Classification of Putative Regulatory Enhancers from DNA Sequences

Inbar Naor and **Tommy Kaplan**

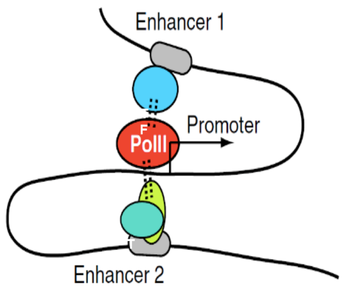
School of Computer Science and Engineering, The Hebrew University of Jerusalem

THE HEBREW UNIVERSITY OF JERUSALEM

i-CORE Israeli Centers of Research Excellence

1 Enhancers

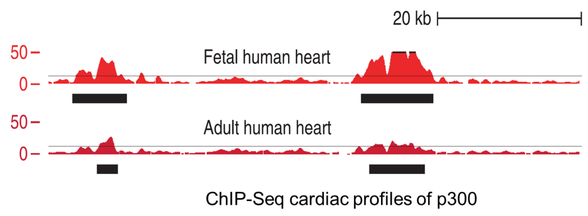
- Regulatory DNA sequences that control the transcriptional activity of their target genes.
- Involved in many developmental and disease-relevant processes.
- Hard to identify due to variety of lengths and distances from affected genes.



Goal: develop computational model for detecting novel enhancers and decipher their functional mechanisms.

2 Data Set

- N = 8,789 putative enhancers from the mouse genome.
- **Distal** (>15Kb of genes), with **H3K27ac** and bound by the co-activator **p300**.
- An equal number of negative (non-enhancer) sequences.
- 500 middle nucleotides were selected from each putative enhancer.



ChIP-Seq cardiac profiles of p300

3 Methods

Raw Data

DNA:
ACGGTCATTACG

H3K27ac and p300 markers

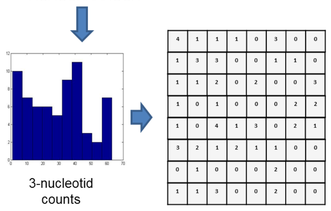
500 middle bp from each sequence

Feature Extraction

K-mer distribution

Example: K=3

AAAACCTGCGGAGGGCTTGAGCGATGGC
AGGTACCAGGTCCTGGAATCCACCTTGG
AGAGCAAAAGCA



3-nucleotid counts

Data Split

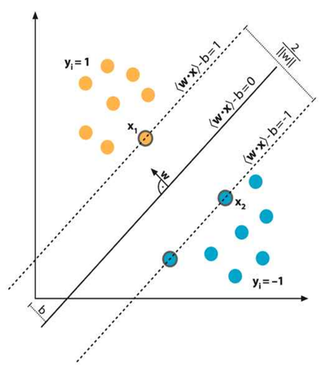
Train 70% / Test 30%

Classifiers

- Support Vector Machines
- K Nearest Neighbors
- Decision Tree
- Random Forest

K-mer distribution: counting the number of occurrences of each k-mer in the sequence, for different k values.

Support Vector Machines



- Support Vector Machines finds the best separating hyperplanes between different classes.
- Classification rule:

Where *w* is a weights vector orthogonal to the separating hyperplane and *b* is a bias term.

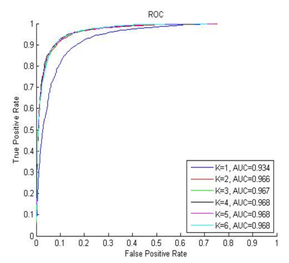
- Kernel functions map the points into (possibly higher dimensional) feature space where linear separation is easier.

4 Results

Linear SVM and k-mer counts for k=6 performance:

False positive rate: 0.09 False negative rate: 0.08

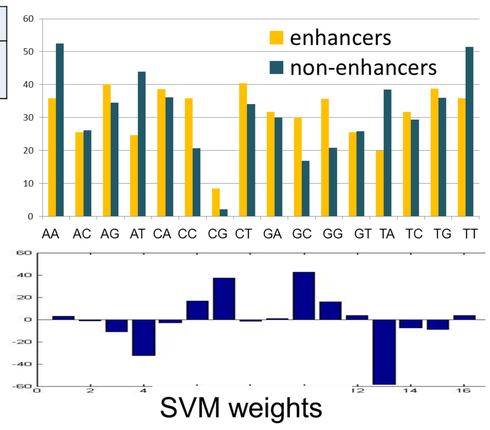
accuracy = 91.2% AUC = 0.97



Linear SVM for various K

- Good results can be obtained using di-nucleotide alone:

	K=1	K=2	K=3	K=4	K=5	K=6
accuracy	86.91	90.73	91.1	91.34	91.11	91.29



- Dinucleotide distribution is different between enhancer and non-enhancer sequences.
- Plotting SVM weights allows detection of important motifs.

SVM weights

5 Features

Beyond K-mer counts:

- K-mer counts concatenation for different k's.

K=1	K=2	K=3	K=4	K=5	K=6

- Motifs proximity – extracting features from 100bp windows.
- Structural properties: Minor Groove Width, Melting Temperature.
- Feature Selection: only differential motifs.

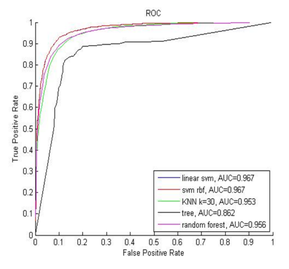
No significant improvement in the results

6 Different Classifiers

We compared performances of different classifiers on Bag of Words representation with k=3.

Linear SVM	SVM rbf	KNN k=30	Tree	Random Forest
91.12	91.12	88.85	84.04	89.56

Accuracy of different classifiers



8 Future Directions

- Feature Learning and Feature Selection
- Multiclass learning of different tissues
- Identifying tissue-specific features
- Structured Hidden Markov model to allow higher interpretability