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

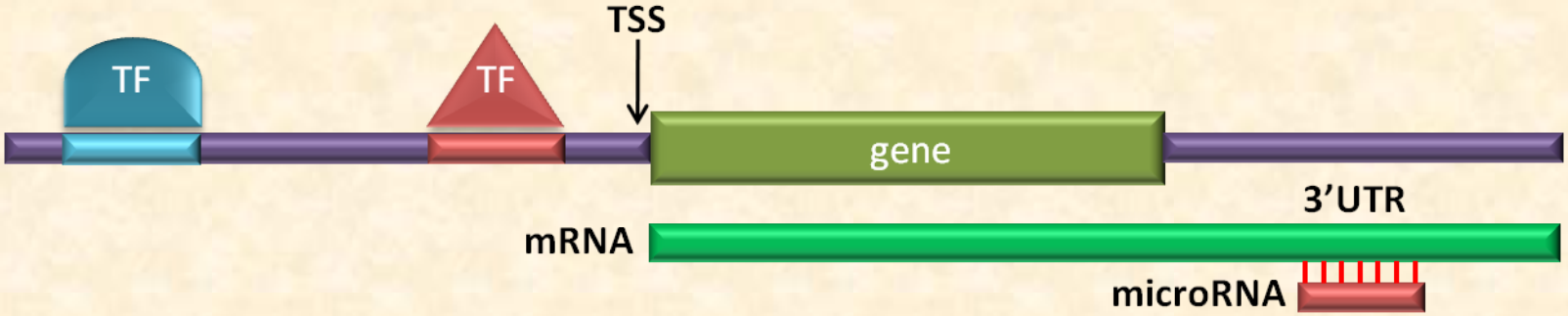
Yaron Orenstein, Chaim Linhart,
Yonit Halperin, Igor Ulitsky, Ron Shamir

AMADEUS

Allegro

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**)

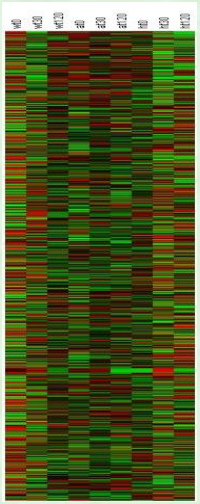


Motif discovery:
The typical two-step pipeline

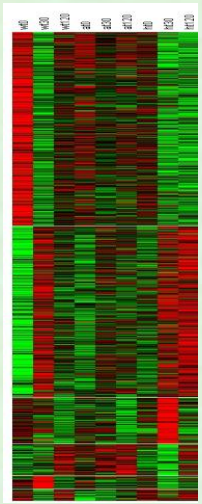
Co-regulated gene set

Promoter/3'UTR sequences

Gene expression microarrays

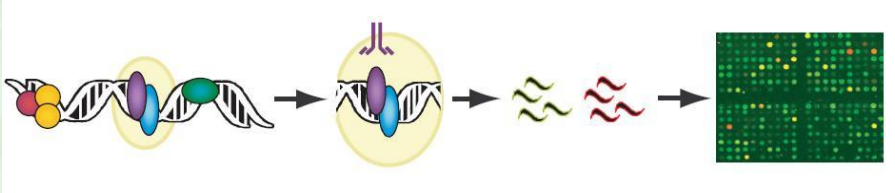


Clustering

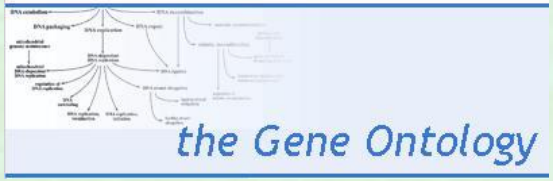


Motif discovery

Location analysis (ChIP-chip, ...)



Functional group (e.g., GO term)



Amadeus

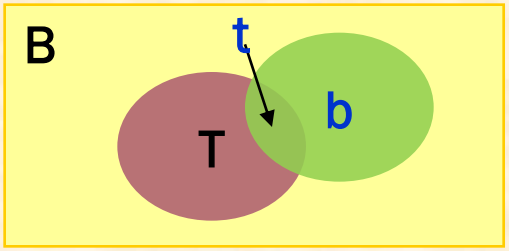
A Motif Algorithm for Detecting Enrichment in mUltiple Species

- **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**

- <u>**Input:**</u> Target set (*T*) = co-regulated genes
  Background (**BG**) set (*B*) = entire genome

  No sequence model is assumed!
- <u>Motif scoring:</u>
  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

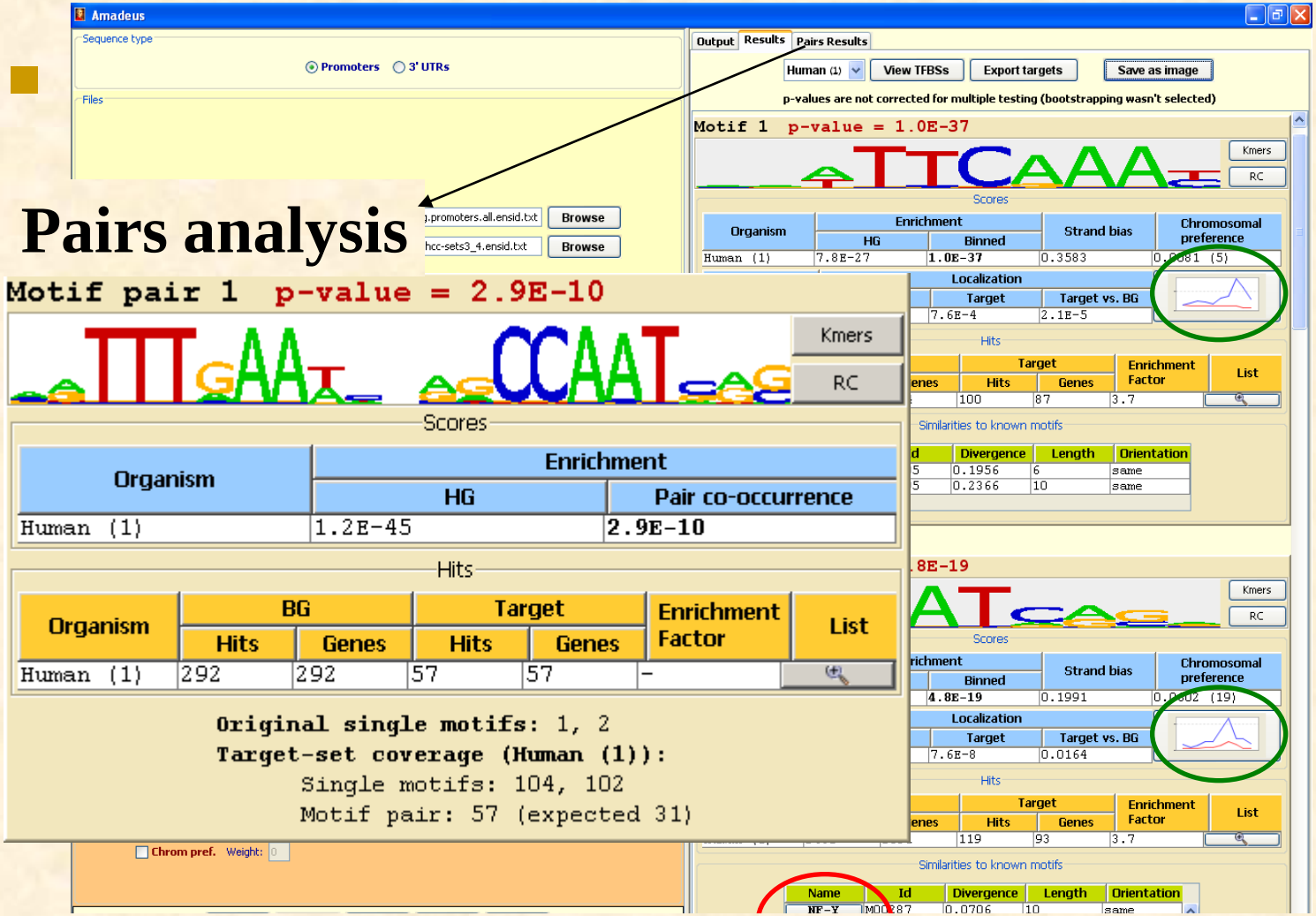
Drawback of the HG score

- 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

- 

CHR

NF-Y

(CCAAT-box)

- 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

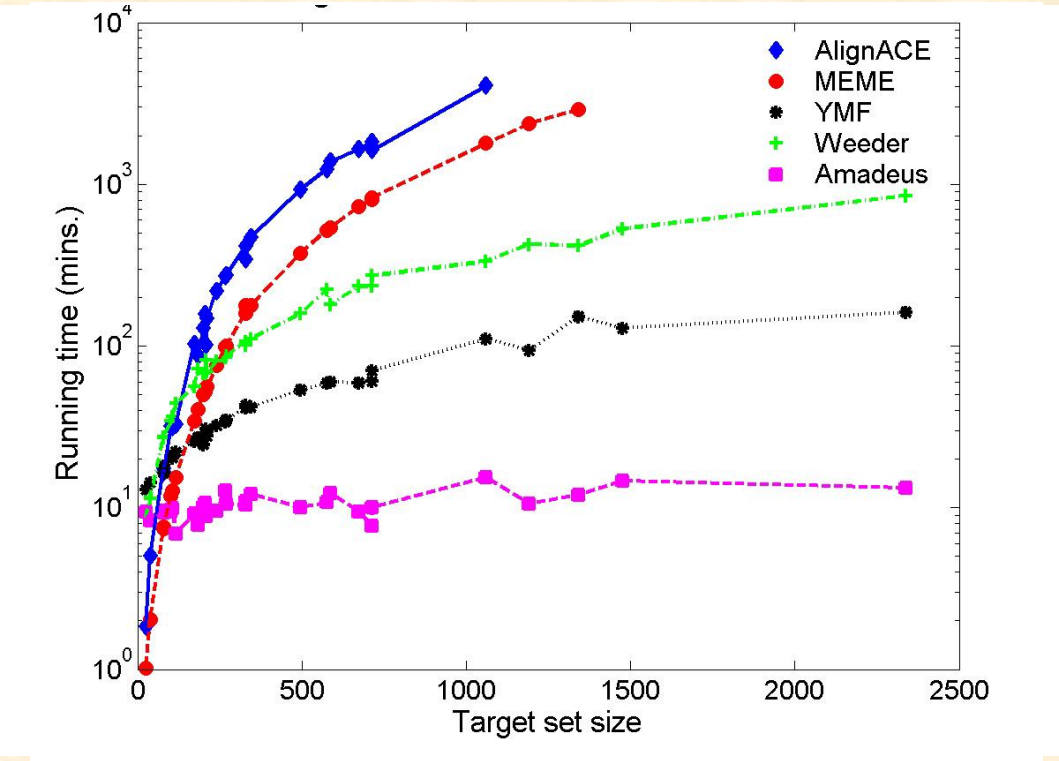
Source: Various (expression, ChIP-chip, Gene Ontology, …)

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)



ed score
vement

Amadeus

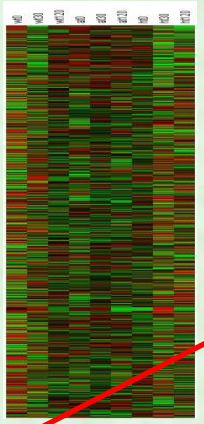
A Motif Algorithm for Detecting Enrichment in mUltiple Species

- **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.

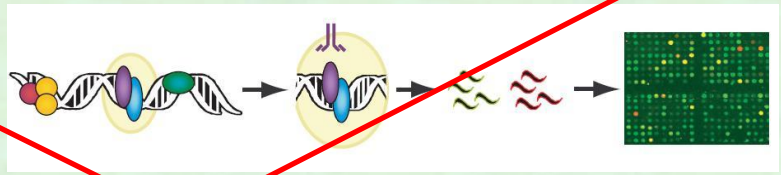
Amadeus – Global spatial analysis

Co-regulated gene set

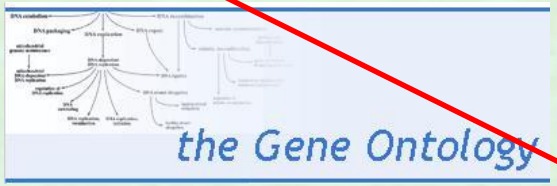
Gene expression microarrays



Location analysis (ChIP-chip, ...)



Functional group (e.g., GO term)



Promoter sequences

Output

Motif(s)

Task II: Global analyses

Scores for spatial features of motif occurrences

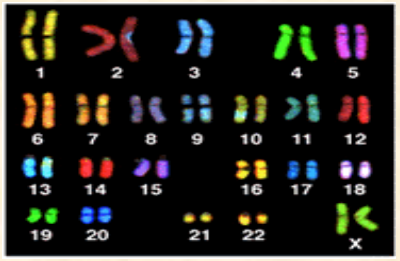
Input: Sequences (no target-set / expression data)

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

Name	Motif logo	Localization		Strand bias	Chrom pref.
		peak	*p*-val	*p*-val	*p*-val
A. Known TFs					
SP1		-60	10^{-129}	-	-
NF-Y		-90...-60	10^{-145}	-	10^{-4} (19)
GABP		-30...0	10^{-113}	-	-
TATA		-30	10^{-61}	10^{-15}	10^{-3} (6)
NRF1		-30	10^{-48}	-	-
ATF/CREB		-60	10^{-27}	-	-
MYC		-60...-30	10^{-27}	-	-
RFX1		-60	10^{-9}	-	-
B. Novel					
ACTACAWYTC		-90...-60	10^{-21}	10^{-8}	10^{-4} (19)
CTCGCGAGAT		-60...-30	10^{-7}	-	-
C. Other					
Splice donor site		+30...	10^{-23}	10^{-8}	-

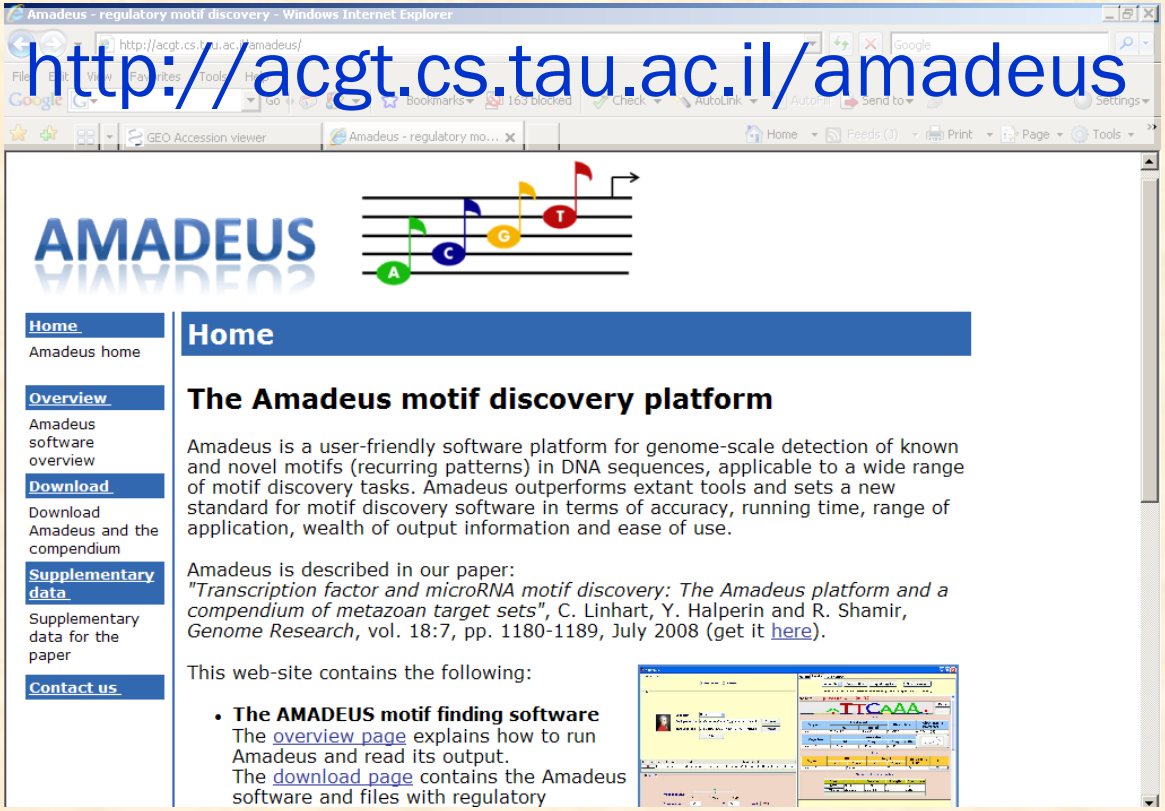
Amadeus is available at:

"Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets",

C. Linhart*, Y. Halperin*, R. Shamir,

Genome Research 18:7, 2008

(*equal contribution)



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 → Co-regulation → 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

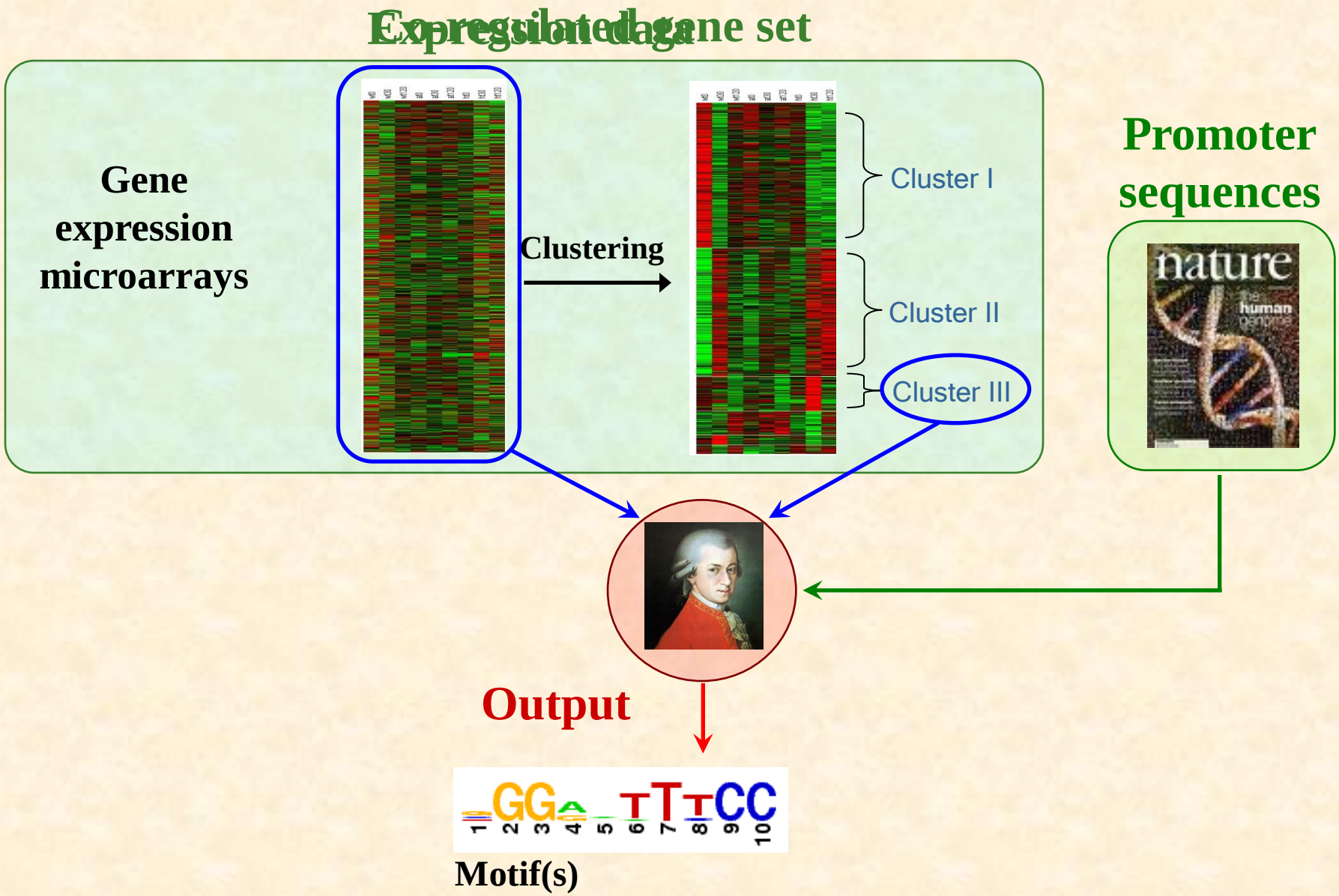
PRIMA – General description

- *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



Allegro: expression model

- Discretization of expression values

e_1=Up (U) ≥1.0
e_2=Same (S) (-1.0, 1.0)
e_3=Down (D) ≤-1.0

Expression pattern

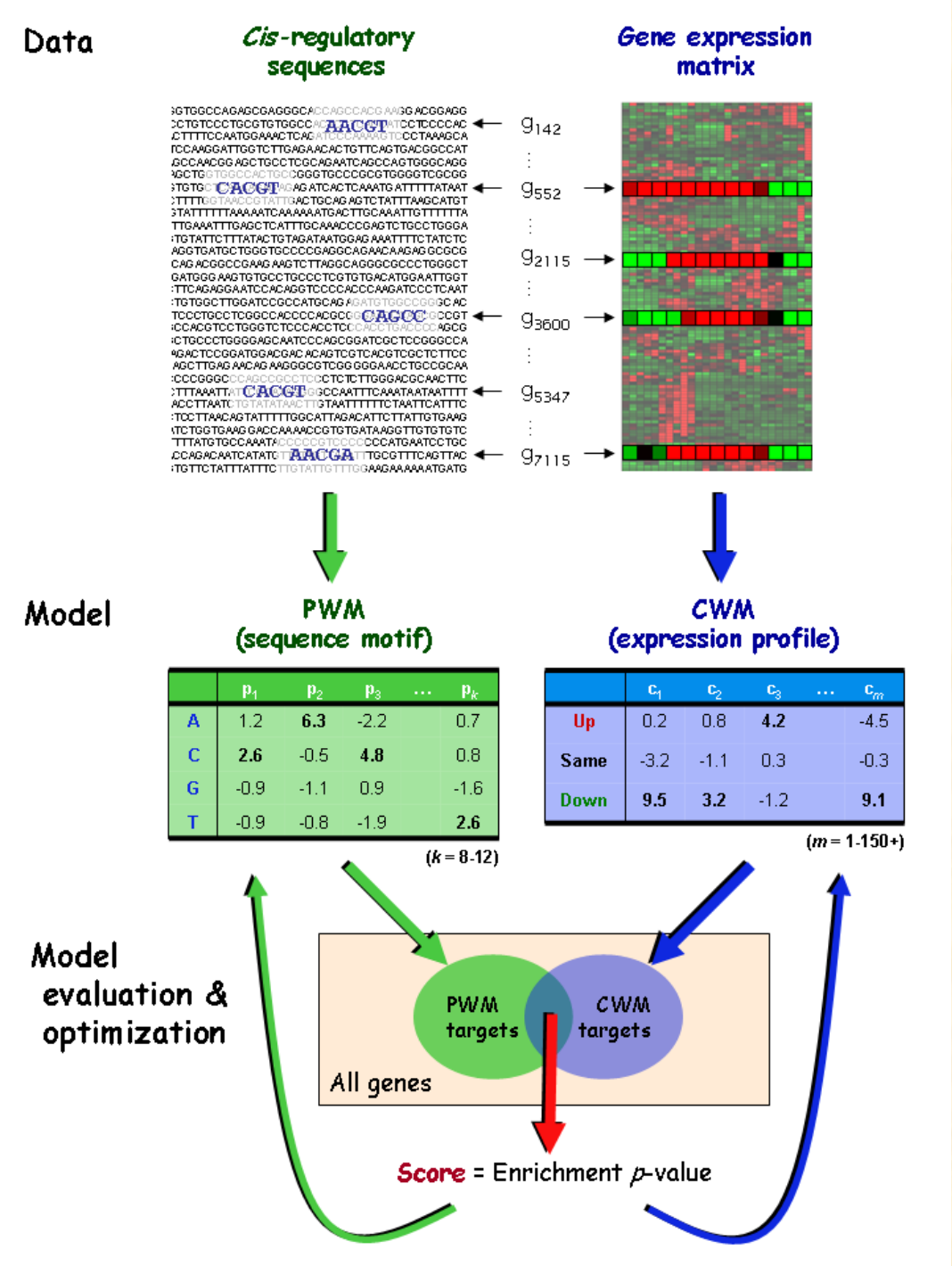
	c_1	c_2	...	c_m
g	-2.3	-0.8		1.5

Discrete expression Pattern (DEP)

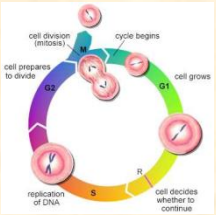
	c_1	c_2	...	c_m
g	D	S	...	U

- Expression data should be (partially) pre-processed, e.g.:
  - Time series → log ratio relative to time 0
  - Several tissues/mutations/... → 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

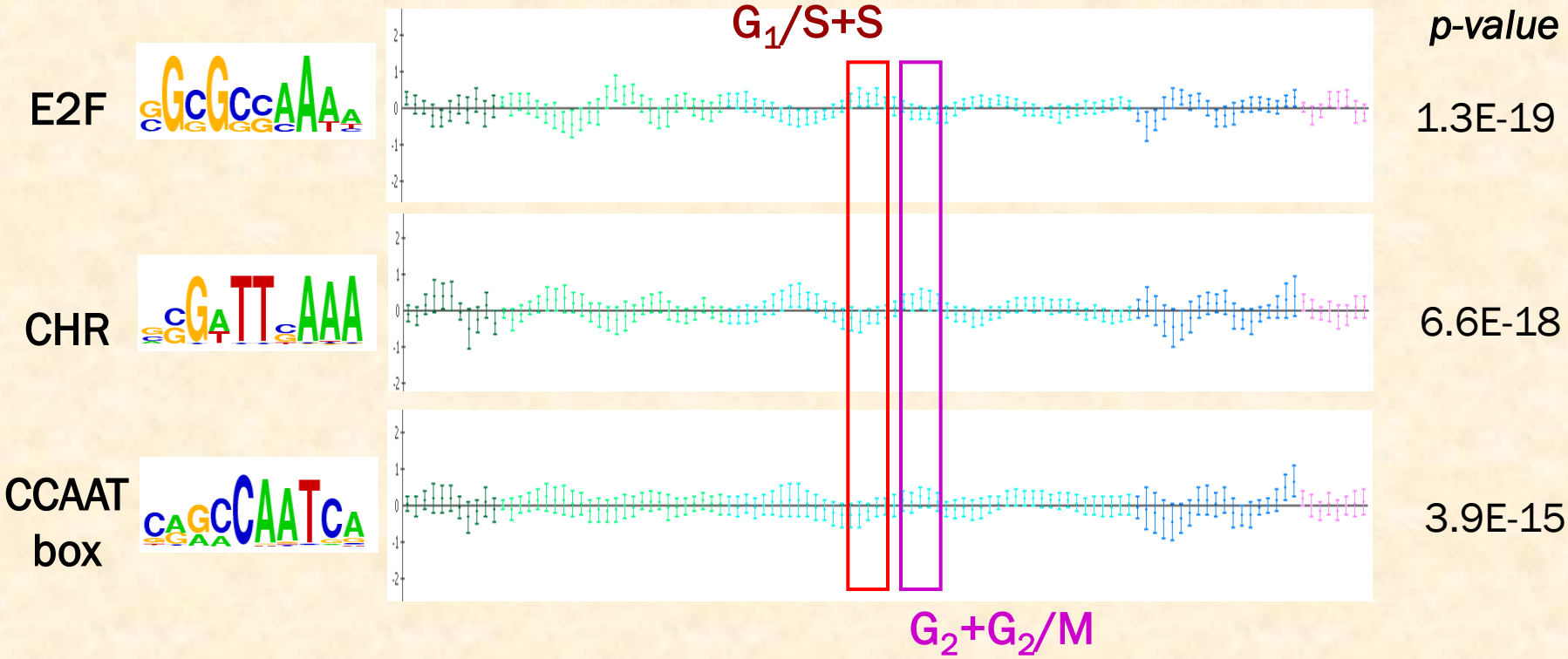
Allegro overview



Human cell cycle [Whitfield et al., ’02]

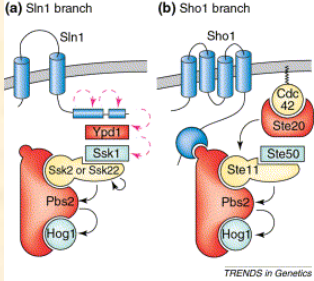


Large dataset: ~15,000 genes, 111 conditions,
promoters region: -1000…200 bps

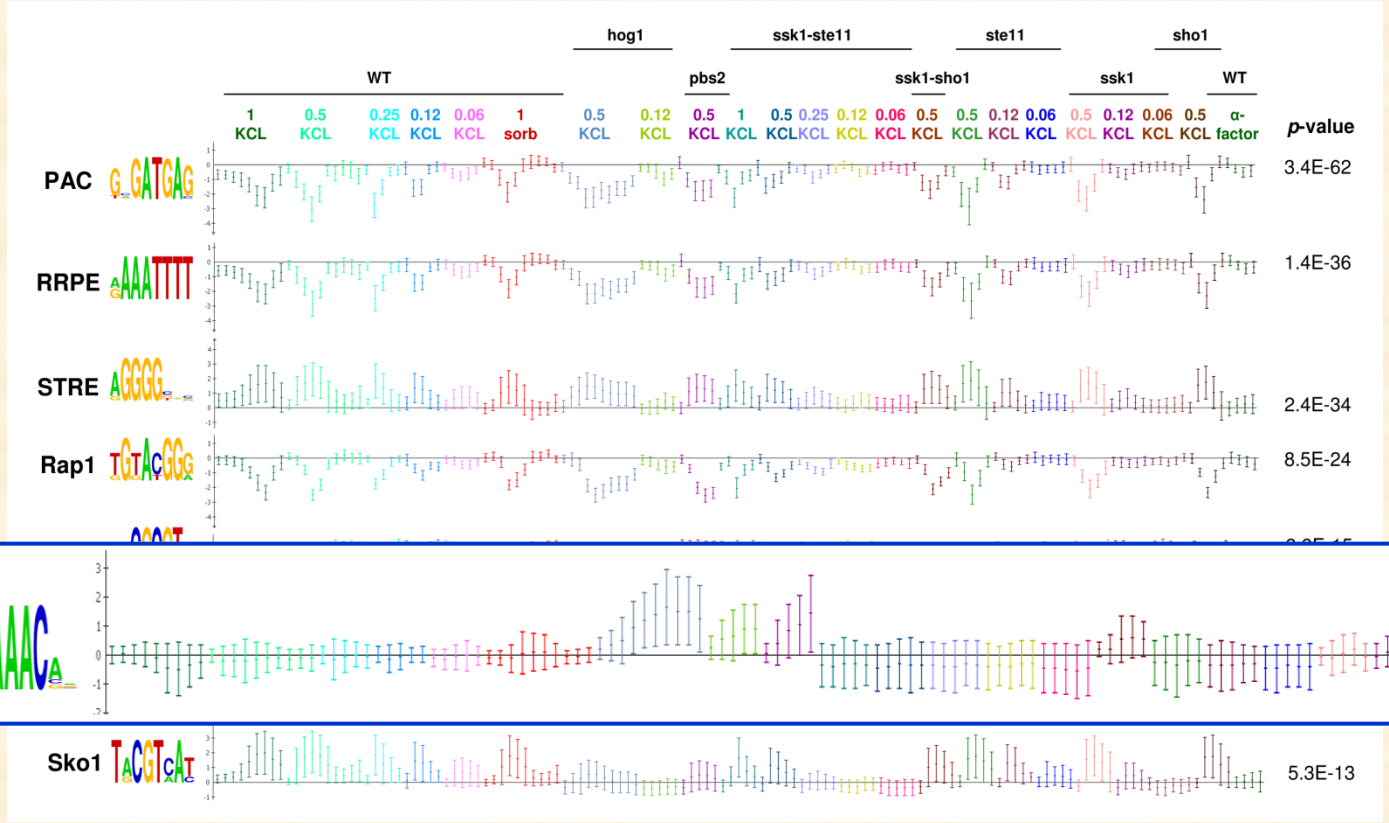


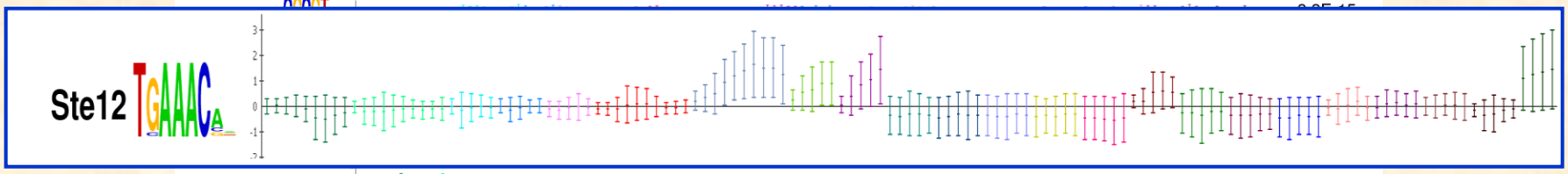
Allegro recovers the major regulators of the human cell cycle
[Elkon et al. ’03; Tabach et al. ’05; Linhart et al. ’05].

Yeast HOG pathway [O'Rourke et al. '04]



- ~6,000 genes, 133 conditions





- 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

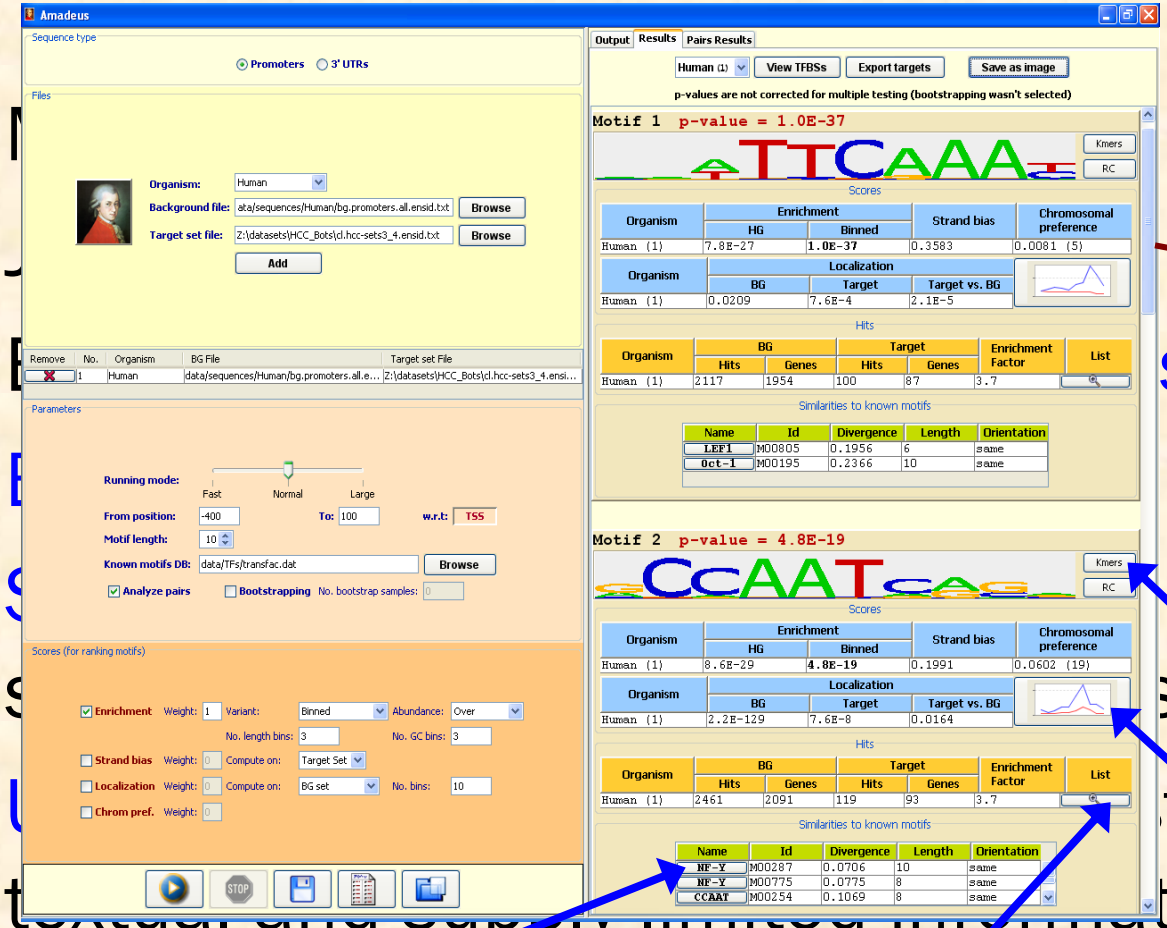
- M
- J
- E
- E
- S
- U

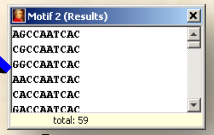
scores

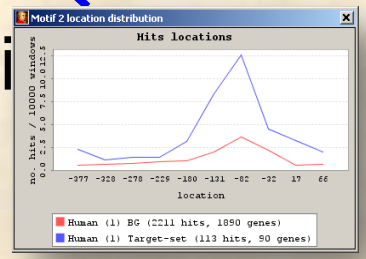
– ignore

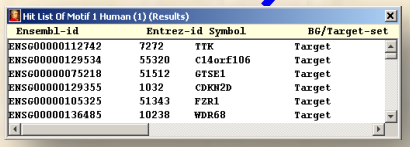
seq

t tools are







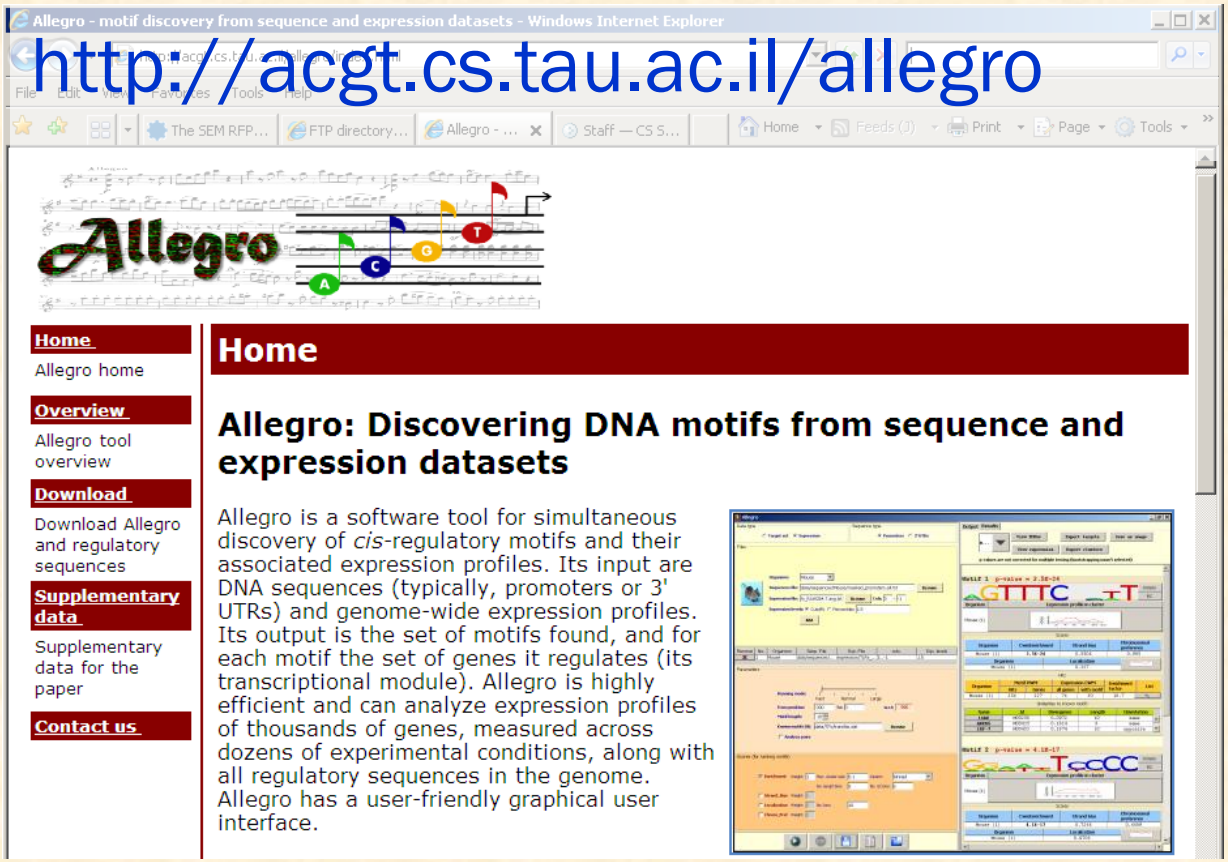


Z

Allegro is available at:

"Allegro: Analyzing expression and sequence in concert to discover regulatory programs",
Y. Halperin*, C. Linhart*, I. Ulitsky, R. Shamir,
Nucleic Acids Research, 2009

(*equal contribution)



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