A weighted two-stage sequence alignment framework to identify DNA motifs from ChIP-exo data

Cankun Wang
Biomedical Informatics Specialist
Department of Biomedical Informatics
The Ohio State University

July 18, 2023

The Ohio State University
WEXNER MEDICAL CENTER

BMBL
Bioinformatics and Mathematical Biosciences Lab

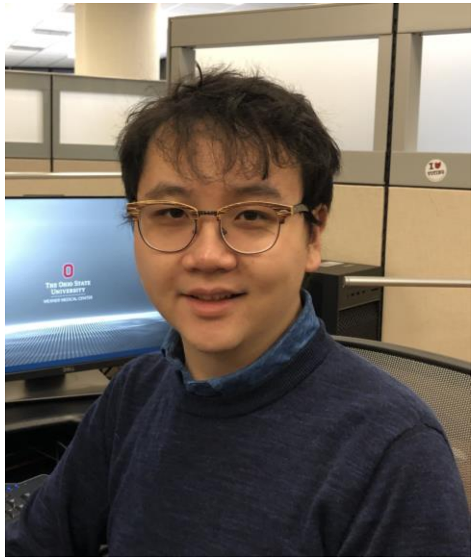
About myself

Research interest:

Inference of gene regulatory mechanisms across various organisms

Developing tools and benchmarking pipelines for next-generation sequencing data

Developing cloud-native biomedical applications for webservers and databases



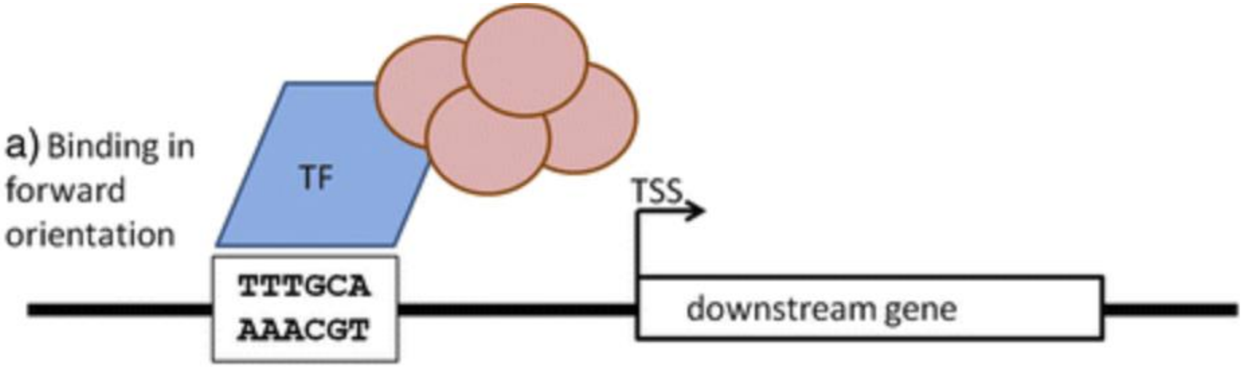
Cankun Wang
Biomedical Informatics Specialist

BMBL
Bioinformatics and Mathematical Biosciences Lab

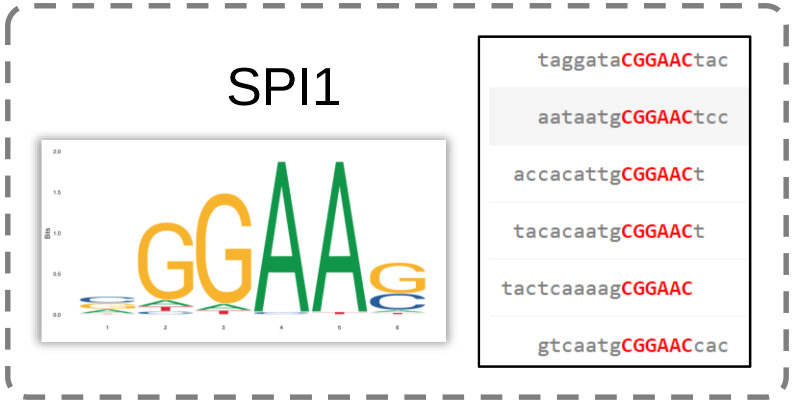
Advisor: Dr. Qin Ma

Background of prediction of transcription factor binding sites

Transcription factor
(Regulatory protein)



Binding sites
(DNA motifs)



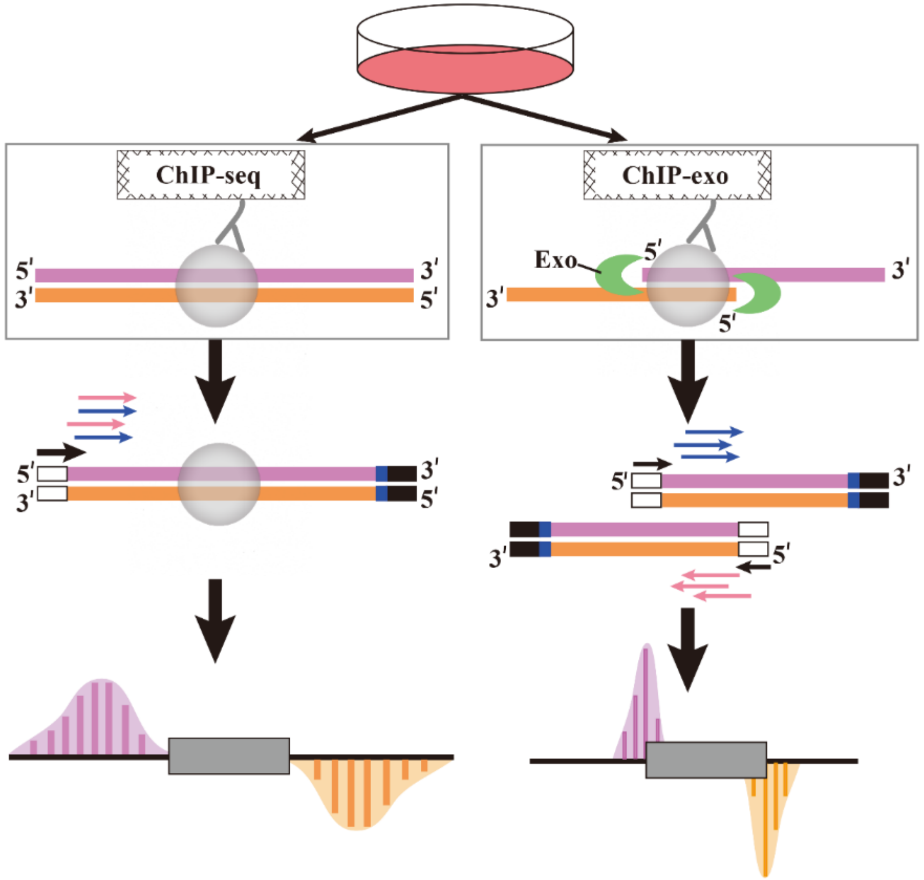
taggataCGGAACtac
aataatgCGGAACtcc
accacattgCGGAACt
tacacaatgCGGAACt
tactcaaaagCGGAAC
gtcaatgCGGAACcac

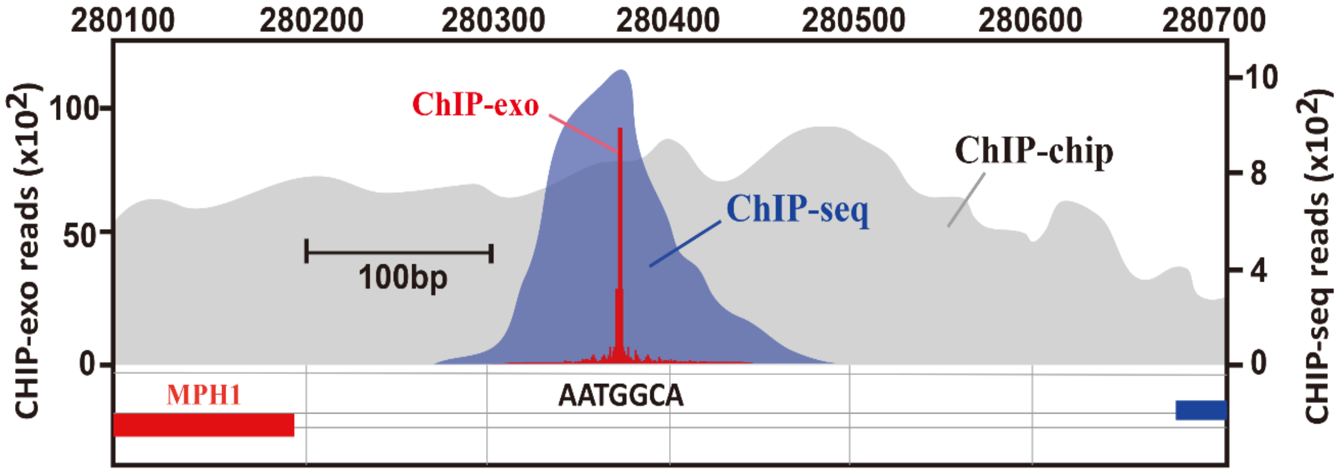
Transcription factor (TF) --- a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence, i.e., TF binding sites (TFBSs).

Bailey, *et al.*, *Nucleic Acids Res.* (2015)
Liu, *et al.*, *Briefs. Bioinformatics* (2018)

Why ChIP-exo Over ChIP-seq for Motif Finding

- ChIP-seq provides genome-wide map of protein-DNA interactions, yet resolution could be insufficient for accurate binding site identification
- Compared with ChIP-seq, Chromatin Immunoprecipitation combined with lambda exonuclease digestion followed by high-throughput sequencing (ChIP-exo) has relatively low noise and achieves near-base pair resolution.





Peter J. Park, *Nature Reviews Genetics.* (2009)
Rossi, *et al.*, *Nature Communications* (2018)

Why new motif discovery tool is needed for ChIP-exo data?

1. **High Sensitivity of ChIP-exo**: Traditional tools may not handle the increased sensitivity to experimental conditions in ChIP-exo data adequately.
2. **Peak Position Variance:** Traditional motif finding tools often assume the binding site is at the peak center, which does not hold true for ChIP-exo data, necessitating a more flexible algorithm.
3. **Sharpness of Peaks**: ChIP-exo data generates sharper peaks than ChIP-seq, making it difficult for traditional tools to distinguish between adjacent binding events.
4. **Need for Advanced Techniques**: New tool is required to account for these intricacies, enhancing signal-to-noise ratio and providing a more precise motif discovery.

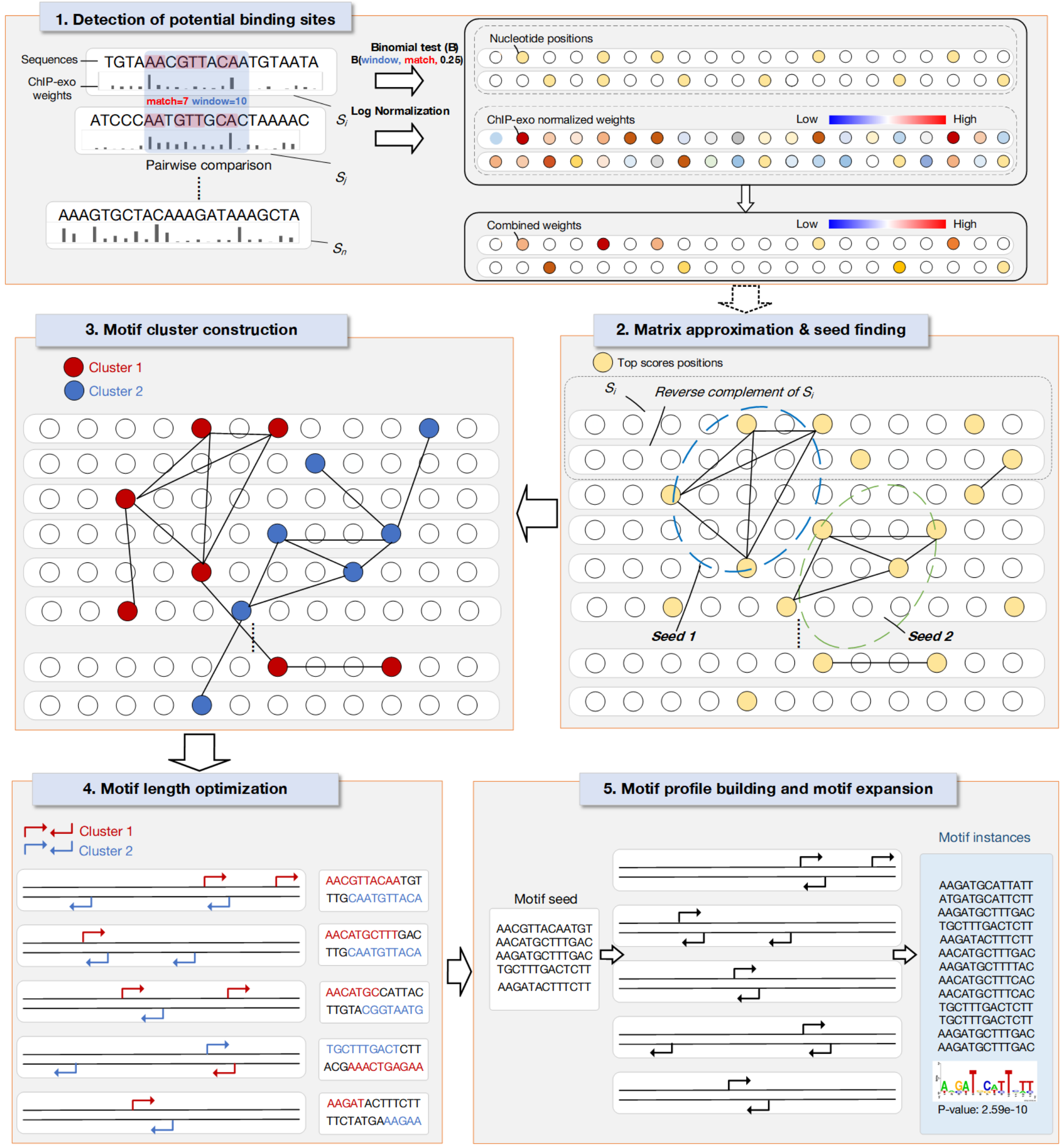
TESA is a novel motif finding tool designed for ChIP-exo data

TESA (**A weighted two-stage sequence alignment framework to identify DNA motifs from ChIP-exo data**) was designed specifically to address the unique challenges presented by ChIP-exo data.

Unique features:

- Base-level Signal Extraction: TESA leverages ChIP-exo's high-resolution, base-level data, enhancing signal-to-noise ratio and providing precise motif discovery
- Dynamic Motif Length Optimization: Through a bookend model, TESA automatically adjusts motif lengths based on TFBS clustering
- Precision in Distinguishing Adjacent Binding Events: Using a binomial test, TESA determines whether potential TF binding site clusters should be combined or treated separately

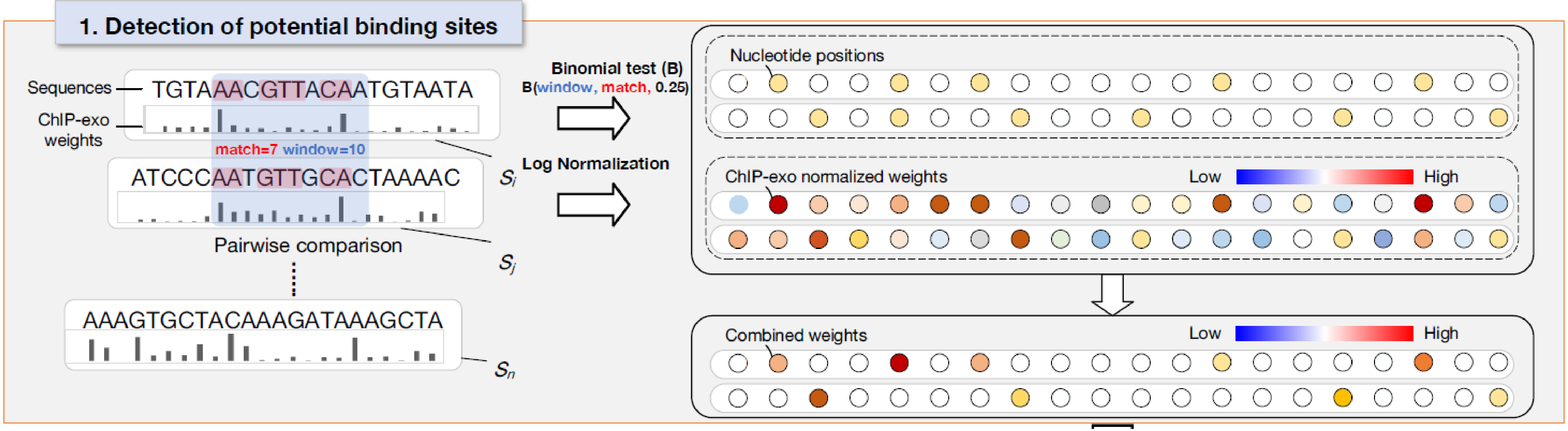
Overview of TESA (weighTEd two-Stage Alignment tool)



Steps of TESA:

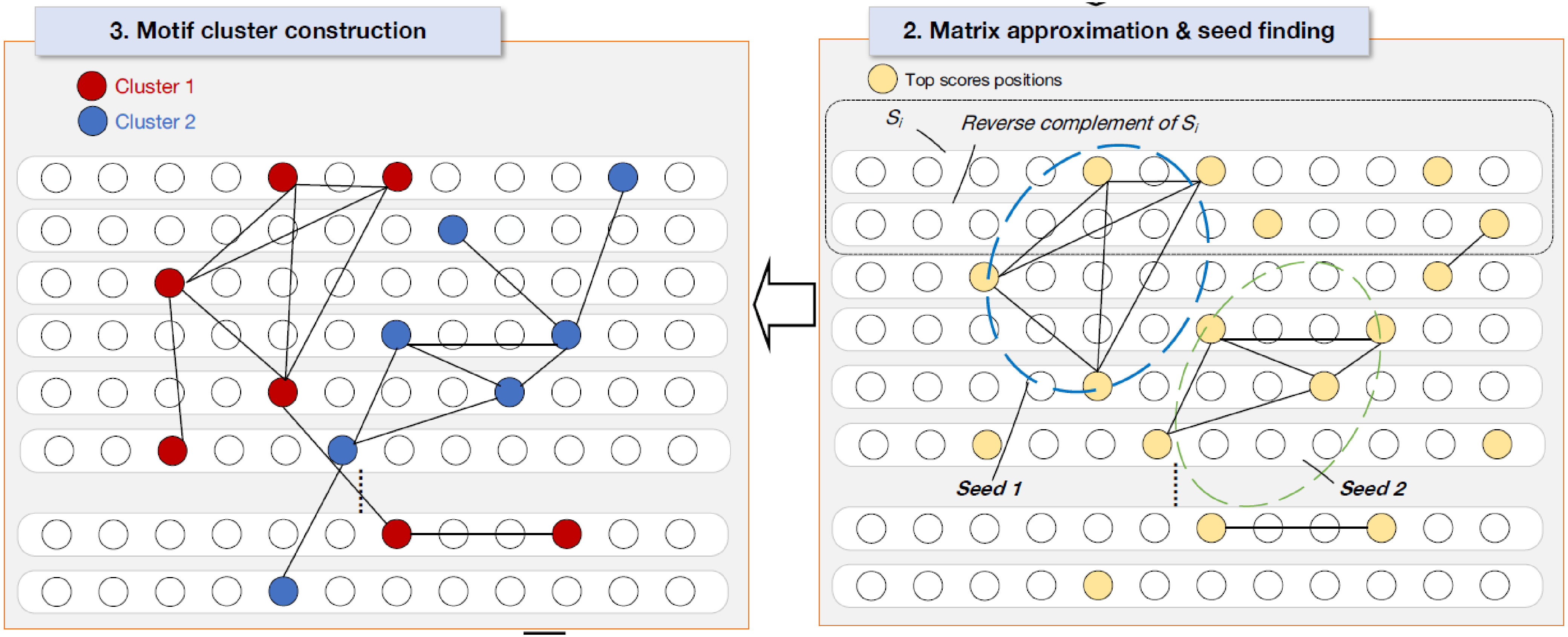
1. Detection of potential binding sites
2. Matrix approximation and seed finding
3. Motif cluster construction
4. Motif length optimization
5. Motif profile building and motif expansion

Detection of potential binding sites



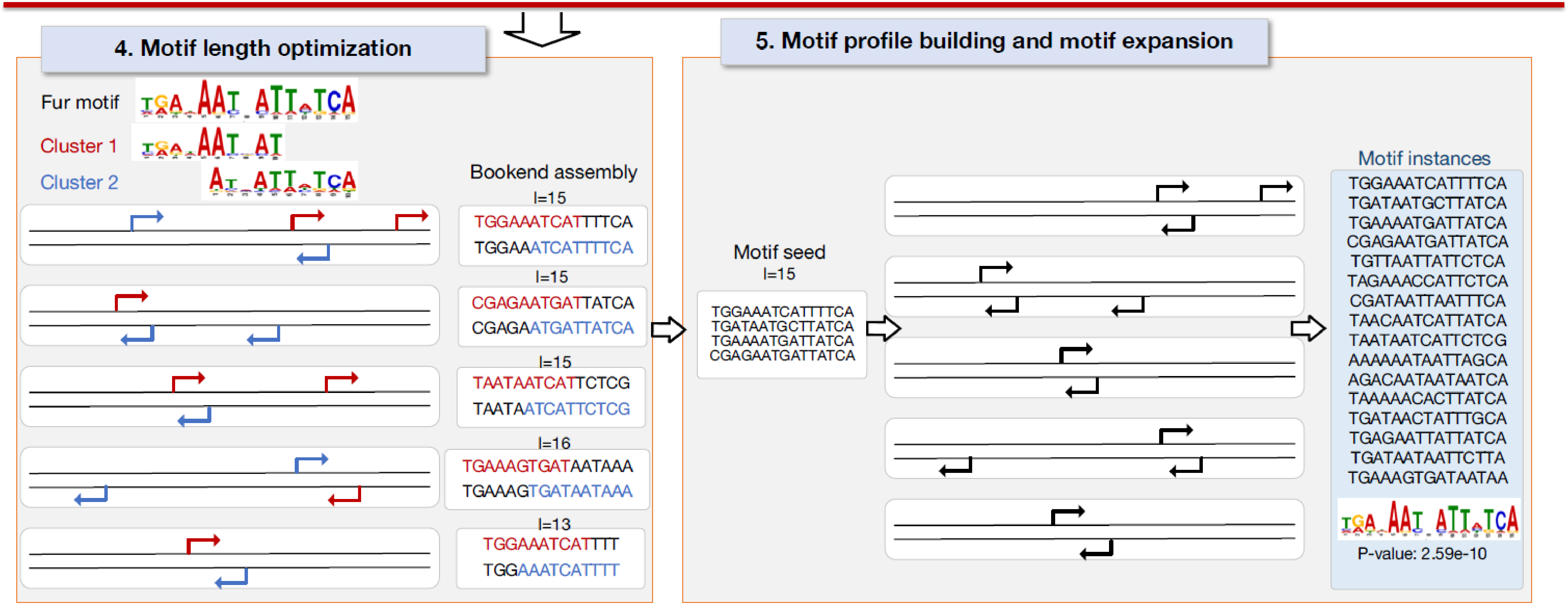
Compared with ChIP-seq, Chromatin Immunoprecipitation combined with lambda exonuclease digestion followed by high-throughput sequencing (ChIP-exo) has relatively low noise and achieves near-base pair resolution.

Matrix Approximation and Cluster Finding


3. Motif cluster construction
Cluster 1
Cluster 2
2. Matrix approximation & seed finding
Top scores positions
S_i
Reverse complement of S_i
Seed 1
Seed 2

- Initial data matrix is refined to highlight potential binding site locations.
- constructed network with points representing possible binding sites.
- Connections are made to identify similar potential binding sites.
- Final clusters are formed by linking these seeds, signifying potential motifs.

Refinement of motif and length



4. Motif length optimization

Fur motif

Cluster 1

Cluster 2

Bookend assembly

l=15

TGGAAATCATTTTCA

TGGAAATCATTTTCA

l=15

CGAGAATGATTATCA

CGAGAATGATTATCA

l=15

TAATAATCATTCTCG

TAATAATCATTCTCG

l=16

TGAAAGTGATAATAAA

TGAAAGTGATAATAAA

l=13

TGGAAATCATTTT

TGGAAATCATTTT

5. Motif profile building and motif expansion

Motif seed

l=15

TGGAAATCATTTTCA
TGATAATGCTTATCA
TGAAAATGATTATCA
CGAGAATGATTATCA

Motif instances

TGGAAATCATTTTCA
TGATAATGCTTATCA
TGAAAATGATTATCA
CGAGAATGATTATCA
TGTTAATTATTCTCA
TAGAAACCATTCTCA
CGATAATTAATTTCA
TAACAATCATTATCA
TAATAATCATTCTCG
AAAAAATAATTAGCA
AGACAATAATAATCA
TAAAAACACTTATCA
TGATAACTATTTGCA
TGAGAATTATTATCA
TGATAATAATTCTTA
TGAAAGTGATAATAA

P-value: 2.59e-10

- Optimized Motif Lengths: TESA prevents premature truncation and enhances accuracy.
- Expanded Motifs: Sequence segments with high similarity scores are added to the motifs.
- Iterative Refinement: Continual reassessment and inclusion ensure highly refined motifs.

Methods for benchmarking the performance of motif finding

Methods to compare:

- **TESA,** BoBro, MEME-ChIP, and Weeder, DiNAMO, Dipartitle, MFMD
- 20 Escherichia coli ChIP-exo datasets downloaded from the proChIPdb database
- Validating the motifs discovered by TESA by comparing them with known motifs from the DPInteract database

BoBro: The segment alignment algorithm as basis of TESA (Li et al. 2011)

MEME-ChIP: This popular tool has been cited 1551 times, widespread use and recognition in the community (Bailey 2011)

Weeder: With 627 citations, Weeder stands as another well-recognized tool (Pavesi et al. 2014)

Added during revision:

DiNAMO: An exhaustive and efficient algorithm for motif discovery (Saas et al. 2018)

Dipartitle: A tool for detecting motifs by considering base interdependencies (Vehed et al. 2018)

MFMD: Semi-greedy constructive heuristics as a local optimizer for motif finding (Caldonazzo et al. 2019)

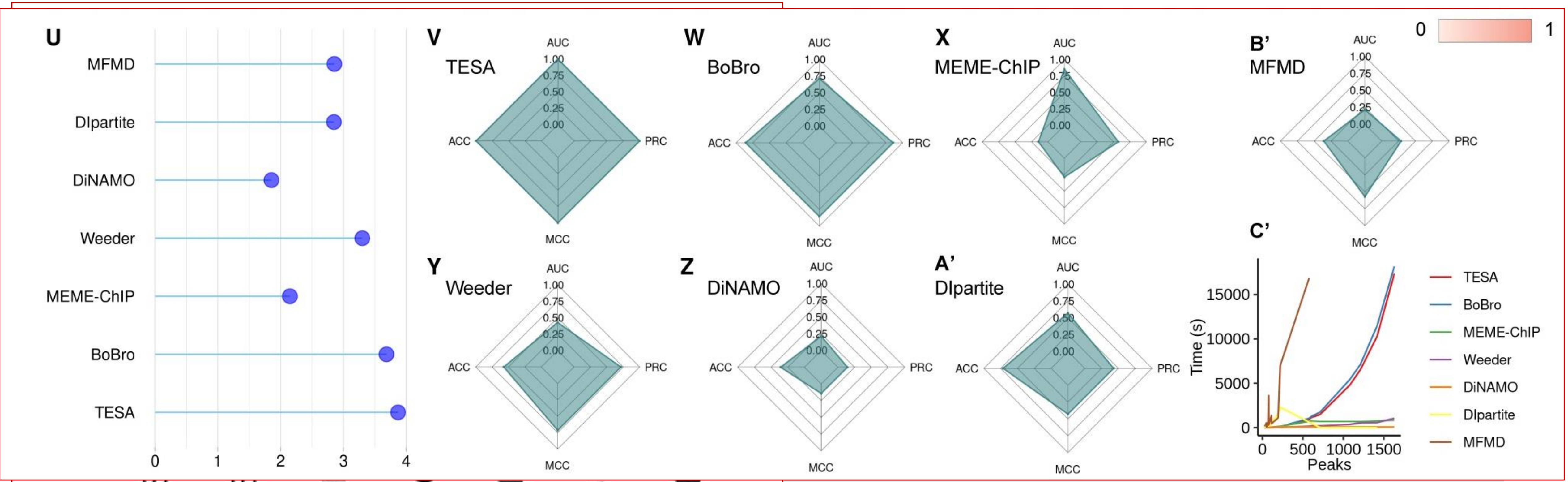
Evaluation metrics of the performance of motif finding

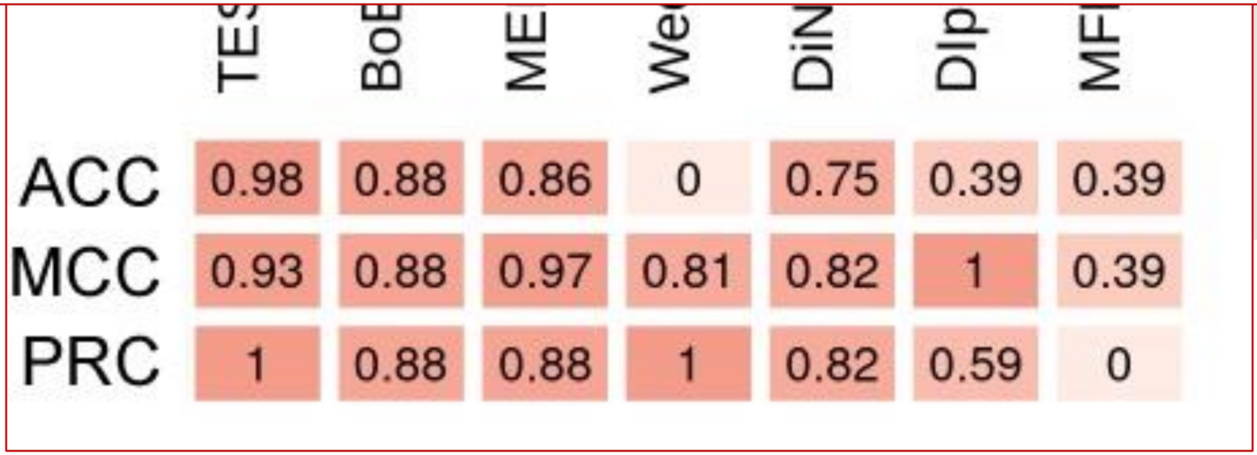
- Utilization of standard performance metrics: PRC, AUC, MCC
- 20 Escherichia coli ChIP-exo datasets downloaded from the proChIPdb database
- Comparison with known motifs from DPInteract, leveraging TOMTOM for similarity computation and significance assessment

pro ChIPdb

Tomtom
Motif Comparison Tool

Overall performance of TESA



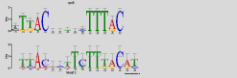
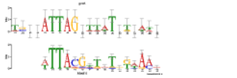
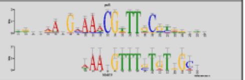
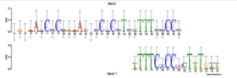
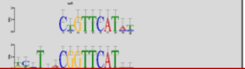


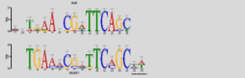
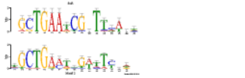
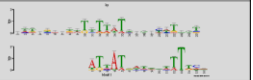
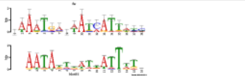
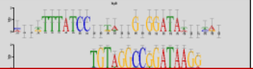
	TES	BoE	ME	We	DiN	DIp	MFI
ACC	0.98	0.88	0.86	0	0.75	0.39	0.39
MCC	0.93	0.88	0.97	0.81	0.82	1	0.39
PRC	1	0.88	0.88	1	0.82	0.59	0

Lollipop plot: Overall scores of each algorithm summed across 20 datasets

Spider plots: averaged scores of each algorithm

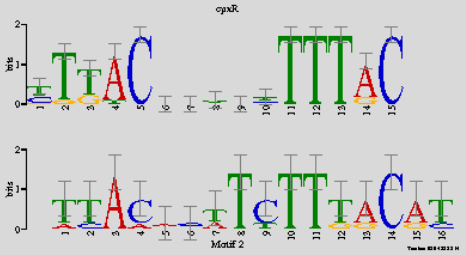
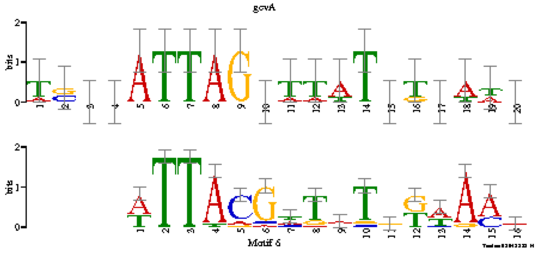
Similarity between target motifs and that discovered by TESA

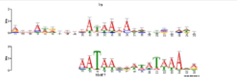
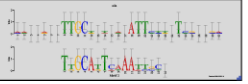
Dataset	Rank	TF	D	L1	L2	Q.value	Logo alignment
cpxr_EtOH	2	cpxR	-	16	15	2.61E-05	
cpxr_EtOH	6	gcvA	-	16	20	2.39E-02	
cpxr_EtOH	9	purR	+	16	26	5.65E-03	
flhc_LB	7	flhCD	-	16	31	2.82E-02	
flia_LB	9	torR	+	16	10	3.25E-04	

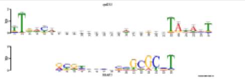
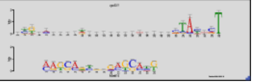
Dataset	Rank	TF	D	L1	L2	Q.value	Logo alignment
mixed-tfs-pool1a_M9+glu	1	fruR	-	16	16	3.93E-07	
mixed-tfs-pool1a_M9+glu	2	fruR	+	16	16	2.74E-06	
mixed-tfs-pool1c_M9+glu	1	lrp	+	16	25	1.55E-02	
mixed-tfs-pool1c_M9+glu	6	fur	-	16	18	3.68E-02	
mixed-tfs-pool2_M9+glu	1	hipB	-	16	30	4.93E-02	

Motif similarity:

- TESA discovers motifs with significant similarity with target motifs
- Some TFs are co-factors of the ChIP-ed one

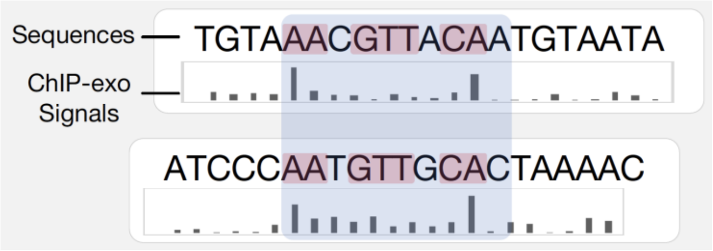
Dataset	Rank	TF	D	L1	L2	Q.value	Logo alignment
cpxr_EtOH	2	cpxR	-	16	15	2.61E-05	
cpxr_EtOH	6	gcvA	-	16	20	2.39E-02	

Dataset	Rank	TF	D	L1	L2	Q.value	Logo alignment
h-ns_expo	7	lrp	-	16	25	7.70E-03	
kdpe_KCl	2	ada	-	16	31	5.83E-03	

Dataset	Rank	TF	D	L1	L2	Q.value	Logo alignment
yiag_M9	1	rpoD15	+	16	27	7.45E-03	
yiag_M9	6	rpoS17	+	16	29	4.60E-02	

significance of motif alignment (the seventh column), and the eighth column showcases the alignment of motif logos.

Summary of TESA



TESA

A weighted two-stage sequence alignment framework to identify DNA motifs from ChIP-exo data

- Effective Graph Construction from base-resolution ChIP-exo
  - Optimal Motif Length Determination
  - High precision in motif identification

Limitations of TESA

- Computational resources needs
- Multiple parameter dependency
- Contemplation of alternative negative controls

Acknowledgement

Dr. Qin Ma

Dr. Bingqiang Liu

Yang Li

Yizhong Wang

Anjun Ma

BMBL

Bioinformatics and Mathematical Biosciences Lab

https://u.osu.edu/bmbl/

@Wang-Cankun
@QinMaBMBL

Cankun.Wang@osumc.edu

THANK YOU

Cankun.Wang@osumc.edu