Super Information

Manav Mishra and Jai Kumar, IISER Bhopal

16th November, 2018

1 Introduction

A gene is basically a sequence of DNA that contains hereditary data (Genetic information). Several of the varied techniques have been explored in the past so as to analyze the information content in the DNA. A well known statistical feature of Shannon entropy gives us the entropy of DNA. But is it enough? Does it really give us the whole information content of the DNA?

In this project, we have implemented an alternate measure of information, known as Super information, which was found to be an effective tool for analyzing the coding and non-coding region of a DNA sequence. It is a measure of 'randomness of randomness'. It has been found out to be a highly accurate in classifying coding and non-coding regions.

We have also implemented two other alternate measure of information which also helps in classification of coding/non-coding region. To see the overall performance of all three methods, we implemented a machine learning algorithm to get a high accuracy classifier.

2 Motivation for Super Information[1]

In all previous attempts to analyze the information content of genetic data, Shannon's definition of entropy has been used. Its mathematical definition is given as:

$$H(X) = -\sum_{i=1}^{N} p(x_i) log_2 p(x_i) \tag{1}$$

Randomness (disorder) implies a higher information content. But in computing the frequencies of bases in a sequence, we do not take the position of the bases into account. As what the name suggests, it gives us the information in the average sense and may not provide the true picture in many cases.

Consider the sequences,

S_1: AAAAGGGGTTTTCCCC,
S_2: ATCGATCGATCGATCG,
S_3: ATTGACCCTGTCGAGA.

Here, S_1 and S_2 have a recurring pattern, while S_3 looks quite disordered. However, for all three cases, the Shannon entropy is the same. This is a serious limitation of Shannon entropy in analyzing DNA sequence.

This motivates us to look for some alternate information measurement techniques. We need to look for a new way to define randomness.

3 Super Information

We saw $H(X)$ is not a good measure for analyzing sequences of symbols with repeating patterns, as is common in DNA sequences. So we start by sub diving the sequence into blocks, and then find out the information content of each block. Depending on the sequence under consideration, different blocks will have different measure for randomness.

Thus, there is some randomness in the measure of randomness. This measurement quantity is what we term as Super-Information. Intuitively, Super information represents the variation in the entropy content of different portions of a given sequence. If there is a large variation in the entropy of different segments, the super information will be higher.

Mathematical Definition
Let us divide the entire sequence of symbols, X into N blocks of length B each. Let's call the i^{th} sequence X_i and the Shannon entropy associated with it as $H(X_i)$.

- Construct the histogram of H(X_i), i.e.

$$H_j(X_i, M) = \text{histogram } H(X_i))$$

i=1,2,3....N j=1,2,3....M

where, $H_j(X_i, M)$ is the number of element in j^{th} element.

- Normalize the H_j to get the probabilities.

$$p_j(X_i, M) = \frac{H_j(X_i,M)}{\sum_{k=1}^{M} H_j(X_i,M)}$$

- Super-Information is defined as :

$$H_s(X_i, B, M) = -\sum_{j=1}^{M} p_j(X_i, M) ln(p_j(X_i, M))$$

4 Analysis of Coding and Non-coding region

We have used Super information as a mathematical tool to study the coding and non-coding regions of DNA. We have taken the different prokaryotic genomes as input data for this analysis.

What we observe is that $H_s(B, M)$ has a maximum for certain values of B. This can be explained as follows. For smaller values of B, a smaller number of entropy values are possible. So, when calculating the entropy of entropy, $H_s(S, B)$ comes out to be small. At large values of B, the spread of the entropy function is less, leading to a smaller $H_s(S, B)$.

B $\in$ (80, 100) gives the maximum value for most cases. Hence, we fix B=90.

5 Results

5.1 Super Information [1]

After getting all the coding and non-coding regions and concatenating each in two coding and non-coding sequences, the plot gives us the following Super-Information with increasing B value.

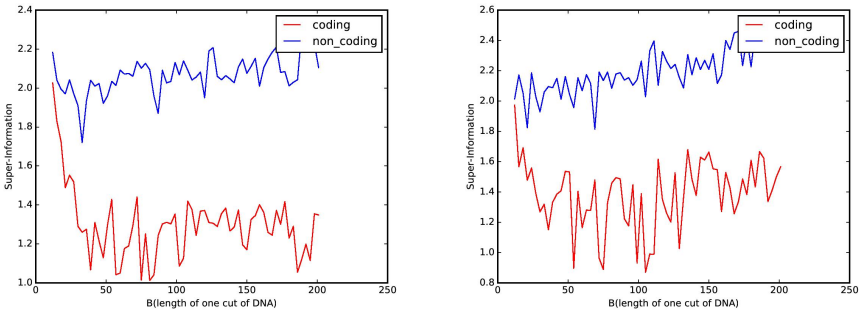

Figure 1: NC_000911 and NC_002570

Super-Information of some separate coding and non-coding regions at different B values out of different genomes (without concatenating any parts).

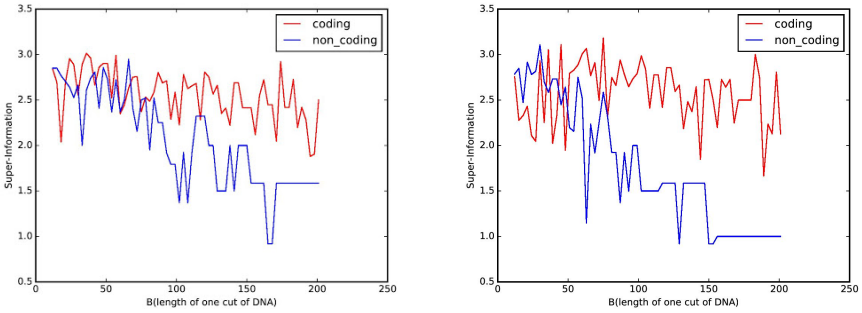

Figure 2: NC_000911 and NC_000915

5.2 P-value [2]

Apart from the Super Information measure, we have used the Fourier transform method as well to classify the given sequence as coding/non-coding.

P-value of codons and non-codons sequences in genome.

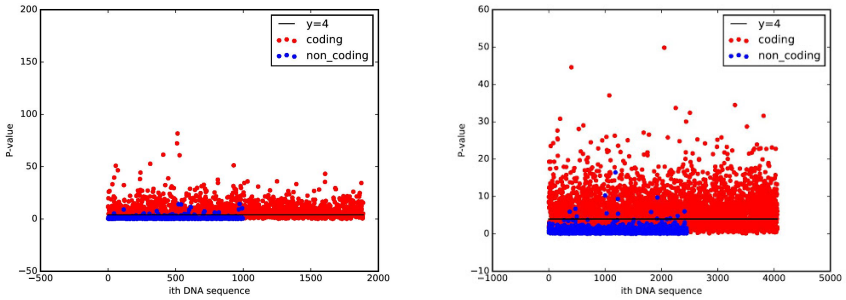

Figure 3: NC_003098 and NC_002570

5.3 F_3-measure [3]

Another useful method that we have used for information analysis is the F_3-measure, which is again a powerful tool for DNA sequence analysis.

F3 value for coding and non-coding regions of sequence of genome.

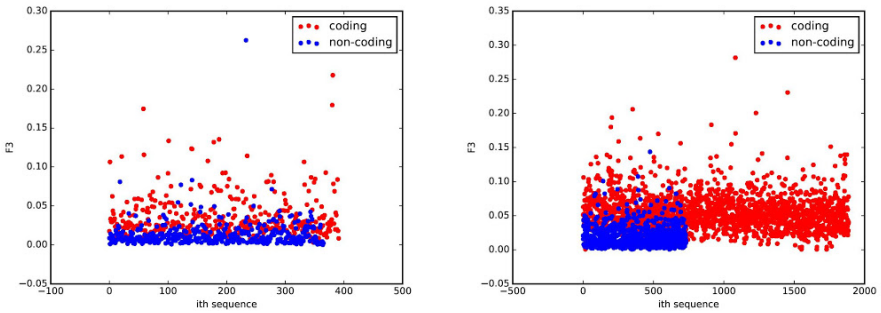

Figure 4: NC_001142 and NC_003098

6 Further analysis by Machine Learning

On obtaining such interesting results, it was necessary to understand the role of these three techniques and how can we effectively use them to classify a given DNA sequence as a coding or a non-coding region.

Goal: Classification of Coding and Non-coding region for a given base-pair sequence.

Input Features:

- H_s value [1]
- P-value [2]
- F_3 value [3]

We have generated a data-set of about 40,500 DNA sequences using the Main_ML.py program. The data-sets that we are considering is of prokaryotic genome only.

6.1 ML using Tensorflow module

We have employed a ML module in Python known as Tensorflow for the binary classification. In there, we used a backend API known as Keras to implement the ML algorithm.

We were successful in making a classifier giving us an accuracy of 88%. This is quite remarkable for the limited set of input features we have used.

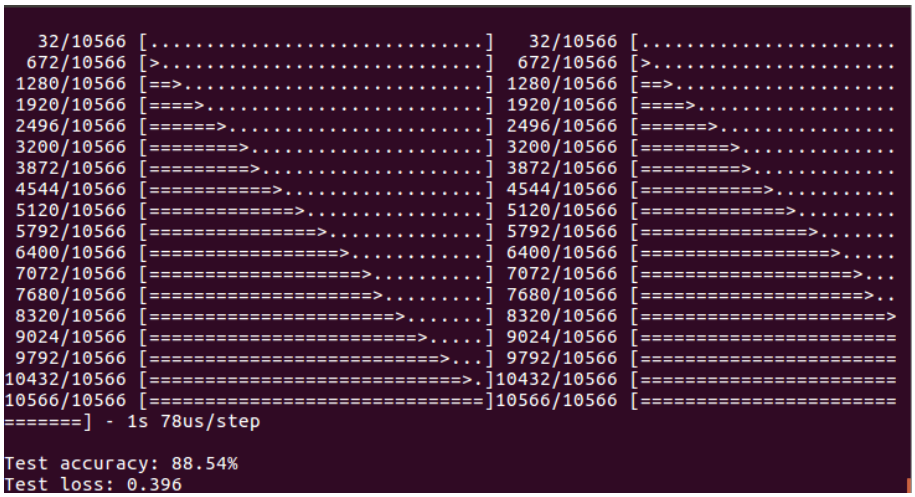

Figure 5: ML output

6.2 ML using WEKA toolkit

We ran self created .arff file on WEKA toolkit having P-value, H_s value, F_3 value as Real values and classifiers as coding and non-coding. We have used Random Forest Tree algorithm. We got different accuracy for different B-values (different length cut of DNA sequence to get H_s).

```
=== Summary ===

Correctly Classified Instances       27970               91.7139 %
Incorrectly Classified Instances      2527                8.2861 %
Kappa statistic                          0.7964
Mean absolute error                      0.117
Root mean squared error                  0.2549
Relative absolute error                 28.798  %
Root relative squared error             56.565  %
Total Number of Instances            30497

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.941    0.143    0.943      0.941    0.942      0.796    0.956     0.977     coding
                 0.857    0.059    0.852      0.857    0.854      0.796    0.956     0.884     non_coding
Weighted Avg.    0.917    0.119    0.917      0.917    0.917      0.796    0.956     0.951

=== Confusion Matrix ===

     a     b   <-- classified as
 20563  1287 |     a = coding
  1240  7407 |     b = non_coding
```

Figure 6: For B=50

```
=== Summary ===

Correctly Classified Instances       35693               87.9875 %
Incorrectly Classified Instances      4873               12.0125 %
Kappa statistic                          0.7378
Mean absolute error                      0.1683
Root mean squared error                  0.3035
Relative absolute error                 36.8595 %
Root relative squared error             63.5256 %
Total Number of Instances            40566

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.903    0.162    0.911      0.903    0.907      0.738    0.934     0.961     coding
                 0.838    0.097    0.824      0.838    0.831      0.738    0.934     0.861     non_coding
Weighted Avg.    0.880    0.139    0.880      0.880    0.880      0.738    0.934     0.925

=== Confusion Matrix ===

     a     b   <-- classified as
 23713  2554 |     a = coding
  2319 11980 |     b = non_coding
```

Figure 7: For B=20

We have found different accuracy when we juggle around with no. of attributes as input.

Attributes	**Accuracy(%)**
All	87.98
F_3	64.75
F_3, P-value	87.09
F_3, H_s	82.56
H_s	78.64
H_s, P-value	85.97
P-value	81.75

Table 1: At B=20

7 Conclusion

We were successful in understanding and replicating the results of Super information as an alternate measure of classifying a DNA sequence as coding/non-coding. We have provided the steps for determining this quantity. Superinformation has been found to be very effective for classifying coding and non-coding regions.

We implemented ML classification using Tensorflow having an accuracy of 88%. Similar implementation was done on WEKA toolkit giving us an accuracy of 90%. The accuracy can be further increased if we get more input features (measures of Information in DNA).

From different combination of attributes we can see 'P-value' is having high distinctive ability than H_s and F_3.

8 Future goals

1. To implement the Super information measurement on Eukaryotic genome sequences.
2. To finally use data of human chromosomes to get more useful complex results.
3. Adding more input features to improve the ML classifier accuracy.

9 Acknowledgements

We would like to thank Dr. Kushal Shah for giving us his valuable time and knowledge during this course project.

Our sincere thanks to Dr. Nagarjuna Vijay for his inputs and suggestions for our future endeavours.

A special mention to Dr. Rajakrishnan R. for giving us this opportunity to start this project, for making this course super awesome, and for his chai-samosa treat.

References

[1] R. Bose and S. Chouhan. *Alternate measure of information useful for DNA sequences.* Physical Review E 83, 051918, 2011.

[2] A. Bhattacharya S. Bhattacharya S. Tiwari, S.Ramachandran and R. Ramaswamy. *Prediction of probable genes by Fourieranalysis of genomic sequences.* Computer applications in the biosciences : CABIOS, 1997.

[3] C. Yin and S-T. Yau. *A Fourier Characteristic of Coding Sequences: Origins and a Non-Fourier Approximation.* J. Computational Biology 12, 1153, 2005.