

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

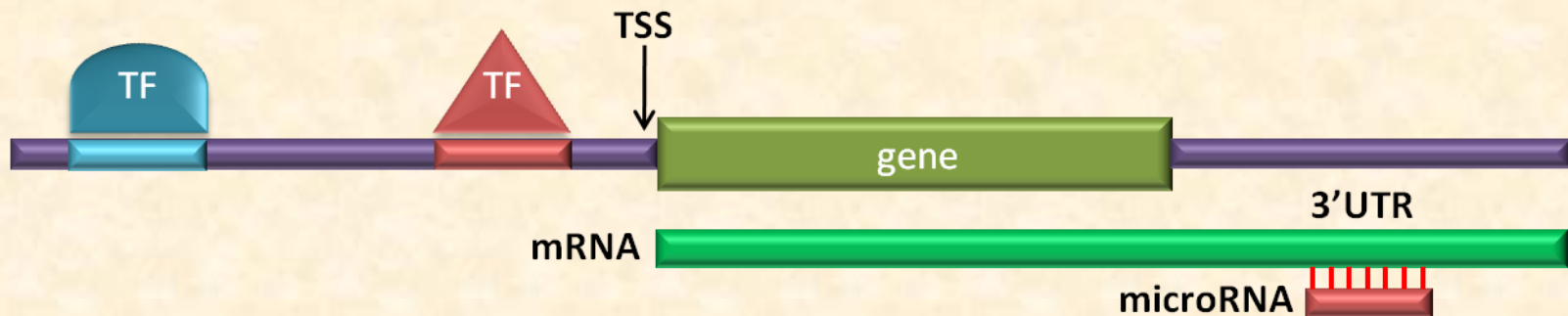
Yaron Orenstein, Chaim Linhart,
Yonit Halperin, Igor Ulitsky, Ron Shamir

AMADEUS



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**)

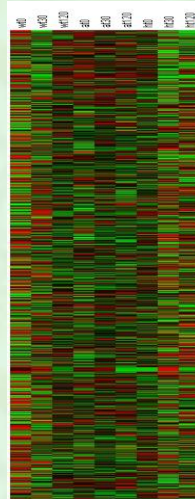


Motif discovery: The typical two-step pipeline

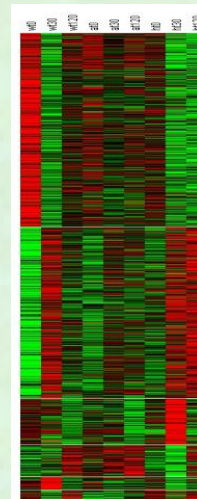
Co-regulated gene set

Promoter/3'UTR
sequences

Gene expression
microarrays



Clustering
→

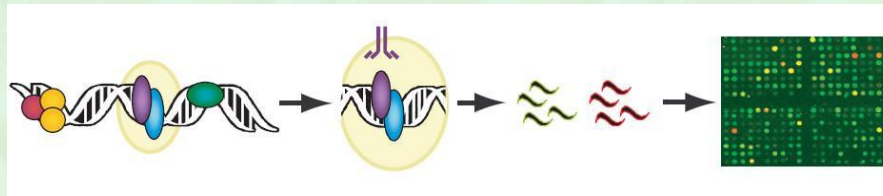


Cluster I

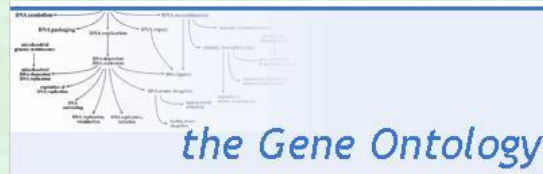
Cluster II

Cluster III

Location analysis
(ChIP-chip, ...)



Functional group
(e.g., GO term)



Motif
discovery



1 GG 2 3 4 5 TT 6 7 8 9 10 CC



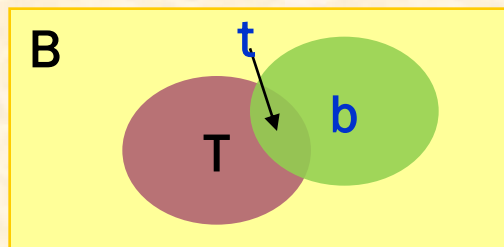
Amadeus

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

Motif pair 1 p-value = 2.9E-10

Original single motifs: 1, 2
 Target-set coverage (Human (1)):
 Single motifs: 104, 102
 Motif pair: 57 (expected 31)

Chrom. pref. Weight: 0

CHR

NF-Y
(CCAAT-box)

Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Pair co-occurrence		
Human (1)	1.2E-45	2.9E-10	0.3583	0.0081 (5)

Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Binned		
Human (1)	7.8E-27	1.0E-37	0.3583	0.0081 (5)

Organism	BG		Target		Enrichment Factor	List
	Hits	Genes	Hits	Genes		
Human (1)	292	292	57	57	-	

Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Binned		
Human (1)	4.8E-19	1.0E-19	0.1991	0.0002 (19)

Organism	BG		Target		Enrichment Factor	List
	Hits	Genes	Hits	Genes		
Human (1)	119	93	93	93	3.7	

Name	Id	Divergence	Length	Orientation
NF-Y	IM00287	0.0706	10	same

- These motifs form a module associated with G₂+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**

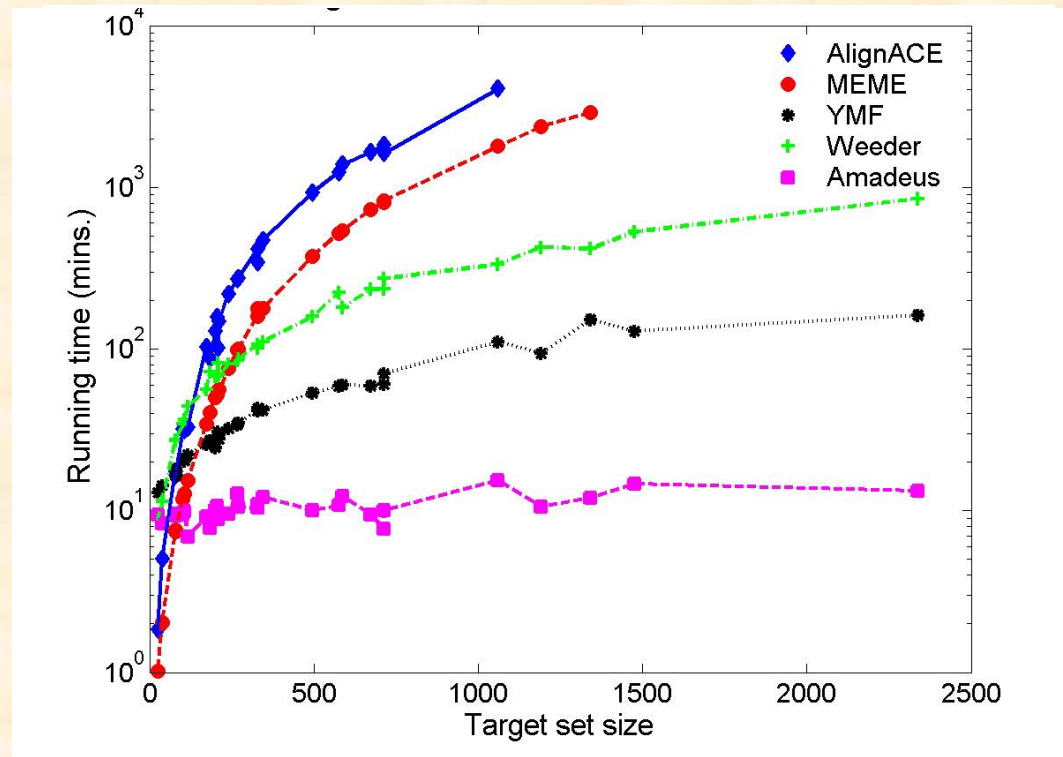
Source: Various (expression, ChIP-chip, Gene Ontology, ...)

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
ovement



Amadeus

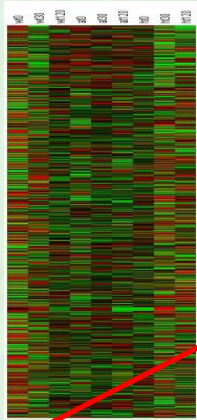
A Motif Algorithm for Detecting Enrichment in mUltiple Species

- 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.

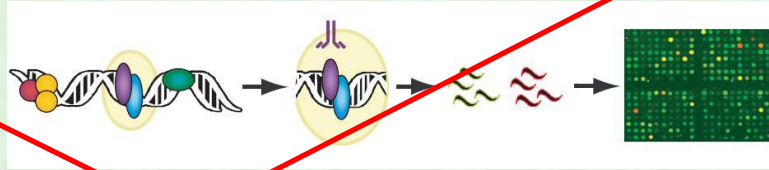
Amadeus – Global spatial analysis

Co-regulated gene set

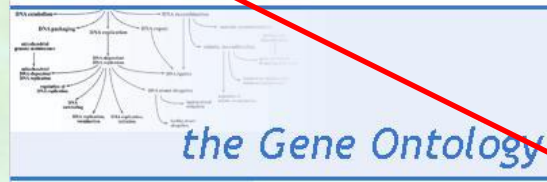
Gene expression
microarrays



Location analysis (ChIP-chip, ...)



Functional group (e.g., GO term)



Promoter
sequences



Output

1 2 3 4 5 6 7 8 9 10
GGTTC

Motif(s)

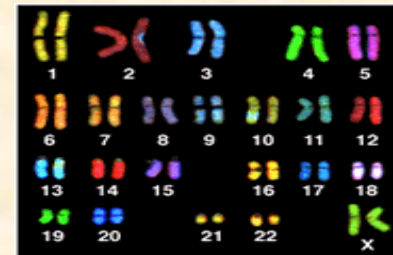
Task II: Global analyses

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**



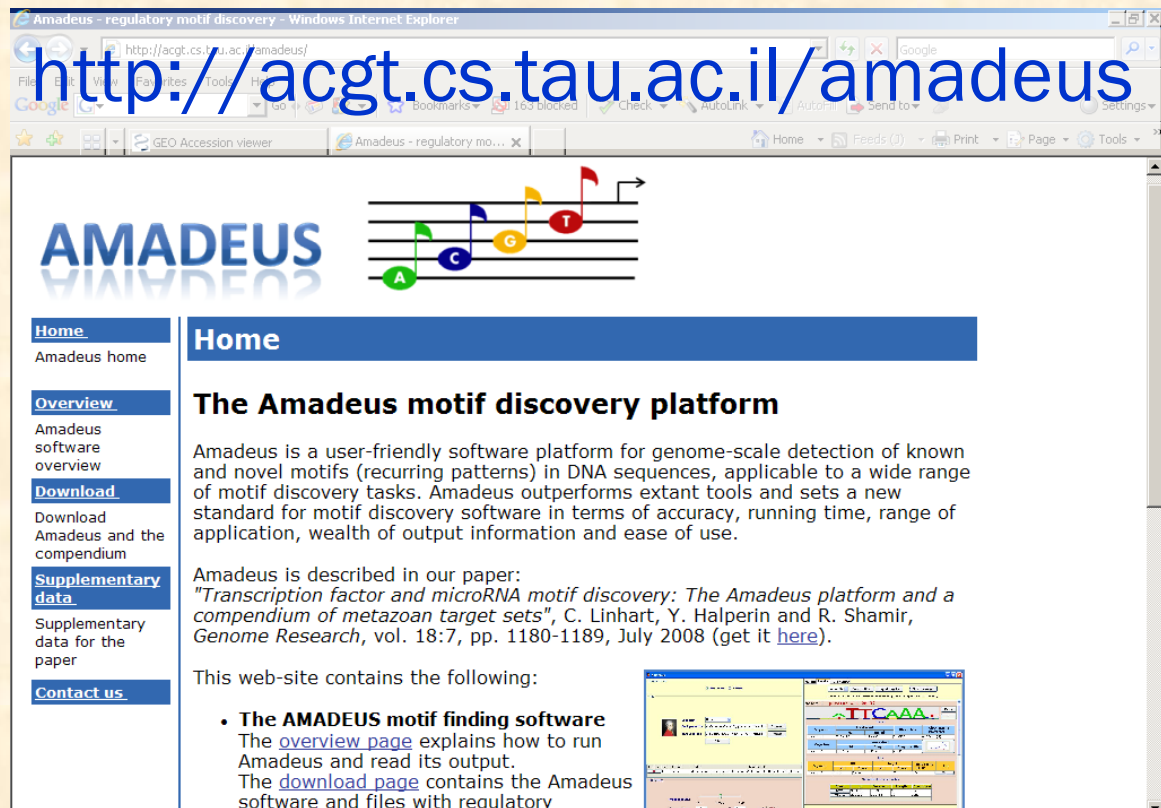
Name	Motif logo	Localization		Strand bias	Chrom pref.
		peak	p-val	p-val	p-val
A. Known TFs					
SP1	GC <u>CC</u> CCC	-60	10 ⁻¹²⁹	-	-
NF-Y	CCAAT	-90...-60	10 ⁻¹⁴⁵	-	10 ⁻⁴ (19)
GABP	GGAAGTG	-30...0	10 ⁻¹¹³	-	-
TATA	TATAA	-30	10 ⁻⁶¹	10 ⁻¹⁵	10 ⁻³ (6)
NRF1	TGC GC	-30	10 ⁻⁴⁸	-	-
ATF/CREB	ACGTCACAGA	-60	10 ⁻²⁷	-	-
MYC	GT CAC TG	-60...-30	10 ⁻²⁷	-	-
RFX1	CTG C AAC	-60	10 ⁻⁹	-	-
B. Novel					
ACTACAWYTC	ACTACA TC	-90...-60	10 ⁻²¹	10 ⁻⁸	10 ⁻⁴ (19)
CTCGCGAGAT	CTC CGAGAT	-60...-30	10 ⁻⁷	-	-
C. Other					
Splice donor site	GGT AG	+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)



<http://acgt.cs.tau.ac.il/amadeus>

AMADEUS

Home
Amadeus home

Overview
Amadeus software overview

Download
Download Amadeus and the compendium

Supplementary data
Supplementary data for the paper

Contact us

Home


The Amadeus motif discovery platform

Amadeus is a user-friendly software platform for genome-scale detection of known and novel motifs (recurring patterns) in DNA sequences, applicable to a wide range of motif discovery tasks. Amadeus outperforms extant tools and sets a new standard for motif discovery software in terms of accuracy, running time, range of application, wealth of output information and ease of use.

Amadeus is described in our paper: *“Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets”*, C. Linhart, Y. Halperin and R. Shamir, *Genome Research*, vol. 18:7, pp. 1180-1189, July 2008 (get it [here](#)).

This web-site contains the following:

- **The AMADEUS motif finding software**
The [overview page](#) explains how to run Amadeus and read its output.
The [download page](#) contains the Amadeus software and files with regulatory





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

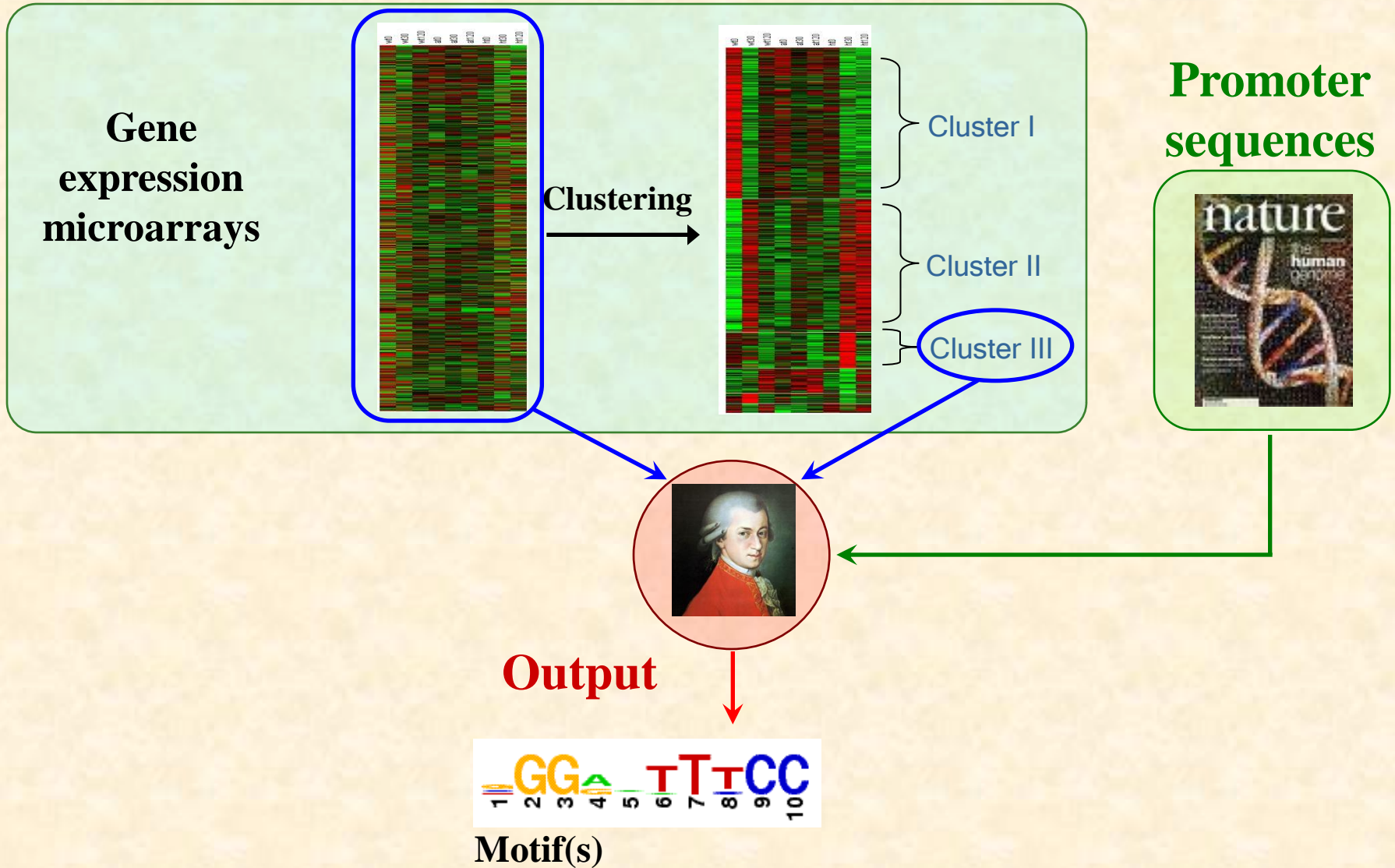
- *Co-expression* → *Co-regulation* → *common cis-regulatory promoter elements*
- Identification of co-expressed genes using microarray technology (clustering)
- Computational identification of *cis*-regulatory elements that are over-represented in promoters of the co-expressed genes

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.

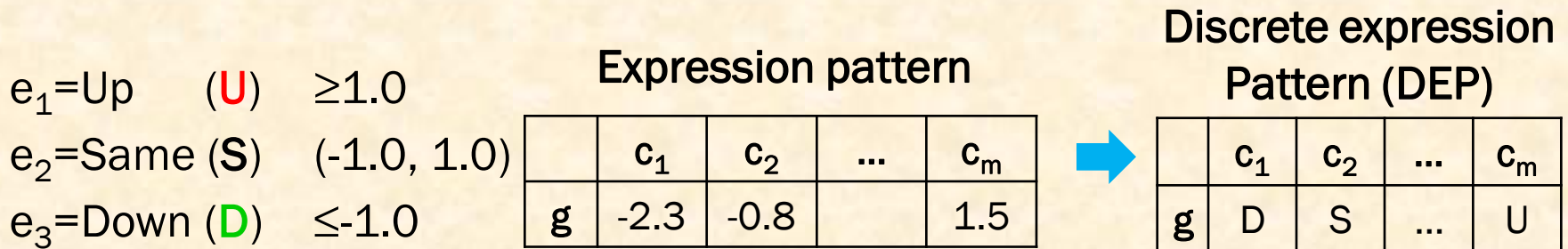
Amadeus - Allegro

Expressed data gene set



Allegro: expression model

- Discretization of expression values



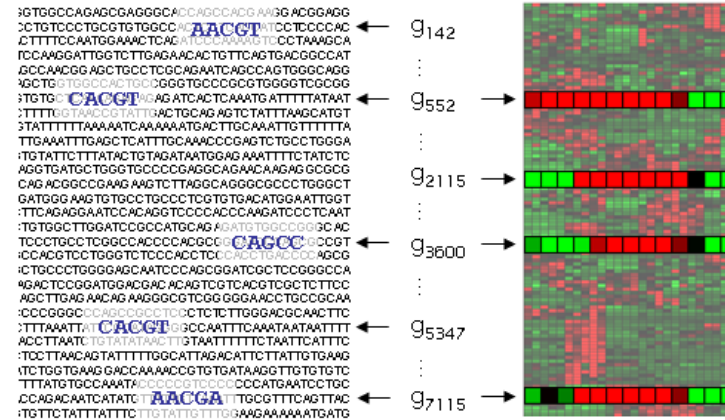
- Expression data should be (partially) **pre-processed**, e.g.:
 - Time series \rightarrow log ratio relative to time 0
 - Several tissues/mutations/... \rightarrow standardization
 - Do NOT filter out non-responsive genes
- Expression model: **CWM** = Condition Weight Matrix
 - Non-parametric, log-likelihood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

	p_1	p_2	p_3	...	p_k
A	1.2	6.3	-2.2		0.7
C	2.6	-0.5	4.8		0.8
G	-0.9	-1.1	0.9		-1.6
T	-0.9	-0.8	-1.9		2.6

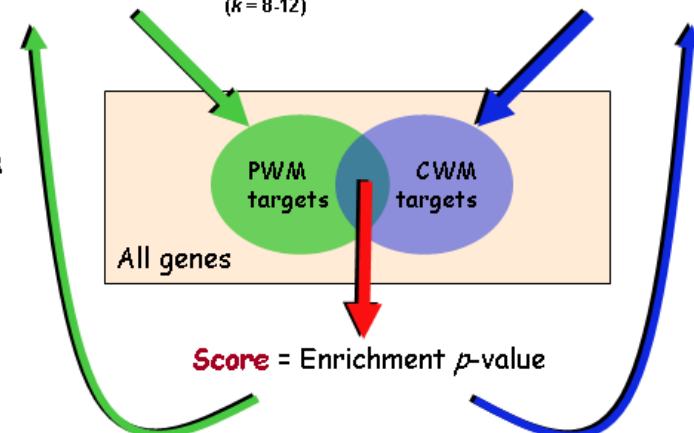
($k = 8-12$)

CWM
(expression profile)

	c_1	c_2	c_3	...	c_m
Up	0.2	0.8	4.2		-4.5
Same	-3.2	-1.1	0.3		-0.3
Down	9.5	3.2	-1.2		9.1

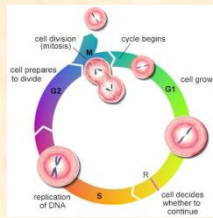
($m = 1-150+$)

Model evaluation & optimization

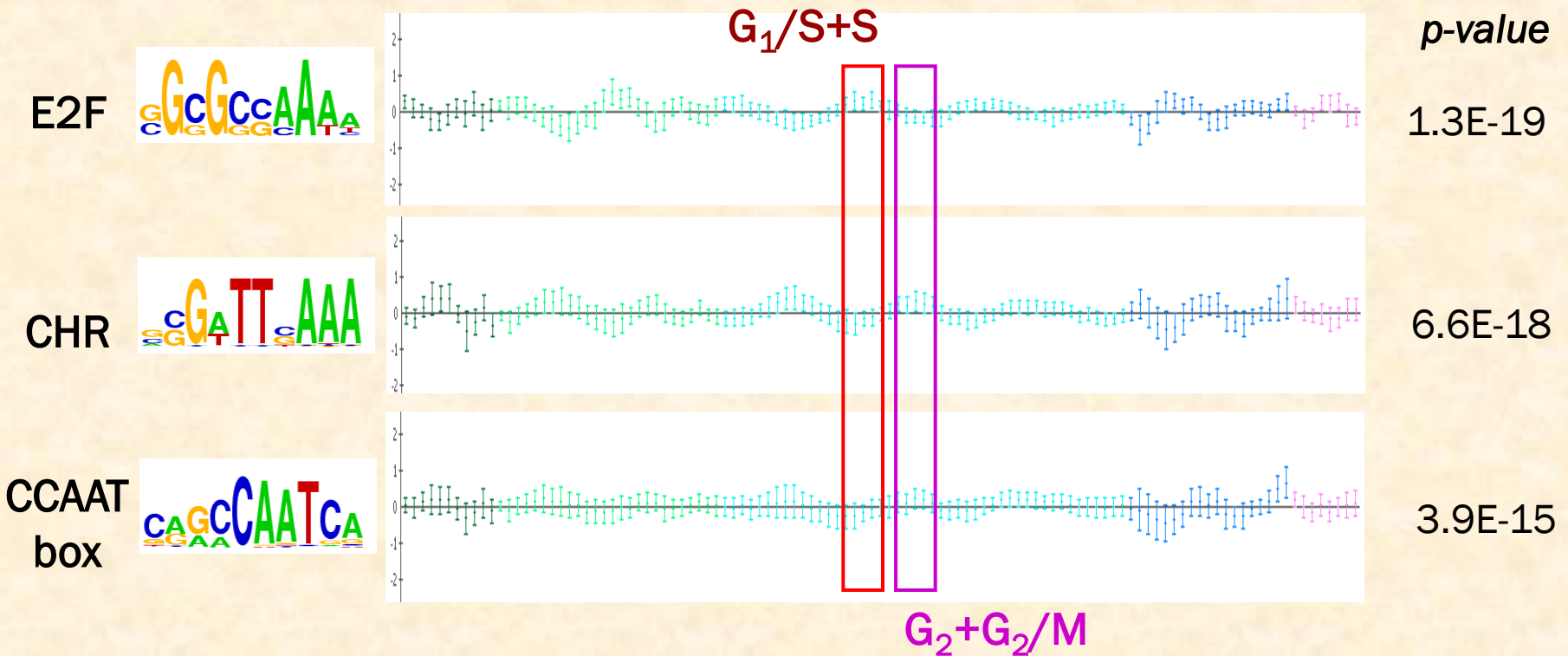


Score = Enrichment p -value

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



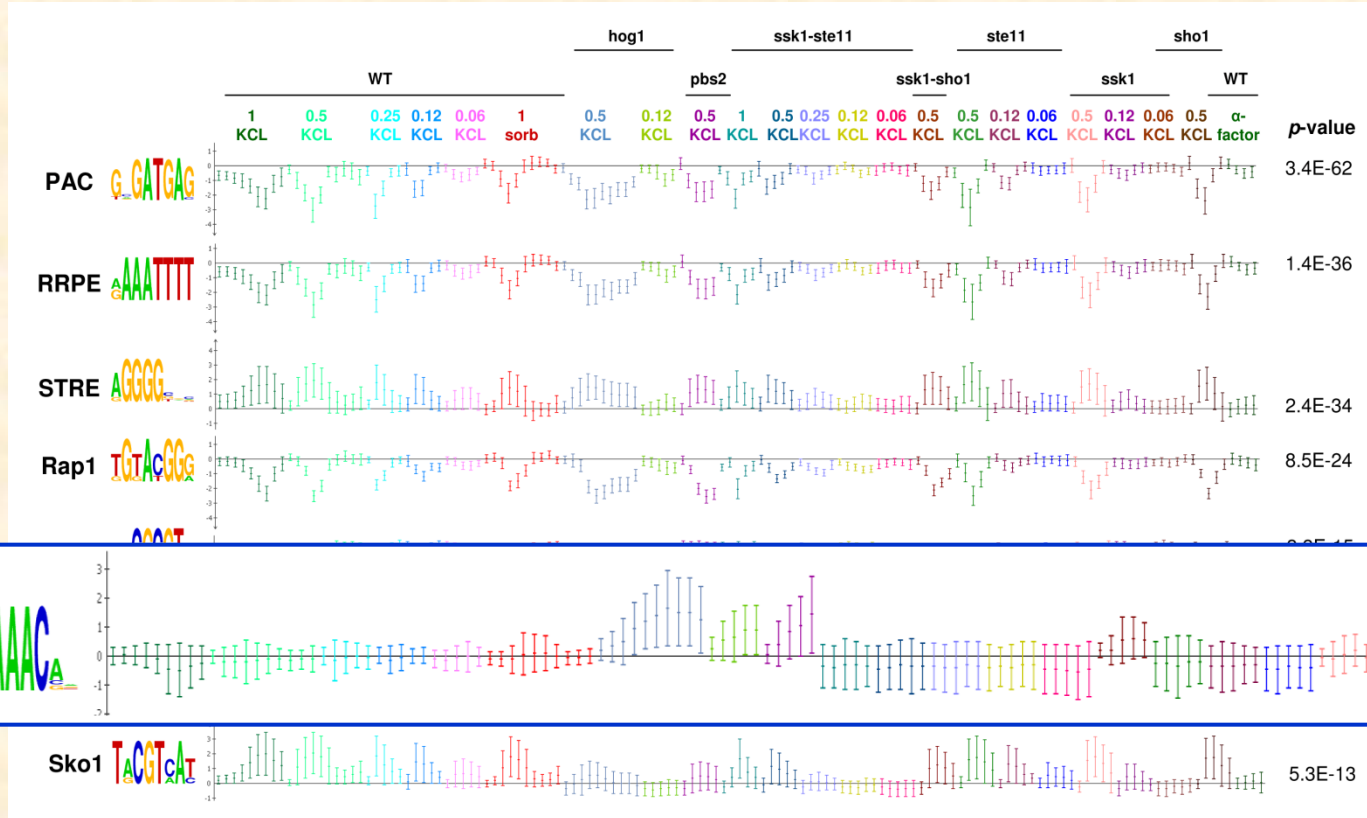
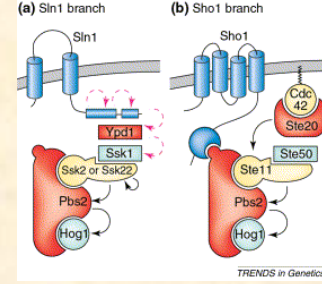
Large dataset: ~15,000 genes, 111 conditions,
promoters region: -1000...200 bps



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

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

The screenshot shows the Amadeus software interface. On the left, the 'Parameters' section includes 'Running mode' (Fast, Normal, Large), 'From position' (-400), 'To' (100), 'Motif length' (10), and 'Known motifs DB' (data/TFs/transfac.dat). The 'Scores (for ranking motifs)' section has checkboxes for 'Enrichment', 'Strand bias', 'Localization', and 'Chrom. pref.', each with a weight and various options.

The main 'Output Results' panel shows two motifs. Motif 1 has a p-value of 1.0E-37 and the sequence TTCAAA. Motif 2 has a p-value of 4.8E-19 and the sequence CCAAT. Each motif section includes a 'Scores' table with columns for Organism, Enrichment (HG, Binned), Strand bias, and Chromosomal preference. Below this is a 'Localization' table with columns for Organism, BG, Target, and Target vs. BG. Further down is a 'Hits' table with columns for Organism, Hits, Genes, and Enrichment Factor. At the bottom of each motif section is a 'Similarities to known motifs' table with columns for Name, Id, Divergence, Length, and Orientation.

scores → Z

ignore
seq
t tools are

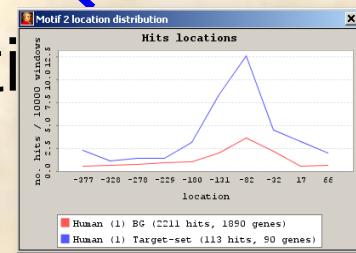


Hit List of Motif 1 Human (1) (Results)

Ensembl-id	Entrez-id	Symbol	BG/Target-set
ENS00000112742	7272	TTK	Target
ENS00000129534	55320	C14orf106	Target
ENS00000075218	51512	GTSE1	Target
ENS000000129355	1032	CDKN2b	Target
ENS000000105325	51343	FZK1	Target
ENS00000136405	10238	WDR69	Target

Motif 2 (Results)

AGCCAATCAC
CGCCAATCAC
GGCCAATCAC
AGCCAATCAC
CACCAATCAC
GACCAATCAC
total: 59



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)

<http://acgt.cs.tau.ac.il/allegro>

Home
Allegro home

Overview
Allegro tool overview

Download
Download Allegro and regulatory sequences

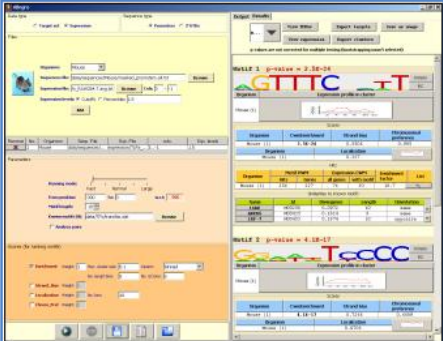
Supplementary data
Supplementary data for the paper

Contact us

Home

Allegro: Discovering DNA motifs from sequence and expression datasets

Allegro is a software tool for simultaneous discovery of *cis*-regulatory motifs and their associated expression profiles. Its input are DNA sequences (typically, promoters or 3' UTRs) and genome-wide expression profiles. Its output is the set of motifs found, and for each motif the set of genes it regulates (its transcriptional module). Allegro is highly efficient and can analyze expression profiles of thousands of genes, measured across dozens of experimental conditions, along with all regulatory sequences in the genome. Allegro has a user-friendly graphical user interface.



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

Handout:

Section 1 and 2

The Hebrew University of Jerusalem

Gidi Weber

