In his paper “Locating Protein Coding Regions in Human DNA using a Decision Tree Algorithm,” Salzberg describes how decision trees were used to classify regions in DNA sequences. Several different coding measures were used to provide a higher accuracy then was attainable with previous approaches. This algorithm is efficient and can be easily be adapted to handle various lengths of base pairs.

With the Human Genome Project and other DNA sequencing efforts, more and more data is being generated. Much of this data has not been classified to see whether it codes a protein or not. As with other areas, it would be helpful if this could be done automatically by a computer. Previous efforts have focused on linear discriminants and neural networks. Measures previously used in the identification of DNA sequences include: codon frequencies, dicodon frequencies, fractal dimensions, repetitive hexamers, as well as other features. The set of nineteen features used in the experiment consist of:

1. Dicodon usage
2. Hexamer-1
3. Hexamer-2
4. Open reading frame
5. Diamino acid usage
6. Codon usage
7. Run
8-11. Position asymmetry (a, c, g, t)
12-19. Fourier coef. for periods 2-9

The difficulty of this problem lies in the classification of short sequences.

This study used the OC1 algorithm and 21 coding measure to divide the data into two
classes: coding and noncoding. The OC1 can use both univariate tests, which depend on only one variable, and oblique splits, which are a multivariate combination of features. OC1 will find the best univariate test first and then consider if an oblique split will work better. Salzberg also created his own measure of the goodness of the split called “average accuracy.” This measure averages the accuracy of each class from a given split instead of looking only at the accuracy of the larger class.

The data set used for these tests included only sequences that were all coding, or that contained no coding. Also included in the data set was the reverse of every sequence. This was done to make identification harder. Context was not used in these experiments. All that was used were the sets of sequences that either contained coding or did not.

Decision trees were found to be more accurate and more consistent regardless of the length of the sequence than previous methods. The results show that as few as six features may be sufficient to be able to classify DNA sequences with a high degree of accuracy.