Discriminative de novo motif discovery from high-throughput data

Jan Grau¹, Stefan Posch¹, Ivo Grosse¹, and Jens Keilwagen²

¹Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany
²Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany

CSC ChIP- and DNase-seq data analysis workshop
Biological question

- reasons for phenotypic observations
- regulation of gene expression
- first step: transcriptional regulation

⇒ transcription factor binding sites

De-novo motif discovery without knowledge of

- motif
- exact location of sites from set of input sequences

[Based on Robert Tjian, “Molecular Machines that Control Genes”]

Motif discovery with Dimont

Jan Grau et al.
Experimental techniques - ChIP-seq

Data

- ChIP-seq peaks: approximate binding regions
  ⇒ extract sequences under peaks
- ChIP-seq peak statistics: information about TF abundance at binding region

[Szalkowski & Schmid, Brief Bioinform, 2010]
Experimental techniques - PBM

Data

- PBM probes: contain all possible DNA 10-mers
  ⇒ probe sequences (length 35 bp + linker)
- Probe intensities: information about TF binding frequency

[Geertz & Maerkl, Brief Func Genom, 2010]
Requirements for a novel approach

**Use all sequences**
thresholding to extract top peaks/probes arbitrary
⇒ use all peaks and probe sequences, respectively

**Use all information** present in the data
  
  ChIP-seq sequence under peak
  peak statistics
  binding more likely around peak center

  PBM probe sequence (including part of linker)
  probe intensities

**Use discriminative learning principle**
which often yield better results than generative principles

**Allow for flexible choice of motif models**
e.g., position weight matrices, weight array matrices, ...

**Retain acceptable runtime**
below 1h for majority of data sets
Weighting schema for integrating ChIP and PBM data allows for using ChIP peak statistics and PBM probe intensities in a common approach

\[
w_{fg}^n := \frac{1}{1 + \frac{h_n}{1-h_n} \cdot \frac{1-q}{q}}, \quad w_{bg}^n := 1 - w_{fg}^n
\]

\(h_n\): relative rank of sequence \(x_n\) based on peak statistic or probe intensity, \(q\): weighting factor, i.e., a-priori fraction of foreground sequences
A-priori position distribution represents that binding occurs close to peak center.

**ChIP-seq**

**PBM**
Discriminative learning - Motivation

ChIP-seq positives

over-represented
Discriminative learning - Motivation

ChIP-seq positives

ChIP-seq negatives

over-represented
Discriminative learning - Motivation

ChIP-seq positives

ChIP-seq negatives

over-represented
differentially abundant

⇒ discriminative learning
Discriminative learning - Objective function

Discriminative weighted maximum supervised posterior principle

\[
\hat{\lambda} = \arg\max_{\lambda} \sum_{n=1}^{N} \sum_{c \in C} w_n^c \log \left( \frac{P(c|\lambda)P_c(x_n|\lambda)}{\sum_{\tilde{c} \in C} P(\tilde{c}|\lambda)P_{\tilde{c}}(x_n|\lambda)} \right) + Q(\lambda|\alpha),
\]

where \( C = \{fg, bg\} \): set of classes,

\( Q(\lambda|\alpha) \): prior on the parameters \( \lambda \) given hyper-parameters \( \alpha \),

\( P(c|\lambda) \): a-priori class probability, and

\( P_c(x_n|\lambda) \): class-conditional likelihood, “model”
\[ P_{fg}(x|\lambda) = P(\text{motif}|\lambda) \cdot \frac{1}{|\Sigma|^{L-w}} \sum_{\ell \in \mathcal{L}} P(\ell)P_{\text{motif}}(x_\ell, \ldots, x_{\ell+w-1}|\lambda) \]

\[ + (1 - P(\text{motif}|\lambda)) \cdot \frac{1}{|\Sigma|^L} \]

- Dimont uses standard ZOOPS model \((P_{fg}(x|\lambda))\)
- sequence flanking the motif: uniform, i.e., all nucleotides with equal probability
- motif model: strand model enclosing
  - position weight matrix (PWM): assumes nucleotide independence or
  - weight array matrix (WAM): allows dependencies between neighboring nucleotides or
  - higher-order Markov models
- background model: uniform or Markov model \((P_{bg}(x|\lambda))\)
Speed-up strategies

Idea:
- pre-optimization on reduced data set
- evaluation of only highest-scoring motif occurrences
Speed-up strategies

Idea:
- pre-optimization on reduced data set
- evaluation of only highest-scoring motif occurrences
Benchmark on PBM data

66 PBM data sets (Weirauch et al.)
- protein binding microarray data for 66 TFs
- two different array designs (HK/ME) with different probes

Task:
Learn motif on one design, predict binding intensities for other design

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Pearson corr.</th>
<th>AUC-ROC</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimont</td>
<td>0.695</td>
<td>0.951</td>
<td>1.002</td>
</tr>
<tr>
<td>FeatureREDUCE</td>
<td>0.693</td>
<td>0.949</td>
<td>0.997</td>
</tr>
<tr>
<td>Team_D</td>
<td>0.691</td>
<td>0.938</td>
<td>0.984</td>
</tr>
<tr>
<td>Team_E</td>
<td><strong>0.696</strong></td>
<td>0.906</td>
<td>0.952</td>
</tr>
</tbody>
</table>
Benchmark on ChIP-seq data

26 ChIP-seq data sets (Ma et al.)
- 26 ChIP-seq data sets for TFs with known motifs
- human, mouse, fly

Task:
Discover motif consistent with literature

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Total successes</th>
<th>Average rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimont</td>
<td>26</td>
<td>1.23</td>
</tr>
<tr>
<td>POSMO</td>
<td>23</td>
<td>1.00</td>
</tr>
<tr>
<td>ChIPMunk</td>
<td>23</td>
<td>1.00</td>
</tr>
<tr>
<td>MEME</td>
<td>22</td>
<td>1.32</td>
</tr>
<tr>
<td>DME</td>
<td>22</td>
<td>1.45</td>
</tr>
<tr>
<td>DREME</td>
<td>22</td>
<td>1.45</td>
</tr>
<tr>
<td>HMS</td>
<td>12</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Example motifs

⇒ most motifs fit the literature well
In-vivo vs in-vitro binding

<table>
<thead>
<tr>
<th>Motif</th>
<th>PBM</th>
<th>ChIP-seq</th>
<th>PBM</th>
<th>ChIP-seq/exo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esrrb</td>
<td><img src="image" alt="Esrrb PBM" /></td>
<td><img src="image" alt="Esrrb ChIP-seq" /></td>
<td><img src="image" alt="Esrrb PBM" /></td>
<td><img src="image" alt="Esrrb ChIP-seq/exo" /></td>
</tr>
<tr>
<td>Foxo1</td>
<td><img src="image" alt="Foxo1 PBM" /></td>
<td><img src="image" alt="Foxo1 ChIP-seq" /></td>
<td><img src="image" alt="Foxo1 PBM" /></td>
<td><img src="image" alt="Foxo1 ChIP-seq/exo" /></td>
</tr>
<tr>
<td>Gata4</td>
<td><img src="image" alt="Gata4 PBM" /></td>
<td><img src="image" alt="Gata4 ChIP-seq" /></td>
<td><img src="image" alt="Gata4 PBM" /></td>
<td><img src="image" alt="Gata4 ChIP-seq/exo" /></td>
</tr>
<tr>
<td>Zfx</td>
<td><img src="image" alt="Zfx PBM" /></td>
<td><img src="image" alt="Zfx ChIP-seq" /></td>
<td><img src="image" alt="Zfx PBM" /></td>
<td><img src="image" alt="Zfx ChIP-seq/exo" /></td>
</tr>
<tr>
<td>Tbx5</td>
<td><img src="image" alt="Tbx5 PBM" /></td>
<td><img src="image" alt="Tbx5 ChIP-seq" /></td>
<td><img src="image" alt="Tbx5 PBM" /></td>
<td><img src="image" alt="Tbx5 ChIP-seq/exo" /></td>
</tr>
<tr>
<td>Nr5a2</td>
<td><img src="image" alt="Nr5a2 PBM" /></td>
<td><img src="image" alt="Nr5a2 ChIP-seq" /></td>
<td><img src="image" alt="Nr5a2 PBM" /></td>
<td><img src="image" alt="Nr5a2 ChIP-seq/exo" /></td>
</tr>
<tr>
<td>Phd1</td>
<td><img src="image" alt="Phd1 PBM" /></td>
<td><img src="image" alt="Phd1 ChIP-seq" /></td>
<td><img src="image" alt="Phd1 PBM" /></td>
<td><img src="image" alt="Phd1 ChIP-seq/exo" /></td>
</tr>
<tr>
<td>Rap1</td>
<td><img src="image" alt="Rap1 PBM" /></td>
<td><img src="image" alt="Rap1 ChIP-seq" /></td>
<td><img src="image" alt="Rap1 PBM" /></td>
<td><img src="image" alt="Rap1 ChIP-seq/exo" /></td>
</tr>
<tr>
<td>Tcf3</td>
<td><img src="image" alt="Tcf3 PBM" /></td>
<td><img src="image" alt="Tcf3 ChIP-seq" /></td>
<td><img src="image" alt="Tcf3 PBM" /></td>
<td><img src="image" alt="Tcf3 ChIP-seq/exo" /></td>
</tr>
<tr>
<td>Tbx20</td>
<td><img src="image" alt="Tbx20 PBM" /></td>
<td><img src="image" alt="Tbx20 ChIP-seq" /></td>
<td><img src="image" alt="Tbx20 PBM" /></td>
<td><img src="image" alt="Tbx20 ChIP-seq/exo" /></td>
</tr>
</tbody>
</table>

⇒ good accordance between *in-vivo* and *in-vitro* binding, but notable exceptions
In case of CTCF, the detected dependencies are highly similar for all combinations of data sets. One reason might be that, in contrast to all other cross-technology comparisons considered in this study, the CTCF binding sites are more conserved between position 4 and 5, between position 5 and 6, between position 12 and 13, and between position 13 and 14 for all combinations of training and test data.

In case of Rap1, we find significant dependencies between neighboring positions only in case of the in-vivo comparison (train(ChIP-exo) - test(ChIP-exo)). This is in contrast to the other comparisons, such as train(ChIP-seq) - test(ChIP-seq) and train(ChIP-exo) - test(PBM), where the level of conservation of binding site positions is influenced to a greater degree by the choice of the test data set than by the training data. Interestingly, the level of conservation of binding site positions is influenced by the choice of the training data set, particularly between position 7 and 8, which are consistently detected from ChIP-exo test data. These positions are also among the positions with the greatest MI values.

One reason for this difference might be that, in contrast to all other cross-technology comparisons considered in this study, the Rap1 binding sites are more conserved between position 2 and 3 and between position 6 and 7.
Motif discovery with Dimont

Jan Grau et al.

Dimont is a universal tool for de-novo motif discovery. Dimont has successfully been applied to ChIP-seq, ChIP-exo and protein-binding microarray (PBM) data.

connected to chipster.csc.fi
Galaxy application
- public server
- convenient user interface
- also available in Galaxy Tool-Shed

```
galaxy.informatik.uni-halle.de
```

- Galaxy-Server: 45 registered users, 500 runs (est.)
- Galaxy Tool-Shed: 60 clones
Command line application

- `<key>=<value>` interface
- easily scriptable
- multi-threaded

```java
java -jar Dimont.jar data=myseqs.fa infix=myresult position=peak value=signal threads=8
```

- available from www.jstacs.de/index.php/Dimont
- 290 downloads of command line program
Conclusions

**Dimont, a general approach for motif discovery**
- reliably discovers motifs from ChIP-seq and PBM data
- achieves an acceptable runtime

**In-vitro and in-vivo binding**
- often in good accordance
- but notable exceptions

**Availability**
- Chipster since version 2.11
- public Galaxy at galaxy.informatik.uni-halle.de and Galaxy Tool-Shed
- command line application: www.jstacs.de/index.php/Dimont
Conclusions

**Dimont, a general approach for motif discovery**
- reliably discovers motifs from ChIP-seq and PBM data
- achieves an acceptable runtime

**In-vitro and in-vivo binding**
- often in good accordance
- but notable exceptions

**Availability**
- Chipster since version 2.11
- public Galaxy at galaxy.informatik.uni-halle.de and Galaxy Tool-Shed
- command line application: www.jstacs.de/index.php/Dimont

Thank you for your attention!