shannon entropy and z-score of MFE plotted as 3D coordinates to estimate overall distribution.

On first observation, it appears that the data point groups in figure 1 are almost perfectly separable by means of a diagonal plane with maximized euclidian distance to both groups. This itself indicates a strong separability of instances with the features at hand. However, it also appears as if the features are not contributing equally to the separation of the groups: Visually, both seem to have a roughly identical offset on the z axis, representing the Shannon entropy. Analyzing the value distribution, it can be clearly observed, that both z-score of MFE and structural conservation index by themselves both allow for a reasonably good linear separation of instances (see figure 2 (a) and (b) especially). On the other hand, the Shannon entropy parameter seems almost identically distributed between both native and control set, backing the observation that entropy might not contribute much to a clear classification.

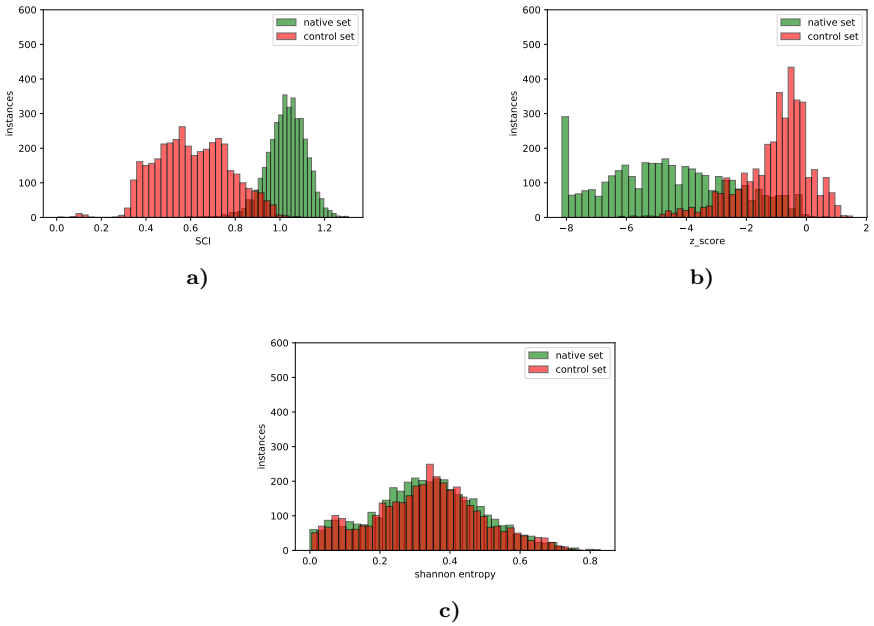



Figure 2: Comparison of feature distribution for native set of RNA alignments and control set. (a) shows distribution of structural conservation indices, (b) shows z-score of MFE and (c) the Shannon entropy.

Figure 3 furthermore shows the absolute value of each feature for all instances in both sets. Here, the interesting observation can be made, that the spreading range of the z-score is much wider in the native set than in the control set, while simultaneously being usually lower in numerical value. This implies, that the discriminatory nature of the z-score lies mainly in it being less than a yet unknown lower limit representing the average lower range of non-ncRNA z-scores of MFE. This however lies within expectations, as this parameter is mainly used to distinguish those RNA alignments with extraordinary minimum free energy. This of course is a property typical for RNA molecules with advanced secondary structure elements, which again are usually found in RNA with distinct biological or chemical functions.

The figure also backs the claim, that the SCI allows for a linear separation of native and control set, as the means of both 'bands' representing data points are offset by a value of approximately 0.5. On the other hand, it can again be observed that no such clear distinction can be made for the Shannon entropy (blue band).

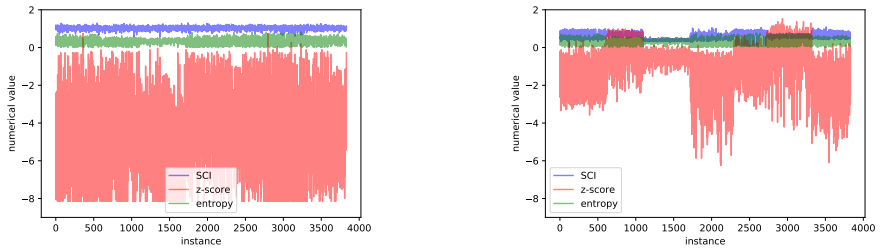
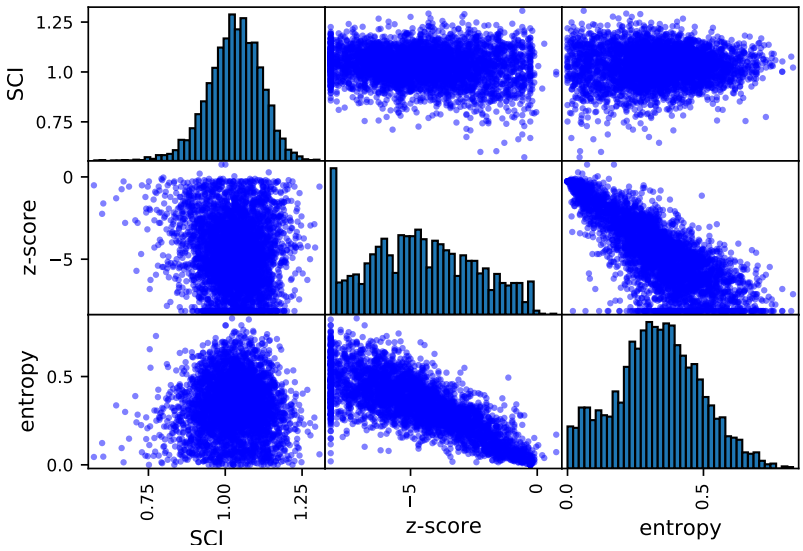
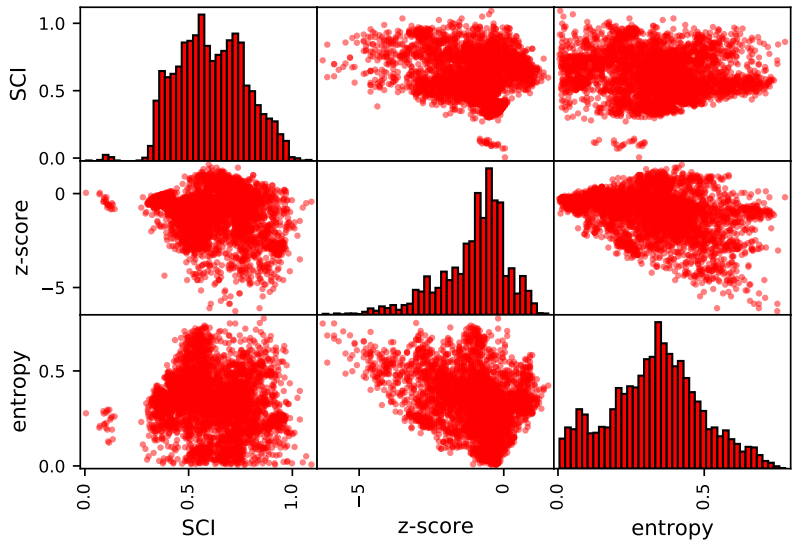



Figure 3: Absolute values of features for all data instances in direct comparison. (a) shows the values for the native alignment group and (b) for the control group. Note: The x axis obviously contains no useful information here, as the data points are not in order. For a better overview of value *distribution*, please refer to figure 2.

Now, after analyzing the overall distribution of each feature by itself, how do these features interact with each other in a two-dimensional space? Investigating the dependency of parameters helps the identification of potential correlation between properties, linear or otherwise, and as such can serve to determine redundant factors in the discrimination process. For this purpose, figure 4 visualizes an estimation of the covariance between all features by means of a scatter matrix. For this first estimation, native and control set are illustrated independently.




**a)**




**b)**

Figure 4: Matrix containing two-dimensional visualizations of all feature pair distributions. The diagonal line visualizes the overall distribution of single features for comparison purposes. (a) visualizes the native set data, (b) the control set.

It can be clearly seen, that the axis representing Shannon entropy is typically very similar between native and control set for all feature pairs, as was already established. From a direct comparison, it is easily observable, that the z-score and SCI feature pair both account for the clearest distinction between both sets. Furthermore, it can be observed, that the overall value range of feature pairs is tighter condensed in the control set for all feature pairs containing the SCI, as was also established above. In regards to the SCI, the data collected here suggests, that the distribution of values in the native set much closer follows as gaussian distribution, while the control set values appear to be spread more evenly in a given range. Interestingly, the opposite appears to be true for the z-score. If this observation holds true or is an artifact of the data at hand should be investigated further in the future.

Now, for a more direct comparison of both sets, a feature pair matrix containing both native and control data points can be seen in figure 5.

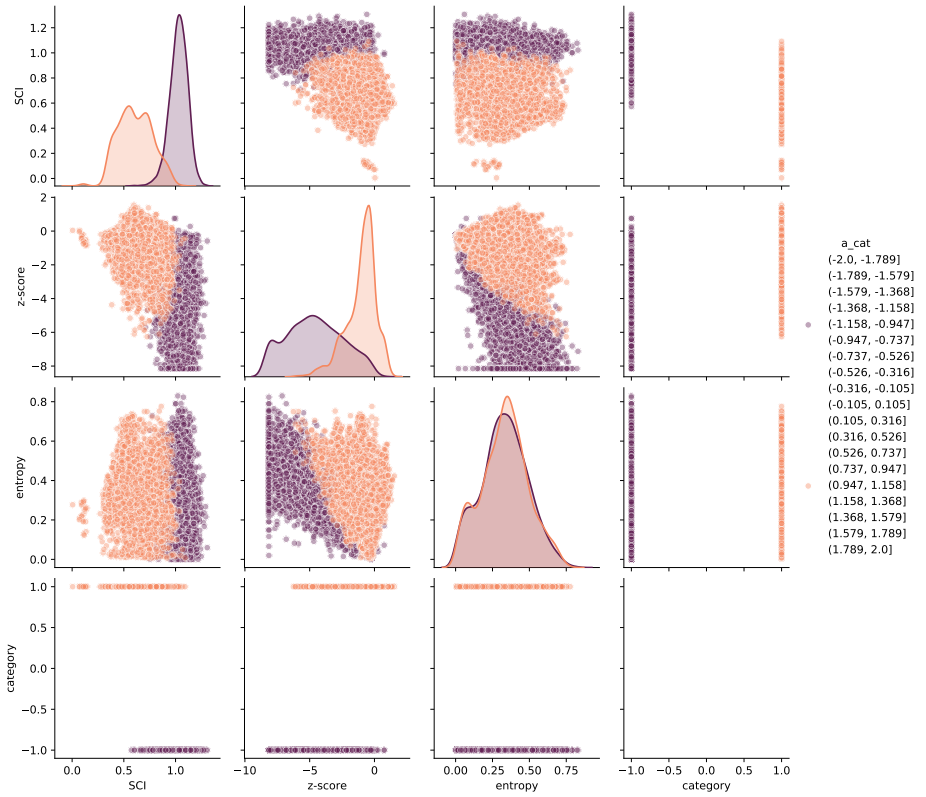



Figure 5: Grid of feature pair visualizations for both native and control set instances. The respective category is also contained to allow for a more natural comparison of single values between both sets. The diagonal contains estimated density functions for added comparison, except of course for the category property, as a density estimation for a binary value for two sets of equal size would be meaningless. Orange represents the negative set here, while violet represents the native alignment data.

Figure 5 supports what has been found so far: SCI and z-score allow for the most clear distinction between data sets, while Shannon entropy is distribution is almost identical between control and native set. In both cases, Shannon entropy closely approximates a normal distribution in density.

Finally, given the data above, it appears that SCI and z-score are mostly independent from each other. However, linear correlation analysis shows a strong covariation between Shannon entropy and z-score (see figure 6) with a Pearson coefficient of -0.77. On the other hand, no significant correlation could be found between the remaining features. Spearman correlation was also investigated, but yields no significantly different results. This is surprising in that the z-score appears to be a much better discriminant from the data analyzed here. It

is, however, very likely, that the previously observed different range of values is relevant here: Since Shannon entropy is confined to a much smaller 'band' of values (see figure 3), it is highly plausible that, while correlated, the numerical difference in entropy is too small to allow for significant discrimination here.

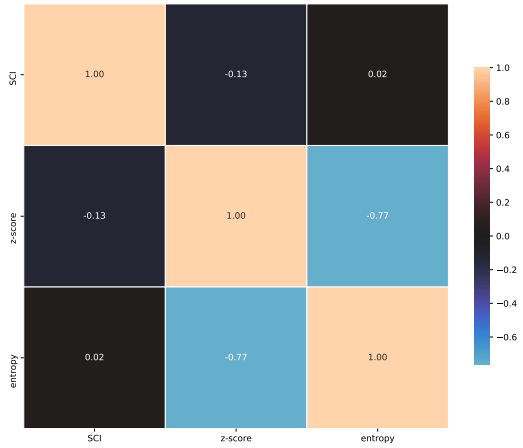



Figure 6: Linear correlation analysis of all features, using the Pearson correlation coefficient as metric.

# 4 Addendum: A potential discriminatory feature as reported by Ulveling et al.

Even though it has already been shown, that the three RNA features discussed up till this point are sufficient to achieve very high classification accuracy on new data sets, other properties with sufficiently high predictive power are obviously a point of interest.
Ulveling et al. reported in 2014 the establishment of a reference database of 'bona fide' long non-coding RNAs (lncRNA). Based on these samples, they introduced the respective CG site abundance in lncRNAs as a discriminatory feature for the identification of ncRNAs. In their test study, they suggested the introduction of the dinucleotide 'signature' of a given sequence, i.e. the percentage increase in CG-dinucleotide frequency when compared to the background frequency of the whole genome. Interesting as this idea is, it poses a certain problem for the RNAz approach: The software was explicitly designed to work without having full knowledge of the genomic composition at hand.

However, to estimate the principal validity of this approach, the data set used in this study was screened in a comparable manner by calculating a relative CG-frequency for each alignment as follows:

$$f = \frac{\sum_n^x \frac{CG_x}{len(seq_x)-1}}{n} \tag{1}$$

where $n$ is the number of sequences $x$ in an alignment and $CG_x$ refers to the number of instances of a G following a C within the sequence. The distribution of frequencies thus obtained is visualized in figure 7.

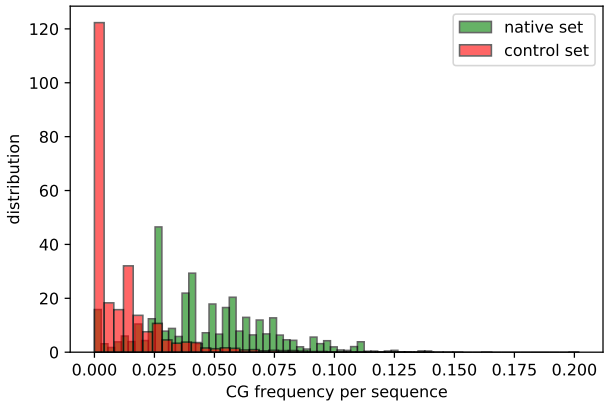



Figure 7: Total distribution of CG site frequencies per alignment in both data sets.

It can be observed, that the differences in distribution are indeed significant. In fact, calculating the mean CG frequency over all alignments yields a value of 0.0466 for the native set and 0.0112 for the control set, indicating that the CG dinucleotide is roughly 4.5 times as likely to be found in an ncRNA test alignment. Even if one completely disregards all the alignments with zero CG sites, the factor is still approximately 2.2 (mean values of 0.0487 and 0.0222 respectively). This difference in distribution is high enough, that further consideration and more investigation seems justified.

# 5 Conclusion

It has been shown, that the contribution of Shannon entropy as a discriminatory factor might be marginal after all. On top of this, the strong linear correlation with the z-score observed here implies a degree of redundancy as a classification component. Obviously, the data set at hand might be insufficient in size to claim representative power. Therefore, the study here will have to be repeated with a larger sample size.
Furthermore, evidence was presented, that the overall difference in CG site distribution even on the scale of the isolated alignment window is very significant. This initial observation should be further validated, as it might form the basis of one or more components usable in the successful categorization of ncRNA elements.