Promotor Detection in ADDNET

Sören Sonnenburg, Olaf Weiss
Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany

FIRST

Fraunhofer Institut
Rechnerarchitektur
und Softwaretechnik

Overview:

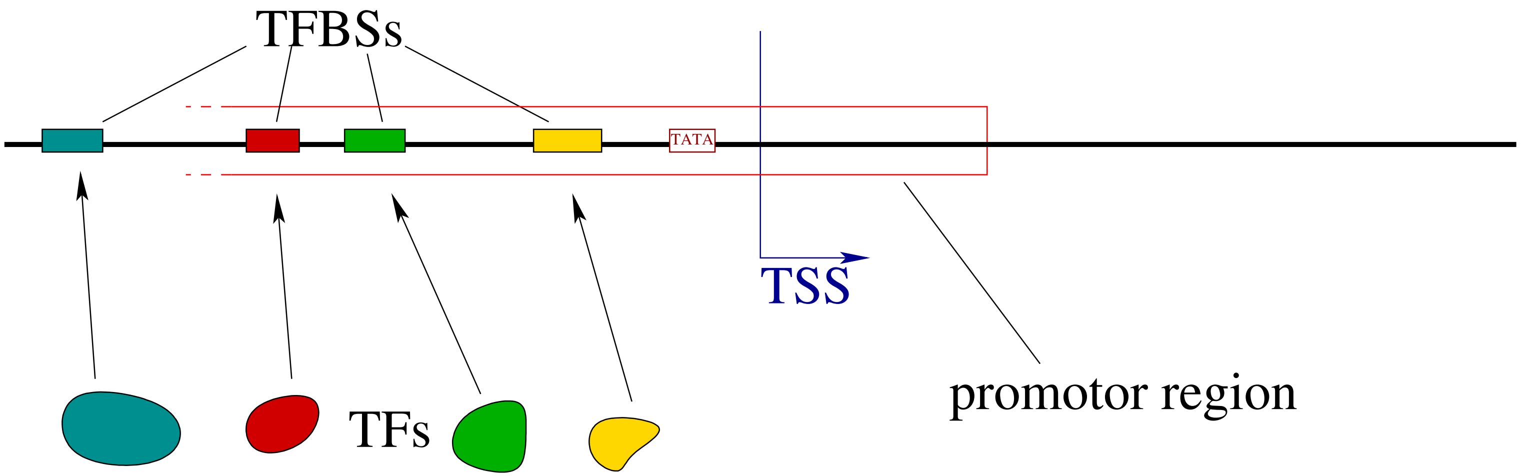
- Relevance for ADDNET
- Promotor ?
- Features to describe a Promotor
- Current Methods
- Approach of G. Rätsch and A. Zien
- TFBS recognition and Outlook

Promotor - Relevance for ADDNET

- find molecular markers of proteinurea by getting **candidates** from the analysis of microarrays of nephrin knock-out mice
- refining this list by **analysis of promotors**
- experimentally identify nephrin-binding proteins
- mass-spectroscopy analysis of urine, serum, and tissue samples
- from animal models
- large scale mass-spectroscopy of urine samples from clinical studies

Promotor - a Definition I

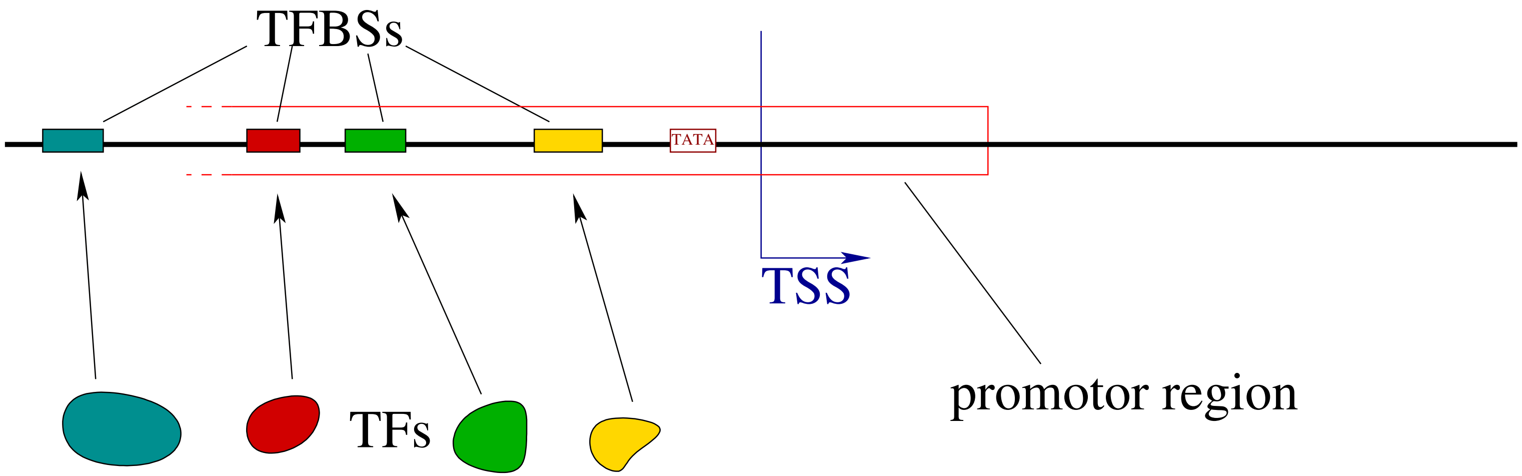
Loose region around the Transcription Start Site (TSS) where Transcription Factors (TFs) bind



Properties:

- Promotor has no exact location, more like a vague region
- no consensus sequence
- consists of core promotor, proximal promotor elements and distal enhancers (can be 10-50kb up/downstream of core promotor)

Promotor - a Definition II



Properties:

- TSS has no exact location, more like a range of $[-20, +20]$ base pairs
- TSS - again no consensus sequence
- position, order and number of TFBS in Promotor region variable

⇒ **Promotor Prediction is non-trivial**

Features to describe the Promotor

- TFBS in Promotor region
- condition: DNA should not be too twisted
- CpG islands (often over TSS/first exon, seem to be 2 general types CpG and non-CpG island promotors)
- TSS with TATA box (≈ -30 bp upstream)
- exon content in UTR 5” region
- distance to first donor splice site

Idea: Combine weak features to build strong promotor predictor

Current Methods

- FirstEF - DA: uses distance from CpG islands to first donor site
- CpGPro - Statistic Model: uses CpG islands
- McPromotor - 3-state HMM: upstream, TATA, downstream
- Eponine - RVM: upstream CpG islands, window upstream of TATA, for TATA, downstream

Good predictor incorporates strongest weak features

Approach of G. Rätsch and A. Zien

internal developmental release, granted to use it

- use SVM classifier
$$f(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{N} y_i \alpha_i \mathrm{k}(\boldsymbol{x}, \boldsymbol{x}_i) + b\right)$$
- key ingredient is kernel $k(\boldsymbol{x}, \boldsymbol{x}')$ – gives means to compare 2 sequences
- they use 5 sub-kernels

The 5 sub-kernels

1. to detect TSS (including parts of core promotor with TATA box) – use Weighted Degree Shift kernel

2. CpG Islands, distant Enhancers and TFBS upstream of TSS – use Spectrum kernel (large window upstream of TSS)

3. model coding sequence TFBS downstream of TSS – use another Spectrum kernel (small window downstream of TSS)

4. stacking energy of DNA – use btwist energy of dinucleotides with linear kernel

5. twistedness of DNA – use btwist angle of dinucleotides with linear kernel

Accuracy

- very recent internal development not yet published
- part of a Genefinder G. Rätsch et.al. are developing
- preliminary results on Drosophila show that this method outperforms McPromotor
- comparison with other PPP ongoing

TFBS Recognition

Weeder - a promising approach:

- can detect short motifs from length 6 to 12 (with mismatches)
- works by enumerating all possible motifs (with a constant, small number of mismatches)
- applied to 6 genes of Olaf's candidate gene list (AY324826, BC042496, J02943, AK172838, X81333, AF035835)

Weeder Results

Weeder finds a match of length 12 (2 mismatches): TTTAAAGAGACA score 15.67 and the following matches of shorter length on (some smaller ones pop-up in TRANSFAC):

10 (2)	8 (1)	6 (0)
GACATAGATT 17.49	TAGGCACT 15.56	ATAGAT 6.45
TAGGCACTAA 16.09	ACATAGAT 15.24	TATCAG 6.45
TTTAGGCACT 15.79	GACATAGA 11.99	ACAAAC 6.13
GGGAACATTA 15.76	CCAAGATA 11.79	AGATAA 6.13
TAGGCACTCA 15.22	GCATGGGG 11.49	GGCACT 5.90
TCAGGTATCA 15.20	CATAGATT 11.00	AACCAA 5.58
CATGGGGTAA 14.44	TCTGTTCC 10.80	AGAGTC 5.35
TATAGGCACT 14.43	ACCTTTAG 10.80	CAAGAA 5.28
CCAAGATAGG 14.08	TTAGGCAC 10.67	AAGAAC 5.11
AGACATAGAT 13.92	AGTACTTG 10.31	GCCATG 5.03

How to further proceed with these results (to be discussed)

Outlook

Todo: Compare promotor detector from Zien & Rätsch with other methods, choose “best” method.

From a clean set of candidate genes

- predict (or lookup) promotor region
- determine candidate TFBS
- get TF from TFBS (wetlab)

Acknowledgements: Alex Zien, Gunnar Rätsch and K.-R. Müller.