

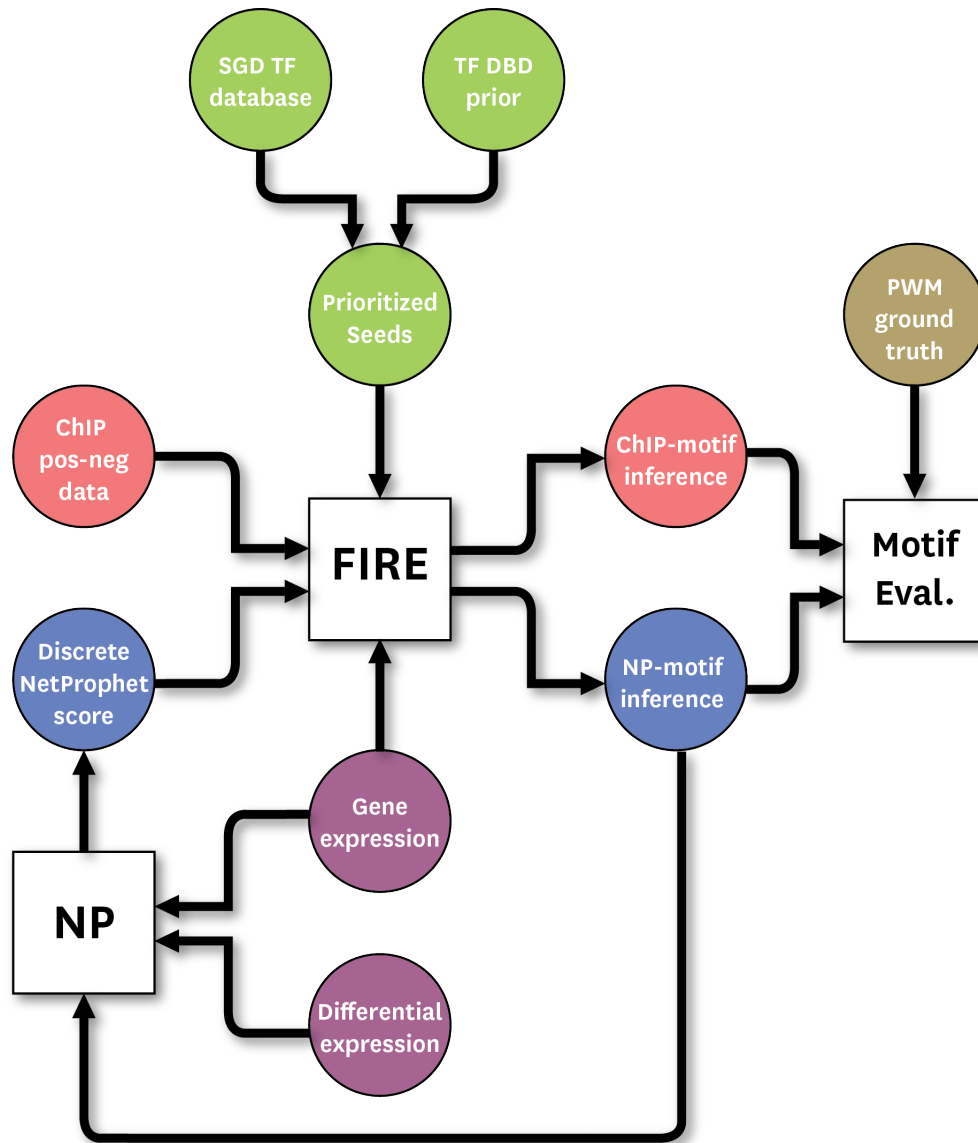
## **Project Overview**

As the gene regulatory network is extremely important in modeling biological systems nowadays, the goal of this project is to ultimately optimize an informative transcriptional network. Thanks to the state-of-the-art NetProphet algorithm<sup>[1]</sup>, we are already able to construct the yeast gene network, which is remarkably more accurate than those bench-work methods such as chromatin immunoprecipitation (ChIP) or protein binding microarray (PBM). In order to advance our novel algorithm onto the next level, incorporating the transcription factor (TF)-target interaction data into NetProphet's networking mapping algorithm is a promising computational approach.

Inferring highly informative TF motifs is the essential intermediate step to build the regulatory edges of all TFs to their target genes. The approach presented in this document, modifies existing motif inference program FIRE<sup>[2]</sup> by constraining the seeding procedure. The constraint allows the input seeds to possess only specific sequence pattern, according to the DNA binding domain (DBD) family that each TF belongs to. To our knowledge, FIRE among the prevalent motif inference programs, is the most feasible software utilizing the direct and functional binding data generated in NetProphet.

The work pipeline is illustrated in Figure 1. DBD family sequence prior is used to modify FIRE k-mer seeds. More specifically, SGD TF database that contains DNA binding domain family is compiled to give information of individual TF; and TF DBD prior is manually curated from literature and online resource; then combination of both datasets yields the prioritized seeds for each protein family. Accordingly, those seeds of k-mer sizes from 5 to 8 are applied to FIRE inference respectively. Meanwhile, FIRE utilizes ChIP positive-negative data and discretized NetProphet data independently for mutual information computation and optimization. NetProphet then constructs a new regulatory network, by using motifs discovered in FIRE using the discrete NetProphet scores, in addition to coexpression from discrete cluster of gene expression data, and difference expression from gene perturbation. And the new network is fed back to FIRE. This process is repeated between motif inference and network mapping algorithms. Ultimately, we expect this pipeline to demonstrate its ability to output the best network and motif.

Both collections of inferred motifs independently using ChIP and NetProphet data are aligned to and evaluated against the ground truth PWMs, which are primarily curated in ScerTF database<sup>[3]</sup> and complementarily in YeTFaSCo database<sup>[4]</sup>. The motif evaluation calculates 2 scores; each is a combination of 3 FIRE scores and 3 TOMTOM alignment scores respectively. Each evaluation score of using DBD family prior is compared with the one without using constraint. Then a compound score of the 2 comparison scores determines whether applying protein family prior information improves the motif inference for each TF.



**Figure 1.** Visualization of the project pipeline. Circles denote the data produced either experimentally or computationally; and squares denote the major algorithms.

### TF DNA Binding Domain Family

Eukaryotic transcription factors in the same family have some common ancestor in evolution, thus they share similar protein structures. The structural pattern results in similarity of TF function, allowing the protein to bind to specific genetic domain. The TFs in most protein families have one binding domain, but there are cases such as Zinc finger domain that have more than one domain, annotated as homodimer or heterodimer. Moreover, there can be formation of multi-protein complex, which accelerates transcription significantly.

The consensus sequence or core sequence of the binding motifs for each protein family denotes the sequence similarity. TF DNA binding domain therefore facilitates the discovery of each TF's sequence specificity, or motif. In this study, there are 11 manually curated TF-DNA binding domain families, which have the distinctive consensus sequences. The curation is based on literatures from Stromo's book [5], Hahn and Young's Yeastbook [6], and other online resources (e.g. Pfam and InterPro). All 11 families and their domain consensus sequences are shown in Table 1. Those core sequences are then utilized to constrain the FIRE k-mer seeds. Those priors essentially prioritize the seeds containing the highly informative binding pattern in the initial mutual information ranking, by eliminating other non-informative k-mers.




























Family	Consensus Sequence
BZIP (basic leucine zipper) domain [7]	GTCAT, GCAAT, GTAAT, GTTAC
FKH (forkhead) domain [8], [9]	[GA][TC]AAA[TC]A, GA[TC]GC
GATA zinc finger [10]	GAT[AC]
BHLH (basic helix-loop-helix) domain [11], [12]	CANNTG, N[GA]CGTG, CACG[AC]G
Homeobox domain [13], [14]	ATT, ATAAAA, TCGTAAA
HSF (heat shock factor) [15]	AGAA
Mlu-1 box domain [16], [17]	ACGCG
Myb domain [18], [19]	AAC[TG]G, G[GT]T[GA], CTCAGCG
SRF (serum response factor) domain [20], [21]	CC([AT] <sub>6</sub> )GG
Zinc finger, C2H2 (Cys2His2) [22], [23]	[CG]-rich
Zinc cluster, Zn2-Cys6 [24], [25], [26]	CGG, CCG (monomer/dimer)

**Table 1.** TF-binding domain families and their consensus sequences.

## Motif Inference Using ChIP Data

Binary ChIP data for individual TF as being listed in SGD TF database are compiled and processed. The binary data denote whether each TF binds to its gene target. All TFs in each DBD family are processed in FIRE using the prioritized seeds. A detailed case study is done on GATA family, in which a range of k-mer sizes ( $k = [5, 8]$ ) is applied for motif inference, shown in Table 2. The result shows  $k = 7$  and  $8$  are the most feasible sizes that facilitates discovery of most amount of binding specificities. In gapped motif discovery, even number of  $k$  is used, allowing a dimer seed to split into even halves. Thus, seed  $k = 6$  substitutes for  $k = 7$  in dimmer inference. If multiple motifs are inferred in FIRE, the top-ranked one containing the highest combined score of mutual information, z-score and robustness is picked as the optimal inference.

GATA family is one of the most representative DBD families. The motif inference constrained by GATA sequence prior are curated in Table 2. Longer k-mers guarantee more motif recoveries than shorter ones do. This is due to the mutual information computed for shorter seeds are less likely to be significant enough for motif optimization. Another reason is that shorter motifs may not be robust to pass the robustness evaluation, a statistical significance test using jack-knife resampling. Transcription factor YCR018C and YFLO21W are the failure cases in robustness test. YIRO13C and YPLO21W (in light grey background) have no motif inference, as they do not bind to any gene target in CHIP experiment.

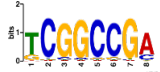







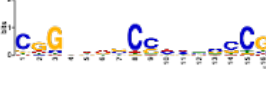





TF Name	ScerTF PWM	5-mer	6-mer	7-mer	8-mer
YER040W		N/A			
YJL110C					
YKR034W		N/A			
YLR013W					
YKL185W		N/A			
YCR018C		N/A	N/A	N/A	N/A
YFLO21W		N/A	N/A	N/A	
YIRO13C		N/A	N/A	N/A	N/A
YPLO21W		N/A	N/A	N/A	N/A

**Table 2.** The logos of inferred GATA motifs using binary ChIP data and various k-mers ( $k = [5, 8]$ ), and the logos of ScerTF PWMs. N/A denotes no motif inference. Note that some motifs are presented as the reverse complements of the PWMs.

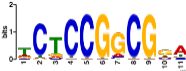

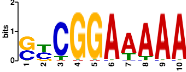

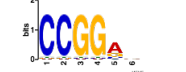


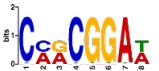
Zinc cluster is a protein family whose TFs bind to targets in form of dimers, as being emphasized in literature. The gap size of those special TFs is an additional parameter used for inferring their dimers, as shown in Table 3(a). The seeds are evenly split into a pair of monomers by definition. For instance, 6-mer CGGCCG is split and form new gapped dimer  $CGG(N_m)CCG$ , where  $m$  is the designated gap size. The sequence prior is applied to the three outer positions of each side of 8-mer seed. In motif optimization, the additional positions are only appended to the outer sides of gapped motif, thus allowing two inner positions to be random acts as adding one position to both sizes of each monomer. There is no 8-mer seed used for the dimer of gap size 0 or 1, since the gap size of 8-mer is

intentionally reduced by two base pairs. Some motifs in Zinc cluster do not have the dimer pattern, therefore sequence prior for monomers is also applied, as shown in Table 4(b).

(a)

TF Name	Gap size	ScerTF PWM	6-mer	8-mer
YCR106W	0			
YLR451W	3			
TDR421W	7			
YKLO15W	10			
YLR098C	13			

(b)

TF Name	ScerTF PWM	6-mer
YDR520C		
YKLO38W		
YMR280C		
YOR172W		

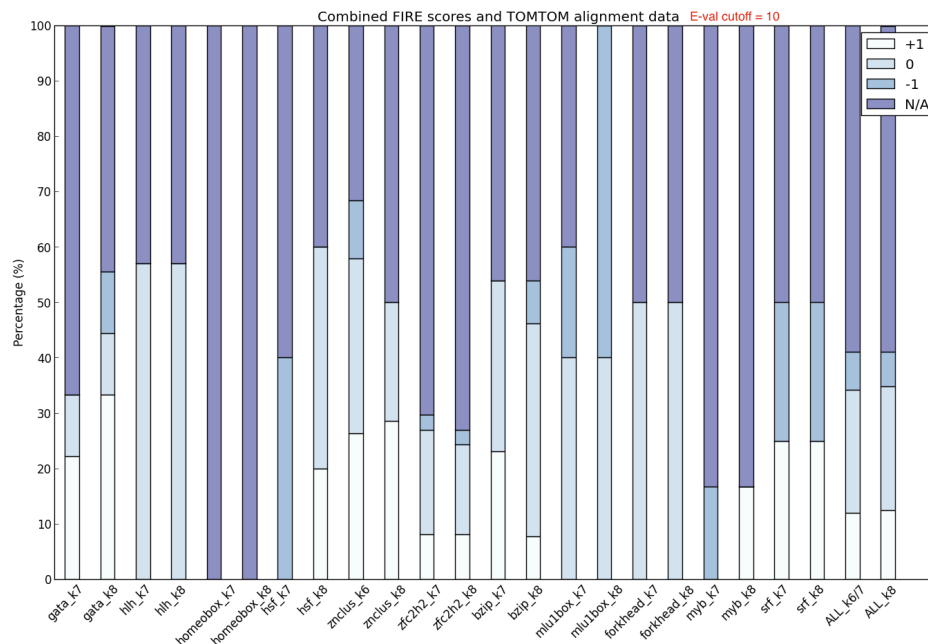
**Table 3.** (a) The logos of inferred dimer Zinc cluster motifs using binary ChIP data and k-mers (k=6, 8) of dimer prior, and the logos of ScerTF PWMs. (b) The logos of inferred monomer Zinc cluster motifs using binary ChIP data and k-mers of monomer prior, and the logos of ScerTF PWMs. Note that some motifs are reverse complements of their ScerTF PWMs.

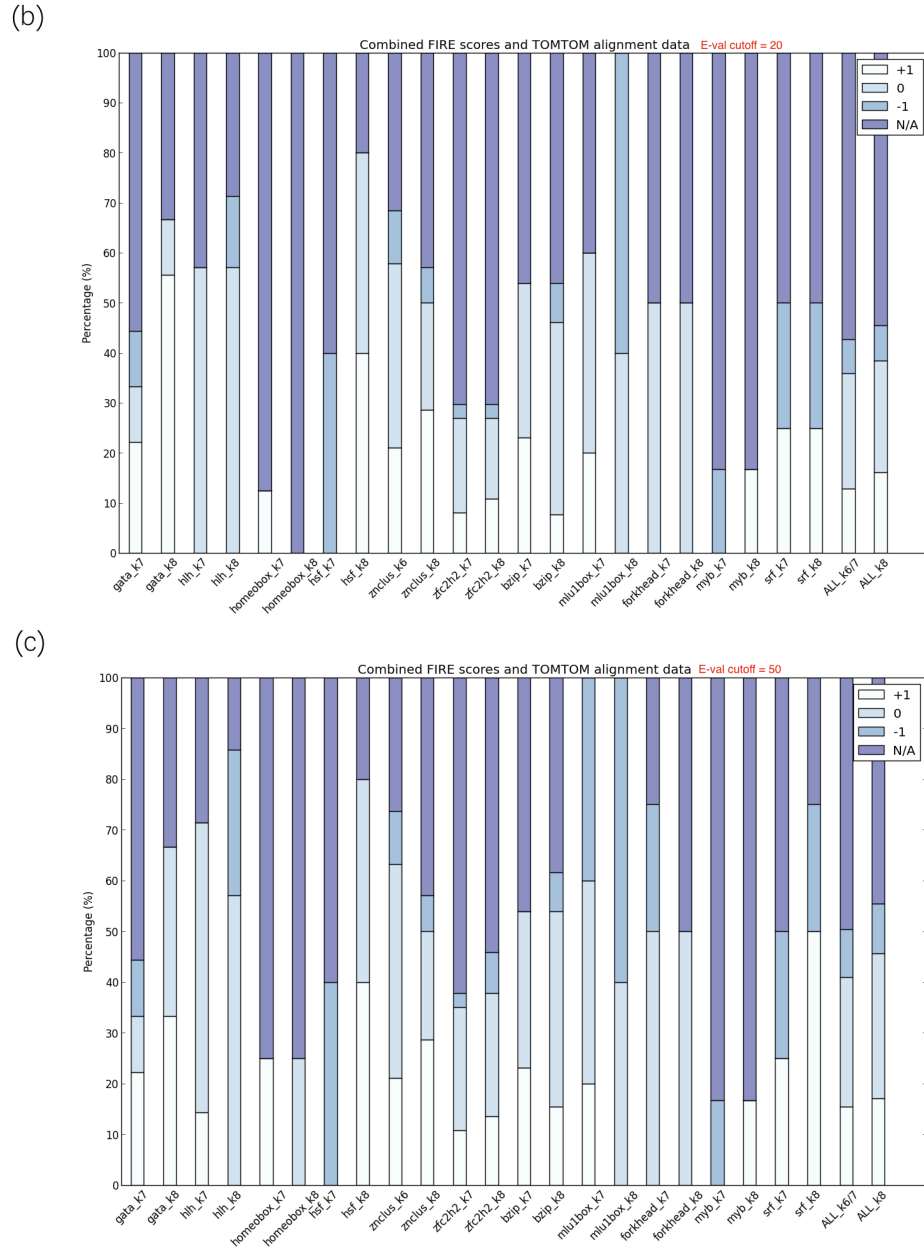
In order to infer motifs in Zinc cluster more effectively, the designated monomer or dimer information are hidden. Alternatively, k-mer seeds in both monomer form and in dimer form (gap size range = [0, 20] are fed into FIRE. For individual TF in this family, all motifs discovered using the 21 types of seeds (1 monomer and 20 dimers) are sorted by rank-sum test, based on 3 independent ranks of the mutual information, z-score, and robustness score. 12 of out 20 dimer motifs inferred are successfully ranked as the optimal ones, which are most informative and matching the gap sizes of PWMs. And 3 out of 19 monomer inferences are successful, as the dimers inferred are indeed much informative than the monomers. Some additional constraints may need to help recover the correct monomer and dimer forms.

An overall evaluation for all motif inferences in the 11 highlighted protein families is computed, as shown in Figure 2. This motif evolution calculates one score combining 3 FIRE scores and one score combining 3 TOMTOM alignment scores; and for each TF, it compares the motif inferred with sequence prior against the one without using prior, based on those 2 scores independently; then it combines those comparison results.

The final evaluation yields 4 possible results: +1 means using DBD family prior improves the motif inference; 0 means using DBD family prior does not change the result; -1 means it worsens the result; N/A means motifs are not comparable. The N/A matches one of those scenarios: no motif inferred for both prior and no-prior cases; no inferred motif can be aligned to the ground truth PWM; no PWM in curated in either ScerTF or YeTFaSCo databases.

(a)





**Figure 2.** Evaluation results of motifs in 11 DBD families inferred using binary ChIP data and the overall comparison results. (a), (b) and (c) present E-value cutoff in TOMTOM alignment at 10, 20 and 50 respectively. Sign +1 denotes improved motif, 0 denotes neutral case, -1 denotes worsened motif, and N/A denotes no motif alignment in TOMTOM or motif inference in FIRE.

The higher TOMTOM E-value cutoff, the higher chance that inferred motif could be aligned with its PWM. The portions of improved, neutral and worsened cases for the comparable motifs conserve through different E-value schemes. It's confident to state that around 83% of the comparable motifs are successfully inferred, and 31% of the comparable motifs are improved. On average, longer k-mers ( $k = 8$ ) help to recover slightly more motifs than shorter k-mers ( $k = 6$  or  $7$ ) do.

There is an argument concerning about the right E-value cutoff to allow proper sequence alignment. Some inferred motifs that can only be recovered in a higher cutoff are specified in Table 4. Those inferred motifs containing sequence prior have reasonable alignments to PWMs, which however have low bits of some core sequences, or do not possess any alleged prior pattern.

E-value 10 → 20			E-value 20 → 50		
Name	E-value	Alignment	Name	E-value	Alignment
YDL1o6C	12.635		YGLo96W	32.6893	
YGR249W	19.5887		YGR044C	43.2628	
YLR013W	10.2656		YDR463	27.6316	

**Table 4.** Special cases demonstrate that motifs can only be aligned in a higher TOMTOM E-value.

### Motif Inference Using NetProphet Scores

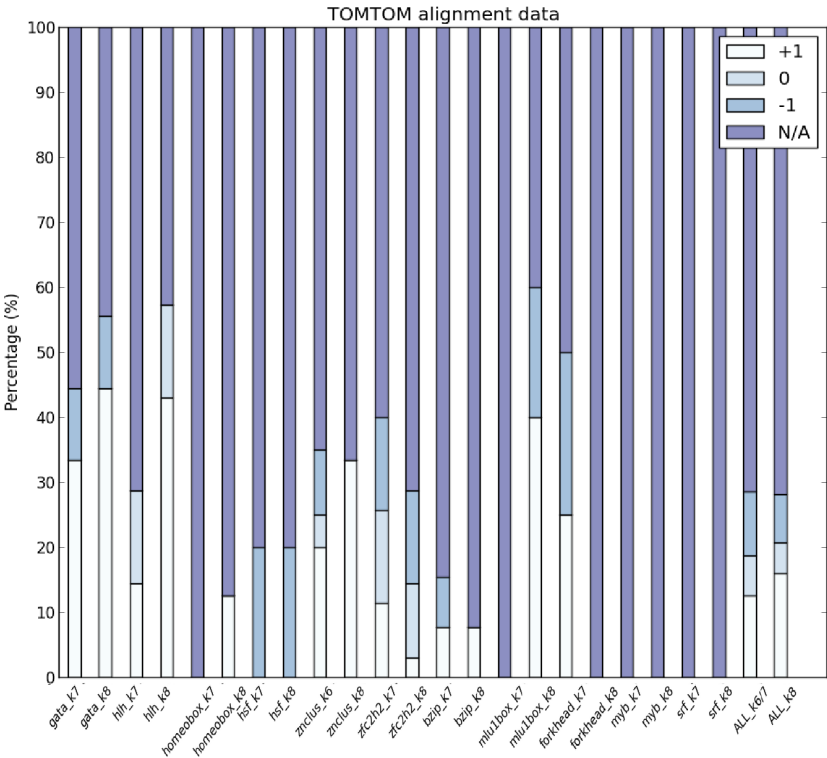
NetProphet (NP) scores in continuous form denote the direct and functional binding of individual TF to its gene targets. NP scores need to be discretized, as the quantization of continuous expression profiles in FIRE performs poorly. Therefore, two quantization strategies are applied and compared: equal-populated binning vs. equal-score-interval binning. Equal-populated binning splits the continuous scores into even number of data in each bin. Equal-score-interval binning clusters the scores into bins, which have the score ranges of equal span, along with a special bin that contains all zero scores. Bin number of 5, 10, 15 and 20 are used for both quantization schemes.

The evaluation of those binning schemes applied to GATA family suggests that equal-score-interval binning outperforms equal-populated binning, as it facilitates to infer more informative motifs. Thus the former using bin number of 5, 10 and 20 is considered a better quantization approach to handle NP scores. The results evaluated at TOMTOM alignment E-value cutoff of 20 are shown in Figure 3.

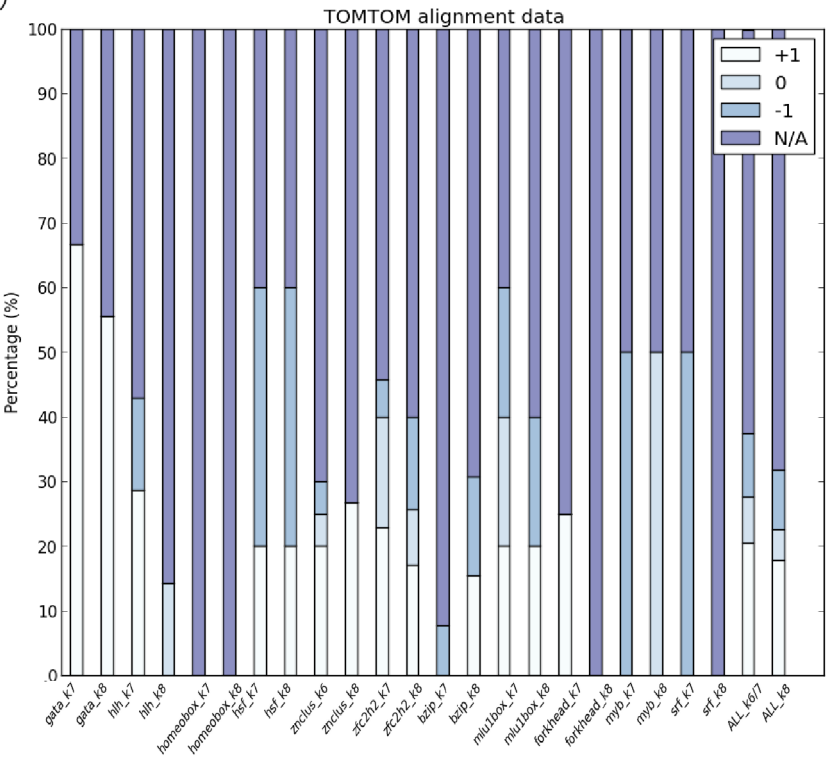


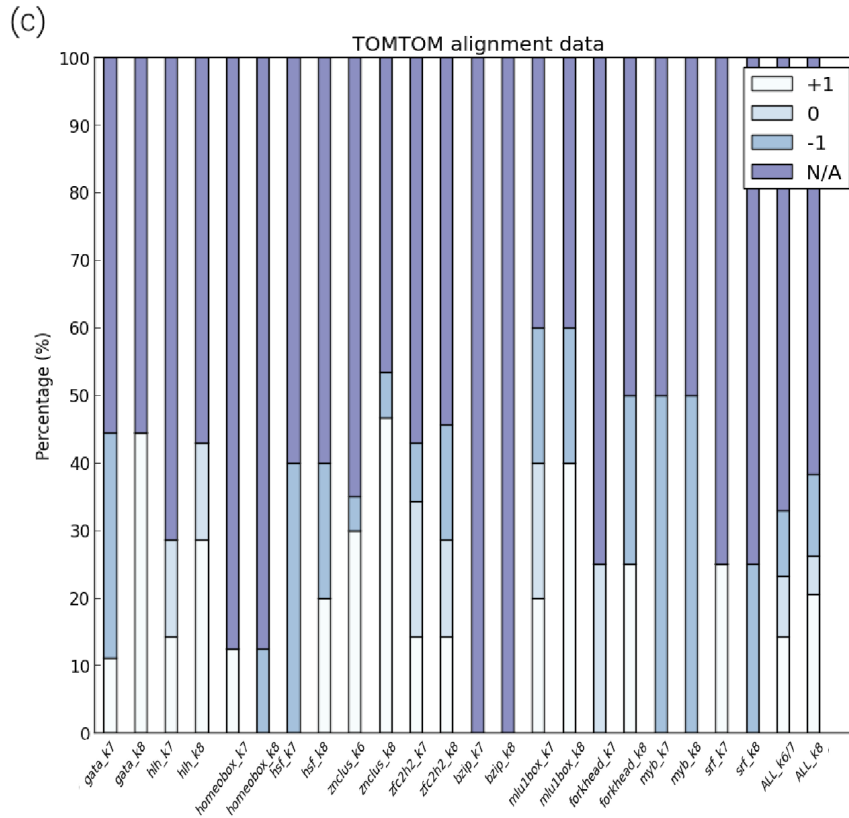
According to previous case studies in Table 4, E-value 20 is a feasible cutoff to show the quality of motif inferences.

(a)



(b)



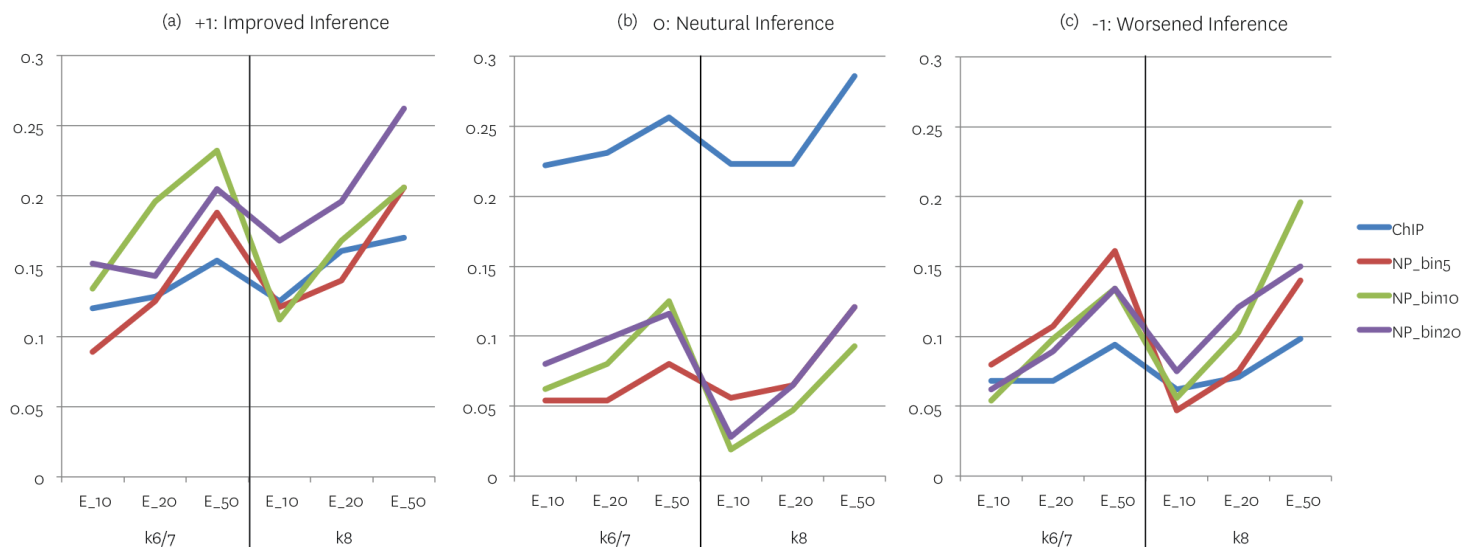


**Figure 3.** TOMTOM evaluation results of motifs in 11 DBD families inferred using discrete NP scores. The TOMTOM E-value cutoff set at 20. (a), (b) and (c) present discretization bin number of 5, 10 and 20 respectively. Sign +1 denotes improved motif, 0 denotes neutral case, -1 denotes worsen motif, and N/A denotes no motif alignment in TOMTOM or motif inference in FIRE.

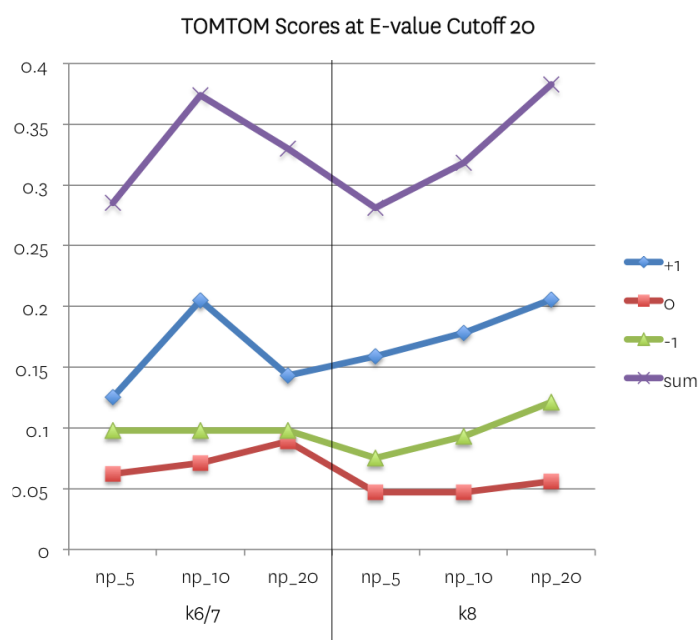
The overall evaluations of the entire motif inferences using both binary ChIP data and quantized NetProphet scores at certain evaluation cutoffs and k-mers are accumulated and re-evaluated across different output sets, as shown in Figure 4.

According to Figure 4(a), motif inference using NP scores shows more improved cases compared to the one using ChIP data. Bin-10 quantization on NP data outperforms other bin types if using shorter k-mers, while bin-20 quantization outperforms other bin types if using longer k-mers. Regardless of the seed's size, bin-20 wins over bin-10 in 4 out 6 inference cases. Figure 4(b) indicates that using ChIP data yields neutral inferences approximately 3 times of using NP scores do. The difference between bin-10 and bin-20 strategies in neutral case is not distinguishable. Figure 4(c) denotes that using ChIP data also introduce the least among of the false-positive evaluation errors. Like neutral results, it is difficult to tell the difference between the amount of errors applying bin-10 and that applying bin-20. Those errors occur primarily due to the fact that the DBD family sequence prior constrains motif inference to specific pattern, while PWM does not have the sequence prior.

To conclude, inference using quantized NP scores in 20 bins is the most feasible solution for now.



**Figure 4.** Overall motif inference evaluation using ChIP and NP data, under various constraints. (a), (b) and (c) illustrate improved, neutral and worsened cases respectively. In each subplot, left data points denote shorter k-mers ( $k=6$  or  $7$ ), and right ones denote longer k-mers ( $k=8$ ). TOMTOM E-value cutoff set at 10, 20 and 50.



**Figure 5.** TOMTOM motif evaluation using quantized NetProphet scores at E-value cutoff 20. Sign +1 denotes improved motif, 0 denotes neutral case, -1 denotes worsened motif, and the sum is a summary of the previous three signs.

## Discussion and Prospective

Improving the gene regulatory network mapping is the ultimate goal of next generation NetProphet algorithm. Applying those inferred motifs using quantized scores back into the network mapping scheme will potentially play an important role, contributing to score the TF-target direct binding profiles. As illustrated in Figure 1, the loop between network mapping and motif inference proceeds till no better regulatory network can be built.

Regarding improving the inference evaluation, it will be necessary to adopt the comprehensive PWM databases instead of the optimal PWM ones. To do so allows biological variance, the existence of multiple binding specificities of each TF to targets; consequently it diminishes the bias introduced from using only one PWM per TF, thus reducing the alignment failure cases. This comprehensive set of ground truth PWMs will curate all ScerTF data as the primary source, and YeTFaSCo data as the complement.

Building subdomains for every DBD family and computing their core sequence independently is a direction to pursue for more informative seeds with sequence priors. In the case study stated previously, some PWMs are the outliers since they possess no such alleged prior in literatures. To reduce those motif outliers, additional core sequence may be learned from clusters of PWMs, and each TF may be classified into a more specific subdomain. Consequently those subdomains will provide more feasible motif seeds, and facilitate to yield higher mutual information in optimization, thus output more confident motifs.

Resolving the gapped motif issue suggests additional improvements to be made. In Zinc cluster family, there are a number of gapped dimers and monomers inferred that fail to possess the expected multimer pattern or gap size. One idea to make the expected motif to be the most informative one is to add an extral score based on the distance from transcription starting site (TSS) to the TF DNA binding site (DBS). This score falls in an exponential decay function of the TSS-DBS distance. And a weighted average is calculated on the distance score and the three existing FIRE scores. Another approach to effectively discover the true monomer or dimer property is to incorporate the annotated protein-protein interactions (PPI) <sup>[27], [28]</sup>, which specify the likelihood of two proteins binding closely, in distance or having no synergy. Currently CYGD is the most comprehensive database of yeast PPI.

## References

- [1] Haynes BC, Maier EJ, Kramer MH, Wang PI, Brown H, Brent MR. Mapping functional transcription factor networks from gene expression data. *Genome Res.* 2013 Aug;23(8):1319-28.
- [2] Elemento O, Slonim N, Tavazoie S. A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell.* 2007 Oct 26;28(2):337-50.
- [3] Spivak AT, Stormo GD. ScerTF: a comprehensive database of benchmarked position weight matrices for *Saccharomyces* species. *Nucleic Acids Res.* 2012 Jan;40(Database issue):D162-8. doi: 10.1093/nar/gkr1180.

[4] de Boer CG, Hughes TR. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.* 2012 Jan;40(Database issue):D169-79. doi: 10.1093/nar/gkr993.

[5] Gary D. Stormo, Introduction to Protein-DNA Interactions: Structure, Thermodynamics, and Bioinformatics. Cold Spring Harbor Laboratory Press. 2013

[6] Hahn S, Young ET. Transcriptional regulation in *Saccharomyces cerevisiae*: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics.* 2011 Nov;189(3):705-36. doi: 10.1534/genetics.111.127019.

[7] Haas NB, Cantwell CA, Johnson PF, Burch JB. DNA-binding specificity of the PAR basic leucine zipper protein VBP partially overlaps those of the C/EBP and CREB/ATF families and is influenced by domains that flank the core basic region. *Mol Cell Biol.* 1995 Apr;15(4):1923-32.

[8] Pierrou S, Hellqvist M, Samuelsson L, Enerbäck S, Carlsson P. Cloning and characterization of seven human forkhead proteins: binding site specificity and DNA bending. *EMBO J.* 1994 Oct 17;13(20):5002-12.

[9] Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML. DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proc Natl Acad Sci U S A.* 2013 Jul 23;110(30):12349-54. doi: 10.1073/pnas.1310430110.

[10] Ko LJ, Engel JD. DNA-binding specificities of the GATA transcription factor family. *Mol Cell Biol.* 1993 Jul;13(7):4011-22.

[11] Ma PC, Rould MA, Weintraub H, Pabo CO. Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell.* 1994 May 6;77(3):451-9.

[12] Longo A, Guanga GP, Rose RB. Crystal structure of E47-NeuroD1/beta2 bHLH domain-DNA complex: heterodimer selectivity and DNA recognition. *Biochemistry.* 2008 Jan 8;47(1):218-29. Epub 2007 Dec 11.

[13] Kissinger CR, Liu BS, Martin-Blanco E, Kornberg TB, Pabo CO. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell.* 1990 Nov 2;63(3):579-90.

[14] Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, Khalid F, Zhang W, Newburger D, Jaeger SA, Morris QD, Bulyk ML, Hughes TR. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell.* 2008 Jun 27;133(7):1266-76. doi: 10.1016/j.cell.2008.05.024.

[15] Vuister GW, Kim SJ, Orosz A, Marquardt J, Wu C, Bax A. Solution structure of the DNA-binding domain of *Drosophila* heat shock transcription factor. *Nat Struct Biol.* 1994 Sep;1(9):605-14.

[16] Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science.* 1993 Sep 17;261(5128):1551-7.

[17] Bean JM, Siggia ED, Cross FR. High functional overlap between MluI cell-cycle box binding factor and Swi4/6 cell-cycle box binding factor in the G1/S transcriptional program in *Saccharomyces cerevisiae*. *Genetics.* 2005 Sep;171(1):49-61.

[18] Biedenkapp H, Borgmeyer U, Sippel AE, Klempnauer KH. Viral myb oncogene encodes a sequence-specific DNA-binding activity. *Nature.* 1988 Oct 27;335(6193):835-7.

[19] Romero I, Fuertes A, Benito MJ, Malpica JM, Leyva A, Paz-Ares J. More than 80R2R3-MYB regulatory genes in the genome of *Arabidopsis thaliana*. *Plant J.* 1998 May;14(3):273-84.

- [20] Pollock R, Treisman R. Human SRF-related proteins: DNA-binding properties and potential regulatory targets. *Genes Dev.* 1991 Dec;5(12A):2327-41.
- [21] Huang K, Louis JM, Donaldson L, Lim FL, Sharrocks AD, Clore GM. Solution structure of the MEF2A-DNA complex: structural basis for the modulation of DNA bending and specificity by MADS-box transcription factors. *EMBO J.* 2000 Jun 1;19(11):2615-28.
- [22] Krishna SS, Majumdar I, Grishin NV. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* 2003 Jan 15;31(2):532-50.
- [23] Lam KN, van Bakel H, Cote AG, van der Ven A, Hughes TR. Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Res.* 2011 Jun;39(11):4680-90.
- [24] Reece RJ, Ptashne M. Determinants of binding-site specificity among yeast C6 zinc cluster proteins. *Science.* 1993 Aug 13;261(5123):909-11.
- [25] Hellauer K, Rochon MH, Turcotte B. A novel DNA binding motif for yeast zinc cluster proteins: the Leu3p and Pdr3p transcriptional activators recognize everted repeats. *Mol Cell Biol.* 1996 Nov;16(11):6096-102.
- [26] MacPherson S, Larochelle M, Turcotte B. A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiol Mol Biol Rev.* 2006 Sep;70(3):583-604.
- [27] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature.* 2000 Feb 10;403(6770):623-7.
- [28] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A.* 2001 Apr 10;98(8):4569-74.