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.



I. Identification of Promoter Regions

II. Promotor Analysis - Detection of TFBS

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

STEP I - IDENTIFICATION OF PROMOTER REGIONS

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 (\approx 4), rest highly variable.
- Appears in modules



• Is conserved between species (mouse, rat)

DATA SETS

FIRST

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.

TOOLS

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



- slide known PWMs over promotor sequences
- score according to

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

• threshold at $S \geq 0.7$

FIRST

TRANSFAC RESULTS





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

further information needed (conservation/modules)

WEEDER



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



ACATCGTTACGTATGG ACATCGTTACGT

CATCGTTACG

TCGTTACG

ATCGTTACGT

CATCGTTACGTA

TACG<mark>ATCG</mark>

TCGTTACGTA

TCGTTACGTATG

C<mark>ACATCGTTACG</mark>

<mark>CGTTACGTATGG</mark>

ATCGTTACGTAT

CGCATATACGCG

T<mark>AT</mark>A<mark>G</mark>CGCGC<mark>TA</mark>

 \Rightarrow works on toy data

Weeder Results II



supersimple5	
supersimple3	
supersimple4	
supersimple7	
supersimple6	
supersimple1	
supersimple3	
supersimple2	
supersimple9	
supersimple8	
supersimple33	
supersimple28	
supersimple29	
supersimple32	
supersimple24	
supersimple25	
supersimple26	
supersimple27	
supersimple20	<u> </u>
supersimple21	
supersimple22	
supersimple22 supersimple23	
supersimple22 supersimple23 supersimple30	
supersimple22 supersimple23 supersimple30 supersimple37	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34 supersimple31	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34 supersimple31 supersimple15	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34 supersimple31 supersimple15 supersimple14	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34 supersimple31 supersimple15 supersimple14 supersimple17	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34 supersimple31 supersimple15 supersimple14 supersimple16	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34 supersimple31 supersimple15 supersimple14 supersimple16 supersimple11	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34 supersimple31 supersimple15 supersimple14 supersimple16 supersimple10	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple31 supersimple15 supersimple14 supersimple16 supersimple10 supersimple13	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple31 supersimple15 supersimple14 supersimple17 supersimple16 supersimple11 supersimple13 supersimple12	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple35 supersimple34 supersimple31 supersimple14 supersimple17 supersimple16 supersimple10 supersimple13 supersimple12 supersimple39	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple35 supersimple34 supersimple31 supersimple14 supersimple17 supersimple16 supersimple10 supersimple13 supersimple12 supersimple38	
supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34 supersimple31 supersimple15 supersimple17 supersimple16 supersimple10 supersimple13 supersimple12 supersimple39 supersimple38 supersimple19	

WEEDER RESULTS II

FIRST

supersimple5 supersimple4 supersimple7 supersimple6 supersimple1 supersimple3 supersimple2 supersimple9 supersimple8 supersimple33 supersimple28 supersimple29 supersimple32 supersimple24 supersimple25 supersimple26 supersimple27 supersimple20 supersimple21 supersimple22 supersimple23 supersimple30 supersimple37 supersimple36 supersimple35 supersimple34 supersimple31 supersimple15 supersimple14 supersimple17 supersimple16 supersimple11 supersimple10 supersimple13 supersimple12 supersimple39 supersimple38 supersimple19 supersimple18





- 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



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



- **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



NM_003688	[ਗ਼ਗ਼ [੶] ੑੑੑੑੑੑੑੑੑ੶ਗ਼ਗ਼੶ਗ਼੶੶੶ਗ਼੶ਗ਼੶ਗ਼੶੶੶੶੶੶੶੶੶੶੶੶੶
U97519	
NM_002859	┝┉╗┥╄╍┯╌┍╌╾╌┰╔┉╔┟┎┲╍╌╌╌╌╌┰┲┰┲┰╌╌╌╌┎┍┲┲┲╍╌╌╌╌┙╓╧╍╌╌╱ <mark>┈╢┈</mark> ┝╢ <mark>╏╢╢╗</mark> ┝╱╋ <mark>┨╔╌┉╖╢╢╽╢┿┲┈</mark> ╗║┫ [╝] ╲╢╖╴╴╴
NM_001904	<u>└────────────────────────────────────</u>
NM_001903	<u>᠆᠆᠇᠆ᢇᢉᡊᡀᢛᡊᢐ᠆ᡎᡊ᠆᠆ᡎ᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆</u> ᢉ᠓᠋ᠧᡪᠰ᠆ᡎ᠆᠆᠆ <mark>᠆</mark> ᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆᠆
D89980	
NM_032531	<u>┟┲┲┲┯┯┲┲┲┉╱╌</u> ┫ <mark>╹╱╌┨╹╌╱┫╲╌╌┨╹┍╱╢╔╎╔╎╔╢╔┦╔╝╱┈╝╱╌╱╴</mark> ╗┍┷╔╋╢╔╝╝╖┲ ^{╼╱╱┲} ╲ <mark>╔╲╌┦╲</mark> ╲╢╽┝╱┟╲╅┍╝╽╲╸ <mark>╴┈╔╴</mark> ╢┍╢
NM_004621	<u>└ᡣ╱╴┫╱╢╱╶┰╍╈┙┨╾┉╾┥┰┲┚┺╱╱┉╌╫╓╱╌╓╶┰╌┯╌┲┲┲┲╼┙┎╢╢┉╖╢┝┉╓╢╟┉╢┈╚┲┉┲┲┉┲╤┉╾┲╢┉╟┉┠╓╟┙</u> ┎┲╱╌┲┷╱╌╍ <u>┎┠╖╱╌┯┙╢┡╇╶╢</u>
NM_014954	
NM_002204	
NM_016081	╧╢╗┶┲┶┶┶┶┶┶┶┶┶┶┶┶┶┶┶┶┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙
NM_004517	
AY243535	┟ <u>┲╔╾╗</u> ╗╋╤╾┲┉┲╲┲┲╊╌┉╾╤┉ <mark>╊┉╔┉╔┉╔┉╔┉╔┉╔┉┉┉┉┍╖┙<mark>╊</mark>╼╔┉┉┉┍╖<mark>╱</mark>╖┉┉┉┍╖<mark>┢</mark>┝╓┉┉┉┉┍╖<mark>┢</mark>╖┈┝╖╔┉╌╻╍┠┰</mark>
AF035835	
AJ279254	
AF078828	<u>└₼₽₩₩₩₽₽₽₽₽₽₩₽₩₽₩₽₩₽₩₽₩₽₩₽₩₽₩₽₩₽₩₽₩₽₩₽₩₽</u>
NM_003870	
M33308	╱ <mark>╹╹^{╶╸}╹╶║</mark> ╹ <mark>┣╺┰╌┰╗╔┎╌╗╴╖╖╔┰╼┎╌╖┶┰╌╖╖╖┙╖╴╖╌╖╴╖╖┙╖╴╖╖┫╌╌╌╓╴╱╖╖╴╹╖╌╻╢╖╢╖╢╖╌╖╖╢╖┯╌╌╽╢╹╢╢╢╹<mark>╵╴╴╴╸</mark>╽╩╶<mark>╵</mark>╢</mark>
AF434715	╧╨╌╱╢╲╌╌╖╎┟╌┼┡╌┙┾╅╱┝╾┉╅┙╢┉╠┉╱┝╼┉┽┥╪╫╌╬╌╬╌╬╢╢╗╌╱╢╴╌┟╴╌╖╴╢┙╌╢┙╱╢╴╱┟╴╌╴
XIVI_290903	<u>॑ᡧᠲᡎᡲᡭ᠆᠆᠆᠆ᡏᠲ᠇᠆᠆᠆ᡍᡎ᠆᠆ᡍ᠆᠆ᢔ᠆ᢉᡛ᠇᠋᠋᠋ᡟᢔᡜᢜᢊ᠋᠋ᡀᠧᢇ᠇᠆᠆᠆᠆᠆᠆᠆᠆᠃᠆᠆ᡢᡢ᠋ᡀᡳ᠆ᡳᡢ᠓ᡎ᠆᠆᠆᠃᠋ᡎᢔ᠆ᢛ᠆ᡎ᠋᠋ᢚ᠆᠓᠆᠆᠃᠆᠁ᠺᢛ᠆ᠱᢤ᠆᠆ᡎᢔᡆ</u>
U80754	┢┷╍╍╍╫╗┙╍┲┙┇┚╢╢┙╗╌╗╴╗╴╗╴╗╴╗╴╗╴╗╴┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙┙
NIVI_002866	└──────────────────────────────────────
BC013903	
NM 002257	<u>└────────────────────────────────────</u>
NIVI_003257	
NM 004525	<u>┢┲╨╌┰╹┝┉╱┯┰╱╶┲╲┉┈┯╪╤╌┲┉╢╌┟┶╖╓┈╱┯┈╱┿╅┙┥┈╱╢╖╉╌┈┢╌┈╹┝┯┲┉┚┈╱╌╢╖┰╓╌╖╖╢╖┰┶╍┾╓┉╟┈╱┞╱║╍</u> ┲┅┉ <u></u>
NM 012120	
NM 005721	
NM 005721	
ΔΥ017369	
AF035771	
BC008799	
NM 007286	┟┉┉┽┿╫╫┯╢┝┥╲┉╫╿┉┉┟┶┝┈┉╫┥║┉╨┺╠╡┵╸╢┝╎┠╴┉┈┯┿┉┈╢║ <mark>╿</mark> ┉┉┿╸╺┿╾╸╸┝┈╺╢┽┝╪┉╲╽╫┿┾┍┿┿╾┿╫┝┈┉┽╢╽╴┶
1 19711	
NM 007124	
U20489	

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)



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