Nick Till

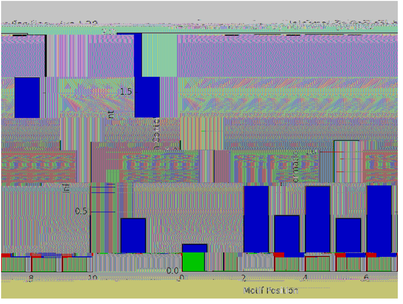
NRPS Code Project Summary

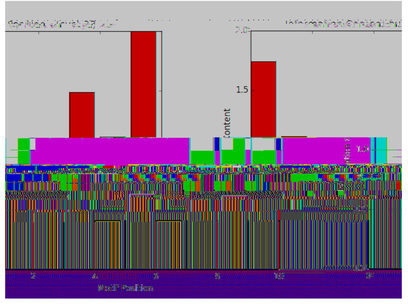
A. Data formatting/trimming

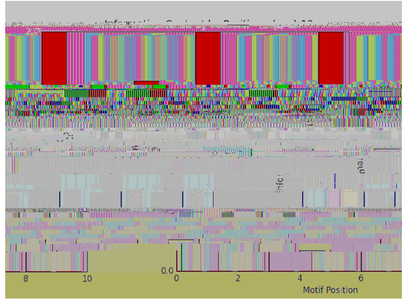
The data used in this project was obtained from a paper which detailed a machine-learning approach to the prediction of amino-acids encoded by a given A-domain in non-ribosomal peptide synthesis. The training data used for this approach was taken from this paper, in .xls format. Then, all unnatural amino acid residues were stripped, leaving 19 amino-acid encoding sequence types (each type contained 1-~30 sequences). No methionine encoding A-domains were in the training data, leaving only 19 amino-

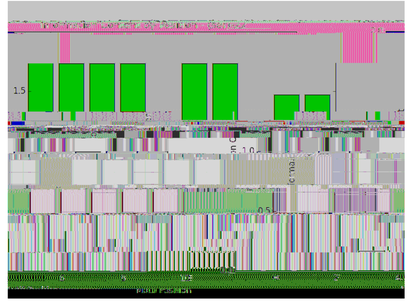
4. Calculate entropy at each position of each consensus (Entropy)
5. Calculate information content at each position of each consensus (IC)
6. Plot bar charts of sequence position vs information content (plotBarChart)
7. Plot bar charts for all A-domain sequences regardless of amino-acid specificity to see overall conservation
8.

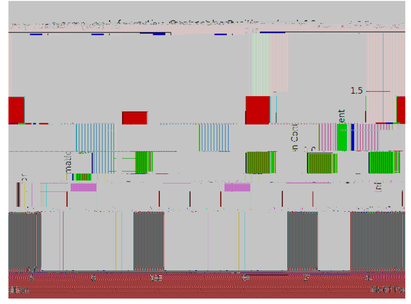
The most obvious problem with the current predictor is the necessity for extensive preparation of the data beforehand. The 34 residues taken as a

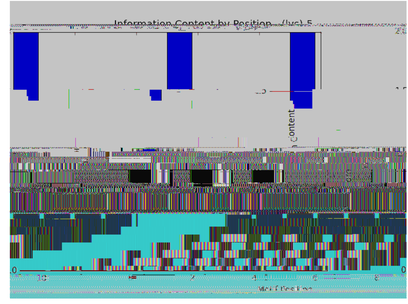


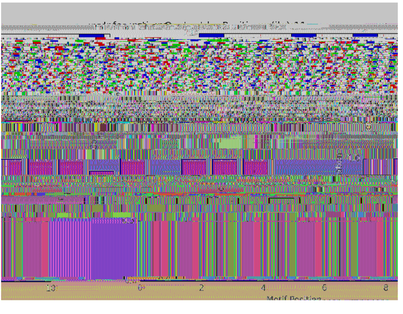


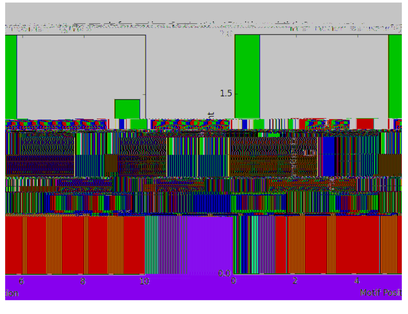


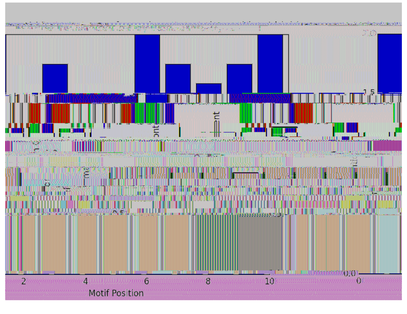


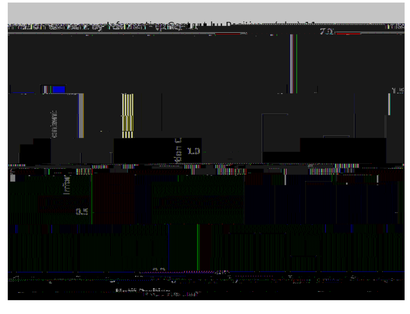


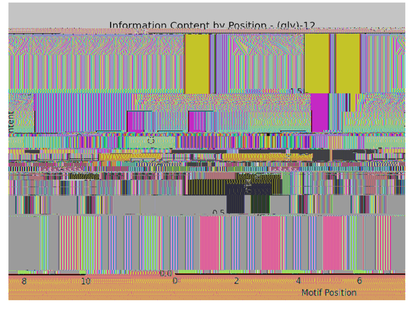


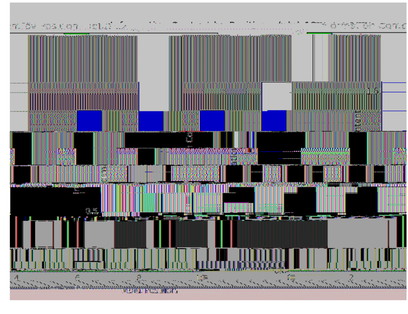


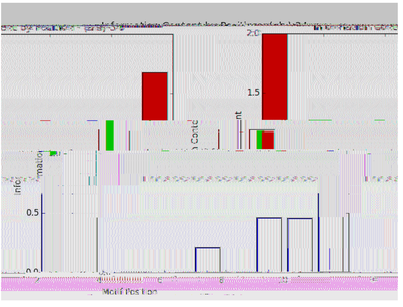


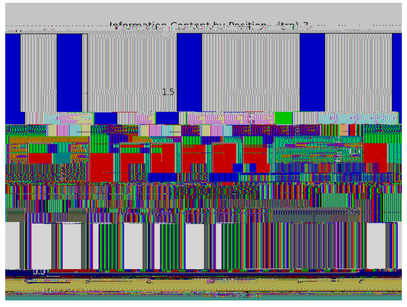


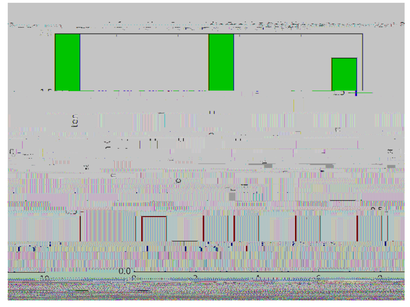


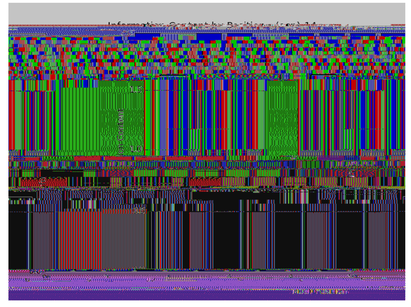


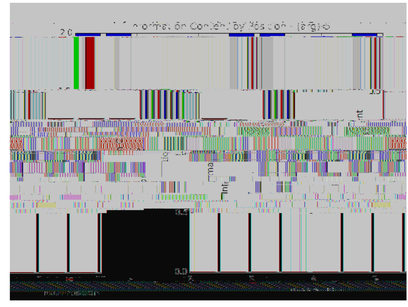


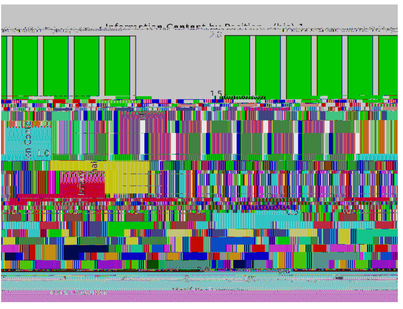




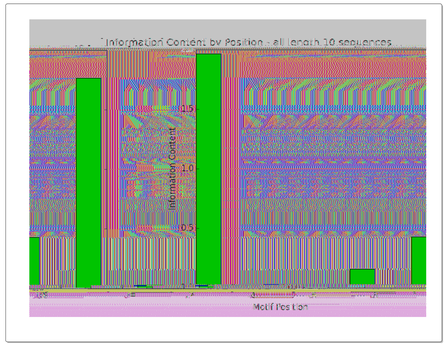




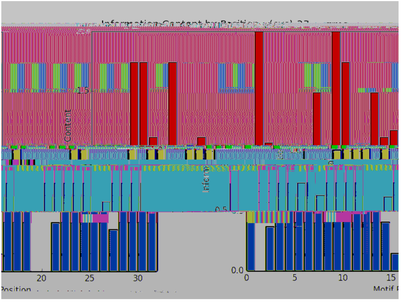


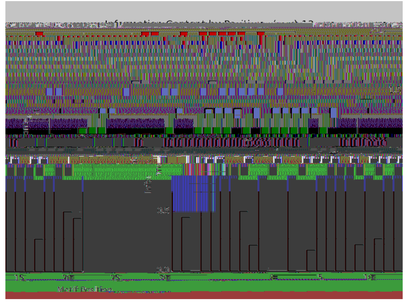


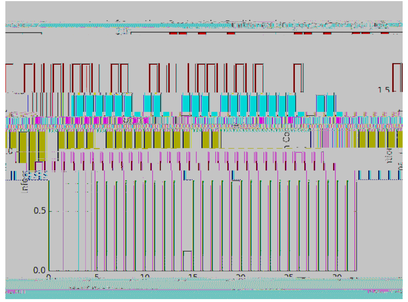
REFERENCE

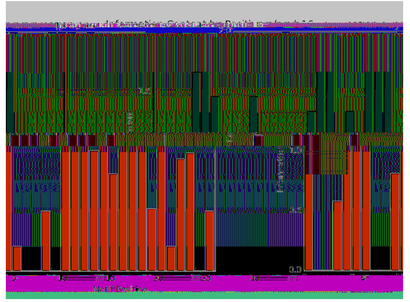


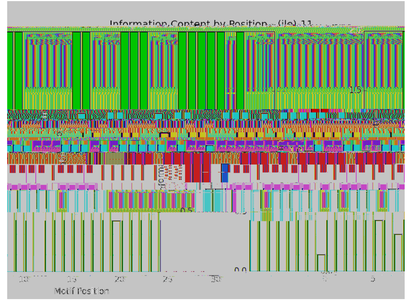
Information content by position for all 19 characterized amino acid A-domains, and a reference information content for all sequences (with length 10), regardless of A-domain specificity.

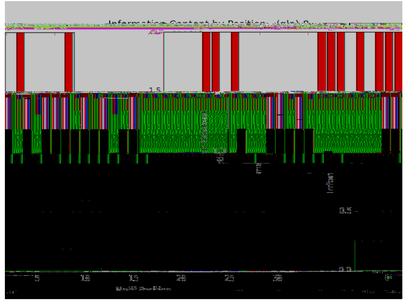


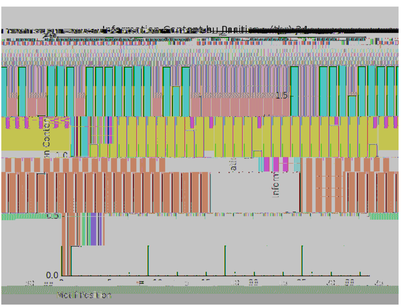


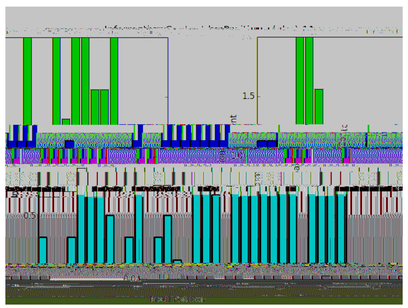


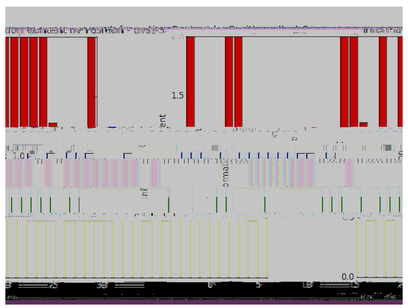


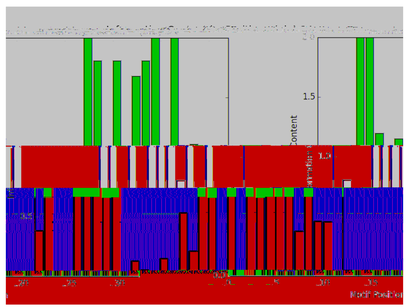


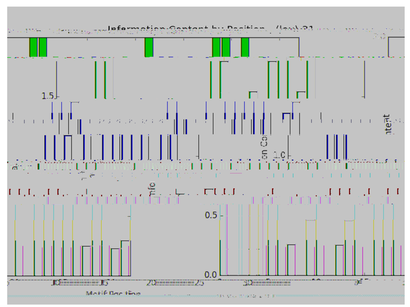


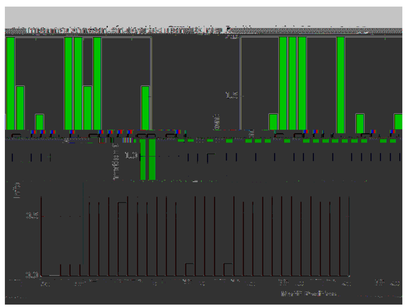


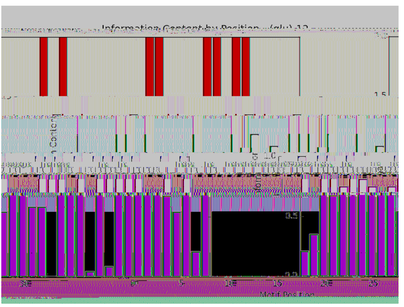


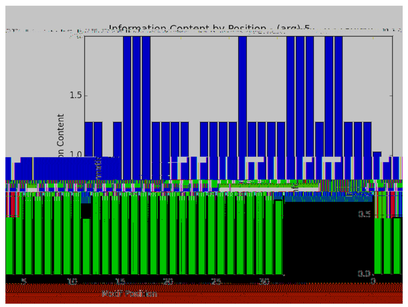


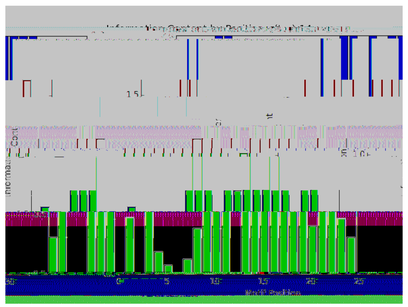


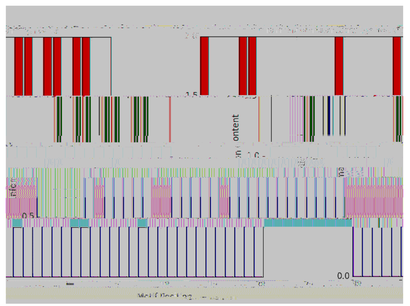


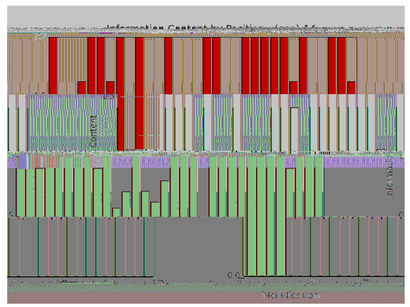


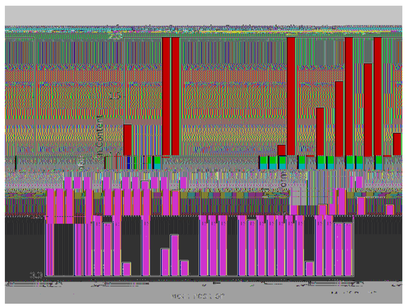


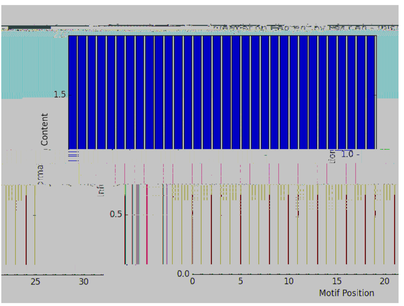




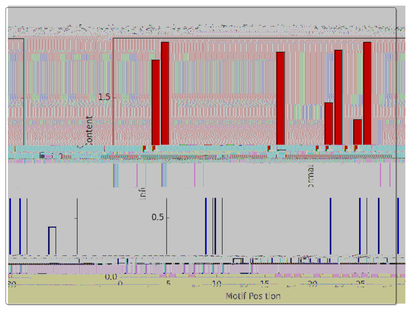








REFERENCE



Information content by position for all 19 characterized amino acid A-domains, and a reference information content for all sequences (length 32), regardless of A-domain specificity.

Table 1. Prediction of AA for a 34-mer sequence randomly selected from each known AA sequence set. Match indicates prediction success.

34-mer sequence	Known AA	Predicted AA	Full Match?	34-mer Match?

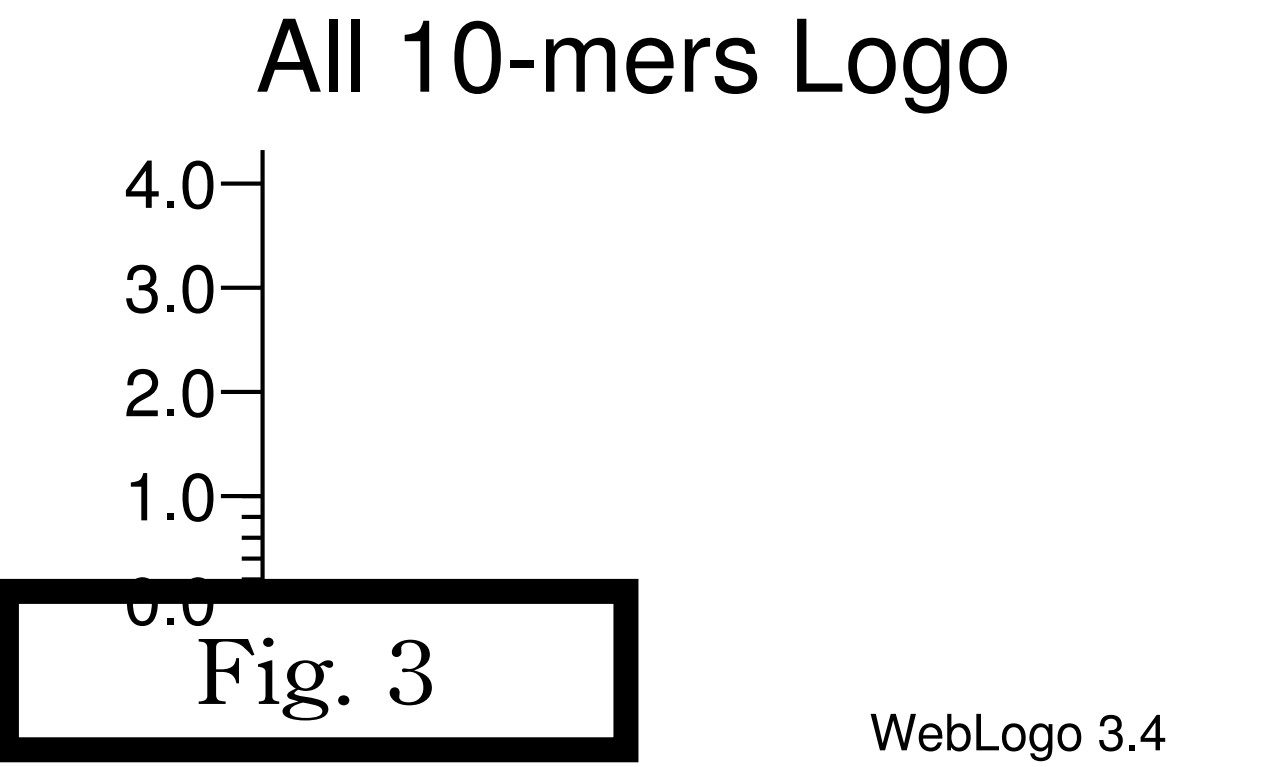


Fig. 3

Figure 4

All 32-mers Logo