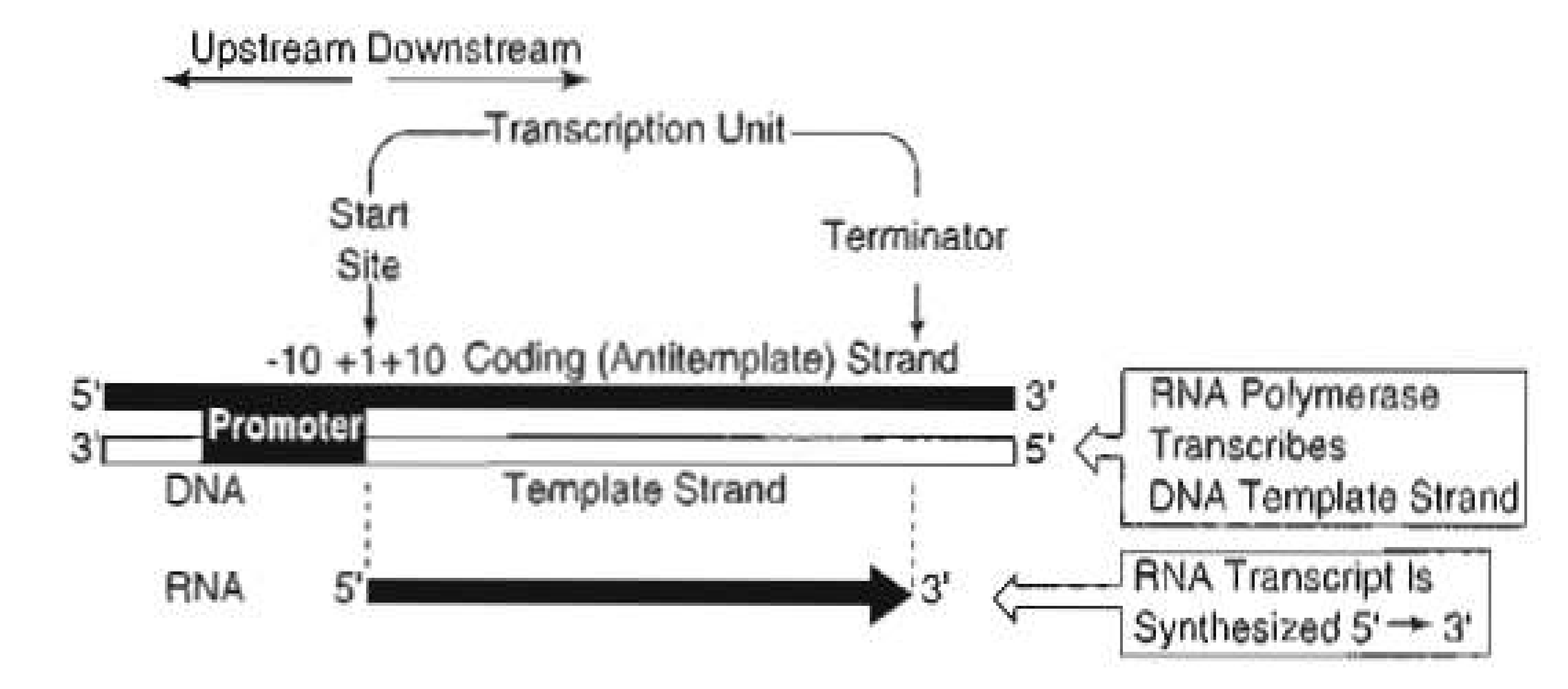
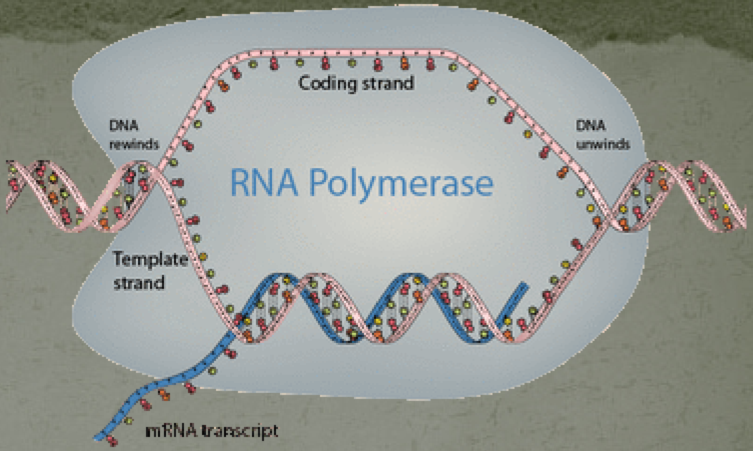
Предсказание и анализ промотерных последовательностей

Татьяна Татаринова


Upstream Downstream
Transcription Unit
Start Site
Terminator
-10 +1+10
Coding (Antitemplate) Strand
5'
3'
Promoter
3'
5'
DNA
Template Strand
RNA
5'
3'
RNA Polymerase Transcribes DNA Template Strand
RNA Transcript Is Synthesized 5' → 3'

Initiation



- Promoter: the DNA sequence that initially binds the RNA polymerase
- The structure of promoter-polymerase complex undergoes structural changes to proceed transcription
- DNA at the transcription site unwinds and a “bubble” forms
- Direction of RNA synthesis occurs in a 5’-3’ direction (3’-end growing)

Elongation

- Once the RNA polymerase has synthesized a short stretch of RNA (~ 10 nt), transcription shifts into the elongation phase.
- This transition requires further conformational change in polymerase that leads it to grip the template more firmly.
- Functions: synthesis RNA, unwinds the DNA in front, re-anneals it behind, dissociates the growing RNA chain

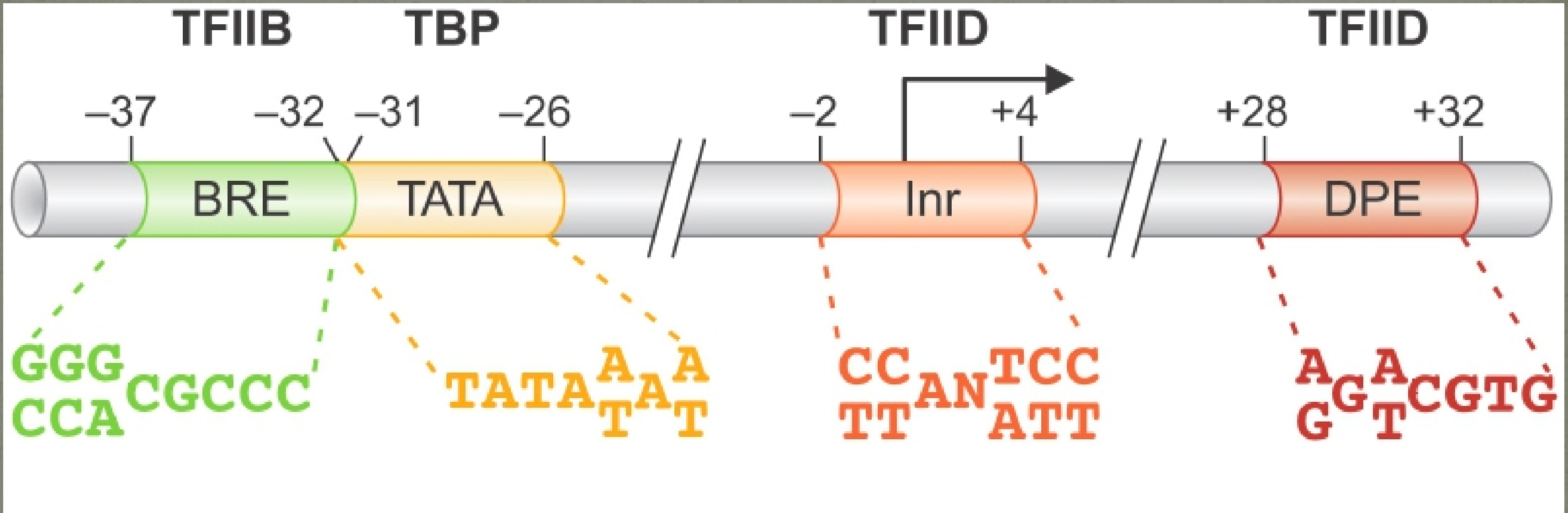
Termination

- After the polymerase transcribes the length of the gene, it will stop and release the RNA transcript.
- In some cells, termination occurs at the specific and well-defined DNA sequences called terminators. Some cells lack such termination sequences.

Control of Transcription initiation in Eukaryotes

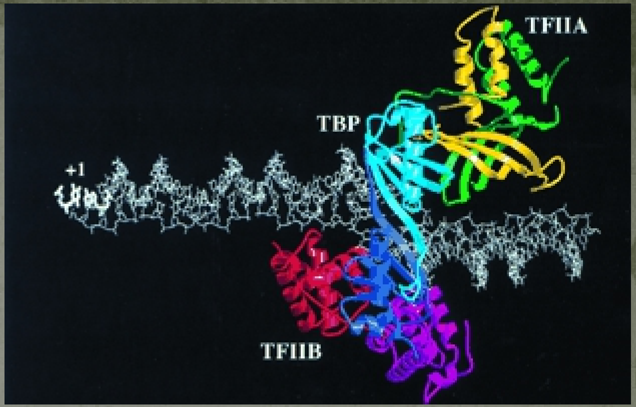
- Three types of RNA polymerases in eukaryotes
  - RNA pol I – transcribes ribosomal RNA genes
  - RNA pol II – transcribes all protein-coding genes (mRNAs) and micro-RNAs
  - RNA pol III – transcribes transfer RNA genes and some small regulatory RNAs
- Transcription initiation needs promoter and upstream regulatory regions.

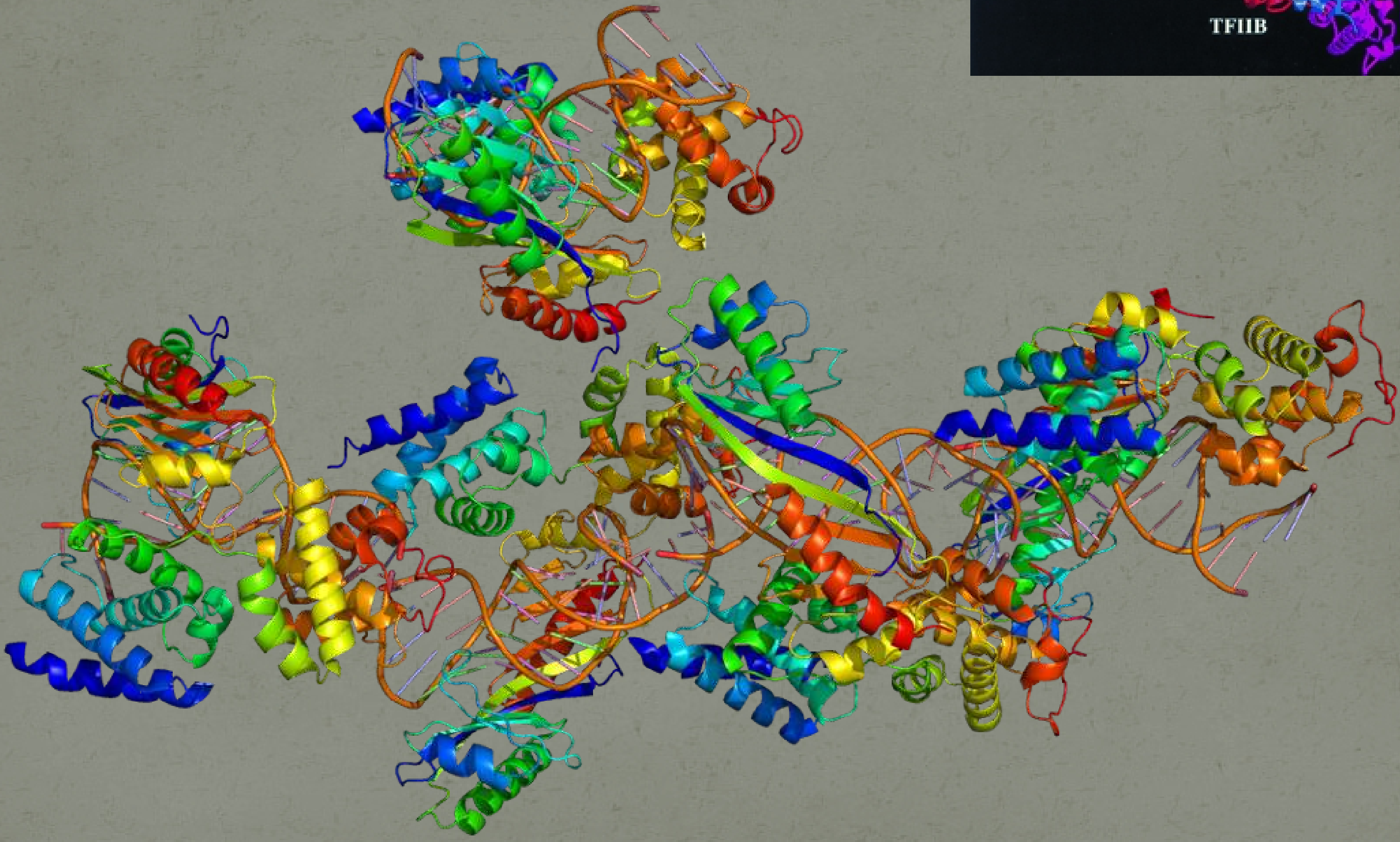
Pol II core promoter



- TFIIB recognition element (BRE)
- The TATA element/box
- Initiator (Inr)
- The downstream promoter element (DPE)

Transcription factor II B





TATA-binding protein (TBP)

- TBP is involved in DNA melting (double strand separation) by bending the DNA by 80° (the AT-rich sequence to which it binds facilitates easy melting)

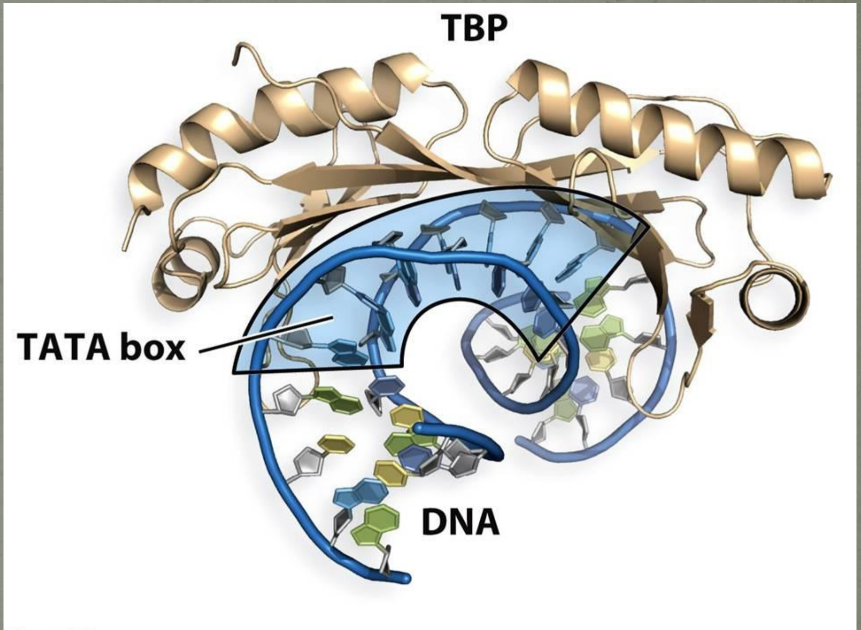


Figure 15-23
Molecular Biology: Principles and Practice
© 2012 W. H. Freeman and Company

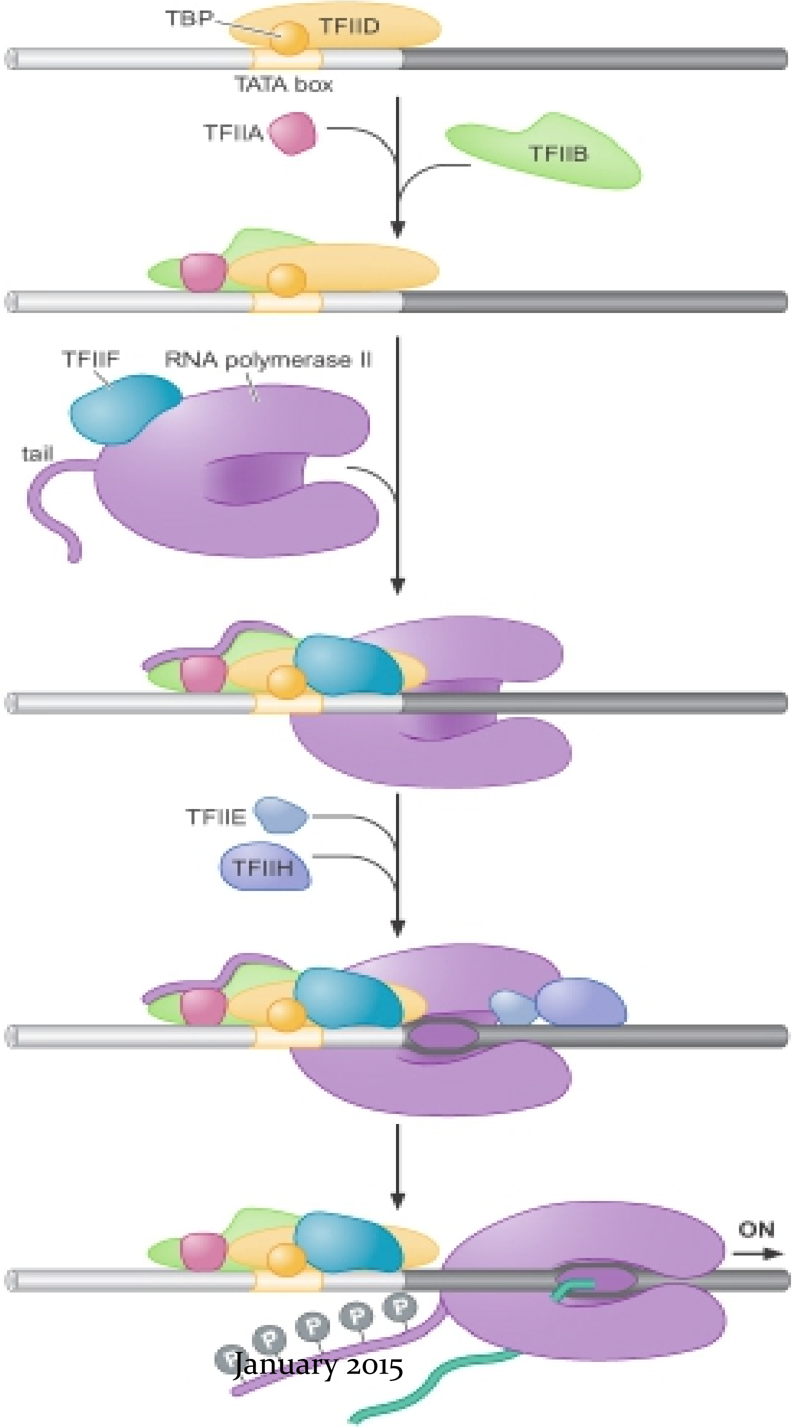
Regulatory sequences

The sequence elements other than the core promoter that are required to regulate the transcription efficiency

Those increasing transcription:

- Promoter proximal elements
- Upstream activator sequences (UASs)
- Enhancers

Those repressing elements: silencers, boundary elements, insulators



1. TBP in TFIID binds to the TATA box
2. TFIIA and TFIIB are recruited with TFIIB binding to the BRE
3. RNA Pol II-TFIIF complex is then recruited
4. TFIIE and TFIIH then bind upstream of Pol II to form the pre-initiation complex
5. Promoter melting using energy from ATP hydrolysis by TFIIH
6. Promoter escapes after the phosphorylation of the CTD tail

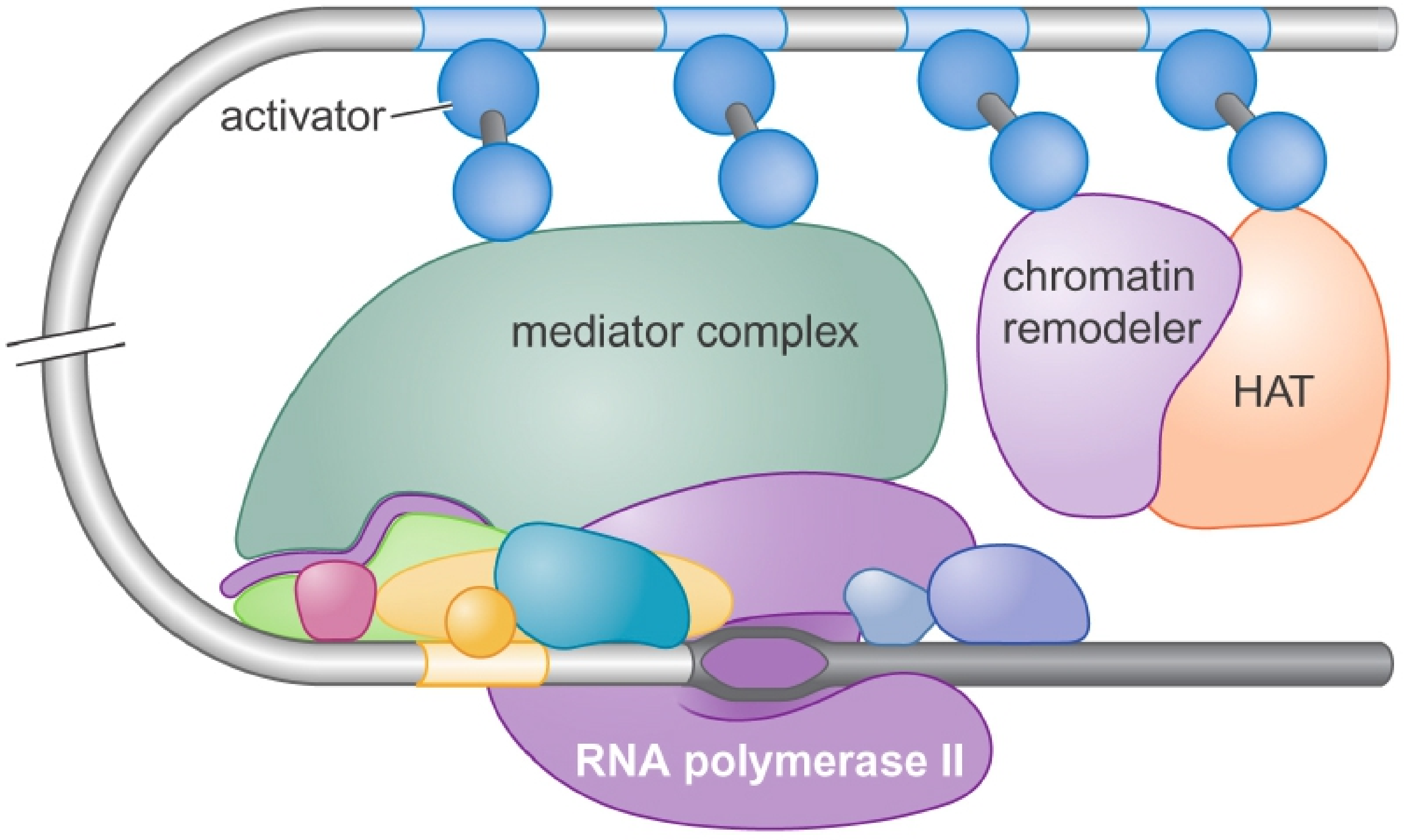
The transcription in eukaryotes

in vivo, transcription initiation requires additional proteins

- The mediator complex
- Transcriptional regulatory proteins
- Nucleosome-modifying enzymes

To counter the real situation that the DNA template *in vivo* involves chromatin

Assembly of the pre-initiation complex in presence of mediator, nucleosome modifiers and remodelers, and transcriptional activators



Eukaryotic Transcription

Mediator consists of many subunits, some conserved from yeast to human

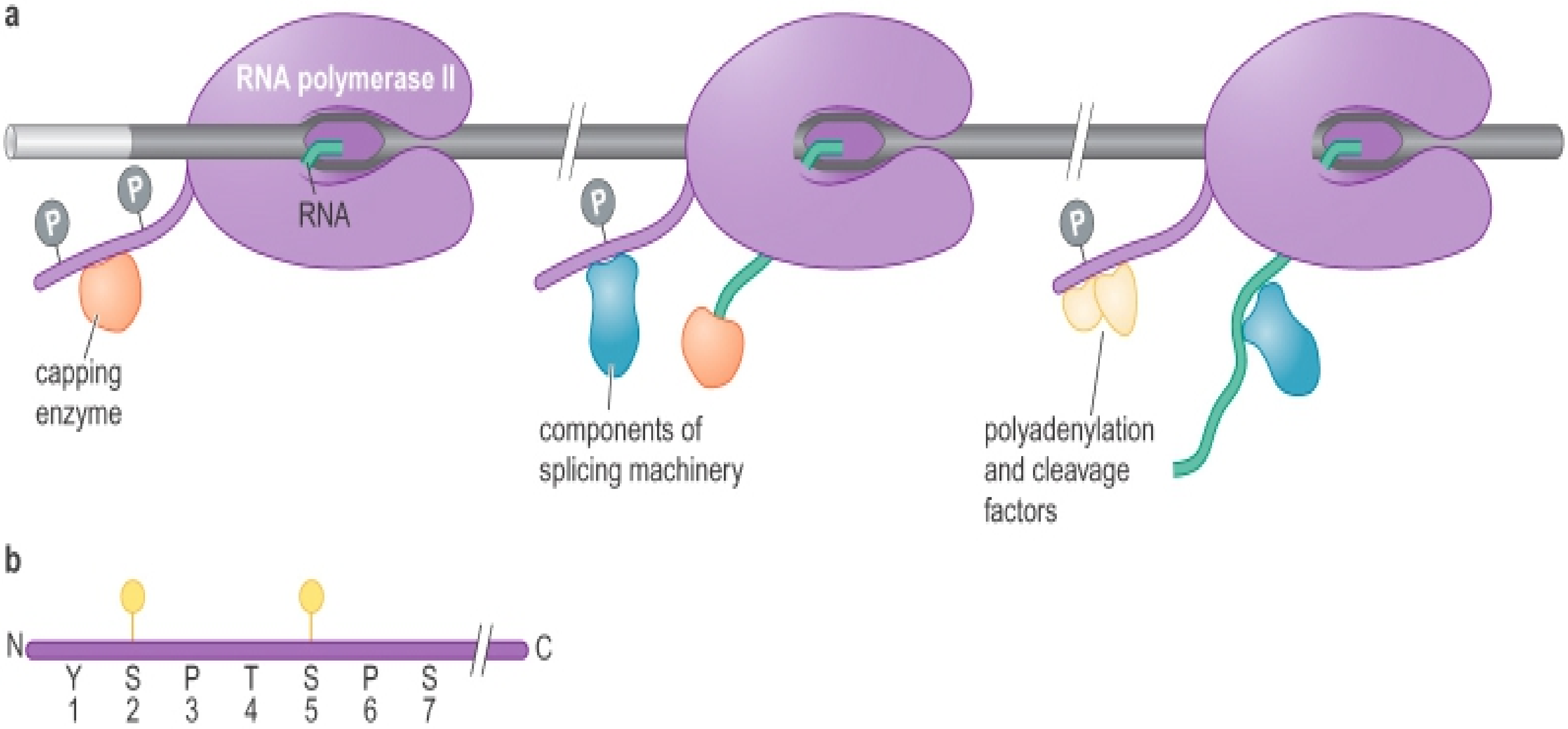
- More than 20 subunits
- 7 subunits show significant sequence homology between yeast and human
- Only subunit Srb4 is essential for transcription of essentially all Pol II genes *in vivo*
- Organized in modules

Transition from the initiation to elongation involves the Pol II enzyme shedding most of its initiation factors (GTF and mediators) and recruiting other factors:

(1) Elongation factors: factors that stimulate elongation, such as TFIIS and hSPT5.

(2) RNA processing factors

Recruited to the C-terminal tail of the CTD of RNAP II to phosphorylate the tail for elongation stimulation, proofreading, and RNA processing like splicing and polyadenylation.

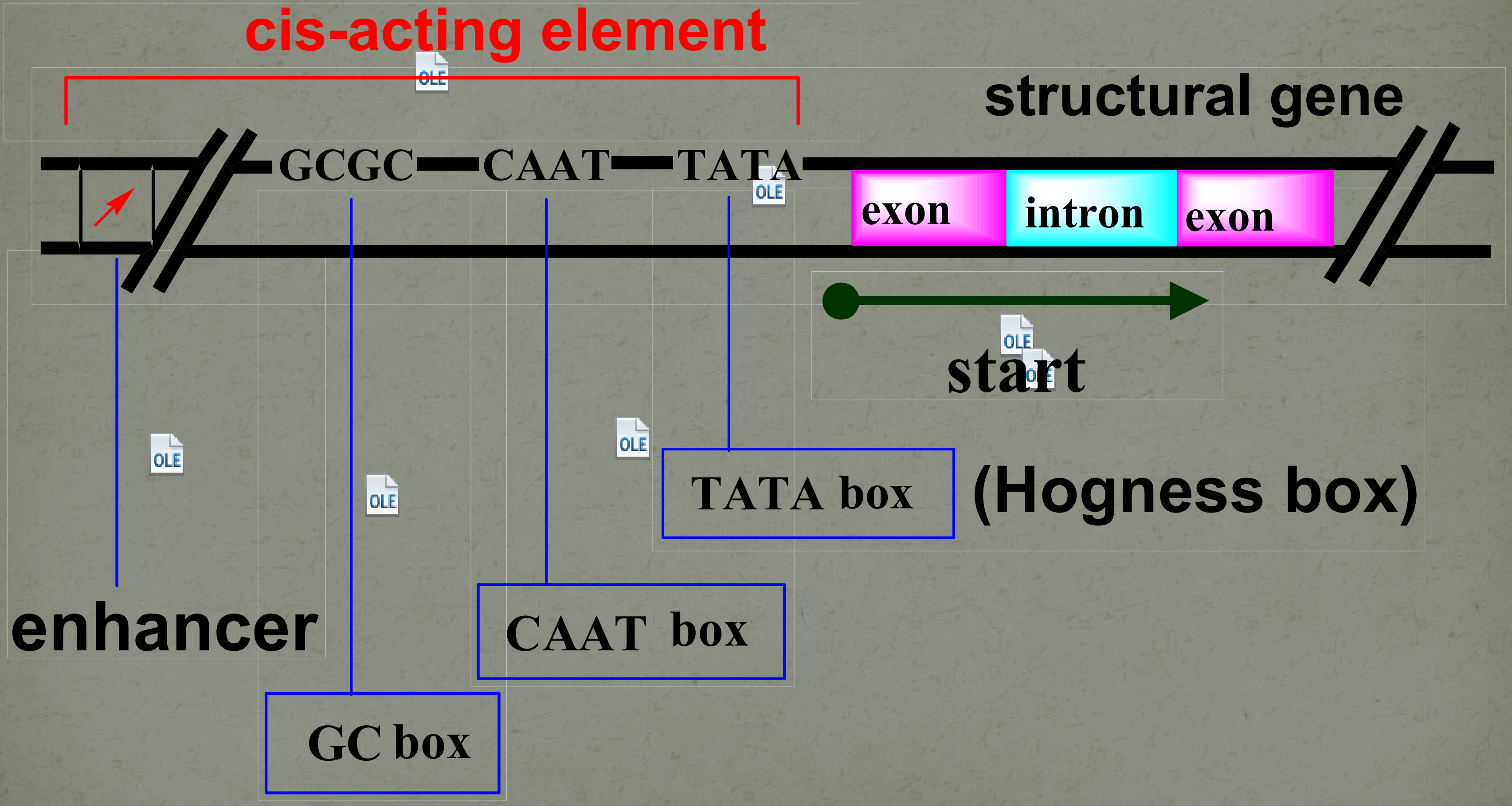

a
RNA polymerase II
P
P
RNA
capping enzyme
P
components of splicing machinery
P
polyadenylation and cleavage factors
b
N
C
Y 1
S 2
P 3
T 4
S 5
P 6
S 7

RNA processing enzymes are recruited by the tail of polymerase

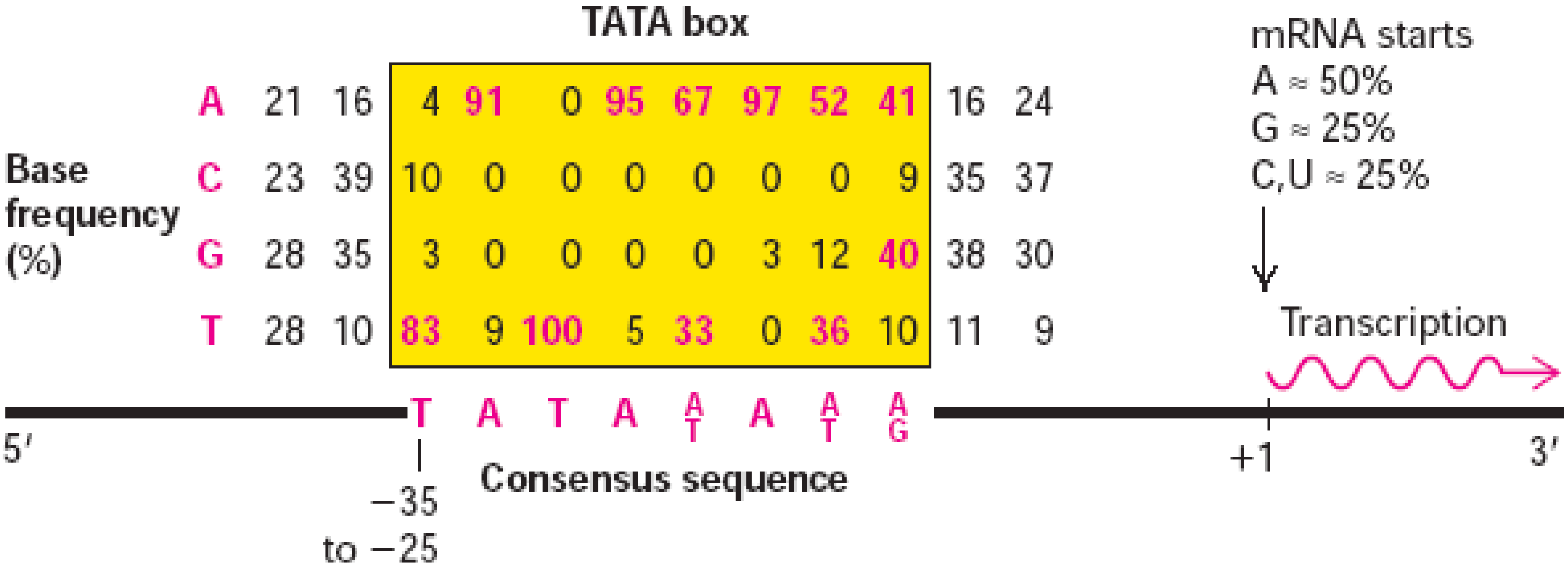
cis-Regulatory Elements (CREs)

- regions of non-coding DNA which regulate the transcription of nearby genes
- typically regulate gene transcription by functioning as binding sites for TF
- Types of CREs: enhancers, promoters, silencer…

Cis-acting regulatory elements



TATA box



TATA box

Base frequency (%)

A	21	16	4	91	0	95	67	97	52	41	16	24
C	23	39	10	0	0	0	0	0	0	9	35	37
G	28	35	3	0	0	0	0	3	12	40	38	30
T	28	10	83	9	100	5	33	0	36	10	11	9

Transcription factors

- RNA-pol does not bind the promoter directly
- RNA-pol II associates with six transcription factors, TFII A - TFII H.
- trans-acting factors - proteins that recognize and bind directly or indirectly to cis-acting elements and regulate its activity.

cis-acting factors: promoters and enhancers

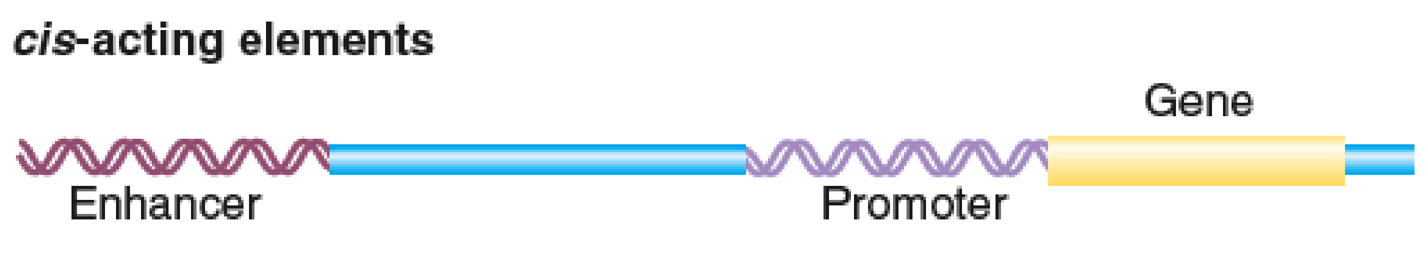
Promoters – usually directly adjacent to the gene

- Include transcription initiation site
- Often have TATA box:
- Allow basal level of transcription

$$TATA\frac{A}{T}A\frac{A}{T}$$

Enhancers – can be far away from gene

- Augment or repress the basal level of transcription



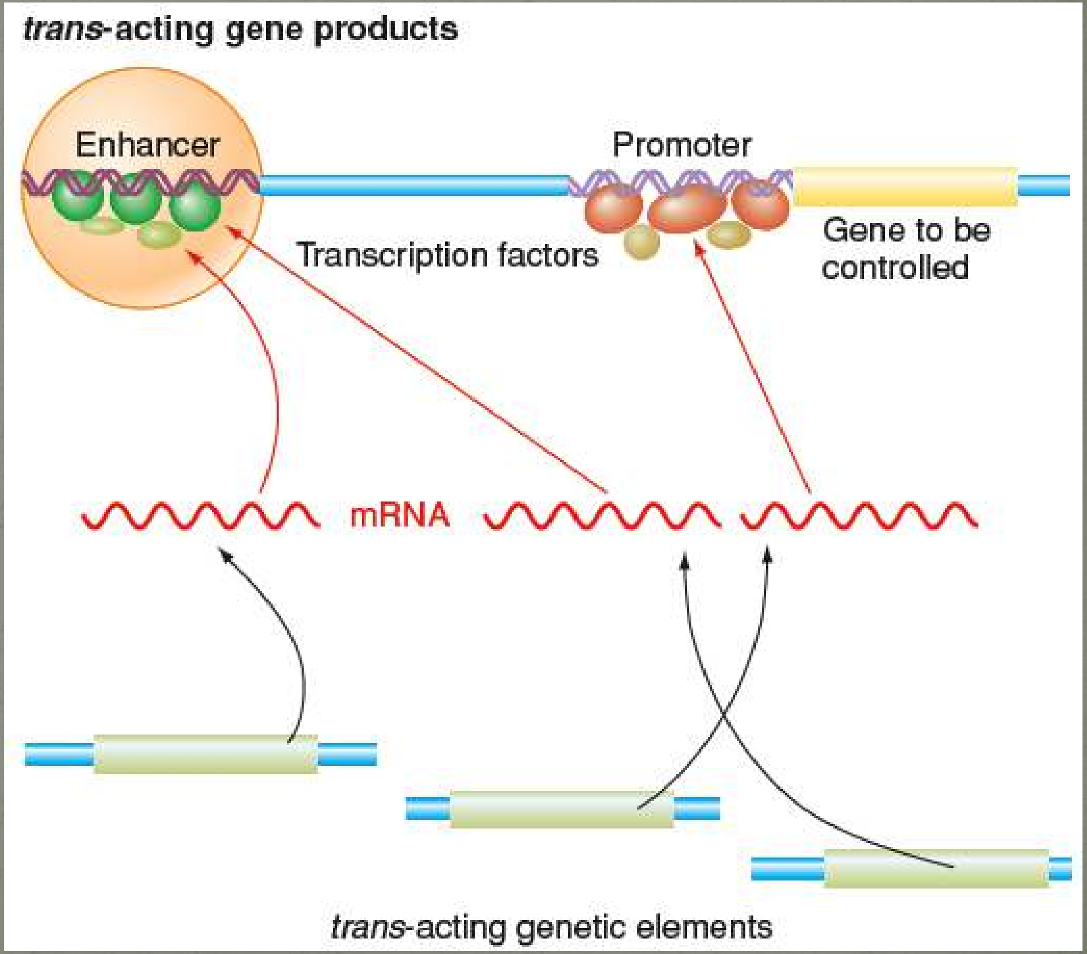
trans-acting factors interact with *cis*-acting elements to control transcription initiation

Direct effects of transcription factors:

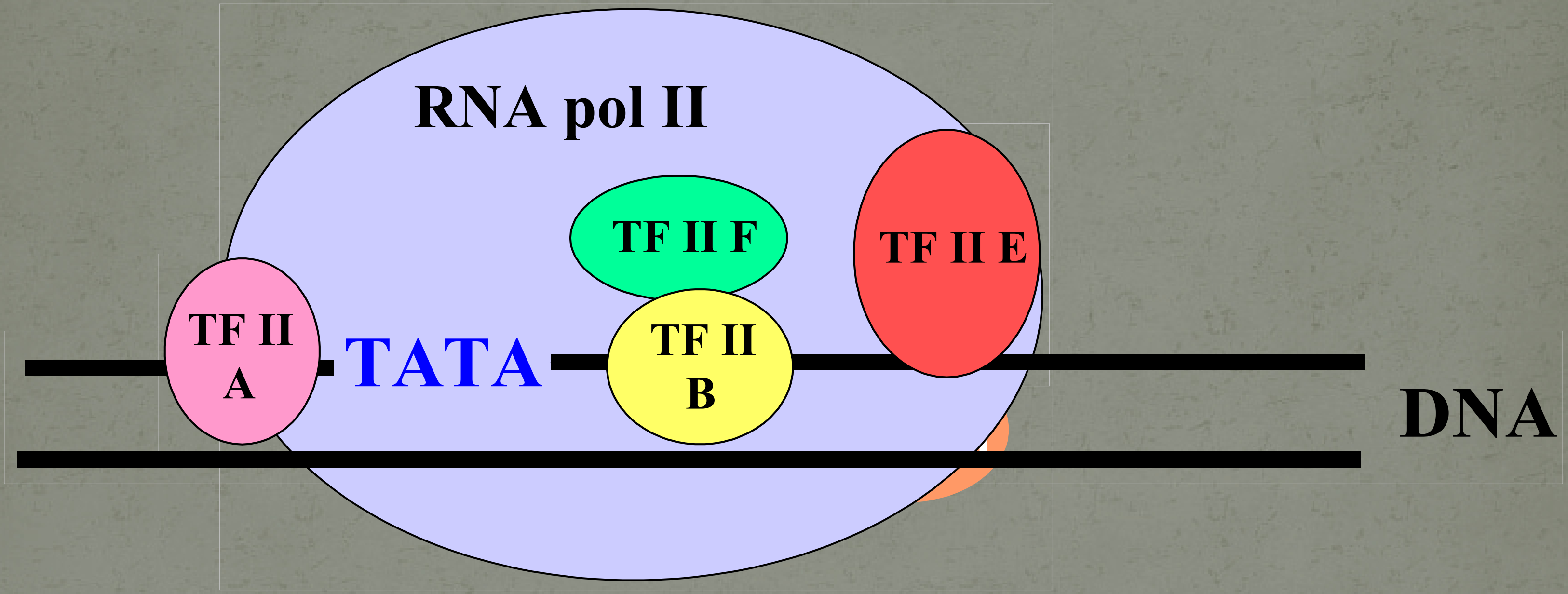
- Through binding to DNA

Indirect effect of transcription factors:

- Through protein-protein interactions



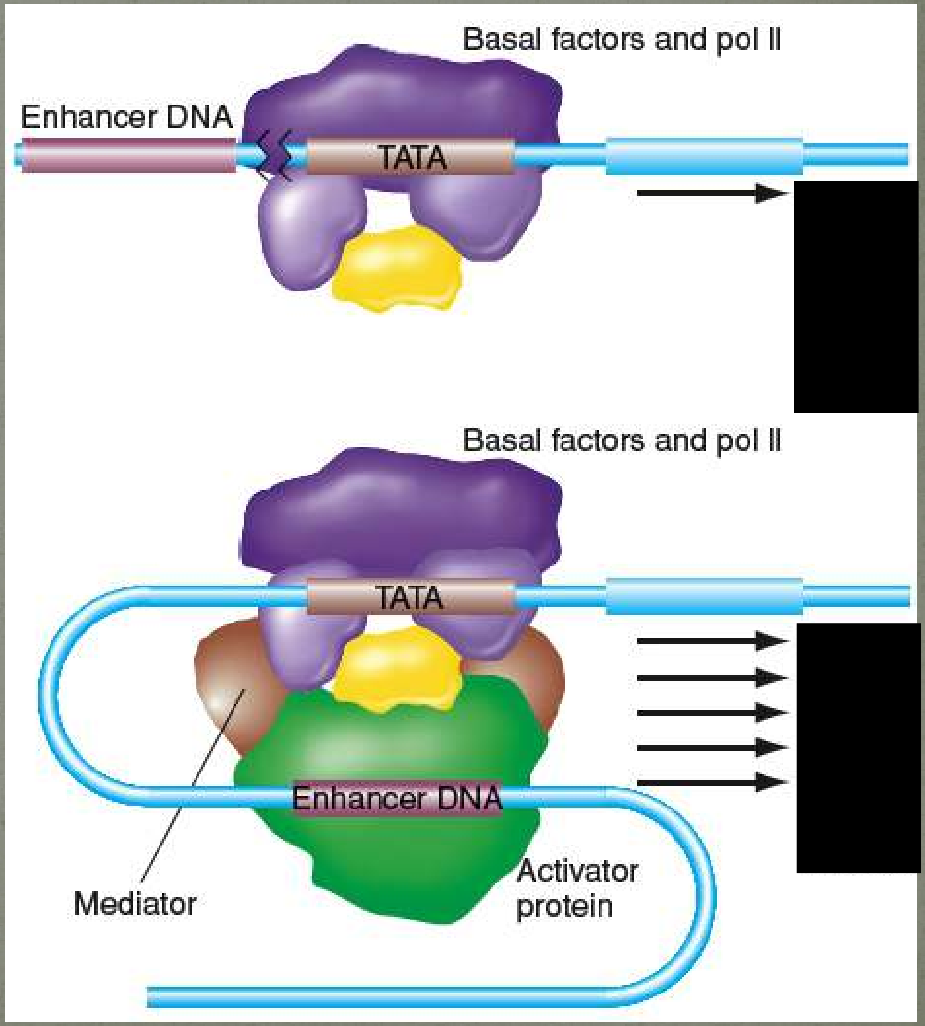
Pre-initiation complex (PIC)



Binding of activators to enhancers increases transcriptional levels

Low level transcription occurs when only basal factors are bound to promoter

When basal factors and activators are bound to DNA, rate of transcription increases



[redacted]

[redacted]

Repressor proteins suppress transcription initiation through different mechanisms

Some repressors have no effect on basal transcription but suppress the action of activators

- Compete with activator for the same enhancer

  OR
- Block access of activator to an enhancer

Some repressors eliminate virtually all basal transcription from a promoter

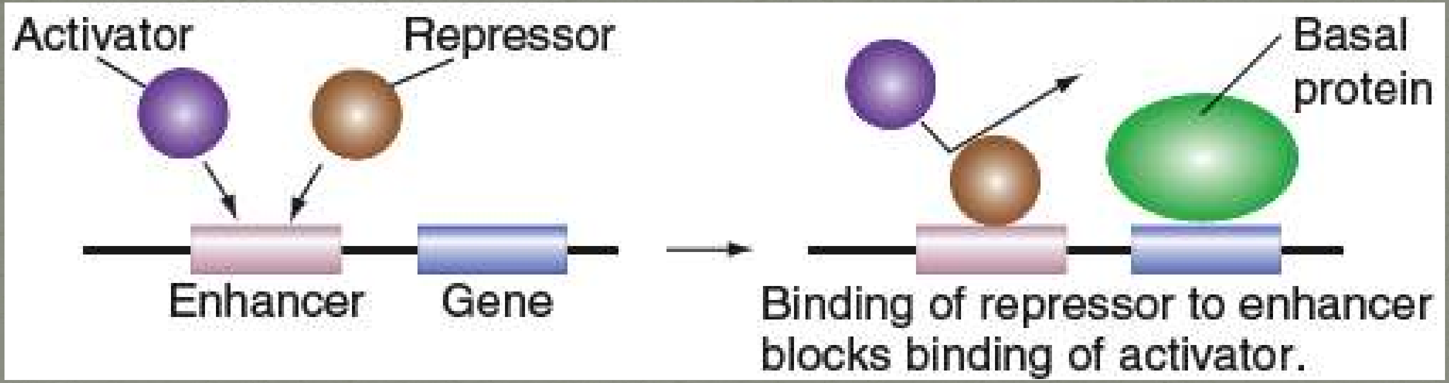
- Block RNA pol II access to promoter

  OR
- Bind to sequences close to promoter or distant from promoter

Repressor proteins that act through competition with an activator protein

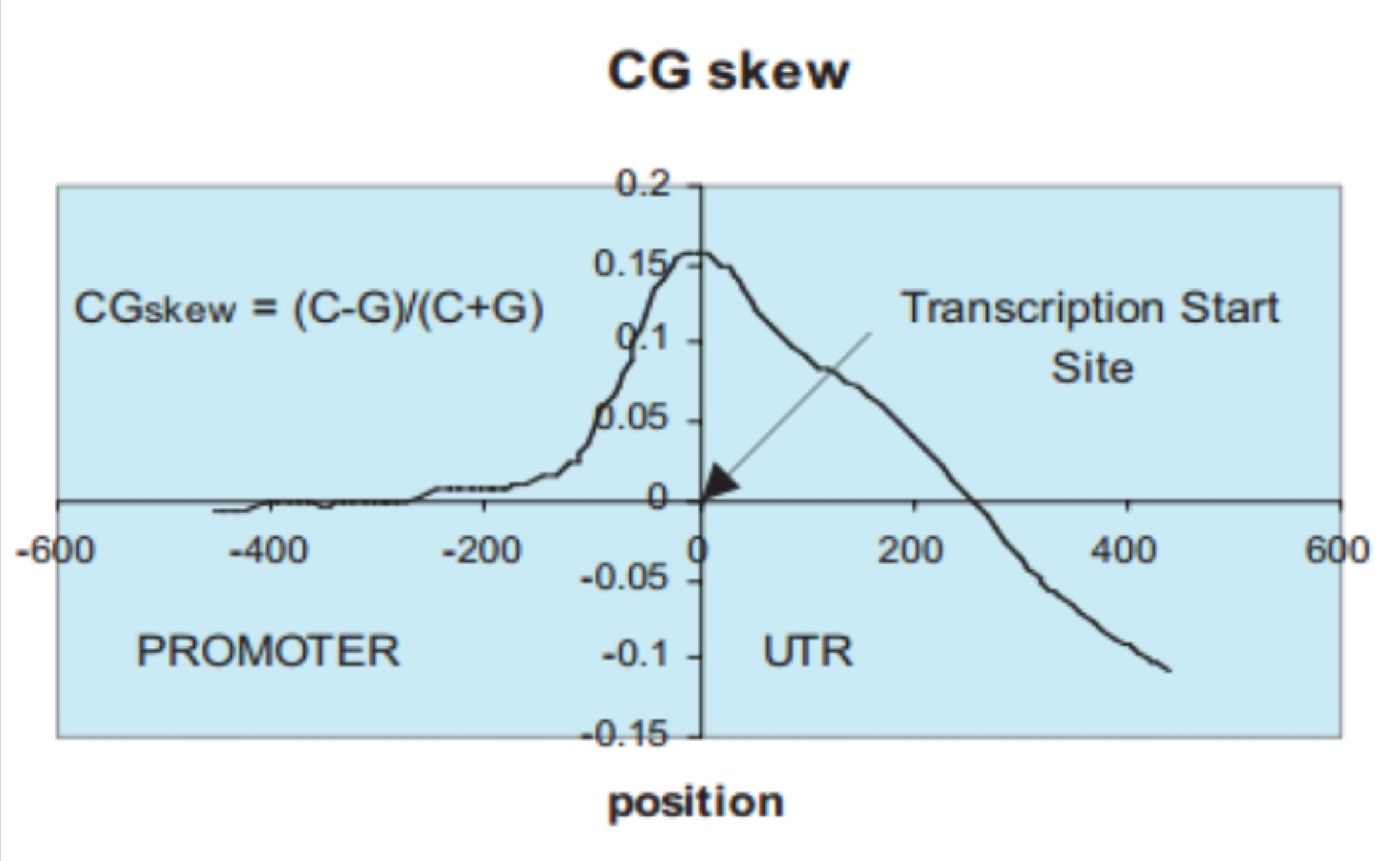
Repressor binds to the same enhancer sequence as the activator

- Has no effect on the basal transcription level



Binding of repressor to enhancer blocks binding of activator.

TSS and DNA features



Tatarinova et al, 2003 Bioinformatics

Promoter Databases and sites for analysis, prediction and search

- http://genetools.us/genomics/Promoter%20databases%20and%20prediction%20tools.htm

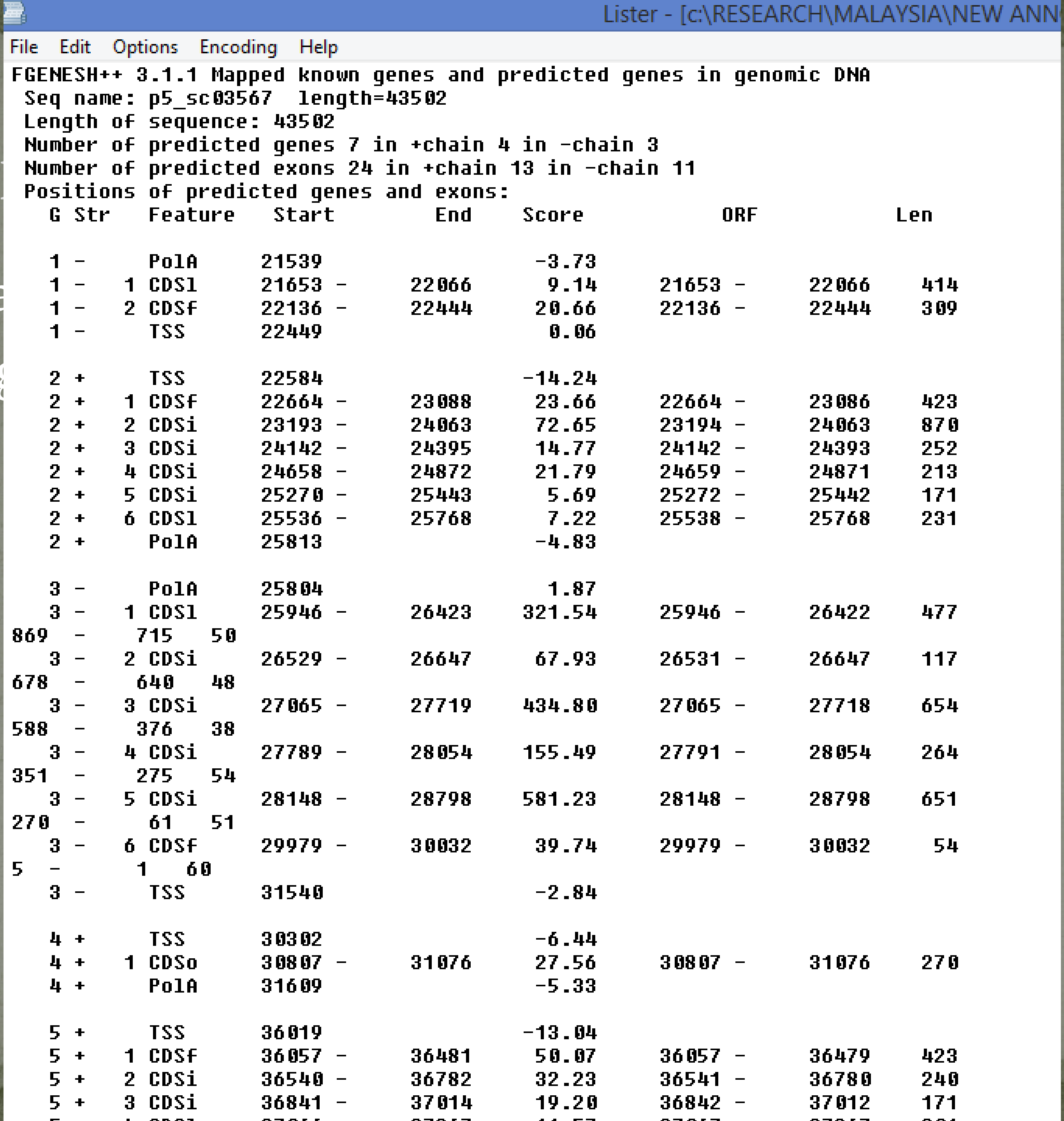
Проблемы с моделированием

Почему это сложно?

- Шумные эксперименты
- Специфичность промотеров в клетках
- Альтернативные промотеры
- Устаревший код и догмы

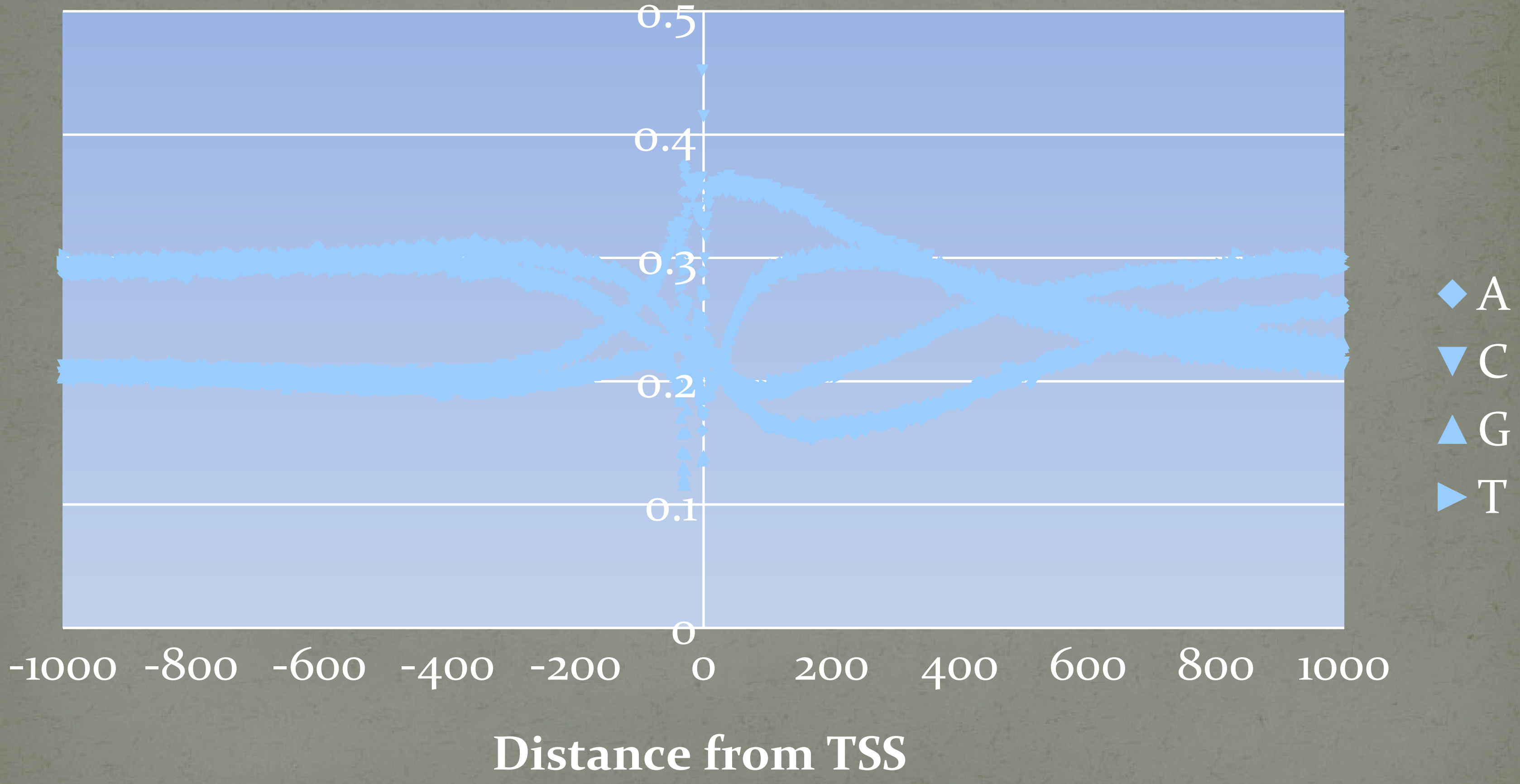
Pro

- As a
- Targ



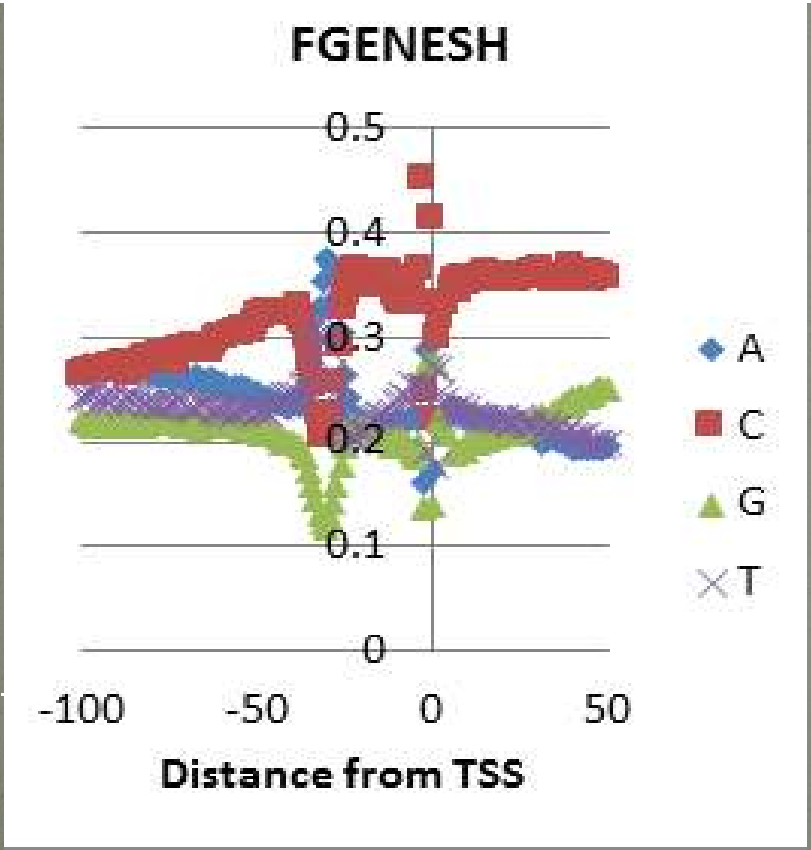
```
FGENESH++ 3.1.1 Mapped known genes and predicted genes in genomic DNA
 Seq name: p5_sc03567  length=43502
 Length of sequence: 43502
 Number of predicted genes 7 in +chain 4 in -chain 3
 Number of predicted exons 24 in +chain 13 in -chain 11
 Positions of predicted genes and exons:
   G Str    Feature   Start         End     Score           ORF            Len

   1 -      PolA     21539                  -3.73
   1 -    1 CDSl     21653 -      22066      9.14     21653 -      22066     414
   1 -    2 CDSf     22136 -      22444     20.66     22136 -      22444     309
   1 -      TSS      22449                   0.06

   2 +      TSS      22584                 -14.24
   2 +    1 CDSf     22664 -      23088     23.66     22664 -      23086     423
   2 +    2 CDSi     23193 -      24063     72.65     23194 -      24063     870
   2 +    3 CDSi     24142 -      24395     14.77     24142 -      24393     252
   2 +    4 CDSi     24658 -      24872     21.79     24659 -      24871     213
   2 +    5 CDSi     25270 -      25443      5.69     25272 -      25442     171
   2 +    6 CDSl     25536 -      25768      7.22     25538 -      25768     231
   2 +      PolA     25813                  -4.83

   3 -      PolA     25804                   1.87
   3 -    1 CDSl     25946 -      26423    321.54     25946 -      26422     477
869   -    715   50
   3 -    2 CDSi     26529 -      26647     67.93     26531 -      26647     117
678   -    640   48
   3 -    3 CDSi     27065 -      27719    434.80     27065 -      27718     654
588   -    376   38
   3 -    4 CDSi     27789 -      28054    155.49     27791 -      28054     264
351   -    275   54
   3 -    5 CDSi     28148 -      28798    581.23     28148 -      28798     651
270   -     61   51
   3 -    6 CDSf     29979 -      30032     39.74     29979 -      30032      54
5   -       1   60
   3 -      TSS      31540                  -2.84

   4 +      TSS      30302                  -6.44
   4 +    1 CDSo     30807 -      31076     27.56     30807 -      31076     270
   4 +      PolA     31609                  -5.33

   5 +      TSS      36019                 -13.04
   5 +    1 CDSf     36057 -      36481     50.07     36057 -      36479     423
   5 +    2 CDSi     36540 -      36782     32.23     36541 -      36780     240
   5 +    3 CDSi     36841 -      37014     19.20     36842 -      37012     171
```

Nucleotide consensus at TSS

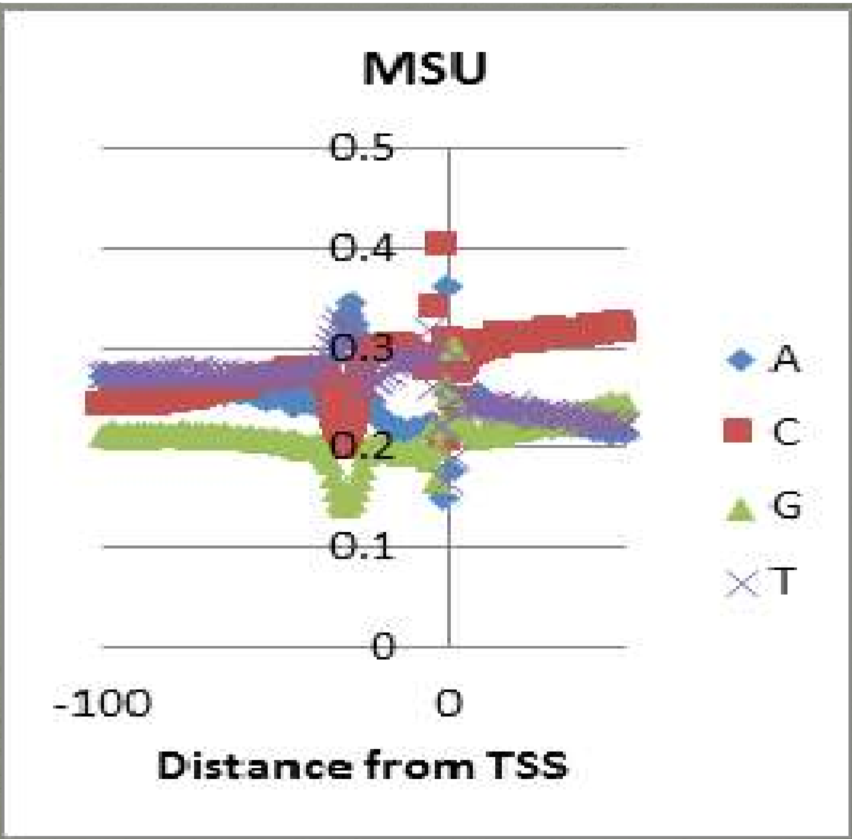


2 sets gene predictions 55K loci, 49K of them non-TE

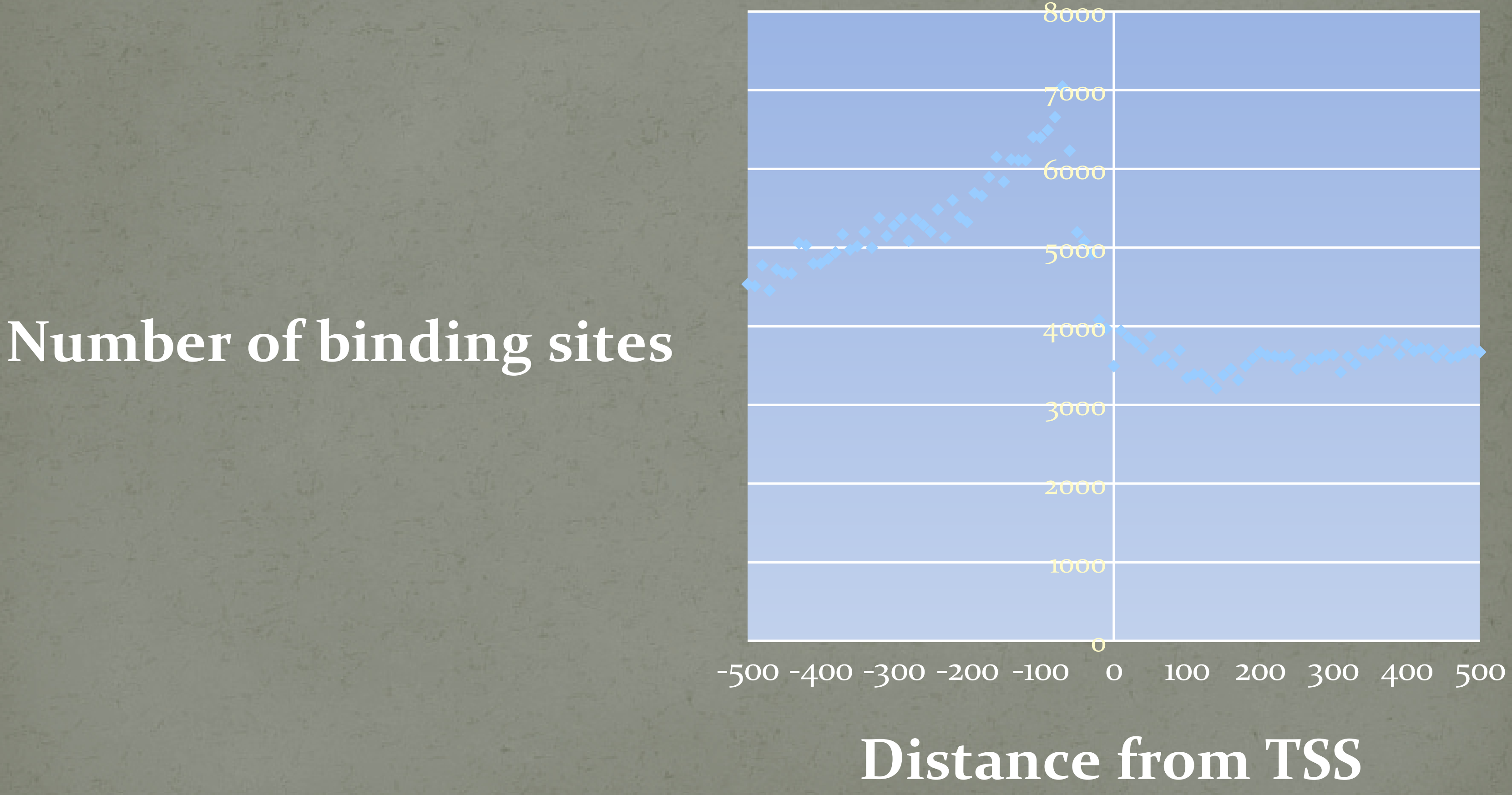
Solovyev (FGENESH)



MSU rice

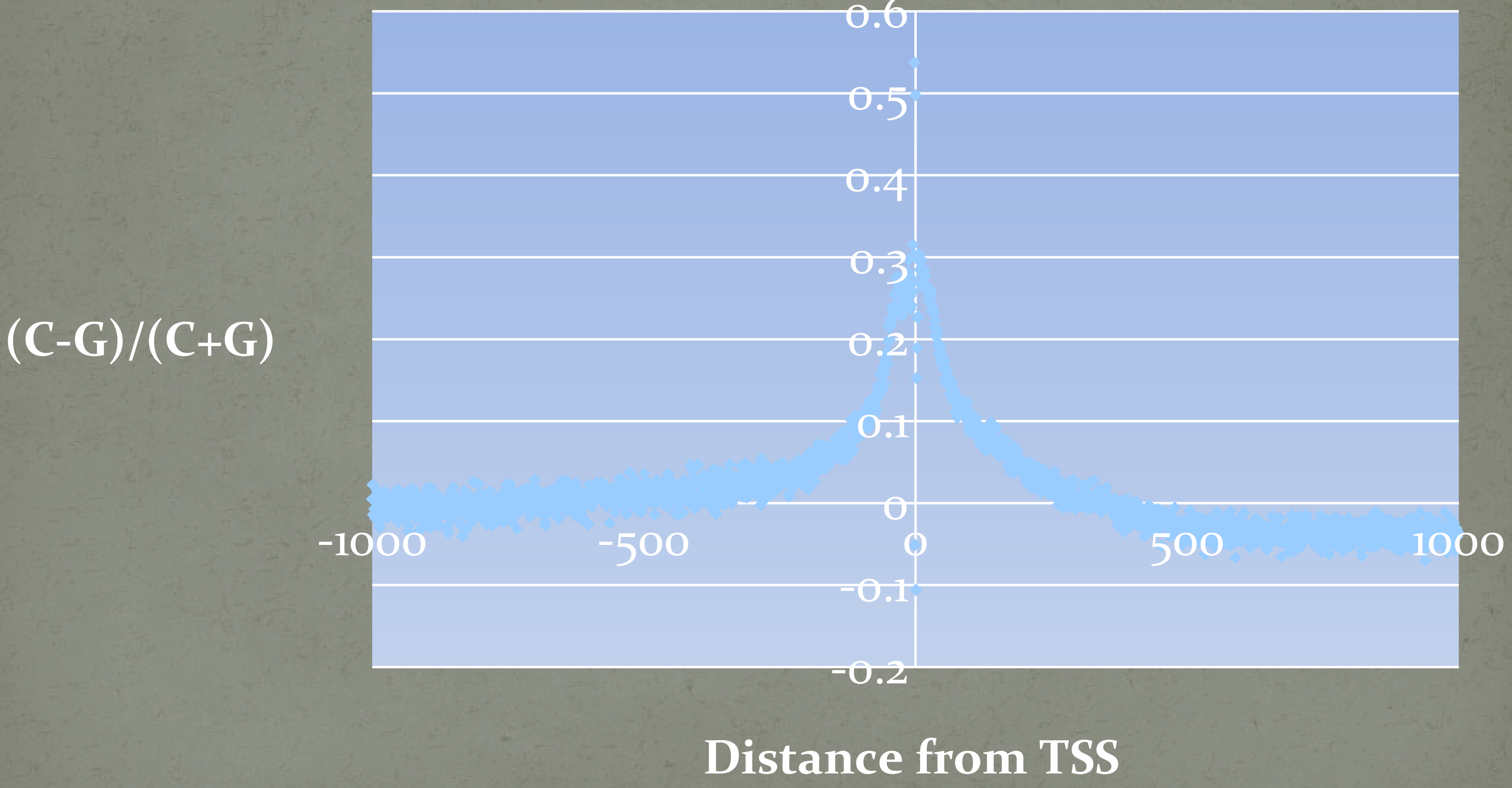


TRANSFAC binding sites

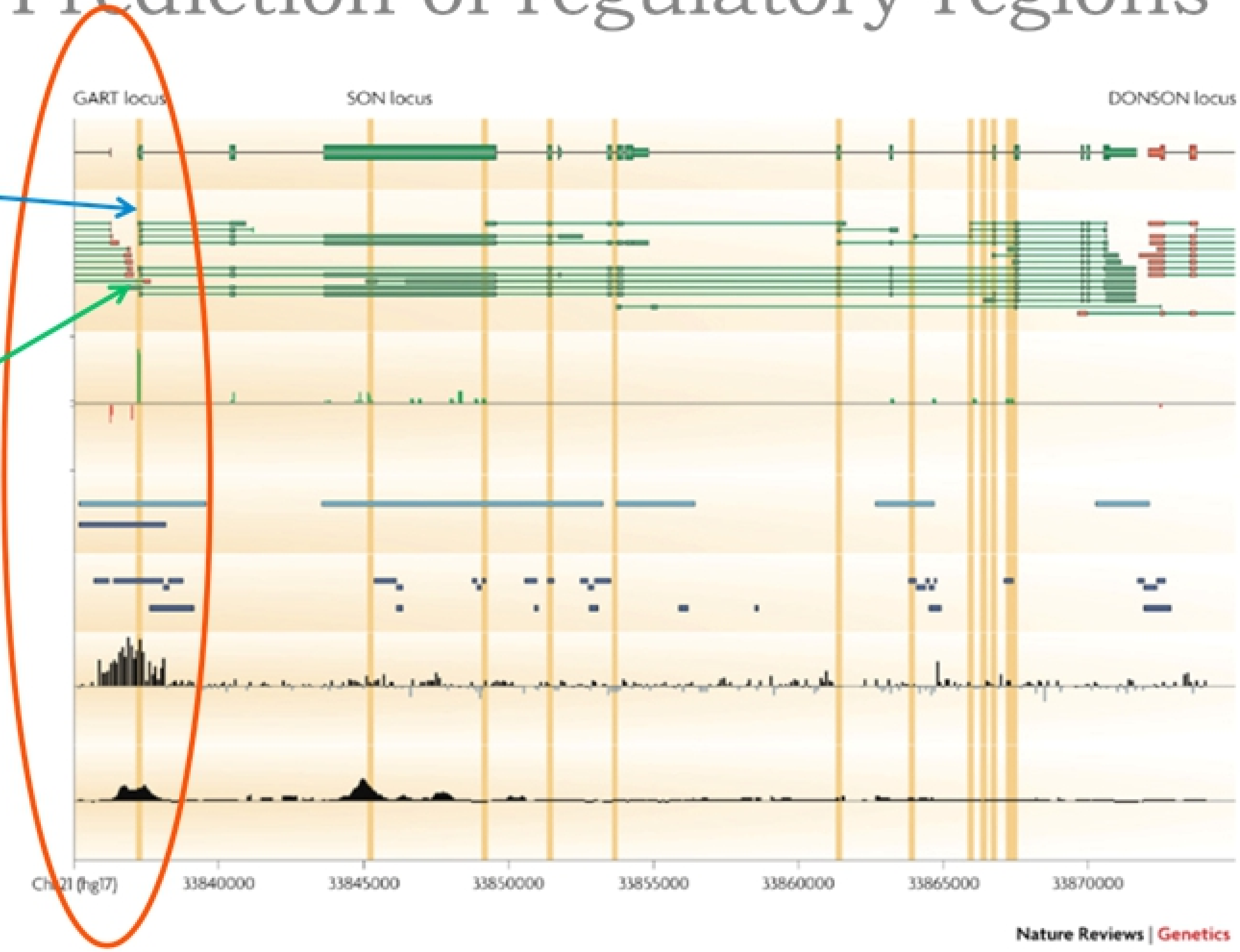


CG skew at TSS

$$\frac{C-G}{C+G}$$



Prediction of regulatory regions



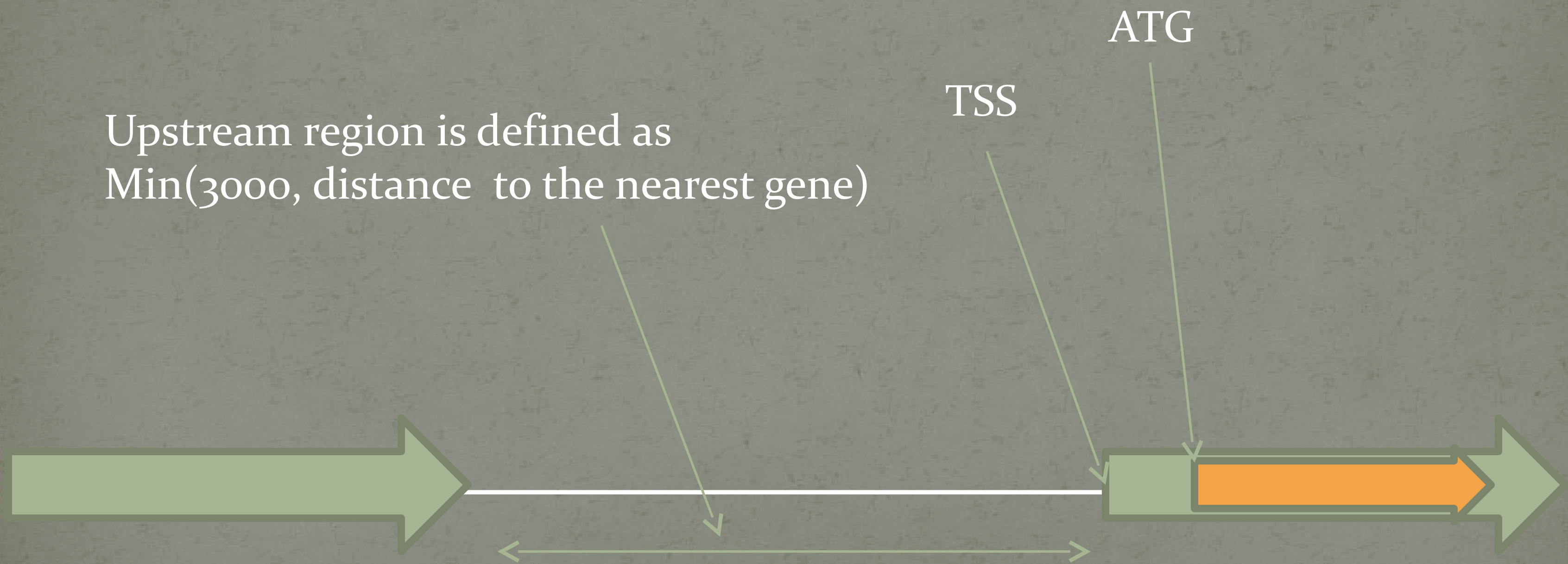
Mode/mean/median of TSS distribution

Longest transcript

Nature Reviews | Genetics

Goal

- Using experimental evidences of TSS, predict location of promoter



ATG

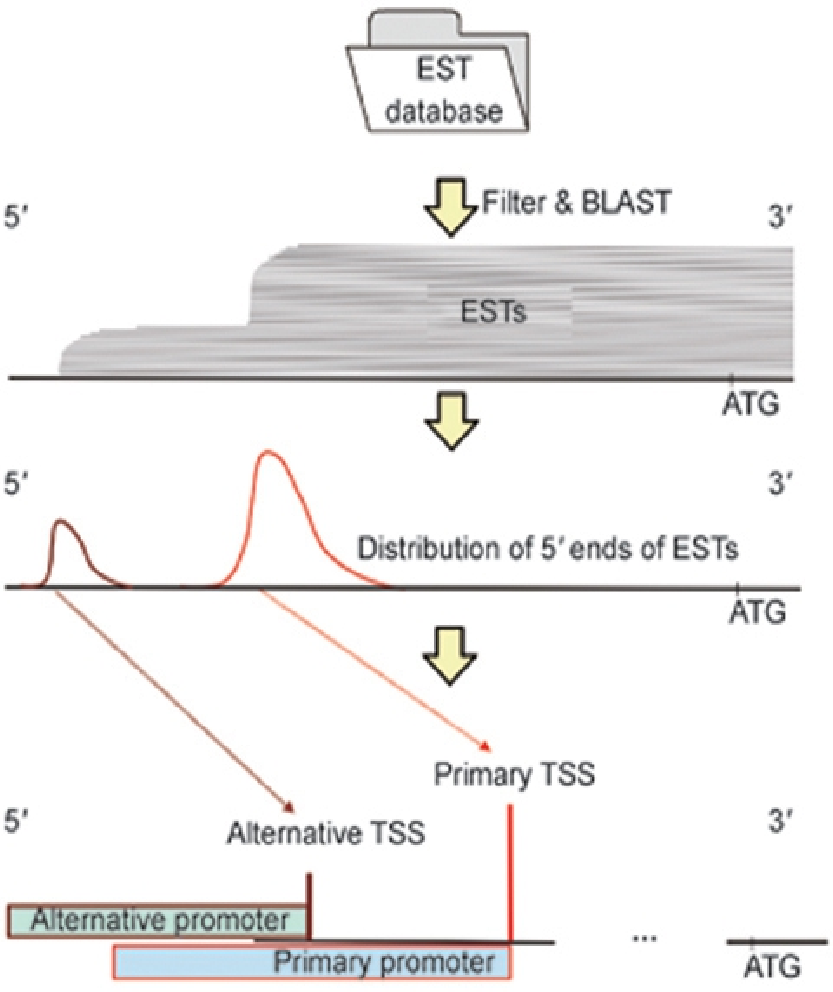
TSS

Upstream region is defined as
Min(3000, distance to the nearest gene)

Procedure

- Get 3000 nucleotides upstream from ATG from TAIR (27K sequences)
- Truncate sequences based on the position of the nearest upstream locus.
- 290K EST sequences were obtained from NCBI and TAIR and mapped onto the upstream sequences using BLAST+ (minimum identity percent: 95%, maximum query start of alignment: 5, only plus strand alignments were used). Using the text search we removed ESTs annotated as or *partial*.

Procedure



- Using Nonparametric Maximum Likelihood method estimate distribution of 5' ends of EST on the genome
- Find the number of modes
- Classify TSS into primary (highest) and alternatives

Model

- Population model:

$$Y_i \sim p_i(Y_i \mid \theta_i) \sim Binomial(n, p_i = \frac{\theta_i}{n}),\ i = 1,\ldots,N$$

$$\theta_i \sim F$$

Y_i - is a position of 5’ end of ESTs starting from ATG

$\theta_i = np_i$ - an unknown position of TSS. We consider θ_i’s to be iid with common but unknown distribution function F

p_i - is the probability of success, where success is considered to be an absence of EST at a given nucleotide of the n nucleotide-long promoter

n - is the length of the upstream region

N – is a number of EST per locus

- Problem: Estimate F given $Y_1,\ldots,Y_N$

Nonparametric Maximum Likelihood Estimation

- F is any distribution function on Ω

- Log-Likelihood Function:

$$l(F) = \sum_{i=1}^{N} \log(p_i(Y_i \mid F))$$

where $p_i(Y_i \mid F) = \int p_i(Y_i \mid \theta)\, dF(\theta)$, and $p_i(Y_i \mid \theta) = N(h_i(\theta), \sigma_i^2 I)$

- MLE problem: F^{ML} maximizes $l(F)$ over all probability distributions on Ω

NPEST

- Equivalent problem (Lindsay, 1983; Mallet, 1986):

  It is shown by Lindsay and Mallet that F^{ML} belongs to the set of discrete distributions with support at no more than N points, i.e.

$$F^{ML} \in \{F = \sum_{k=1}^{K} w_k \delta_{\phi_k} : \quad K \le N,\ \phi_k \in \Omega,\ w_k \ge 0,\ \sum_{k=1}^{K} w_k = 1\}$$

- We use iterative method based on EM algorithm, which is described in (Schumitzky, 1991) to obtain the numerical solution of MLE

- NPEST

  Let $\lambda = (\theta_1, \ldots, \theta_K, w_1, \ldots, w_K)$

  Step 1. Initiate: $\lambda = \lambda^{(0)}$

  Step 2. Update: for all $k = 1, \ldots, K$

$$\theta_k^{(n+1)} = \arg\max\{\sum_{i=1}^{N} p(\theta_k^{(n)} \mid Y_i, \lambda^{(n)}) \log(p_i(Y_i \mid \theta)) : \theta \in \Theta\}$$

$$w_k^{(n+1)} = \frac{1}{N} \sum_{i=1}^{N} p(\theta_k^{(n)} \mid Y_i, \lambda^{(n)})$$

where $$p(\theta_k^{(n)} \mid Y_i, \lambda^{(n)}) = \frac{w_k^{(n)} p_i(Y_i \mid \theta_k^{(n)})}{\sum_{k=1}^{K} w_k^{(n)} p_i(Y_i \mid \theta_k^{(n)})}$$

Step 3. If $| l(\lambda^{(n)}) - l(\lambda^{(n+1)}) | < \varepsilon$, stop. Otherwise, go to step 2.

Convergence results

- Theorem 1 (Lindsay, 1983):

  Define the function

  $$D(\theta, F) = \sum_{i=1}^{N} \frac{p(y_i|\theta)}{p(y_i|F)} - N$$

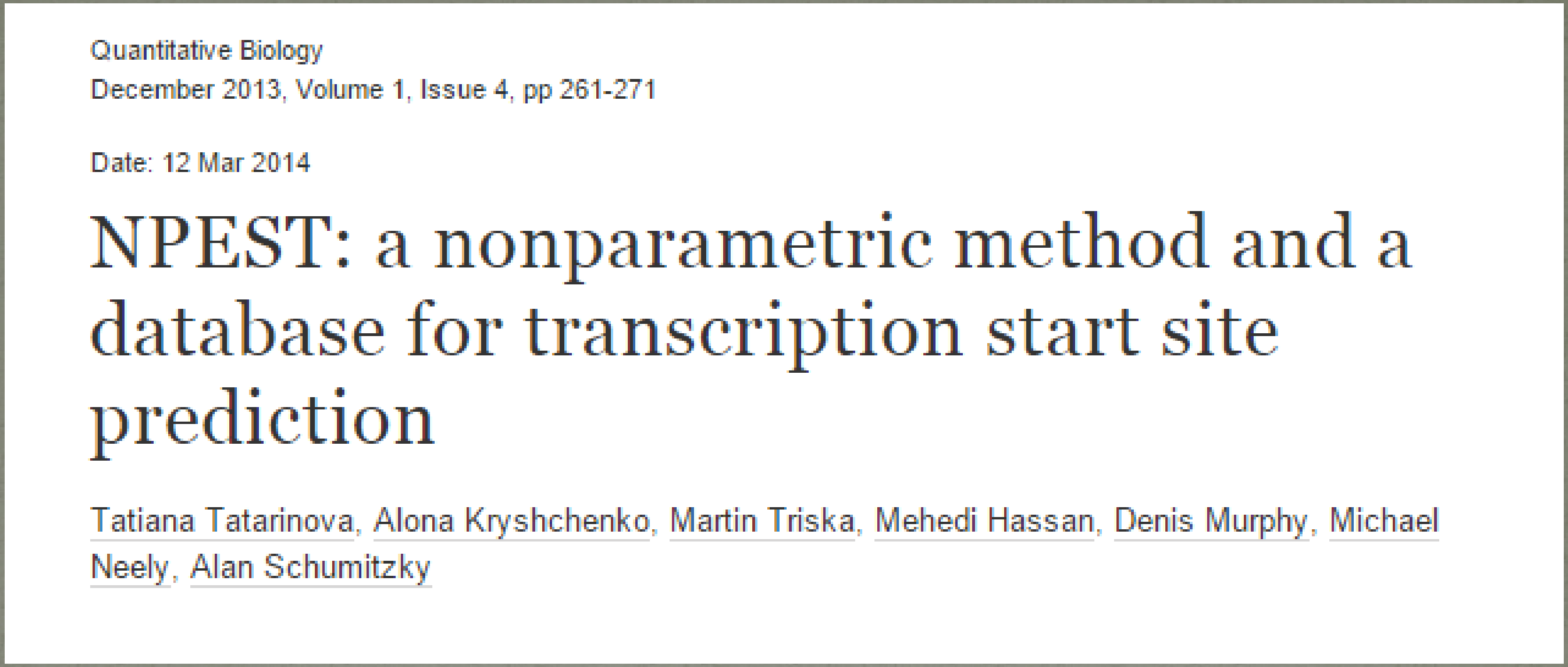
  Then

  1. F^{ML} *maximizes* $l(F)$ *if and only if* $\max\{D(\theta, F^{ML}) : \theta \in \Omega\} = 0$

  2. The support of F^{ML} *is contained in the set* $\{\theta : D(\theta, F^{ML}) = 0\}$

- Post-processing of the results

  The goal of this step is to obtain smoothed versions of FML, find the number of peaks, and remove peaks that are supported by less than the preset fraction of ESTs. The R routine findpeaks from the package pracma is applied to this smoothed distribution of FML(φ) to identify the number and positions of peaks.

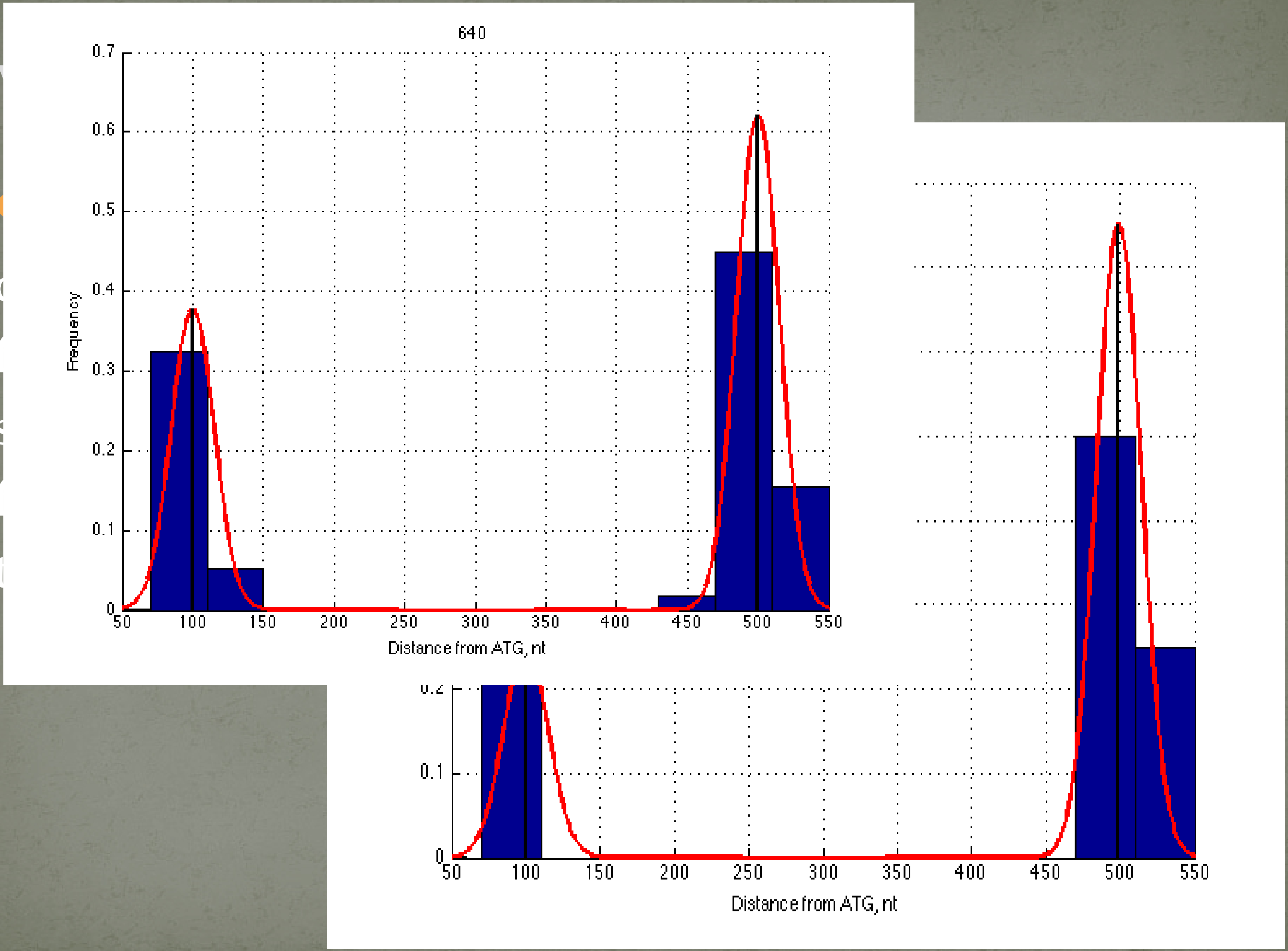


Quantitative Biology
December 2013, Volume 1, Issue 4, pp 261-271

Date: 12 Mar 2014

NPEST: a nonparametric method and a database for transcription start site prediction

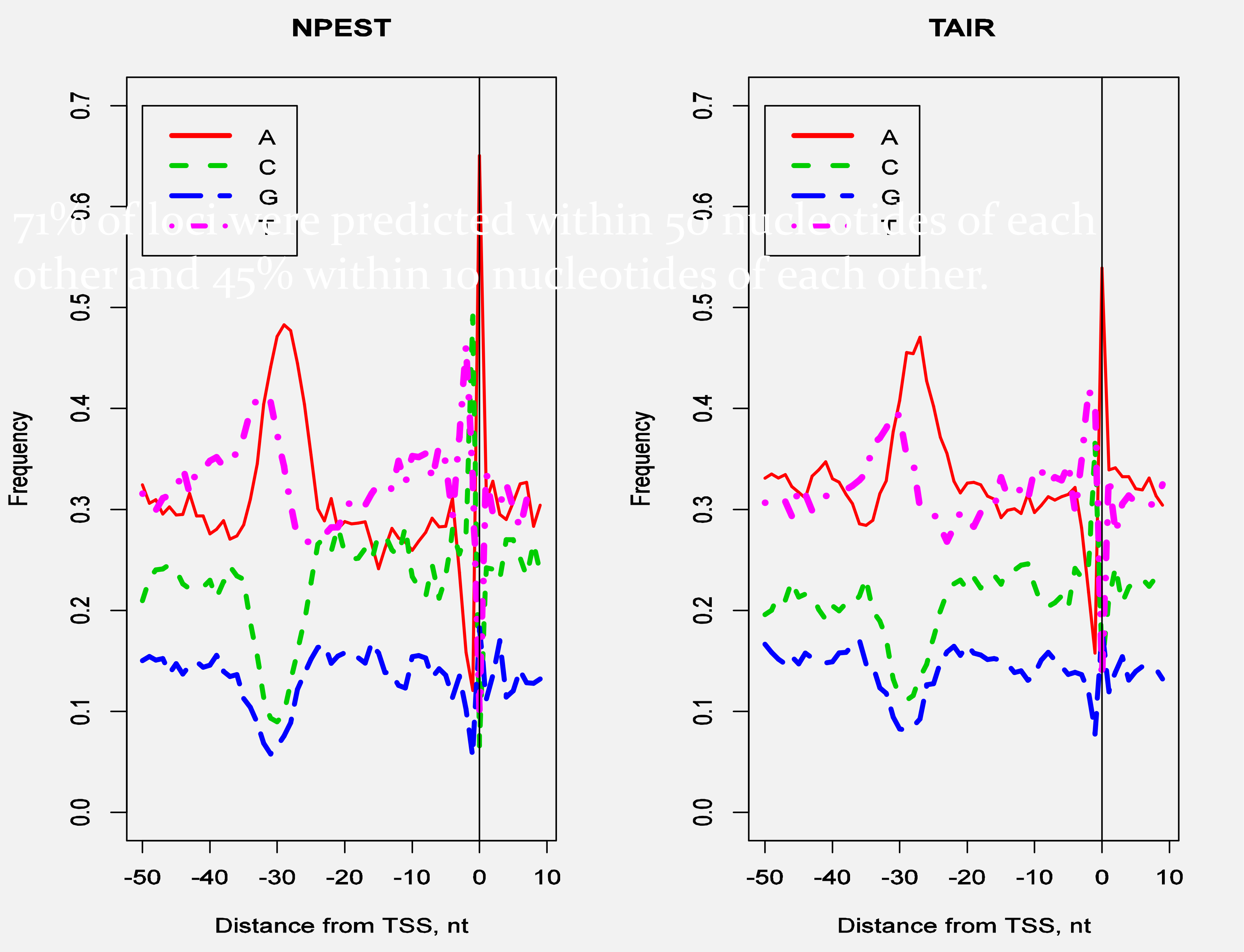
Tatiana Tatarinova, Alona Kryshchenko, Martin Triska, Mehedi Hassan, Denis Murphy, Michael Neely, Alan Schumitzky


640
Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
300
350
400
450
500
550
Distance from ATG, nt
0.2
0.1
0
50
100
150
200
250
300
350
400
450
500
550
Distance from ATG, nt

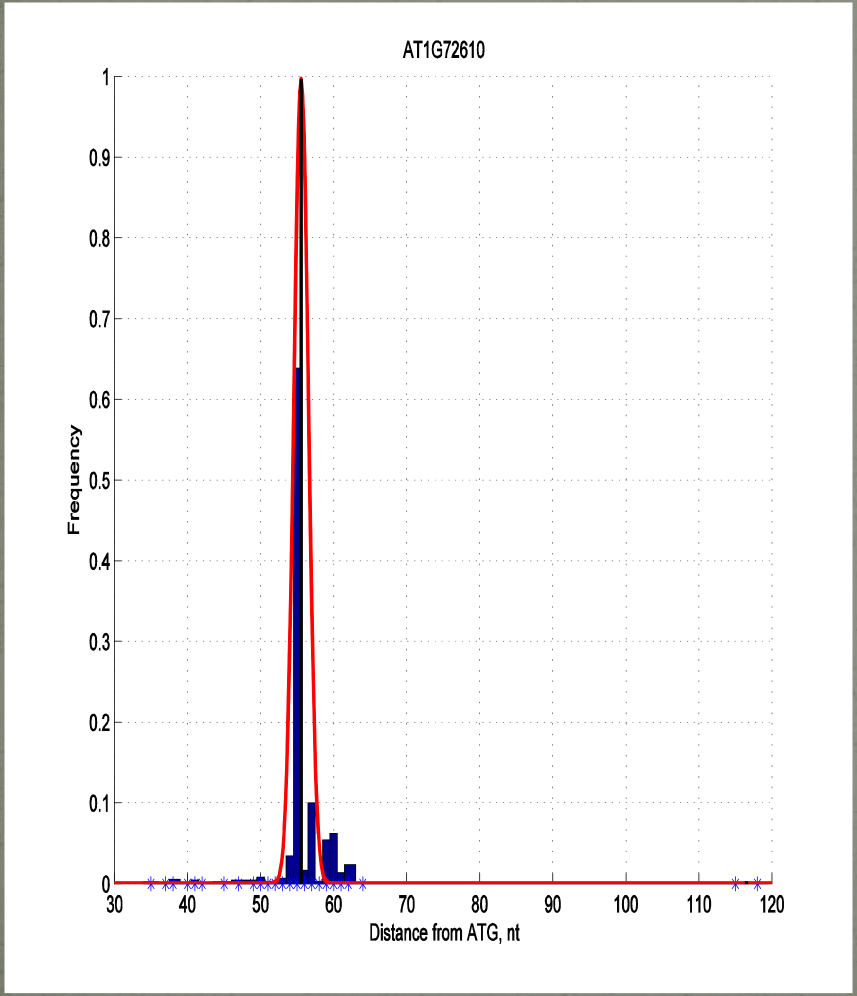
Next, biological validation

- We applied NPEST to *Arabidopsis thaliana*. Using NPEST, we predicted TSS for 16,520 loci in *Arabidopsis*.
- Two aspects for assessment:
  - presence of characteristic motifs at TSS (e.g. TATA-box at -30 and CA di-nucleotide at TSS)
  - Agreement between NPEST and previously published results, such as TAIR, PlantProm DB, PlantPromoter Database, and Pol II occupancy data. Use the "main" TSS for comparison.

71% of loci were predicted within 50 nucleotides of each other and 45% within 10 nucleotides of each other.



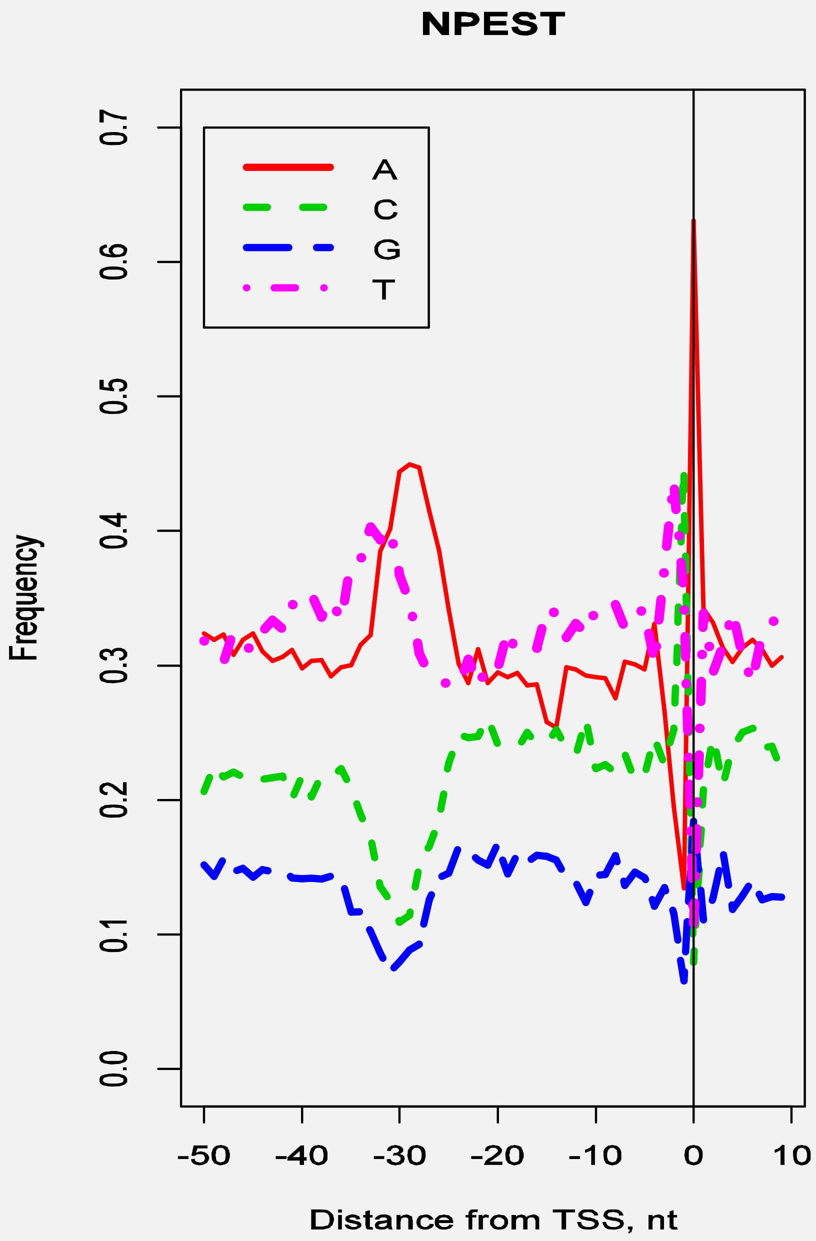
Example *AT1G72610*

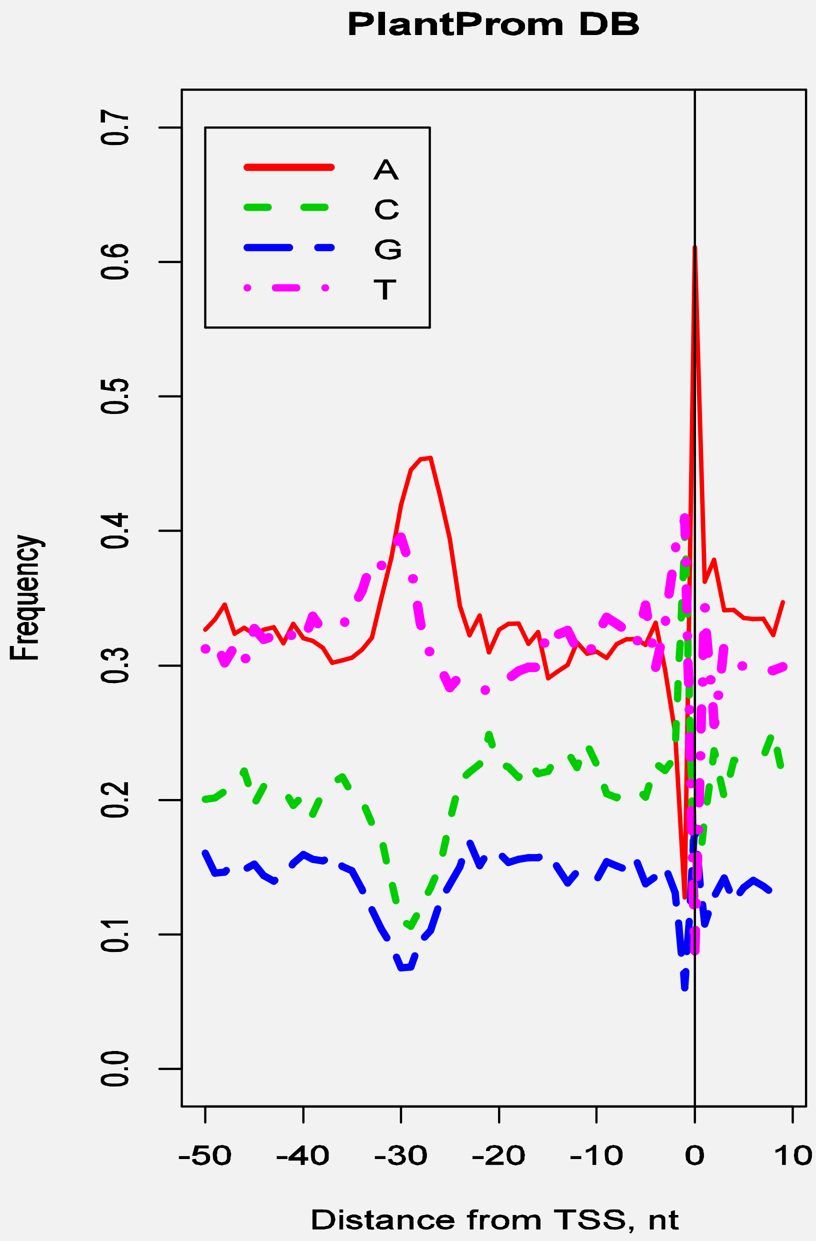


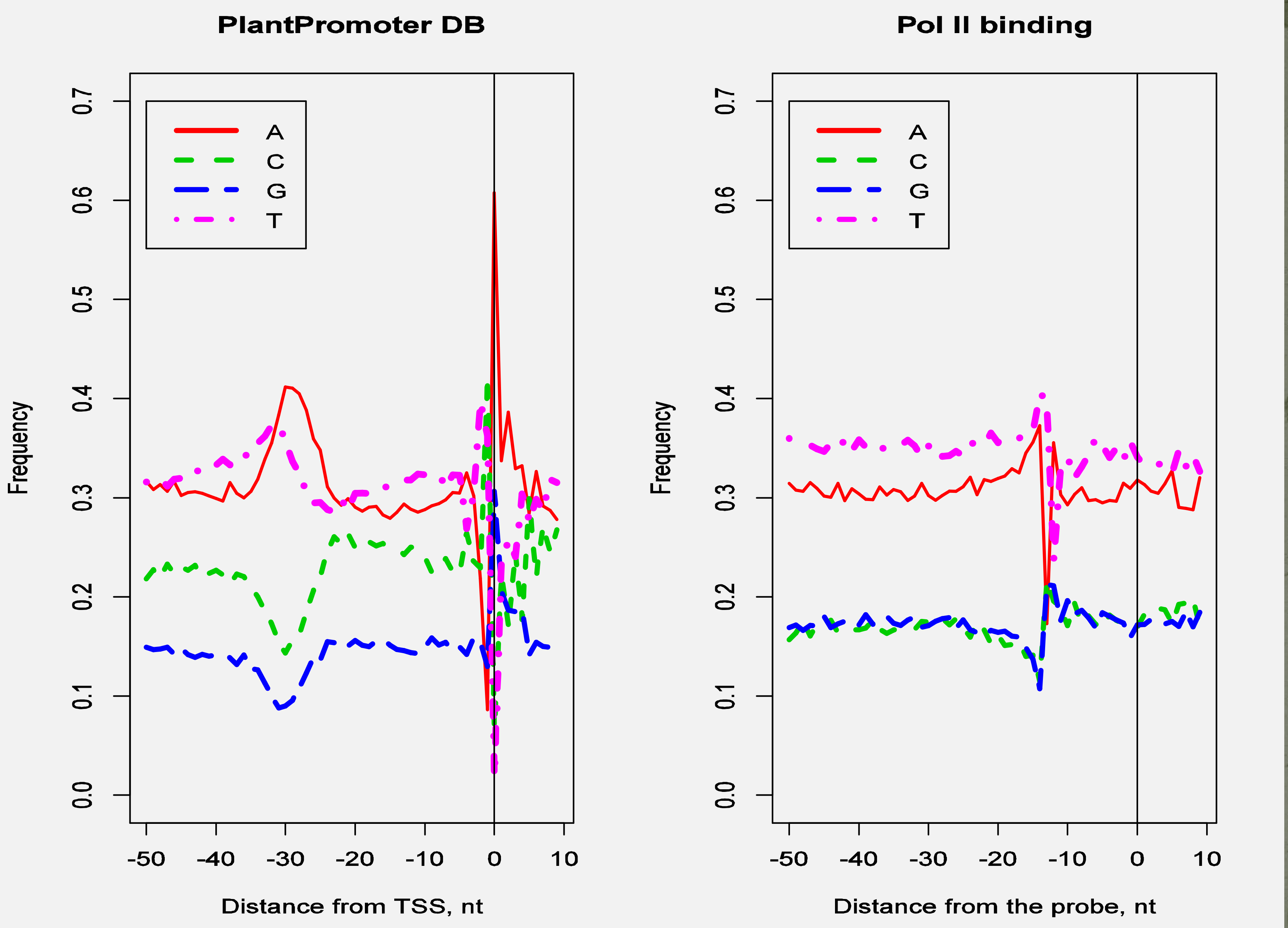
According to TAIR, the 5' UTR is 116 nucleotides long; according to NPEST, there are two peaks. The major peak is 55 nucleotides (as supported by 64% of the ESTs mapped to this locus) and the minor peak 116 nucleotides upstream from ATG

Example AT1G72610

- For the TSS at 55 nt, the sequence has a very strong canonical TATA-box (“CTATATAAA”) at -37 nt upstream from the TSS:
- *tcccacacctctCTATATAAAcacccgagaccgagaggagtgagaagagtagggaaaaag*
- For the TSS at 116 nt, the sequence is equipped with the TATA-like motif “CTAAAA” at position -33:
- *gacgtccataatggtttCTAAAAgcttatctccgtctttcgaatgttcaccacacagttt*
- Note, that there are 3 versions of the TATA-box (“TATA”, “CTAT” and “TAAA”) in the first case


NPEST
A
C
G
T
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
-50
-40
-30
-20
-10
0
10
Distance from TSS, nt


PlantProm DB
A
C
G
T
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
-50
-40
-30
-20
-10
0
10
Distance from TSS, nt

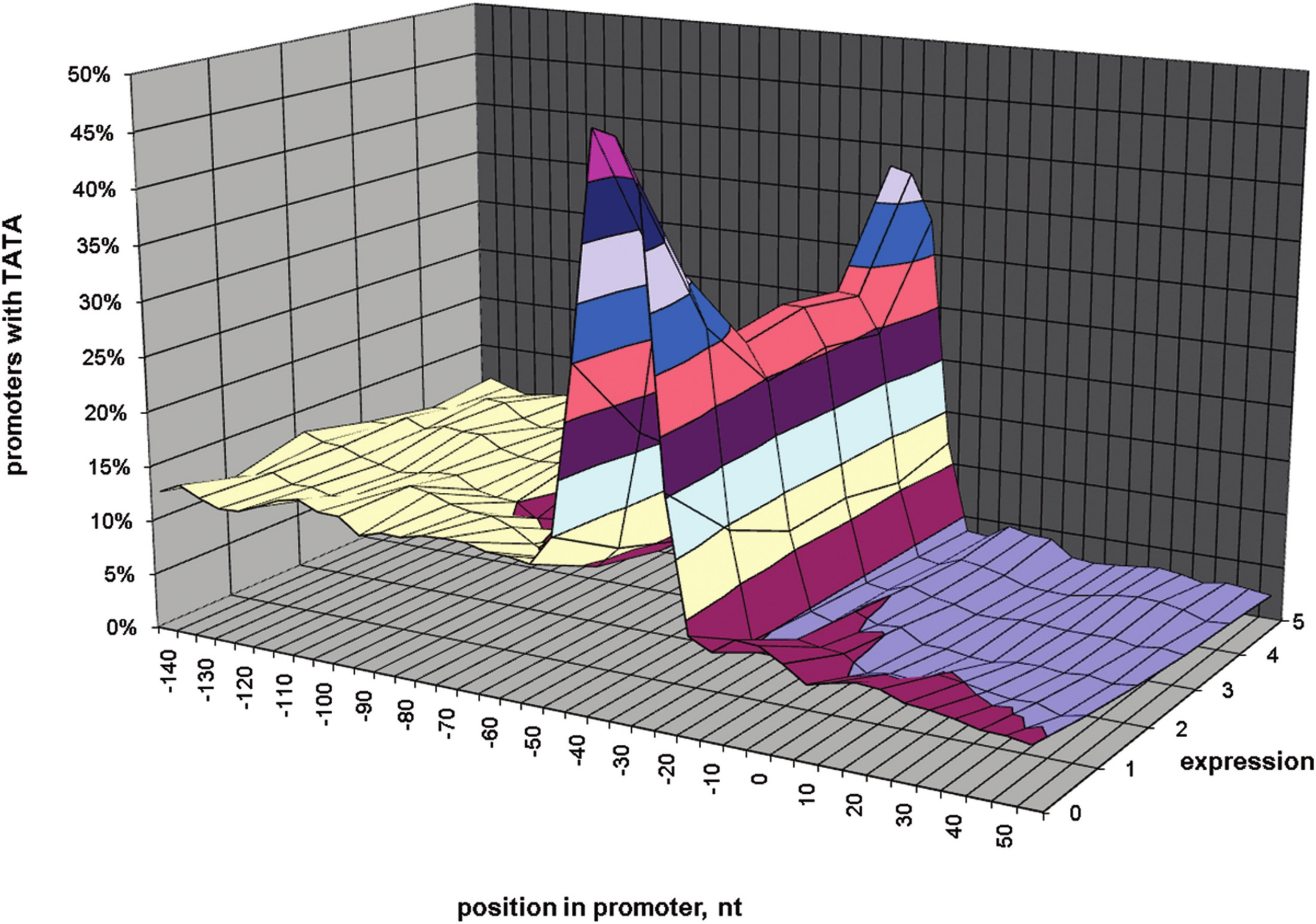

PlantPromoter DB
Pol II binding
A
C
G
T
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
-50
-40
-30
-20
-10
0
10
Distance from TSS, nt
Distance from the probe, nt

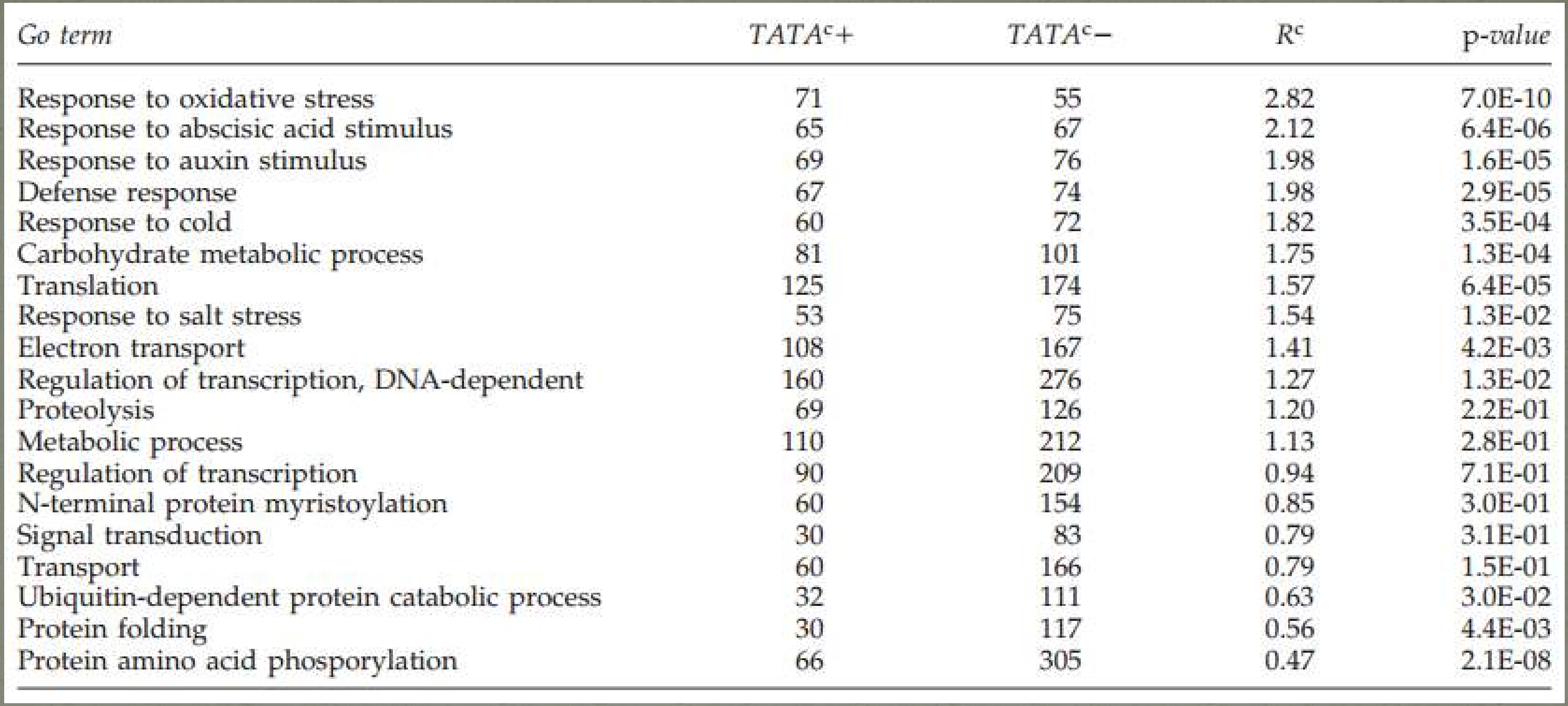
Alternative TSSs

- Using EST library annotation information, we have assigned each EST to one of the 40 categories based on the library (e.g. “Shoots”, “Roots”, “Drought” etc.). A separate category was reserved for ESTs without library information. There are 7,549 loci that have one category of EST assigned to them and 5,281 with two categories. On average, one-category loci have 1.43 alternative TSSs and two-category loci have 2.43 TSSs.

Possible bias of the method

- TATA-box is more prevalent in stress-related genes
- Stress-related genes have more TATA-boxes than house-keeping ones.
- Stress-related genes have more defined TSS, and house-keeping genes may have TSR


50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
promoters with TATA
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
0
10
20
30
40
50
position in promoter, nt
0
1
2
3
4
5
expression

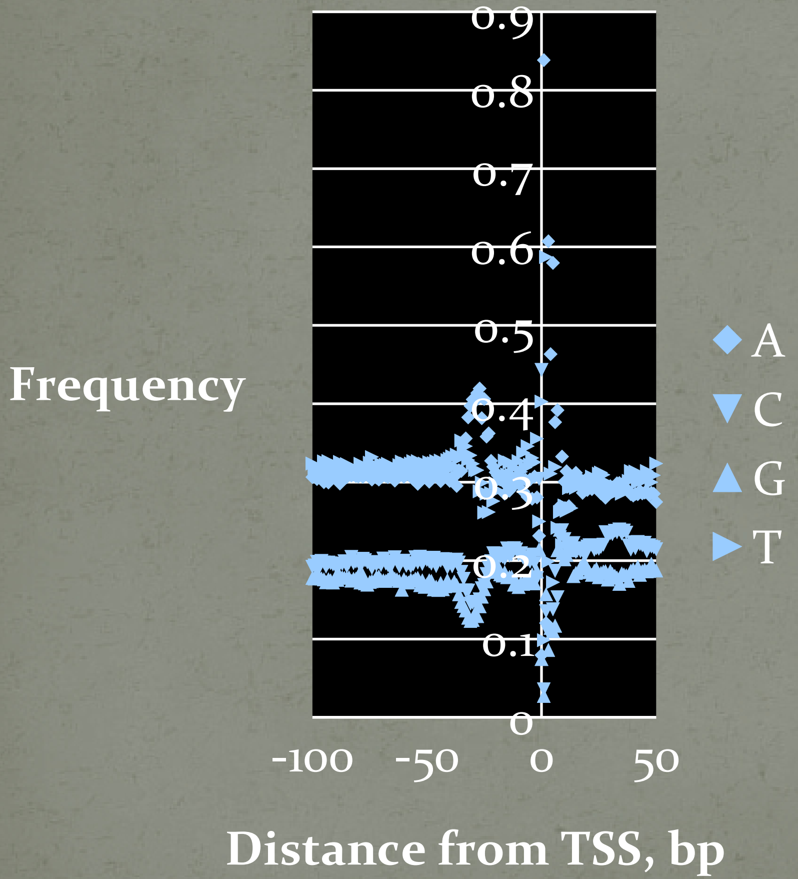


Go term	TATA^c+	TATA^c−	R^c	p-value
Response to oxidative stress	71	55	2.82	7.0E-10
Response to abscisic acid stimulus	65	67	2.12	6.4E-06
Response to auxin stimulus	69	76	1.98	1.6E-05
Defense response	67	74	1.98	2.9E-05
Response to cold	60	72	1.82	3.5E-04
Carbohydrate metabolic process	81	101	1.75	1.3E-04
Translation	125	174	1.57	6.4E-05
Response to salt stress	53	75	1.54	1.3E-02
Electron transport	108	167	1.41	4.2E-03
Regulation of transcription, DNA-dependent	160	276	1.27	1.3E-02
Proteolysis	69	126	1.20	2.2E-01
Metabolic process	110	212	1.13	2.8E-01
Regulation of transcription	90	209	0.94	7.1E-01
N-terminal protein myristoylation	60	154	0.85	3.0E-01
Signal transduction	30	83	0.79	3.1E-01
Transport	60	166	0.79	1.5E-01
Ubiquitin-dependent protein catabolic process	32	111	0.63	3.0E-02
Protein folding	30	117	0.56	4.4E-03
Protein amino acid phosporylation	66	305	0.47	2.1E-08

For every category "c"

$$R^c = \frac{TATA_+^c}{TATA_-^c}\frac{TATA_-}{TATA_+}$$

Other species: oil palm



- 38% of promoters have a canonical TATA-box (“tata”) around between TSS-40, TSS-20, and 62% have different variations of thee TATA-box (i.e. “taaa”, “ctat”, etc). These frequencies are typical for high-quality TSS predictions and provide additional indication that the annotation is of high quality

What do we want to do next?

- Combine other evidences of TSS positions using Naïve Bayes approach
- In addition to EST distributions with
  - Pol II binding
  - DNA methylation
  - Nucleosome position
  - Distribution of UTR length
  - Function of the gene
  - Known motifs in promoter

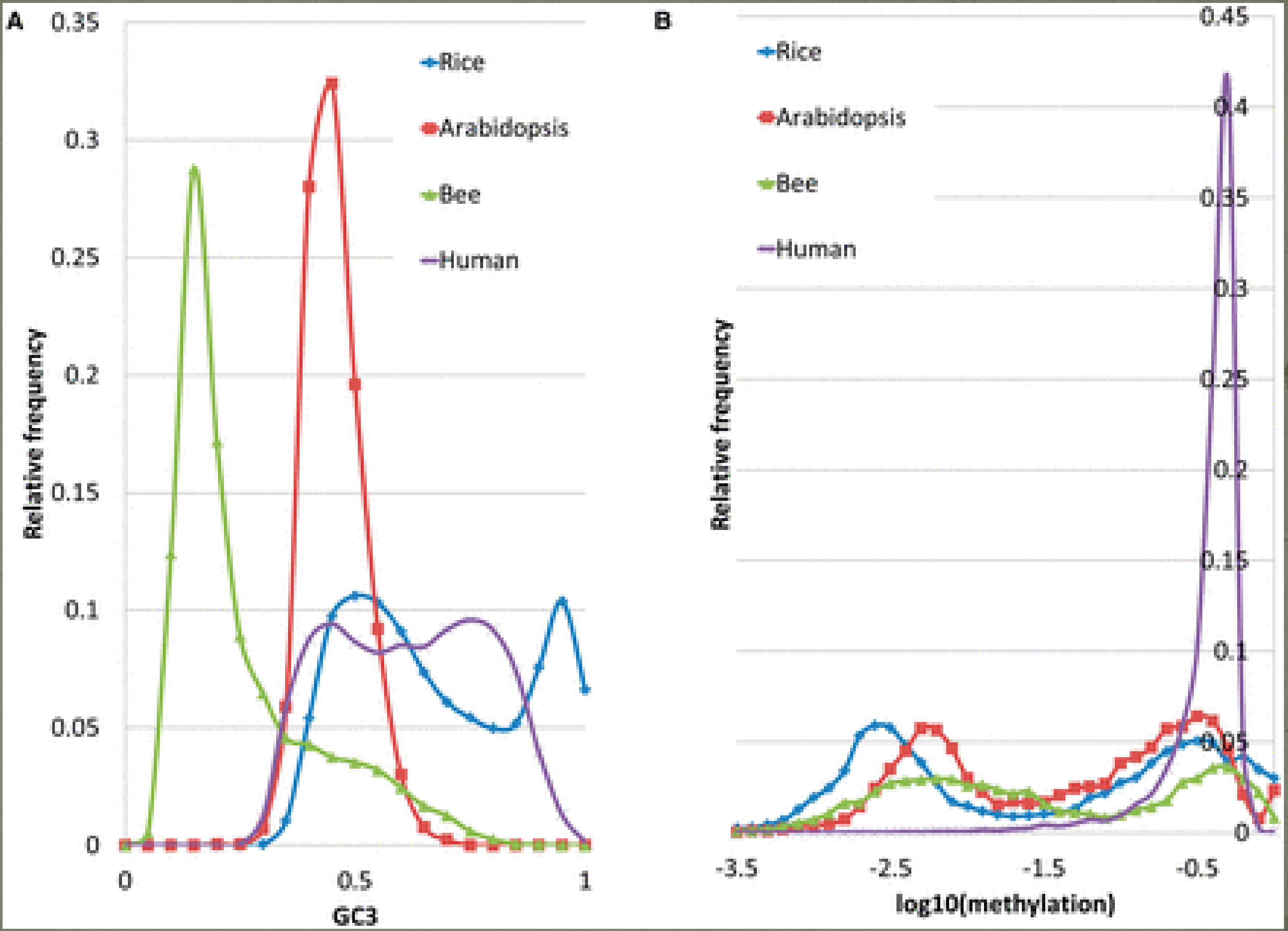

A
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Relative frequency
0
0.5
1
GC3
Rice
Arabidopsis
Bee
Human
B
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Relative frequency
-3.5
-2.5
-1.5
-0.5
log10(methylation)
Rice
Arabidopsis
Bee
Human

Figure S3: *Oryza sativa (A), Arabidopsis thaliana (B), Homo sapiens (C)* and *Apis mellifera (D)*. Gradient of CpG methylation along CDS. Blue curve corresponds to GC$_3$-rich genes and red curve corresponds to GC$_3$-poor genes. 10% of genes from two extreme ends of GC$_3$ distribution selected to represent GC$_3$-rich and -poor datasets. Every point represents an average across approximately 1000 genes.

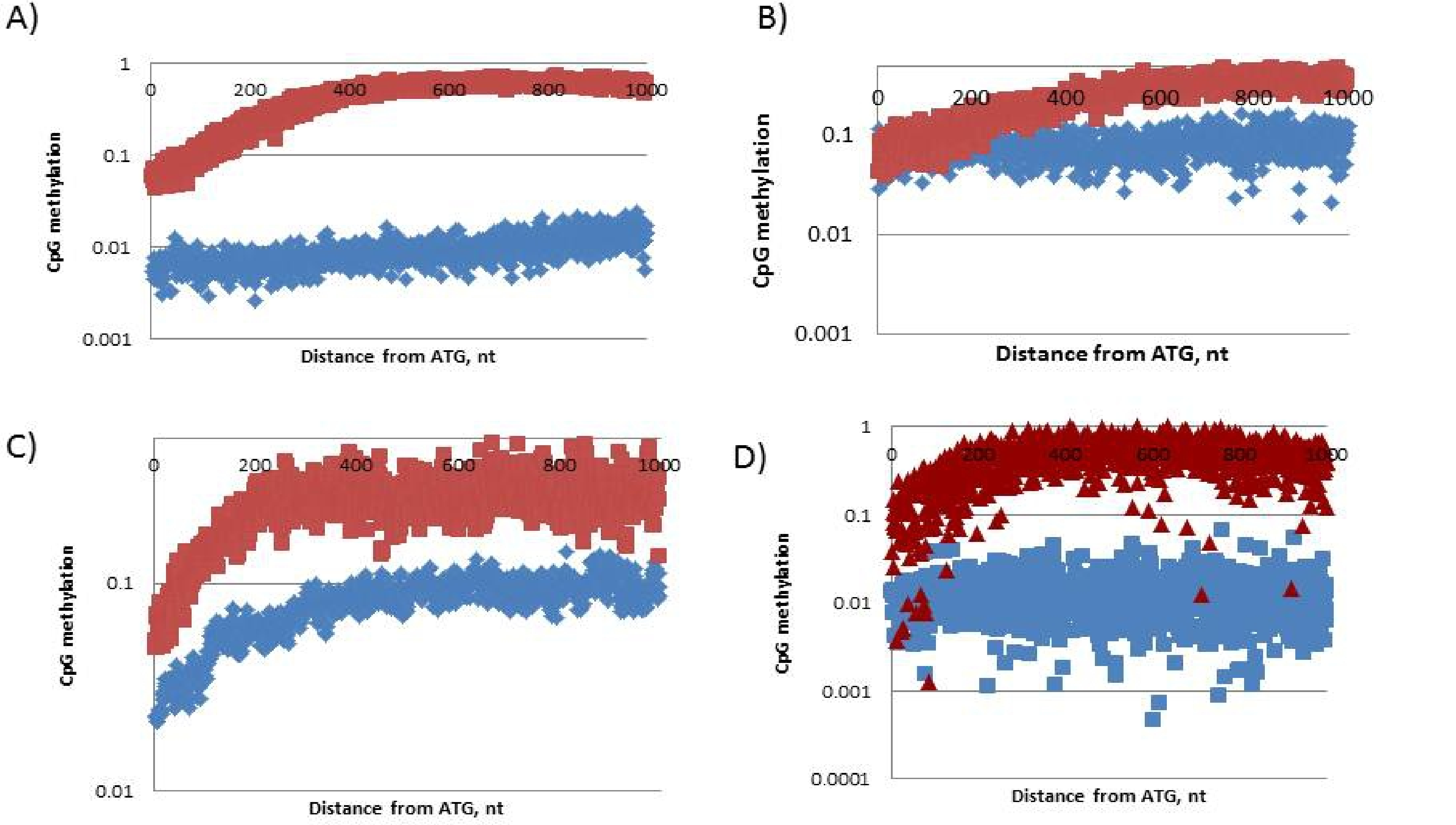
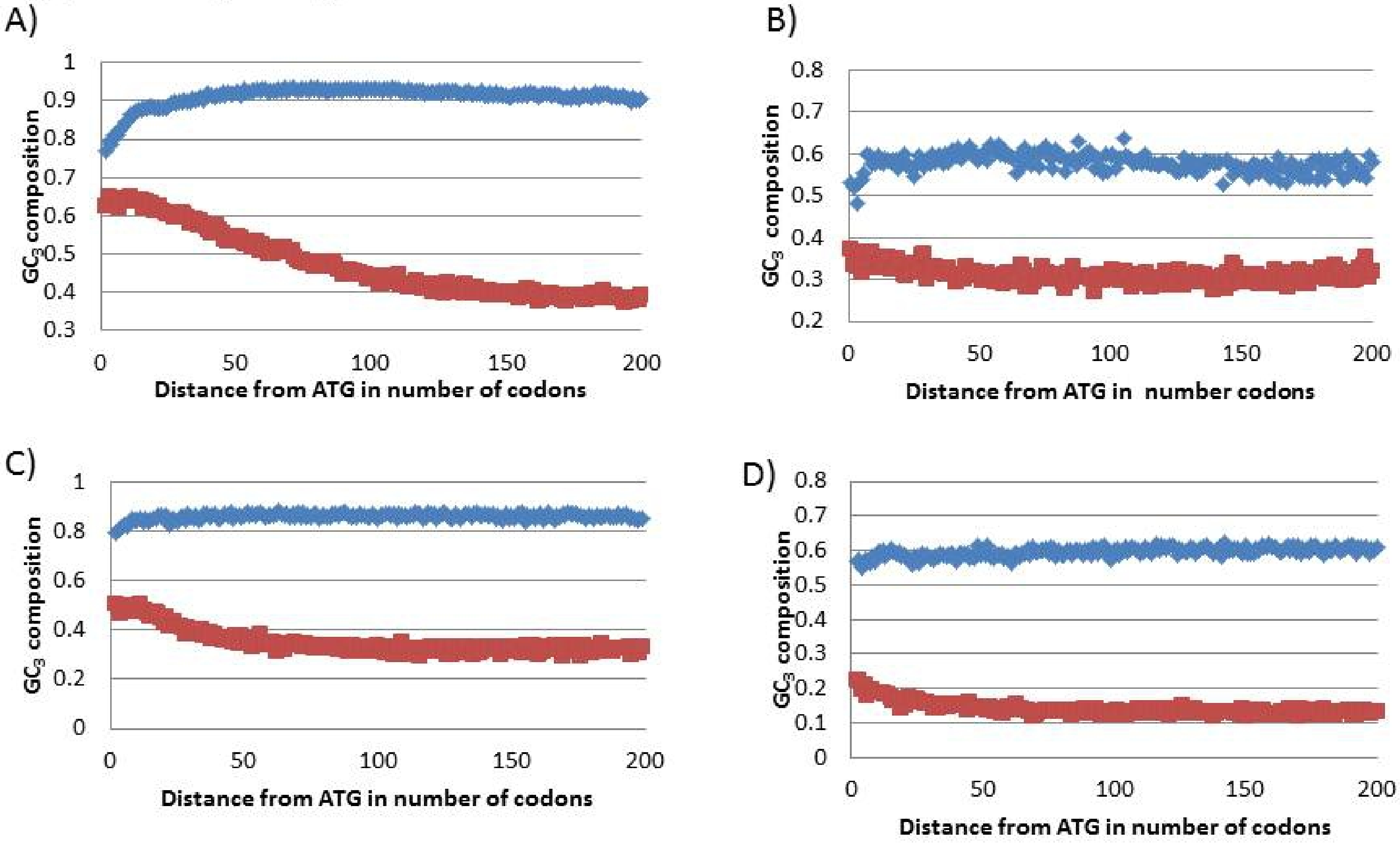
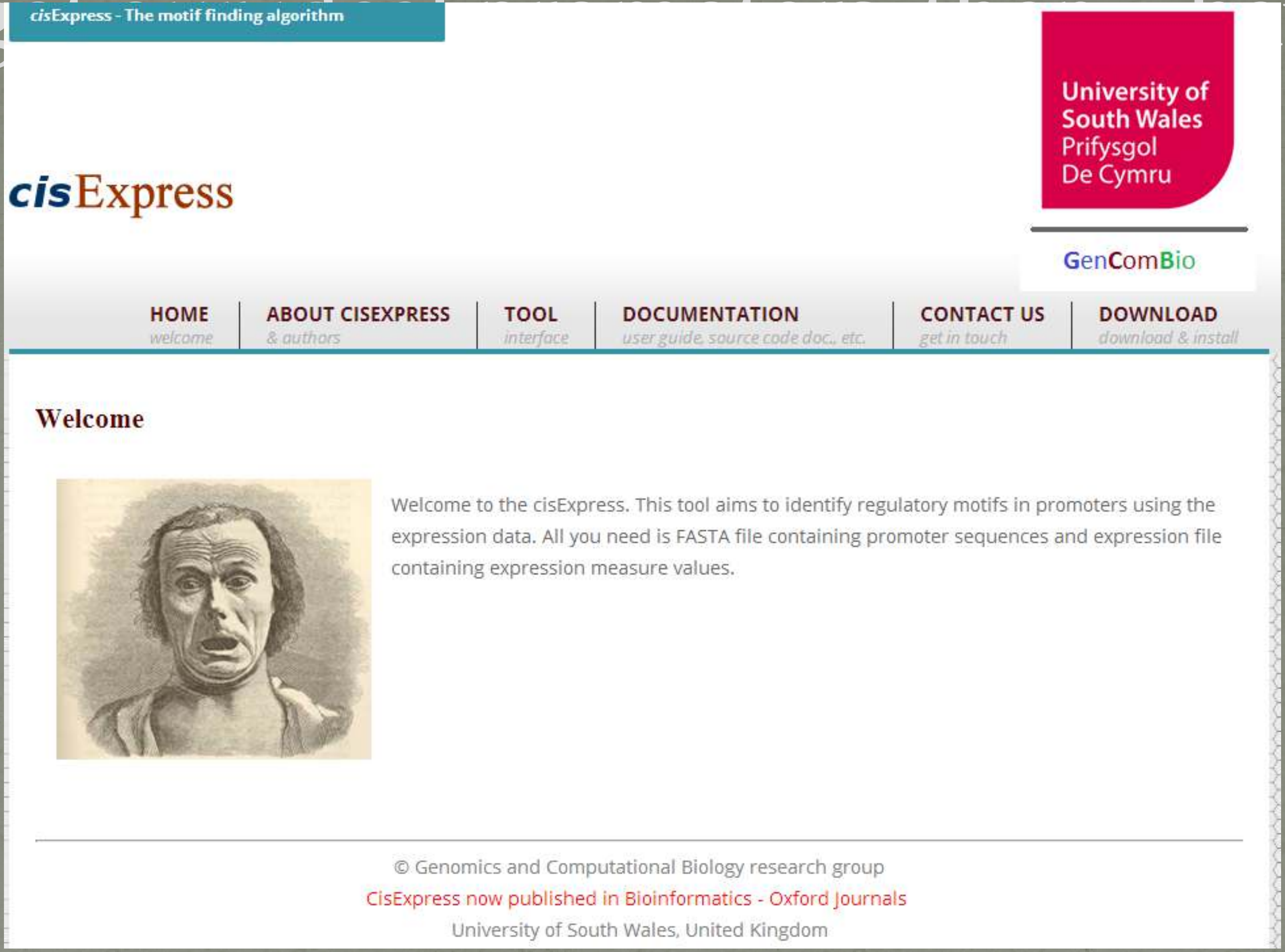


Figure S1*: Oryza sativa (A), Arabidopsis thaliana (B), Homo sapiens (C)* **and** *Apis mellifera (D)*. **GC$_3$ gradient from 5' to 3' ends of coding regions. Blue curve corresponds to GC$_3$-rich genes and red curve corresponds to GC$_3$-poor genes. Every point represents an average across approximately 1000 genes.**



So, we g t?



*cis*Express - The motif finding algorithm

cisExpress

University of South Wales
Prifysgol De Cymru

GenComBio

Welcome

Welcome to the cisExpress. This tool aims to identify regulatory motifs in promoters using the expression data. All you need is FASTA file containing promoter sequences and expression file containing expression measure values.

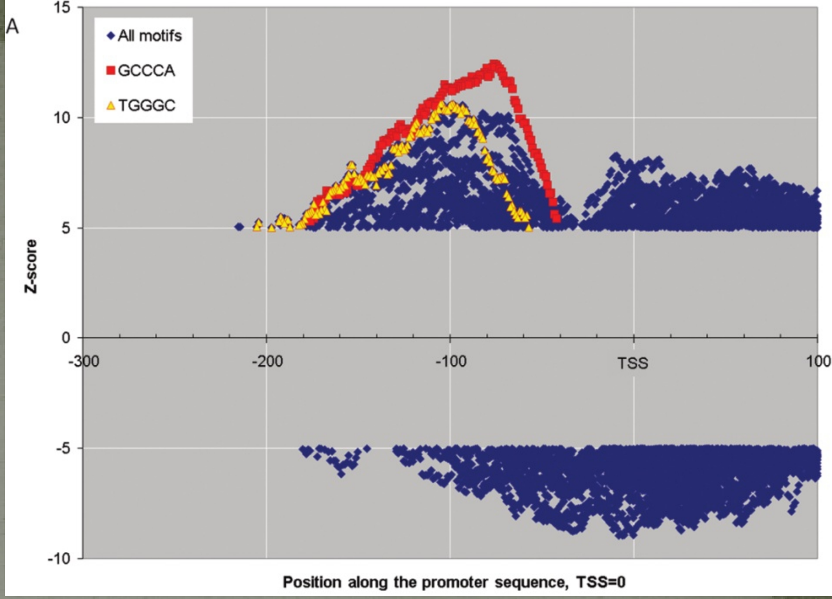
© Genomics and Computational Biology research group
CisExpress now published in Bioinformatics - Oxford Journals
University of South Wales, United Kingdom

http://chcb.saban-chla.usc.edu/cisExpress/

Types of motif finding algorithms

- Cluster/alignment based
- Statistical overrepresentation
- Database driven
- Functional

What is *cis*Express



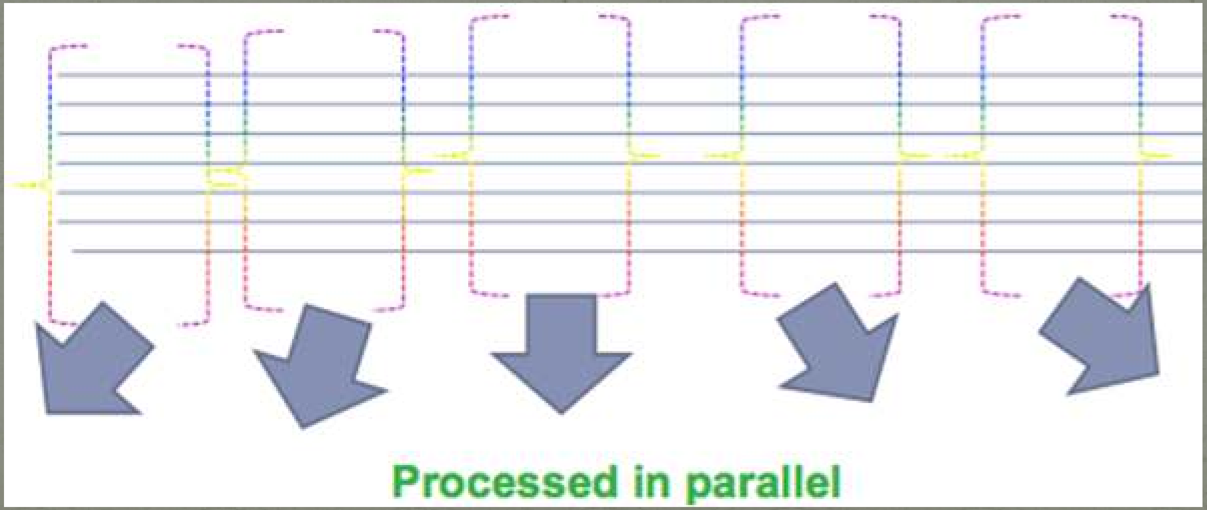
1. Motifs are position-specific
2. Influence of a motif w in position k onto gene expression e is given by:

$$Z_{score}(w,k) = \frac{e_{with}(w,k) - e_{without}(w,k)}{\sqrt{\dfrac{Stdev^2_{with}(w,k)}{n_{with}(w,k)} + \dfrac{Stdev^2_{without}(w,k)}{n_{without}(w,k)}}}$$

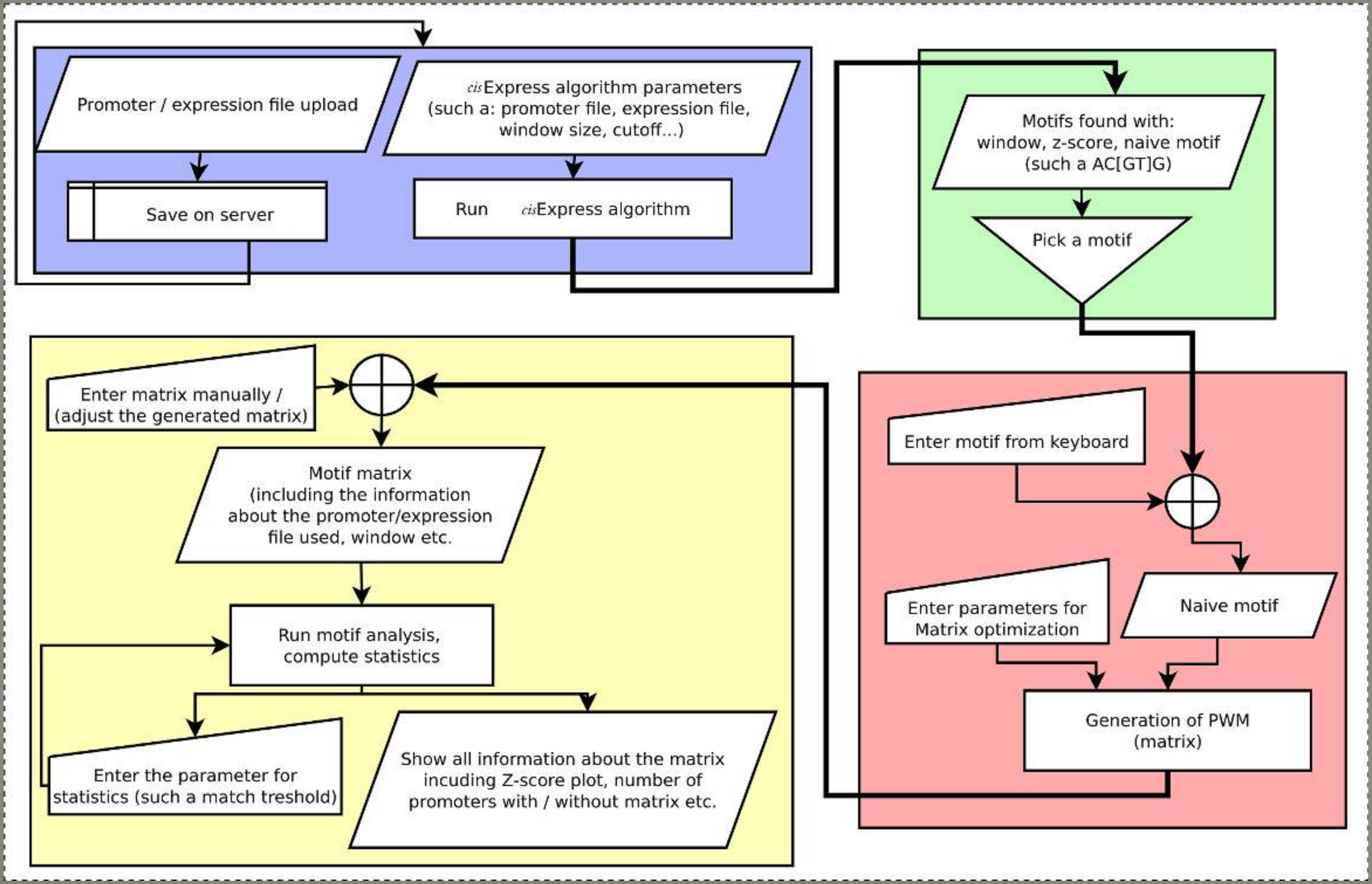
How does *cis*Express work?

Two principal stages

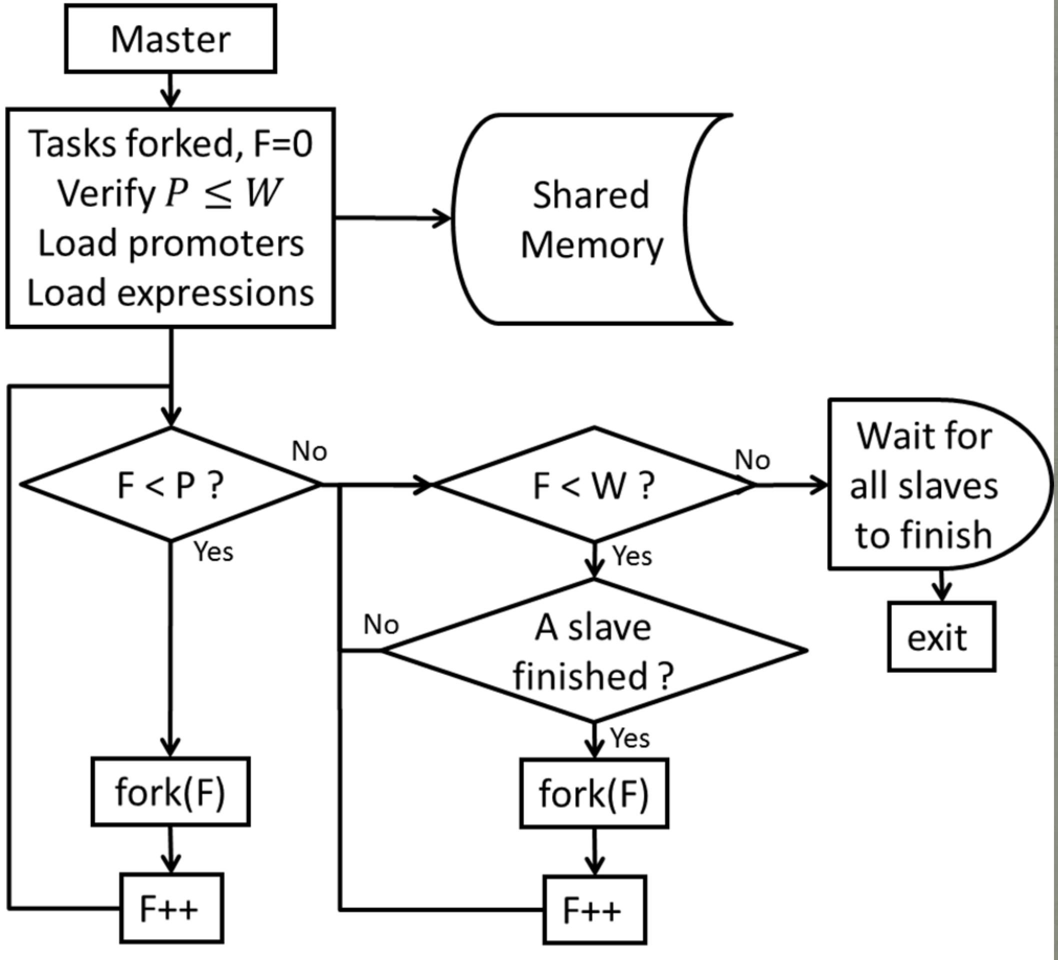
- Stage 1: Detecting 'seed' motifs, using Z-score
- Stage 2: Optimizing the previously obtained motifs using a genetic algorithm producing motif matrices. Similar motifs are merged.



cisExpress web interface flowchart

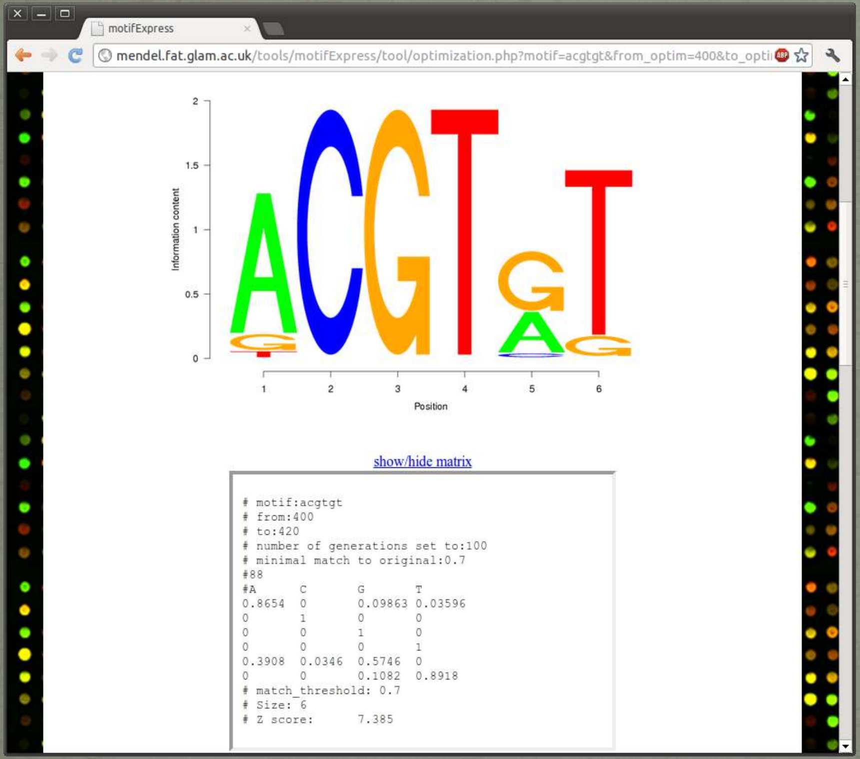

Promoter / expression file upload
cisExpress algorithm parameters (such a: promoter file, expression file, window size, cutoff...)
Save on server
Run cisExpress algorithm
Motifs found with: window, z-score, naive motif (such a AC[GT]G)
Pick a motif
Enter matrix manually / (adjust the generated matrix)
Motif matrix (including the information about the promoter/expression file used, window etc.
Run motif analysis, compute statistics
Enter the parameter for statistics (such a match treshold)
Show all information about the matrix incuding Z-score plot, number of promoters with / without matrix etc.
Enter motif from keyboard
Enter parameters for Matrix optimization
Naive motif
Generation of PWM (matrix)

Parallelization



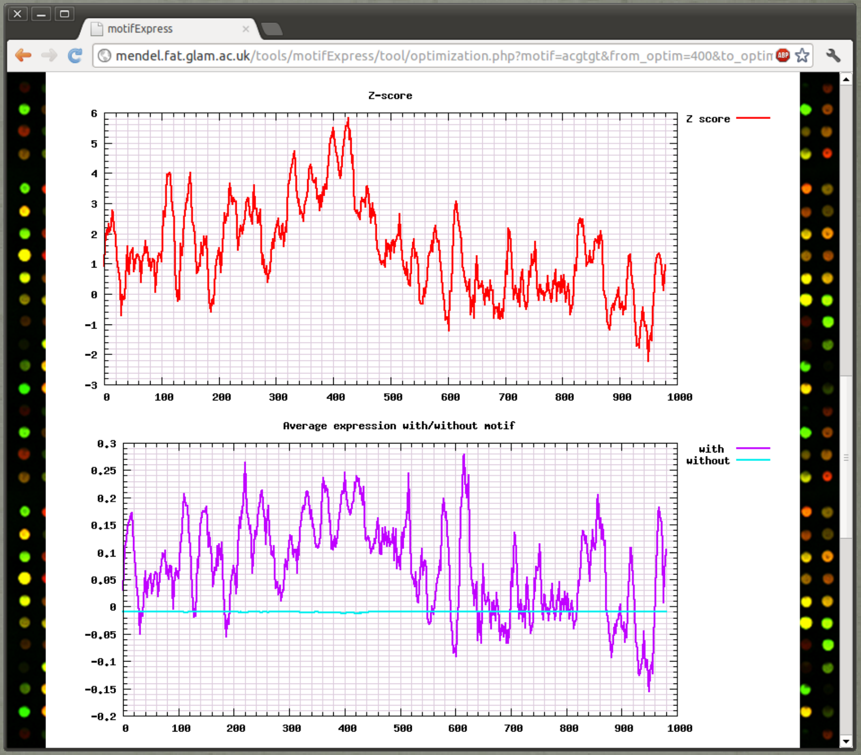
This is the original version, now we use OpenMPI

Screenshots



mendel.fat.glam.ac.uk/tools/motifExpress/tool/optimization.php?motif=acgtgt&from_optim=400&to_opti

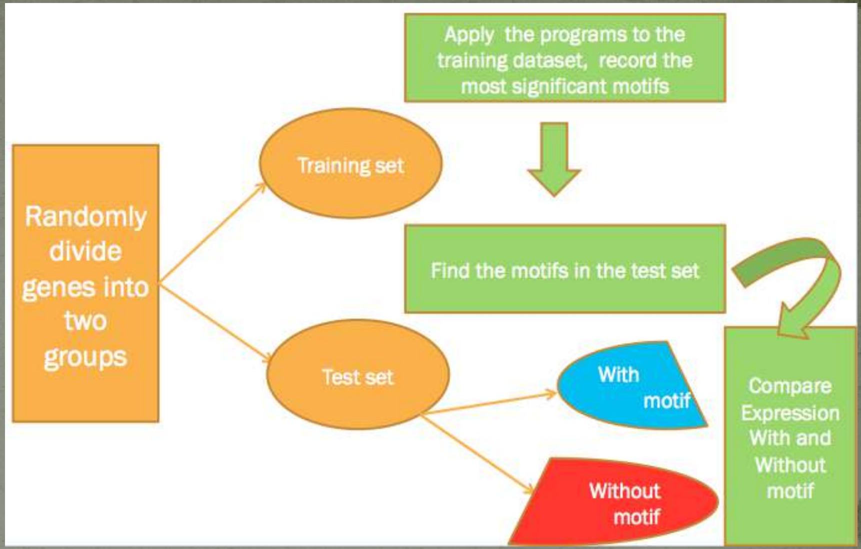
```
# motif:acgtgt
# from:400
# to:420
# number of generations set to:100
# minimal match to original:0.7
#88
#A      C       G       T
0.8654  0       0.09863 0.03596
0       1       0       0
0       0       1       0
0       0       0       1
0.3908  0.0346  0.5746  0
0       0       0.1082  0.8918
# match_threshold: 0.7
# Size: 6
# Z score:      7.385
```



mendel.fat.glam.ac.uk/tools/motifExpress/tool/optimization.php?motif=acgtgt&from_optim=400&to_optin

Benchmark

$$t_{n_1+n_2-2} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \text{ where } \sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}}$$



Randomly divide genes into two groups

Training set

Test set

Apply the programs to the training dataset, record the most significant motifs

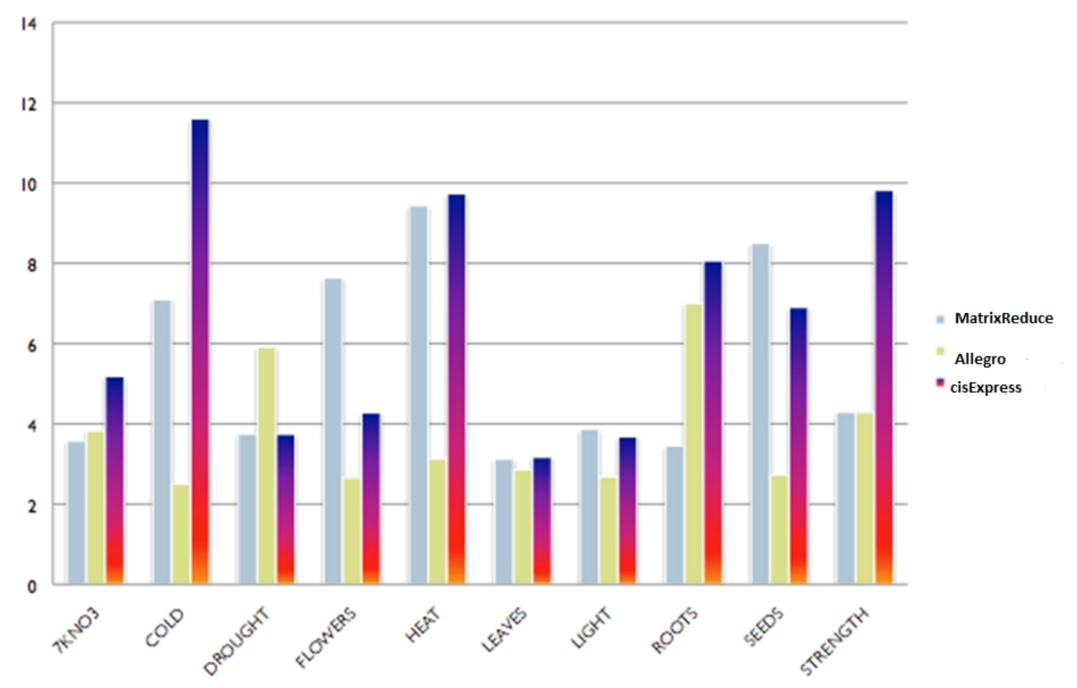
Find the motifs in the test set

With motif

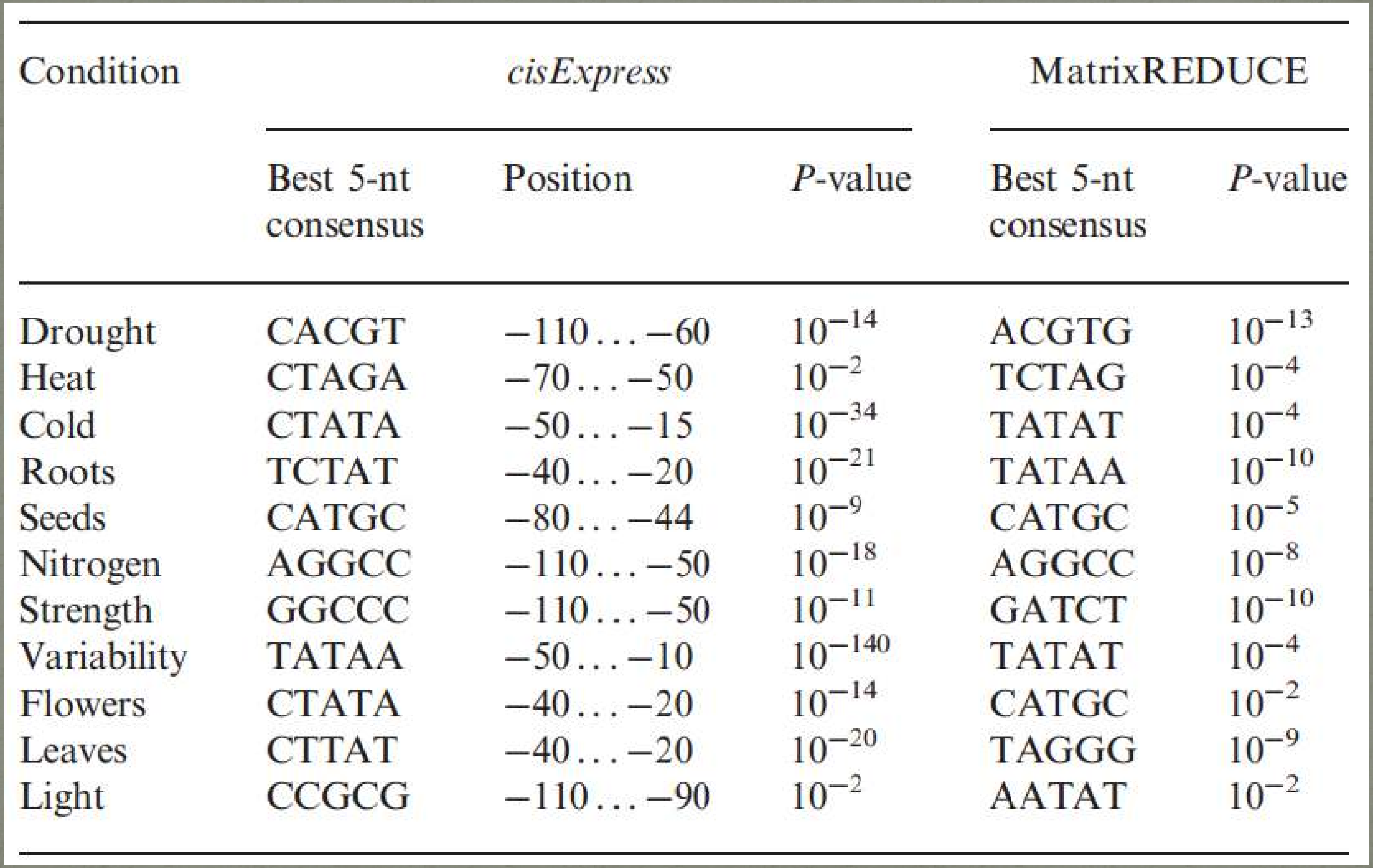
Without motif

Compare Expression With and Without motif

6-mers test



Benchmark



Condition	*cisExpress*			MatrixREDUCE	
	Best 5-nt consensus	Position	*P*-value	Best 5-nt consensus	*P*-value
Drought	CACGT	$-110\ldots-60$	10^{-14}	ACGTG	10^{-13}
Heat	CTAGA	$-70\ldots-50$	10^{-2}	TCTAG	10^{-4}
Cold	CTATA	$-50\ldots-15$	10^{-34}	TATAT	10^{-4}
Roots	TCTAT	$-40\ldots-20$	10^{-21}	TATAA	10^{-10}
Seeds	CATGC	$-80\ldots-44$	10^{-9}	CATGC	10^{-5}
Nitrogen	AGGCC	$-110\ldots-50$	10^{-18}	AGGCC	10^{-8}
Strength	GGCCC	$-110\ldots-50$	10^{-11}	GATCT	10^{-10}
Variability	TATAA	$-50\ldots-10$	10^{-140}	TATAT	10^{-4}
Flowers	CTATA	$-40\ldots-20$	10^{-14}	CATGC	10^{-2}
Leaves	CTTAT	$-40\ldots-20$	10^{-20}	TAGGG	10^{-9}
Light	CCGCG	$-110\ldots-90$	10^{-2}	AATAT	10^{-2}

Future directions

- Make TSS prediction/motif finding pipeline
- Add promoter/expression datasets for multiple species
- Improve *cis*Express by adding co-location of motifs and gene expression time-series
- Take advantage of methylation information
- Enhance functionality of the web-tool, such as motif scanning

Suggestions/collaborations and cool datasets are appreciated