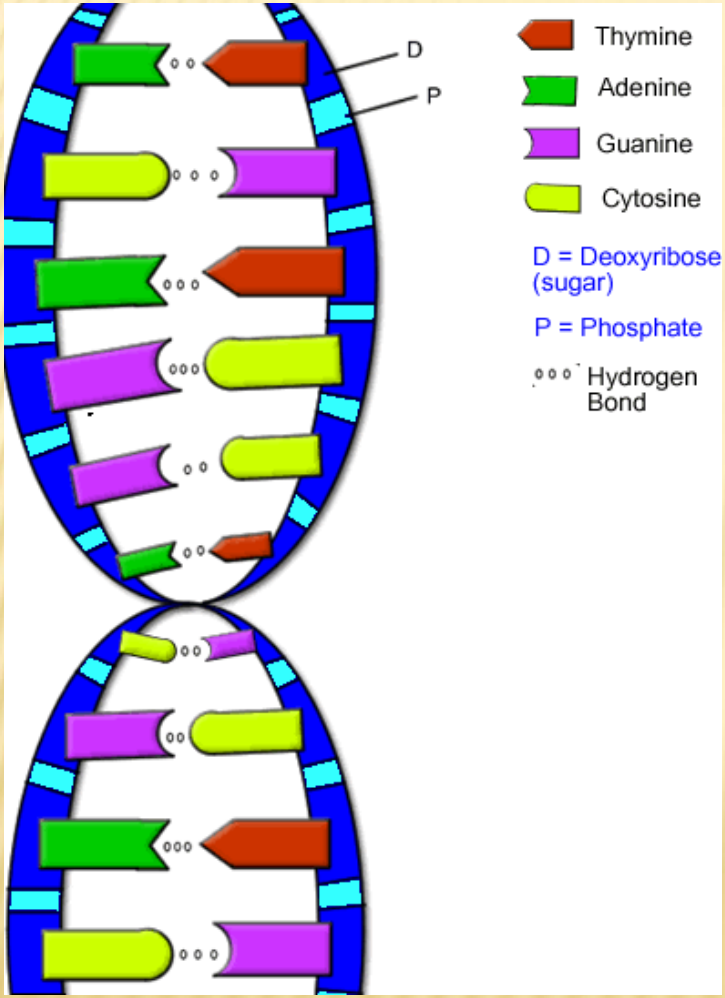
Tatiana Tatarinova

POWER OF Z-SCORES, OR EASY DOES IT (SOMETIMES)

NERDS CAN BE COOL!

DNA BASICS A,C,G,T

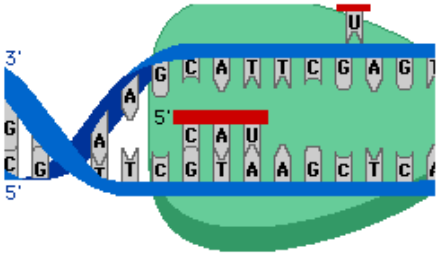


- The shape of the DNA molecule is a double-helix (like a twisted ladder). The sides of the ladder are composed of alternating sugars (deoxyribose) and phosphates.

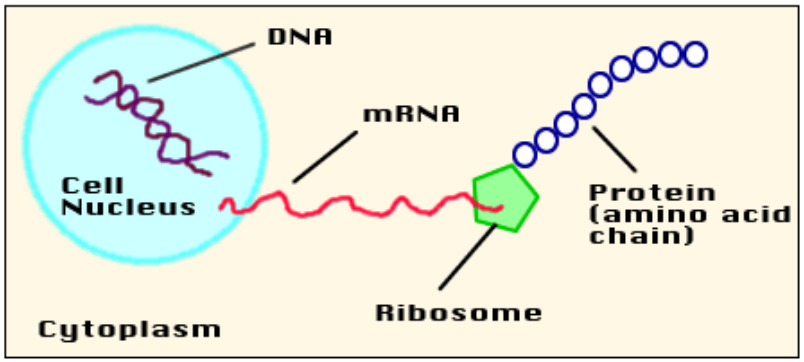
GENES AND PROTEINS

- **DNA** contains instructions for building an organism and ensuring that organism functions correctly.
- **Gene** - a segment of DNA that codes for a protein, which in turn codes for a trait (e.g. skin tone, eye color), a gene is a stretch of DNA.

Transcription - RNA is made from DNA



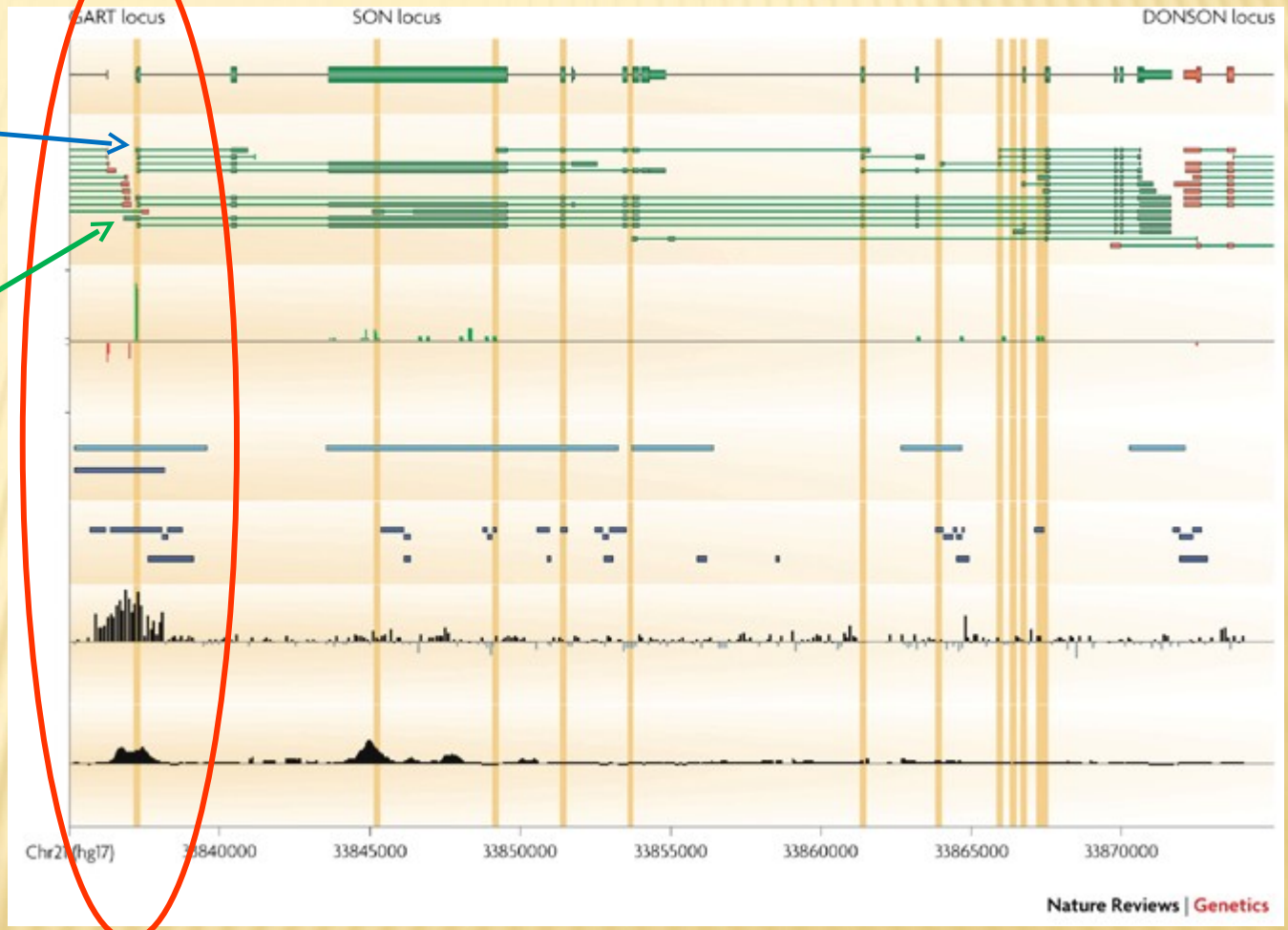
Translation - Proteins are made from the message on the RNA



GENOME ANNOTATION

Mode/mean/median of TSS distribution

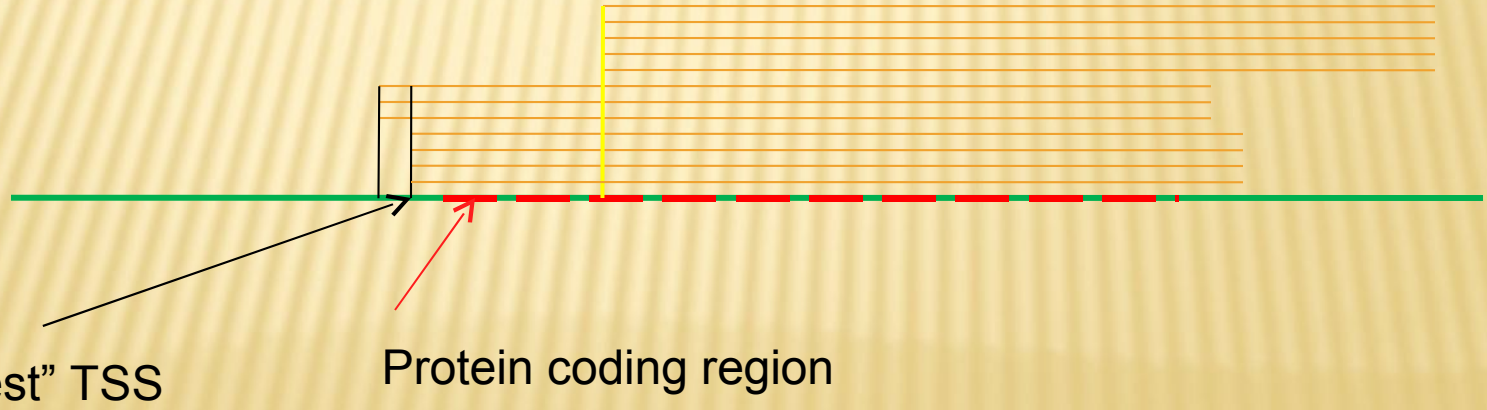
Longest transcript



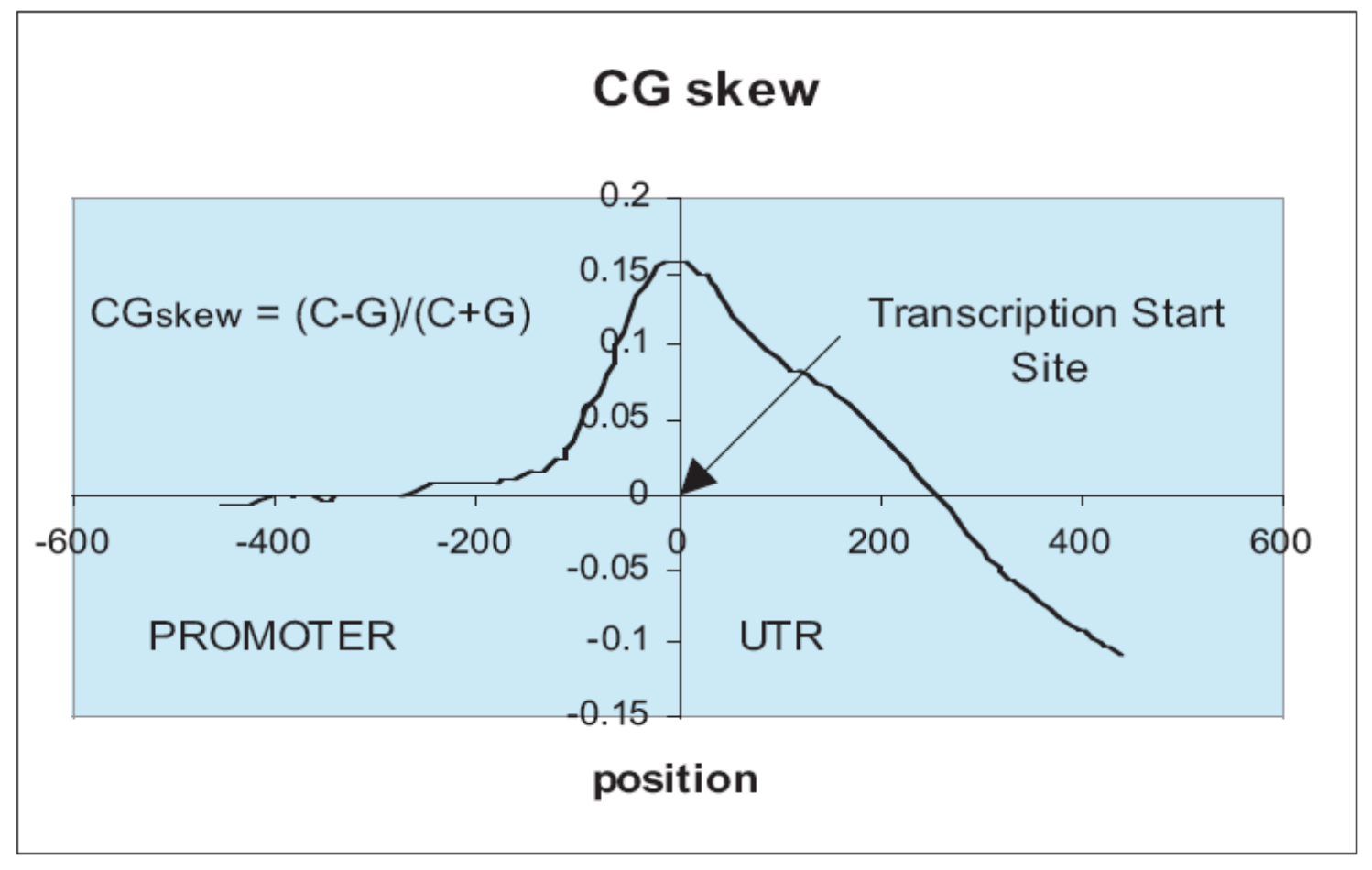
TSS =
Transcription Start Site

TSSER

- TSSer is a method to predict transcription start sites:
  - Align mRNAs to the genome
  - Compute distribution of putative TSSs
  - Find the mode of this distribution
  - If the most frequent TSS does not contradict the gene model, it is designated to be the TSS for the given gene.



CG SKEW



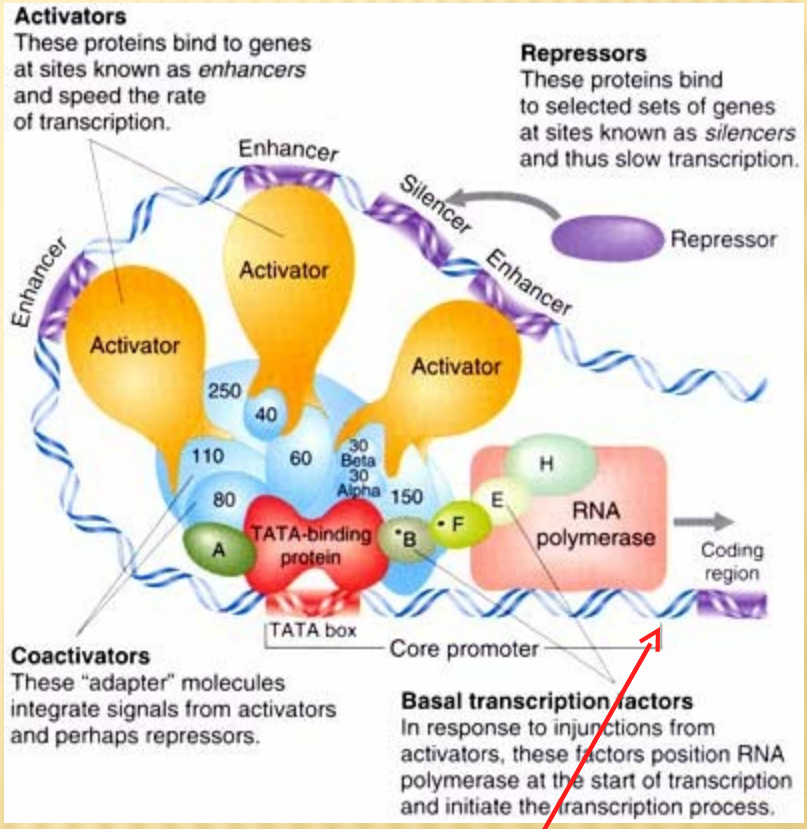
CG-skew is related to the bendability of DNA.

ASSESSMENT OF TRANSCRIPTION START SITE PREDICTION

Use known features, namely:

2.TATA is common at -30 nucleotides from TSS. Presence of TATA box indicates that the DNA denaturates more easily in this position.

3.CA is the most common di-nucleotides at TSS. Presence of this element increases the efficiency of stability-instability transition.



Compare three approaches:

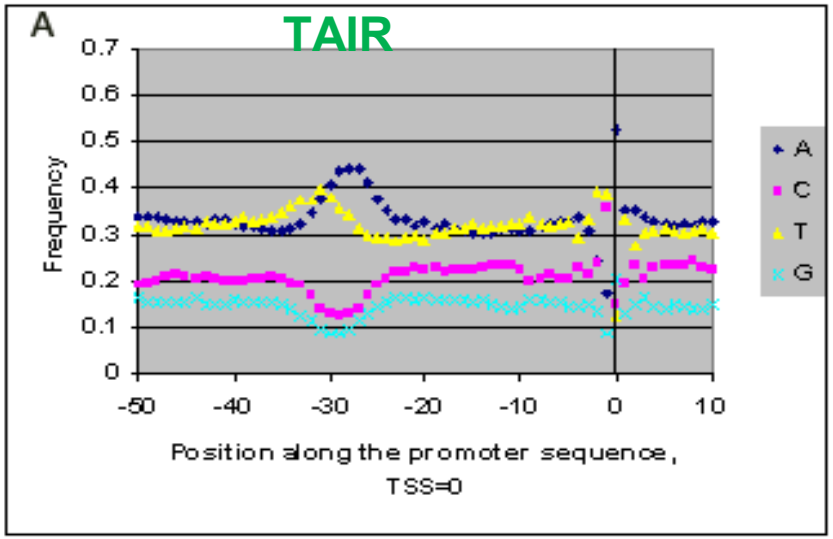
TSSer – mode

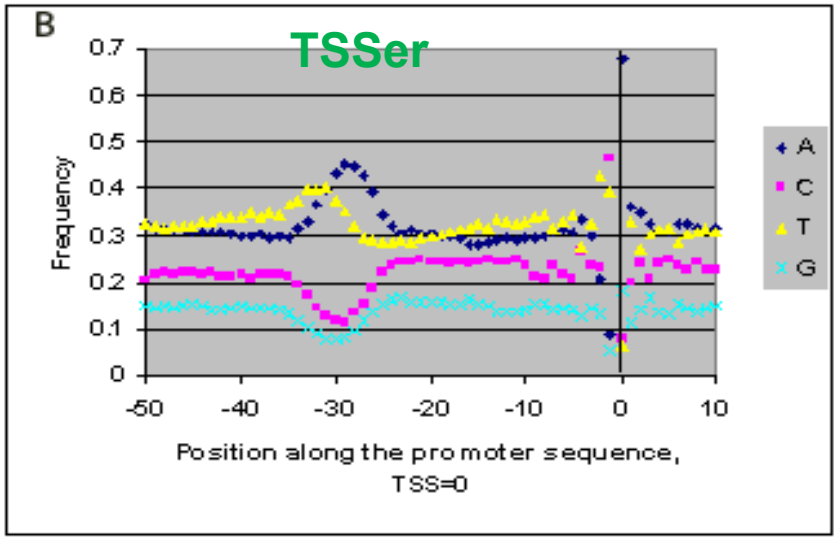
TAIR – longest

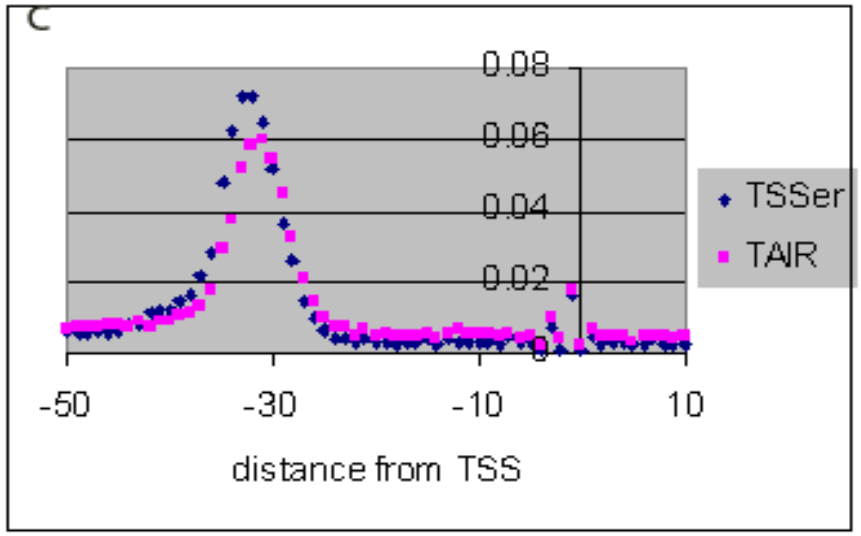
EP3 - structural profile based on base stacking energy

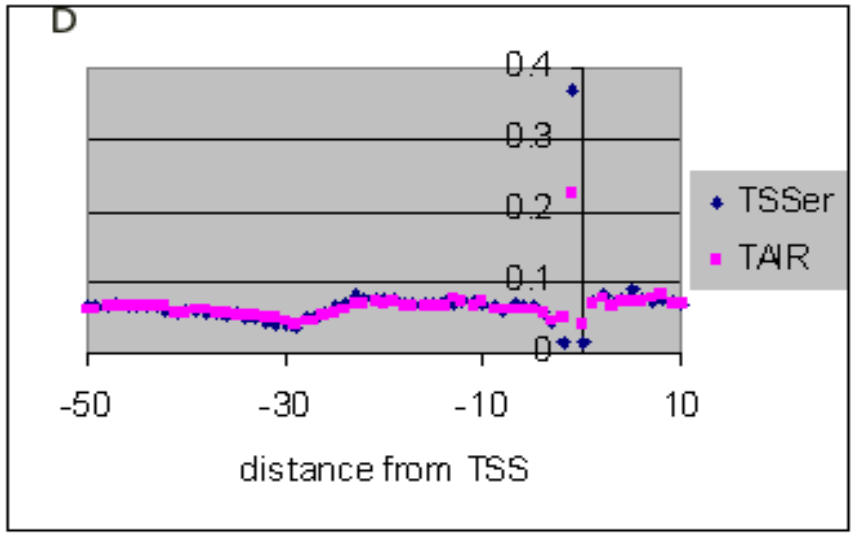
Transcription Start Site

TAIR VS TSSER

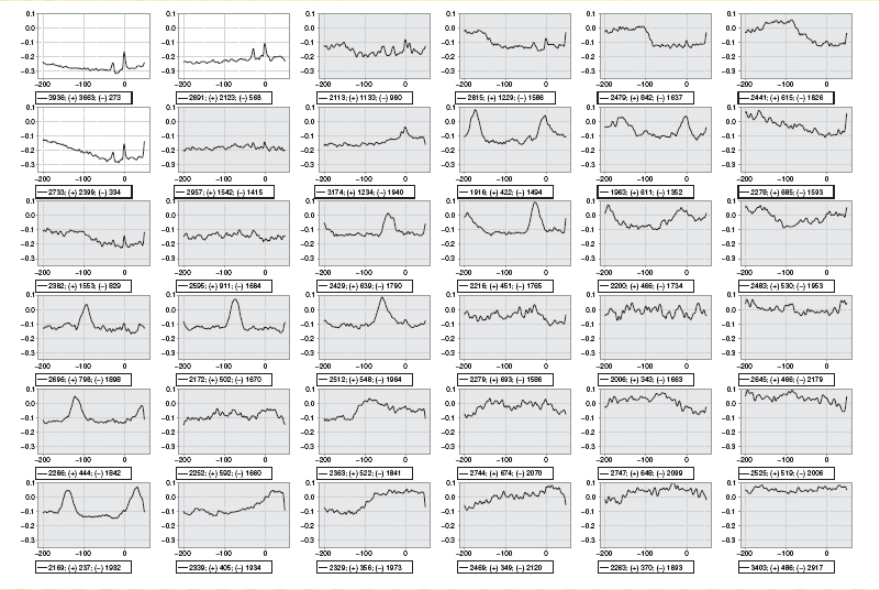








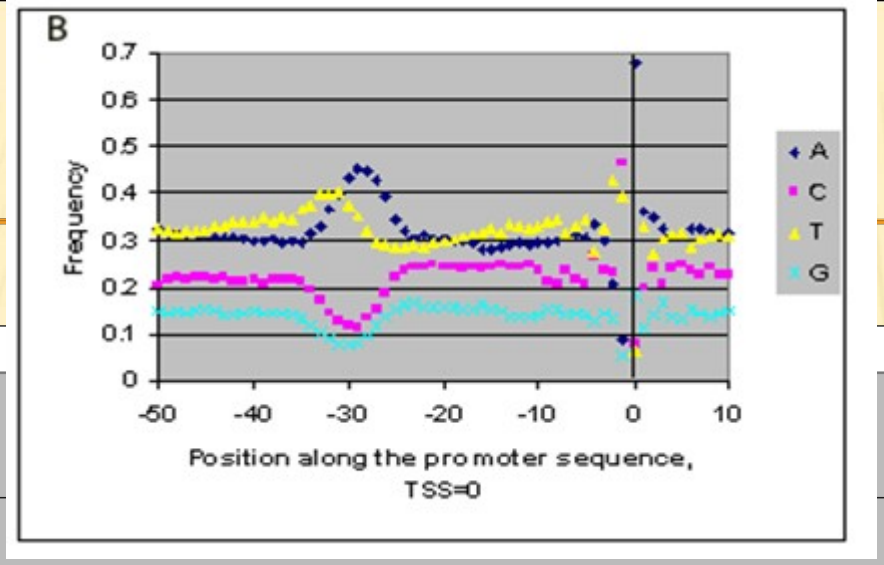
EP3 (ABEEL, ET AL., 2008)

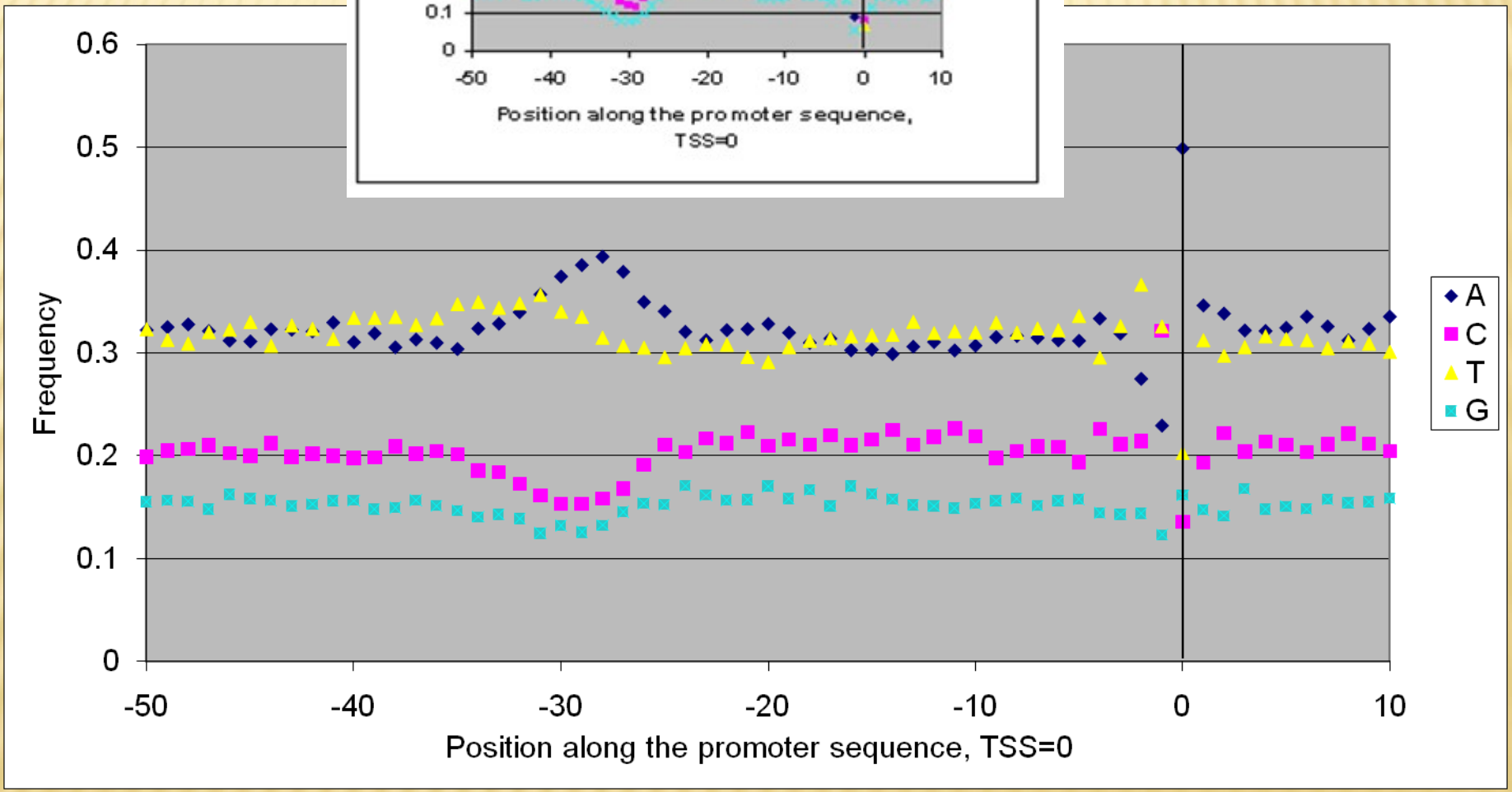


Downside: Computation of these profiles involves smoothing over few hundreds base pairs. Therefore it is too coarse to pinpoint the location of a TSS. It predicts a window of 400 nucleotides where TSS is likely to occur.

- Use conversion table for di- and tri- nucleotides, the following properties are computed
  - + Bendability
  - + Stabilizing energy
  - + DNA denaturation values
  - + Stabilizing energy
  - + Protein-induced deformability
  - + Duplex-free energy

EP3


B
Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-50
-40
-30
-20
-10
0
10
Position along the promoter sequence, TSS=0
A
C
T
G

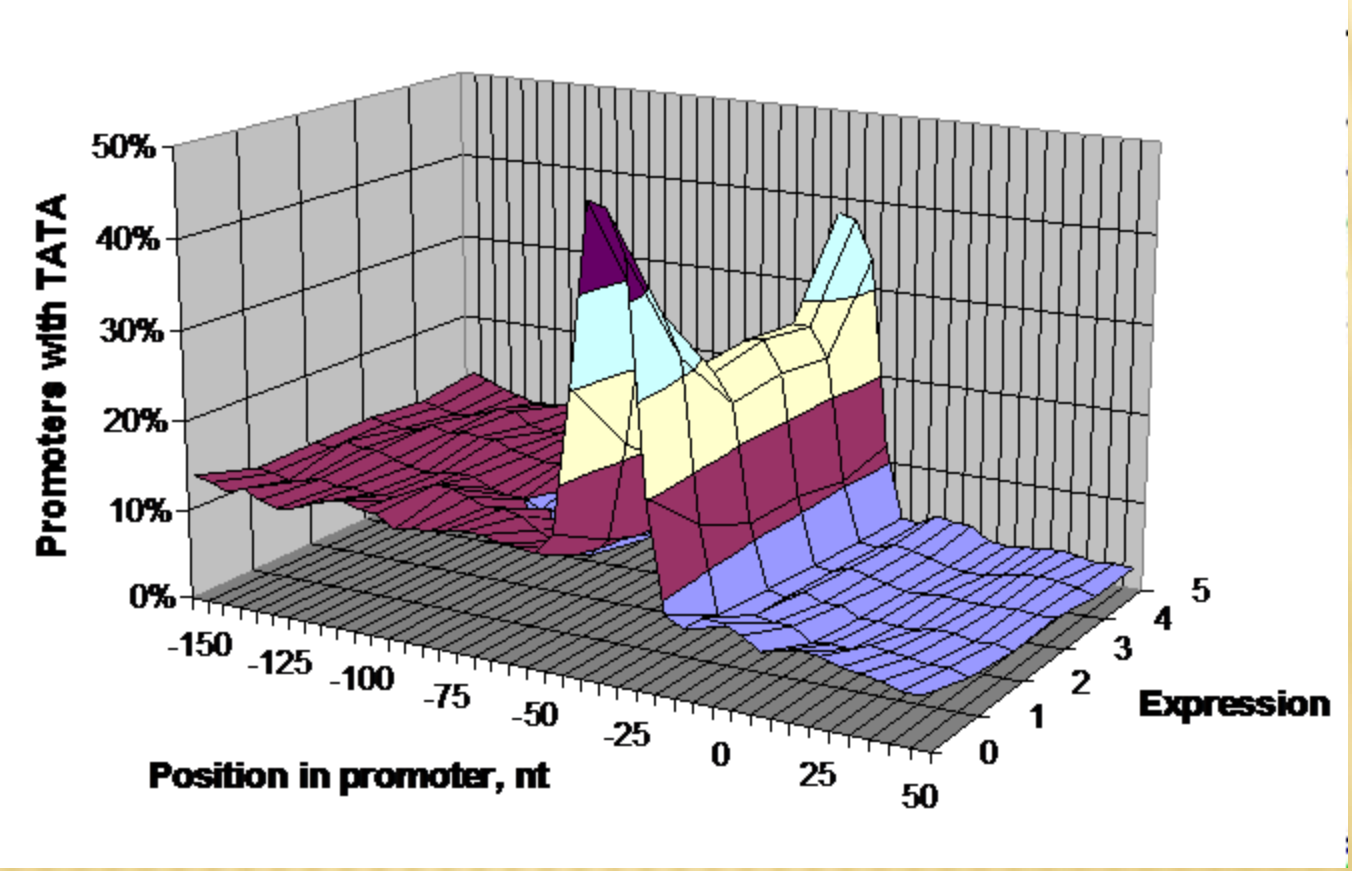

Frequency
0.6
0.5
0.4
0.3
0.2
0.1
0
-50
-40
-30
-20
-10
0
10
Position along the promoter sequence, TSS=0
A
C
T
G

POSSIBLE GENOME ANNOTATION PIPELINE

- Take a newly sequenced genome
- Map known proteins from other species (blast)
- Examine upstream regions for presence of typical structural features – 400 nucleotide potential promoter areas
- Create mini-tiling array for potential promoter regions
- Sequence 5’ ESTs
- Find TSS

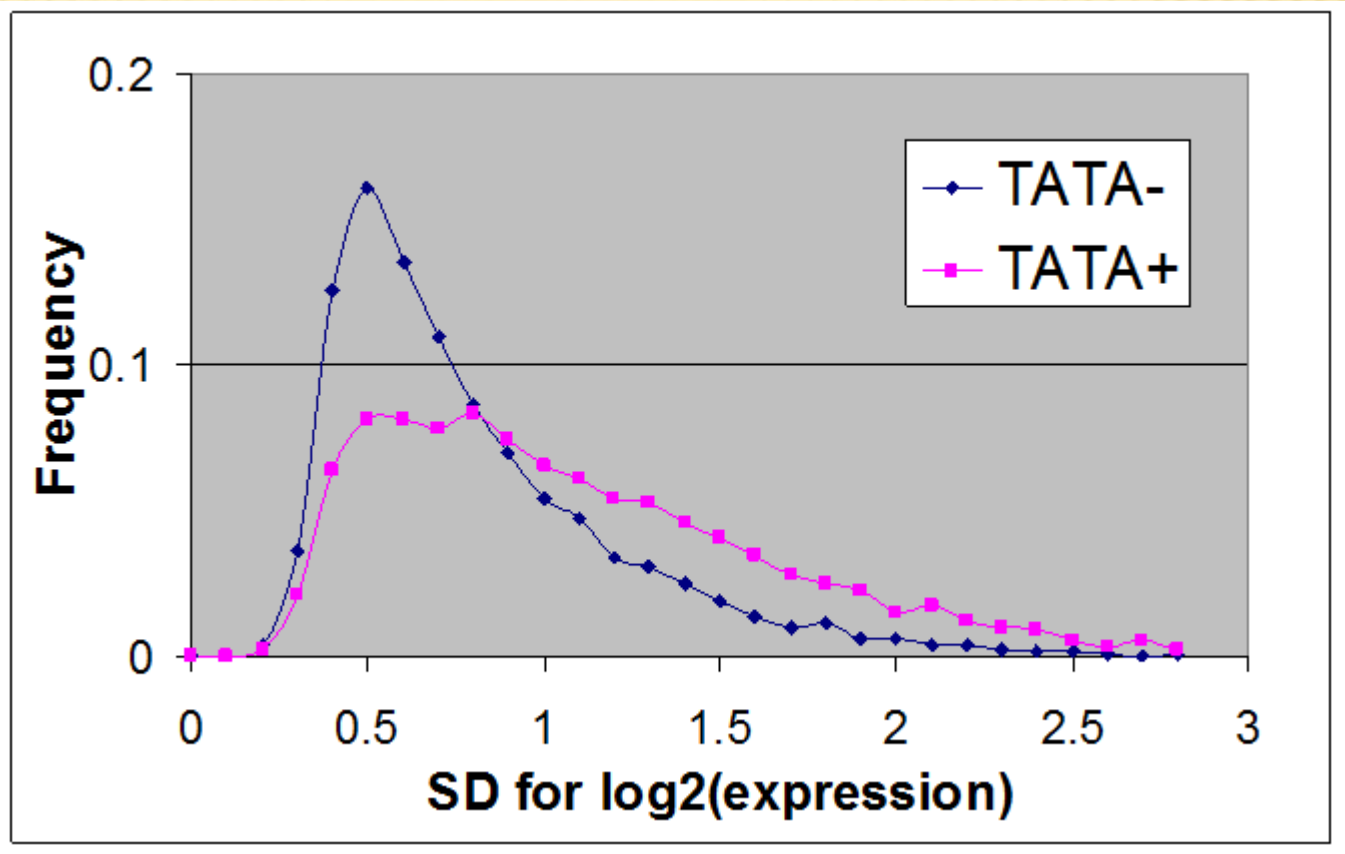
PROMOTER MOTIFS AND GENE EXPRESSION

Arabidopsis thaliana



TATA-box is more prevalent in weak and strong promoters

TATA+ PROMOTERS SHOW LARGER VARIABILITY



Arabidopsis thaliana

MORE ABOUT TATA BOX

Mus musculus