Transfer String Kernel for Cross-Context Sequence Specific DNA-Protein Binding Prediction

by

Ritambhara Singh

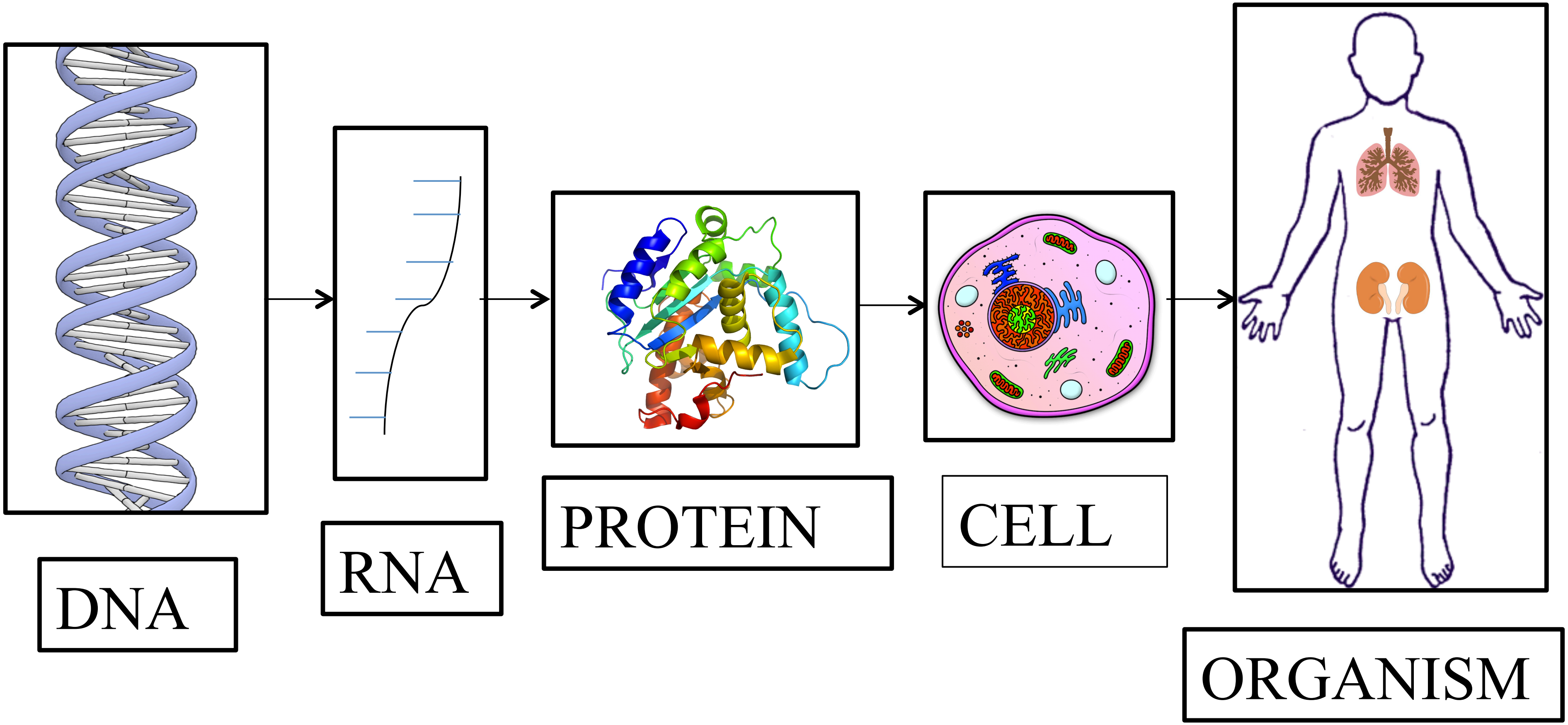
IIIT-Delhi

June 10, 2016

Computer Science at the University of Virginia

University of Virginia Engineering

Biology in a Slide



DNA

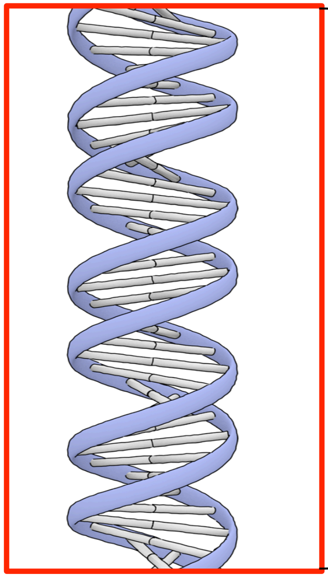
RNA

PROTEIN

CELL

ORGANISM

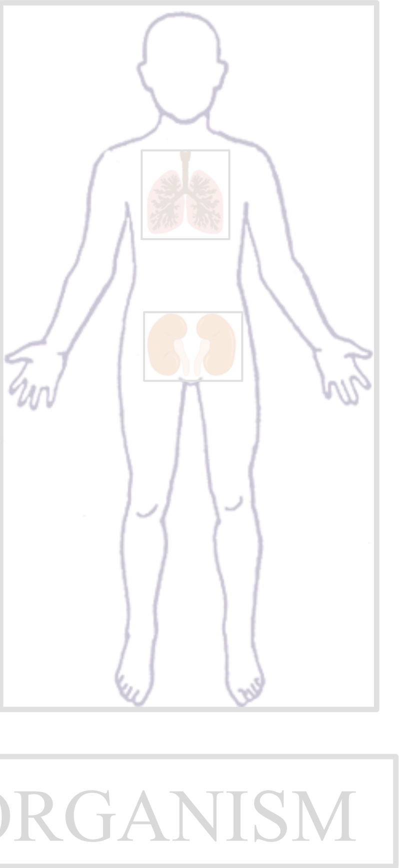
DNA and Diseases



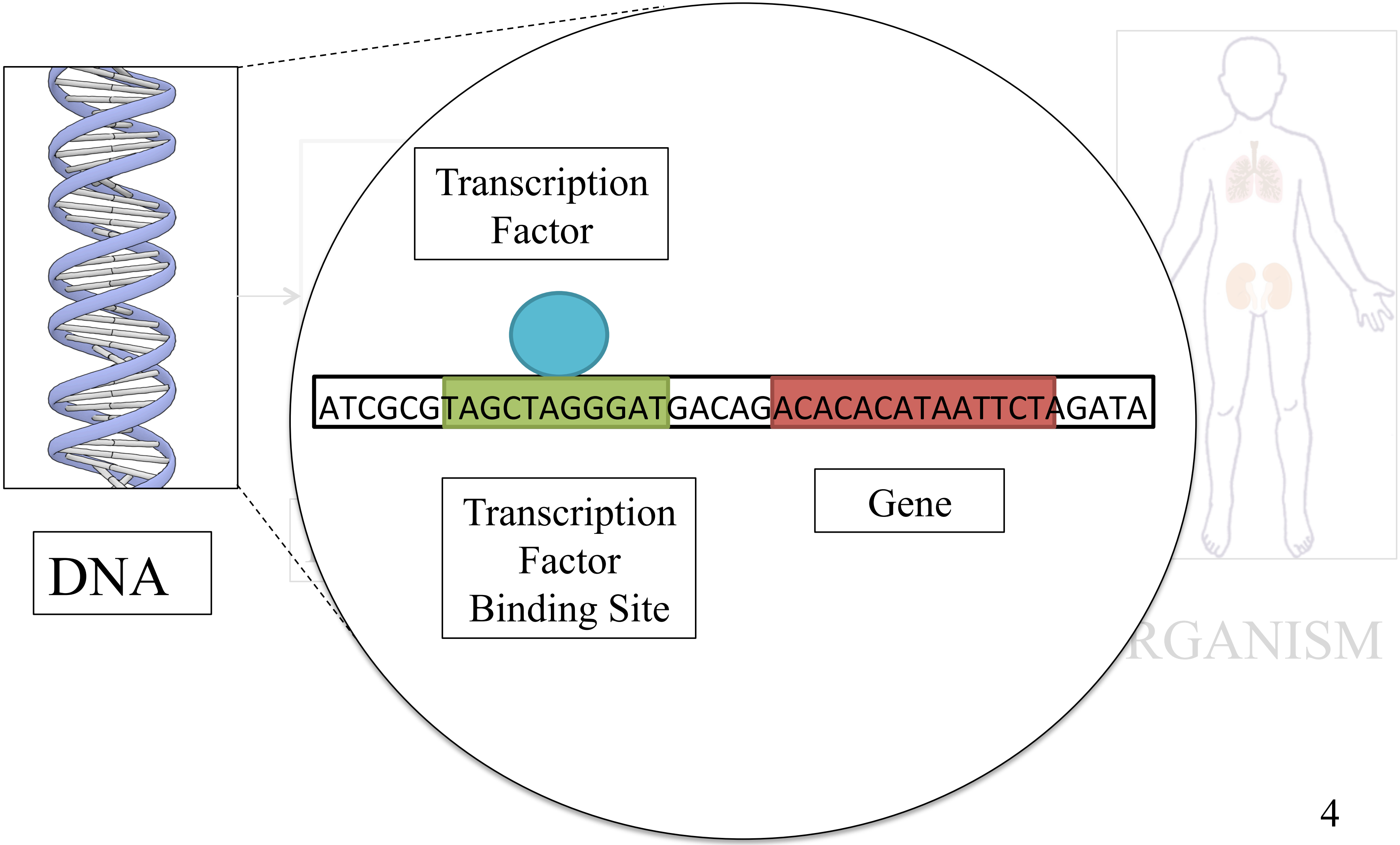
- Down Syndrome
- Parkinson’s Disease
- Autism
- Muscular Atrophy
- Sickle Cell Disease

…………

…………..



Transcription Factors



Transcription Factor

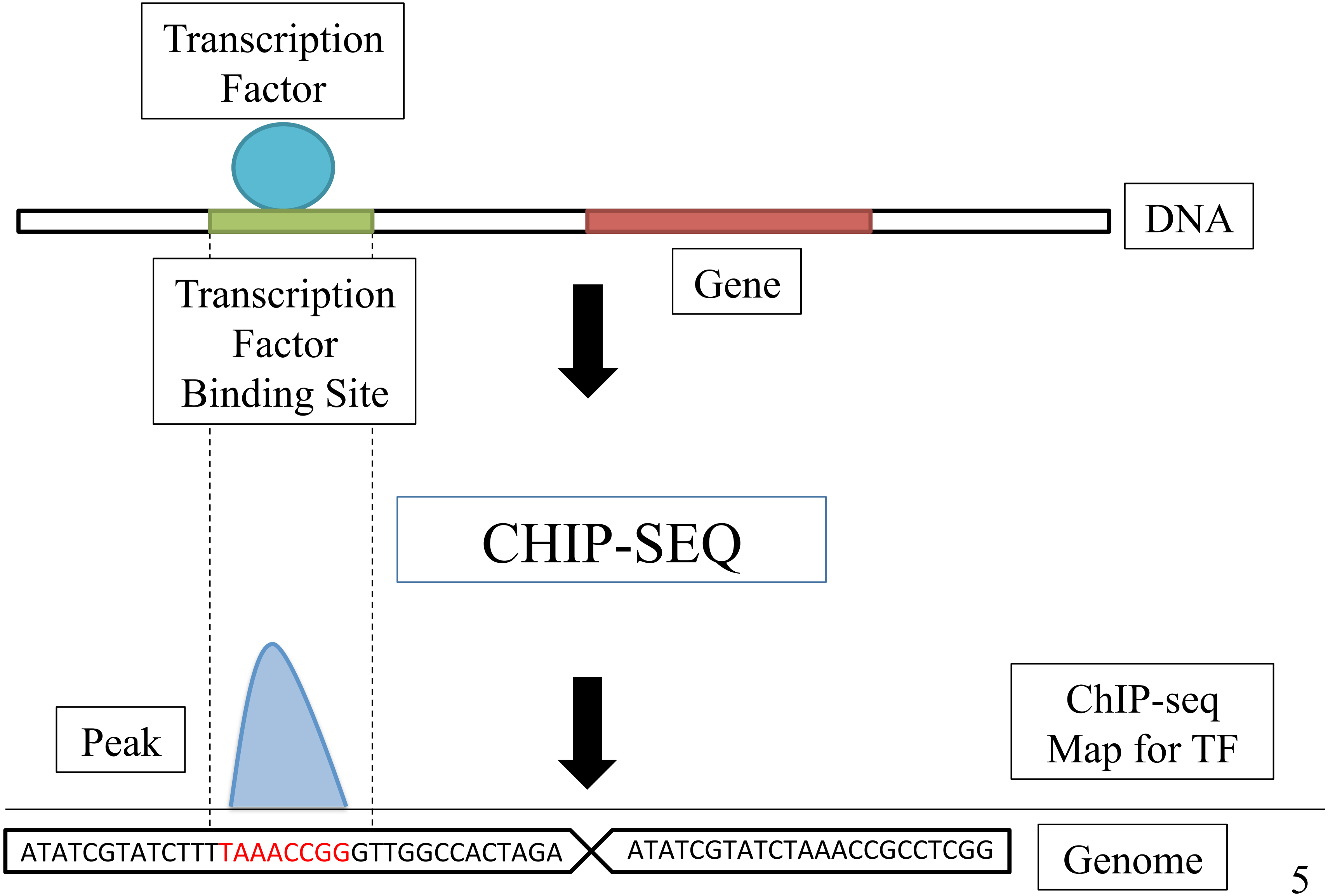
ATCGCGTAGCTAGGGATGACAGACACACATAATTCTAGATA

Transcription Factor Binding Site

Gene

DNA

ChIP-seq Maps TF binding



Transcription Factor

DNA

Transcription Factor Binding Site

Gene

CHIP-SEQ

Peak

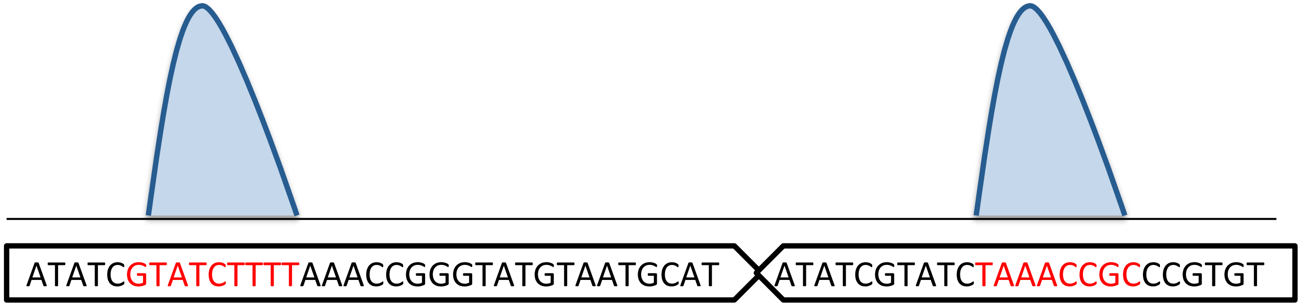
ChIP-seq Map for TF

ATATCGTATCTTTTAAACCGGGTTGGCCACTAGA

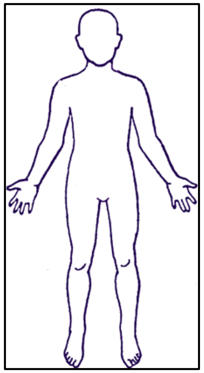
ATATCGTATCTAAACCGCCTCGG

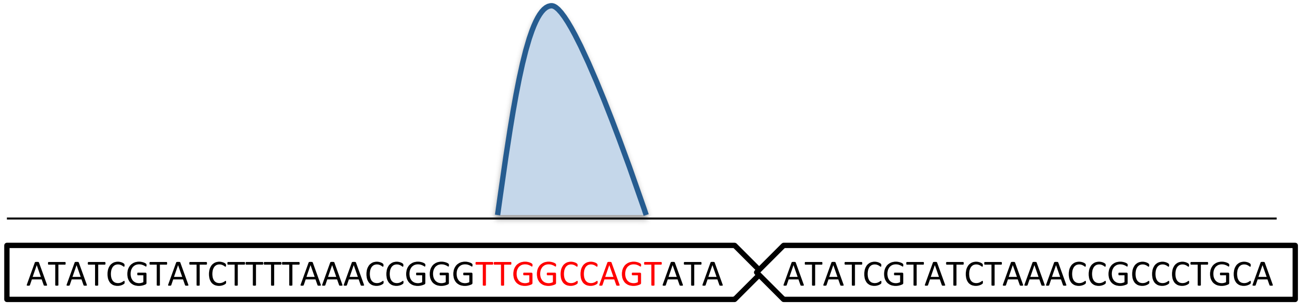
Genome

TF Binding Differs Across Contexts



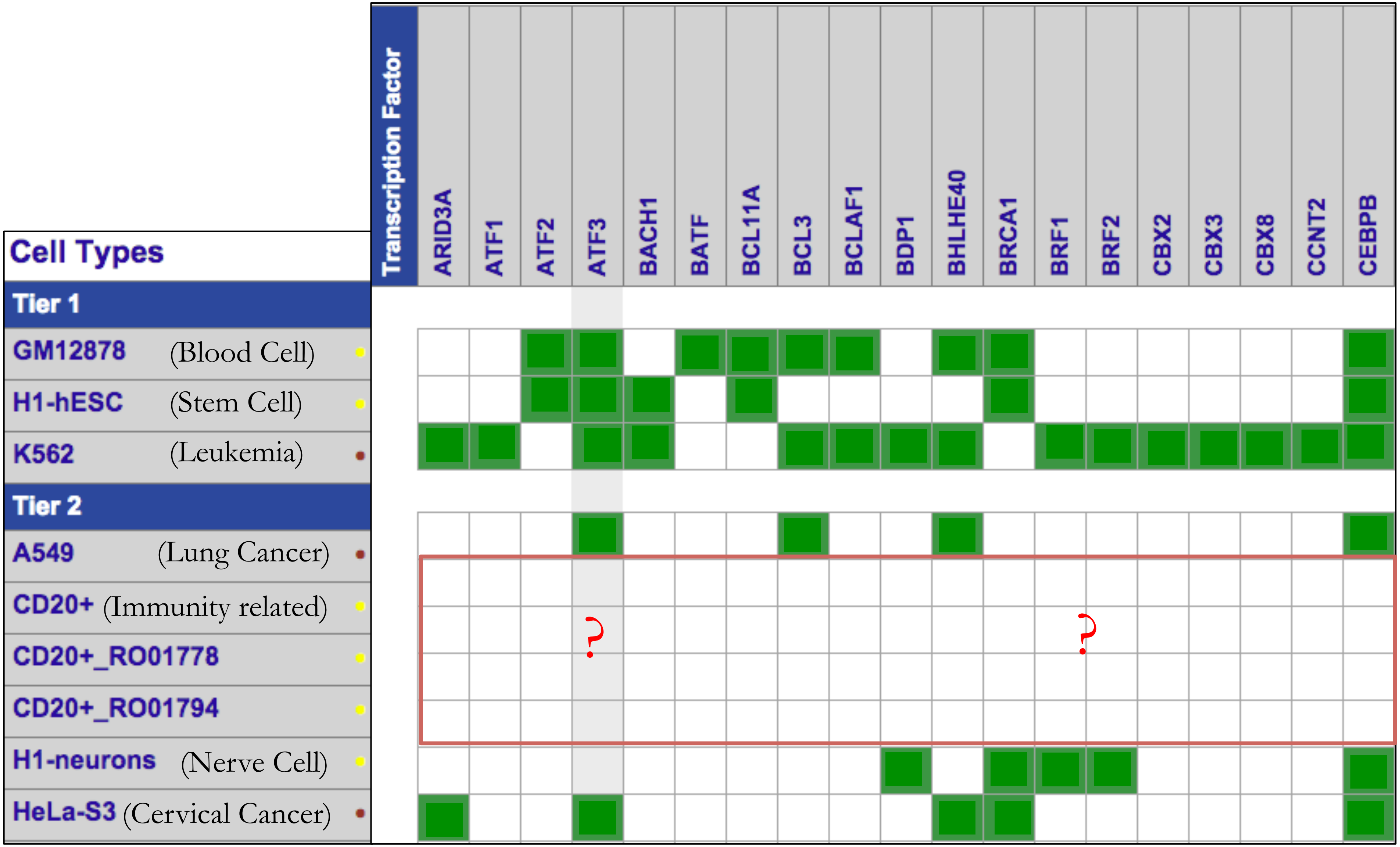
ATATCGTATCTTTTAAACCGGGTATGTAATGCAT ATATCGTATCTAAACCGCCCGTGT





ATATCGTATCTTTTAAACCGGGTTGGCCAGTATA ATATCGTATCTAAACCGCCCTGCA

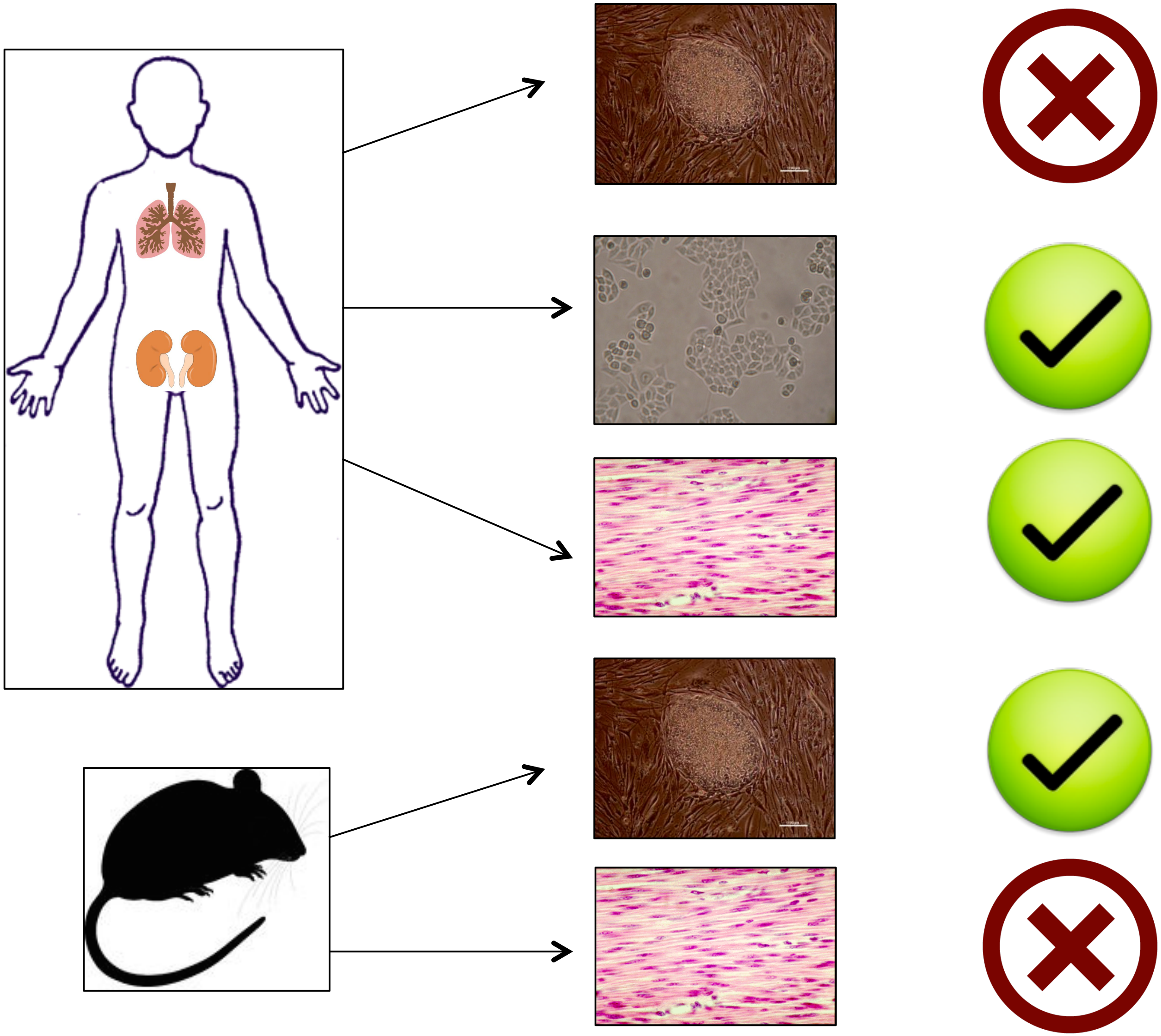
Current Challenge: ENCODE Data Gap



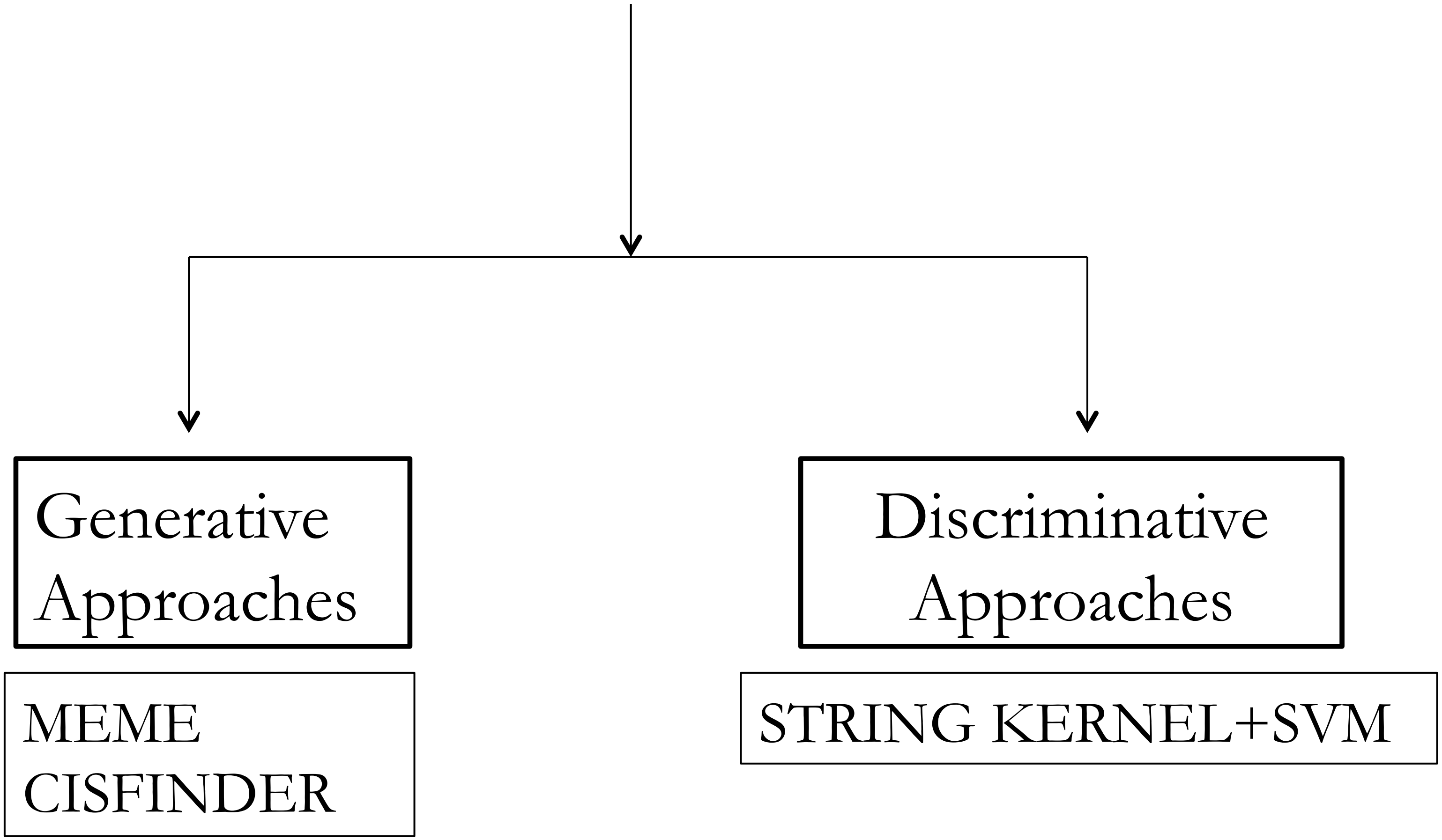
Cell Types	ARID3A	ATF1	ATF2	ATF3	BACH1	BATF	BCL11A	BCL3	BCLAF1	BDP1	BHLHE40	BRCA1	BRF1	BRF2	CBX2	CBX3	CBX8	CCNT2	CEBPB
Tier 1																			
GM12878 (Blood Cell)			🟩	🟩		🟩	🟩	🟩	🟩		🟩	🟩							🟩
H1-hESC (Stem Cell)			🟩	🟩	🟩		🟩					🟩							🟩
K562 (Leukemia)	🟩	🟩		🟩	🟩			🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩
Tier 2																			
A549 (Lung Cancer)				🟩				🟩			🟩								🟩
CD20+ (Immunity related)																			
CD20+_RO01778																			
CD20+_RO01794																			
H1-neurons (Nerve Cell)										🟩		🟩	🟩	🟩					🟩
HeLa-S3 (Cervical Cancer)	🟩			🟩							🟩	🟩							🟩

Source : http://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html

Case for Computational Tools



Existing Computational Tools



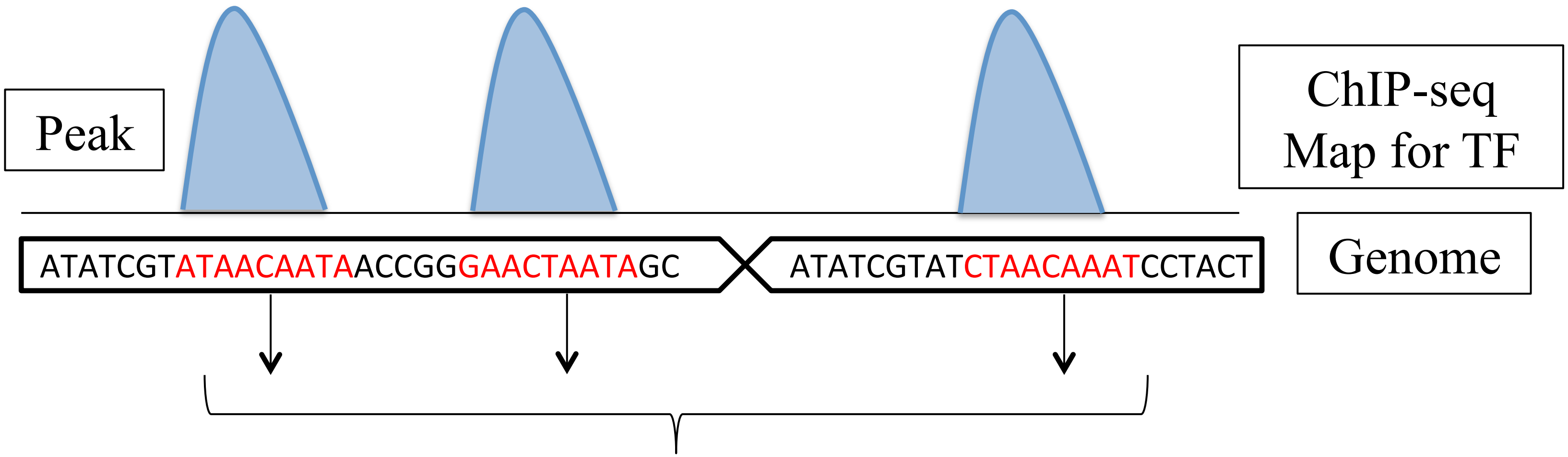
Generative Approaches

MEME
CISFINDER

Discriminative Approaches

STRING KERNEL+SVM

Generative : PWM Based approach



Peak

ChIP-seq
Map for TF

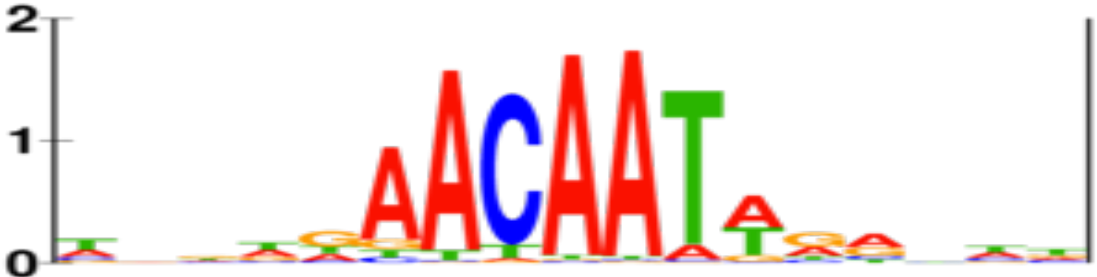
Genome

ATATCGTATAACAATAACCGGGAACTAATAGC

ATATCGTATCTAACAAATCCTACT

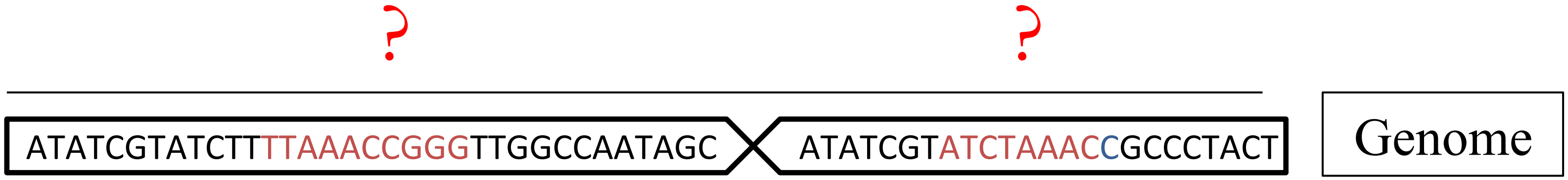
	1	2	3	4	5	6	7	8	9	10	11	12
A	14	0	0	14	28	40	9	45	42	13	15	9
T	12	3	4	12	11	10	9	6	5	38	12	3
C	3	0	1	8	2	2	36	2	2	0	1	0
G	0	1	0	16	10	1	2	3	2	0	7	11

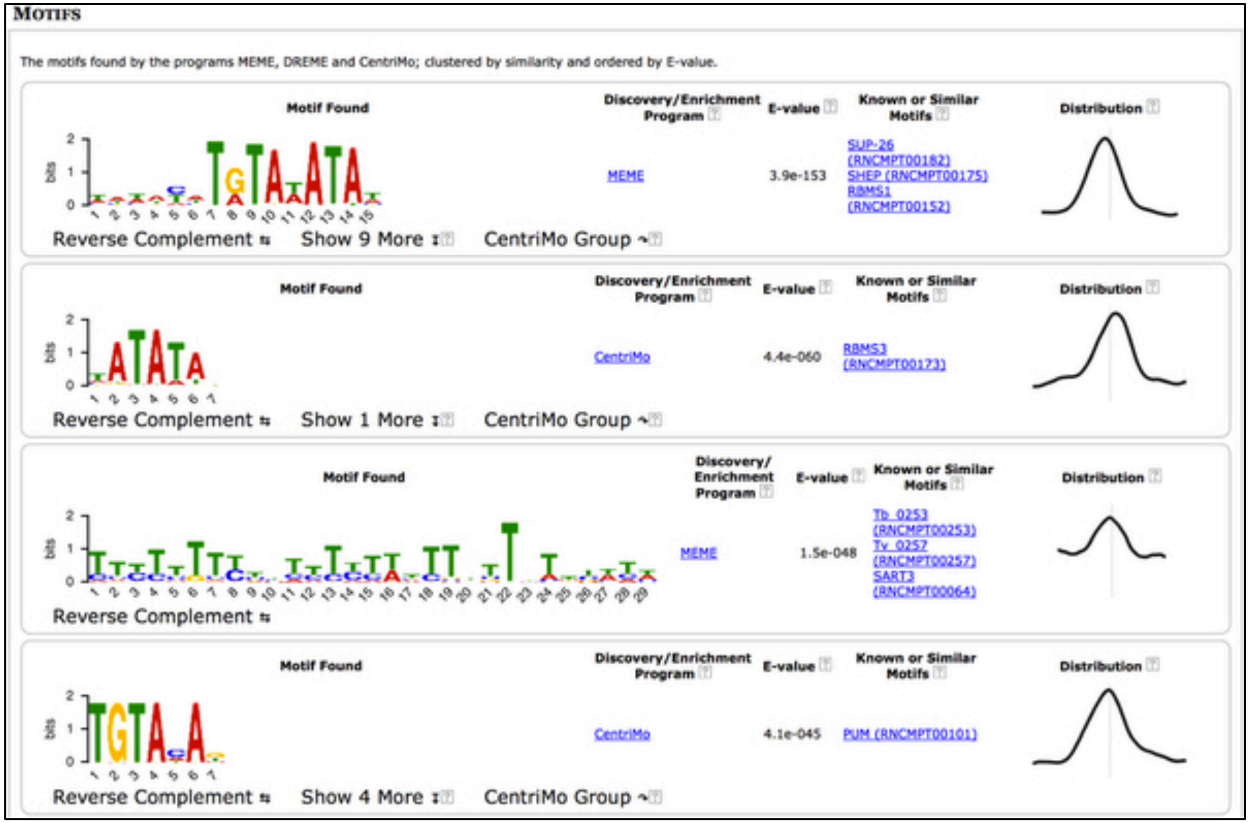
Position
Weight
Matrix



Sequence
Logo

Generative Approach : Output





Motifs

The motifs found by the programs MEME, DREME and CentriMo; clustered by similarity and ordered by E-value.

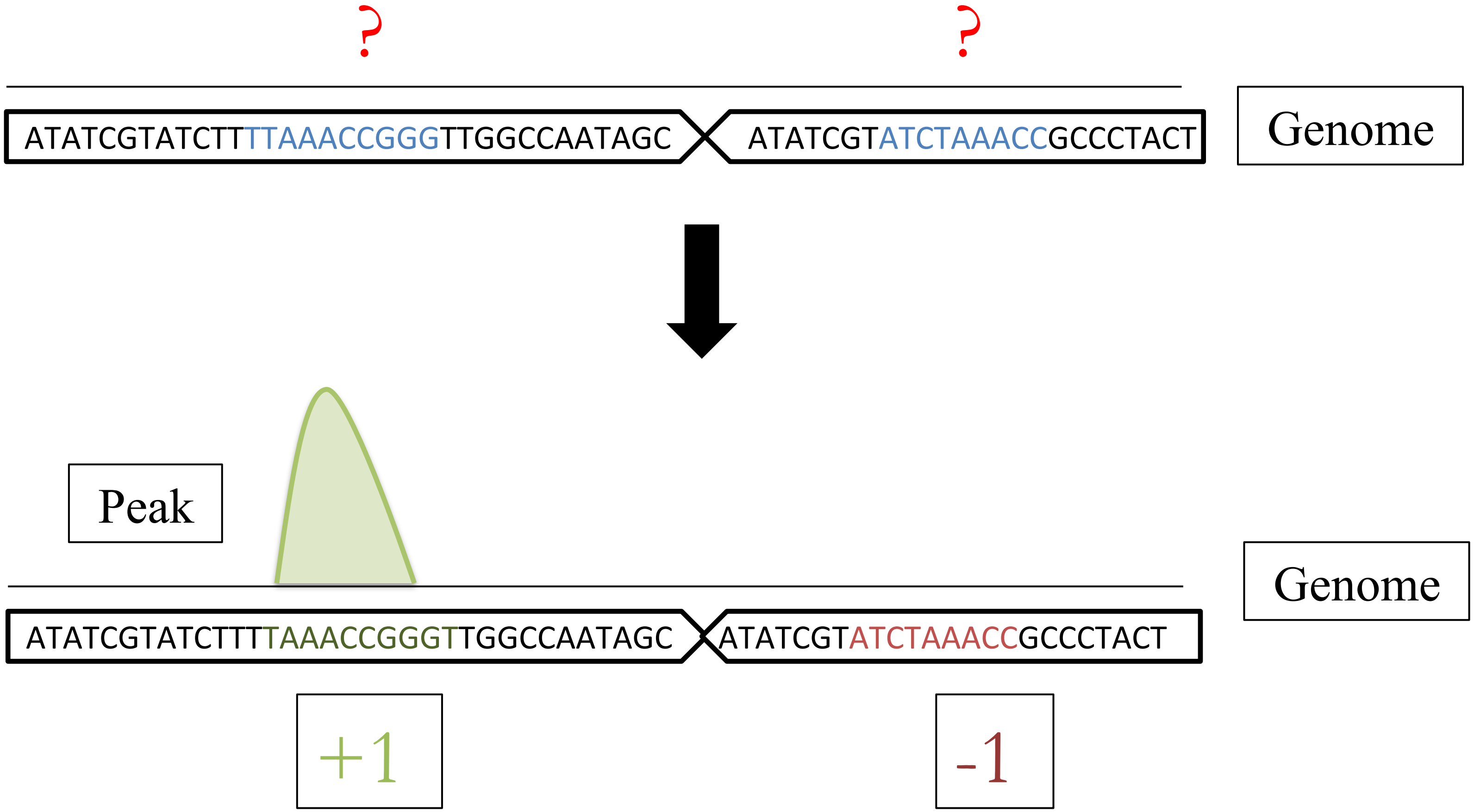
Motif Found	Discovery/Enrichment Program	E-value	Known or Similar Motifs	Distribution
Reverse Complement ⇆ Show 9 More ↕ CentriMo Group	MEME	3.9e-153	SUP-26 (RNCMPT00182) SHEP (RNCMPT00175) RBMS1 (RNCMPT00152)	
Reverse Complement ⇆ Show 1 More ↕ CentriMo Group	CentriMo	4.4e-060	RBMS3 (RNCMPT00173)	
Reverse Complement ⇆	MEME	1.5e-048	Tb_0253 (RNCMPT00253) Tv_0257 (RNCMPT00257) SART3 (RNCMPT00064)	
Reverse Complement ⇆ Show 4 More ↕ CentriMo Group	CentriMo	4.1e-045	PUM (RNCMPT00101)	

Source : http://www.cbil.upenn.edu/EpoDB/release/version_2.2/meme/meme-output.html#sample

Generative Approach: Limitations

- Output: **Long list** of potential TFs
- Work well for only **well preserved motifs or large training datasets**
- PWMs for all ~2000 TFs **not available**
- **Lower** prediction performance than discriminative approaches

Discriminative Approach : Output



? ?

ATATCGTATCTTTTAAACCGGGTTGGCCAATAGC ATATCGTATCTAAACCGCCCTACT

Genome

Peak

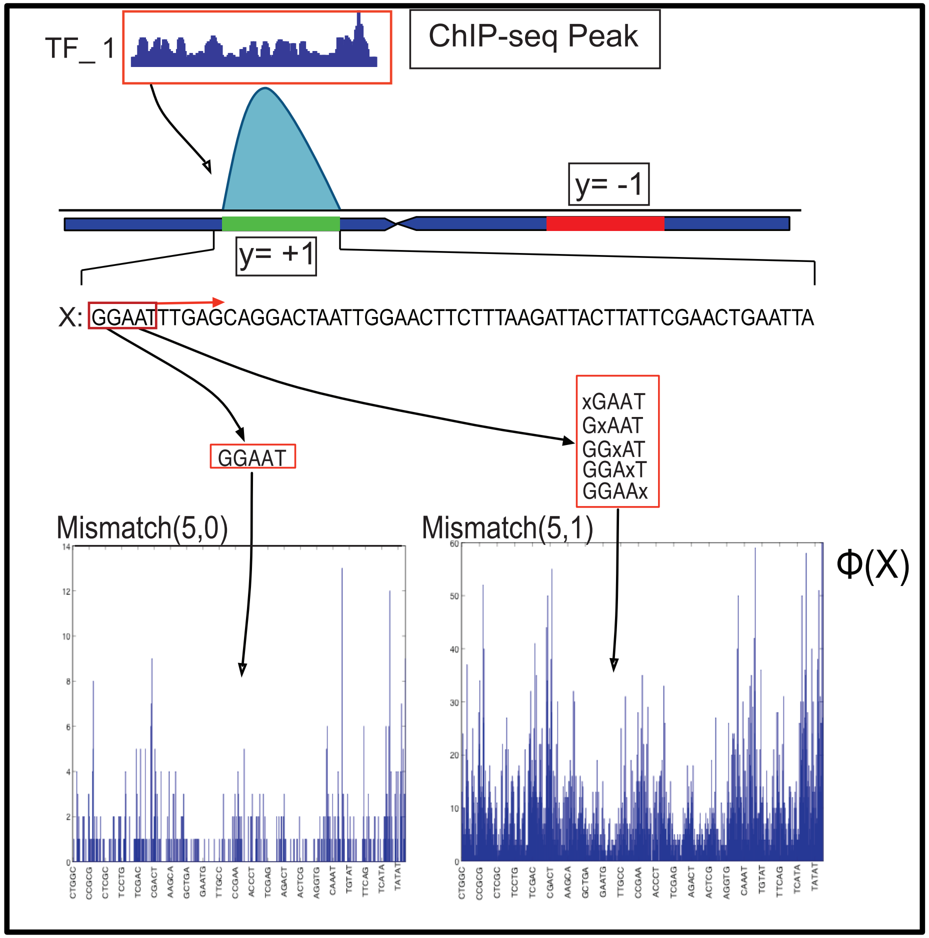
ATATCGTATCTTTTAAACCGGGTTGGCCAATAGC ATATCGTATCTAAACCGCCCTACT

Genome

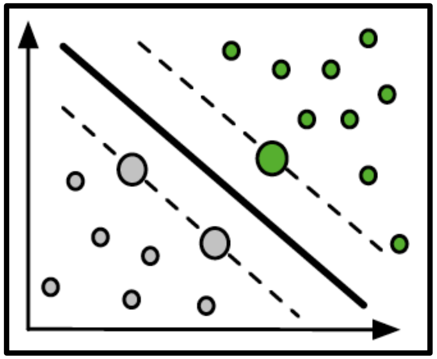
+1

-1

Discriminative : String Kernel Approach



Support Vector Machine



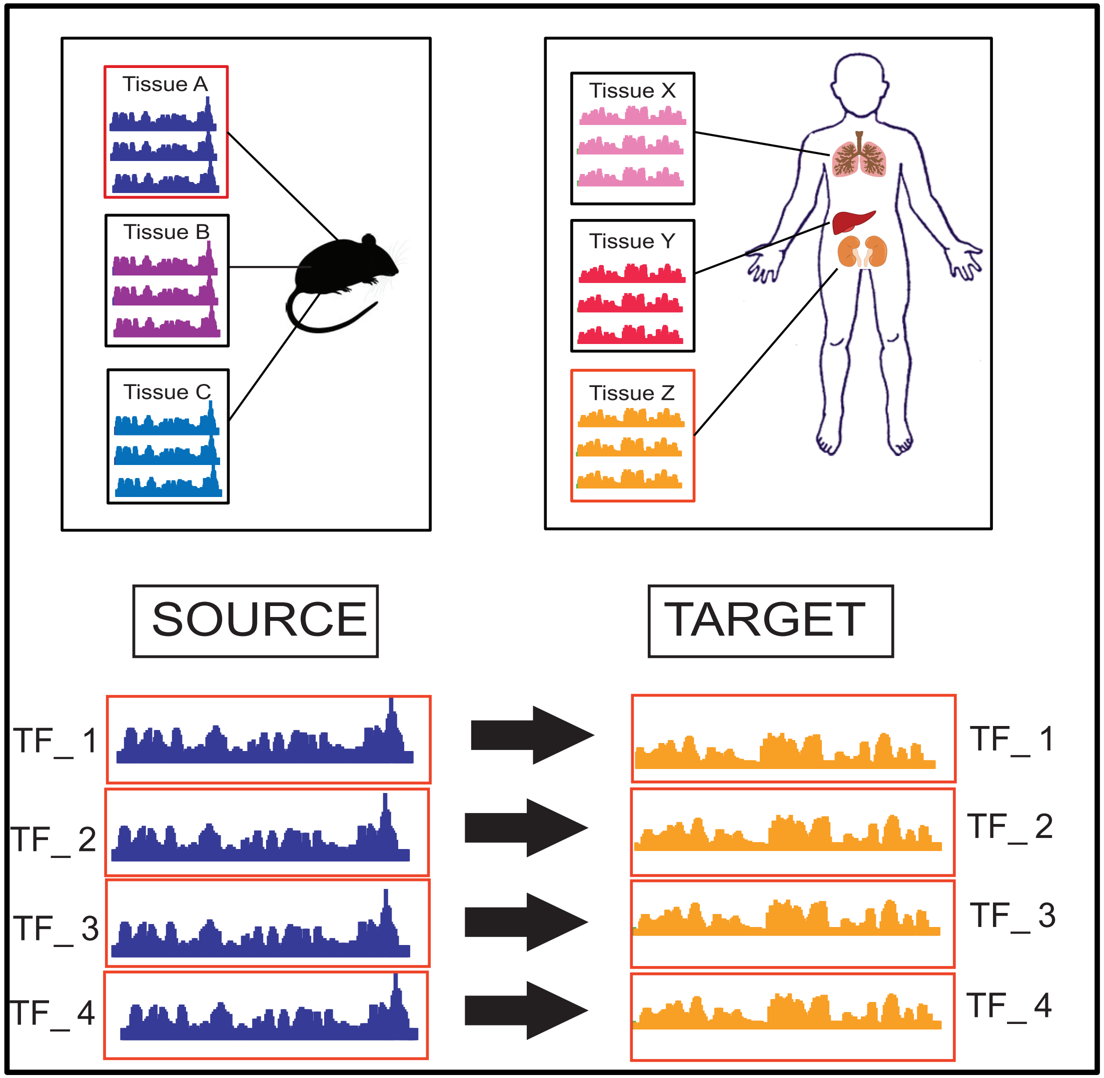
Discriminative Approach : Limitation

Assumption: Training/test data follow **same distribution regardless of context.**

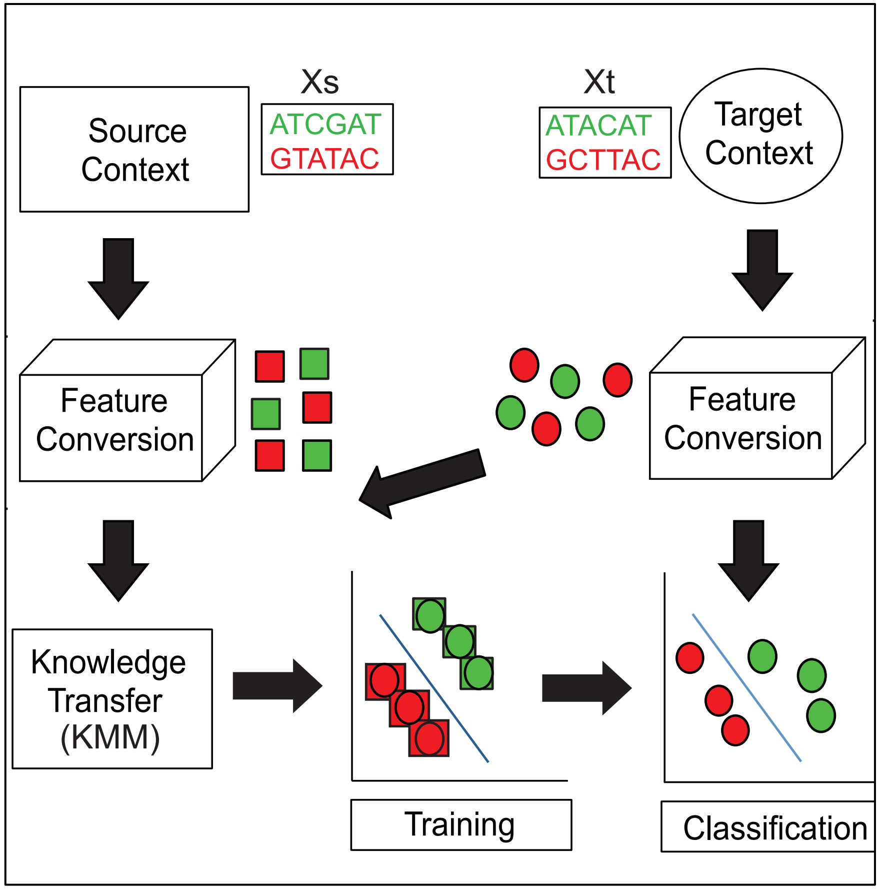
Aim

- **Improve** prediction of Transcription Factor Binding sites across contexts using knowledge transfer.

Proposed Solution : Cross-Context Knowledge Transfer



Transfer String Kernel : Overview



Outline

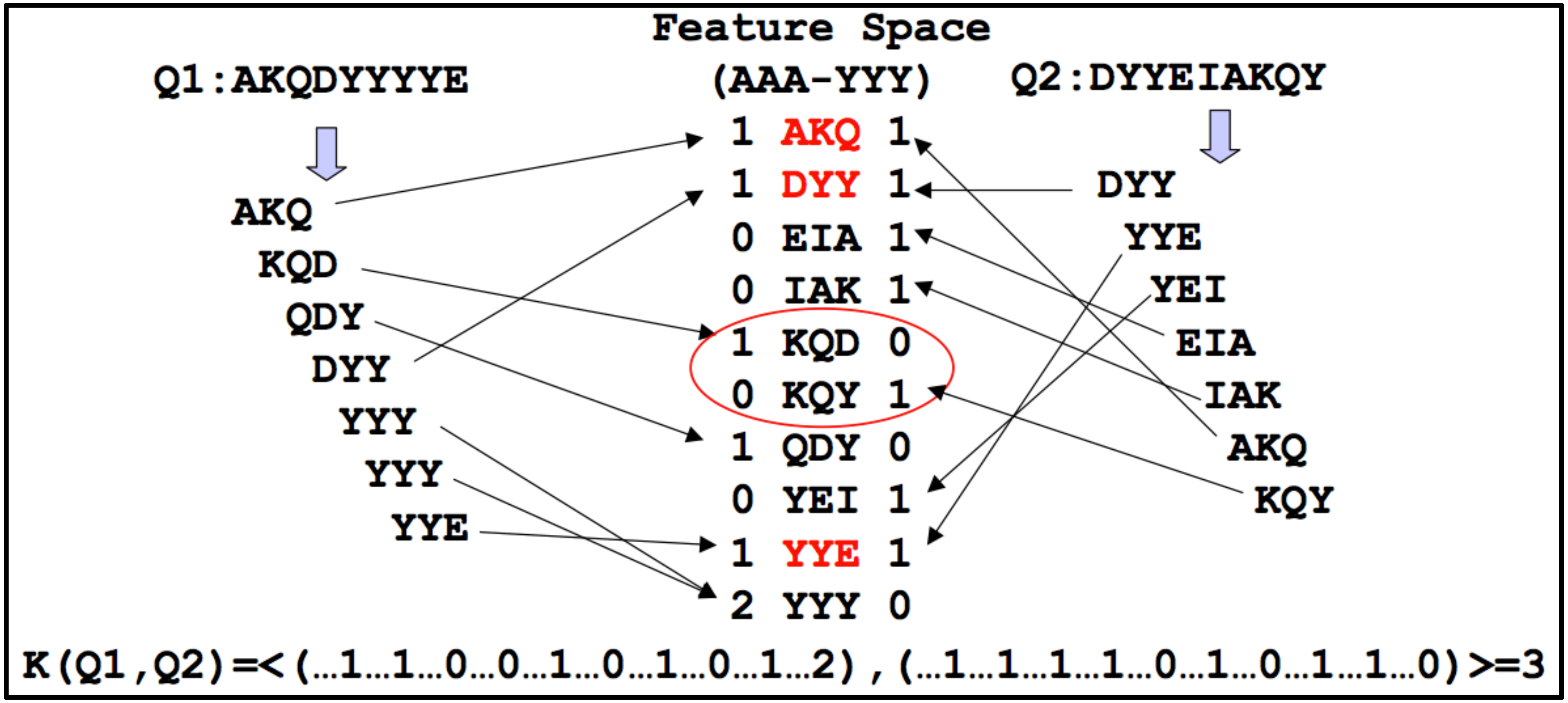
- Method
  - String Kernel
  - Support Vector Machine
  - Transfer Learning (KMM)
  - Importance re-weighting
  - Transfer String Kernel
- Evaluation
  - Experimental Setup
  - Cross-context TFBS prediction
  - Cross-context Protein Binding prediction

Outline

- Method
  - String Kernel
  - Support Vector Machine
  - Transfer Learning (KMM)
  - Importance re-weighting
  - Transfer String Kernel
- Evaluation
  - Experimental Setup
  - Cross-context TFBS prediction
  - Cross-context Protein Binding prediction

String Kernel : Spectrum Kernel

Feature map indexed by all k-length subsequences ("k-mers") from alphabet Σ of amino acids, |Σ|=20



Feature Space
(AAA-YYY)

Q1:AKQDYYYYE

AKQ
KQD
QDY
DYY
YYY
YYY
YYE

Q2:DYYEIAKQY

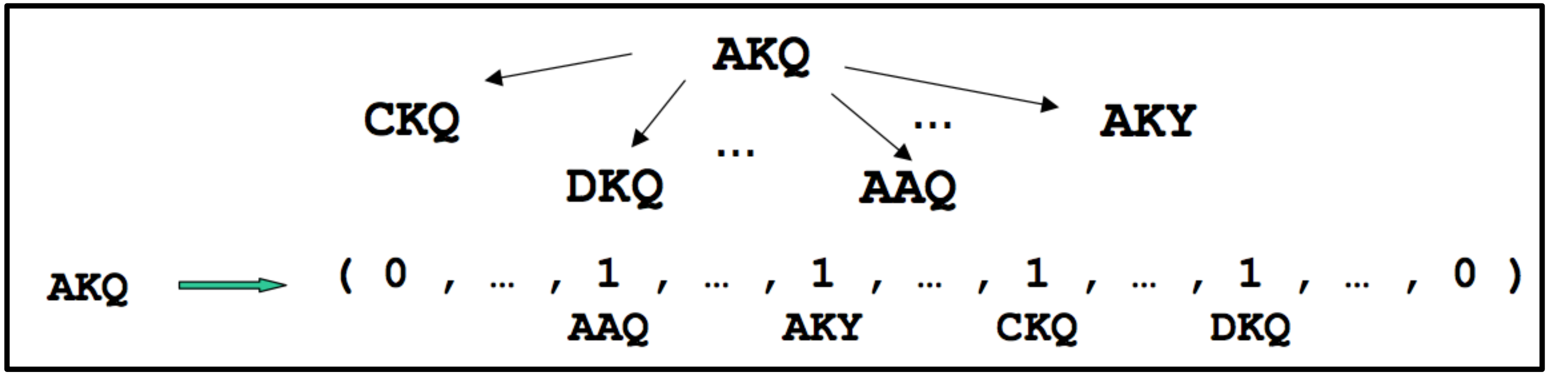
DYY
YYE
YEI
EIA
IAK
AKQ
KQY

1	AKQ	1
1	DYY	1
0	EIA	1
0	IAK	1
1	KQD	0
0	KQY	1
1	QDY	0
0	YEI	1
1	YYE	1
2	YYY	0

K(Q1,Q2)=<(…1…1…0…0…1…0…1…0…1…2),(…1…1…1…1…0…1…0…1…1…0)>=3

String Kernel : Mismatch Kernel

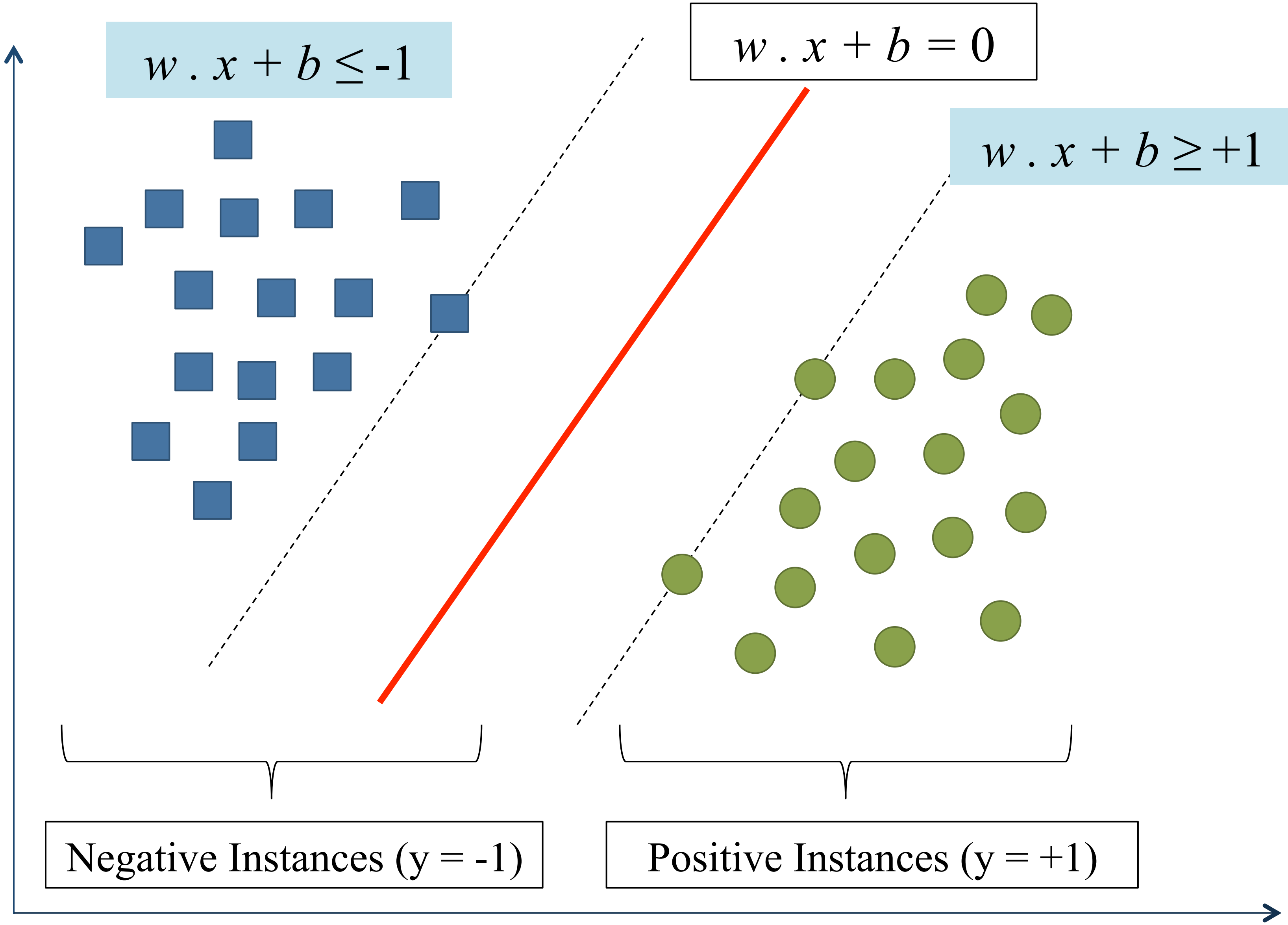
For k-mer **s**, the mismatch neighborhood $N_{(k,m)}(s)$ is the set of all k-mers *t* within *m* mismatches from **s**.



Outline

- Method
  - String Kernel
  - Support Vector Machine
  - Transfer Learning (KMM)
  - Importance re-weighting
  - Transfer String Kernel
- Evaluation
  - Experimental Setup
  - Cross-context TFBS prediction
  - Cross-context Protein Binding prediction

Support Vector Machine



$w . x + b \leq -1$

$w . x + b = 0$

$w . x + b \geq +1$

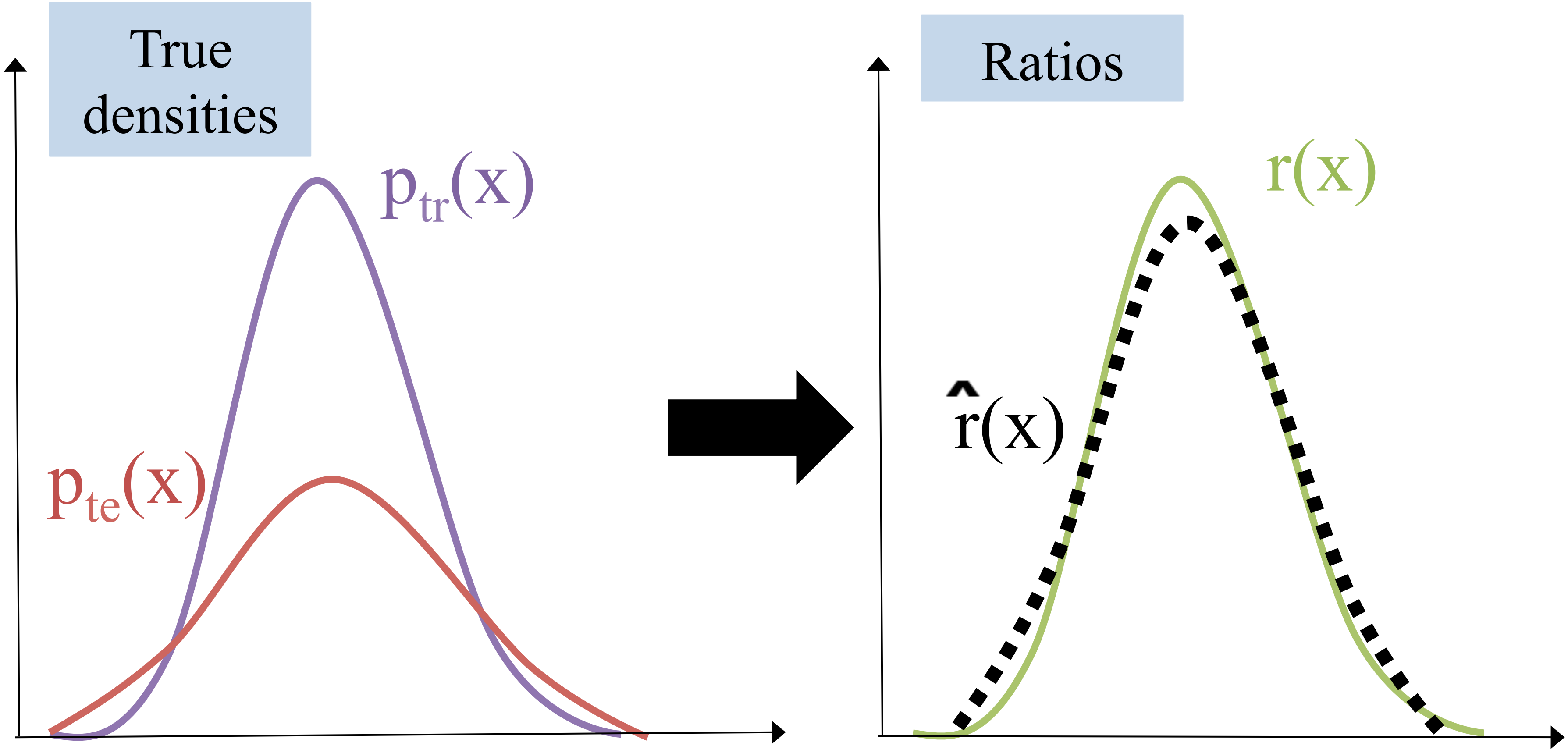
Negative Instances (y = -1)

Positive Instances (y = +1)

Outline

- Method
  - String Kernel
  - Support Vector Machine
  - Transfer Learning (KMM)
  - Importance re-weighting
  - Transfer String Kernel
- Evaluation
  - Experimental Setup
  - Cross-context TFBS prediction
  - Cross-context Protein Binding prediction

Transfer Learning (KMM)



True densities

Ratios

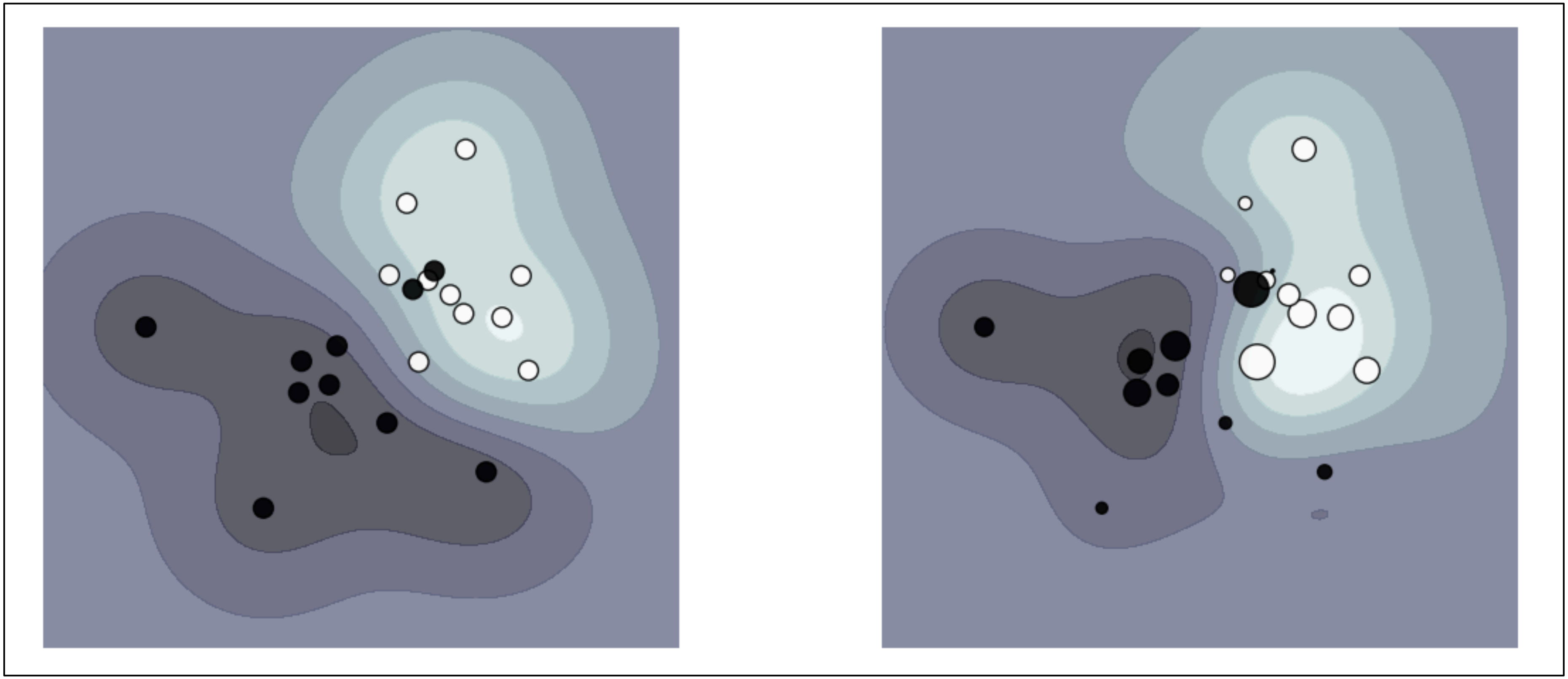
Outline

- Method
  - String Kernel
  - Support Vector Machine
  - Transfer Learning (KMM)
  - Importance re-weighting
  - Transfer String Kernel
- Evaluation
  - Experimental Setup
  - Cross-context TFBS prediction
  - Cross-context Protein Binding prediction

Importance Re-weighting

Original Weights

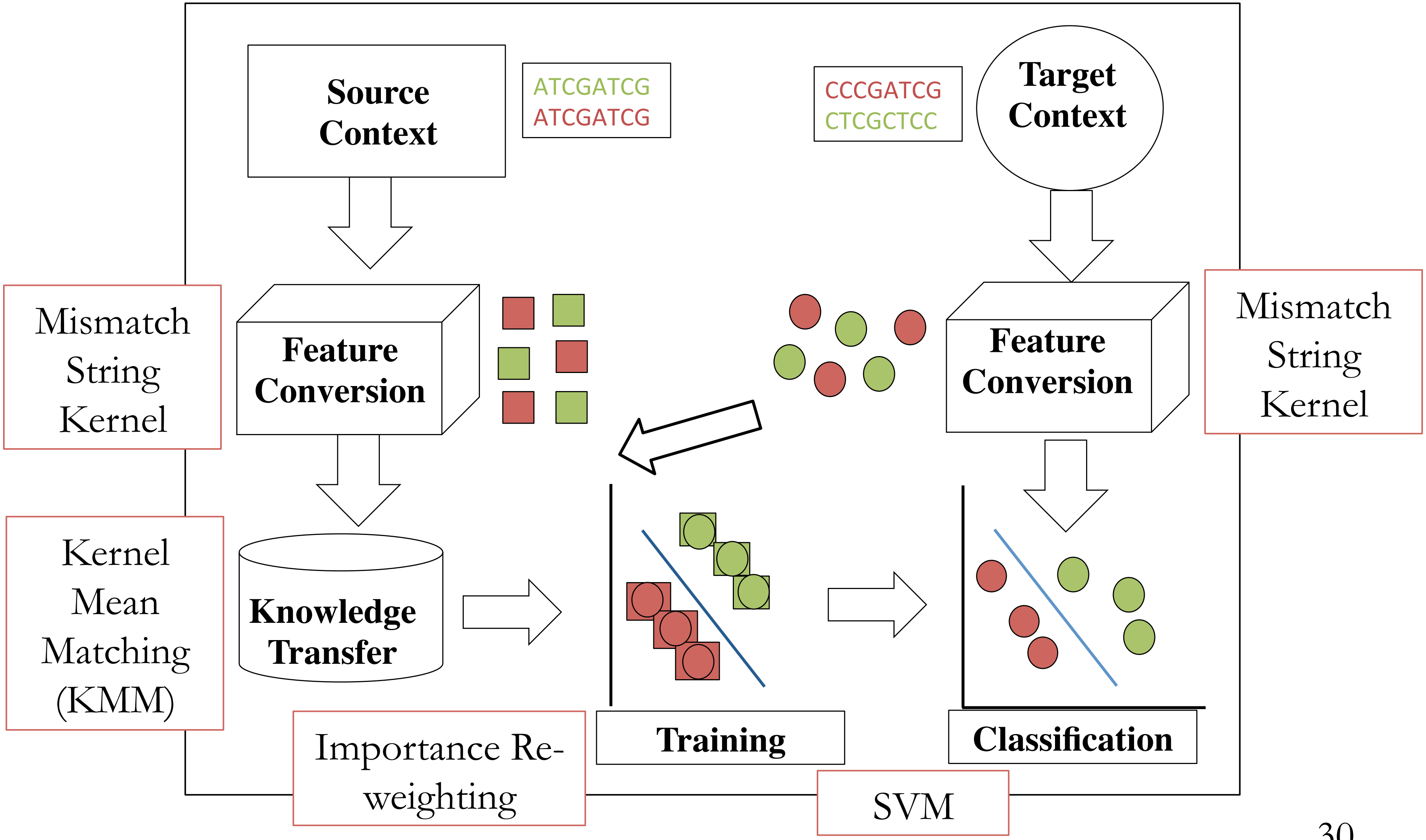
KMM Weights



Outline

- Method
  - String Kernel
  - Support Vector Machine
  - Transfer Learning (KMM)
  - Importance re-weighting
  - Transfer String Kernel
- Evaluation
  - Experimental Setup
  - Cross-context TFBS prediction
  - Cross-context Protein Binding prediction

Transfer String Kernel (TSK)



Source Context

ATCGATCG
ATCGATCG

CCCGATCG
CTCGCTCC

Target Context

Mismatch String Kernel

Feature Conversion

Feature Conversion

Mismatch String Kernel

Kernel Mean Matching (KMM)

Knowledge Transfer

Training

Classification

Importance Re-weighting

SVM

Outline

- Method
  - String Kernel
  - Support Vector Machine
  - Transfer Learning (KMM)
  - Importance re-weighting
  - Transfer String Kernel
- Evaluation
  - Experimental Setup
  - Cross-context TFBS prediction
  - Cross-context Protein Binding prediction

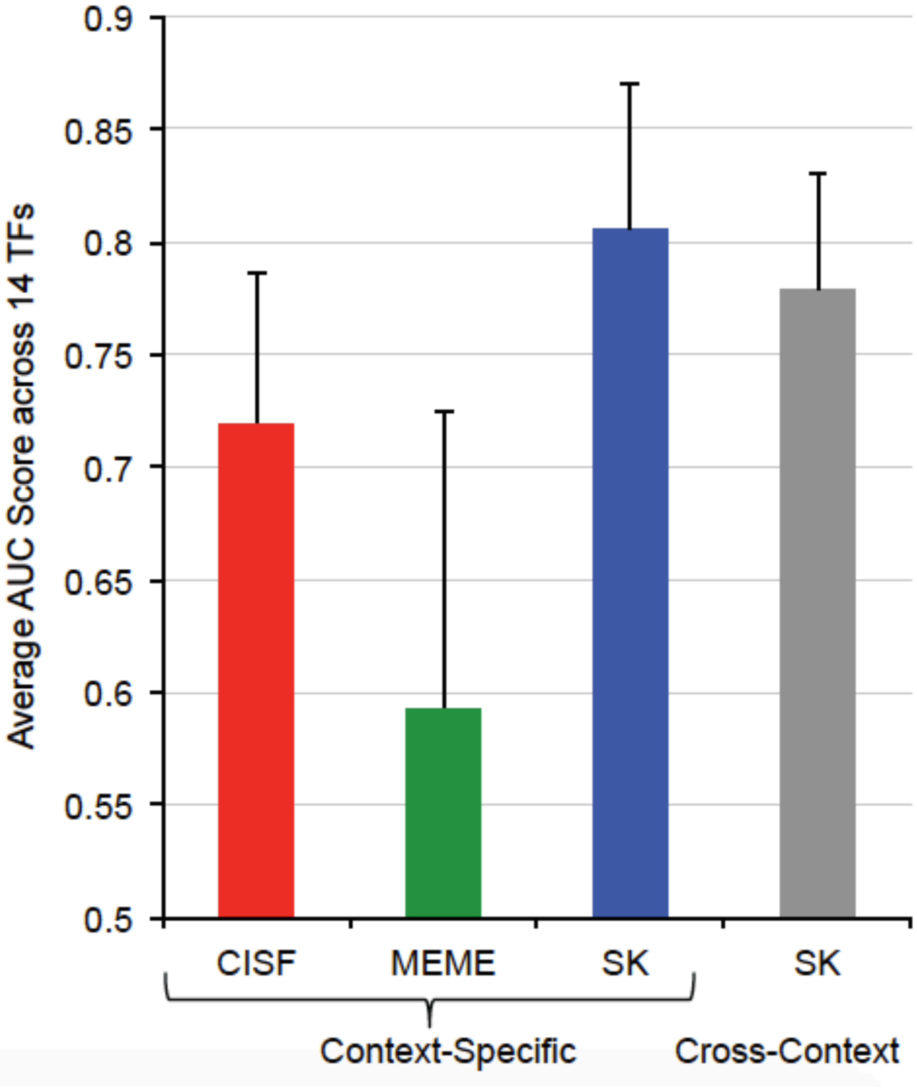
Experimental Setup

- 14 Transcription Factors (ENCODE ChIP-seq)
- Top 1000 positive sequences (500 training and 500 testing)
- 1000 random negative sequences
- Hyper-parameter tuning for k=(8,10,12) and m=(1,2,3)
- Dictionary size = 4 {A,T,C,G}

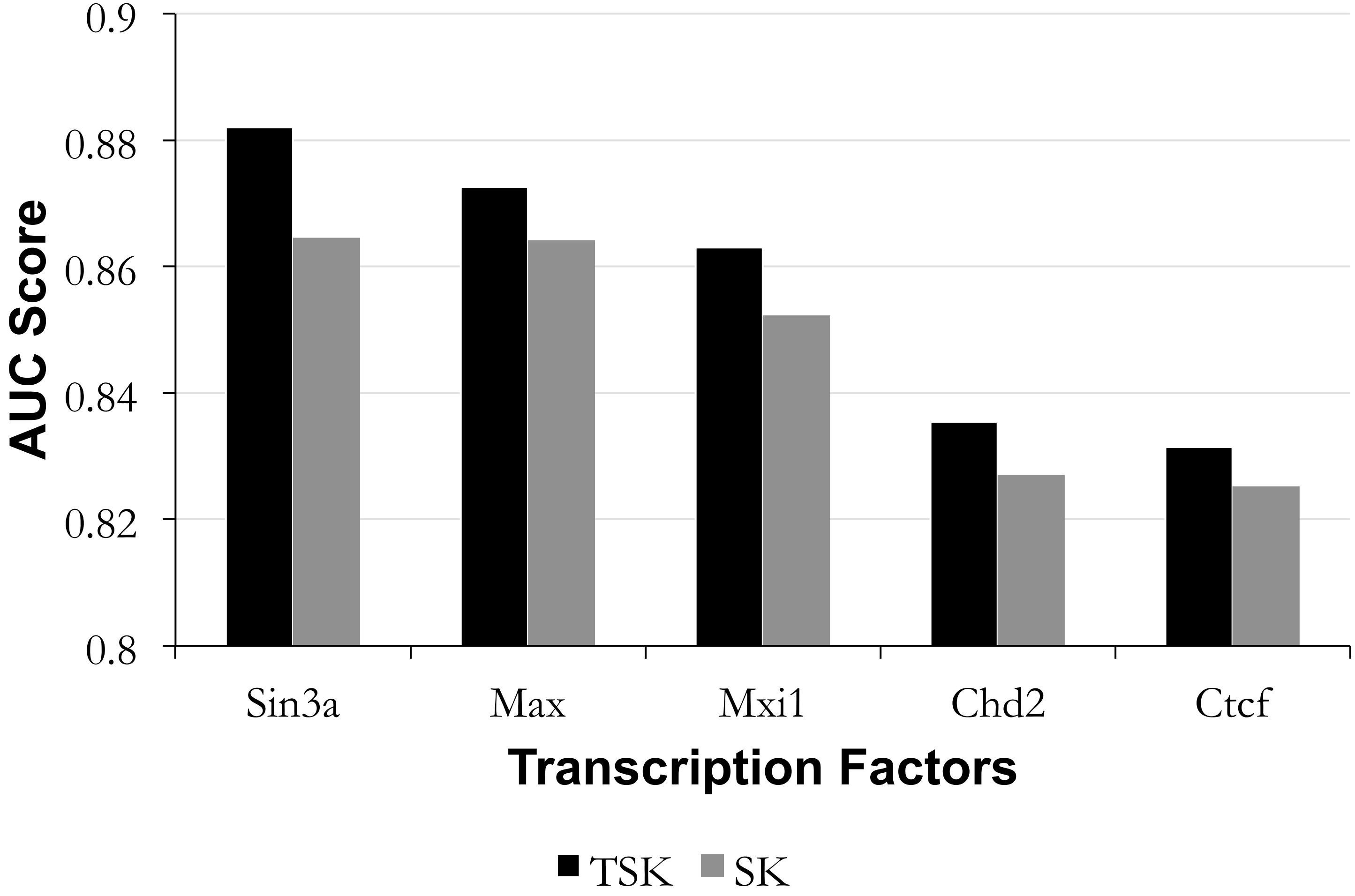
Outline

- Method
  - String Kernel
  - Support Vector Machine
  - Transfer Learning (KMM)
  - Importance re-weighting
  - Transfer String Kernel
- Evaluation
  - Experimental Setup
  - Cross-context TFBS prediction
  - Cross-context Protein Binding prediction

Results


0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Average AUC Score across 14 TFs
CISF
MEME
SK
SK
Context-Specific
Cross-Context

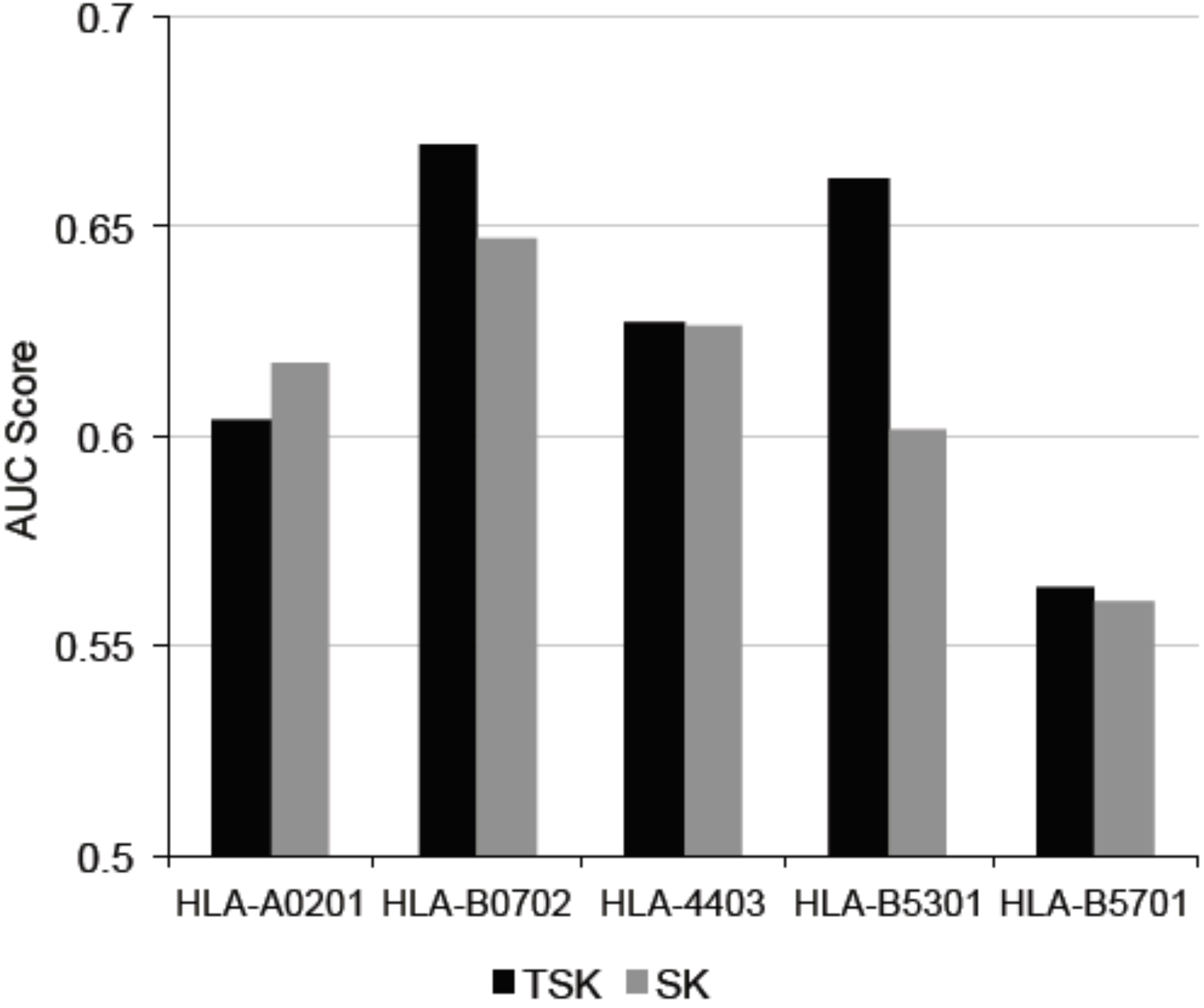
Results – Cross Context



Outline

- Method
  - String Kernel
  - Support Vector Machine
  - Transfer Learning (KMM)
  - Importance re-weighting
  - Transfer String Kernel
- Evaluation
  - Experimental Setup
  - Cross-context TFBS prediction
  - Cross-context Protein Binding prediction

Results – Cross context



Summary

- TSK overall improves the cross-context TFBS predictions;

- String kernel based approaches perform better than the state-of-the-art Position Weight/Frequency Matrix based TFBS tools;

- TSK approach is generalizable for performance improvement of any cross-context sequence prediction task.

Presented in BIOKDD ’15

Acknowledgements

Dr. Mazhar Adli

Adli Lab : Department of Biochemistry and Molecular Genetics @Uva

Nipun Batra

IIIT-Delhi

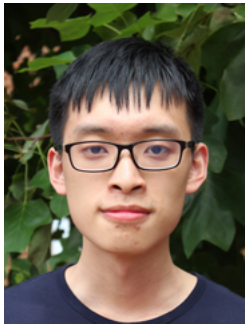
Machine Learning Lab @ UVa

Dr. Yanjun Qi (Advisor)

Beilun Wang

Weilin Xu

Jack Lanchantin



Ji Gao

Future Directions

- Deep Learning :
  - Gene expression prediction using histone modification data (ECCB 2016)
  - Improving TFBS prediction using DNA sequences (ICLR Workshop 2016, ICML Workshop 2016)

- String Kernels: Improving efficiency!! (on-going work)

Thank You