Detecting Transcription Factor Binding Sites

Sören Sonnenburg, Mikio Braun
Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany

Fraunhofer Institut
Rechnerarchitektur
und Softwaretechnik

Overview:

- Problem Description
- Roadmap
- Tools
- Experiments
- Outlook and Results

Problem Description

Given:

- Candidate lists of genes

  1. top ranked genes from previous microarray experiments
  2. based on literature

Assumption:

- Subset of genes are simultaneously expressed (which are (de-)activated by the same transcription factors).

Method:

- Find common TFBS on DNA
  ↳ shared motifs in promoter regions of these genes.

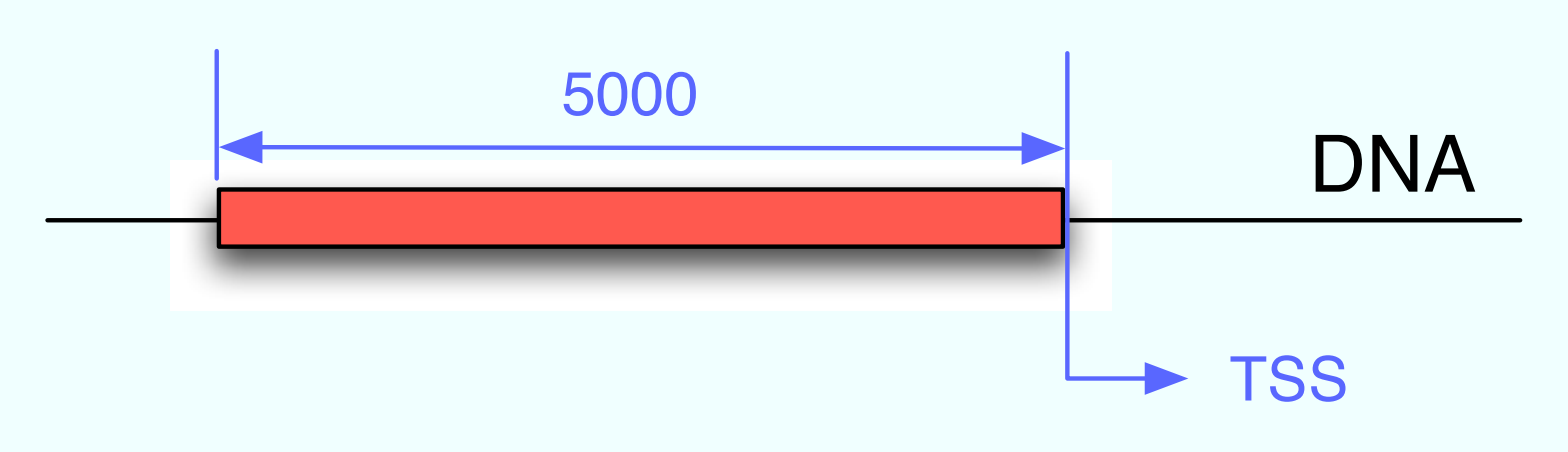
FIRST

I. Identification of Promoter Regions

II. Promotor Analysis - Detection of TFBS

- Database matching vs. de-novo motifs.
- Generation of background model.
- Modules.
- Phylogenetic Footprints

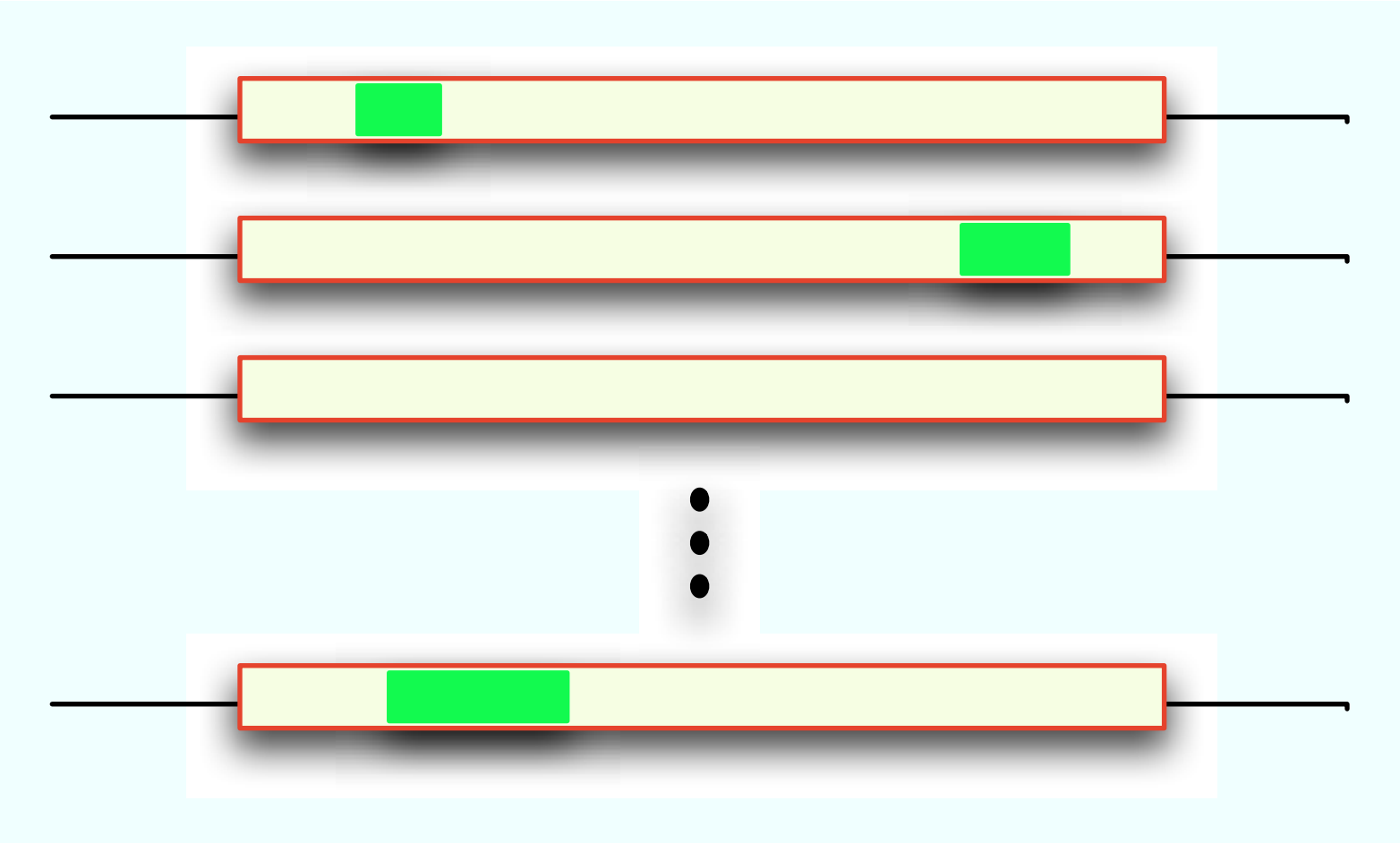
Lucky: Genes are known.



Steps:

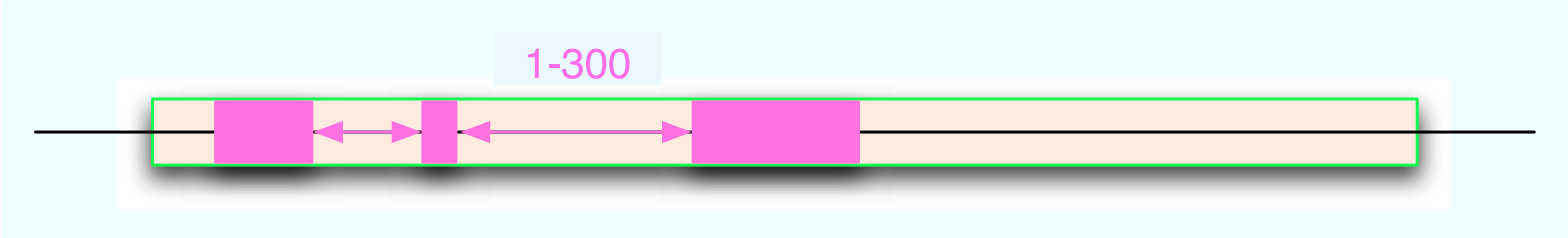
- obtain mRNA
- match to human (mouse/rat) genome (check with genome browser)
- extract 5kb upstream of TSS

Step II - Detection of TFBS



Properties of TFBS

- Length 4-26.
- Core motifs small (≈ 4), rest highly variable.
- Appears in modules



- Is conserved between species (mouse, rat)

Data Sets

1) Artificial promoter sequences (5k length): 39 random sequences, uniformly sampled

- NOISE
- NOISE/MOTIF: insert 'ACATCGTTACGTATGG' into 8 sequences
- NOISE/MODULE: insert module 'ACATCGTTACGTATGG' – 'TTTTACGATGGTAG' (4 mutations, 1-100 distance) into 8 sequences

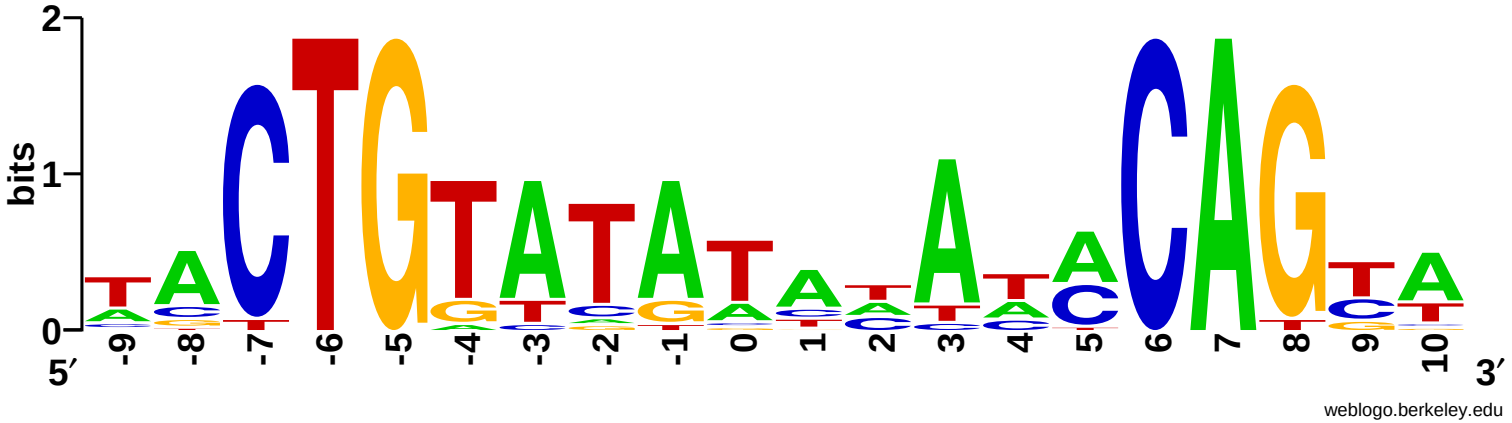
2) Real promoter sequence (5k length): 39 random sequences extracted randomly from 17k promoter regions

- NEGATIVE
- POSITIVE: insert into 8 sequences 3 modules of 2 motifs sampled from TBFS PWMs (0-100 distance)

3) WT1 associated promoters (5k length): 8 WT1 assoc. + 31 random real promoter sequences

Tasks with increasing complexity.

Transfac database of known TFBS sequences and their position-weighted matrices (762 TFBS PWMS)

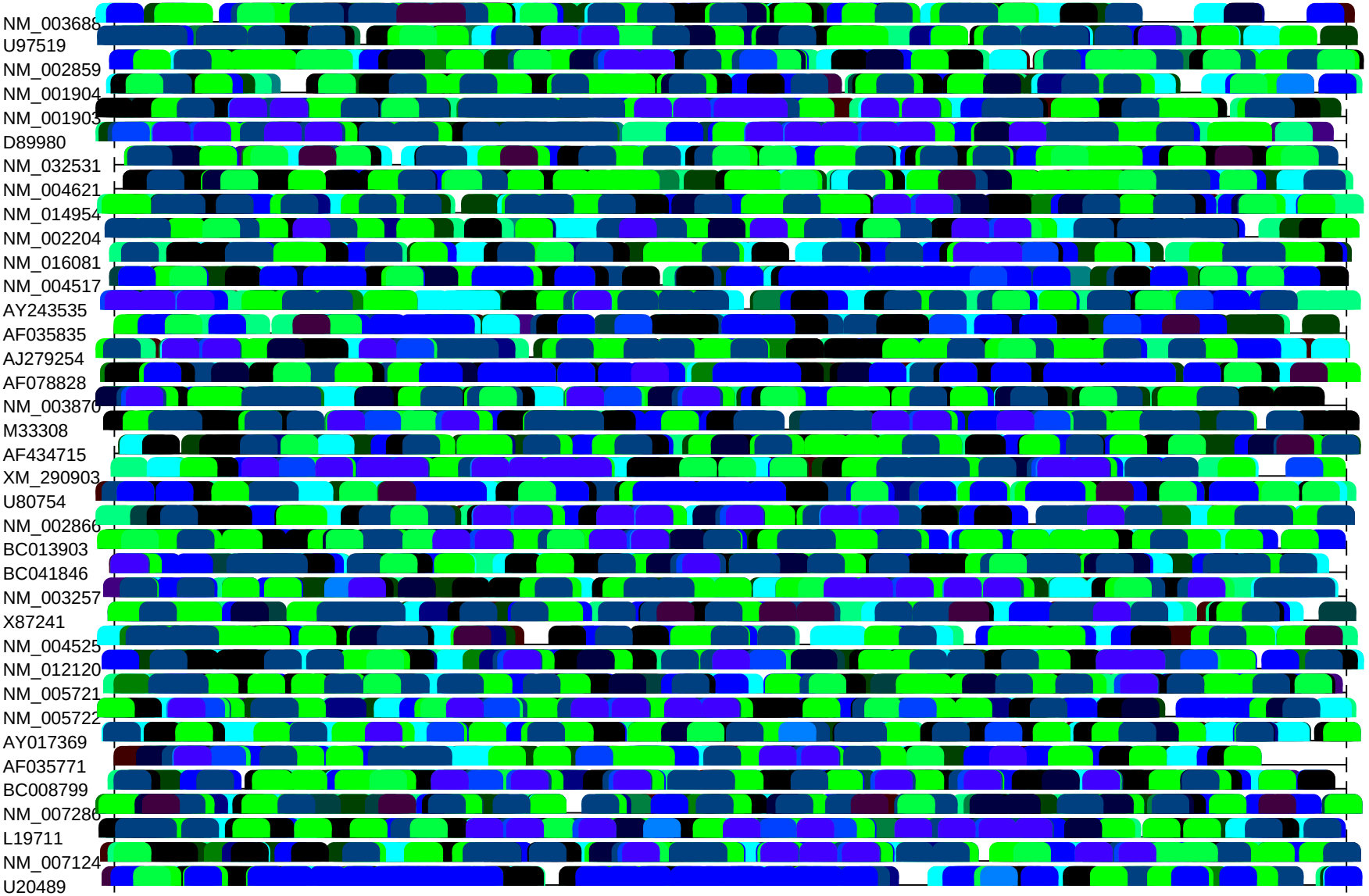


- slide known PWMs over promotor sequences
- score according to

$$S = \frac{current - min}{max - min}, \quad score = \sum_i I_i f_{i,b(i)}, \quad I_i = \sum_{b \in \{A,C,G,T\}} f_{i,b} \log\left(4 f_{i,b}\right)$$

- threshold at $S \geq 0.7$

Transfac Results



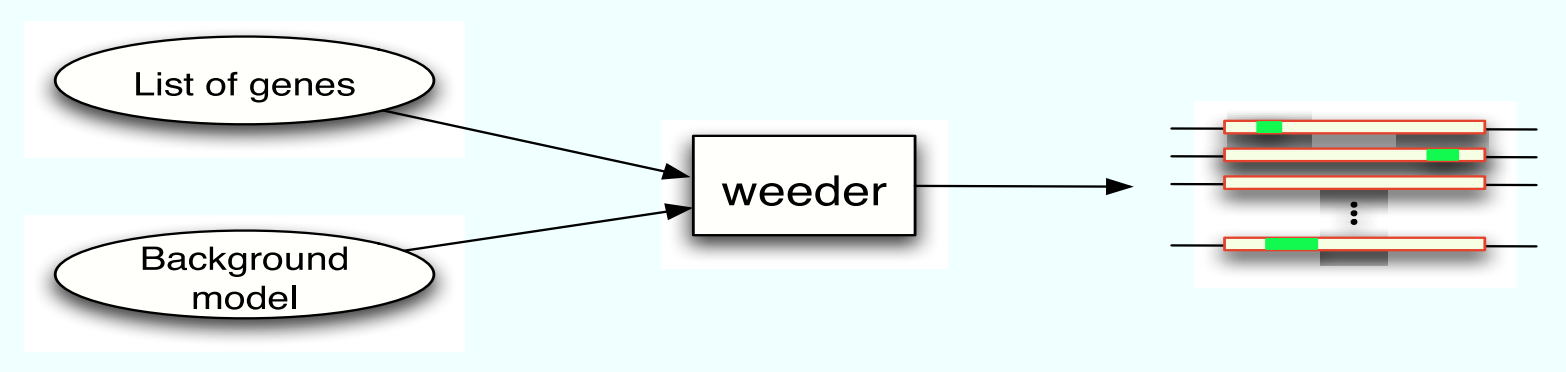
Transfac Discussion

- similar results on artificial datasets
- PWMs too unspecific (match everywhere)

further information needed (conservation/modules)

Weeder

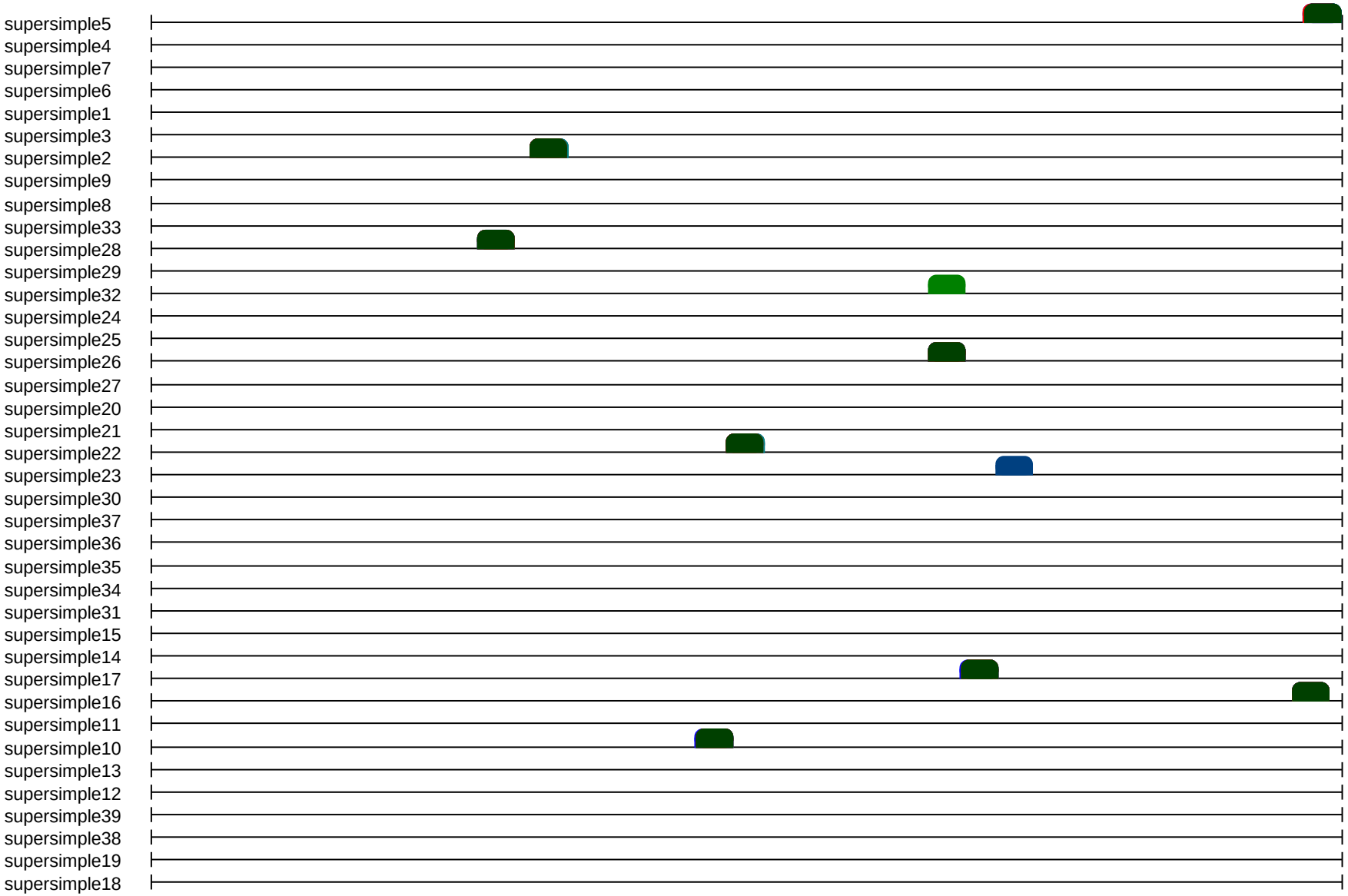
- Winner **Nature Biotechnology (2005).**
- Program for finding novel motifs (TFBS) conserved in a set of sequences.
- find overrepresented k-mers w.r.t. "background noise"
  - extract background from all 17,000 promotor regions (5kb long)
  - compute frequencies of 6-mers/8-mers.
- detects up to 12mers with up to 4 mismatches.



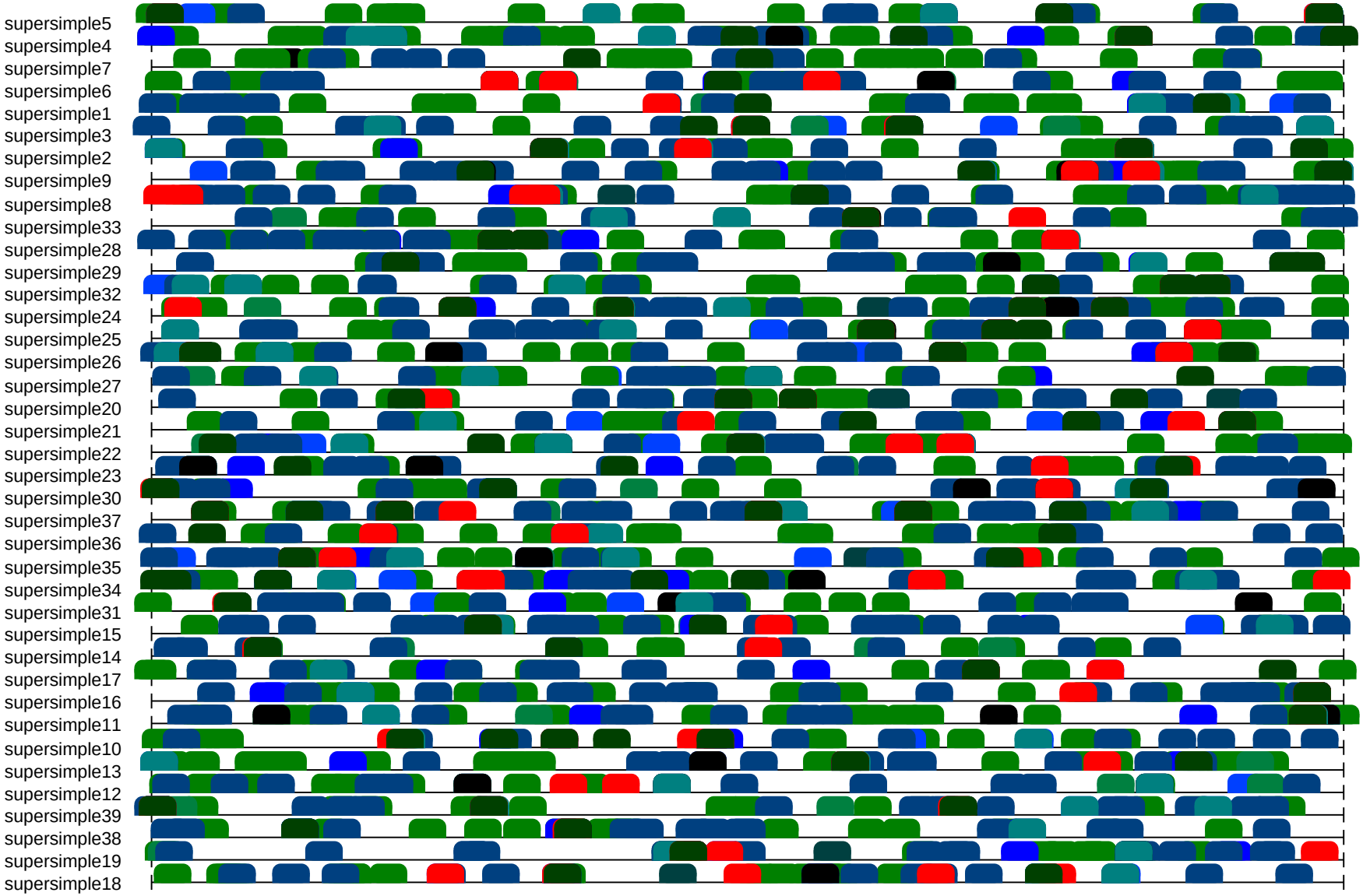
Weeder Results I

$\Rightarrow$ works on toy data

Weeder Results II



Weeder Results II



Weeder Discussion

- in our data (39 sequences each 5kb) one finds shared 8-mers on 8 sequences with probability ≈ 1

  Weeder cannot find them stand-alone!

- further information needed like
  - conservation
  - modules
  - which TF
  - shorter sequences ($< 5kb$)
  - require motifs to appear in more sequences than just 8 of 39 sequences

Module Detectors

Two candidates:

- cis-module
  - promising approach: *de-novo* motif and module predictor (in a single step)
  - based on the Bayesian framework
- cluster buster
  - takes e.g. transfac PWMs as input and identifies clusters of potential TFBS
  - computes scores w.r.t. a probabilistic model of clusters of binding sites

Results: Module Detectors

- cis-module
  - does not find anything on toy data
  - finds motifs in random promotors
  - fixed, built-in background model
- cluster buster
  - not applicable to toy data
  - finds motifs in random promotors

Methods have to be combined with further information

Conservation

- **Conjecture:** Functional BS are conserved between species.
- **Approach:** Generation of cross-species aligned sequences.

But:

- Level of conservation also varies greatly between different genes
- even when using conservation weeder and transfac-scoring still detect *many* potential TFBS

Conservation Human—Mouse

Summary - Difficulties

- BS are highly variable and short
- one finds motifs of length 8 with probability ≈ 1 in 8 of 39 sequences of length $5kb$
- conservation not a reliable indicator - we still find to many potential BS (and BS we are looking for must not be conserved)

Improvements

It would help to...

- know which TFs are potentially involved.
- use shorter/fewer sequences ($< 5kb$) (group them ?)
- require motifs to appear in more sequences than just 8 of 39 sequences
- know the length of TFBS
- know whether BS appear in consisten modules

Develop Integrated approach:

Existing methods show promising approaches, but lack in one respect or the other.

Approach should combine all available information:

- known binding sites (transfac)
- over-expressed motifs
- conservation information
- clusters