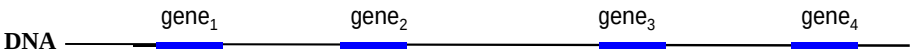
Discovery of transcription factor binding motifs from large sequence sets

Esko Ukkonen
University of Helsinki
ABS4NGS Workshop 23 June 2015 Paris

Outline

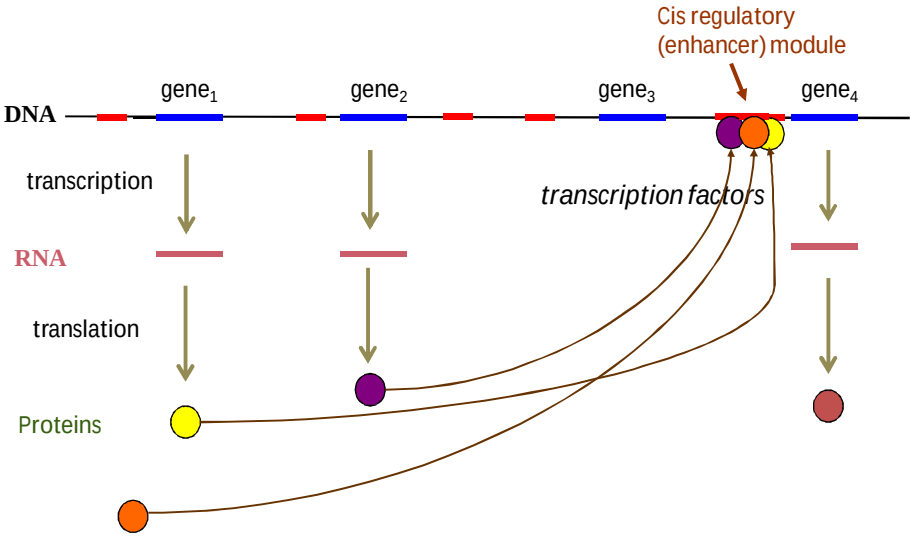
1. *Cis* regulatory modules
2. Alignment-based approach: EEL
3. Alignment-free approach
4. HT-Selex
5. Dimeric binding models from Selex data

Gene regulatory modules (cis-regulation)



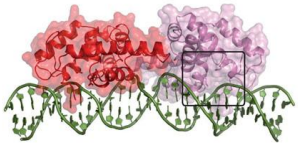
3

Gene regulatory modules (cis- regulation)

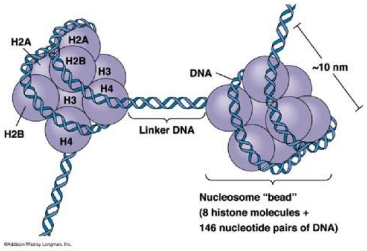


4

Structure of regulatory module: dimer, nucleosome, …



Structure of the moss LEAFY dimer bound to DNA sequence. (Sayou et al, Science 343 (7 Feb 2014)).



Characterization of a regulatory module?

- A regulatory module (*cis-regulatory module*) is a collection of transcription factor (TF) binding sites on DNA; no precise definition available
- properties of a module:
  - consists of several good binding sites of TFs (TF-DNA interaction)
  - the sites are spatially clustered together (TF-TF interaction)
  - the pattern of sites is conserved

6

PWM model of TF-DNA binding

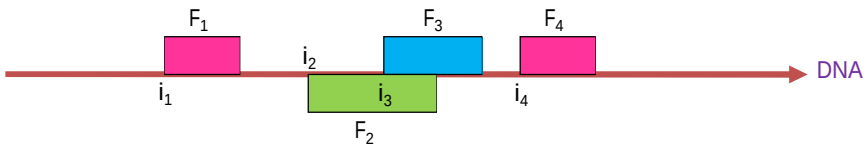
A	9	11	49	51	0	1	1	4
C	19	3	0	0	0	45	25	16
G	5	1	2	0	17	0	4	21
T	18	36	0	0	34	5	21	10

Fast search of PWM hits:

C Pizzi, P Rastas & E Ukkonen: Finding Significant Matches of Position Weight Matrices in Linear Time. *IEEE/ACM Trans. Comput. Biology Bioinform* (2011)

J Korhonen et al: MOODS: ... , *Bioinformatics* (2010)

Module structure



$(i_1, F_1, \text{Nor}), (i_2, F_2, \text{Rev}), (i_3, F_3, \text{Nor}), (i_4, F_4 = F_1, \text{Nor})$

OBS: Overlaps are allowed

8

Chain model for regulatory modules

- Problem: Given a DNA sequence, predict the regulatory modules on it.
- Formalization: Given a DNA sequence $s = s_1 s_2 \ldots s_n$ and a set T of transcription factors, a regulatory module of s is sequence

$$(i_1, F_1, o_1), (i_2, F_2, o_2), \ \ldots \ , (i_t, F_t, o_t)$$

chain model

where

$1 \leq i_1 < i_2 \ldots < i_t < n$,

each $F_i \in T$ and each $o_i \in$ {Normal,Reversed}

9

Predicting regulatory modules

A. Alignment-based approach: find conserved motifs (chains) of TF binding sites by comparing the DNA of different organisms by alignment

B. Alignment-free:
   - find over-represented TF binding-site pairs (dimers) from Selex data
   - Combine dimer models into a 'scanner' that finds chains of sites

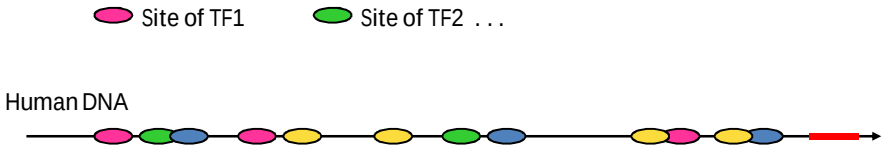
10

A. Finding conserved chains of TF binding sites – *alignment-based approach*

- looking at one genome gives too many positives
- Comparative (conservation-based) approach:
  - take the regions surrounding the same genes (paralogs and orthologs) of different species: human, mouse, chicken, ...
  - Align and find preserved clusters (motifs) of binding sites
- Smith-Waterman type algorithm with a novel scoring function

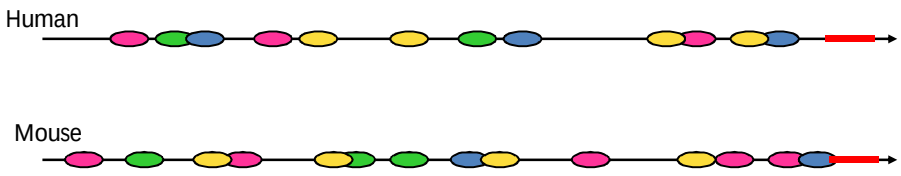
11

Good binding sites



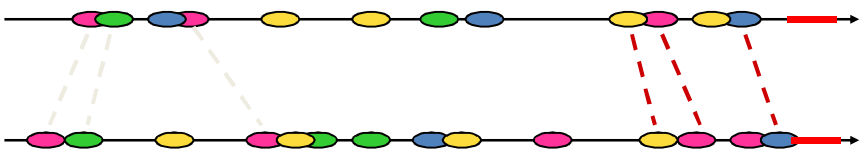
12

Clustering and conservation



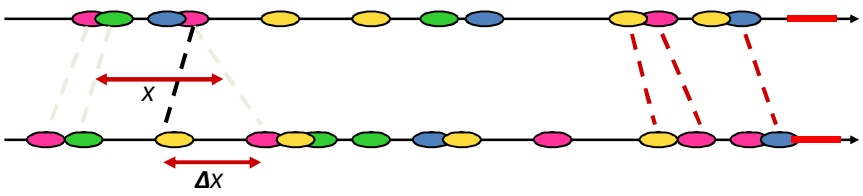
13

Clustering and conservation



14

Alignment scoring



High binding affinity at the aligned sites: bonus

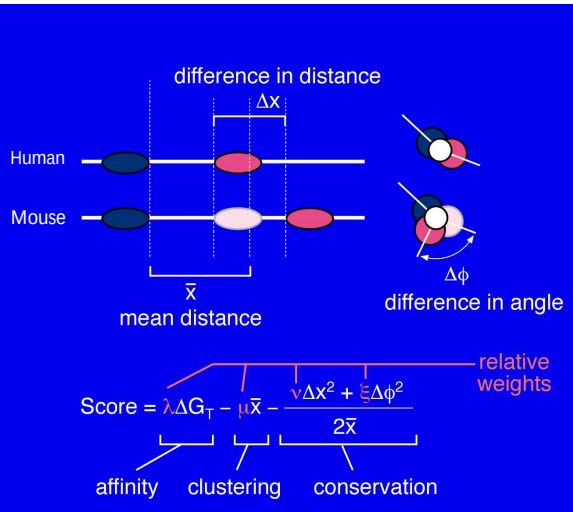
Long distance *x* between aligned sites: penalty

Non-conserved distance *(Δx > 0)* between aligned pairs: penalty

15

Scoring function for alignments

- Preserved in evolution:
  - Affinities of functional cis-elements.
  - Spatial arrangement of elements within a module.



16

Dynamic programming

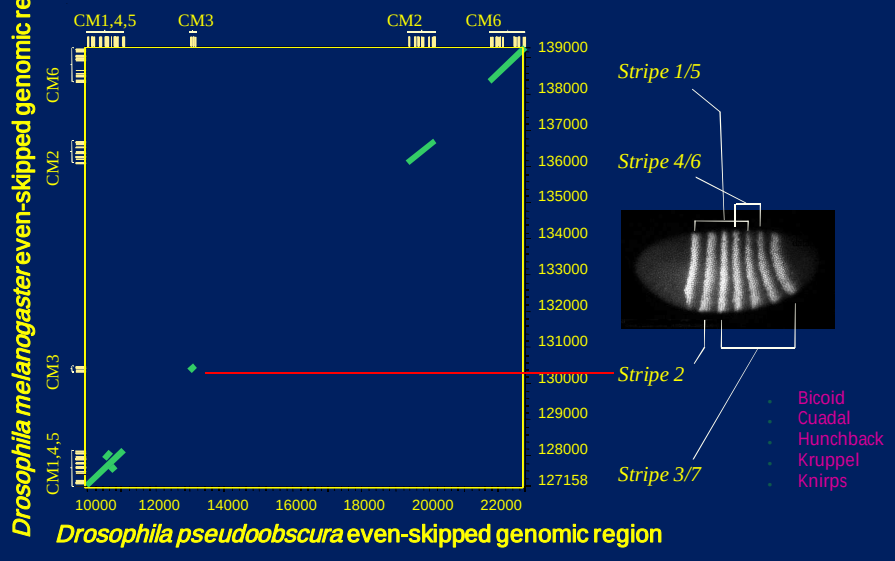
$$D_{ij} = \begin{cases} \max \{ \lambda w_{ij},\ D_{k,l} + \lambda w_{ij} - F(p_i - q_k\ ,\ p'_j - q'_l\ ) \\ \quad | \quad 0 < p_i - q_k < 1000,\ 0 < p'_j - q'_l < 1000 \}, \\ \quad \text{if } f_i = f'_j \text{ (i.e., the same TF aligned)} \\ -\infty, \quad \text{otherwise} \end{cases}$$

$O(n^4)$

w_{ij} = sum of the binding affinities of the sites of the TF at i and j in the two sequences

$F(\Delta i, \Delta j)$ = penalty for the non-conservation and the length of the distances between adjacent sites

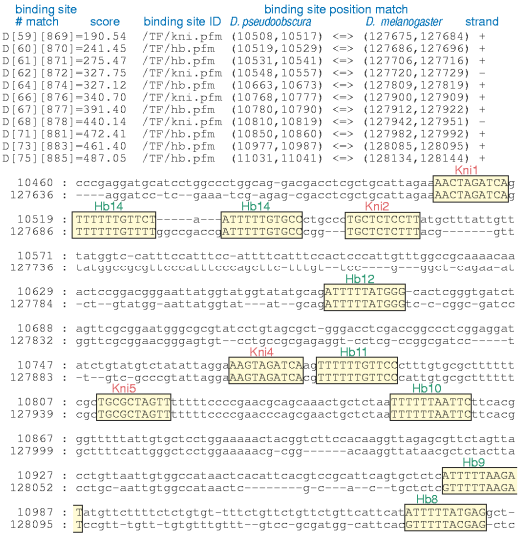
Enhancer prediction in *Drosophila*



18

Drosophila enhancer by EEL

- Output from Enhancer Element Locator (EEL) program: *Drosophila* even-skipped gene stripe 2 enhancer
- Hallikas, Palin *et al*, Genome-wide prediction …, *Cell* 124 (2006), 47-59.
- Palin, Taipale, Ukkonen, *Nature Protocols* (2006)
- Tuupanen et al, *Nature Genetics* (2009)

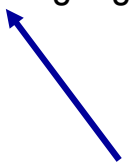


19

Disease-associated rSNP: colorectal cancer (Tuupanen et al, *Nat Gen* 2009)

Disease-associated rSNP: colorectal cancer

AGATAAGATAATGTAGTCAAAGG/Tgcaggttaacc



Homozygosity for the G allele of rs6983267 at 8q24 increases colorectal cancer risk approximately 1.5 fold.

Europe: 46% of people have G

China: 39% of people have G

SNP rs6983267 occurs in an EEL module

Alignment-free approach: model the local steps of regulatory modules

Scoring of regulatory modules

$$\text{Score(Regulatory module)} = \sum_{\text{chain}} \text{TF•DNA interaction} + \sum_{\text{chain}} \text{TF•TF interaction}$$

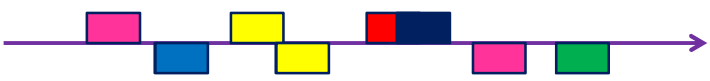
- TF•DNA interaction score of $(i_j, F_j, o_j) \approx$
  DNA binding affinity of factor F_j in orientation o_j at location i_j of s
- TF•TF interaction score of $((i_j, F_j, o_j), (i_{j+1}, F_{j+1}, o_{j+1})) \approx$
  co-operative binding of factors F_j and F_{j+1} in orientation (o_j, o_{j+1}) and at distance $i_{j+1} - i_j$.

23

High-scoring modules

- Find the highest scoring regulatory module of s
  - Highest scoring subsequence -> Smith-Waterman style dynamic programming (+ Karlin-Altschul conditions)

$$W(i,F,o) = \max_{i'<i,F',o'} \begin{cases} W(i',F',o') + \text{TF•DNA}(i,F,o) + \text{TF•TF}(F,o,F',o',i-i') \\ \text{TF•DNA}(i,F,o) \end{cases}$$

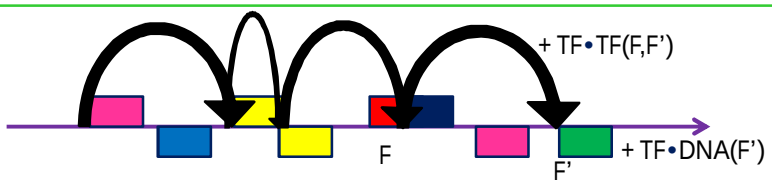


24

High-scoring modules

- Find the highest scoring module of *s*
  - Highest scoring subsequence -> Smith-Waterman style dynamic programming (Karlin-Altschul conditions)

$$W(i,F,o) = \max_{i'<i,F',o'} \begin{cases} W(i',F',o') + \text{TF•DNA}(i,F,o) + \text{TF•TF}(F,o,F',o',i-i') \\ \text{TF•DNA}(i,F,o) \end{cases}$$

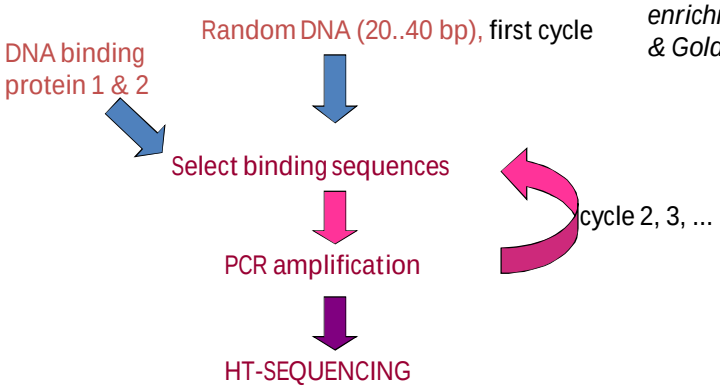


25

SELEX

SELEX protocol

Systematic Evolution of Ligands by EXponential enrichment (Turk & Gold 1990)



A. Jolma, T. Kivioja et al.: Multiplexed massively parallel SELEX ..., *Genome Res.* 2010

27

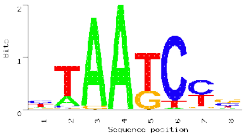
SELEX data

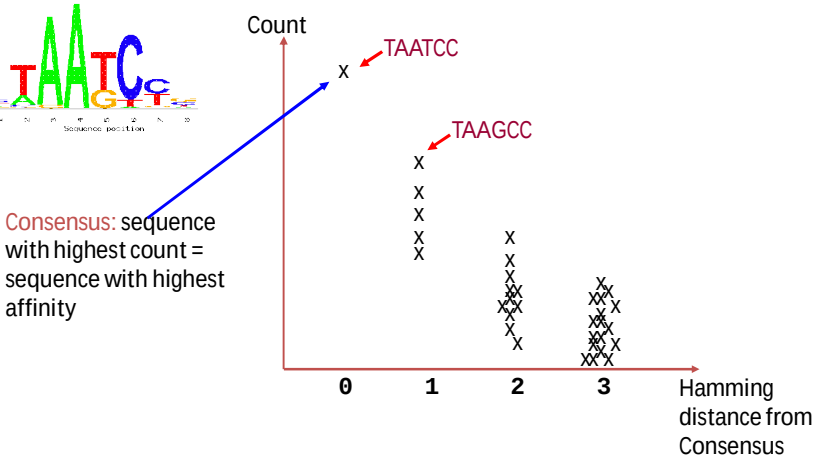
- Input
  - Transcription factor(s)
  - Pool of oligonucleotides containing 40 bp of *randomized sequence* flanked by constant regions
- 3-6 selection cycles
- Output
  - On average 10 000 – 150 000 sequencing reads per cycle, selected according to the binding affinity to the TF or a pair of TFs

TTCACAGCTCTCAAGGACTTTTCACAGCTCTCAAGGACTT

28

Enrichment by SELEX

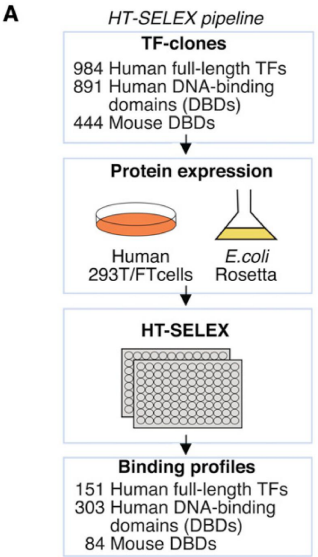




29

HT-Selex for TFs and TF-TF pairs (Taipale lab)

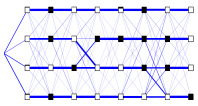
- Hundreds of TFs
- Thousends of pairs of TFs



Jolma et al, Cell 152, Jan 2013)

TF-DNA and TF-TF binding affinity models using Selex data

TF-DNA binding affinity models

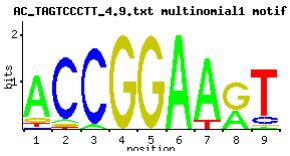
- Here: Classic Position Weight Matrix (PWM)
- But as we have big data, also possible
  - HMM 
  - Models with long-distance dependencies
  - Plain list of highly affine subsequences

k-mer cover as a TF_DNA binding affinity model

9-mer	count	cumulative count	%
ACCGGAAGT	25359	25359	15.1
ACCGGAAAT	15703	41062	24.5
ACCGGAAAC	6611	47673	28.4
CGGATATCC	6281	53954	32.2
ACCGGAAGC	3791	57745	34.4
CGGATTTCC	3500	61245	36.5
ACATCCGGT	3204	64449	38.4
ACTTCCGGC	2833	67282	40.1
ACTTCCGGA	1985	69267	41.3
CATTTCCGG	1859	71126	42.4
ACAGGAAGT	1191	72317	43.1
AACGGAAGT	1114	73431	43.8
ACTTCCGCT	1028	74459	44.4
ACCGGAAGG	1007	75466	45.0
ACTTCCGGG	856	76322	45.5

EWS–FLI1

(Greedy) covering
by 9-mers
167789 lines from SELEX

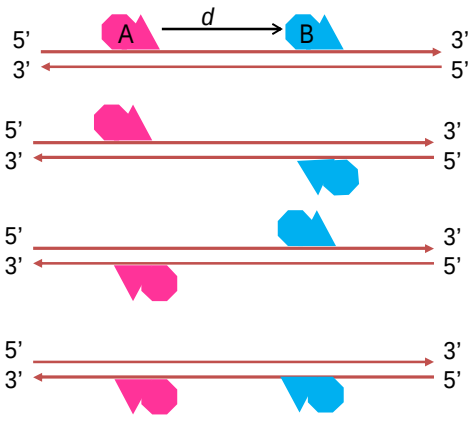


Dimer structure: orientations (4) and distances

A: binding sequence
A': reverse complement of A

d = distance between components (negative if overlap)

OBS: only 3 types for homodimer A = B



A B
B' A'
Type I: HT

A B'
B A'
Type II: HH

A' B
B' A
Type III: TT

A' B'
B A
Type IV: TT

34

Model structure: homodimer

- R: transcription factor λ: mixing parameters
- Probability of a Selex read X as a mixture of events
  - Binding site of monomer R occurs in X
  - Binding site of a homodimer (R,R) occurs in X
    - *O*: relative orientation (out of three possibilities) of (R,R)
    - *d*: distance in [min, max] between dimer components in X
  - q: X is from background

=> Probability of X:

$$P(X) = \lambda_R P_R(X) + \sum_{o,d} \lambda_{R,R,o,d} P_{R,R,o,d}(X) + \lambda_q P_q(X)$$

Model structure: heterodimer

- R, S: transcription factors λ: mixing parameters
- Probability of a Selex read X as a mixture of events
  - Binding site of monomer R occurs in X; binding site of S occurs in X
  - Binding site of a homodimer (R,R) [homodimer (S,S)] occurs in X
    - *O*: relative orientation (out of *three* possibilities) of (R,R) [of (S,S)]
    - *d*: distance in [min, max] between dimer components in X
  - Binding site of a heterodimer (R,S) occurs in X
    - *O*: relative orientation (out of *four* possibilities) of (R,S)
    - *d*: distance in [min, max] between dimer components in X
  - q: X is from background

=> Probability of X:

$$P(X) = \lambda_R P_R(X) + \lambda_S P_S(X)$$
$$+ \sum_{o,d} \lambda_{R,R,o,d} P_{R,R,o,d}(X) + \sum_{o,d} \lambda_{S,S,o,d} P_{S,S,o,d}(X) + \sum_{o,d} \lambda_{R,S,o,d} P_{R,S,o,d}(X) + \lambda_q P_q(X)$$

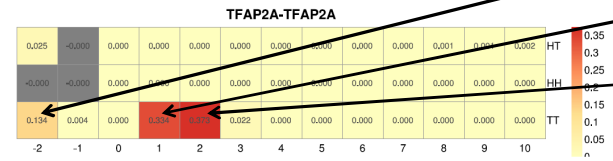
Model learning

- Maximum likelihood model by EM algorithm (c.f. MEME)
  - PWM learned for each monomer (possibly several variants) and each dimer variant specified by *o* and *d*
  - Values for the mixing parameters λ
- Good initial values essential: learn monomer models separately and use them as seeds for the general case (pipeline still under development)
- To improve convergence trim away obviously uninformative cases of dimers

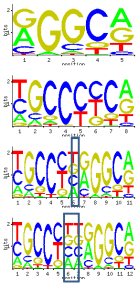
Example: TFAP2A homodimer

- Result file: ../seed_GGGCA_learning_dist_1.out
- EM-algorithm took 30 iterations
- Convergence criterion cutoff is 0.010000
- Excluded cob cases: 3
- Initial and final seeds:
  - GGGCA
  - GGGCA

Model	Lambda
TFAP2A	0.0518912
none	0.0
TFAP2A-TFAP2A	0.8974
none	0.0
none	0.0
Background	0.050747
Sum	1.0000382



Seed: GGGCA



TFAP2A

TT -2

TT 1

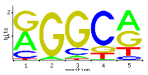
TT 2

Example: TFAP2A homodimer - validation

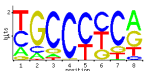
- Result file: ../seed_GGGCA_learning_dist_1.out
- EM-algorithm took 30 iterations
- Convergence criterion cutoff is 0.010000
- Excluded cob cases: 3
- Initial and final seeds:
  - GGGCA
  - GGGCA

Model	Lambda
TFAP2A	0.0518912
none	0.0
TFAP2A-TFAP2A	0.8974
none	0.0
none	0.0
Background	0.050747
Sum	1.0000382

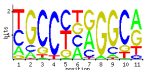
Seed: GGGCA



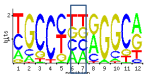
TFAP2A



TT -2

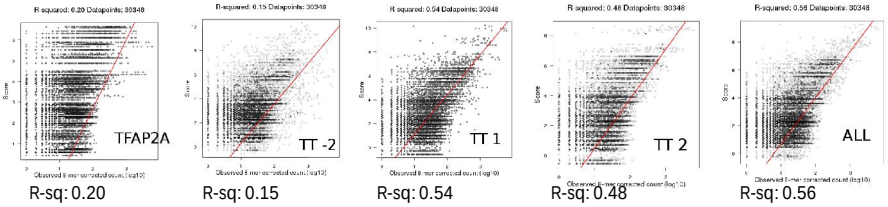


TT 1



TT 2

Explaining the enrichment of gapped 8-mers (4-mer -gap- 4-mer). Red lines indicate least-squares fit. Plots in all panels have logarithmic axes to facilitate visualization; the R2 values are from the corresponding linear data.

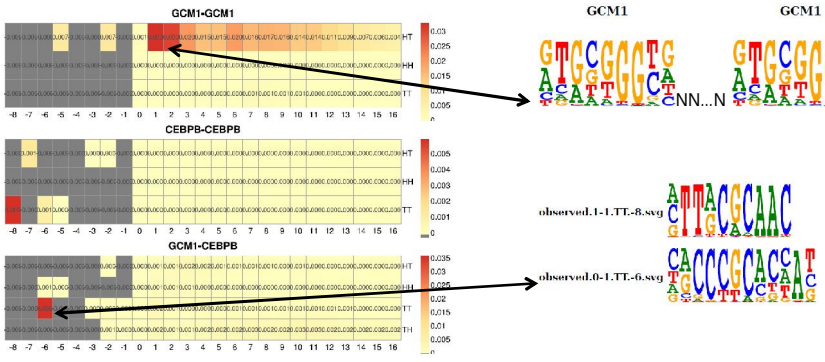


R-sq: 0.20 R-sq: 0.15 R-sq: 0.54 R-sq: 0.48 R-sq: 0.56

Example: GCM1-CEBPB heterodimer

- Result file: ../dinuc_all_bg.out
- EM-algorithm took 33 iterations
- Convergence criterion cutoff is 0.010000
- Excluded cob cases: 63
- Initial and final seeds:
  - ATGCGGGTA TTRCGCAAY
  - ATGCGGGTA TTRCGCAAY

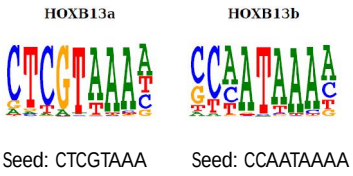
Model	Lambda
GCM1	0.0223766
CEBPB	0.0158677
GCM1-GCM1	0.2654
CEBPB-CEBPB	0.0088
GCM1-CEBPB	0.0972
Background	0.590389
Sum	1.0000333



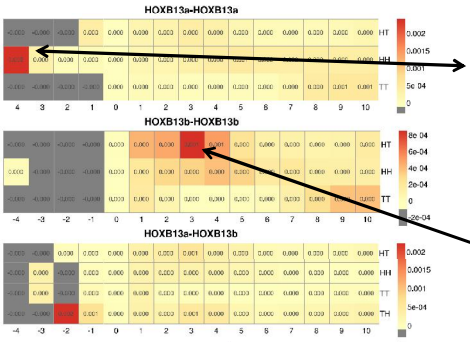
Example: HOXB13 – a monomer with several PWMs

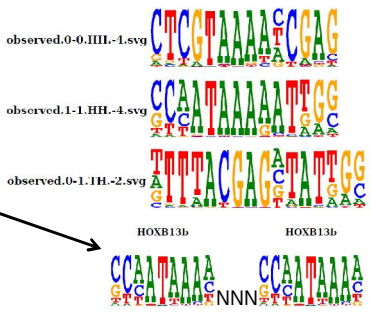
- Result file: ../testing_for_article2.out
- EM-algorithm took 36 iterations
- Convergence criterion cutoff is 0.010000
- Excluded cob cases: 26
- Initial and final seeds:
  - CTCGTAAAA CCAATAAAA
  - CTCGTAAAA CCAATAAAA

Model	Lambda
HOXB13a	0.350924
HOXB13b	0.139309
HOXB13a-HOXB13a	0.0114
HOXB13b-HOXB13b	0.0070
HOXB13a-HOXB13b	0.0108
Background	0.480642
Sum	1.000075

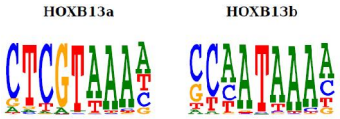


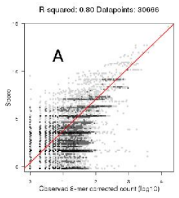
Seed: CTCGTAAA Seed: CCAATAAAA



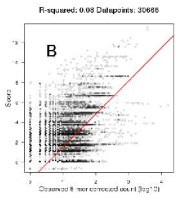


HOXB13: validation

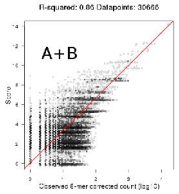




R-sq: 0.80

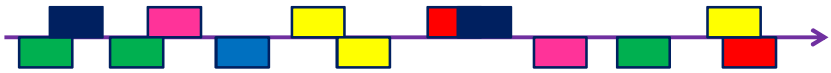


R-sq: 0.08



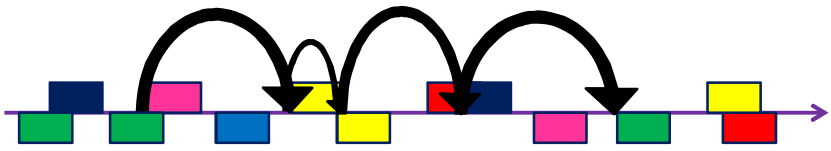
R-sq: 0.86

Scanner (HMM) for regulatory modules



43

Scanner (HMM) for regulatory modules



TF-TF models needed for selecting the best
steps through the available TF-DNA binding sites

44

Conclusion

- Validation (some preliminary results)
- A scanner that combines 'all' TF binding models is likely to produce too many false positives
- Cell-type specific subsets of TFs, nucleosome

- Collaborators: Jussi Taipale (KI), Jarkko Toivonen, Arttu Jolma, Teemu Kivioja, Kimmo Palin, Pasi Rastas

- Funding:

45