Предсказание положения промотора в геноме *Oryza sativa*

Анастасия Данчурова
Арина Дробышева
Павел Калинин
Артём Мулюков
Алена Титова
Ольга Черникова

Используемый алгоритм - random forest

Преимущества random forest:

- отсутствие метрики
- интерпретируемость

Данные:

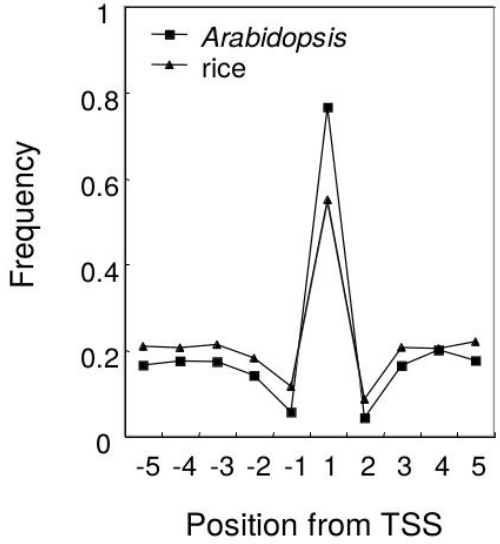
- риды 2000bp с известным расположением TSS

Обучающая выборка:

- случайные 80% данных

Параметры обучающей выборки

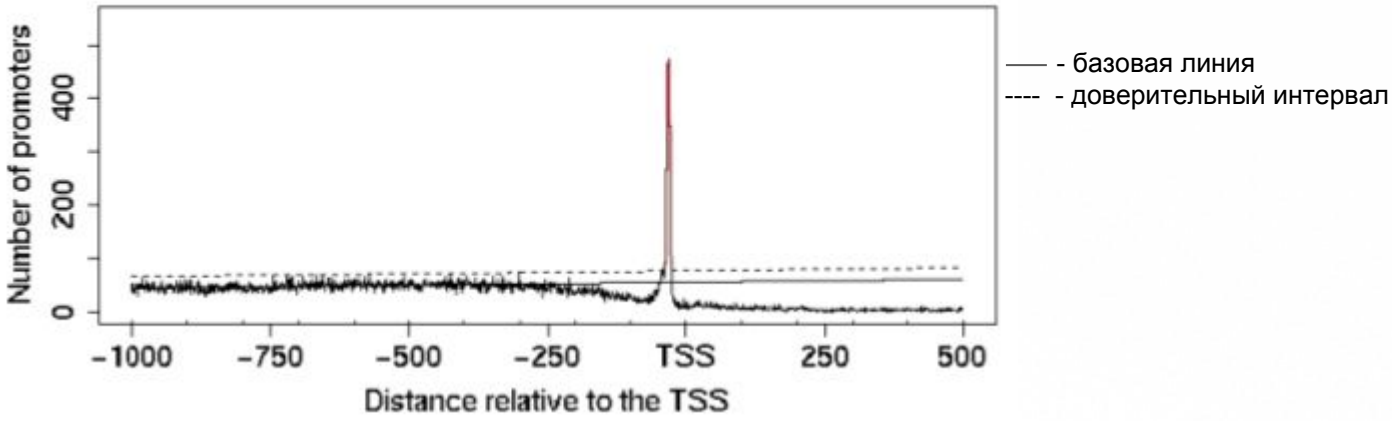
- димерный мотив YR rule: (C/T A/G) на позиции (-1/+1)



Yamamoto, Yoshiharu Y et al. “Identification of Plant Promoter Constituents by Analysis of Local Distribution of Short Sequences.” *BMC Genomics* 8 (2007): 67. *PMC*. Web. 4 Aug. 2017

Параметры обучающей выборки

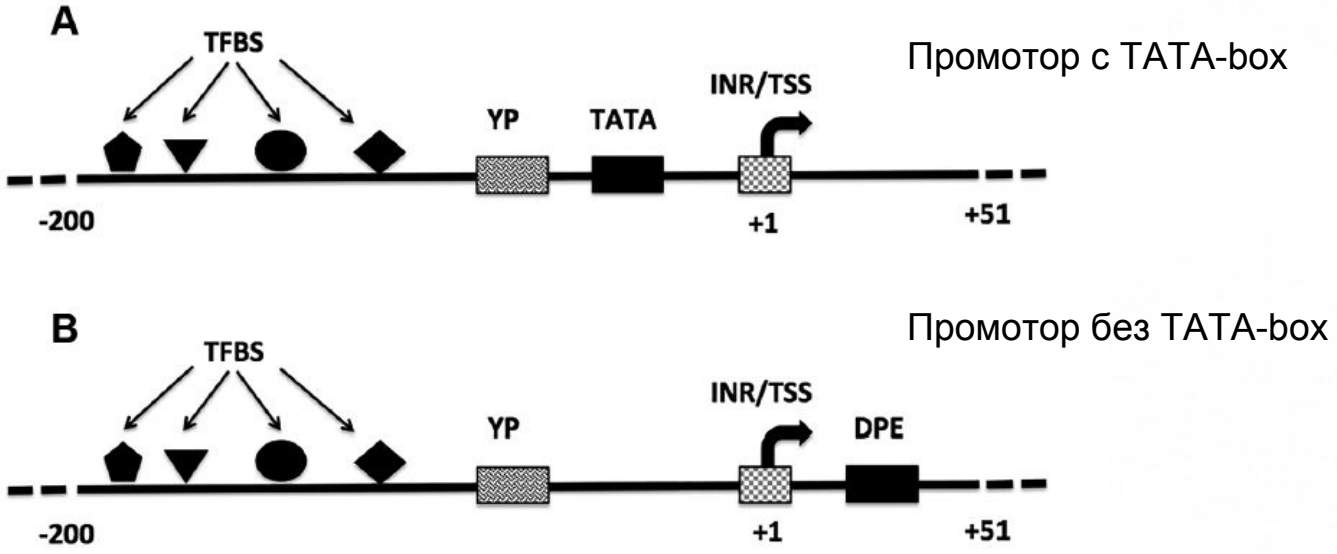
- TATA-box: поиск последовательности TATAWA на промежутке [-40;-20] от потенциального TSS



V.Bernard, V.Brunaud and A.Lecharny “TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation” BMC Genomics 2010

Параметры обучающей выборки

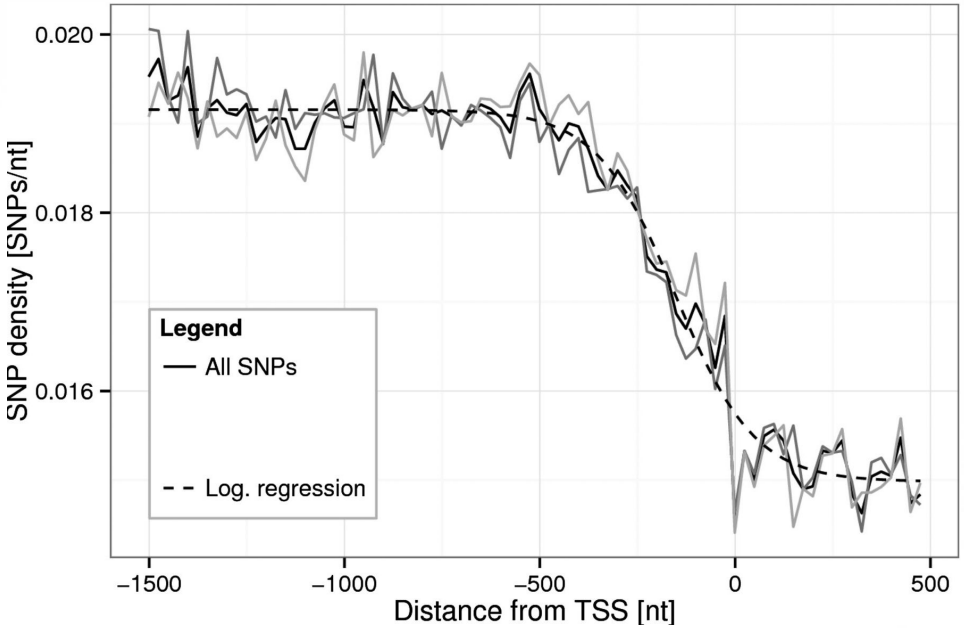
- Y Patch: TCTTCT / TTTCTT / TTCTTC на промежутке [-40;-25]



Shahmuradov IA, Umarov RK, Solovyev VV. TSSPlant: a new tool for prediction of plant Pol II promoters. Nucleic Acids Res. 2017 May 5;45(8):e65.

Параметры обучающей выборки

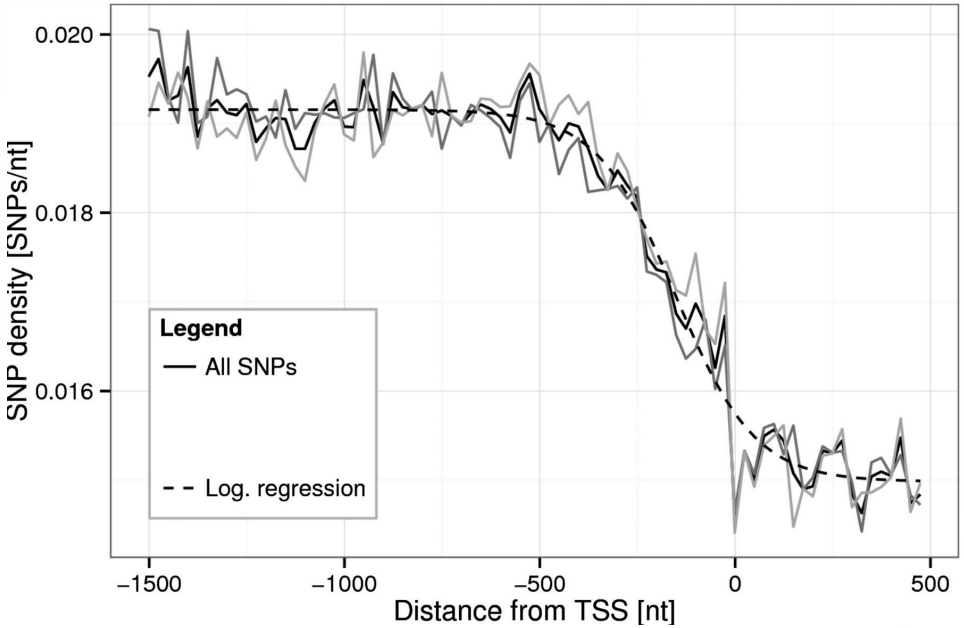
- SNP: отсутствие SNP на промежутке [-30;+10]



P.Korkuć, J.H.M.Schippers, D.Walther “Characterization and Identification of cis-Regulatory Elements in Arabidopsis Based on Single-Nucleotide Polymorphism Information” Published January 2014

Параметры обучающей выборки

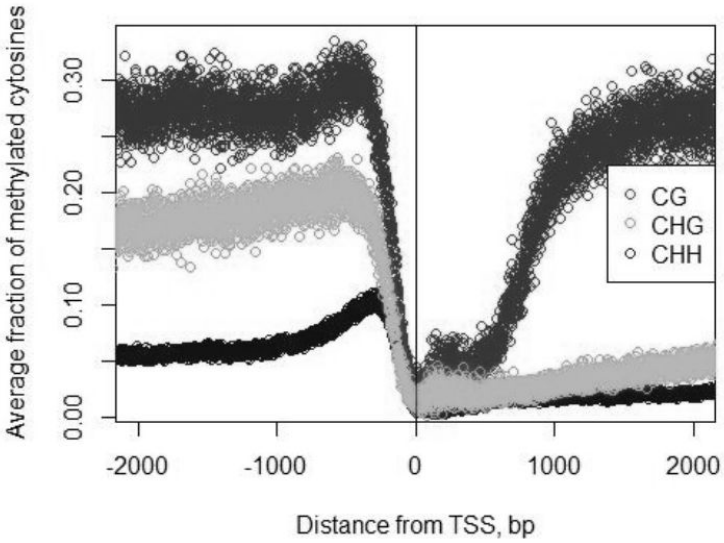
- SNP: отсутствие SNP на промежутке [-30;+10]



P.Korkuć, J.H.M.Schippers, D.Walther “Characterization and Identification of cis-Regulatory Elements in Arabidopsis Based on Single-Nucleotide Polymorphism Information” Published January 2014

Параметры обучающей выборки

- Метилирование: минимум метилированных CG на [-60;+60]



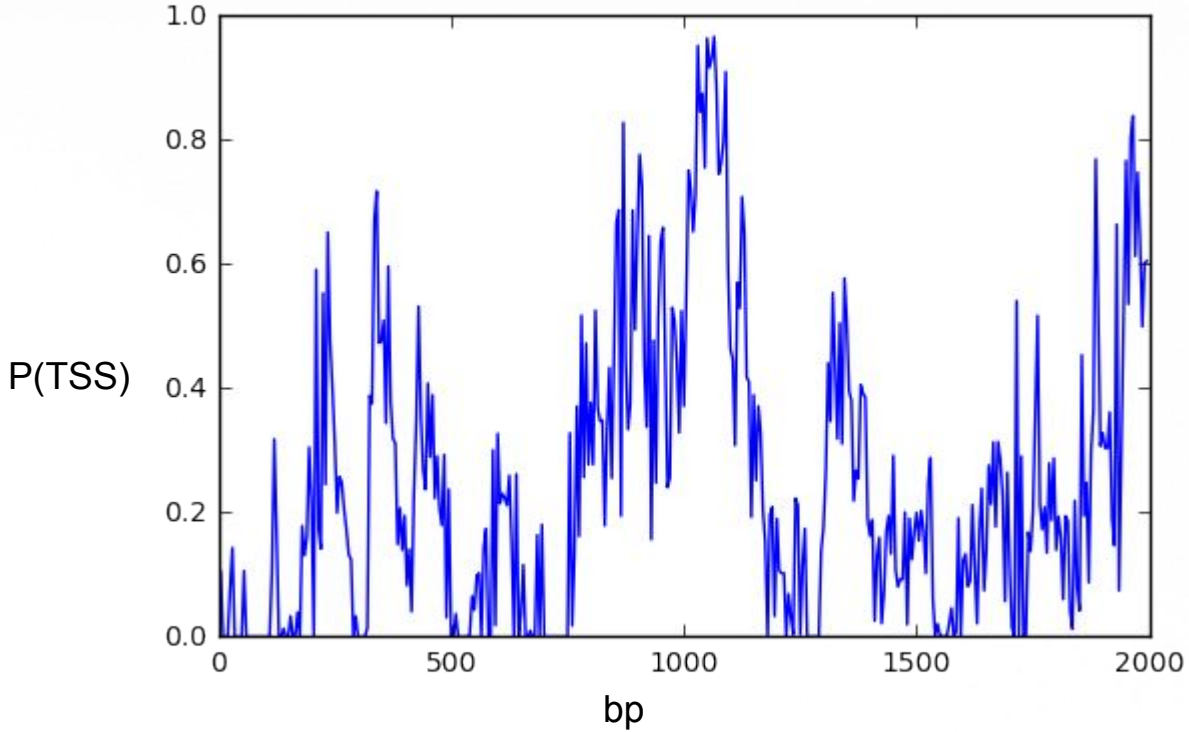
M.Triska et al.”Nucleotide patterns aiding in prediction of eukaryotic promoters” PONE-D-17-24929

Параметры обучающей выборки

- CAAT- и GC-box: GGGCGG и GGCCAATCT на [-100;-60] [1]
- увеличение %C на [-25;-17] при отсутствии TATA-box [2]
- уменьшение %GC на [-35;+35]

[1] https://en.wikipedia.org/wiki/CAAT_box, https://en.wikipedia.org/wiki/GC_box
[2] M.Triska et al.”Nucleotide patterns aiding in prediction of eukaryotic promoters” PONE-D-17-24929

Результаты



Спасибо за внимание!