Discovering cis-regulatory motifs using genome-wide sequence, expression and protein binding microarray data

Yaron Orenstein, Chaim Linhart, Yonit Halperin, Igor Ulitsky, Ron Shamir
Gene expression regulation

- Transcription is regulated mainly by transcription factors (TFs) - proteins that bind to DNA subsequences, called binding sites (BSs).
- TFBSs are located mainly in the gene’s promoter – the DNA sequence upstream the gene’s transcription start site (TSS).
- TFs can promote or repress transcription.
- Other regulators: micro-RNAs (miRNAs)
Motif discovery: The typical two-step pipeline

Co-regulated gene set

Gene expression microarrays

Clustering

Cluster I

Cluster II

Cluster III

Location analysis (ChIP-chip, ...)

Functional group (e.g., GO term)

Promoter/3’UTR sequences

Motif discovery

The Gene Ontology
Amadeus
A Motif Algorithm for Detecting Enrichment in multiple Species

- Supports diverse motif discovery tasks:
  1. Finding over-represented motifs in one or more given sets of genes.
  2. Identifying motifs with global spatial features given only the genomic sequences.
  3. Simultaneous inference of motifs and their associated expression profiles given genome-wide expression datasets.

- How?
  - A general pipeline architecture for enumerating motifs.
  - Different statistical scoring schemes of motifs for different motif discovery tasks.
Task I: Over-represented motifs in given target set

- **Input:** Target set \((T)\) = co-regulated genes
  Background \((BG)\) set \((B)\) = entire genome

No sequence model is assumed!

- **Motif scoring:**
  Hypergeometric \((HG)\) enrichment score

- \(b, t\) = BG/Target genes containing a hit

BG set should be of the same “nature” as the target set, and much larger
E.g., all genes on microarray
Drawback of the HG score

- **Length/GC-content** distribution in the **target set** might significantly differ from the distribution in the **BG set**
  - Very common in practice due to correlation between the expression/function of genes and the length/GC-content of their promoters and 3’ UTRs
  - The HG score might **fail** to discover the correct motif or detect many **spurious** motifs

→ **Use the binned enrichment score**
  - Slightly less sensitive than HG score...
  - ... but takes into account length/GC-content biases
Test case: Human G2+M cell-cycle genes

Pairs analysis

These motifs form a module associated with G_{2}+M [Elkon et al. ’03, Tabach et al. ’05, Linhart et al. ’05]
Benchmark:
Real-life **metazoan** datasets

We constructed the first motif discovery benchmark that is based on a **large compendium of experimental studies**

**Source:** Various (expression, ChIP-chip, Gene Ontology, ...)

**Data:** 42 target-sets of 26 **TFs** and 8 **miRNAs** from 29 publications

**Species:** human, mouse, fly, worm

**Average set size:**
400 genes (=383 Kbps)
Amadeus
A Motif Algorithm for Detecting Enrichment in mUltiple Species

- Supports diverse motif discovery tasks:
  1. Finding over-represented motifs in one or more given sets of genes.
  2. Identifying motifs with global spatial features given only the genomic sequences.
  3. Simultaneous inference of motifs and their associated expression profiles given genome-wide expression datasets.

- How?
  - A general pipeline architecture for enumerating motifs.
  - Different statistical scoring schemes of motifs for different motif discovery tasks.
Amadeus — Global spatial analysis

Co-regulated gene set

Gene expression microarrays

Location analysis (ChIP-chip,...)

Functional group (e.g., GO term)

the Gene Ontology

Promoter sequences

Output

Motif(s)
Task II: Global analyses

Scores for spatial features of motif occurrences

**Input:** Sequences (no target-set / expression data)

**Motif scoring:**
- Localization w.r.t the TSS
- Strand-bias
- Chromosomal preference
Global analysis I:
Localized human + mouse motifs

Input:
- All human & mouse promoters (2 x ~20,000)
- Score: localization

<table>
<thead>
<tr>
<th>Name</th>
<th>Motif logo</th>
<th>Localization</th>
<th>Strand bias</th>
<th>Chrom pref.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>peak</td>
<td>p-val</td>
<td>p-val</td>
</tr>
<tr>
<td>A. Known TFs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP1</td>
<td>GC_C_C_C_C_C_C_C</td>
<td>-60</td>
<td>10^{-129}</td>
<td>-</td>
</tr>
<tr>
<td>NF-Y</td>
<td><em>C_CA_A_AT</em></td>
<td>-90...-60</td>
<td>10^{-145}</td>
<td>-</td>
</tr>
<tr>
<td>GABP</td>
<td><em>G_G_A_A_A_G_T_G</em></td>
<td>-30...0</td>
<td>10^{-113}</td>
<td>-</td>
</tr>
<tr>
<td>TATA</td>
<td>T_A_T_A_A_A</td>
<td>-30</td>
<td>10^{-61}</td>
<td>10^{-15}</td>
</tr>
<tr>
<td>NRF1</td>
<td><em>T_G_C_G_C_G_T</em></td>
<td>-30</td>
<td>10^{-48}</td>
<td>-</td>
</tr>
<tr>
<td>ATF/CREB</td>
<td><em>A_C_G_T_C_A_C_A_G_A</em></td>
<td>-60</td>
<td>10^{-27}</td>
<td>-</td>
</tr>
<tr>
<td>MYC</td>
<td><em>G_T_C_A_C_C_T_G</em></td>
<td>-60...-30</td>
<td>10^{-27}</td>
<td>-</td>
</tr>
<tr>
<td>RFX1</td>
<td><em>C_T_G_G_A_A_C_G</em></td>
<td>-60</td>
<td>10^{-9}</td>
<td>-</td>
</tr>
<tr>
<td>B. Novel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTACAWYTC</td>
<td>ACTAC_A_A_T_C</td>
<td>-90...-60</td>
<td>10^{-21}</td>
<td>10^{-8}</td>
</tr>
<tr>
<td>CTCGGCGAGAT</td>
<td>C_T_C_G_C_G_A_G_A_T</td>
<td>-60...-30</td>
<td>10^{-7}</td>
<td>-</td>
</tr>
<tr>
<td>C. Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splice donor site</td>
<td><em>G_G_T_S_A_G</em></td>
<td>+30...</td>
<td>10^{-23}</td>
<td>10^{-8}</td>
</tr>
</tbody>
</table>
Amadeus is available at:

“Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets”,
(*equal contribution)

http://acgt.cs.tau.ac.il/amadeus
Amadeus
A Motif Algorithm for Detecting Enrichment in multiple Species

- Supports diverse motif discovery tasks:
  1. Finding **over-represented** motifs in one or more given sets of genes.
  2. Identifying motifs with **global spatial features** given only the genomic **sequences**.
  3. Simultaneous inference of **motifs** and their associated **expression profiles** given genome-wide **expression datasets**.

- **How?**
  - A general **pipeline architecture** for enumerating motifs.
  - Different statistical **scoring schemes** of motifs for different motif discovery tasks.
**PRIMA – GOAL: ‘Reverse engineering’ of transcriptional networks**

- Co-expression $\rightarrow$ Co-regulation $\rightarrow$ common cis-regulatory promoter elements

- Identification of co-expressed genes using microarray technology (clustering)

- Computational identification of cis-regulatory elements that are over-represented in promoters of the co-expressed genes
**PRIMA – General description**

- **PRIMA** identifies transcription factors (TFs) whose binding sites (BSs) are enriched in a given ‘target set’ of promoters with respect to a ‘background set’ of promoters.

- Required ‘databases’:
  - Promoter sequences on a genomic scale
  - ‘Models’ for binding sites recognized by TFs

- Implemented in Expander.
Amadeus - Allegro

Gene expression microarrays

Clustering

Cluster I
Cluster II
Cluster III

Expressed gene set

Promoter sequences

Output

Motif(s)
Allegro: expression model

- **Discretization** of expression values

  \[ e_1 = \text{Up (U)} \geq 1.0 \]
  \[ e_2 = \text{Same (S)} (-1.0, 1.0) \]
  \[ e_3 = \text{Down (D)} \leq -1.0 \]

<table>
<thead>
<tr>
<th>( g )</th>
<th>( c_1 )</th>
<th>( c_2 )</th>
<th>( c_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.3</td>
<td>-0.8</td>
<td>1.5</td>
<td></td>
</tr>
</tbody>
</table>

Expression pattern

<table>
<thead>
<tr>
<th>( g )</th>
<th>D</th>
<th>S</th>
<th>( c_1 )</th>
<th>( c_2 )</th>
<th>( c_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>S</td>
<td>( c_1 )</td>
<td>( c_2 )</td>
<td>( c_m )</td>
<td></td>
</tr>
</tbody>
</table>

Discrete expression Pattern (DEP)

- Expression data should be (partially) **pre-processed**, e.g.:
  - Time series \( \rightarrow \) log ratio relative to time 0
  - Several tissues/mutations/... \( \rightarrow \) standardization
  - Do NOT filter out non-responsive genes

- **Expression model:** **CWM** = Condition Weight Matrix

- Non-parametric, log-likelihood based model, analogous to PWM for sequence motifs

- Sensitive, robust against extreme values, performs well in practice
Allegro overview

Data

Cis-regulatory sequences

Gene expression matrix

Model

PWM (sequence motif)

CWM (expression profile)

Model evaluation & optimization

Score = Enrichment $p$-value
Human cell cycle [Whitfield et al., ’02]

Large dataset: ~15,000 genes, 111 conditions, promoters region: -1000...200 bps

Allegro recovers the major regulators of the human cell cycle [Elkon et al. ’03; Tabach et al. ’05; Linhart et al. ’05].
Allegro can discover *multiple motifs with diverse expression patterns*, even if the response is in a *small fraction of the conditions*

*Extant two-step techniques* recovered only 4 of the above motifs:

- **K-means/CLICK + Amadeus/Weeder**: RRPE, PAC, MBF, STRE
- **Iclust + FIRE**: RRPE, PAC, Rap1, STRE
Amadeus/Allegro - Additional features

- Motif pairs analysis
- Joint analysis of multiple datasets
- Evaluation of motifs using several scores
- Bootstrapping – get fixed p-value
- Sequence redundancy elimination – ignore sequences with long identical subsequences
- User-friendly and informative (most tools are textual and supply limited information!)
Allegro is available at:


(*equal contribution)

http://acgt.cs.tau.ac.il/allegro
Summary

• Developed *Amadeus* motif discovery platform:
  • Broad range of applications:
    - Target gene set
    - Spatial features (sequence only)
    - Expression analysis - *Allegro*
  • Sensitive & efficient
  • Easy to use, feature-rich, informative
• New over-representation score to handle biases in length/GC-content of sequences
• Novel expression model - *CWM*
• Constructed a large, real-life, heterogeneous benchmark for testing motif finding tools
Acknowledgements

Tel-Aviv University
Chaim Linhart
Yonit Halperin
Igor Ulitsky
Adi Maron-Katz
Ron Shamir

The Hebrew University of Jerusalem
Gidi Weber

Handout:
Section 1 and 2

THANK YOU