Exploring the Connection between Sequence and Coordinated Gene Activity for Adjacent Promoter Pairs

Ameen Salem and Marc S. Halfon
Department of Biochemistry
NYS Center of Excellence Bioinformatics & Life Sciences
University at Buffalo

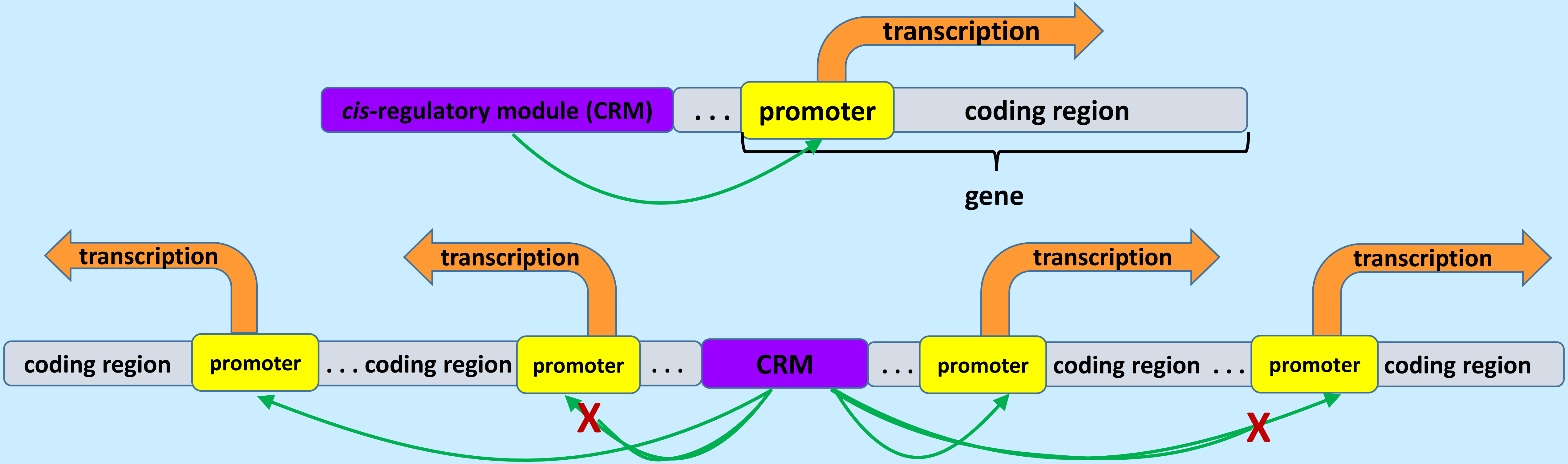
What is a genome?

A – Adenosine
C – Cytosine
G – Guanine
T – Thymine

Promoter-CRM specificity

some DNA sequence

TTGGCCGCTCCCAGCGACGGCGG . . . TAAAAGGGCGCTGAGAGCAGCACAC



transcription

cis-regulatory module (CRM) . . . promoter coding region

gene

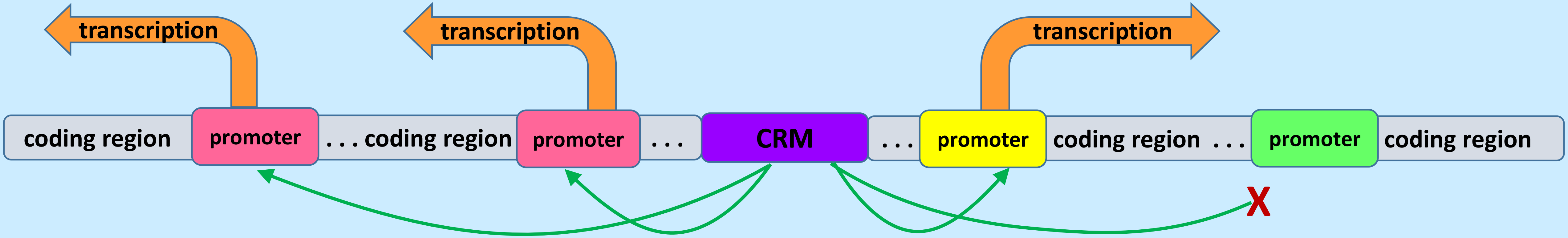
transcription transcription transcription transcription

coding region promoter . . . coding region promoter . . . CRM . . . promoter coding region . . . promoter coding region

X X

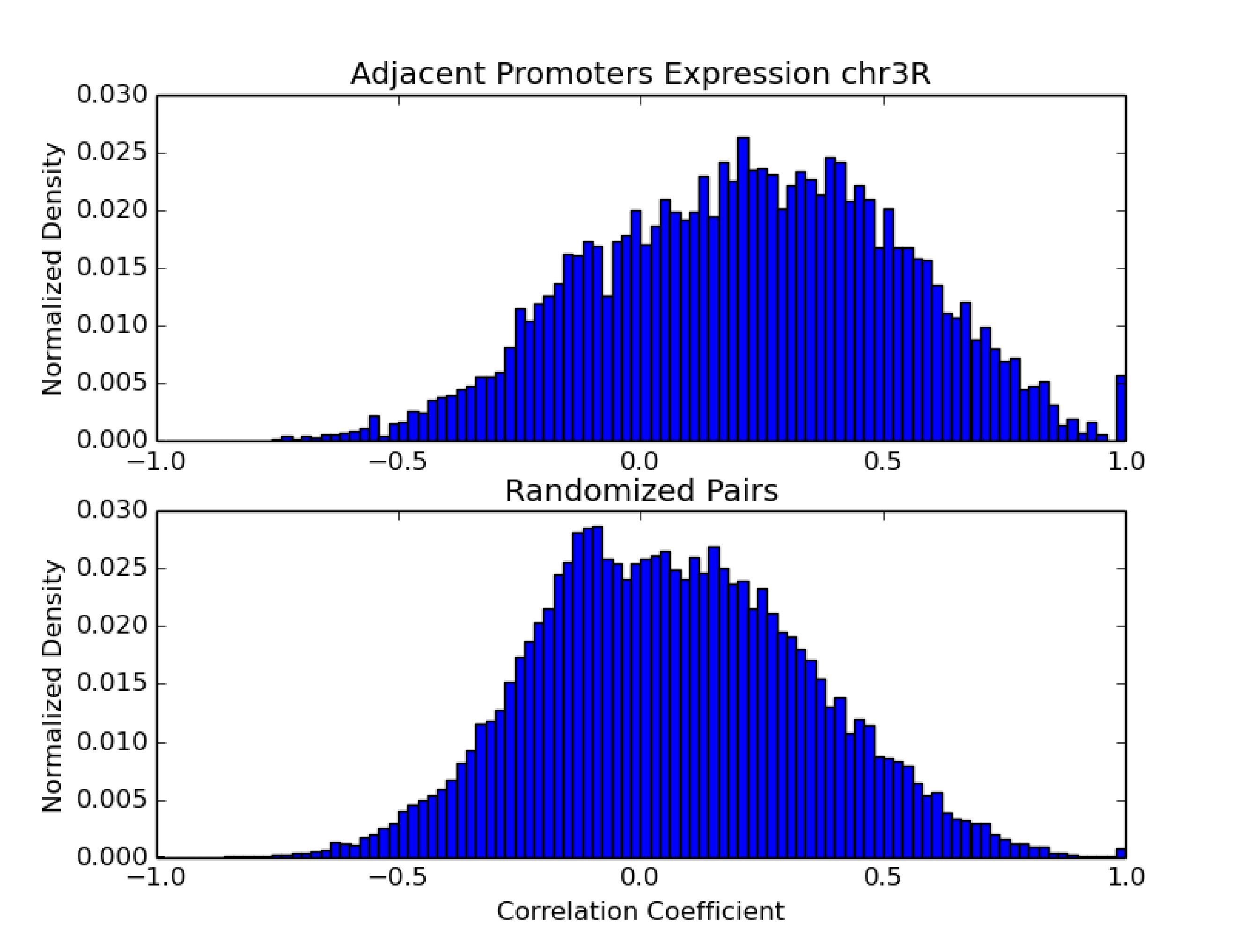
Hypothesis

Adjacent promoters with coordinated control will have similar sequences whereas those not coordinately regulated will have dissimilar sequences



- Interactions with CRMs simply based on proximity would lead us to expect that adjacent promoters are always co-regulated
- Cases show specfcity of CRM-promoter interactions

Activity of Adjacent Promoters



RAMPAGE profiling (Genome Res. 23:169)

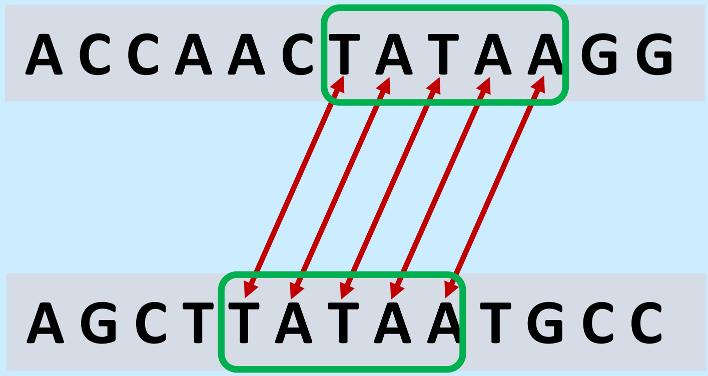
Observations: 6967
Range : -0.74, 1.0
Mean : 0.22
Median : 0.23

p value : 1.2e-283

Observations : 28502
Range : -0.87, 1.0
Mean : 0.07
Median : 0.06

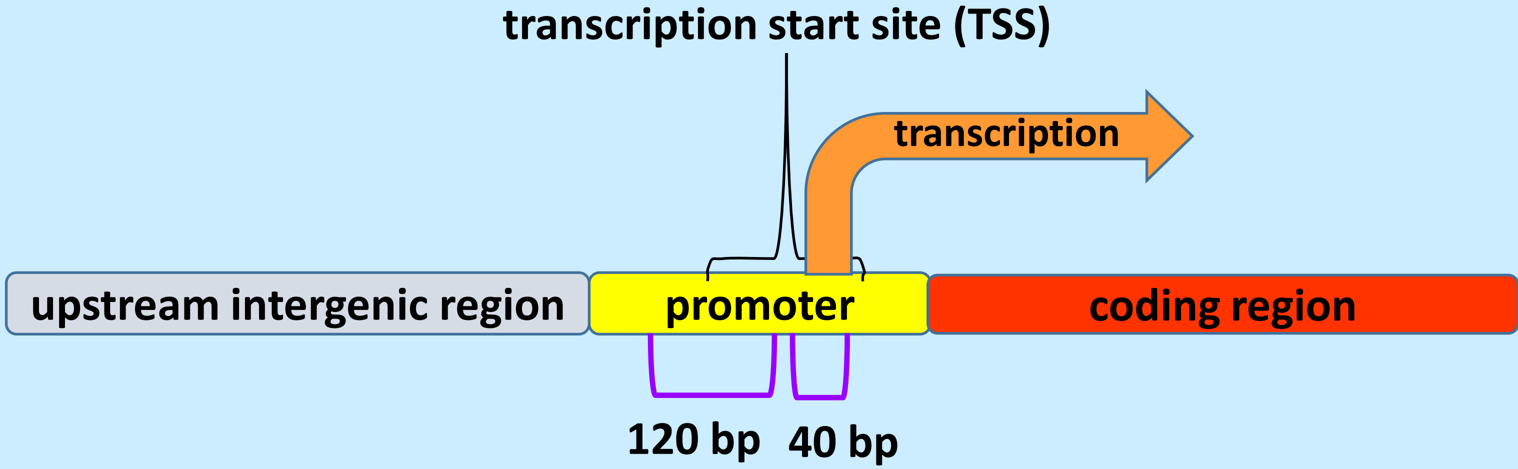
Measuring Sequece Similarity - N2 Scoring

- An alignment-free sequence similarity measure
- Used to compare promoter sequences
- Scoring : 1 is most similar -1 is most dissimilar
- Promoter length measurement:
  - From transcription start site (-120bp, +40bp)



ACCAACTATAAGG

AGCTTATAATGCC



transcription start site (TSS)

transcription

upstream intergenic region

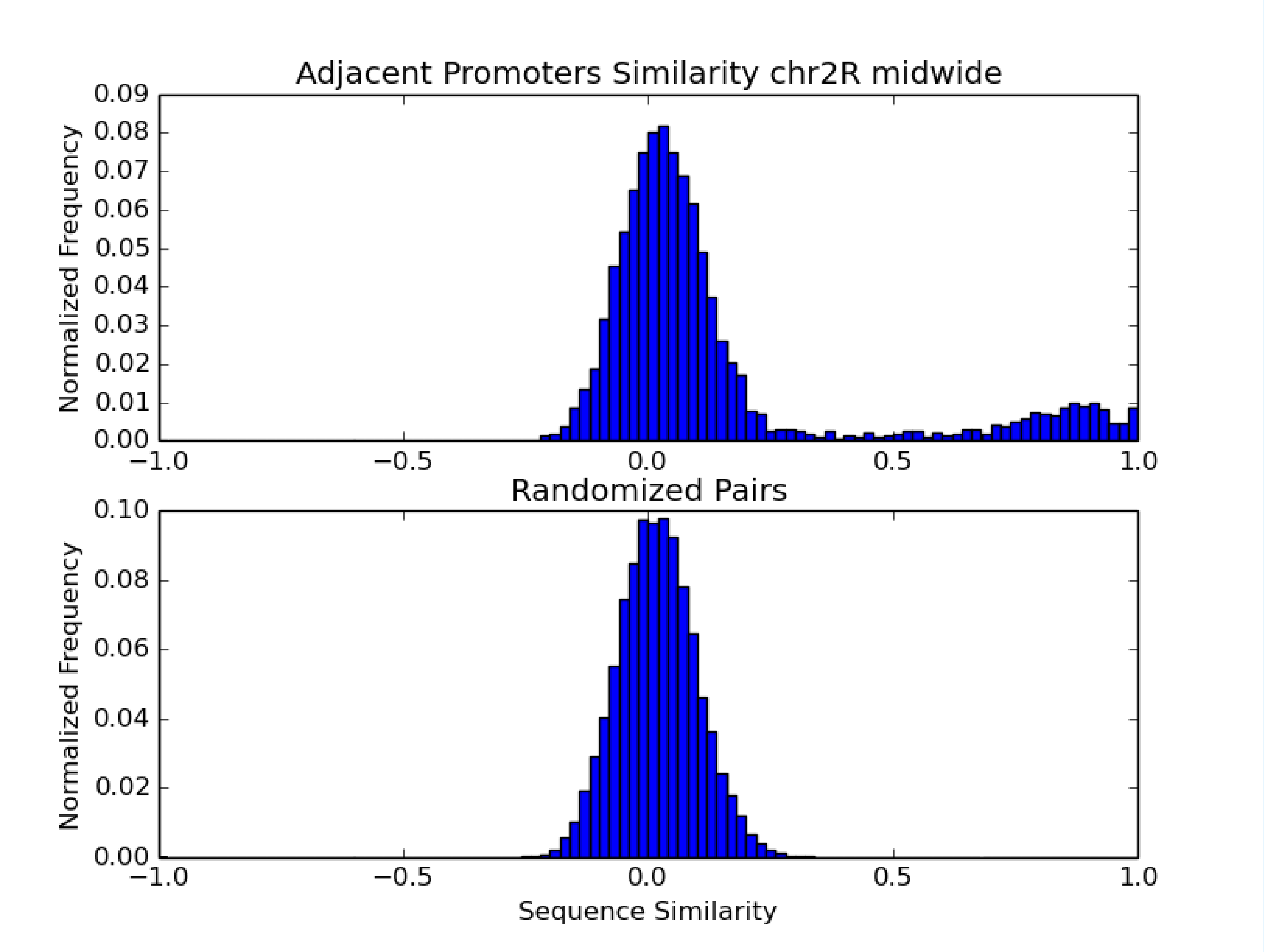
promoter

coding region

120 bp

40 bp

Sequence Similarity of Adjacent Promoters

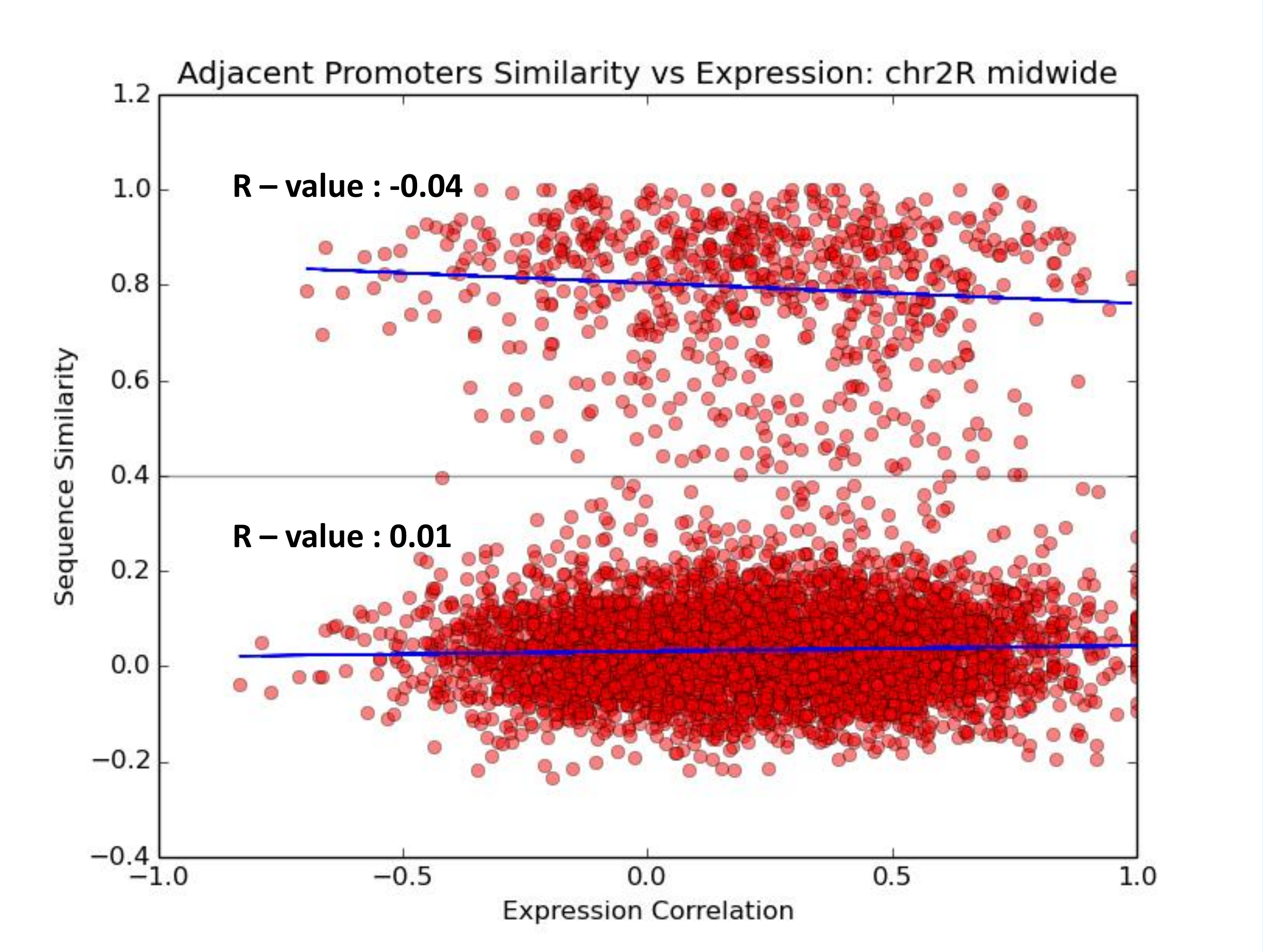


Observations : 6023
Range : -0.23, 1.00
Mean : 0.13
Median : 0.05

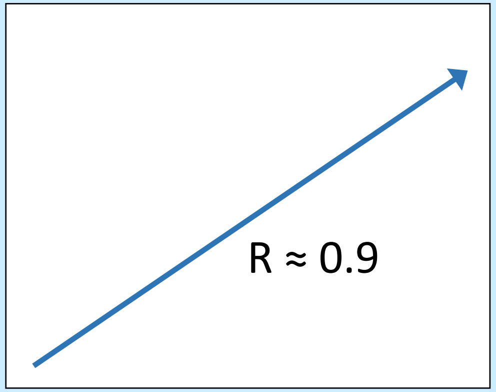
p value : 1.24e-207

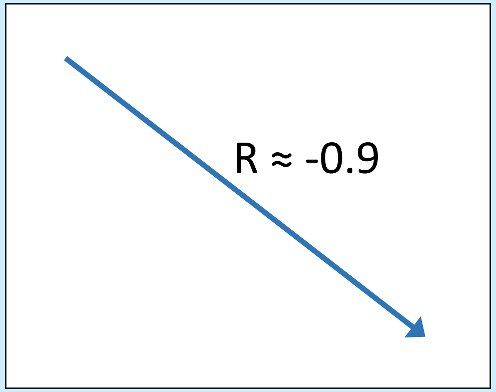
Observations : 22383
Range : -0.27, 0.78
Mean : 0.02
Median : 0.02

Sequence Similarity vs Expression Correlation



Strong Correlations

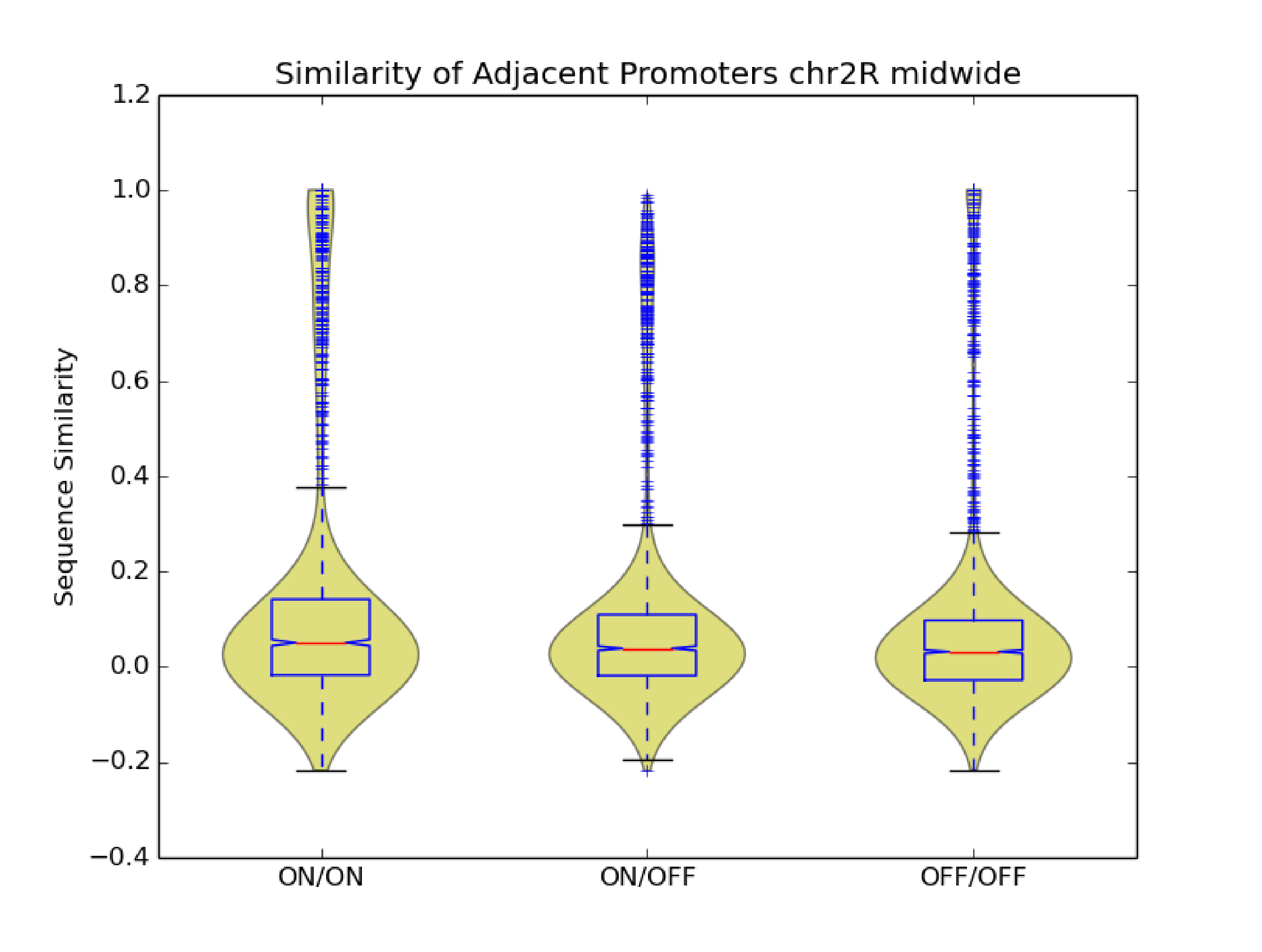




Data from a Homogeneous Cell Type

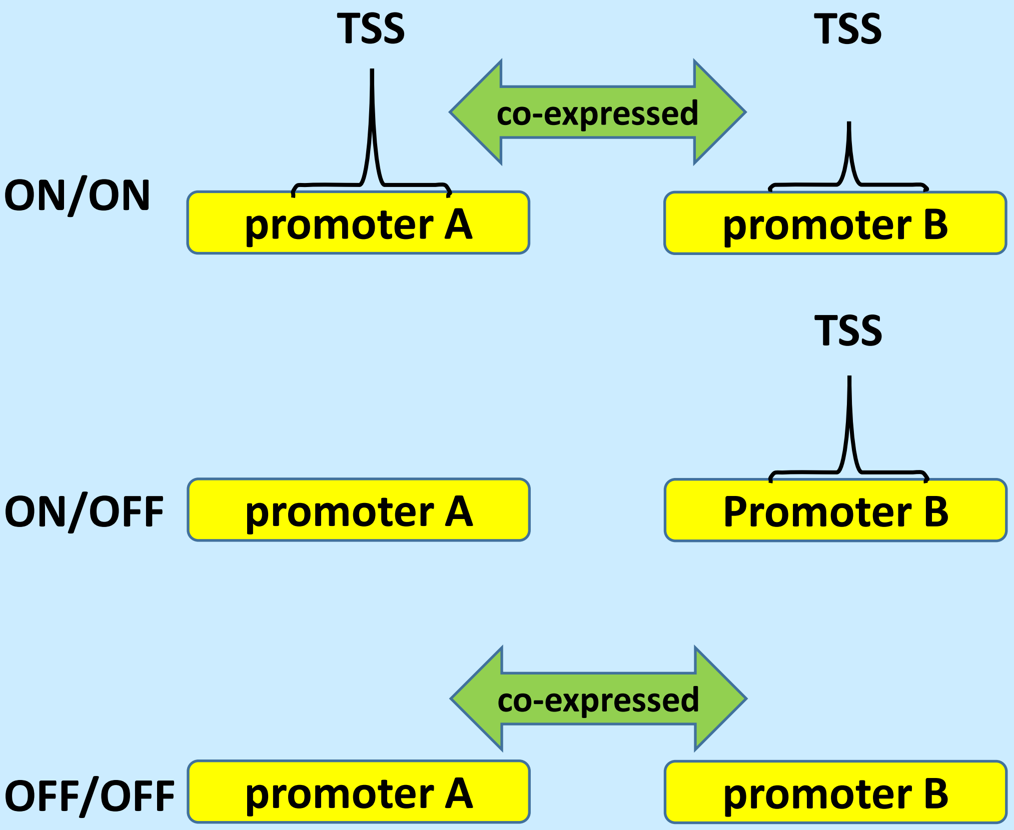
- S2 cells, a popular *Drosophila* cell line
  - **Cell line** : a population of cells derived from a single cell
- All cells therefore are expressing the same genes (no cell differentiation)
- Expression levels measured at single discrete instance (no time course)
- Expression level measurements:
  - **Digital** : Any measurable transcription signifies promoter activity
  - **Continuous** : Define a measure to evaluate expression

Digital Description – Co-expression vs Sequence Similarity



S2 cells (Cell Rep. 2:1025)

Groupings are as follows:



Defining Continuous Measure

- A pair of promoters with TSS scores that vary greatly should have a different co-expression description than a pair with identical peaks
- This measure ranges from 0 to 1, with 0 being the highest level of co-expression and 1 being the lowest
- Scoring formula :

$$m = \frac{|\ \text{Difference of maximum TSS scores}\ |}{\text{sum of maximum TSS scores}}$$

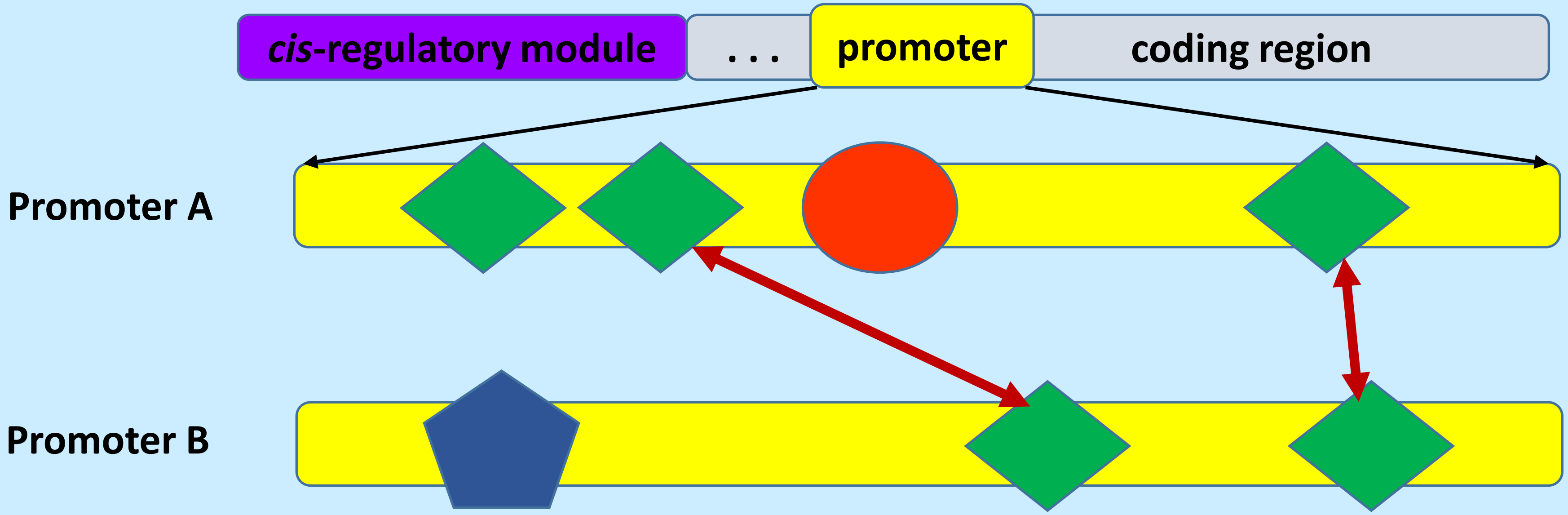
- No additional conclusions were drawn from continuous measure results

Conclusions

- We have not yet found support for our hypothesis that co-regulated promoters share similar sequences
- We used different data in an effort to more accurately describe promoter activity
- Using N2 scoring to measure sequence similarity may be more apt for comparing sets of sequences as opposed to individual pairs
- Shift in focus now is to explore alternatives to quantifying promoter similarity

New Directions

- Motifs are short recurring patterns of DNA in promoters
- Often indicate sequence-specific transcription start sites, within promoter
- Presence can classify promoter classes



Acknowledgements

Marc S. Halfon, Ph.D., Department of Biochemistry, University at Buffalo

John Ringland, Ph.D., Department of Mathematics, University at Buffalo

This work was supported by the NSF CSUMS program,
grant DMS 0802994 to the University at Buffalo.

URGE NSF to Compute
Undergraduate Research Group Experiences
in Computational Mathematics 2014

New York State Center of Excellence
in Bioinformatics
& Life Sciences