A Brief History

• Bootstraping

• Bagging

• Boosting (Schapire 1989)

• Adaboost (Schapire 1995)
What’s So Good About Adaboost

• Improves classification accuracy

• Can be used with many different classifiers

• Commonly used in many areas

• Simple to implement

• Not prone to over-fitting

Ex. “How May I Help You?”
easy to find “rules of thumb” that are “often” correct
• e.g.: “IF ‘card’ occurs in utterance
  THEN predict ‘Calling Card’ ”
• hard to find single highly accurate prediction rule
Bootstrap Estimation

• Repeatedly draw $n$ samples from $D$
• For each set of samples, estimate a statistic
• The bootstrap estimate is the mean of the individual estimates
• Used to estimate a statistic (parameter) and its variance
Bagging - Aggregate Bootstrapping

• For $i = 1 .. M$
  – Draw $n^* < n$ samples from $D$ with replacement
  – Learn classifier $C_i$

• Final classifier is a vote of $C_1 .. C_M$

• Increases classifier stability/reduces variance
Boosting (Schapire 1989)

- Randomly select $n_1 < n$ samples from $D$ without replacement to obtain $D_1$
  - Train weak learner $C_1$

- Select $n_2 < n$ samples from $D$ with half of the samples misclassified by $C_1$ to obtain $\tilde{D}_2$
  - Train weak learner $C_2$

- Select all samples from $D$ that $C_1$ and $C_2$ disagree on
  - Train weak learner $C_3$

- Final classifier is vote of weak learners
Adaboost - Adaptive Boosting

• Instead of sampling, re-weight
  – Previous weak learner has only 50% accuracy over new distribution

• Can be used to learn weak classifiers

• Final classification based on weighted vote of weak classifiers
Adaboost Terms

• Learner = Hypothesis = Classifier

• Weak Learner: < 50% error over any distribution

• Strong Classifier: thresholded linear combination of weak learner outputs
Basic idea of boosting technique

In general Boosting Bagging Single Tree. “AdaBoost · · · best off-the-shelf classifier in the world” — Leo Breiman
AdaBoost (Freud & Schapire, 1996)

1. Initialize the observation weights
   \[ w_i = 1/N, \; i = 1, 2, \ldots, N. \]

2. For \( m = 1 \) to \( M \) repeat steps (a)–(d):
   (a) Fit a classifier \( G_m(x) \) to the training data using weights \( w_i \).
   (b) Compute
   \[ err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}. \]
   (c) Compute \( \alpha_m = \log((1 - err_m)/err_m) \).
   (d) Update weights for \( i = 1, \ldots, N \):
   \[ w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))] \]
   and renormalize to \( w_i \) to sum to 1.

3. Output \( G(x) = \text{sign} \left[ \sum_{m=1}^{M} \alpha_m G_m(x) \right] \).
weak classifiers = vertical or horizontal half-planes
Round 1

\[ h_1 \]

\[ \varepsilon_1 = 0.30 \]
\[ \alpha_1 = 0.42 \]

\[ D_2 \]
Round 2

\[ \varepsilon_2 = 0.21 \]
\[ \alpha_2 = 0.65 \]
Round 3

\[ \varepsilon_3 = 0.14 \]
\[ \alpha_3 = 0.92 \]
Final Classifier

\[ H_{\text{final}} = \text{sign}(0.42 + 0.65 + 0.92) \]
Training and testing errors

Nested spheres in $\mathbb{R}^{10} \quad -$ Bayes error is 0%.

Deterministic decision boundary

Stumps

Nested Gaussians in $\mathbb{R}^{10} \quad -$ Bayes error is 25%.

Noisy boundary

Stumps
Over-fitting resistance

Boosting fits additive logistic models, where each component (base learner) is simple. The complexity needed for the base learner depends on the target function.

\[ f(x) = \log \frac{\Pr(Y = 1|x)}{\Pr(Y = -1|x)} = \sum_{m=1}^{M} \alpha_m G_m(x) \]

by stagewise fitting using the loss function

\[ L(y, f(x)) = \exp(-y f(x)). \]

Boosting generalizes because it pushes the training margins well above zero, while keeping the VC dimension under control (also Vapnik, 1996). With \( Pr \geq (1 - \delta) \)

\[ P_{Test}(M(X) \leq 0) \leq P_{Train}(M(X) \leq \theta) \]

\[ + O \left( \frac{1}{\sqrt{N}} \left( \frac{\log N \log |\mathcal{H}|}{\theta^2 + \log 1/\delta} \right)^{1/2} \right) \]
Sequence feature + chromatin feature?

- **DNA sequence**
  - TTCTATAAAGCCGG
- **TFBS motif**
- **DNA energy and flexibility properties**
  - \ldots
- **Histone**
  - Protein Complex
  - TF
  - Pol II
  - RNA
  - TSS
- **Histone modification**
  - GC
  - TATA
  - TF
  - SP1
  - TBP
  - TF

**Proximal promoter**
- TF
- TF
- TF
- TF

**Distal promoter**
- TF
- TF
- TF
- TF

**Histone**
- Histone
Histone modification markers show characteristic pattern around gene promoter.
Code the histone feature

- Dot product: \( X \cdot Y = \sum_{i=1}^{n} x_i y_i \)

- Correlation \( \rho = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \)
Data set

• Histone modification data from CD4+ T-cell:
  – 20 different histone methyllations
  – one histone variant H2A.Z (Barski A, et al., Cell, 2007)
  – 18 different histone acetylations (Wang ZB, et al., NG, 2008)

• Gene annotation
  – ~12,000 Refseq genes with known expression information in CD4+ T Cell
  – Use EPD and DBTSS database annotation of the transcription start sites.

• Training and test set
  – Training set: 4533 CpG related and 1774 non-CpG related prompters respectively. Without alternative TSSs within 2kb.
  – Test set: 1642 non-overlap gene promoters, each containing one or multiple TSS, 2619 independent core-promoters according to EPD and DBTSS annotation.
Boosting with stumps

- Boosting is a supervised machine learning algorithm combing many weak classifiers to create a single strong classifier.

- Denote the training data as \((x_1, y_1), \ldots, (x_N, y_N)\) where \(x_i\) is the feature vector and \(y_i\) is the class label \{\(-1,1\)\}. We define \(f_m(x)\) as the \(m\)-th weak binary classifier producing value of +1 or -1, and as the ensemble of a series \((M)\) weak classifiers, where \(c_m\) are constants and \(M\) is determined by cross validation.

\[
F(x) = \sum_{m=1}^{M} c_m f_m(x)
\]

Stump

\(-1\)

\(1\)
(a) Initialize weight \( w_i^{(0)} = \frac{1}{n}, F^{(0)}(x_i) = 0 \), and probability \( p^{(0)}(x_i) = \frac{1}{2} \), \( i = 1, \ldots, n \). \( p(x) \) is the probability of \( y^* = 1 \).

(b) For \( m = 1, \ldots, M \),

(b.1) Compute the working response and weight for all \( i = 1, \ldots, n \),

\[
\begin{align*}
w_i^{(m)} &= p^{(m-1)}(x_i)(1 - p^{(m-1)}(x_i)), \\
z_i^{(m)} &= \frac{y_i^* - p^{(m-1)}(x_i)}{w_i^{(m)}}.
\end{align*}
\]

(b.2) Fit a regression tree \( f^{(m)}(x) \) minimizing the weighted least squares

\[
\sum_{i=1}^{n} w_i^{(m)} (z_i^{(m)} - f(x_i))^2.
\]

(b.3) Update for \( i = 1, \ldots, n \),

\[
\begin{align*}
F^{(m)}(x_i) &= F^{(m-1)}(x_i) + \frac{1}{2} f^{(m)}(x_i), \\
p^{(m)}(x_i) &= \frac{\exp(F^{(m)}(x_i))}{\exp(F^{(m)}(x_i)) + \exp(-F^{(m)}(x_i))}.
\end{align*}
\]

(c) Let \( F(x) = F^{(M)}(x) = \sum_{m=1}^{M} f_m(x) \). Output classifier \( \text{sign}(F(x)) \) and class probability \( p^M(x_i) \).

**Figure 5**

LogitBoost algorithm with trees.
Classification tree

CoreBoost/CoreBoostHM

Error Rate: 0.073
Performance evaluation

Maximal prediction scores to the annotated TSS

\[
Sensitivity = \frac{TP}{TP + FN},
\]

\[
PPV = \frac{TP}{TP + FP},
\]

\[
F = \frac{2}{(1/Sensitivity)+(1/PPV)} = \frac{2\times Sensitivity\times PPV}{Sensitivity + PPV},
\]

TP: true positives, TN: true negatives, FP: false positives, and FN: false negatives. F-score is the harmonic average of sensitivity and PPV.
Use histone features to predict gene core-promoter

Wang XW et al., Genome Res., 2009
How many histone markers do we need for promoter predictions?

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of Histone markers&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Maximum allowed distance from true TSSs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>50 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sen</td>
</tr>
<tr>
<td>Top one</td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>Top three</td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>Top ten</td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>Non-CpG</td>
<td>All 39 Markers</td>
<td>0.13</td>
</tr>
<tr>
<td>Top one</td>
<td></td>
<td>0.12</td>
</tr>
<tr>
<td>Top three</td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>Top ten</td>
<td></td>
<td>0.26</td>
</tr>
<tr>
<td>CpG</td>
<td>All 39 Markers</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Top Histone features contribute to promoter prediction

### Table 1. Top histone markers contribute to promoter prediction

<table>
<thead>
<tr>
<th>Methylation</th>
<th>Acetylation</th>
<th>All histone markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>CpG&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CpG&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CpG&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>H3K18ac</td>
<td>H3K4me3</td>
</tr>
<tr>
<td>H3K4me1</td>
<td>H2AK5ac</td>
<td>H3K4me1</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>H4K91ac</td>
<td>H3K4me2</td>
</tr>
<tr>
<td>H3K79me3</td>
<td>H3K18ac</td>
<td>H3K4me3</td>
</tr>
<tr>
<td>H3K79me1</td>
<td>H3K18ac</td>
<td>H3K79me3</td>
</tr>
<tr>
<td>H4R3me2</td>
<td>H4K16ac</td>
<td>H3K79me1</td>
</tr>
<tr>
<td>H4K20me1</td>
<td>H4K91ac</td>
<td>H2AK5ac</td>
</tr>
<tr>
<td>H3K27me1</td>
<td>H2BK5ac</td>
<td>H4R3me2</td>
</tr>
<tr>
<td>H3K9me2</td>
<td>H3K9ac</td>
<td>H2A.Z</td>
</tr>
<tr>
<td>H4K20me3</td>
<td>H2BK12ac</td>
<td>H4K91ac</td>
</tr>
<tr>
<td>H3K36me1</td>
<td>H2AK9ac</td>
<td>H3K18ac</td>
</tr>
<tr>
<td>H3K9me1</td>
<td>H3K36ac</td>
<td>H3K23ac</td>
</tr>
</tbody>
</table>

These markers were sorted according to the order they were selected by the boosting classifier.

<sup>a</sup>For CpG-related and non-CpG-related promoters, respectively.
CoreBoost_HM: boosting with stumps for predicting core-promoter using histone modification signal

Features used in CoreBoost_HM:

- Sequence
  TTCTATAAAGCCGG

- TFBS motif
  AATGAGTCA

- DNA energy and flexibility properties
  ...+

- Histone modification signal
  ...+

Promoter structure

DNA sequence

Proximal promoter

Distal promoter

Protein Complex

Pol II

RNA

Cap

Proximal promoter

Histone

Histone modification

Histone modification signal
Comparison with other programs

- CoreBoost (Zhao XY, et al. Genome Biology, 2007)
- McPromoter (Ohler, et al. Bioinformatics, 2001)

Table 1. The performance of all current promoter prediction programs capable of analyzing the whole human genome

<table>
<thead>
<tr>
<th>Program</th>
<th>Reference</th>
<th>Recall</th>
<th>Prec.</th>
<th>F</th>
<th>Recall</th>
<th>Prec.</th>
<th>F</th>
<th>Recall</th>
<th>Prec.</th>
<th>F</th>
<th>Recall</th>
<th>Prec.</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP3(Base stacking)</td>
<td>Bajic et al. (2003)</td>
<td>0.42</td>
<td>0.46</td>
<td>0.44</td>
<td>0.46</td>
<td>0.56</td>
<td>0.51</td>
<td>0.49</td>
<td>0.64</td>
<td>0.56</td>
<td>0.34</td>
<td>0.66</td>
<td>0.45</td>
</tr>
<tr>
<td>DragonGF</td>
<td>Scherf et al. (2000)</td>
<td>0.45</td>
<td>0.63</td>
<td>0.53</td>
<td>0.50</td>
<td>0.74</td>
<td>0.60</td>
<td>0.54</td>
<td>0.80</td>
<td>0.64</td>
<td>0.31</td>
<td>0.75</td>
<td>0.44</td>
</tr>
<tr>
<td>PromoterInspector</td>
<td>Davuluri et al. (2001)</td>
<td>0.38</td>
<td>0.70</td>
<td>0.49</td>
<td>0.42</td>
<td>0.78</td>
<td>0.55</td>
<td>0.44</td>
<td>0.82</td>
<td>0.57</td>
<td>0.29</td>
<td>0.81</td>
<td>0.43</td>
</tr>
<tr>
<td>FirstEF</td>
<td>Down and Hubbard (2002)</td>
<td>0.36</td>
<td>0.51</td>
<td>0.42</td>
<td>0.38</td>
<td>0.65</td>
<td>0.48</td>
<td>0.40</td>
<td>0.74</td>
<td>0.52</td>
<td>0.28</td>
<td>0.75</td>
<td>0.41</td>
</tr>
<tr>
<td>N-Scan</td>
<td>Gross and Brent (2006)</td>
<td>0.55</td>
<td>0.51</td>
<td>0.53</td>
<td>0.60</td>
<td>0.55</td>
<td>0.57</td>
<td>0.63</td>
<td>0.58</td>
<td>0.60</td>
<td>0.33</td>
<td>0.45</td>
<td>0.38</td>
</tr>
<tr>
<td>CpgProD</td>
<td>Ponger and Mouchiroud (2002)</td>
<td>0.50</td>
<td>0.36</td>
<td>0.42</td>
<td>0.59</td>
<td>0.42</td>
<td>0.49</td>
<td>0.65</td>
<td>0.46</td>
<td>0.54</td>
<td>0.34</td>
<td>0.41</td>
<td>0.37</td>
</tr>
<tr>
<td>PromoterExplorer</td>
<td>Xie et al. (2006)</td>
<td>0.55</td>
<td>0.24</td>
<td>0.33</td>
<td>0.66</td>
<td>0.28</td>
<td>0.39</td>
<td>0.73</td>
<td>0.32</td>
<td>0.44</td>
<td>0.39</td>
<td>0.30</td>
<td>0.34</td>
</tr>
<tr>
<td>McPromoter (0.0)</td>
<td>Ohler et al. (2000)</td>
<td>0.24</td>
<td>0.61</td>
<td>0.34</td>
<td>0.29</td>
<td>0.71</td>
<td>0.41</td>
<td>0.32</td>
<td>0.77</td>
<td>0.45</td>
<td>0.20</td>
<td>0.68</td>
<td>0.28</td>
</tr>
<tr>
<td>PromFD</td>
<td>Chen et al. (1997)</td>
<td>0.55</td>
<td>0.14</td>
<td>0.22</td>
<td>0.61</td>
<td>0.16</td>
<td>0.25</td>
<td>0.68</td>
<td>0.18</td>
<td>0.28</td>
<td>0.44</td>
<td>0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>DragonPF</td>
<td>Bajic et al. (2002)</td>
<td>0.65</td>
<td>0.11</td>
<td>0.19</td>
<td>0.71</td>
<td>0.13</td>
<td>0.22</td>
<td>0.79</td>
<td>0.16</td>
<td>0.27</td>
<td>0.51</td>
<td>0.11</td>
<td>0.18</td>
</tr>
<tr>
<td>ARTS</td>
<td>Sonnenburg et al. (2006)</td>
<td>0.77</td>
<td>0.08</td>
<td>0.14</td>
<td>0.82</td>
<td>0.09</td>
<td>0.16</td>
<td>0.88</td>
<td>0.10</td>
<td>0.18</td>
<td>0.63</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td>PromoterScan</td>
<td>Pretridge (1995)</td>
<td>0.19</td>
<td>0.08</td>
<td>0.11</td>
<td>0.27</td>
<td>0.11</td>
<td>0.16</td>
<td>0.36</td>
<td>0.15</td>
<td>0.21</td>
<td>0.16</td>
<td>0.09</td>
<td>0.12</td>
</tr>
<tr>
<td>Promoter2.0 (medium)</td>
<td>Knudsen (1999)</td>
<td>0.68</td>
<td>0.03</td>
<td>0.06</td>
<td>0.91</td>
<td>0.04</td>
<td>0.08</td>
<td>0.99</td>
<td>0.04</td>
<td>0.08</td>
<td>0.63</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>NNPP 2.2(0.99)</td>
<td>Reese (2001)</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.06</td>
<td>0.04</td>
<td>0.05</td>
<td>0.10</td>
<td>0.08</td>
<td>0.09</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
</tbody>
</table>

The F-measure of the CAGE dataset with strictest maximum allowed mismatch was used to rank the programs (in bold). The FProm program (Solovyev et al. 2006) was omitted from the analyses because it is not for free for academic use. Prec., precision.
Prediction Performance

Ten-fold cross validation on training set
Comparison with the other state-of-the-art algorithms

CoreBoost_HM provides significantly higher sensitivity and specificity at high resolution, and can be used to identify both active and repressed promoters.
Performance on active or silent genes

<table>
<thead>
<tr>
<th>Promoter Type</th>
<th>CpG related</th>
<th>Non-CpG related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related gene expression level</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>High</td>
<td>H3K4me3</td>
<td>H3K4me3</td>
</tr>
<tr>
<td>Low</td>
<td>H3K4me2</td>
<td>H3K79me3</td>
</tr>
<tr>
<td>High</td>
<td>H3K4me1</td>
<td>H2A.Z</td>
</tr>
<tr>
<td>Low</td>
<td>H3K79me3</td>
<td>H4R3me2</td>
</tr>
</tbody>
</table>

Top Histone Markers

- H3K4me3
- H3K4me2
- H3K79me3
- H2A.Z
- H3K27ac
- H2BK120ac
- H3K27me3
- H3K79me3
- H4K20me3
- H3K4me2
- H4K16ac
- H3K36me3
- H3K4R3me2
- H3K9me2
- H4K5ac
Performance on test set

- 1642 nonoverlap gene promoters, each containing one or multiple TSS, 2619 independent core-promoters according to EPD and DBTSS annotation.
- 10 kb region from 5 kb to 5 kb relative to the refseq gene 5’-end

<table>
<thead>
<tr>
<th>Program</th>
<th>50 bp</th>
<th></th>
<th></th>
<th>200 bp</th>
<th></th>
<th></th>
<th>500 bp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>PPV</td>
<td>F</td>
<td>Sensitivity</td>
<td>PPV</td>
<td>F</td>
<td>Sensitivity</td>
<td>PPV</td>
<td>F</td>
</tr>
<tr>
<td>CoreBoost_HM-HM</td>
<td>0.26</td>
<td>0.43</td>
<td>0.32</td>
<td>0.45</td>
<td>0.56</td>
<td>0.50</td>
<td>0.57</td>
<td>0.72</td>
<td>0.64</td>
</tr>
<tr>
<td>CoreBoost_HM-high*</td>
<td>0.34</td>
<td>0.37</td>
<td>0.36</td>
<td>0.50</td>
<td>0.54</td>
<td>0.52</td>
<td>0.66</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td>CoreBoost_HM-low^</td>
<td>0.26</td>
<td>0.29</td>
<td>0.27</td>
<td>0.42</td>
<td>0.48</td>
<td>0.45</td>
<td>0.59</td>
<td>0.62</td>
<td>0.60</td>
</tr>
<tr>
<td>BHMF</td>
<td>0.20</td>
<td>0.28</td>
<td>0.23</td>
<td>0.38</td>
<td>0.53</td>
<td>0.44</td>
<td>0.52</td>
<td>0.72</td>
<td>0.60</td>
</tr>
<tr>
<td>CoreBoost</td>
<td>0.19</td>
<td>0.20</td>
<td>0.20</td>
<td>0.38</td>
<td>0.39</td>
<td>0.38</td>
<td>0.55</td>
<td>0.56</td>
<td>0.55</td>
</tr>
<tr>
<td>McPromoter</td>
<td>0.17</td>
<td>0.19</td>
<td>0.18</td>
<td>0.43</td>
<td>0.32</td>
<td>0.37</td>
<td>0.67</td>
<td>0.54</td>
<td>0.60</td>
</tr>
<tr>
<td>EP3</td>
<td>0.12</td>
<td>0.11</td>
<td>0.11</td>
<td>0.40</td>
<td>0.37</td>
<td>0.38</td>
<td>0.62</td>
<td>0.61</td>
<td>0.62</td>
</tr>
</tbody>
</table>

*a*CoreBoost_HM performance on the promoters of highly expressed genes.
*b*CoreBoost_HM performance on the promoters of low expressed genes.

PPV, positive predictive value; F, harmonic average of sensitivity and PPV.
Top features used in CoreBoost_HM

<table>
<thead>
<tr>
<th>CpG related promoters</th>
<th>non-CpG related promoters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-likelihood ratios from third order Markov chain</td>
<td>DNA sequence energy profile</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>H3K4me3</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>Log-likelihood ratios from third order Markov chain</td>
</tr>
<tr>
<td>GC-box score</td>
<td>TATA box score using weight matrix model</td>
</tr>
<tr>
<td>Difference between average energy score and TSS energy score</td>
<td>Inr score using consensus model</td>
</tr>
<tr>
<td>H4K20me1</td>
<td>TATA box score using consensus model</td>
</tr>
<tr>
<td>weighted score of transcription factor NFY, KROX, MAZ</td>
<td>H2AZ</td>
</tr>
<tr>
<td>H3K4me1</td>
<td>weighted score of transcription factor ELK1, MAZ</td>
</tr>
<tr>
<td>H2AZ</td>
<td>weighted score of transcription factor NFY</td>
</tr>
<tr>
<td>weighted score of transcription factor ELK1</td>
<td>GCbox score using weight matrix model</td>
</tr>
<tr>
<td>H3K79me3</td>
<td>Combined score of TATA box, Inr and background segment</td>
</tr>
<tr>
<td>H4K5ac</td>
<td>weighted score of transcription factor HNF1</td>
</tr>
<tr>
<td>H4K91ac</td>
<td>Inr score using weight matrix model</td>
</tr>
<tr>
<td>Inr score using weight matrix model</td>
<td>weighted score of transcription factor CREBATF</td>
</tr>
<tr>
<td>Difference between average energy score and TSS energy score</td>
<td>H3K79me3</td>
</tr>
</tbody>
</table>
Predict microRNA promoters

Jeffrey SS, Nat. Biotechnol., 2008

A microRNA component of the p53 tumour suppressor network

Lin Ho¹, Xingyue Hu², Lee P. Lim³, Elisa de Stanchina¹, Zhenyu Xuan¹, Yu Liang¹, Wen Xue¹, Lars Zender¹, Jill Magnus¹, Dana Ridzon⁴, Aimée L. Jackson⁵, Peter S. Linsley⁵, Cailiu Chen⁶, Scott W. Lowe⁷, Michele A. Cleary⁶ & Gregory J. Hannon¹

He et al., Nature, 2007
CoreBoost_HM can accurately predict human microRNA promoters

Intronic miRNA has its own promoter!

Wang XW et al., Genome Res., 2009