Finding regulatory elements shared by a set of genes

Matthieu Defrance - Hélène Touzet - LIFL

January 2006

Overview

- Motif over-representation in regulatory regions
- A fast algorithm to extract significant local over-representation
- Example of Rel/NF-κB target genes and Muscle specific genes

Biological questions

- Understanding gene transcriptional regulation in higher eukaryotes
- Detecting Transcription Factors involved in regulatory mechanisms

Over-represented motifs & regulation

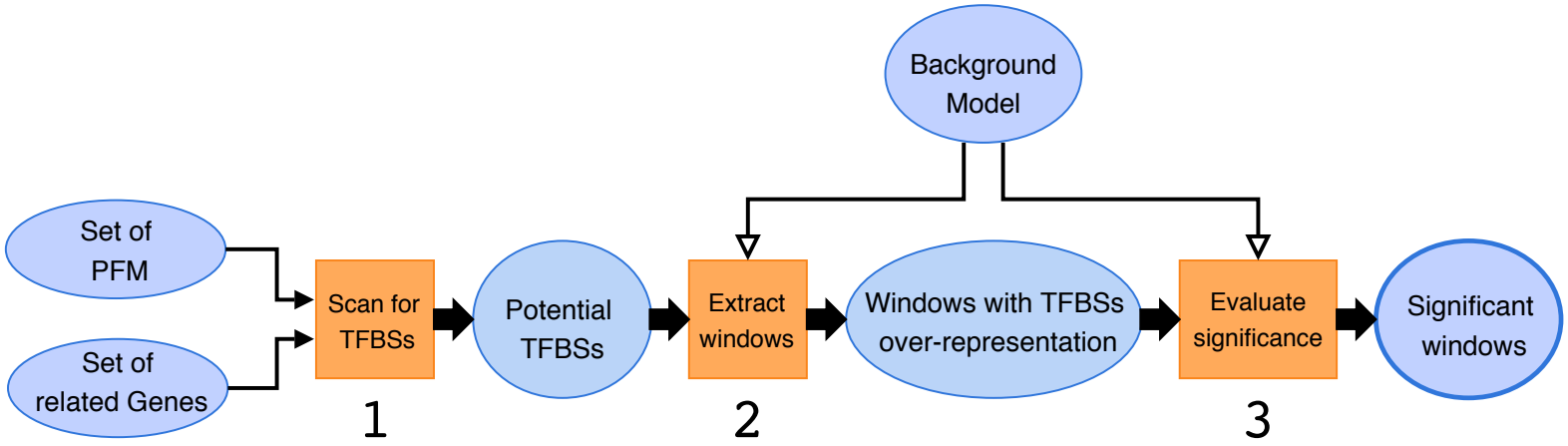
Hypothesis: over-represented motifs are involved in regulation.

- Working with a set of genes that (are assumed to) share regulatory mechanisms
  - Functionally related genes
  - Clusters of genes derived from DNA array analysis
  - ...
- A motif can be:
  - Aligo-nucleotide
  - Motif whose profile is known: Position Frequency Matrix, HMM, regular expression
- Need a background model to evaluate over-representation
  - Markov model
  - Empiric model

Finding motifs over-representation

- Regulatory motifs are highly degenerated in higher eukaryotes
- In order to provide accurate predictions we choose to:
  - Restrict motif search to known profiles
  - Use motif conservation across multiple species

Finding motifs over-representation: TFM-Explorer



1. **Scan for all potential TFBSs** (exhaustively)
2. **Extract regions** where predicted TFBSs are over-represented
3. **Evaluate significance** of extracted windows (P-value, E-value)

1. Scan for potential TFBSs

Input:

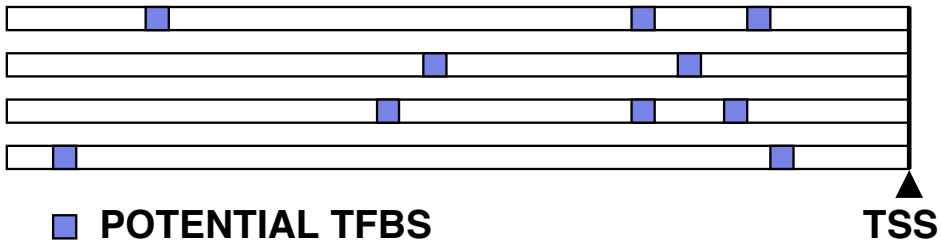
- Set of regulatory sequences
- Set of PFM (for example all TRANSFAC matrices)

Output:

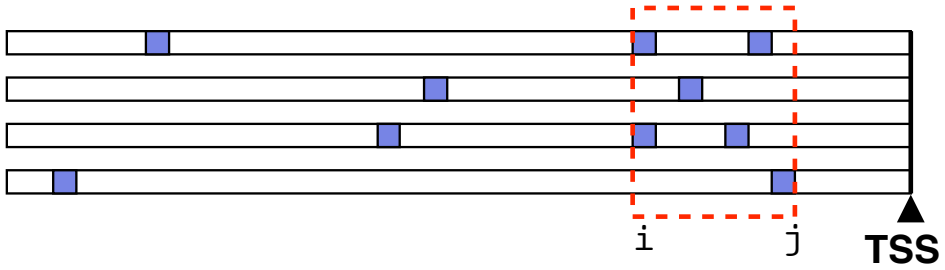
- Exhaustive list of potential Transcription Factor Binding Site
  - Overlapping sites (for a TF) are cut
  - Forward and reverse strands are scanned
  - Use an uniform cutting threshold based on P-value

2. Extract windows

- Sequences are aligned on Transcription Start Site
- All TFBSs found in the previous stage are used



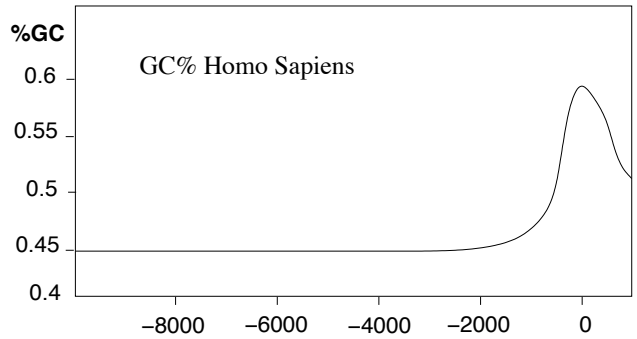
- Search for common regulatory elements in promoter regions of these genes?



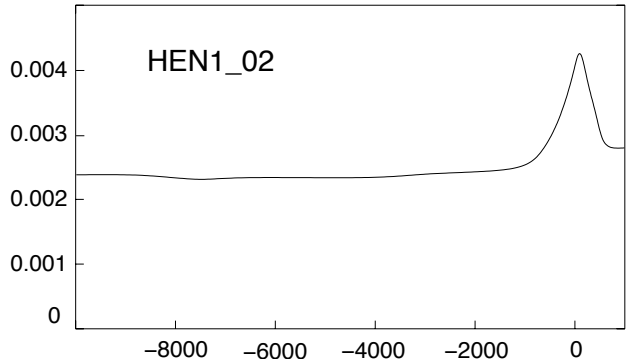
Detect regions where predicted binding sites are locally over-represented

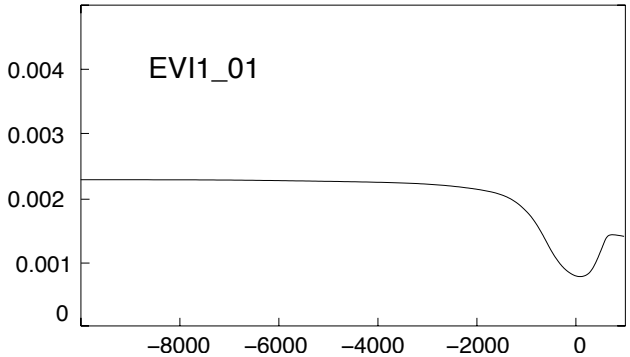
2. Extract windows (2) - Background Model?

GC% variations in regulatory region



Density of predicted hits depends on PFM (GC or AT rich) and position:

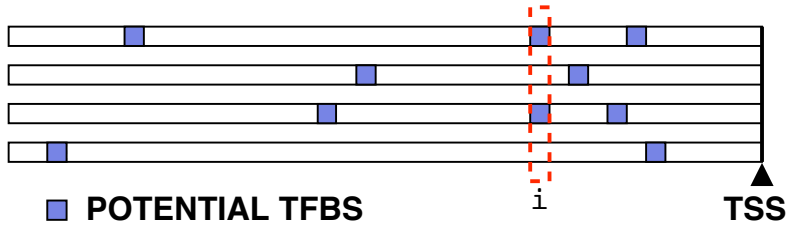




We need to determine a local distribution of predicted TFBSs for each matrix

2. Extract windows (3) - Scoring function

- Scan on the fly with a scoring function



$$s_i = k_i \ln \frac{\lambda_i}{\lambda_i^b} + |E|(\lambda_i^b - \lambda_i)$$

- E sequences set
- k_i number of hits at position i
- λ_i^b Poisson parameter in the background model at position i
- λ_i Poisson parameter in the expected model at position i

2. Extract windows (3) - Scoring function

Extension to heterogeneous sequences (multiple organisms)

$$s_i = \left(\frac{\sum_{j=1}^{|E|} \lambda_{i,j}}{\sum_{j=1}^{|E|} \lambda^b_{i,j}} \right)^{k_i} e^{|E| \left(\sum_{j=1}^{|E|} \lambda^b_{i,j} - \sum_{j=1}^{|E|} \lambda_{i,j} \right)}$$

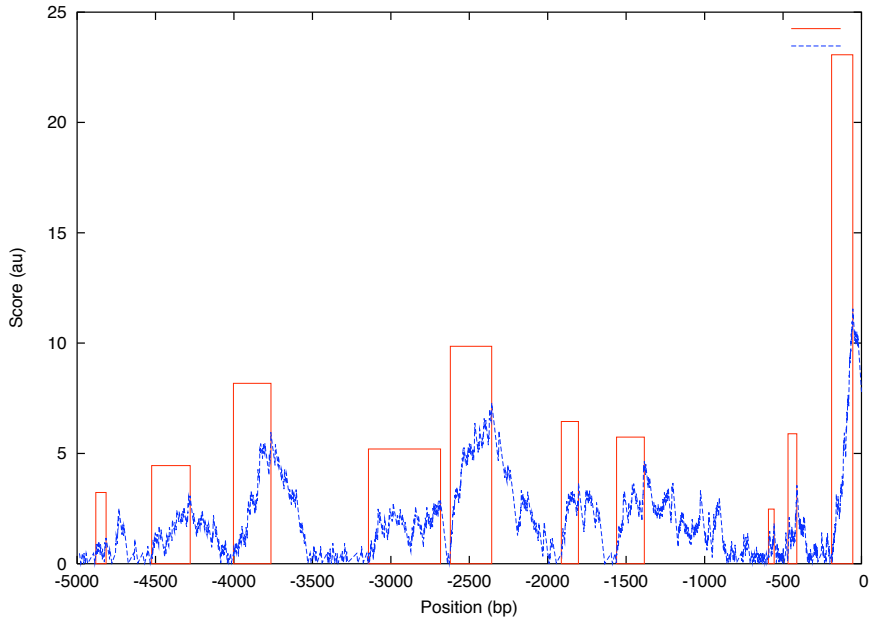
- E sequences set
- k_i number of hits at position i
- $\lambda^b_{i,j}$ Poisson parameter in the background model at position i for sequence j
- $\lambda_{i,j}$ Poisson parameter in the expected model at position i for sequence j

2. Extract windows (4) - Cumulative Score

- Global score S_i is cumulative and bounded by zero

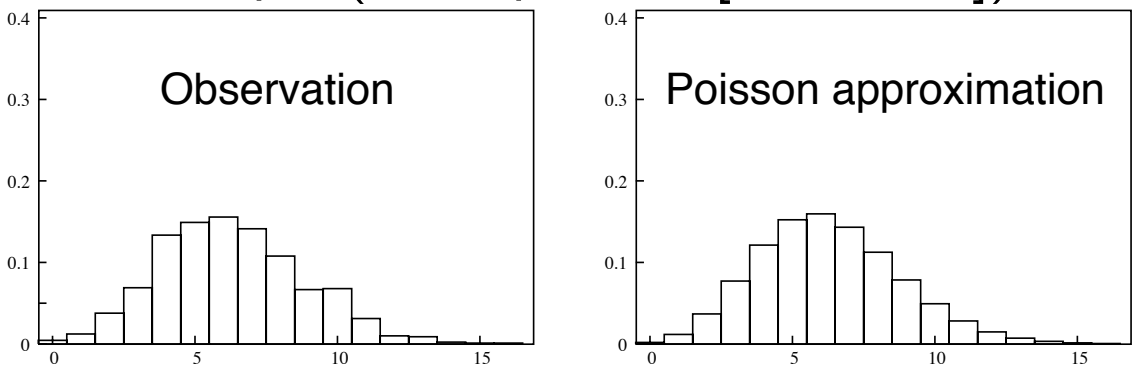
$$S_i = max \begin{cases} S_{i-1} + s(i) \\ 0 \end{cases}$$

- Extract regions with positive score without a priori knowledge on their size



3. Evaluate significance of window

- P-value: Probability $P(k, \alpha, \beta, N)$ to observe $X \geq k$ hits in window $[\alpha, \beta]$ for $|E|$ sequences
- Approximate hit count distribution in a window by a **Poisson** distribution
- IK3_01 PFM example (*20 sequences* $[-200 - 1]$):



(Hit count in window [-200,-1] for sets of 20 sequences)

3. Evaluate significance of window (2)

Heterogeneous sequences set case

$$P(X \geq k) = 1 - \sum_{z=0}^{k-1} \frac{(|w| \sum_{j=1}^{|E|} \lambda_s)^z}{z!} e^{-|w| \sum_{j=1}^{|E|} \lambda_j}$$

- E sequences set
- w window
- k number of hits in w ($[\alpha, \beta]$)
- λ_j Poisson parameter of his count distribution in w for the sequence j

3. Evaluate significance of window (3) - Limitations

Is this (Poisson approximation) correct for all matrices?

Fitting Observation/Poisson with χ^2 test			
α *(error type I)*	$[100\%, 5\%]$	$]5\%, 0.5\%]$	$]0.5\%, 0.0]$
% PFM	70%	10%	20%

Application example NF-κB target genes set

[Karo Gosselin, Corinne Abbadie IBL]

- NF-κB control expression of genes involved in
  - Inflammation and immunity
  - Stress responses, including apoptosis
- Gene set compiled from literature data
- Gene considered as true NF-κB target when
  - A NF-κB was map
  - The functionality of motif was experimentally validated

NF-κB target genes set (2)

- Set of 102 human genes
  - Promoter sequences retrieved from University California Santa Cruz Genome Browser (region $[-10000 + 1000]$)
  - List of NF-κB motif with exact position, exact sequence (for validation)
- Empiric Background model
- All TRANSFAC matrices

NF-κB target genes set (3) - Results of TFM-Explorer

Factor	Matrix ID	Location	Hits	Sequences	P-Value
TATA	V$TATA_01	[-0074:-0009]	042	042 (41.18%)	3.57e-14
NF-kappaB	V$NFKB_C	[-0507:-0016]	206	087 (85.29%)	4.89e-14
NF-kappaB	V$NFKAPPAB65_01	[-0520:-0013]	192	084 (82.35%)	1.93e-13
NF-kappaB	V$NFKAPPAB_01	[-0227:-0017]	112	075 (73.53%)	2.71e-13
NF-kappaB	V$NFKB_Q6	[-0230:-0020]	096	069 (67.65%)	3.12e-11
c-Rel	V$CREL_01	[-0511:-0017]	175	076 (74.51%)	5.37e-11
RREB-1	V$RREB1_01	[-4382:-3850]	246	089 (87.25%)	1.13e-10
TATA	V$TATA_C	[-0060:+0042]	039	035 (34.31%)	3.61e-10
NF-AT	V$NFAT_Q6	[-0251:-0016]	107	066 (64.71%)	4.34e-08
SRY	V$SRY_01	[-9915:-9736]	110	059 (57.84%)	1.00e-07
CdxA	V$CDXA_02	[-5849:-5523]	099	050 (49.02%)	2.23e-07

e-value < 1.0

TFM-Scan is able to delineate promoter areas that share relevant over-represented TF binding sites

Muscle data set

[Wasserman]

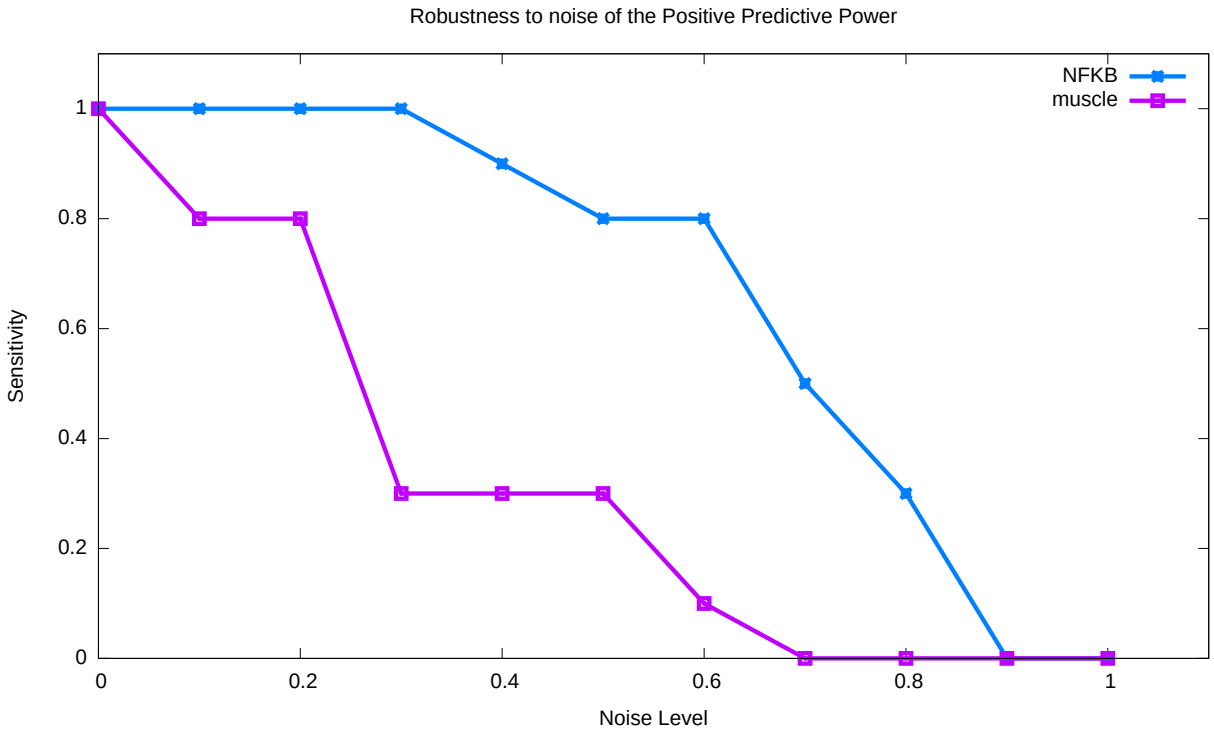
- Set of 27 genes that have skeletal muscle-specific expression (13 human, 17 mouse and 7 rat genes)
  - Promoter sequences retrieved from University California Santa Cruz Genome Browser (region $[-2000 + 200]$)
  - 5 factors that have muscle-specific expression are known: MyF, MEF-2, SRF, TEF (for validation)
- Empiric Background model
- All TRANSFAC matrices

Muscle data set (2) - Results

Rank	Transcription Factor (and PSSM)	Score
TFM-Explorer		
1	SRF* ([-0241:-0027])	6.459e-08
2	MyF* ([-0123:-0024])	9.965e-07
3	MEF2* ([-0073:-0026])	1.003e-05
4	p50 ([-0089:-0058])	3.637e-05
5	Hen-1 ([-0489:-0331])	1.498e-04
OTFBS		
1	MYOD_Q6*	3.076e-08
2	AP4_Q5	1.644e-07
3	TAL1BETAE47_01	1.468e-06
4	E47_01	3.599e-06
5	FOXJ2_01	5.798e-06
Toucan		
1	TGIF_01	3.002e-05 (2.29)
2	SRF_C*	7.748e-05 (1.878)
3	E47_02	2.194e-04 (1.426)
4	RFX1_02	2.462e-04 (1.386)
5	LMO2COM_01	3.232e-04 (1.258)
oPOSSUM		
1	MEF2*	1.663e-05
2	SRY*	4.190e-04
3	c-MYB_1	5.022e-04
4	S8	9.329e-04
5	SP1*	1.023e-03
6	Hen-1	1.034e-03

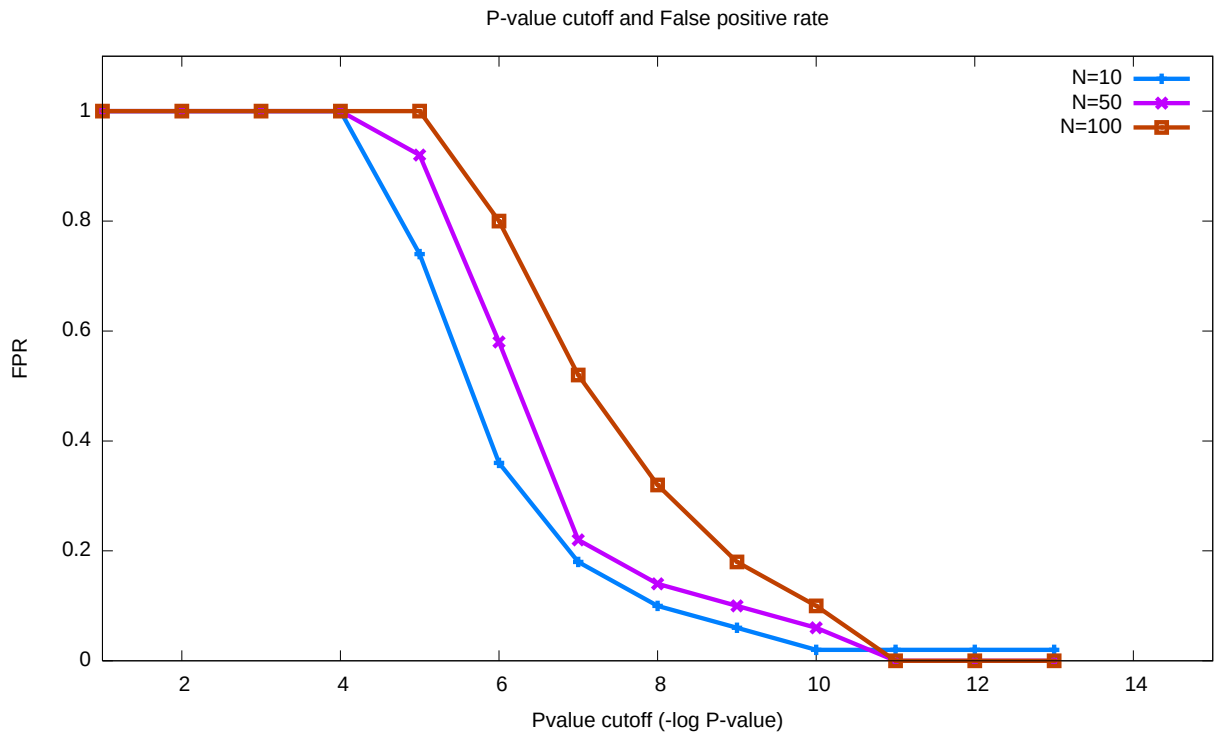
Rank	Transcription Factor (and PSSM)	Score
TFM-Explorer		
1	SRF* ([-0241:-0027])	6.459e-08
2	MyF* ([-0123:-0024])	9.965e-07
3	MEF2* ([-0073:-0026])	1.003e-05
4	p50 ([-0089:-0058])	3.637e-05
5	Ahr-ARNT ([-1342:-1268])	4.505e-05
OTFBS		
1	TAL1BETAITF2_01	3.244e-10
2	TAL1BETAE47_01	5.304e-09
3	YY1_02	1.506e-08
4	TAL1ALPHAE47_01	7.534e-08
5	AP4_Q5	3.401e-07
6	MYOD_Q6*	7.808e-07
Toucan		
1	E47_02	5.414e-02 (-0.969)
2	MEF2_01*	1.128e-01 (-1.288)
3	TAL1ALPHAE47_01	1.586e-01 (-1.436)
4	MEF2_02*	2.080e-01 (-1.554)
5	MEF2_03*	2.080e-01 (-1.554)
6	CEBP_C	2.196e-01 (-1.577)
oPOSSUM		
1	MEF2*	1.663e-05
2	SRY*	4.190e-04
3	c-MYB_1	5.022e-04
4	S8	9.329e-04
5	SP1*	1.023e-03
6	Hen-1	1.034e-03

Robustness to noise



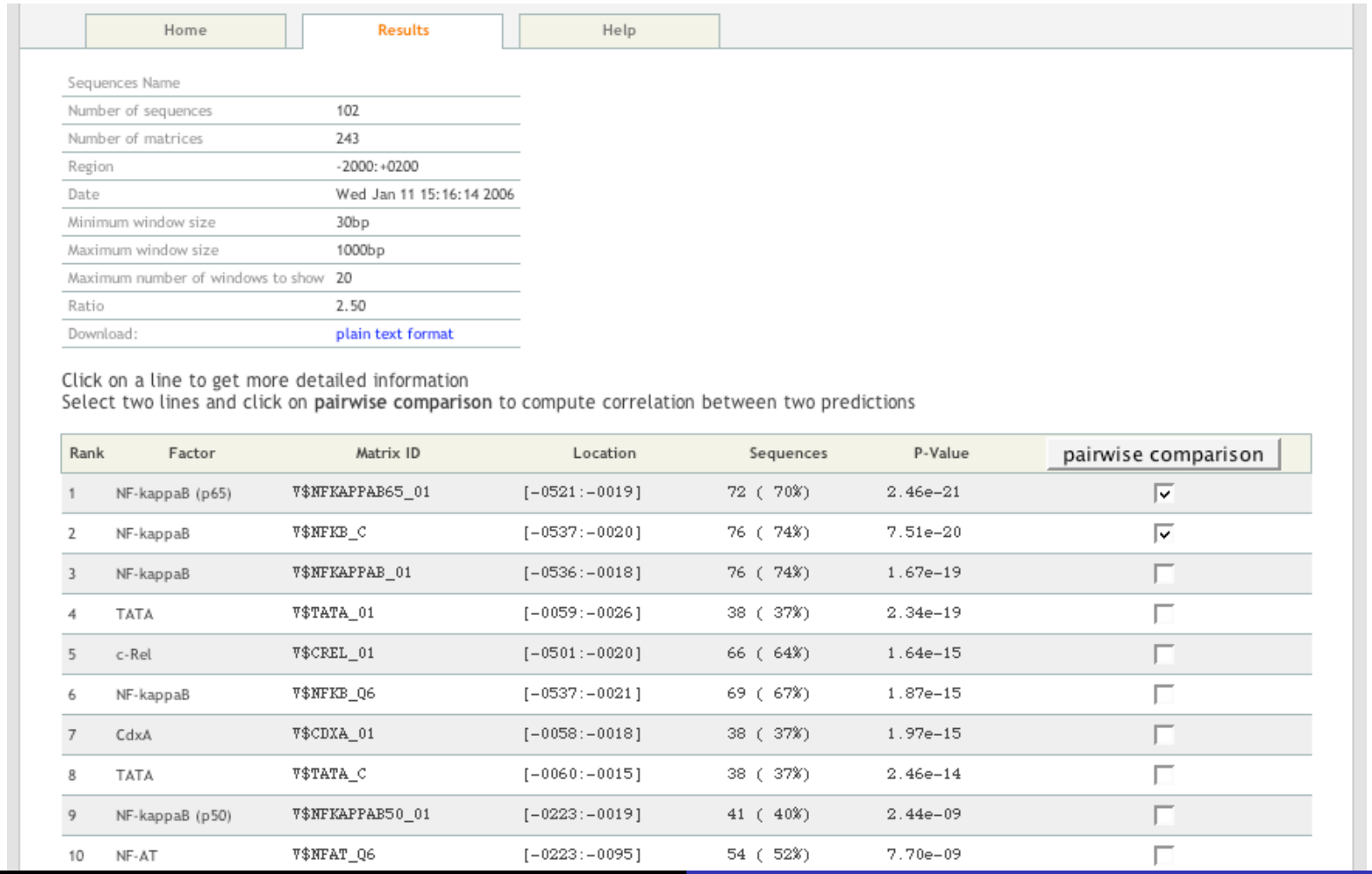
P-value cutoff and False Positive Rate

$$FPR = FP/Actual-$$



Web interface

http://bioinfo.lifl.fr



Home | Results | Help

Sequences Name	
Number of sequences	102
Number of matrices	243
Region	-2000:+0200
Date	Wed Jan 11 15:16:14 2006
Minimum window size	30bp
Maximum window size	1000bp
Maximum number of windows to show	20
Ratio	2.50
Download:	plain text format

Click on a line to get more detailed information
Select two lines and click on **pairwise comparison** to compute correlation between two predictions

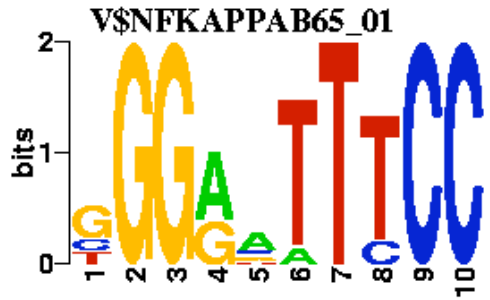
Rank	Factor	Matrix ID	Location	Sequences	P-Value	pairwise comparison
1	NF-kappaB (p65)	V$NFKAPPAB65_01	[-0521:-0019]	72 (70%)	2.46e-21	☑
2	NF-kappaB	V$NFKB_C	[-0537:-0020]	76 (74%)	7.51e-20	☑
3	NF-kappaB	V$NFKAPPAB_01	[-0536:-0018]	76 (74%)	1.67e-19	☐
4	TATA	V$TATA_01	[-0059:-0026]	38 (37%)	2.34e-19	☐
5	c-Rel	V$CREL_01	[-0501:-0020]	66 (64%)	1.64e-15	☐
6	NF-kappaB	V$NFKB_Q6	[-0537:-0021]	69 (67%)	1.87e-15	☐
7	CdxA	V$CDXA_01	[-0058:-0018]	38 (37%)	1.97e-15	☐
8	TATA	V$TATA_C	[-0060:-0015]	38 (37%)	2.46e-14	☐
9	NF-kappaB (p50)	V$NFKAPPAB50_01	[-0223:-0019]	41 (40%)	2.44e-09	☐
10	NF-AT	V$NFAT_Q6	[-0223:-0095]	54 (52%)	7.70e-09	☐

Web interface (2)

Full results for matrix V$NFKAPPAB65_01 and region [-0521:-0019]

Matrix information

Matrix name	V$NFKAPPAB65_01
Transcription Factor name	NF-kappaB (p65)
Source	TRANSFAC OR JASPAR
Information Content	14.76 bits

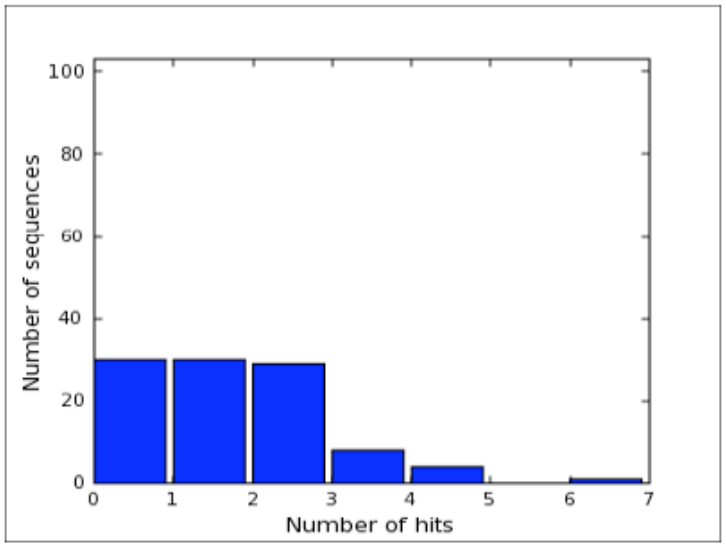


WebLogo representation of the matrix V$NFKAPPAB65_01 (NF-kappaB (p65))

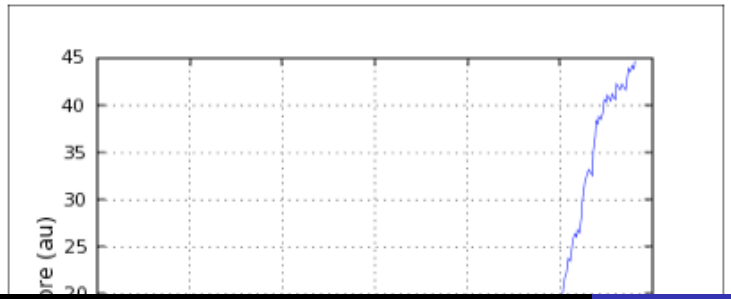
Web interface (3)

Sequences information

Total number of hits in region [-0521:-0019]	134
Total number of sequences	102 (2.61e-03 TFBS per sequences per bp)
Number of sequences having at least one putative TFBS in region [-0521:-0019]	72 (3.70e-03 per sequence per bp)
P-value	2.46e-21
E-value	2.81e-15



Hits per sequence distribution



Web interface (4)

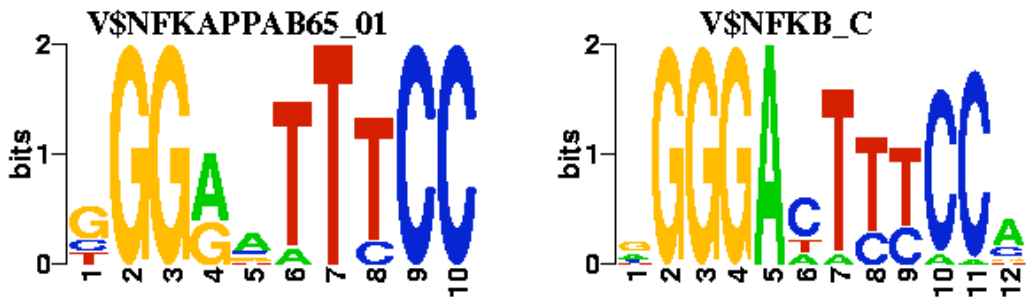
TFM-Explorer :: Correlation

Home | Results | Help

Comparison between matrix V$NFKAPPAB65_01 - region [-0521:-0019] **and** matrix V$NFKB_C - region [-0537:-0020]

Matrices

Matrix name	V$NFKAPPAB65_01	V$NFKB_C
Transcription Factor name	NF-kappaB (p65)	NF-kappaB
Database:	JASPAR OR TRANSFAC	JASPAR OR TRANSFAC
Information Content	14.76 bits	16.58 bits
Average GC%	0.57 bits	0.58 bits



WebLogo WebLogo

Sequences

Correlation coefficient	0.862745098039	
Number of sequences common to both regions	67	
Number of overlapping TFBSs	?	
Region	[-0521:-0019]	[-0537:-0020]

Conclusion

- Extract promoter areas that share relevant over-represented TF binding sites
  - No a priori knowledge of areas size or location is needed
  - Any kind of TF profile can be used
- Use regulatory motifs conservation across species
- Run on the fly